UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Physics of Data

Final Dissertation

Anomaly detection on trigger-less muon data streams

Thesis supervisor:

Prof. Marco Zanetti

Candidate:

Nicolò Lai

Academic Year 2022/2023

Physics is becoming so unbelievably complex that it is taking longer and longer to train a physicist. It is taking so long, in fact, to train a physicist to the place where he understands the nature of physical problems that he is already too old to solve them.

Eugene Wigner

Every great and deep difficulty bears in itself its own solution. It forces us to change our thinking in order to find it.

Niels Bohr

What we observe is not nature itself, but nature exposed to our method of questioning.

Werner Heisenberg

The most profound discoveries are not necessarily beyond the next mountain range, but often within the range of our own minds.

E. T. Bell

We are in a position similar to that of a mountaineer who is wandering over uncharted spaces, and never knows whether behind the peak which he sees in front of him and which he tries to scale there may not be another peak still beyond and higher up.

Wolfgang Pauli

Abstract

This thesis investigates the challenges and limitations in detecting New Physics phenomena in particle collider experiments, with a focus on the CMS experiment at the LHC. The research evaluates the role of trigger systems in potentially introducing bias by filtering data based on established theories, thereby limiting the discovery potential for New Physics. To address this, we employ the CMS data scouting method for more unbiased data extraction from various trigger chain levels. identifies limitations in traditional statistical analyses, which typically rely on predefined theoretical models to be tested against the Standard Model predictions. To counteract these issues, we introduce the New Physics Learning Machine (NPLM), a machine learning framework designed for model-independent anomaly detection. We validate NPLM by integrating it with unfiltered muon data, facilitated by prototypes of the front-end electronics for the CMS Drift Tubes muon chambers, expected to be operational during the High Luminosity LHC phase. An experimental demonstrator was constructed using a small-scale replica of the CMS muon drift tubes, equipped with on-detector board prototypes for unfiltered muon data extraction. In this setup, we combined data scouting and (quasi-)online anomaly detection using the NPLM. Preliminary results indicate that this integrated approach is promising, particularly for the forthcoming High Luminosity LHC phase Online data quality monitoring experiments were conducted on the unfiltered muon streams. Utilizing the model-independent and multivariate nature of the NPLM algorithm, we achieved significant improvements over standard machine learning techniques commonly used for such tasks. Quantitative evaluations include metrics on pipeline execution time and scalability, achieved through parallelized GPU computing and FPGAs. The analysis results are reported with respect to their sensitivity to anomalies in the data and their scalability for high-throughput scenarios. The results indicate applicability in both online data quality monitoring and quasi-online New Physics searches, especially relevant for the upcoming High Luminosity LHC phase. As the CMS detector readies for this next phase, the thesis lays the groundwork for enhanced data analysis methods by integrating data scouting and machine learning-based anomaly detection. Leveraging heterogeneous computing resources, the work contributes to more efficient and unbiased data analysis, holding potential for advancing both New Physics searches in the High Luminosity phase.

Acknowledgments

I wish to begin by expressing my deep appreciation to my family for their unwavering support and understanding throughout the course of my academic endeavors. Their comprehension of the difficulties associated with maintaining a balance between academic pursuits and personal life has served as a consistent source of motivation and fortitude.

I would like to acknowledge Prof. Marco Zanetti for his exemplary role as a supervisor, advisor, and mentor over the past two years. His insightful conversations have significantly contributed to my understanding of the scientific world, thereby clarifying my academic pathway. I look forward to starting my Ph.D. journey under his expert guidance. Furthermore, I extend a note of gratitude to the entire BoostLab group for crafting a work environment that manages to make even the most routine meetings somewhat bearable. I am pleased to be a part of the team.

My most sincere gratitude goes to Dr. G. Grosso and Dr. M. Letizia for their instrumental roles in shaping my research methodology and imparting practical expertise during our two-year collaboration, which has culminated in this thesis. The experience has provided me with an invaluable educational opportunity that stands as a highlight in my academic journey.

To Giacomo, Alberto, and Pietro: CittaRomanze. My most fulfilling academic experiences have been shared with you. The summer project sessions, accompanied by our unique selection of music, remain unforgettable. Not to mention, our sticker collection stands as perhaps the finest ever assembled by humankind. *Stochasticity.* Additional acknowledgments go to Andrea and Tommaso. The contributions of each individual have significantly enriched this academic journey in ways that are difficult to encapsulate in words. The experiences and memories formed will continue to be a lasting part of my personal and professional life.

Reflecting on the past summer, I extend my profound appreciation to the exceptional group of friends I met during the CERN summer school. In what was undoubtedly a demanding work period, your fellowship transformed it into an enjoyable experience. Special acknowledgments are directed to my supervisors, Dr. T. James and Rocco, who focused on fostering my learning experience rather than on the results I produced during these hectic months. Manz couldn't have scored better supervisors, BMT bruv. I also wish to recognize Michele, a friend and fellow summer student at CERN, for shared experiences and challenges. Your steadfast support during the late-night thesis-writing sessions in Building 40 has been indispensable. The countless cigarette breaks and those unforgettable bad dinners at R1 definitely made the summer a lot more fun and memorable. As for Matteo, my office mate—your presence, perseverance, and resolve have served as vital sources of motivation throughout this journey. Your straightforward manner and the practical insights you have shared have been invaluable to my learning process.

A shout-out to my colleagues from my Bachelor's and Master's programs. Those casual conversations and coffee breaks were often the perfect antidote to the stresses of academia, making tough times more bearable. Special nods go to Marcello and Davide; while I cannot spell out all my thoughts here without turning some heads, know that I am deeply thankful.

While it is not feasible to individually acknowledge each friend or colleague, it should be understood that every shared experience and contribution to my academic journey is highly valued and deeply appreciated. Each of you has served as a fundamental pillar in this academic journey, and for that, I extend my deepest sense of gratitude.

> If I have seen further, it is by standing
> on the shoulders of giants.
>
> *Sir Isaac Newton*

Contents

Chapter 1

Introduction

In the vast canvas of our universe, particle physics remains a complex domain that challenges even the most advanced scientific inquiries. The celebrated Standard Model has stood the test of time and has been the cornerstone of our understanding of high-energy physics for decades. It is a testament to our most advanced comprehension, confirmed with exceptional precision through rigorous experimental verifications. The Standard Model has proven effective in numerous applications, yet it remains limited in its comprehension of certain phenomena. Areas of uncertainty include the source of the electroweak scale, neutrino masses, and the flavor structure of quarks and leptons. Additionally, the model fails to account for enigmatic enigmas such as dark matter. The unresolved questions in particle physics point us towards unexplored areas, often categorized under 'New Physics'. While not yet fully understood, this domain might provide explanations that extend our current scientific understanding. It offers the potential for new perspectives on the foundational principles of the universe.

In high-energy physics, the Large Hadron Collider (LHC) stands as a significant achievement. Its capabilities are exemplified by its capacity to record proton-proton collisions at an astounding frequency of every 25 ns. This produces a substantial amount of data: every interaction, captured by an extensive array of sensors, results in about 1.5 MB of information for detailed encoding. When faced with a large amount of data, the scientific community could have been overwhelmed. Leveraging our foundational knowledge of particle physics, primarily rooted in the Standard Model, is the solution to this challenge. By building upon these established theoretical understandings, researchers have managed to efficiently process and analyze the LHC data. Yet, the same foundational principles that facilitated the management of LHC's massive data throughput and culminated in the monumental discovery of the Higgs boson in 2012 have also imposed limits on advancing our knowledge further. While detecting the Higgs boson solidified the Standard Model's robustness, it raised a perplexing issue. Supersymmetry emerged as a leading candidate to go beyond the Standard Model after discovering the Higgs boson. Still today, it remains among numerous theoretical frameworks attempting to expand our fundamental understanding. Despite using analysis techniques analogous to those that uncovered the Higgs boson, these theories have undergone exhaustive testing, and as yet, no compelling evidence has emerged. The problem is complex, with endless theories and no clear way to prioritize or confirm the correct one. Furthermore, even if such a theory emerges, the conventional datasets and methodologies might prove inadequate for its validation.

Confronted with these challenges, the physics community acknowledges the need for a paradigm shift in our quest for New Physics. Rather than pinning hopes on specific theories, there is a growing emphasis on *model-independent* approaches. Such methods do not exclusively pursue a particular 'Beyond the Standard Model' theory; instead, they scrutinize the data for any deviations from Standard Model predictions. Machine learning has heralded a new era in this domain, empowering researchers with tools for a signal-agnostic examination of the data. With the physics community adopting model-independent strategies aided by machine learning, the insufficiency of current datasets used for analysis becomes apparent. At the heart of the LHC's operational design is the online trigger system, an indispensable tool designed to manage the deluge of data by filtering events in real-time at a staggering rate of 40 MHz. The trigger ensures manageability by only retaining "interesting" events, but the definition of "interesting" is crucial. Historically, this criterion has been shaped by our understanding of particle physics, inherently based on the Standard Model. While the trigger system applies a selection that permits studies of Beyond the Standard Model theories, the inherent rarity of these events means they are often elusive and challenging to detect in practice.

Given the enormous throughput of data produced at the LHC, outright removing the trigger system is not an option, as reading out all the information is technically impossible. However, relying solely on triggered data for model-independent strategies is suboptimal, as it could potentially mask subtle deviations from the Standard Model. This situation suggests a need for an astute workaround that allows us to capture and analyze significant portions of data without being hindered by traditional trigger constraints. In response to these challenges, data scouting and online analysis studies have been developed to navigate the trigger system's intrinsic constraints and enhance our pursuit of New Physics. The exploration and implications of these strategies form the core of this thesis. Data scouting has been a long-standing practice in particle physics, especially in the context of LHC. Despite this method's complex and sophisticated nature, it has been widely employed for its valuable implications. The Compact Muon Solenoid (CMS) experiment stands out as a notable proponent of this technique, having utilized it for over a decade. Here, coarsely reconstructed events are prized due to their reduced memory footprint, enabling storing a larger number of events and amplifying our ability to scrutinize rarer processes.

Significant upgrades are required with the upcoming transition to the High Luminosity LHC (HL-LHC) era. The HL-LHC's objective to increase the data produced every 25 ns necessitates reevaluations and enhancements in the hardware, electronics, and triggering systems. Coinciding with these upgrades, there is an opportunity to refine data scouting techniques. Such refinements could enable data extraction that aligns more closely with the detector's immediate outputs, yielding unfiltered and unbiased datasets. These advancements, however, come with challenges. The near-raw data will generate a colossal throughput that necessitates an astutely crafted computing infrastructure and processing strategy. The prize? The capacity to deploy sophisticated, machine-learning-based model-independent anomaly detection on these unbiased data streams in real time. By doing so, the range and depth of searches for New Physics phenomena could be significantly expanded. The core of this thesis is to demonstrate the feasibility of this approach: employing data scouting combined with machine learning to probe data for deviations from our theoretical expectations.

A crucial component of this research is a small-scale version of the CMS muon drift tubes situated at the Legnaro National Laboratories in Padova, Italy. This configuration is advantageous because it offers a more manageable replica of the CMS muon drift tubes. Furthermore, when operated as a cosmic muon telescope, the muon interaction rate with the detector is considerably lower. This provides a more favorable environment for initial studies compared to the demanding conditions posed by the LHC's collision events. An integral aspect of this setup is the incorporation of front-end electronics prototypes, slated for installation at the CMS experiment during the High Luminosity phase. These advanced boards facilitate data scouting directly from the detector's front-end, providing a continuous 40 MHz data stream without any filtering. To validate the applicability of our setup in the context of collider experiments, we have developed a scalable processing and analysis pipeline utilizing heterogeneous computing and commercial big-data tools. This architecture incorporates an analysis algorithm specifically designed to operate on GPUs, ensuring efficient processing. As a result, the pipeline permits prompt assessments to determine the presence or absence of anomalies in specific data batches.

In our study, we have applied our data acquisition, processing, and analysis pipeline to the drift tube setup at the Legnaro National Laboratories, effectively simulating an online data quality monitoring system. It is important to note that monitoring data quality has similarities with searches for New Physics, as both aim to detect deviations from expected data patterns. Although the contexts in which they operate may differ, the foundational objective is consistent: to detect anomalies that diverge from expected distributions. This thesis elucidates these similarities while highlighting the distinct aspects, ensuring our proof of concept maintains its rigor and validity.

This thesis provides an extensive analysis of the demonstrator and its potential application to the CMS experiment in the search for New Physics. In Chap. 2, we establish the fundamental theoretical framework, focusing on the anomaly detection algorithm and detailing the methodology that characterizes this model-independent analysis approach. In Chap. 3, the discourse shifts to a focused analysis of the CMS experiment. This chapter explains the challenges of the CMS trigger system for investigating New Physics without a target model dependence. Furthermore, we give a comprehensive account of data scouting within the CMS context, charting its development from its foundational stages to its current state, with projections on its role after the High-Luminosity upgrades. Chapter 4 starts with a thorough examination of data quality monitoring in collider experiments. This provides a backdrop for our subsequent examination of the demonstrator at the Legnaro National Laboratories. The chapter provides a comprehensive description, covering the experimental setup design, online processing infrastructure adaptability, and presentation of our DQM anomaly detection technique. It concludes with a summary of our results and a discussion of their potential applicability and scalability.

Chapter 2

A model-independent strategy to search for New Physics

In our quest to understand the mysteries of the universe, we strictly adhere to the scientific method. This method operates on a foundation of confutation: a given null hypothesis stands as the preferred phenomenon description until evidence suggests an alternative hypothesis is more fitting. In particle physics, the Standard Model (SM) stands as a testament to this process. Its predictions, precise and consistent, have been validated by numerous experiments. Nevertheless, as with all scientific models, continuous scrutiny of the SM is essential.

Statistics provides a rigorous framework to quantify the likelihood of observations under set hypotheses. Nonetheless, the decision to replace an established model with a new one is rooted in the scientific community's accumulated wisdom, collective insight, and domain expertise. To assess whether some Beyond the Standard Model theory should replace the SM, we proceed via hypothesis testing:

1. We identify the null hypothesis representing our expectations about a specific phenomenon.
2. We choose a criterion to decide whether to reject the null hypothesis.
3. After collecting a data sample representing the tested phenomenon, the observations are compared to the expectations, and based on the predefined rejection criterion, we draw our conclusions.

Typically, the testing procedure relies on the definition of a test statistic. This random variable is a function of the observed data, and its distribution depends on the chosen hypothesis. A detailed examination of this test statistic emphasizes the complexities of deriving robust conclusions. Neyman and Pearson, pioneers in the field, identified two types of errors within the testing process: the Type I error, denoting the rate of rejection of the null hypothesis when it is true, and the Type II error, which highlights the rate of failures in rejecting the alternative when the null is wrong (see the left panel in Fig. 2.1). The delicate balance of minimizing both errors provides the backdrop for many hypothesis-testing strategies. Neyman and Pearson introduced the accepted standard practice of predefining a threshold for the type I error, known as the significance level of the test (denoted as α), and then searching for the optimal test statistic that minimizes the type II error. After defining a test statistic t, experimental observations are used to evaluate its value t_{obs} and compare it with the test statistic distribution given the assumptions of the null hypothesis. We define the p-value as the probability of observing a value of t as extreme or more extreme than the observed value, assuming the null hypothesis is true. The p-value is then compared to the predefined significant level α and, if lower, the null hypothesis is rejected (see the right panel in Fig. 2.1).

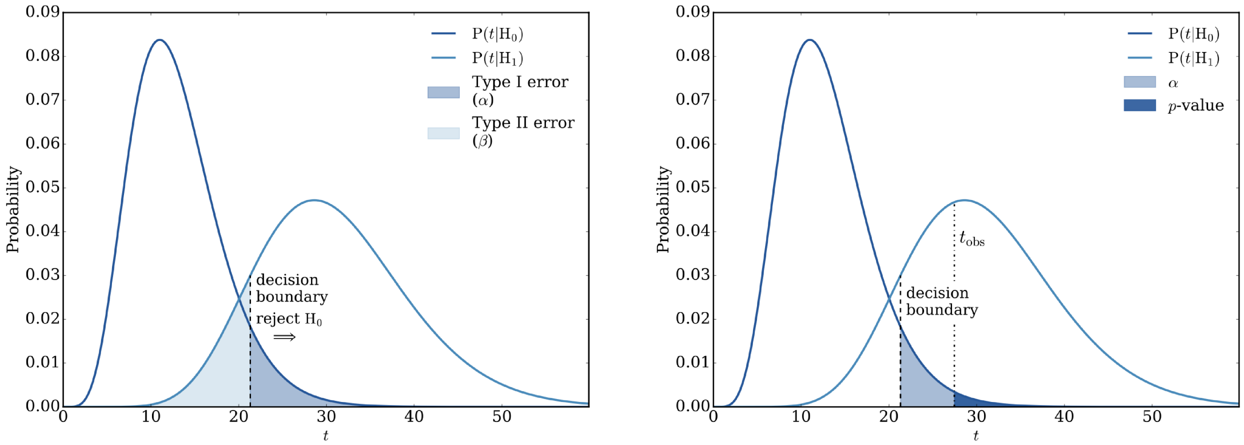


Figure 2.1: Illustration of the concepts of type I and II errors (left panel) and the concept of p-value (right panel). (Figures from [1]).

In 1933, Neyman and Pearson introduced the concept of the optimal test statistic [2]. They concluded that it is the ratio between the likelihood of the data under the null and true hypotheses. This is now known as the Neyman-Pearson lemma, valid for simple hypotheses where models do not depend on free parameters. However, in particle physics, models often become complicated with numerous parameters, some directly extracted from data.

Nevertheless, the likelihood of the data is the most powerful tool at our disposal to extract meaningful information about the underlying truth. The core of available information for statistical analysis is the experimental measurements. Any layer of summarization they are subjected to makes them lose some of the information they carry. To preserve potentially lost information, we need analysis techniques that can examine large sets of raw, unsummarized observations characterized by a high dimensional set of properties without relying on theoretical assumptions. Unbiased, model-independent searches have a trade-off of reduced precision when fewer assumptions are imposed. If the observed data contained complete knowledge about the physics process governing them, we could infer precise statements about the system. Instead, our unavoidable ignorance about the intrinsic nature of the data leads to less exclusive results.

There are two approaches to seeking optimality: a top-down approach that begins with a theoretical assumption of the true data model and constructs the optimal test statistic under that assumption and a bottom-up approach that analyzes the statistical behavior of the data to formulate the best hypothesis testing approach. Two different questions are being addressed by the two approaches, which means that two separate ideas of what is optimal are being pursued. The former approach assumes the theory to be true, and statistical fluctuations may distort the data. Therefore, the optimal result can only be achieved if the data adheres to the proposed hypothesis. This notion of optimality is the one defined by the Neyman-Pearson lemma, which states that, given a null hypothesis and a fixed significance level for its rejection, the test statistic based on the likelihood ratio produces the highest power when the model hypothesized by the alternative coincides with the true model of the data. The second approach focuses on experimental observations and views models as imperfect attempts to capture reality. Each model, or hypothesis, is "learned" from the data by maximizing the likelihood, and thus, it fluctuates around the true hypothesis. The bottom-up approach aims for inclusiveness by detecting broad families of alternatives in one attempt rather than guaranteeing optimality as the Neyman-Pearson lemma intended. The first approach relates optimality to a theoretical model, while the second relates to the experimental evidence provided by the collected data.

The construction of the New Physics Learning Machine (NPLM) algorithm follows the second approach, seeking to address the challenges of statistical analyses, especially when theoretical priors are not readily available for experiments. The primary objective of the NPLM is to compute the optimal test statistic while minimizing the assumptions about the actual hypothesis that might explain the observed event. The NPLM identifies discrepancies from a Reference model prediction across multiple dimensions without any predisposed bias about the potential signal causing the variation. Machine learning is exploited to build the log-likelihood-ratio hypothesis test, and determining the maximum likelihood is equivalently viewed as minimizing a specific loss function.

2.1 Conceptual foundations

Suppose we conduct an experiment and observe d properties, labeled $x_1, \ldots, x_d$, of a specific physical system. We can represent these outcomes using the vector $x = (x_1, \ldots, x_d)$, which belongs to the space $\mathcal{X} \subseteq \mathbb{R}^d$. The outcome x is thus a d-dimensional random variable following a probability distribution function whose true (T) form $p(x|\mathrm{T})$ is unknown. We assume we have a detailed understanding of the physics laws behind the observed phenomenon, which we term the 'Reference model'. Our goal is to see how closely this model aligns with our data.

We need a sample $\mathcal{D}$ made up of $\mathcal{N}_{\mathcal{D}}$ repeated observations from the physical system to draw statistical conclusions. We will work under the assumption that each observation in $\mathcal{D}$ is statistically independent and originates from the same true source.

In this thesis, we will consider the case of a perfectly known Reference model (R) in which the effects of systematic uncertainties are negligible. However, a rigorous treatment of the more realistic case in which R is affected by systematics for the approach that is about to be presented has been developed and tested to be effective [1,3]. If we denote a generic alternative hypothesis depending on some free parameters $\mathbf{w}$ as $\mathrm{H}_{\mathbf{w}}$, the log-likelihood-ratio hypothesis test that has maximum power given the dataset is the one for which the likelihood of the data under $\mathrm{H}_{\mathbf{w}}$ is maximized over the space of free parameters Ω:

$$t(\mathcal{D}) = \max_{\mathbf{w}\in\Omega}\left[\log\frac{\mathcal{L}(\mathcal{D}\,|\,\mathrm{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D}\,|\,\mathrm{R})}\right] . \tag{2.1}$$

2.1.1 Universal approximators to model data distributions

To compute $t(\mathcal{D})$ by making minimal assumptions about the nature of the alternative hypothesis, our primary concern is to build an alternative model so generic that no physics-motivated bias can be read into it. Our approach involves using universal functional approximations.

2.1.1.1 Histograms

The most commonly adopted approach to model data distributions uses piece-wise constant functions, known as histograms. Histograms are universal approximations whose accuracy increases with finer binning choices as long as the data resolution limits are not exceeded. Let the d-dimensional space $\mathcal{X}$ of the outcomes be divided into m bins. Denote the number of expected events in the i-th bin under the generic hypothesis H as $n_i(\mathrm{H})$ and the number of actual observed outcomes as o_i. The likelihood of the binned data is the product over the bins of the Poissonian probability of observing the counting o_i given the expected one $n_i(\mathrm{H})$:

$$\mathcal{L}(\mathcal{D}|\mathrm{H}) = \prod_{i=1}^{m} P_{\mathrm{pois}}(o_i;\, n_i(\mathrm{H})) = \prod_{i=1}^{m}\left(\frac{n_i(\mathrm{H})^{o_i} e^{-n_i(\mathrm{H})}}{o_i!}\right) = e^{-\mathrm{N(H)}}\prod_{i=1}^{m}\left(\frac{n_i(\mathrm{H})^{o_i}}{o_i!}\right) . \tag{2.2}$$

Here, N(H) represents the expected count summed over all bins. The model aligning most closely with the actual data distribution is the one where expected counts match the observed ones. This is often referred to as *saturated model*[1], and we will denote it as S. Hence, $n_i(\mathrm{S})$ would be equal to o_i and $\mathrm{N(S)} = \sum_{i=1}^{m} o_i = O_i$. Using the Reference hypothesis as the null and the saturated model as an alternative, we can build the log-likelihood ratio test statistic

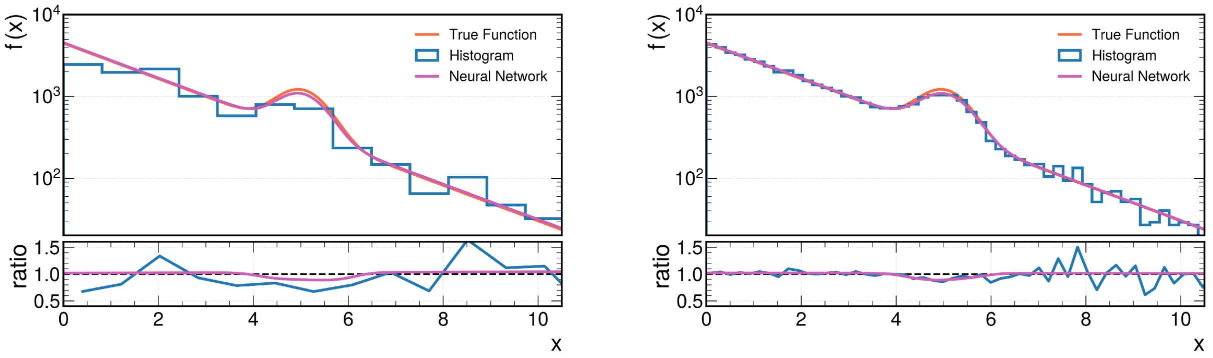
$$\bar{t}_{\text{bins}}(\mathcal{D}) = -2\log\frac{\mathcal{L}(\mathcal{D}|\mathrm{R})}{\mathcal{L}(\mathcal{D}|\mathrm{S})} = 2\left[\mathrm{N(R)} - O_i + \sum_{i=1}^{m} o_i \log\frac{o_i}{n_i(\mathrm{R})}\right]. \tag{2.3}$$

Using histograms for model-independent tests, however, presents some challenges. The first issue comes from the arbitrariness of the binning choice. Although the quality of the model approximation increases using a higher number of bins with reduced width, arbitrarily adding bins could reduce the power of the test statistic. We can intuitively understand this argument by studying the asymptotic properties of $\bar{t}_{\text{bins}}$ in the limit of large statistics samples (i.e., $o_i \gg 1\,\forall i$). In this case, the Poissonian probability distribution describing each bin can be approximated by a Gaussian distribution with mean $n_i(\mathrm{H})$ and standard deviation $\sqrt{n_i(\mathrm{H})}$. Under this condition, $\bar{t}_{\text{bins}}$ reduces to a sum of squared gaussian-distributed random variables and thus under the Reference distribution $\bar{t}_{\text{bins}}$ follows a χ^2 distribution with a number of degrees of freedom equal to the number of bins. Bins in which the observations are highly discrepant with the Reference expectations will contribute with a positive shift in the value of $\bar{t}_{\text{bins}}$ while bins that are in accordance with the expectations do not contribute significantly to the sum, as they would do under the null hypothesis. If the number of bins impacting the final value of $\bar{t}_{\text{bins}}$ is few with respect to the number of non-discrepant bins, then the discrepant effect could be shadowed and found not statistically significant. Removing bins that are in accordance with the expectations improves the sensitivity of the test. This is possible in model-dependent applications because the prior assumption about the signal nature allows us to restrict the region of the analysis down to the space where the signal shows up. For model-independent approaches, this is impossible unless some assumptions are made to identify potentially interesting regions, thus introducing a bias and reducing the test's inclusiveness level.

The second issue comes from the curse of dimensionality. By binning data, we are categorizing sample elements into clusters (usually hyper-boxes in multi-dimensional problems) and condensing the information carried by the elements in each subset in a single point, the *centroid* of the bin. The properties of the centroid are assigned to all the elements in the bin, losing the information on the position of each element within the bin and thus inducing a resolution uncertainty. To keep the resolution uncertainties under control, the bin width should not be much larger than the experimental resolution. In multi-dimensional problems, the number of bins scales with the power of the number of dimensions d. Binning the dataset becomes, therefore, computationally expensive as soon as the number of dimensions of the problem becomes greater than two or three. Furthermore, for the binning process to be meaningful, the bins should be actually populated with a reasonable amount of data. This demands a dataset whose size grows exponentially with the power of d. This is also practically impossible if the number of dimensions is higher than two or three.

In high-energy physics experiments, we have very large statistics samples at our disposal. Nevertheless, the number of observed properties is so large that the only way to carry out a binned analysis is to reduce the data to two or three relevant dimensions. To address this limitation, machine learning techniques can be used to accurately and efficiently approximate data distribution, even with multi-dimensional datasets.

[1] See Cousins' note at `physics.ucla.edu/~cousins/stats/ongoodness6march2016.pdf`



(a) Histogram with 20 bins

(b) Histogram with 100 bins

Figure 2.2: Illustration of the differences between histograms and neural networks as function approximators. The left panel shows the comparison between a histogram with 20 bins (in blue) and the neural network approximation (in purple). Similarly, the right panel shows the histogram with 100 bins instead. The true underlying function is also displayed in orange.

2.1.1.2 Neural networks

Seeking a more scalable alternative to histograms for the NPLM algorithm, we focus on artificial neural networks (NN), specifically fully connected feed-forward neural networks. These networks have proven to be universal approximations (refer to [4] for a review). In many applications, neural networks have responded satisfactorily to the curse of dimensionality. Using neural networks also allows us to avoid problems related to the binning choice (refer to Fig. 2.2 for a visual comparison between histograms and neural networks as function approximators). Nonetheless, while they solve some problems, they bring in new ones—mainly the introduction of additional parameters linked to the architecture and training process. These parameters, often referred to as *hyper-parameters*, demand careful selection, typically based on model performance.

Given a specific choice of the NN training hyper-parameters, we can define a family of models $\mathcal{F}$ constituted by the NN models (f_{NN}) obtained for different configurations of the trainable parameters defined in a space Ω:

$$\mathcal{F}_{\text{NN}} = \{f_{\text{NN}}(x; \mathbf{w}), \, \forall \mathbf{w} \in \Omega\} \, . \tag{2.4}$$

Neural networks are essentially compositions of elementary blocks called neurons. Each neuron (v) is a real-valued function that takes as input m features and depends on a set of m weights (w) and a bias (b) and produces a real value. This value is derived from the dot product of inputs x and weights w, which is then offset by b and passed through a non-linear function σ (also known as *activation function*):

$$\begin{aligned} v_{\text{w,b};\,\sigma} \, : \, & \mathbb{R}^{\text{m}} \to \mathbb{R} \\ & x \mapsto \sigma(x \cdot \text{w} + \text{b}) \, . \end{aligned} \tag{2.5}$$

In feed-forward neural networks, neurons are organized into layers (l). A layer function with n neurons is represented as:

$$\begin{aligned} l_{\text{w,b};\,\sigma} \, : \, & \mathbb{R}^{\text{m}} \to \mathbb{R}^{\text{n}} \\ & x \mapsto \sigma(x \cdot \text{w} + \text{b}) \, . \end{aligned} \tag{2.6}$$

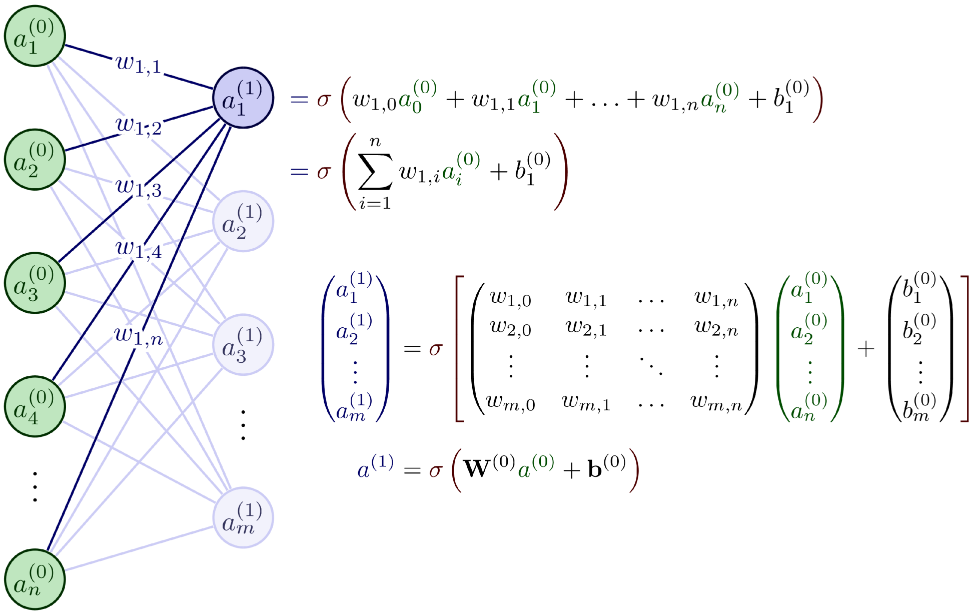

$a_1^{(0)}$
$a_2^{(0)}$
$a_3^{(0)}$
$a_4^{(0)}$
$a_n^{(0)}$
$w_{1,1}$
$w_{1,2}$
$w_{1,3}$
$w_{1,4}$
$w_{1,n}$
$a_1^{(1)}$
$a_2^{(1)}$
$a_3^{(1)}$
$a_m^{(1)}$
$= \sigma \left(w_{1,0}a_0^{(0)} + w_{1,1}a_1^{(0)} + \ldots + w_{1,n}a_n^{(0)} + b_1^{(0)} \right)$
$= \sigma \left(\sum_{i=1}^{n} w_{1,i}a_i^{(0)} + b_1^{(0)} \right)$
$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \sigma \left[\begin{pmatrix} w_{1,0} & w_{1,1} & \ldots & w_{1,n} \\ w_{2,0} & w_{2,1} & \ldots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \ldots & w_{m,n} \end{pmatrix} \begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix} \right]$
$a^{(1)} = \sigma \left(\mathbf{W}^{(0)} a^{(0)} + \mathbf{b}^{(0)} \right)$

Figure 2.3: Visualization of two layers of a generic feed-forward neural network and the mathematical formulation of the forward pass.

In Eq. 2.6, the weight w has dimensions m × n, the bias b has dimension n and σ is applied to each of the n layers output. Notice that neurons within a layer do not cross-talk.

Layers connect sequentially, meaning the output of layer i serves as the input for layer $(i+1)$. Given L layers, a neural network can be mathematically described as

$$f_{\mathrm{NN}}(\,\cdot\,;\,\mathbf{w}) = \left(l^{L}_{\mathrm{w,b};\,\sigma} \circ l^{L-1}_{\mathrm{w,b};\,\sigma} \circ \ldots \circ l^{1}_{\mathrm{w,b};\,\sigma} \right)(\,\cdot\,)\,, \tag{2.7}$$

where here $\mathbf{w}$ denotes all the weights and biases of the neurons in the model. A visualization of the mathematical formulation of neural networks is displayed in Fig. 2.3, along with the matrix representation of the forward propagation from one layer to the next.

Several studies support the idea that neural networks are superior to histograms; some even consider neural networks to be improved binning models. One of the main advantages of neural network models is their smoothness. Histograms are discontinuous by construction at the edges of the bins. The model, therefore, varies rapidly on the edges, and adjacent bins could also have gradients of opposite signs just because of statistical fluctuations in the data sample. This profile is unrealistic for the alternative hypothesis, which is expected to be a continuous function. Neural networks are smoother and more expressive than histograms, requiring fewer parameters to replicate specific data features (e.g., bumps).

2.1.1.3 Kernel methods

Kernel models represent another class of universal approximator functions. They are versatile, not requiring any predefined assumptions about the type of data representation, whether vectors, graphs, or images, to formulate an optimization algorithm. Furthermore, these problems can be simplified into convex optimization problems with strong theoretical guarantees. Kernel models are based on a pair-wise real-valued "comparison function" that takes inputs from the input space $\mathcal{X}$.

$$\begin{aligned} K \;:\; & \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\ & (x,\, x') \mapsto k(x,\, x')\,. \end{aligned} \tag{2.8}$$

Kernel methods are algorithms that operate on a dataset $\mathcal{D} = \{x_i \in \mathcal{X};\, i = 1, \ldots, \mathcal{N}_\mathcal{D}\}$ and use the kernel function values for every pair of points in that dataset as input. Those can be synthetically expressed as a matrix K whose entries are defined as

$$\mathrm{K}_{i,j} := K(x_i,\, x_j) \qquad\qquad \forall\, i,\, j = 1, \ldots, \mathcal{N}_\mathcal{D}\,. \tag{2.9}$$

The kind of kernel models relevant for building kernel methods are positive-definite. A kernel model is classified as positive-definite if the function is symmetric for all pairs $(x,\, x') \in \mathcal{X}^2$, meaning $K(x,\, x') = K(x',\, x)$, and if the relation

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j\, K(x_i,\, x_j) \geq 0 \tag{2.10}$$

holds for all pairs of series $(x_1, \ldots, x_N) \in \mathcal{X}^N$ and $(a_1, \ldots, a_N) \in \mathbb{R}^N$ that can be built taking any $N \in \mathbb{N}$. Furthermore, positive-definite kernels can always be defined as an inner product over a Hilbert space $\mathcal{H}$

$$k(x,\, x') = \langle \Phi(x),\, \Phi(x') \rangle \qquad\qquad \forall\, x,\, x' \in \mathcal{X}\,, \tag{2.11}$$

where $\Phi : \mathcal{X} \to \mathbb{R}^d$ maps the input features into $\mathcal{H}$. This property enables the kernel method algorithm to be defined and solved in a Hilbert space $\mathcal{H}$, where all calculations can be expressed as inner products and solved with linear operations alone. This is preferable to defining and solving the algorithm in the original space $\mathcal{X}$, which can be arbitrarily complex and difficult to treat mathematically. Moreover, working with linear forms in the Hilbert space implicitly induces non-linearities in the original space $\mathcal{X}$, making the approach highly adaptable.

The Reproducing Kernel Hilbert space (RKHS or RK Hilbert space) is the specific Hilbert space associated with a positive definite kernel model used to solve machine learning problems. It contains all functions $f : \mathcal{X} \to \mathbb{R}$ that satisfy the reproducing property $f(x) = \langle f,\, K(x,\, \cdot) \rangle_\mathcal{H}$ for all the elements $x \in \mathcal{X}$. Each point x can be represented by a function $f(x)$ in RK Hilbert space. In $\mathcal{H}_{\mathrm{RK}}$, the initial representation of the data is no longer relevant, so any type of data can be treated the same way with no need for additional tailoring. The absence of an assumption on the nature of the input space $\mathcal{X}$ makes kernel methods flexible and universal tools for data analysis.

The literature presents several possible kernel functions. A few examples include

- **Polynomial kernels:** $k(x,\, x_i)_{c,\,p} = \left(x^{\mathrm{T}} \cdot x_i + c\right)^p$, where $p \in \mathbb{N}$ and $c \geq 0$;
- **Gaussian kernels:** $k(x,\, x_i)_\sigma = \exp\left[-\frac{||x-x_i||^2}{2\sigma^2}\right]$;
- **Laplace kernels:** $k(x,\, x_i)_\sigma = \exp\left[-\frac{||x-x_i||}{2\sigma}\right]$.

A more extensive illustration can be found in [5].

To conclude, we will introduce two important theoretical aspects of kernel methods: the *kernel trick* and the representer theorem.

The kernel trick allows any algorithm based on pair-wise inner products on finite-dimensional vectors to be extended to infinite-dimensional vectors by replacing the inner product's evaluation with the corresponding kernel's calculation. The kernel trick's crucial practical benefit is that the elements in the feature space are never manipulated explicitly but only through

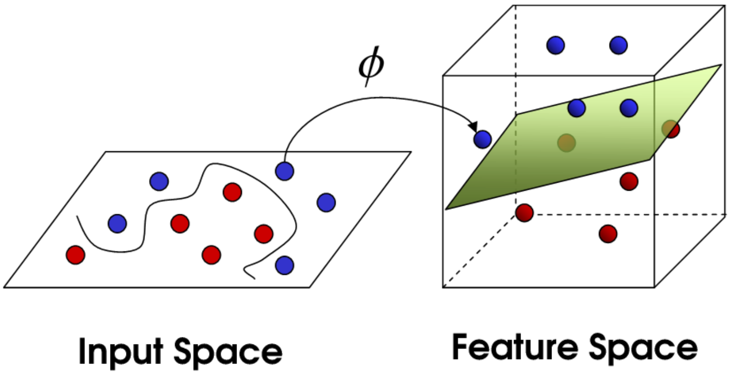


Figure 2.4: Visualization of the mapping from the input space to the Hilbert space at the basis of kernel methods.

the inner products defined in the RK Hilbert space. Therefore, it can replace classical algorithms with less standard ones, for example, modifying metric definitions or transferring valid definitions in the Euclidean space into spaces of non-vectorial data.

The representer theorem provides a limit to the dimension of the solution for kernel methods defined on a finite set of points $\mathcal{D} = \{x_1, \ldots, x_{\mathcal{N}_\mathcal{D}}\} \subseteq \mathcal{X}$. Namely, let us consider the problem of finding the function f over an RKHS $\mathcal{H}$ such that it minimizes a function $\psi : \mathbb{R}^{\mathcal{N}_\mathcal{D}} \times \mathbb{R} \to \mathbb{R}$

$$\psi\left(f(x_1), \ldots, f(x_{\mathcal{N}_\mathcal{D}}), ||f||_\mathcal{H}\right) , \tag{2.12}$$

which is strictly increasing with respect to the last variable. This minimization problem admits a solution f^* in a finite-dimensional subspace spanned by $\mathcal{N}_\mathcal{D}$ kernels $(K(x_1, \cdot), \ldots, K(x_{\mathcal{N}_\mathcal{D}}, \cdot), \cdot)$, and f^* can be written as

$$f^*(x; \mathbf{w}) = \sum_{i=1}^{\mathcal{N}_\mathcal{D}} w_i\, K(x_i, x)\,. \tag{2.13}$$

Often, in machine learning applications, the function ψ that we are interested in minimizing has the form

$$\psi\left(f(x_1), \ldots, f(x_{\mathcal{N}_\mathcal{D}}), ||f||_\mathcal{H}\right) = L\left(f(x_1), \ldots, f(x_{\mathcal{N}_\mathcal{D}})\right) + \lambda\, U\left(||f||_\mathcal{H}\right) , \tag{2.14}$$

where L is a “loss function” measuring the goodness of fit of the function f to a specific problem and U is a regularization term depending on the norm enforcing a certain degree of smoothness to f. Knowing that the solution f^* exists in a subspace of the RKHS of finite dimension $\mathcal{N}_\mathcal{D}$ allows the definition of efficient algorithms, although the RKHS could be infinite-dimensional.

The three families of universal approximators just described (histograms, NNs, and kernel methods) are not the only existing ones. Others, like Legendre polynomials, are not treated in this thesis and have not been explored in this work. Now that we have a set of potential definitions for f, our next step is to approach computing the log-likelihood-ratio test statistic as a machine learning machine task.

2.1.2 Maximum likelihood from minimal loss

The NN-based NPLM approach starts with using the exponential of a universal approximator $f(\cdot, \mathbf{w})$ as a local scale factor to parameterize the departure of the alternative hypothesis from the Reference hypothesis. In simpler terms, the alternative hypothesis describes a locally rescaled version of the differential distribution predicted by the Reference hypothesis:

$$n(x \,|\, \mathrm{H}_{\mathbf{w}}) = e^{f(x;\, \mathbf{w})}\, n(x \,|\, \mathrm{R})\,. \tag{2.15}$$

In this formulation, while the Reference hypothesis R is assumed to be simple (i.e., with no free parameters), the alternative hypothesis $\mathrm{H}_{\mathbf{w}}$ is composite, and its only free parameters are the trainable parameters of the neural network $\mathbf{w}$.

The exponential parametrization ensures that $n(x \,|\, \mathrm{H}_{\mathbf{w}})$ is positive, thereby enabling the use of $f(x;\, \mathbf{w})$ as a direct approximation for the log-ratio of the distributions:

$$f(x;\, \mathbf{w}) = \log\left[\frac{n(x \,|\, \mathrm{H}_{\mathbf{w}})}{n(x \,|\, \mathrm{R})}\right]. \tag{2.16}$$

It should be noted that both Eq. 2.15 and Eq. 2.16 are defined on the differential number of expected observations under each hypothesis. Those are easily turned into probability distributions by applying a normalization factor:

$$p(x \,|\, \mathrm{H}) = \frac{1}{\mathrm{N(H)}}\, n(x \,|\, \mathrm{H}) \qquad \text{where} \qquad \mathrm{N(H)} = \int \mathrm{d}x\, n(x \,|\, \mathrm{H})\,. \tag{2.17}$$

Here, N(H) is the expected number of observations under the generic hypothesis H, and it generally depends on the probability of the observed process, the integrated data acquisition time, and the acceptance and efficiency of the acquisition apparatus. The effective number of collected observations is a random Poissonian variable distributed around this number.

Leveraging Eq. 2.15, we can construct the Neyman-Pearson log-likelihood-ratio test statistic as previously done for histograms. This time, the likelihood is defined as the product of the probabilities over the $\mathcal{N}_{\mathcal{D}}$ observations in $\mathcal{D}$ rather than over the bins. We can interpret this likelihood as a binned likelihood in the limit case of a number of bins equal to the number of data points. Each bin contains a single point (i.e., $o_i = 1$ for all i), and the expected number of events in the bin is precisely the value of the differential counting distribution in that data point, $n(x_i \,|\, \mathrm{H})$:

$$\mathcal{L}(\mathcal{D}|\mathrm{H}) = e^{-\mathrm{N(H)}} \prod_{i=1}^{\mathcal{N}_{\mathcal{D}}} \left(\frac{n_i(\mathrm{H})^{o_i}}{o_i!}\right) = e^{-\mathrm{N(H)}} \prod_{i=1}^{\mathcal{N}_{\mathcal{D}}} n(x_i \,|\, \mathrm{H})\,. \tag{2.18}$$

This way, the stochastic nature of the total number of observations is automatically accounted for by the exponential factor $e^{-\mathrm{N(H)}}$. This likelihood formulation is called *extended likelihood* [6]. The optimal test statistic, according to Neyman and Person, is obtained by maximizing the likelihood of the data under the alternative $\mathcal{L}(\mathcal{D} \,|\, \mathrm{H}_{\mathbf{w}})$ over the space Ω of the possible configurations of $\mathbf{w}$. This allows us to select the best simple hypothesis within the family of alternatives. The test can, therefore, be written as

$$\bar{t}(\mathcal{D}) = 2 \max_{\mathbf{w} \in \Omega} \left\{ \mathrm{N(R)} - \mathrm{N(H_{\mathbf{w}})} + \sum_{i=1}^{\mathcal{N}_{\mathcal{D}}} \log\left[\frac{n(x_i \,|\, \mathrm{H}_{\mathbf{w}})}{n(x_i \,|\, \mathrm{R})}\right]\right\}. \tag{2.19}$$

Notice that the argument of the sum is precisely the definition of $f(x;\, \mathbf{w})$ coming from the parametrization in Eq. 2.16 and thus can be straightforwardly replaced. The first term, N(R), represents the number of expected observations under the Reference hypothesis. It is assumed to be precisely known from the initial assumptions of a Reference model that is free of uncertainties. On the other hand, $\mathrm{N(H_{\mathbf{w}})}$ represents the expected number of observations in the alternative hypothesis, which is not known a priori and needs to be estimated from the data.

When we integrate Eq. 2.15 over x, we can calculate the expected number of observations in the alternative hypothesis:

$$\mathrm{N}(\mathrm{H}_{\mathbf{w}}) = \int \mathrm{d}x \, e^{f(x;\,\mathbf{w})} \, n(x\,|\,\mathrm{R}) = \mathrm{N}(\mathrm{R}) \cdot \mathbb{E}\left[e^{f(x;\,\mathbf{w})}\right]_{p(x\,|\,\mathrm{R})} , \tag{2.20}$$

where the last equality is obtained by multiplying and dividing the integral argument by the normalization factor $\mathrm{N}(\mathrm{R})$ to make the probability explicit. Instead, $\mathbb{E}\left[e^{f(x;\,\mathbf{w})}\right]_{p(x\,|\,\mathrm{R})}$ denotes the mathematical expectation value of $e^{f(x;\,\mathbf{w})}$ given the probability distribution $p(x\,|\,\mathrm{R})$. If the analytic description of $p(x\,|\,\mathrm{R})$ were available, we could calculate the integral in Eq. 2.20 in its exact form. In high-energy physics, probability density distributions are estimated using Monte Carlo techniques by drawing simulated samples of observations since they are rarely known analytically. The integral is then approximated with a discrete sum calculated over a set of observations $\mathcal{R}$ of size $\mathcal{N}_{\mathcal{R}}$, which are simulated according to the Reference hypothesis. We will refer to such sample as 'Reference sample', $\mathcal{R}$. Therefore, Eq. 2.20 is approximated by

$$\mathrm{N}(\mathrm{H}_{\mathbf{w}}) = \mathrm{N}(\mathrm{R}) \cdot \sum_{x\in\mathcal{R}} \mathrm{w}_x \, e^{f(x;\,\mathbf{w})} \,. \tag{2.21}$$

The elements in $\mathcal{R}$ should be reweighted with a factor of ${\mathcal{N}_{\mathcal{R}}}^{-1}$ to preserve normalization, assuming an initial weight of 1. Furthermore, the approximation depends on the size of the reference sample. To keep the approximation error negligible compared to the characteristic scale of the problem, the size of the $\mathcal{R}$ sample, $\mathcal{N}_{\mathcal{R}}$, should be much larger than the expected number of observations $\mathrm{N}(\mathrm{H}_{\mathbf{w}})$.

Using this approximation and the parametrization in Eq. 2.16, we can express the test statistic as

$$\begin{aligned} \bar{t}(\mathcal{D}) &= 2 \max_{\mathbf{w}\in\Omega} \left\{ \mathrm{N}(\mathrm{R}) - \frac{\mathrm{N}(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}} \sum_{x\in\mathcal{R}} e^{f(x;\,\mathbf{w})} + \sum_{x\in\mathcal{D}} f(x;\,\mathbf{w}) \right\} \\ &= 2 \max_{\mathbf{w}\in\Omega} \left\{ \frac{\mathrm{N}(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}} \sum_{x\in\mathcal{R}} \left(1 - e^{f(x;\,\mathbf{w})}\right) + \sum_{x\in\mathcal{D}} f(x;\,\mathbf{w}) \right\} . \end{aligned} \tag{2.22}$$

Notice that Eq. 2.22 is the maximization of the functional $\bar{t}(\,\cdot\,;\,\mathbf{w})$ over the space of the trainable parameters of a neural network, and it is computed using two data samples $\mathcal{R}$ and $\mathcal{D}$ with assigned labels $y = 0$ and $y = 1$, respectively. Thus, the problem is equivalent to minimizing a loss function:

$$\bar{t}(\mathcal{D}) = -2 \min_{\mathbf{w}\in\Omega} \bar{L}(\,\cdot\,;\,\mathbf{w}) \,, \tag{2.23}$$

where, for our purposes, the loss function is

$$\begin{aligned} \bar{L}(\,\cdot\,;\,\mathbf{w}) &= \frac{\mathrm{N}(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}} \sum_{x\in\mathcal{R}} \left(1 - e^{f(x;\,\mathbf{w})}\right) + \sum_{x\in\mathcal{D}} f(x;\,\mathbf{w}) \\ &= \sum_{(x,\,y)} \left[(1-y) \, \frac{\mathrm{N}(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}} \left(e^{f(x;\,\mathbf{w})} - 1\right) - y \, f(x;\,\mathbf{w}) \right] . \end{aligned} \tag{2.24}$$

We have shown one possible method to solve Eq. 2.19 exploiting machine learning. Indeed, there are other solutions available. For instance, when using kernel methods instead of neural networks as function approximations, we use a reweighted version of the binary cross-entropy loss. A discussion on this approach can be found in Sec. 2.3.

2.2 The New Physics Learning Machine

The formulation in Eq. 2.24 aligns with a semi-supervised machine learning problem, where labels are assigned to different training datasets. These samples are, however, not purely populated by one unique data category. Precisely, the Reference sample consists solely of background events. On the other hand, the data sample's composition is unknown. It is presumed to comprise background events as predicted by the Reference hypothesis and might include a small deviation due to New Physics signals. It is also possible that the data sample lacks any distinguishing signal, thus aligning it with the distribution of the Reference sample.

As mentioned earlier, for the results to be accurate, it is imperative that $\mathcal{N}_{\mathcal{R}} \gg \mathcal{N}_{\mathcal{D}}$. This scenario is referred to as having imbalanced classes in machine learning language. Imbalanced classes can negatively impact training performance, as the dominant class may overshadow the training process. This is because each event typically weighs equally in the contribution to the loss function. Model updates that better fit the dominant class have a greater impact on overall loss minimization than those targeting the less populated class. In Eq. 2.24, however, the term related to the Reference sample is weighted by the factor $\mathrm{N(R)}/\mathcal{N}_{\mathcal{R}}$ so that N(R) contributes similar to the one computed over $\mathcal{D}$. This way, the NPLM loss function definition automatically addresses the issue of imbalanced classes.

The first step in deploying the NPLM test statistic is to select a suitable set of universal approximators. This section will concentrate only on neural networks, and the implementation of kernel-based NPLM will be discussed in the following Sec. 2.3.

2.2.1 Machine learning implementation

When using neural networks, we prefer feed-forward networks that use a sigmoid activation function for input and hidden layers and a linear activation function for the output layer. The neural network model $f(x;\mathbf{w})$ should approximate a logarithm, and therefore, its output should span all real values, as motivated by the parametrization choice in Eq. 2.16. The sigmoid activation function is not motivated by any specific reason and can be replaced by other non-linear functions, such as the hyperbolic tangent. The choice of architecture is arbitrary and depends on the complexity of the problem. We privilege small architectures with $\mathcal{O}(10)$ parameters for one-dimensional problems and $\mathcal{O}(10^2)$ parameters for $1 < d < 30$. For now, we will assume that a specific configuration of the model hyper-parameters has been selected, and we focus only on the strategy to extract a p-value in a frequentist manner.

The model trains on all events from $\mathcal{D} \cup \mathcal{R}$ in an unbinned manner with d input properties (*features*) and associated weights and labels. Full-batch gradient updates are used for training, with one epoch corresponding to a single update based on all available events. This approach differs from conventional machine learning methods in two ways. First, it slows down the learning process due to the low rate of updates. Second, batching helps prevent local minima by introducing noise in the gradient-based descent path.

Conversely, segmenting the training sample into smaller fractions effectively lowers the statistics of the observed sample used for decision-making. This is not usually an issue for balanced or almost balanced training datasets. However, the problems we tackle in high-energy physics are characterized by small numbers of signal events on top of a core set of background events. The signal fractions we are interested in studying go from a few percent down to zero. Therefore, splitting the data into batches would result in further rare signals, increasing the risk of misinterpretation as a background fluctuation, which could completely undermine the sensitivity of our analysis algorithm.

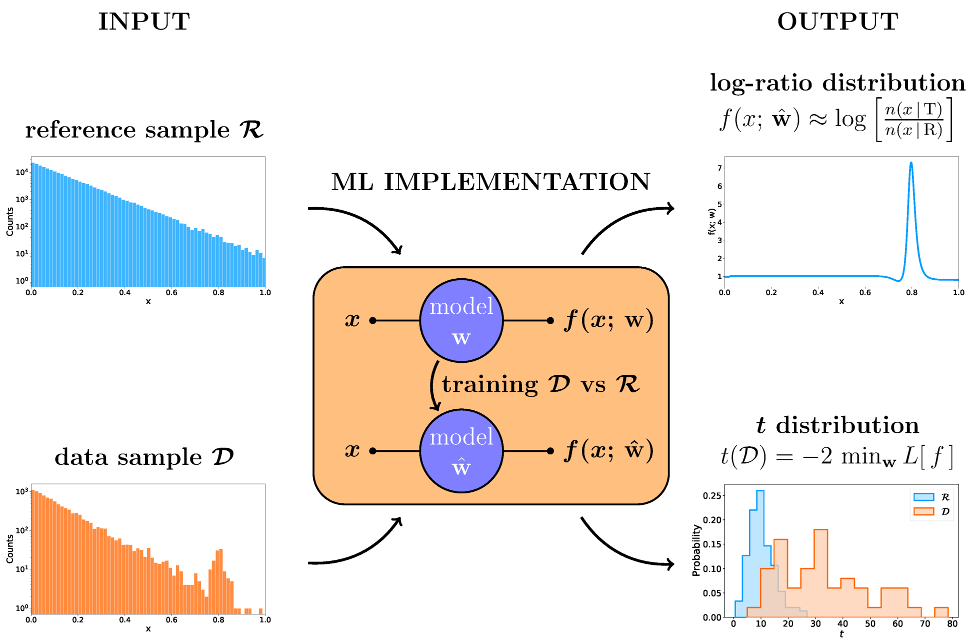

INPUT
OUTPUT
reference sample $\mathcal{R}$
data sample $\mathcal{D}$
ML IMPLEMENTATION
x
model w
$f(x;\,\mathrm{w})$
training $\mathcal{D}$ vs $\mathcal{R}$
model $\hat{\mathrm{w}}$
$f(x;\,\hat{\mathrm{w}})$
log-ratio distribution
$f(x;\,\hat{\mathrm{w}}) \approx \log\left[\frac{n(x\,|\,\mathrm{T})}{n(x\,|\,\mathrm{R})}\right]$
t distribution
$t(\mathcal{D}) = -2\,\min_{\mathrm{w}} L[\,f\,]$
Counts
x
f(x; w)
Probability
t
$\mathcal{R}$
$\mathcal{D}$

Figure 2.5: Schematic view of the NPLM algorithm.

Finally, input features for high-energy physics problems can have very different natures. Typically, inputs are kinematic variables describing a collision event: momentum, angles, energy of the produced particles or combinations of those. Discrete variables like ID or time labels are also considered in some cases. Some variables, like labels or angles, are generally defined in $\mathcal{O}(1)$ ranges. Others, like momenta, energies and invariant masses, span many orders of magnitude instead and carry a unit of measure that can be arbitrarily chosen. Ensuring inputs are on a consistent scale facilitates uniform training across all input space dimensions. This is valid for the NPLM implementation, too. As a result, we consistently standardize the input features before feeding them to the model. For variables with negative and positive values (mainly angles), we center the distribution to zero and rescale it so that the final standard deviation is one. For positive definite variables, like momenta and energies, we instead rescale the distribution by its original mean so that the mean of the distribution at the end of the standardization procedure is centered at one.

2.2.2 Extracting a p-value

The computation of a test statistic (e.g., Eq. 2.19) returns, in general, a scalar number that represents a measure of the analyzed dataset $\mathcal{D}$ in some summary statistics. Being $\mathcal{D}$ a set of random variables, the test statistic is also a random variable, and its distribution depends on the distribution of the observations x constituting $\mathcal{D}$. Once a test statistic t is defined, it is evaluated on the experimental observations, and its value $\bar{t}(\mathcal{D})$ is compared with the distribution of the test statistic under the null hypothesis, $P(t\,|\,\mathrm{R})$. We define the p-value as the probability of observing a value of t at least as incompatible with the null as the observed one, assuming the null hypothesis to be true (right panel of Fig. 2.1). This means we imagine repeating the experiment multiple times and asking ourselves what the probability distribution would be if the collected data were drawn from the null distribution. The p-value is then computed as the right tail of the distribution exceeding the observed value:

$$p\,[\,\bar{t}(\mathcal{D})\,] = \int_{\bar{t}(\mathcal{D})}^{\infty} P(t\,|\,\mathrm{R})\,\mathrm{d}t\,. \tag{2.25}$$

Repeating the experiment many times and controlling the data source so that the null hypothesis is guaranteed to be correct is not generally possible. Therefore, in most high-energy physics applications, repeated experiments are obtained using simulations. More specific applications, such as the one covered in Chap. 4 of this thesis, allow instead clever strategies to ensure the data is collected under the assumptions of the null hypothesis. In any case, several pseudo-datasets are built based on the null hypothesis, and the test is run on them. The test outcome for each pseudo-dataset is then used to fill a histogram, which builds an empirical distribution of the test statistic under the null hypothesis.

In the case of the NPLM test statistic, a new maximum likelihood fit and neural network training is required for each dataset. The same training scheme is used for each pseudo-dataset, including the number of epochs, hyper-parameters, and optimizer. Notice that the value of $\bar{t}(\mathcal{D})$ depends on the Reference sample $\mathcal{R}$, which is used along with $\mathcal{D}$ as an input to the neural network. However, assuming a much larger sample size for $\mathcal{R}$, its fluctuations become negligible compared to those of the dataset of interest. Therefore, we can consider $\mathcal{R}$ as an infinite statistics sample that represents our exact knowledge of the Reference hypothesis and is a constant input to the algorithm.

2.2.3 Asymptotic formula for the test statistic distribution

The NPLM test statistic approximates the most powerful log-likelihood-ratio test statistic under the Neyman-Pearson lemma. When the null hypothesis is a subset of the alternative, Wilks and Wald's studies state that, under the null hypothesis and in the limit of large sample size, the test statistic follows a chi-squared distribution with degrees of freedom equal to the difference in dimensionality between the space of the parameters describing the alternative and those describing the null hypothesis. It is not always clear when the conditions for the asymptotic limit are met, and the minimum sample size required depends on the problem's complexity and the network model's complexity used to calculate the test.

Verifying that the test is independent of the Reference sample is important. This is not an issue if $\mathcal{N}_{\mathcal{R}} \gg \mathcal{N}_{\mathcal{D}}$. It could be relevant if the Reference size is only slightly larger than the data samples. Additionally, the training procedure instabilities can disrupt the emergence of a χ^2. The divergence of the training due to isolated points that the model tries to fit with spikes, for instance, can create outliers on the right tail of the test statistic distribution.

It is unclear what the effective number of degrees of freedom of the putative χ^2 is in the NPLM test statistic. However, in the case of neural networks, the number of trainable parameters of the model is an upper bound to it. However, the effective number of degrees of freedom could be lower due to the non-linear relations between the parameters. On the other hand, as will be discussed in the next section, the kernel models are non-parametric. Namely, they cannot be defined by a finite set of parameters. They are often said to be parametrized by an infinite set of parameters or, better, by functions of the data (the kernels). Hence, the amount of information they can capture about the data can grow as the dataset grows. This makes them more flexible but also prevents us from guessing any specific number of degrees of freedom for the final test statistic distribution.

2.2.4 After-training anomaly inspection

Another powerful product of the NPLM algorithm is the model output at the end of training. We have seen that for a given dataset $\mathcal{D}$, a neural network is trained to perform the maximum likelihood fit of the neural network trainable parameters to the data. At the end of the training, the configuration of the model parameters, $\hat{\mathbf{w}}$, is the one for which the model returns the best approximation of the true data distribution (T),

$$f(x;\,\hat{\mathbf{w}}) = \log\frac{n(x\,|\,\hat{\mathbf{w}})}{n(x\,|\,\mathrm{R})} \approx \log\frac{n(x\,|\,\mathrm{T})}{n(x\,|\,\mathrm{R})}\,. \tag{2.26}$$

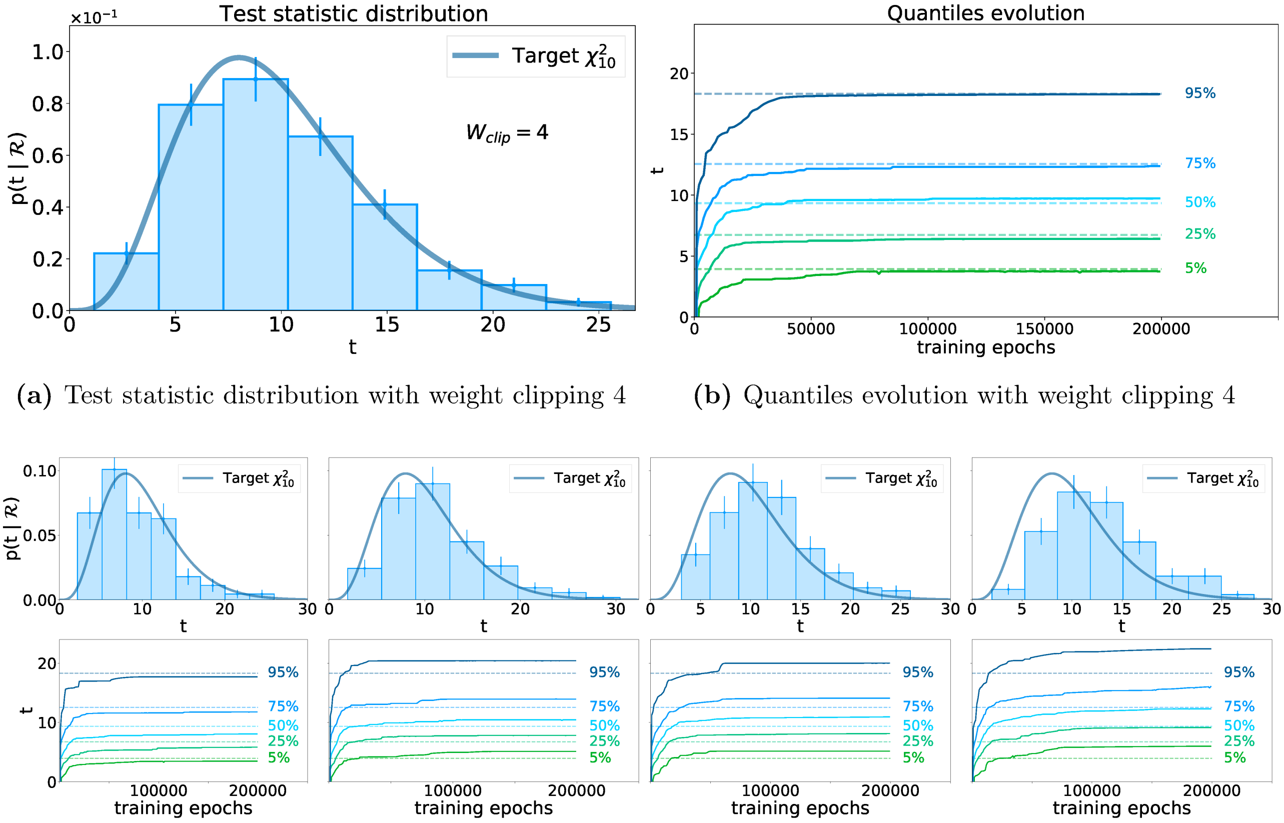
The model characterizes the difference between $\mathcal{R}$ and $\mathcal{D}$ samples as a logarithmic ratio. The neural network output is null when the samples agree, positive for excess detected in the data and negative for present deficits. Furthermore, since the model is a continuous function of the input variables, it is possible to analyze the shape of non-local deviations. There are two main ways of exploiting the information provided by the model. The first and simplest one to implement is to use the model's output as a binary classification metric, defining a score threshold to select the data events and tag them as belonging to the atypical region. Although this can locate the discrepancy, it completely loses the information about its shape. To better understand the shape of the information, one can reconstruct the marginal distributions of the input data variables learned by the neural network model. This can be achieved by using the parametrization in Eq. 2.16 and reweighting the Reference sample with the exponential of the model. For one-dimensional problems, $f(x;\,\hat{\mathbf{w}})$ can be visualized for any interval of x. However, for multi-dimensional problems, the correlation between features must be specified for a meaningful result. Additionally, the distribution of any input variable combination can be reconstructed using the same reweighting scheme for the Reference sample.

2.2.5 A regularization scheme for NN-based NPLM

The NPLM algorithm requires a specific regularization approach for the neural network models used to build it. This is due to the unbounded loss function defined in Eq. 2.24, which takes negative infinite values if $f(x;\,\mathbf{w})$ tends to positive infinite values. The model may try to fit a sharply localized peak from a singular, isolated data point within sample $\mathcal{D}$, leading to such a situation during training. The neural network model's resolution capabilities determine the concept of isolation. Every data point can be seen as distinct or isolated upon closer examination. Therefore, constraints should be imposed to adjust the neural network flexibility to the resolution of the experimental apparatus used to acquire the data. A way to incorporate such physics knowledge into the model definition is by introducing *weight-clipping*. This is a hyperparameter of the neural network training procedure that sets an upper bound on the absolute value that each weight parameter of the neural network can take during training. The weights of a model determine the output changes with respect to input features encoded in x. Limiting weight values sets a maximum slope for building spikes around points, which affects resolution. The ideal solution would be to adjust the weight-clipping locally based on individual needs. However, for simplicity, we currently apply a constant upper bound to all neural network model weights, allowing us to study only one parameter at a time.

A crucial aspect of the algorithm implementation is the choice of the weight clipping parameter for recovering the asymptotic χ^2 distribution for the test statistic under the Reference hypothesis. Our tests reveal that if the chosen weight clipping parameter is overly generous, the test statistic distribution tends to gravitate towards higher numbers. At the same time, it is gradually squeezed to zero if the weight clipping is too small. This trend is observed for problems with varying input features, training sample sizes, and neural network architectures. As a result, the algorithm configuration can include adjusting the weight clipping parameter based on test statistic distribution.

Furthermore, it has been observed that no appropriate weight clipping value can be found for certain architectures, preventing the emergence of a χ^2 distribution. Besides reducing in-



(a) Test statistic distribution with weight clipping 4

(b) Quantiles evolution with weight clipping 4

(c) $\mathrm{w_{clip}} = 1$

(d) $\mathrm{w_{clip}} = 10$

(e) $\mathrm{w_{clip}} = 50$

(f) $\mathrm{w_{clip}} = 100$

Figure 2.6: Example of the weight clipping tuning procedure results for a one-dimensional problem solved using a simple $1 \times 3 \times 1$ neural network architecture yielding ten trainable parameters. The best fit of the test statistic distribution to a χ^2 with 10 degrees of freedom is presented on the two top panels.

stabilities caused by incorrect resolution, weight clipping tuning aimed at the asymptotic χ^2 behavior offers a reliable method for selecting a neural network candidate. The selection of the architecture remains open-ended, as for numerous structures, there exists an ideal weight clipping value that aligns with a χ^2 distribution. This approach, however, filters out unsuitable architectural options, ensuring a level of statistical reliability in the outcomes.

It is important to note that only the Reference hypothesis needs to be specified to study the asymptotic behavior of the test statistic. The Reference sample $\mathcal{R}$ and the pseudo-data samples $\{\mathcal{D}\}$ are drawn from the Reference distribution. This regularization procedure is independent of the alternative hypothesis. It is valid regardless of the data nature of the data as long as the sample size expectation, input variable choice, and pre-processing remain the same.

An example of weight clipping is demonstrated in Fig. 2.6. A one-dimensional problem is solved using a neural network with one hidden layer of three neurons for a total of 10 trainable parameters. The histogram panels display the empirical distribution of the test statistic obtained from 50 pseudo data sets drawn from the Reference model for different values of the weight clipping. Higher weight clipping values correspond to larger test statistic values, which shift the empirical distribution to the right of the x-axis. The other panels use five empirical quantiles to show how the empirical distribution evolves over training time. The test statistic stabilizes over the training epoch for small weight clipping values, enabling a fair comparison with the target χ^2 distribution. If the weight clipping values are larger, the quantiles fail to reach a plateau that matches the target χ^2 quantiles.

Overall, the selection of neural network models involves balancing two principles. The first is maximizing the model's ability to fit complex departures from the Reference model expectation, making it sensitive to the largest possible variety of New Physics scenarios. This means seeking the highest complexity that the available computational resources can handle in a reasonable time. On the other hand, given the finite amount of training data, the model should be simple enough for the distribution of the associated test statistic to be in the asymptotic regime. This condition is enforced by monitoring the compatibility with the χ^2 asymptotic formula for the test statistic distribution under different choices of the weight-clipping.

2.3 Algorithm time optimization with kernel methods

Although the neural network implementation of the New Physics Learning machine meets the requirements for real analysis in LHC experiments, other aspects of the NPLM implementation may impede its application in increasingly complex scenarios.

On the one hand, the algorithm's sensitivity in detecting various kinds of discrepancies between the data and the Reference model should be enhanced. The main determinant factor is the choice of neural network architecture and hyper-parameters. The neural network model is not infinitely flexible, and the model's ability to fit sharp features in input variables is limited by the regularization technique used to maintain stability during training.

On the other hand, the algorithm's execution time is even more pressing. When implementing the NPLM with neural networks, the CPU takes consistently longer to train a single toy than for simple models in standard supervised problems. In this section, we present an alternative implementation of the NPLM based on kernel methods that greatly reduces convergence time. This is achieved through the use of a library called Falkon, which solves a Nyström-approximated version of kernel models exploiting parallelization on GPUs [7].

2.3.1 Time-efficient NPLM with non-parametric models

The execution time is essential to expand the algorithm's possible use case range. To exploit the method in quasi-online applications such as a data quality monitoring (DQM) tool, explained in Chap. 4, experimental observations accumulated over a short time window need to be analyzed. The algorithm must process a given sample of data before a new one is gathered. Furthermore, as explained in the following two chapters, we plan to use the NPLM for New Physics searches in a quasi-online fashion, mirroring DQM tools. Therefore, one potentially limiting factor of this analysis algorithm is its execution time.

Offline analyses are typically less computationally intensive than online analyses because they can apply numerous selection filters and reduce the problem's dimensionality to relevant summary statistics. Additionally, there is no real-time constraint since the data is stored permanently. Nonetheless, approaches like NPLM, which do not target specific models, aim to search for anomalies as inclusively as possible across many observables. This requires efficient handling of large datasets. In addition, the NPLM algorithm requires running several experiments to regularize the model and evaluate the test statistic distribution under the null hypothesis. This increases the time requirements by approximately three orders of magnitude. This issue can be addressed by utilizing a CPU cluster to concurrently operate several processes. Still, the time per single toy can be unsustainable if it is of the order of days.

The absence of a target performance makes it impossible to use standard stopping criteria based on validation metrics. The final value of the loss function is directly related to the test statistics output. Hence, to obtain reliable p-values, all training replicas must reach perfect convergence to a plateau and yield a stable test statistics distribution during training. Furthermore, the current neural network-based NPLM algorithm does not use an efficient batching procedure. Batching could cause the statistical significance of a small fraction of signal events to be lost since splitting the data into batches reduces the chance of being sensitive to small signal injections.

To speed up the execution time, we explore replacing neural networks with kernel methods. Kernel methods solve non-linear problems with non-parametric models. However, their basic form has limited applicability due to stringent computational requirements for large-scale data. Nevertheless, several approaches have been considered to reduce the computational demand of kernel methods, introducing efficient approximations, like the Nyström approximation. The Falkon library efficiently solves kernel methods even when the data sample reaches millions of points. It combines several algorithmic principles, like stochastic subsampling, iterative solvers and preconditioning. Linear algebra operations are re-implemented out-of-core to fully exploit GPU acceleration and parallelization with multiple GPUs.

2.3.2 Machine learning implementation

As mentioned in Section 2.1.2, we have demonstrated a method to employ machine learning for solving Eq. 2.19, which evaluates the log-likelihood ratio test statistic. Replacing neural networks with kernel methods, we adopt a reweighted version of the binary cross-entropy loss

$$\begin{aligned} L_{\mathrm{BCE}}\,[\,f\,] &= \frac{\mathrm{N(R)}}{\mathcal{N}_{\mathcal{R}}} \sum_{x\in\mathcal{R}} \log\left[1+e^{+f(x)}\right] + \sum_{x\in\mathcal{D}} \log\left[1+e^{-f(x)}\right] \\ &= \sum_{(x,\,y)} \left\{ (1-y)\,\frac{\mathrm{N(R)}}{\mathcal{N}_{\mathcal{R}}} \log\left[1+e^{+f(x)}\right] + y\,\log\left[1+e^{-f(x)}\right] \right\} . \end{aligned} \tag{2.27}$$

We can easily show that, in the limit of infinite statistics, the sums over the $\mathcal{D}$ and $\mathcal{R}$ training samples can be seen as the approximations of integrals over the data probability distribution, $p(x\,|\,\mathrm{T})$, and the Reference probability distribution, $p(x\,|\,\mathrm{R})$:

$$L_{\mathrm{BCE}}\,[\,f\,] \simeq \mathrm{N(R)} \int \mathrm{d}x\, p(x\,|\,\mathrm{R}) \log\left[1+e^{+f(x)}\right] + \mathrm{N(T)} \int \mathrm{d}x\, p(x\,|\,\mathrm{T}) \log\left[1+e^{-f(x)}\right] . \tag{2.28}$$

The function f that minimizes L is then

$$f(x,\,\hat{\mathbf{w}}) \simeq \log\frac{n(x\,|\,\mathrm{T})}{n(x\,|\,\mathrm{R})} , \tag{2.29}$$

which coincides with the solution of the loss function previously defined in Eq. 2.24.

We will, in general, describe the universal approximator f using a base of Gaussian kernels:

$$f(x,\,\mathbf{w}) = \sum_{i=1}^{\mathcal{N}_{\mathcal{D}}} \mathrm{w}_i\, k_\sigma(x,\,x_i) , \tag{2.30}$$

where the parameter σ is the standard deviation determining the kernel width.

Two terms characterize the loss function used for the training:

$$L\,[\,f\,] = L_{\mathrm{BCE}}\,[\,f\,] + \lambda\, U(||f||) ; \tag{2.31}$$

where the first term is the weighted cross-entropy loss defined in Eq. 2.27 while the second term is a L_2 penalty, given by

$$U(||f||) = \sum_{i,j} \mathrm{w}_i \, \mathrm{w}_j \, k(x_i, \, x_j) \, , \tag{2.32}$$

which constrains the Lipschitz smoothness of the function.

The L_2 penalty acts as a regularization on the model, preventing the “overfitting” of sharp features and speeding up the convergence of the minimization problem. The parameter λ sets the weight of the regularization term over the total loss function, determining both the model's flexibility and the training time.

Solving this problem takes time and resources that scale cubically with the number of data points $\mathcal{N}_{\mathcal{D}}$ and quadratically with the dimensionality of the features space. Research has been done to optimize time and memory resources for kernel methods [8–11].

The implementation exploited for the NPLM is supported by the Falkon library [7]. The main mathematical simplification introduced in Falkon is called Nyström approximation, and the aim is to reduce the number of points $\mathcal{N}_{\mathcal{D}}$ over which to compute the kernels to a shorter set of M points, randomly extracted from $\mathcal{D}$ and called *centers*. The problem is then solved by an approximate Newton iteration (see [7] for technical details. The number of centers determines the accuracy of the approximation, and it has been shown that $M \sim \mathcal{O}\left({\mathcal{N}_{\mathcal{D}}}^{1/2}\right)$ provides satisfying solutions in general. The number of Nyström centers M, the regularization parameter λ, and the Gaussian kernel width σ are the three main hyper-parameters of the kernel methods implementation of the NPLM.

Another remark should be added regarding how the kernel method training results in the retrieval of the final value of the test statistic. When using the cross-entropy loss, unlike with the original loss function in Eq. 2.24, the test statistic must be explicitly computed at the end of the training using Eq. 2.22.

2.3.3 Heuristic hyper-parameter selection scheme for kernel models

The NPLM implementation based on kernel methods also requires some regularization. We have seen that a L_2 penalty term is considered by default in the machine learning implementation used to solve the kernel method problem for NPLM, and its weight is controlled by a hyper-parameter that we call λ. Imposing a penalty term related to the model norm has the same effect of smoothing the function, preventing it, to some extent, from overfitting isolated points. The action is similar to weight clipping in neural network models. In practice, the L_2 penalty prevents the trainable parameters w_i, which are the coefficients weighting the series of kernels defining f, from taking too large values.

Furthermore, the smoothness of the function can be modified acting on the kernel width σ, which characterizes the Gaussian kernels. The choice of λ and σ depends on the number of centers used for model approximation. To optimize the configuration of the three hyper-parameters M, σ, and λ they all need to be considered together.

Unlike neural networks, kernel methods do not have a target number of degrees of freedom for the expected χ^2. Nonetheless, it was heuristically observed that every configuration of $(M, \, \sigma, \, \lambda)$ produces an empirical distribution of the test statistic under the null hypothesis that fits a χ^2. Hence, although the number of degrees of freedom cannot be known beforehand, it

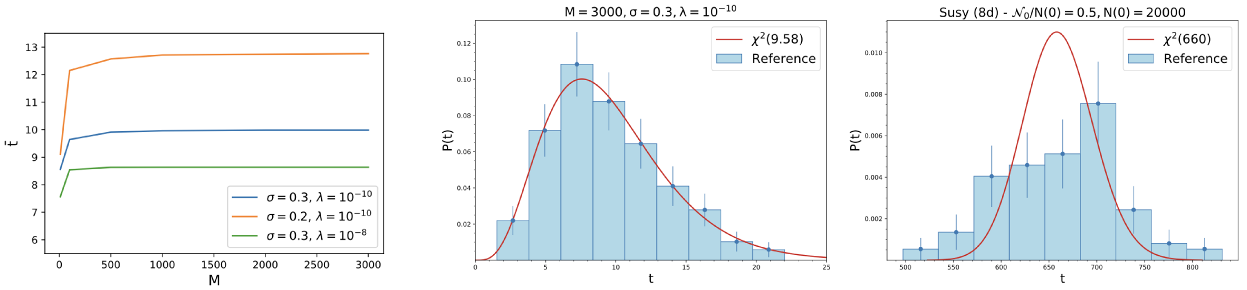


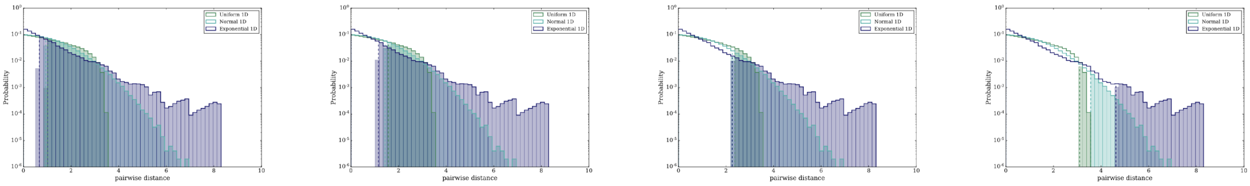
Figure 2.7: Left: evolution of the empirical test statistic distribution median over the number of Nyström centers M. A saturation trend can be observed for different choices of the σ and λ parameters choice. The curves also demonstrate how the distribution gets closer to zero, increasing the values of σ and λ. Center: example of the empirical distribution of the test statistic matched to a χ^2 distribution fitting the number of degrees of freedom to the data. Right: example of missed observation of the χ^2 behavior in the experimental condition of equal size Reference and Data training samples. (Figures from [12]).

can be estimated by fitting the χ^2 to the empirical distribution. In the left panel of Fig. 2.7, we display the median of the empirical distribution for different hyper-parameter configurations and perform a χ^2 compatibility check for each point. We observe that the number of effective degrees of freedom rises with more centers but decreases with stricter regularization using higher values of σ and λ. In the central panel of Fig. 2.7 instead, we display an example of empirical test statistic distribution under the Reference hypothesis, which matches a χ^2 distribution fitting the number of degrees of freedom to the data. Finally, in the right panel of Fig. 2.7, we present an exception to this behavior, which arises when the Reference sample size is equal to or smaller than the size of the (pseudo-)data samples. This is, however, a working condition that never occurs in our experiments.

We can use the saturation pattern of median $\bar{t}(\mathcal{D})$ as a function of M to develop a heuristic procedure for M selection. Any M value within the saturation regime is a good choice, and the higher the number of centers, the better the Nyström approximation. However, if the number of centers is increased, the algorithm's execution time will be slowed down. For binary classification tasks, keeping the number of centers M at $\mathcal{O}(\sqrt{N})$ avoids accuracy losses while keeping computational costs affordable [10]. As a final trade-off prescription, we choose $M \geq \sqrt{N}$ with an upper limit set by computational resources.

The choice of σ and λ is more complicated. Since the binary cross-entropy loss function is a well-defined convex function, the role of the L_2 regularization is less relevant than in the neural networks case. The regularization term is only needed to keep the training process stable[2]. Therefore, the regularization parameter λ can be kept as small as possible while preventing instabilities from occurring. When determining σ, we encode information about relevant scales of the problem in the kernel width in the input feature space. As with weight-clipping, relevant scales differ in different regions of the phase space. Ideally, we model such dependency by making σ a function of x. This is not a trivial task and is not possible in the current implementation of the NPLM algorithm using the FALKON library. As a result, we are restricted to using a global value of σ. To determine this global value of σ, we analyze the distribution of pair-wise Euclidean distances among training sample points. The pair-wise distance between points is commonly used in data analysis to characterize a statistical sample. Samples from different sources are also characterized by other distributions of the pair-wise distance.

[2] See FALKON documentation at `falkonml.github.io/falkon/` for more details.



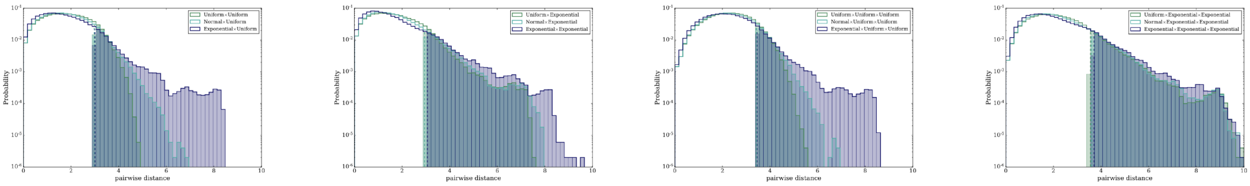
(a) 50% quantile

(b) 70% quantile

(c) 90% quantile

(d) 99% quantile

Figure 2.8: Examples of distributions of the pair-wise distance between the points of each dataset for three one-dimensional toy models. The 50, 70. 90 and 99% quantiles are highlighted in the four panels. The shaded areas represent the corresponding right tails. (Figures from [1]).



(a)

(b)

(c)

(d)

Figure 2.9: Examples of distribution of pair-wise distances for two- and three-dimensional toy models. The sparser dimension dominates the pair-wise distance distribution outcome. A combination of sparse variables broadens the distribution. (Figures from [1]).

We can see the distribution of pair-wise distances for three samples of 1000 points each in Fig. 2.8. The samples include a univariate uniform distribution, standard normal distribution, and exponential distribution with unit average. All samples have been pre-processed to set their standard deviation to unity. The vertical dashed lines for each histogram in the four panels represent the distribution's 50th, 70th, 90th, and 99th percentiles. The shaded areas signify the corresponding right tails. Significant variations in the pair-wise distribution can be observed at the 99th percentile.

Fig. 2.9 displays the 90th percentile position of combined independent random variables in higher dimensional problems (2D and 3D). Panels (a) and (b) exhibit two-dimensional problems, while panels (c) and (d) showcase three-dimensional ones. Each panel shows three examples with one or two independent extra dimensions extending the original uniform, normal, and exponential univariate problems. Firstly, increasing the number of dimensions, the 90th percentile shifts to the right. Secondly, the shape of the distribution is determined by the most sparse variable, which is the exponentially distributed one.

This study can inspire a heuristic process for choosing the appropriate σ. To be completely conservative against overfitting, one would choose the maximum pair-wise distance as the optimal σ value. However, this approach would prevent the model from fitting isolated peaks and potential signals. We have previously observed that in univariate problems, the effects of significant variations in the data distribution are only noticeable in the pair-wise distance distribution beyond the 90th percentile. Therefore, the 90th percentile is a reliable metric for determining data sparsity, allowing for flexibility in fitting data shapes while not relying on the specific data source. However, this only holds when the training sample has been standardized so that the same standard deviation characterizes all input features.

To summarize, the three heuristic rules used to FALKON's hyper-parameters are:

- $M \geq \sqrt{N}$, remembering that increasing M improves accuracy but reduces time efficiency;
- λ as small as possible without compromising the stability of the algorithm;
- σ set to the 90th percentile of the pair-wise (Euclidean) distance distribution (assuming all input variables have been properly standardized).

Chapter 3

The quest for unbiased datasets at the CMS experiment

Particle physics delves into the nature and behaviors of the fundamental particles that compose our universe. An essential subset of particle physics investigations occurs at accelerator experiments. Here, particles, including protons and electrons, are propelled at incredible velocities, colliding with stationary targets (fixed-target experiments) or other accelerated particles (colliders). The resulting collisions produce a multitude of secondary particles, providing insights into the fundamental interactions that generated them. Particle detectors at these collision points detect and reconstruct the properties of the produced particles.

In high-energy physics experiments, the data acquisition system (DAQ) plays an essential role. It receives the information from the particle detector and transforms it into a format suitable for further analysis. Central to the DAQ is the trigger system. The trigger system filters particle interactions that are considered "interesting" in the data and discards background events. The primary motivation for the trigger system is to select only pertinent events for read-out, addressing storage and subsequent data analysis challenges. This selective process is crucial due to bandwidth limitations, largely influenced by thermal constraints.

Hadron colliders, like the Large Hadron Collider (LHC), are often characterized by a background dominated by inelastic proton scattering, with the cross-section of signals of interest being significantly lower. For instance, the total cross-section of proton-proton scattering at $\sqrt{s} = 14$ TeV is around 0.11 barns. The cross-section of the Higgs boson production via vector boson fusion is ten orders of magnitude smaller. In order to produce interesting physics within a reasonable experimental lifetime, hadron colliders must reach high instantaneous luminosity. This is achieved by producing a high rate of events where very focused bunches of protons collide, generating multiple simultaneous interactions. Collision events generate a large number of particles. Typically, only one proton pair in an event leads to an interesting interaction, while the others constitute a background. The interesting interaction is often called "primary vertex", while the background is referred to as "pile-up" (PU). Large amounts of information must be read from the detectors at hadron colliders, as they often require highly segmented devices and precise tracking to reject pile-up interactions and identify the primary vertex. Particle physics demands an astute evaluation of the data, a task where the trigger system proves indispensable. However, this very system, configured based on our current understanding of particle physics, can potentially filter out subtle manifestations of New Physics that do not align with established selection algorithms. There is a genuine concern that groundbreaking signals could be misclassified as background due to their non-conformity to current trigger paradigms. In the context of the methodologies described in the preceding chapter, mainly introducing the New

Physics Learning Machine (NPLM), the harmony between the trigger system and data analysis becomes pivotal. Unbiased datasets amplify the capabilities of the NPLM and any other model-independent anomaly detection algorithm, heightening its potential to detect deviations from the established Standard Model.

The following sections will provide an in-depth overview of the CMS (Compact Muon Solenoid) experiment, one of the four main experiments operating at the LHC. Within the CMS experiment, our attention narrows to the muon system and trigger system, crucial for understanding both the difficulties and the possible advancements in producing unbiased datasets at particle colliders. In this regard, we will introduce *data scouting*, which involves real-time extraction and online data processing at various trigger chain stages. The focus is on data scouting in the context of the muon system and the muon trigger chain. The underlying principle is that the closer the scouting is to the detector within this chain, the lesser the inherent biases, primarily because our knowledge of particle physics dictates the reconstruction and filtering algorithms embedded within the trigger system. In this thesis, we describe and exploit scouting muons directly at the detector's front-end electronics, where preprocessing, reconstruction or selection has yet to be applied. Such an approach ensures an unparalleled minimum bias in the extracted datasets. However, it is crucial to note that the methodologies employed here are based on prototypes. The complete realization of front-end scouting will only come to fruition with the upgrades of the CMS experiment for the High Luminosity LHC (HL-LHC). These upgrades, designed to complement the challenging data-taking environment of the HL-LHC, will significantly augment our scouting capabilities, as detailed in this chapter.

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [14–17] is located in Geneva, Switzerland, at CERN and is the current largest and most powerful particle accelerator in the world. Operational since 2010 and housed within a 27-kilometer tunnel—formerly occupied by the LEP (Large Electron-Positron) collider—this circular accelerator is the most advanced of its kind. It is engineered to accelerate either protons or heavy-ion beams, with the proton mode currently reaching center-of-mass energy ($\sqrt{s}$) of 13.6 TeV. This incredible capability is made possible by the LHC's superconducting magnets, which are cooled to 2.1 K using superfluid helium. This cooling system allows for a magnetic field of 8.3 T, which bends the beam and allows it to circulate at high energies within the ring.

The process of proton injection into the LHC is methodically carried out via a series of pre-existing accelerators. The accelerator chain, illustrated in Fig. 3.1, includes the Linac, the PSB (Proton Synchrotron Booster), the PS (Proton Synchrotron), and the SPS (Super Proton Synchrotron). This intricate cascade ensures an injection energy into the LHC of 450 GeV. Beam filling in the LHC typically takes two hours. To ensure optimal beam intensity and maximize the integrated luminosity achieved, beam circulation and data acquisition typically last for approximately 10 hours. After this, the beams are dumped, a new fill is initiated, and the process recommences.

The nominal beam structure includes 39 trains, each housing 72 bunches containing $N = 1.1 \times 10^{11}$ protons. The configuration, depicted in Fig. 3.2, operates at a crossing frequency of 40 MHz, equating to an inter-collision interval of 25 ns. The periodic time between collisions, called bunch crossing (BX), is a standardized temporal unit.

The LHC functions at a nominal instantaneous luminosity $\mathcal{L}_{\text{inst}} = 2 \times 10^{34}\,\text{cm}^{-2}\text{s}^{-1}$. This specific operating condition leads to an average pile-up of 50, representing the median number

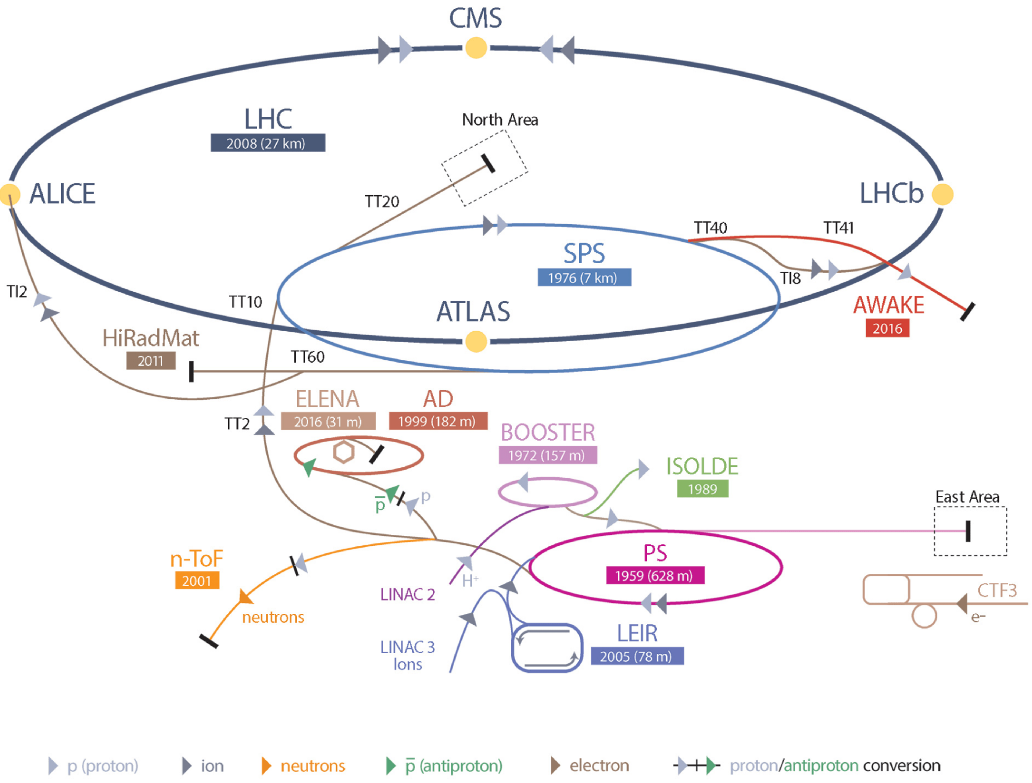


Figure 3.1: The CERN accelerator complex and experiments [13].

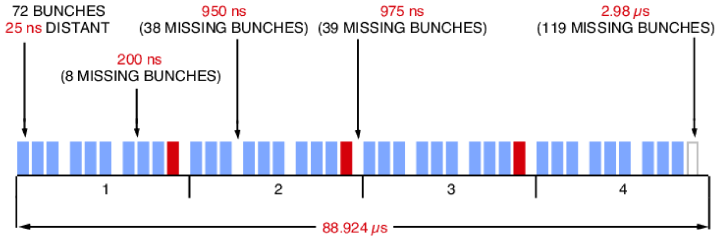


Figure 3.2: The LHC bunched beam structure.

of concurrent proton-proton interactions at each bunch crossing. The concept of pile-up is pivotal in understanding the complex dynamics of collisions within the LHC and plays an integral role in the detector's design and functionality.

Strategically positioned within the LHC are four primary experiments at designated interaction points. The ATLAS [18] and CMS [19] are general-purpose detectors. In contrast, ALICE [20] centers on heavy-ion physics and quark-gluon plasma investigations, and LHCb [21] probes the CP violation in b-physics.

The LHC's is not solely confined to its design but is manifest in its contributions to physics. Notably, the meticulous calibrations and adaptability of ATLAS and CMS were instrumental in the groundbreaking discovery of the Higgs boson in 2012. By leveraging the LHC's unparalleled energy potential, these experimental configurations continually drive pioneering explorations in particle physics.

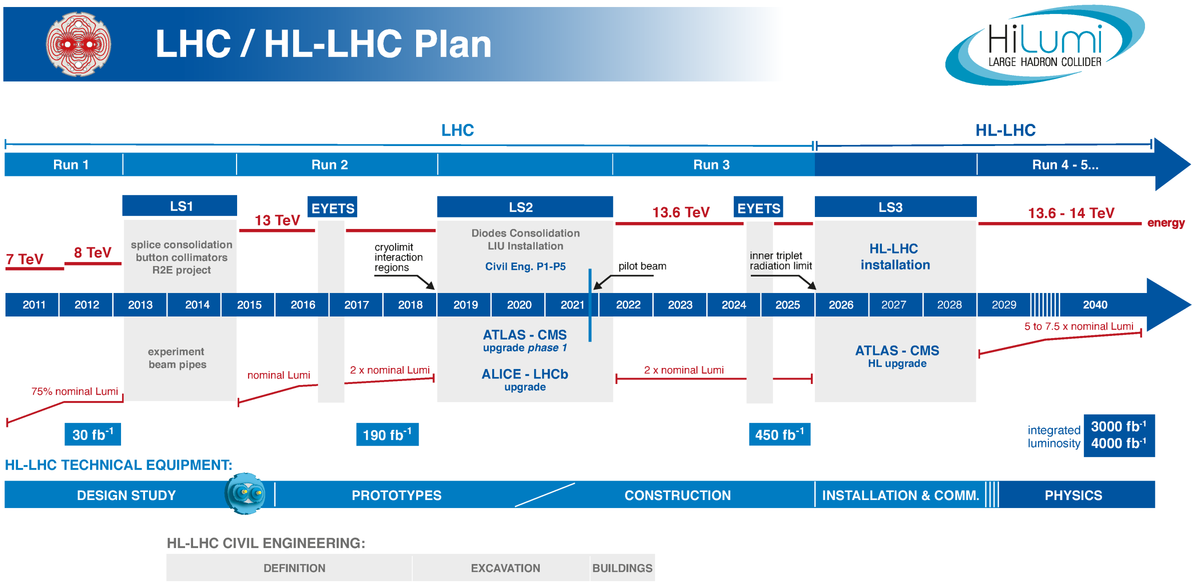


Figure 3.3: The LHC / High-Luminosity LHC plan [22], updated in 2022.

3.1.1 The High-Luminosity LHC upgrade

Both the ATLAS and CMS experiments are expected to collect nearly 300 fb^{-1} of data by the end of Run-3 in 2025. However, maintaining the current instantaneous luminosity of $2 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ beyond this point may not yield considerable statistical advancements in a feasible duration. For context, operating at the current nominal luminosity post-Run-3 would necessitate an extended timeframe, possibly exceeding a decade, to effectively halve the statistical uncertainty of specific measurements of scientific interest [23]. Recognizing these constraints, plans for the LHC post-2028 are geared towards initiating its high-luminosity phase, termed the High-Luminosity LHC (HL-LHC) [23]. With the upgraded accelerator, it is forecasted that the instantaneous luminosity will rise to $5 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, accompanied by an average pile-up of 140. Over the years it is planned to reach an instantaneous luminosity of $7 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, translating to an average of 200 pile-up events. Such a remarkable increase in luminosity is pivotal for ATLAS and CMS to target an integrated luminosity spanning between 3000 fb^{-1} and 4000 fb^{-1}. Achieving CERN's ambitious luminosity goals requires extensive upgrades to the foundational accelerator components, such as the injector systems, kickers, and cryogenic mechanisms. Concurrently, a comprehensive upgrade of all LHC experimental setups is imperative to ensure their detection sensitivities are fine-tuned, effectively countering the complications introduced by the amplified pile-up backgrounds. The strategic roadmap involves deploying the primary hardware for the HL-LHC during the Long Shutdown 3 (LS3) between 2026-2028, as illustrated in Fig. 3.3. The objective is to conclude the hardware commissioning upon machine restart in 2029 and ensure efficient operations until 2040. Such an expansive upgrade relies on innovative technologies: 11-12 T superconducting magnets, ultraprecise superconducting radio-frequency cavities for beam rotation, and 100-m-long zero-dissipation superconducting links. Enhanced luminosities will also necessitate advancements in vacuum, cryogenics, machine protection, collimation, diagnostics, and beam modeling. Novel beam-crossing schemes will be imperative to optimize the physics derived from the collisions. The HL-LHC presents vast potential. This facility will facilitate more in-depth investigations if deviations from the Standard Model or new particles emerge, however, even without such revelations, the ten-fold surge in data acquisition will propel the search for New Physics into unexplored realms.

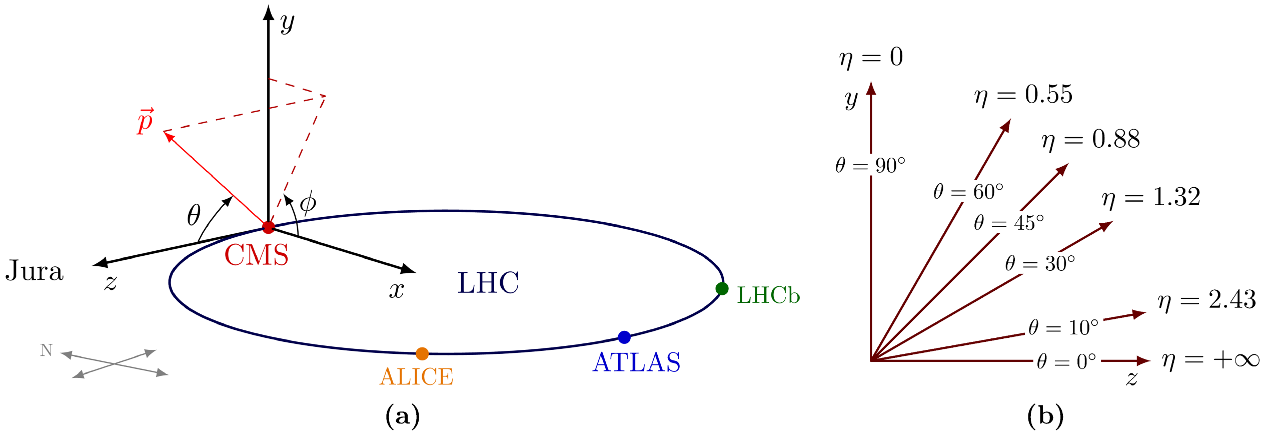


Figure 3.4: Left: the CMS coordinate system. Right: relation between the polar θ and the pseudorapidity η

3.2 The CMS experiment at the LHC

The Compact Muon Solenoid (CMS) [19] detector is situated at the LHC's interaction point five (P5) in Cessy, France. The detector, with its extensive dimensions of 21.6 m in length and 15 m in diameter, weighs approximately 14 thousand tons. As depicted in Fig. 3.5, the CMS features a central cylindrical barrel complemented by two endcaps, giving it a comprehensive design to capture diverse physics processes. A concise breakdown of the CMS detector's structure, from its innermost to its outermost components, is as follows:

- **Silicon tracker:** Comprised mainly of silicon, this component is sensitive to charged particles that pass through, producing a detectable electric signal. Charged particle trajectories can be reconstructed from these signals. The magnetic field induces curvature in these tracks, which gives information about the particle's momentum and charge [24].

- **Electromagnetic calorimeter (ECAL):** Made up of lead tungstate ($PbWO_4$) crystals, the ECAL measures the energy of photons and electrons. When such particles enter the crystals, they produce electromagnetic showers that generate light proportional to their energy, which is then measured. [25].

- **Hadronic calorimeter (HCAL):** Using brass and steel, this calorimeter measures the energy of hadrons. Incoming hadrons induce hadronic showers in these dense materials. Plastic scintillating tiles or fibers then capture the light produced by these showers, translating to an energy measurement for both charged and neutral hadrons [26].

- **Superconducting solenoid:** This component is constructed from the superconducting niobium-titanium (NbTi). It produces a strong magnetic field—3.8 T within its core and a 2 T field in its flux return yoke. The magnetic field influences the trajectories of charged particles, allowing for their properties to be inferred [27].

- **Muon chambers:** Typically made of aluminum or other lightweight metals and featuring gas detectors such as drift tubes, cathode strip chambers, and resistive plate chambers, these chambers are crucial for capturing muon tracks. Working in conjunction with the silicon tracker, they evaluate muon momentum, as muons are the primary particles capable of reaching this detector layer without being absorbed [28].

The CMS coordinate system, depicted in Fig. 3.4a, is right-handed and is defined as follows. Here, the x-axis directs towards the accelerator ring's center, the y-axis ascends upwards,

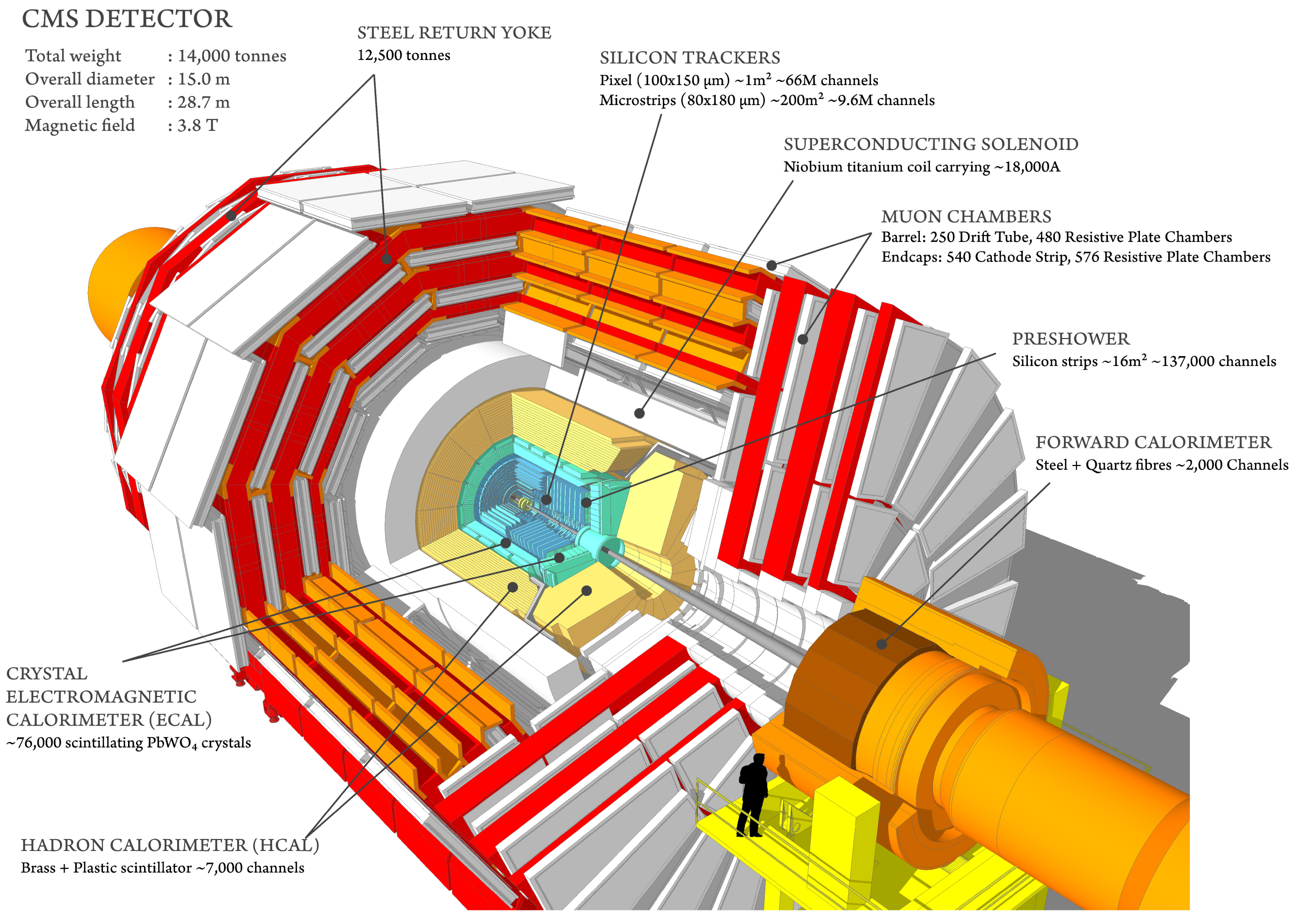

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~1m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO4 crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 3.5: Illustrative depiction of the CMS detector showcasing its diverse subdetectors.

while the z-axis runs parallel to both the beam pipe and the solenoid's magnetic field. Additionally, the system incorporates two angles: the polar angle θ, in reference to the z-axis and ranging from $0 \leq \theta \leq \pi$, and the azimuthal angle ϕ, relative to the y-axis in the $x - y$ plane, spanning from 0 to 2π. In most cases, the polar angle is replaced by the *pseudo-rapidity*, an ultra-relativistic approximation of the rapidity y, defined as $\eta = -\log\left(\tan\frac{\theta}{2}\right)$ (see Fig. 3.4b). CMS provides comprehensive azimuthal angle (ϕ) coverage and a pseudo-rapidity (η) range from -5.2 to $+5.2$, enabling it to monitor a broad spectrum of physics phenomena.

Given the upcoming transition to the HL-LHC, modifications are planned for the CMS detector. This evolved version is termed CMS Phase-2, succeeding the Phase-1 detector used during the LHC's Run-2 and Run-3. While the basic layered structure remains unchanged in Phase-2, changes to some subdetectors are intended to maintain efficient reconstruction performance under the increased pile-up conditions of the HL-LHC. The front-end and back-end electronic systems will also see upgrades to accommodate the expected growth in data transmission rates. A detailed discussion of these upgrades in the context of the High Luminosity phase will be addressed in Sec. 3.3.

3.2.1 The CMS muon system

From the very inception of the CMS experiment, precise and robust muon measurement has been central to its design and objectives. As the name suggests, the Compact Muon Solenoid places the detection of muons at the heart of its endeavors. The CMS muon system, located as the furthest external subdetector, is fundamental to the overarching goals of the experiment. An in-depth diagrammatic representation of the muon system can be viewed in Fig. 3.6.

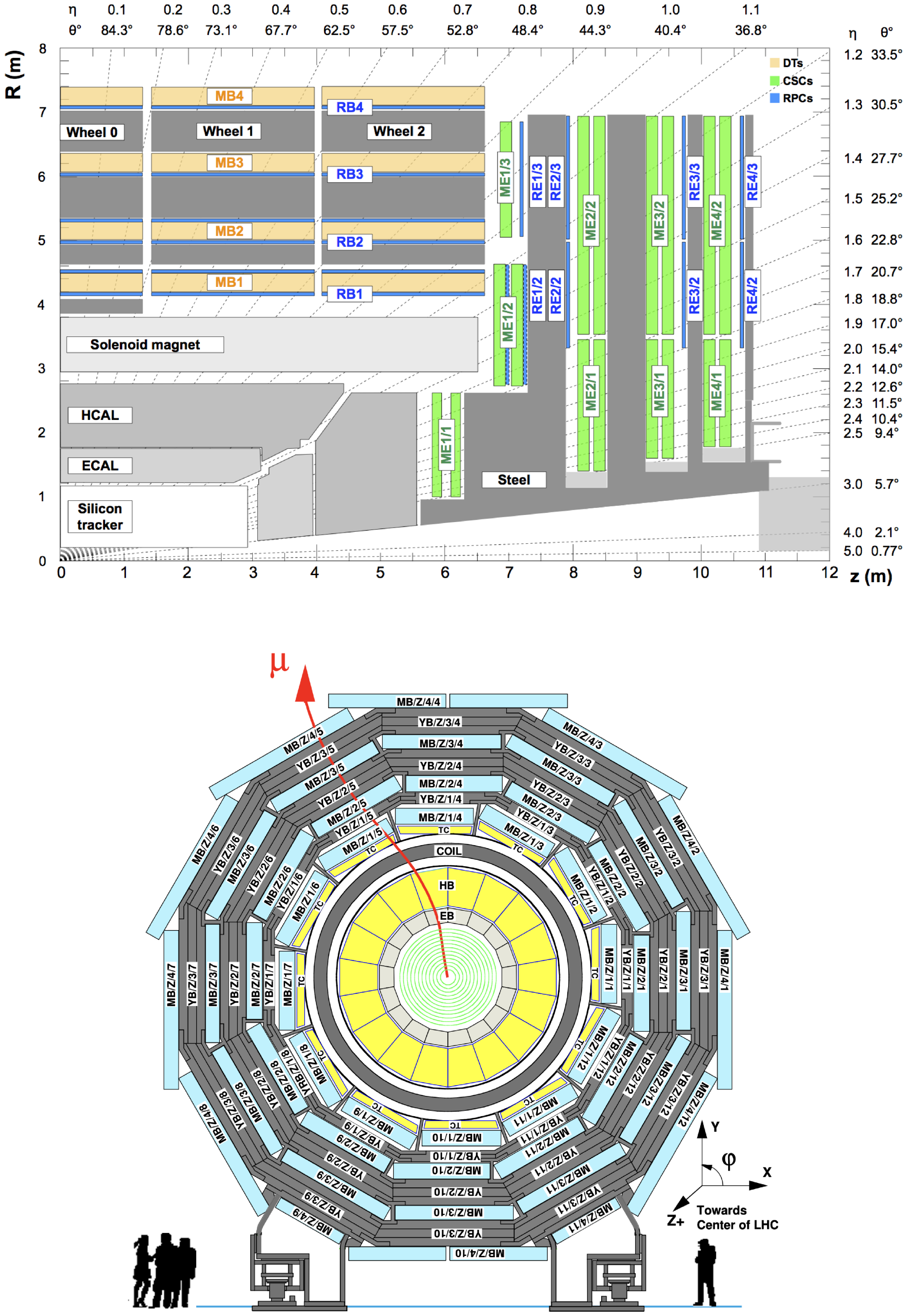


Figure 3.6: Longitudinal (top) and transverse (bottom) slices of the CMS detector.

The muon system is tasked with three primary functions:

1. Identifying muons.
2. Measuring their momentum accurately across a wide kinematic range.
3. Triggering events based on this momentum.

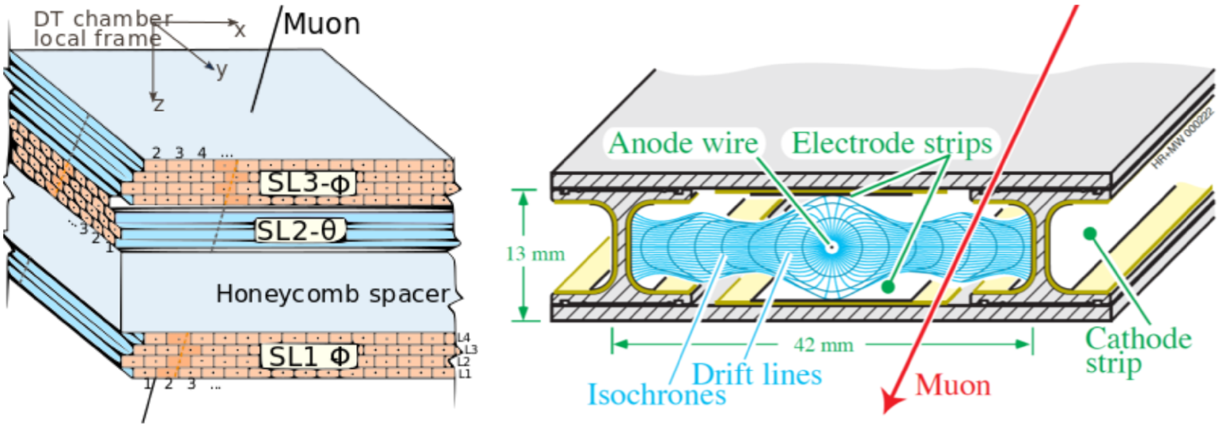


Figure 3.7: Left: CMS muon chamber schematic. Right: drift tube cell.

To fulfill these objectives, the solenoidal magnet and its flux-return yoke play an instrumental role. The high-field solenoidal magnet enables superior muon momentum resolution and triggering capabilities, while the flux-return yoke serves as a dual-purpose tool: it aids in momentum measurement and acts as a barrier against hadrons, thereby assisting in the precise identification of muons. In the barrel region of the CMS, where the background remains minimal, and the muon rate is low, drift tubes (DT) with rectangular cells are employed. The DT chambers cover the pseudorapidity region $|\eta| < 1.2$ and are systematically arranged into four concentric stations interspersed within the layers of the flux return plates. Conversely, the endcap regions of the CMS face higher muon rates and increased background levels. Due to non-uniform magnetic fields, Cathode Strip Chambers (CSC) are the detector of choice in these regions. Their quick response, granular design, and radiation resistance make them apt for tracking muons between $|\eta|$ values of 0.9 and 2.4. In each endcap of the detector, there are four stations of Cathode Strip Chambers (CSCs). The chambers are strategically positioned perpendicular to the beamline and set between the flux return plates. To enhance the triggering mechanism, Resistive Plate Chambers (RPC) are integrated throughout both the central and forward regions ($|\eta| < 1.8$). Despite their slightly lower spatial resolution than DTs and CSCs, RPCs stand out for their rapid response and exemplary time resolution. Their crucial contribution lies in offering an unambiguous BX assignment to muons. These RPCs are thus vital to both triggering and complementing DTs and CSCs in offline reconstruction.

3.2.1.1 The Drift Tube Chambers

The DT muon system is fundamentally composed of drift tube cells. A schematic view of a drift tube cell is depicted in the right panel of Fig. 3.7. Each cell measures 42 mm × 13 mm and houses a stainless steel anode wire. This wire, which varies in length from 2 to 3 meters, has a diameter of 50 µm. These cells are arranged adjacently, separated by aluminum beams configured in an “I” shape. The aluminum beams feature strips on both sides, which function as cathodes and are electrically isolated. The anode wires and cathodes operate under a voltage of +3600 V and −1200 V, respectively, producing the requisite electric field within the cell. As a muon track passes through, its distance from the wire is ascertained by the drift time of ionizing electrons. To further refine this measurement, two strips, biased at +1800 V, are mounted centrally on the inner surfaces of the aluminum planes for optimal field shaping. These adjustments strengthen time-to-distance linearity within the cell. A gas blend of 85% Ar and 15% CO_2 fills the tubes, providing optimal quenching properties. This results in a drift speed of approximately 55 µm/ns, yielding a maximal drift time of roughly 380 ns, which correlates to 15-16 BXs.

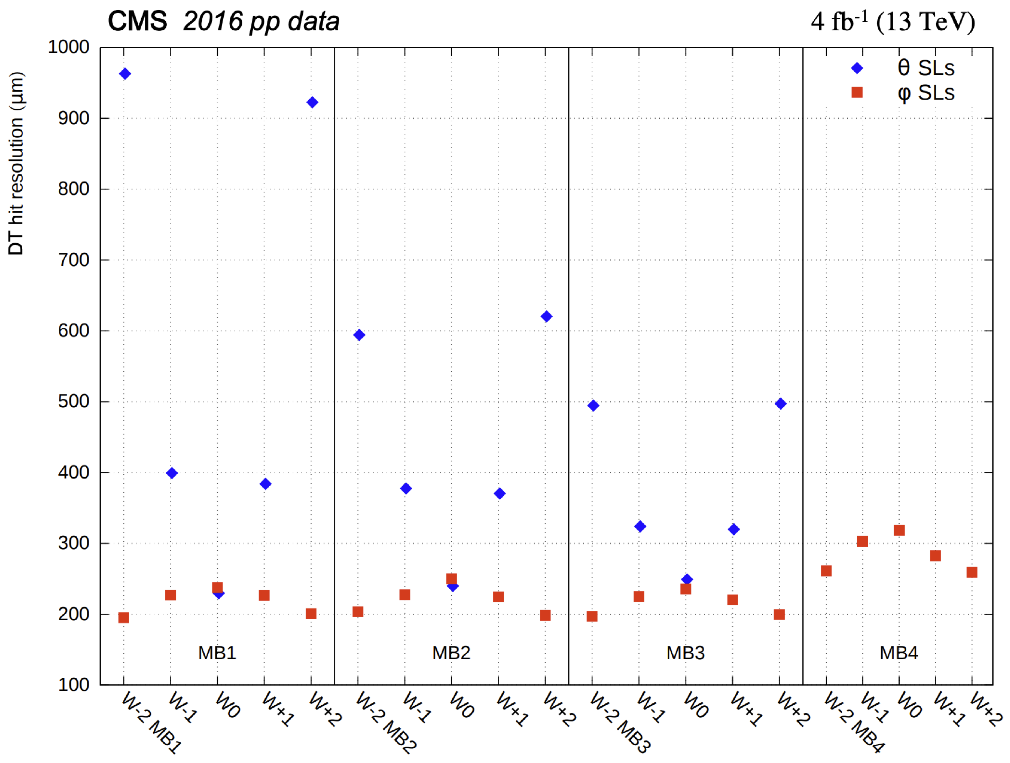


Figure 3.8: Reconstructed DT hit resolution for ϕ (red squares) and θ (blue diamonds) superlayers. Figure from [29].

The DT system's design is organized into five wheels along the z-direction, each spanning approximately 2.5 meters in width. These wheels are divided into 12 azimuthal sectors, each covering close to 30°. Drift tubes are methodically arranged within each wheel in four concentric stations, interspaced with the iron yoke and situated at varying distances from the interaction point. Every DT station hosts 12 chambers in each wheel, except for the outermost station, MB4. In MB4, the top and bottom sectors have an additional chamber each, totaling 14 chambers per wheel. As seen in the left panel of Fig. 3.7, the chambers' DT layers are assembled in groups of four layers, creating three *superlayers* (SL). Two superlayers record the muon's position in the $r - \phi$ bending plane, while the third determines the position along the z-coordinate. However, MB4 chambers have just two ϕ superlayers. In sum, the CMS detector houses 250 DT chambers.

The spatial resolution of the DT hits is detailed in Fig. 3.8 and varies by station, wheel, and wire orientation (ϕ and θ). The resolution for ϕ superlayers is noted to be under 250 μm for MB1, MB2, and MB3, and less than 300 μm for MB4. Meanwhile, θ superlayers' resolution ranges between approximately 250 and 600 μm, except in MB1's outer wheels. The symmetry in detector design ensures that both θ and ϕ superlayers exhibit symmetric behavior concerning the $z = 0$ plane. In the central wheel, where tracks predominantly intersect at right angles to all layers, θ and ϕ superlayers have an equivalent resolution. The increase in $|\eta|$ values for tracks emanating from the interaction region influences the θ and ϕ superlayers differently as one moves from wheel 0 towards the forward area. This difference is attributed to track inclination, ionization charge, and distance-drift time linearity. Notably, the ϕ superlayers of MB4 present a marginally poorer resolution due to the absence of a θ measurement, precluding any corrections for muon time-of-flight and signal propagation time along the wire.

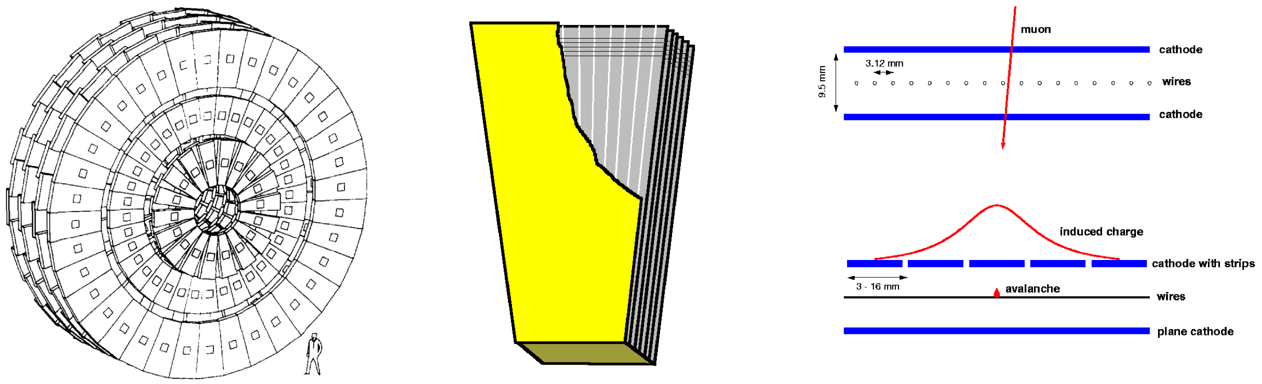


Figure 3.9: Left: layout of the CSC subsystem. Center: drawing of a single CSC. Left: schematic view of CSC chamber signal formation.

3.2.1.2 The Cathode Strip Chambers

In the endcap regions of the muon system, the anticipated high magnetic field and particle flux challenge the efficacy of drift tube detectors, particularly at large $|\eta|$ values. To address this, the CMS employs Cathode Strip Chambers (CSCs) as an alternative solution [30].

The endcap of the CMS houses four distinct station disks of CSCs labeled from ME1 to ME4. These chambers, inherently trapezoidal in shape, operate as multi-wire proportional chambers (MWPC) with segmented cathodes. Their design enables the extraction of precise spatial and temporal information, even in environments characterized by inhomogeneous magnetic fields and high particle rates. This precision is attributed to a limited drift length that facilitates rapid signal acquisition. When a charged particle traverses these layers, several adjacent cathode strips register the signal. Given the radial arrangement of these strips, charge interpolation facilitates a detailed measurement of the ϕ-coordinate. Concurrently, the wire signal analysis aids in determining the orthogonal r-coordinate, and its swift response is invaluable for trigger applications.

Structurally, each CSC comprises six wire layers interspersed between cathode panels. While the wire spacing remains relatively constant, the cathode panels undergo a milling process to accommodate six radially-arranged strip panels, with each gas gap hosting a single strip plane. Consequently, every chamber offers six measurements each for the ϕ-coordinate (using strips) and the r-coordinate (using wires). In the spatial arrangement, ME1 features three concentric rings of CSCs, with each subsequent ring having a greater radius. The other three stations, however, consist of two such rings. Notably, the entirety of ME1's chambers, except for the outermost one, overlap in ϕ. This ensures the formation of rings devoid of any azimuthal dead zones. In contrast, stations 2 through 4 incorporate 36 chambers spanning $10°$ in ϕ for the external ring and 18 chambers covering $20°$ for the internal ring situated proximal to the beam pipe. These are then systematically organized into four disks of concentric rings installed between the endcap's iron yokes.

The spatial resolution of CSCs is contingent on where a muon intersects a strip. Optimal resolution is achieved when a muon intersects near the strip's edge rather than its center. This is because a more significant portion of the induced charge is distributed between the strip in question and its adjacent counterpart, enhancing the accuracy of the charge distribution's center determination. As per design specifications, the CSC system's spatial resolutions are delineated as 75 µm for the chambers located in ME1's inner ring and 150 µm for the remainder.

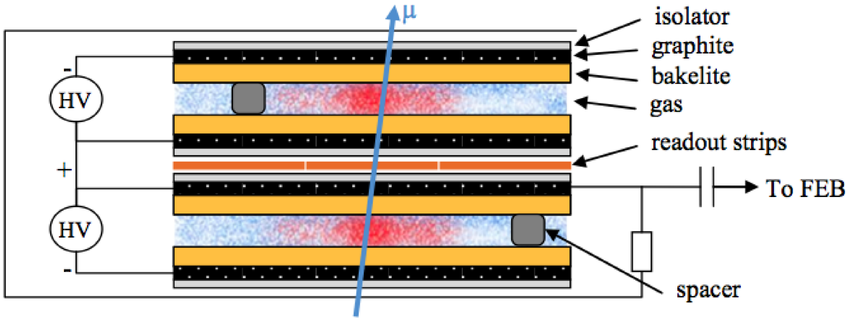


Figure 3.10: Schematic view of a CMS double-gap RPC.

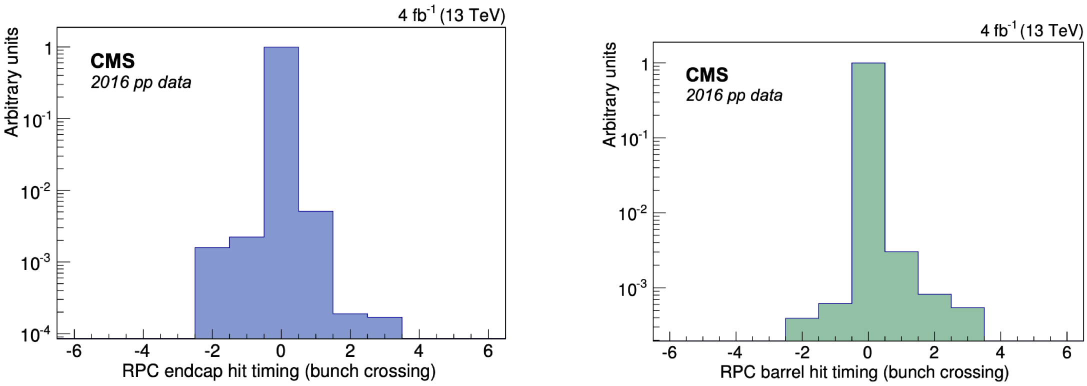


Figure 3.11: The bunch crossing distribution from reconstructed RPC hits in one endcap (left) and in the barrel (right). Figures from [29].

3.2.1.3 The Resistive Plate Chambers

In the muon spectrometer, the use of Resistive Plate Chambers (RPCs) [31] spans both the barrel and the endcap regions, thereby supplementing the DT and CSC systems. This ensures enhanced robustness and redundancy. While RPCs possess a relatively coarse spatial resolution, they demonstrate a rapid time response akin to scintillators. Coupled with high segmentation, they allow for muon momentum measurements at trigger time and facilitate a definitive assignment of the BX.

Structurally, an RPC consists of two planes constructed from high-resistivity material (Bakelite) separated by a 2 mm gap. This gap is filled with freon ($C_2H_2F_4$) and isobutane ($i{-}C_4H_{10}$). As depicted in Fig. 3.10, the exterior of these planes is coated with graphite, serving as the cathode when exposed to a high voltage of 9.5 kV. When a particle traverses this structure, it triggers an electron avalanche. This avalanche, in turn, induces a signal in the insulated aluminum strips outside the graphite cathodes, prepped for read-out. CMS employs a double-gap RPC design: two gas gaps are read by a central set of strips. This approach amplifies the signal on the read-out strip, effectively capturing the aggregate of individual gap signals. In the barrel region, the read-out comprises rectangular strips varying between $1-4$ cm in width and $30-130$ cm in length. Conversely, the endcaps feature trapezoidal strips, covering approximately the range $\Delta\phi = 5-6$ and $\Delta\eta = 0.1$. Within the barrel region, two RPC stations are joined to each flank of the two innermost DT stations in a sector, mirroring DT segmentation. However, only a singular RPC is affixed to the inner facet of the third and fourth DT stations. This configuration is pivotal for accurately detecting low p_T range muons within the barrel trigger. Such muons typically intersect multiple RPC layers before halting within the iron yoke.

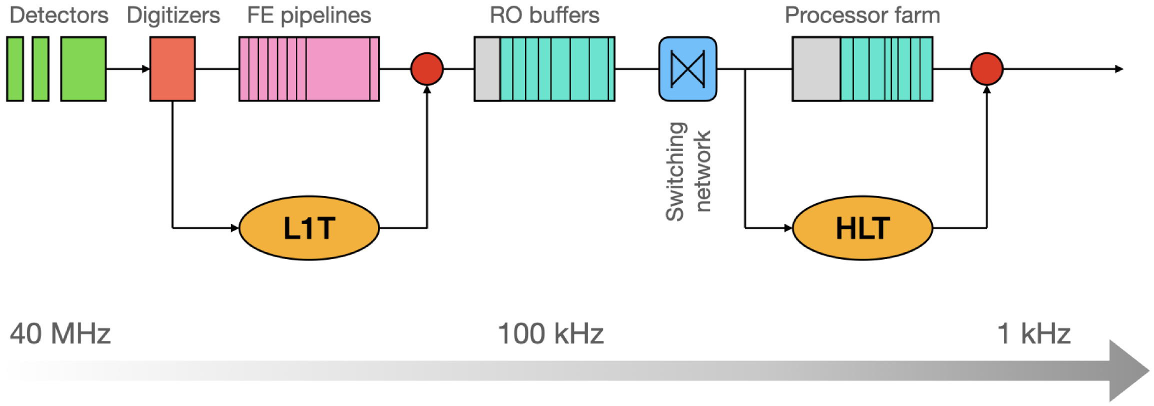


Figure 3.12: Schematic representation of the CMS trigger system.

RPCs, when operating in avalanche mode, have exhibited an intrinsic time resolution close to 2 ns [32]. However, this resolution must be considered in conjunction with other temporal uncertainties. For instance, time propagation along the strip introduces an approximate uncertainty of 2 ns. Likewise, slight variances in the electronics and cable lengths introduce an additional jitter of about $1-2$ ns. When these factors are compounded quadratically, the cumulative time resolution hovers around 3 ns [33], significantly lower than the 25 ns timing window associated with CMS's RPC data acquisition system. A detailed view of the RPC's performance can be discerned from Fig. 3.11. The BX distribution of RPC hits corresponding with global muons in an endcap is showcased on the figure's left. Instead, the right side illuminates a similar distribution within the barrel. In the context of Fig. 3.11, each bin epitomizes the 25 ns bunch separation characteristic of the LHC, with bin 0 signifying the L1T timestamp. Within this representation, approximately 0.5% of RPC hits fall outside bin 0.

3.2.2 The CMS trigger system

The CMS Trigger System is an indispensable tool in the data selection process, ensuring the storage of only the most pertinent data for subsequent analyses. At the LHC, bunches of protons collide every 25 ns. Depending on the instantaneous luminosity delivered by the LHC, multiple collisions might arise during each proton bunch crossing. The challenge arises from the large volume of data produced—storing and processing data at such a magnitude is untenable. To manage this, the trigger system is implemented as the first step in the physics event selection process, reducing the 40 MHz rate down to an offline storage rate of roughly 1 kHz.

This sophisticated system comprises two principal stages: the Level-1 Trigger (L1T) [34] and the High-Level Trigger (HLT) [35]. The experiment's overarching physics objectives inherently guide the design and functionality of the trigger system. Consequently, the system's selection criteria are tuned to efficiently reconstruct various physics objects (such as muons, electrons, gammas, jets, missing energy, and more). The objective is to achieve the necessary rate reduction without diminishing the yield of "interesting" events.

The L1T is primarily constructed using custom electronics, tasked with reducing the event acceptance rate to a maximum of 100 kHz. It utilizes coarse information sourced from muon detectors and calorimeters. In contrast, the HLT is software-based, running on a dedicated farm of commercial processors. Since the L1T already significantly reduces the bandwidth, the HLT can afford more extensive processing times for each event. Events are processed

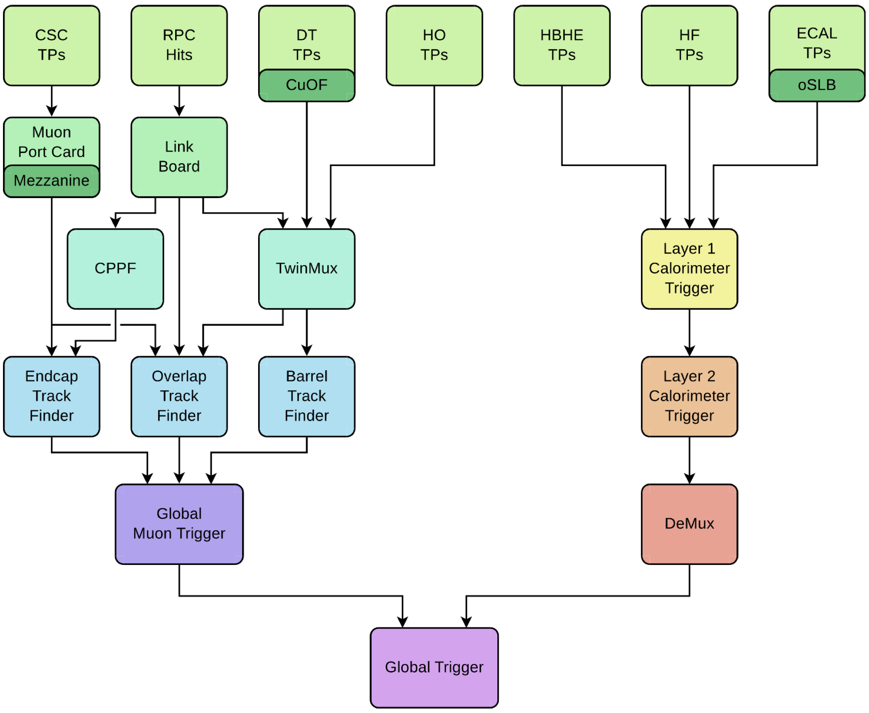


Figure 3.13: Schematic representation of the Phase-1 Level-1 trigger of CMS. CSC: Cathode Strip Chambers; RPC: Resistive Plate Chambers; DT: Drift Tubes; HO: Hadronic Calorimeter – Outer; HF: Hadronic Calorimeter - Forward; ECAL: Electromagnetic Calorimeter; TPs: Trigger Primitives; oSLB: optical Synchronization and Link Board; CuOF: Copper to Optical Fibre; CPPF: Concentration, Pre-processing and Fan-out system.

concurrently by distinct machines within the HLT, enabling comprehensive access to data from all subdetectors, including the silicon inner tracker's data. Such access facilitates an additional event rate reduction by factors ranging from 10^2 to 10^3. In Fig. 3.12, we display a diagram of the CMS trigger system with the event rate that characterizes each step in the trigger chain.

3.2.2.1 The Level-1 Trigger

The L1T [34], depicted in Fig. 3.13, is made of local, regional, and global components:

- **Local Triggers or Trigger Primitive Generators (TPG):** These rely on energy deposits in calorimeter trigger towers, track segments, and hit patterns in muon chambers.
- **Regional Triggers:** These merge information from the TPGs. Based on their spatial positioning, these triggers implement pattern logic to assess and rank trigger objects, such as muon candidates. Their ranking hinges on parameters like energy, momentum, and the quality of their measurements.
- **Global Triggers:** The Global Triggers consolidate vital calorimeter and muon data throughout the CMS. The Global Trigger determines whether to retain an event for further analysis or pass it to the HLT. This decision is based on rigorous algorithmic evaluations, the subdetectors' operational status, and the DAQ system's status.

The real-time processing challenges posed by the L1T necessitate rapid evaluations of every bunch crossing (BX). Given the limited depth of the FE buffers, the system needs to perform non-trivial algorithmic assessments quickly, while FIFO memories retain data from the subdetectors. The trigger logic, segmenting its evaluations into steps, is pipelined such that it can accept data from a new BX every 25 ns. To achieve this, custom-programmable hardware, such as Field Programmable Gate Arrays (FPGA) and Programmable Lookup Tables (LUTs), is crucial. This infrastructure culminates in an event acceptance decision within a bounded timeframe dictated by the FIFO's storage capacity, roughly corresponding to 4 µs.

The L1T system experienced substantial upgrades between LHC Run-1 and Run-2. In Run-2, the L1T faced an event rate increase by nearly a factor of six, necessitating comprehensive enhancements to the L1T process. With these updates, the L1T effectively processed approximately 5 TB/s of data, reducing the detector read-out rate from 40 MHz to a fixed 100 kHz.

Muon Trigger The Muon Trigger, divided into three subsystems targeting distinct η ranges, is central to muon tracking across the detector. As depicted in Fig. 3.13, Trigger Primitives (TP) from the CSCs are routed to the Endcap Muon Track Finder (EMTF) and the Overlap Muon Track Finder (OMTF) via a mezzanine on the muon port card. Endcap RPC hits are channeled via the link board to the Concentrator Pre-Processor and Fan-out (CPPF) card, while barrel RPC hits approach the TwinMux concentrator card. DT trigger primitives reach the TwinMux card through a copper-to-optical fiber (CuOF) mezzanine. The TwinMux, in turn, crafts *superprimitives*, amalgamating the precise spatial resolution of DT trigger segments with the optimal timing characteristics of RPC hits, thereby refining the efficiency and data quality for subsequent phases. Notably, the EMTF absorbs RPC hits via the CPPF card. The OMTF and CSC hits also take into account DT and RPC hits via the CPPF and TwinMux, which also provide the Barrel Muon Track Finder (BMTF) with DT and RPC hits. In its final step, the Global Muon Trigger (GMT) arranges muons, discards duplicates and transmits the top eight muon candidates to the Global Trigger.

Calorimeter Trigger This trigger processes the energy deposited in calorimeter towers. A two-tier structure equipped with time-multiplexing capabilities ensures proficient energy sum calculations.

The global trigger (GT) functions on a "trigger menu", a spectrum of selection criteria from basic single-object p_{T} thresholds to intricate object correlations. The GT can perform up to 512 selection algorithms in parallel, and it takes all the results into account to decide whether to send an acceptance signal, named Level-1 Accept (L1A), based on a global OR condition. The L1A decision is sent to the sub-detectors through the Timing, Trigger, and Control (TTC) system. The L1 Trigger must evaluate every bunch crossing, with a maximum latency of 4 microseconds between a particular bunch crossing and the trigger decision distribution. Therefore, pipelined processing is necessary for near-deadtime-free operation.

3.2.2.2 The High-Level Trigger

The HLT [35] is the final tier in the two-level CMS trigger system. It plays a pivotal role in further refining the event rate, bringing it down from the 100 kHz L1T rate to a more manageable ∼1 kHz—aligning with the requirements of the storage system. This translates to a rate reduction factor of 100, which the HLT achieves by meticulously selecting events with high physics significance and efficiently discarding less pertinent ones. While the L1T is based on FPGAs and ASICs to run fast and relatively simple trigger algorithms, the HLT is software-implemented, running on a farm of commercial computers that includes about 16 000 CPU cores. This expansive setup facilitates the deployment of intricate software similar to the offline reconstruction one. However, the software is optimized to meet online selection's real-time processing demands. Fig. 3.12 offers a schematic representation of the CMS DAQ system. The front-end electronics of the detector are concurrently read by the Front-End System (FES), which arranges and preserves the data in sequentially organized buffers. These buffers are bridged to the HLT farm processors through a significant switch network dubbed the Builder Network. Data progression involves its transfer from the Front End Drivers (FEDs) to the Front End Readout Links (FRLs)—the latter is proficient in acquiring information from two distinct FEDs. Subsequently, multiple FRLs transmit this data to the Event Builder system, tasked with building a complete event. After data assembly, events are preliminarily reconstructed and then sent to CMS surface facilities. From there, each event is directed to the Event Filter. Here, sophisticated HLT algorithms and select Data Quality Monitoring operations are executed. Upon filtering, the data is segregated into specific online streams, with the content contingent on varying trigger configurations. The final data repository is a local storage framework before it is migrated to CERN's expansive storage infrastructure. Two systems aid data movement: the Event Manager for trajectory within the DAQ, and the Control and Monitor System for oversight and monitoring of components. The HLT processes are organized into 'paths', each representing a step-by-step process of selecting specific physics objects or combinations through reconstruction and filtering. These paths consist of blocks of producers and filters, systematically organized by complexity. Preliminary, rapid algorithms are prioritized, and their resultant products are subsequently filtered. Any filter failure preempts the execution of subsequent, more computationally intensive algorithms. The final HLT decision is the logical OR of all the trigger paths within the menu.

3.3 The CMS Phase-2 upgrades

In preparation for the High-Luminosity LHC operational period, the CMS detector is undergoing extensive consolidations and modifications, as detailed in works [36–41]. The anticipated increase in instantaneous luminosity and pile-up will lead to a very high particle multiplicity. The significant hadronic background, with an estimated 200 collisions occurring concurrently per bunch crossing, highlights the need for vital enhancements to the L1T system to ensure consistent performance. The Phase-2 L1T system upgrades aim to enhance physics selectivity. The upgraded trigger and DAQ system retain their two-tiered architecture but with significant modifications. For example, the L1T's rate cap will increase to 750 kHz, and its latency will extend to 12.5 µs. Tracker data will be incorporated during the L1 trigger phase and allow for advanced event reconstruction. Ongoing discussions about deploying advanced algorithms, potentially harnessing particle-flow reconstruction or Machine Learning methodologies, are also underway. In addition to these upgrades, there is anticipation for a 40 MHz scouting system. This system will extract trigger primitives generated by sub-detectors and identify trigger objects at various stages of the trigger hierarchy. A detailed examination of data scouting is reserved for Sec. 3.4.

3.3.1 The CMS detector upgrades

For better physics signature extraction, higher granularity detectors and robust read-out electronics are required. The CMS collaboration is planning to replace the Strip and Pixel tracking detectors with an Inner Tracker featuring small-size pixel sensors and an Outer Tracker equipped with strip and macro pixel sensors, extending their coverage to $|\eta| = 4.0$. The Outer Tracker will implement stacked strip modules, reducing the hit multiplicity and allowing track candidates for the trigger (L1 tracks) to be reconstructed up to $|\eta| = 2.4$. The read-out electronics for the barrel calorimeters will be replaced to achieve finer granularity and provide timing information. The endcap calorimeters will be replaced by the high-granularity calorimeter (HGCAL),

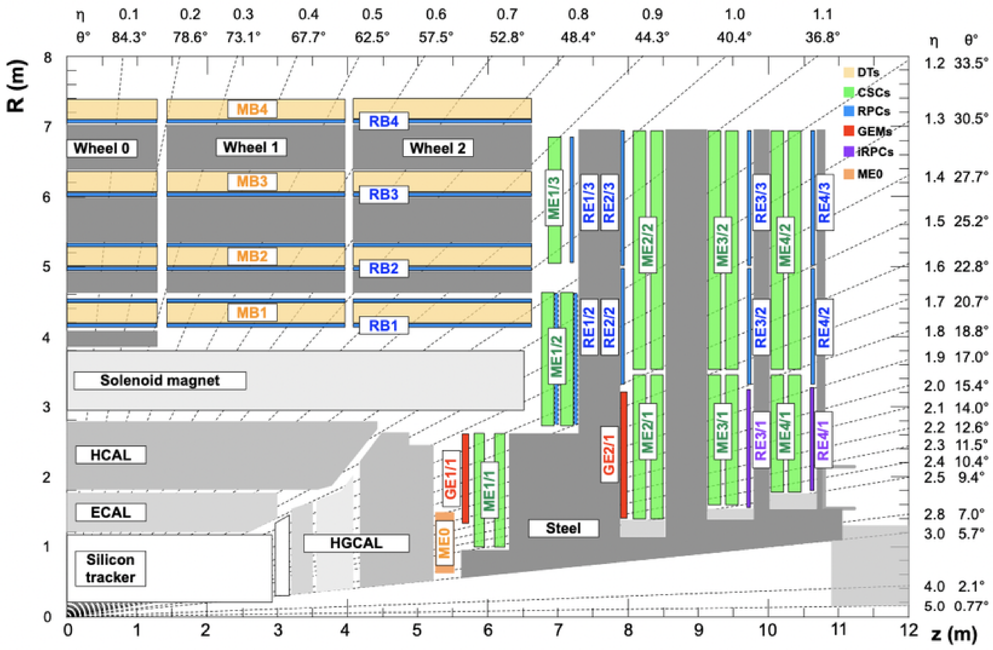


Figure 3.14: Longitudinal view of a quadrant of the CMS Phase-2 muon system. Different colors in the figure refer to different sub-detectors: DT (light orange), RPC (light blue), CSC (green), iRPC (purple), GE (red), and ME0 (orange).

implementing over 6 million read-out channels. This sampling calorimeter will provide shower separation and identification adapted to harsher conditions in the forward region of the detector.

3.3.1.1 Muon system upgrades

The Phase-2 upgrade for the DT system focuses on updating the On-detector Boards for DT (OBDT) [42, 43]e lectronics. DT chambers will remain unchanged for Phase-2, but current components cannot handle the increased L1T rate and radiation resistance needed for HL-LHC conditions. In this new setup, the On-detector Boards for Drift Tubes (OBDT) [42, 43] will send time digitization (TDC) data directly to a new backend electronics system in the service cavern, named the Barrel Muon Trigger Layer-1 (BMTL1). This BMTL1 system, with its modern commercial FPGAs, will be responsible for event building and trigger primitive generation. This improves the BX identification and spatial resolution and reduces the chance of seeing multiple trigger segments for a single crossing muon.

Regarding the existing RPC system, its off-detector electronics, the Link System, will be replaced to ensure a consistent read-out throughout the HL-LHC timeframe. This change includes increasing the read-out frequency from 40 MHz to 640 MHz. Each RPC hit sent to the muon track finders will now have more detailed time information. Additionally, to enhance the Muon System's capability in the forward region, new RPCs (iRPC), chambers will be added in stations 3 and 4 (RE3/1 and RE4/1), pushing the RPC pseudorapidity coverage to $|\eta| < 2.4$. The detector has been optimized to handle the high rates expected in RE3/1 and RE4/1.

For the CSCs, upgrades involve changing the on-chamber cathode boards on the inner rings of chambers in the range of $1.6 < |\eta| < 2.4$. This helps them cope with the increased trigger and output data rates. Additionally, most on-chamber anode boards will get new FPGA mezzanine boards to adjust to the higher L1T latency. Boards that handle trigger and read-out data will also be updated to accommodate the higher data rates.

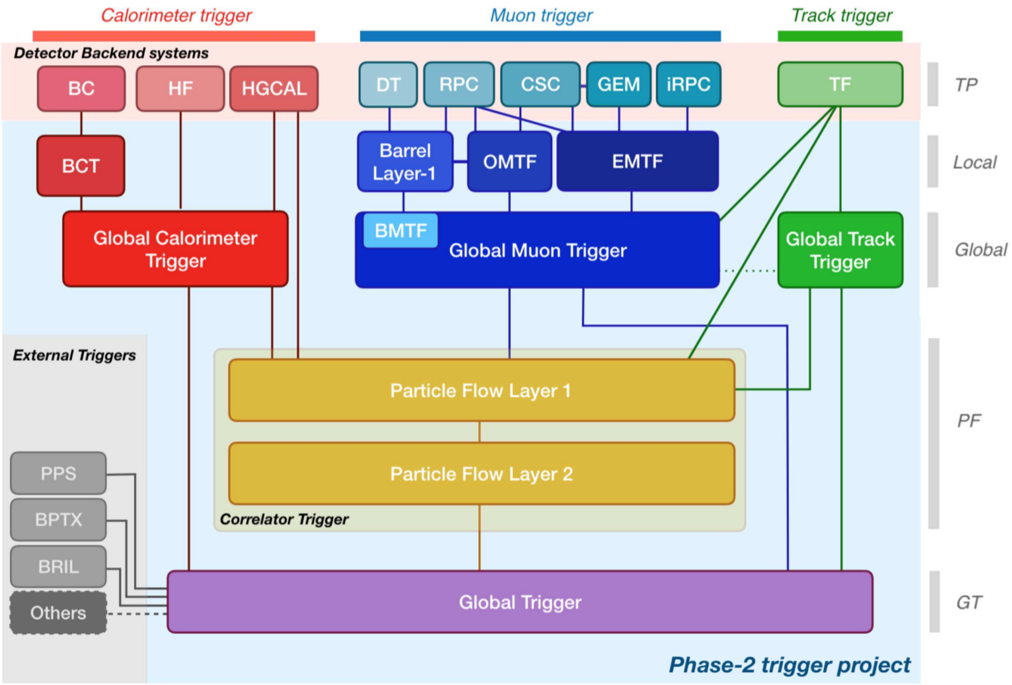


Figure 3.15: Functional diagram of the CMS L1 Phase-2 upgraded trigger design. The calorimeter trigger is composed of the barrel calorimeter trigger (BCT) and the global calorimeter trigger (GCT), receiving inputs from the barrel (BC), endcap (HGCAL) and forward (HF) calorimeters. The muon trigger is composed of a barrel layer-1 and muon track finder processors: BMTF, OMTF and EMTF, for each region (barrel, overlap and endcap, respectively), and receiving inputs from drift tubes (DT), resistive plate chambers (RPC), cathode strip chambers (CSC), and gas electron multipliers (GEM). The global muon trigger (GMT) matches muons with tracks from the track finder (TF). The event vertex is reconstructed in the global track trigger (GTT), and the correlator trigger (CT) implements the particle-flow reconstruction. The global trigger (GT) issues the final L1 trigger decision.

Lastly, the forward region $1.6 < |\eta| < 2.8$ will add Gas Electron Multiplier (GEM) chambers. These chambers provide precise measurements of the muon's bending angle in the first two stations and control the muon trigger rate. Adding these chambers will also improve the system's efficiency and resilience. The GEM foils, selected for the CMS forward region, are organized into chambers. These chambers are then grouped into superchambers, placed in areas GE1/1, GE2/1, and ME0, as illustrated in Fig 3.14.

3.3.2 The Level-1 trigger Phase-2 upgrade

The Phase-2 upgrade of the L1 trigger system is designed not only to maintain the efficiency of the signal selection to the level of the Phase-1 performance but also to enhance or enable the selection of any possible New Physics manifestations that could lead to unconventional signatures [44]. High-precision measurements of physics processes will benefit from the extension of the available phase space, such as enhanced trigger coverage in the forward region of the detector or the ability to exploit fully hadronic final states. Furthermore, a longer latency will enable higher-level object reconstruction and identification, as well as the evaluation of complex global event quantities and correlation variables to optimize physics selectivity. Implementing sophisticated algorithms using traditional reconstruction techniques or machine-learning-based approaches can now be contemplated. In addition, the design includes a dedicated scouting system streaming data from key parts of the trigger at 40 MHz via FPGAs into HPC resources.

The scouting system provides unprecedented flexibility for parasitic debugging and commissioning new ideas and is also being investigated for physics channels that are impossible through traditional triggering techniques.

The conceptual design of the Phase-2 Level-1 Trigger system is the result of several considerations: the design has to efficiently distribute and process the input trigger primitives, provision appropriate resources and interconnections and retain enough headroom for future flexibility and robustness to evolve with running conditions and physics needs. The high-level functional diagram of the system is shown in Fig. 3.15. The system features four distinct trigger processing paths with a calorimeter, muon, track, and particle-flow trigger. This division reflects the need to generate complementary types of trigger objects to achieve the best physics selectivity. The key design feature is the implementation of a correlator trigger combining all detector information and running sophisticated algorithms. The final trigger decision is performed at the global trigger level.

The trigger algorithms are designed to use tracking information to reach near offline resolution. The availability of fully reconstructed tracks translates into sharper turn-on efficiency curves. The trigger object reconstruction performance is close to offline physics object reconstruction with optimized response and resilience to high pileup conditions. Dedicated trigger algorithms can be implemented to select specific physics topologies, including final states with displaced objects coming from New Physics signatures.

3.4 Data scouting at the CMS experiment

The CMS experiment at the LHC produces a vast amount of data. To handle this massive throughput, the CMS trigger system selectively processes and filters data based on established particle physics knowledge. However, while invaluable, this system inherently introduces biases into the dataset and often omits significant amounts of statistics vital for observing rare decay channels. Data scouting emerges as a promising approach, providing a potential pathway to both enhance statistical reach and reduce dataset biases. Historically, data scouting has been used within the CMS experiment to augment traditional analyses, especially in studying rare events. The essence of data scouting is to utilize objects within the trigger chain, extract and process these objects online to ensure efficient storage. By representing events in a more compact format, it becomes feasible to store many more events than typically possible. The approach focuses on obtaining objects with a reduced level of detail, trading off some resolution for greater statistics. While this technique has been prevalent at the HLT level, new opportunities are presented with the LHC's High-Luminosity upgrade. Specifically, the possibility of data scouting at the Level-1 trigger is emerging. This new strategy aims to extract L1 objects at different stages of the L1 trigger chain, with options for direct storage or online processing using heterogeneous computing methods, including FPGAs, GPUs, and big-data tools. A demonstrator is currently used in Run-3 to collect data on scouted muons at different points of the L1 trigger chain. This data will provide insights into the potential physics research possible with this information. It is important to note that data extraction closer to the detector is less biased because it does not undergo trigger algorithms that could skew trigger objects. Furthermore, with the Phase-2 front-end upgrades of the DTs, there is potential for direct scouting at the front-ends at a frequency of 40 MHz. This approach, though promising, comes with the challenge of preprocessing the raw data to convert it into meaningful physical quantities. The emphasis on DT scouting in this thesis arises from its intrinsic unbiased nature. Even though this approach possesses the most significant challenges in data extraction and processing, its potential for delivering unbiased, unfiltered datasets makes it invaluable for comprehensive physics analyses. This thesis introduces the concept of DT scouting combined with advanced heterogeneous computing for preprocessing. While the focus is on the CMS experiment, the current demonstrator (see Chap. 4) operates on a smaller-scale mock-up experiment yet still presents complexities similar to those in CMS. For the CMS experiment, exploring data scouting levels is crucial for understanding its functionality. This section provides an analysis of the three levels of data scouting, their implementation, and solutions to the challenges faced.

3.4.1 Data scouting at the High-Level Trigger

The High-Level Trigger is central to the CMS experiment's data scouting strategy, aiming to optimize LHC data analysis [45–47]. The HLT acts like a precise filter by using reconstruction algorithms that replicate offline techniques. These include a version of the Particle Flow (PF) algorithm involving track finding, clustering of calorimeter energy deposits, and identification of muons, electrons, photons, and hadrons.

Through the HLT, events are reconstructed in real time. A moderate selection is applied to these reconstructed physics objects. Events meeting this criterion see only their HLT-reconstructed physics objects saved to disk while the corresponding raw data is discarded. This streamlined dataset then becomes the foundation for New Physics searches.

Several challenges emphasize the value of this HLT-centric scouting approach:

1. **Bandwidth Limitations:** The CMS data acquisition system has a set bandwidth. This limitation affects the data that can be temporarily stored at LHC Point 5 and the transmission between Point 5 and the primary CERN CMS computing facility.

2. **Reconstruction Time:** All recorded physics data should be reconstructed and available within 48 hours of collection.

3. **Storage Constraints:** Physical storage (tape and disk) is limited. The financial aspect of acquiring more storage also plays a role.

4. **Trigger Decision Time:** The HLT has a tight window, just a few hundred milliseconds, to make a trigger decision.

Given these constraints and the vast difference between the collision rate (up to 40 MHz) and the recording rate (approximately 1 kHz), the HLT scouting method becomes vital. The advantages of this approach over traditional strategies include:

- **Compact Event Format:** Events require significantly less disk space, 100 to 1000 times smaller than standard raw data.

- **Online Reconstruction:** All event processing occurs online, eliminating the need for offline reconstruction.

- **Concurrent Operation:** Scouting trigger paths can run alongside standard HLT paths, preserving physics objects even if the standard paths reject them.

HLT scouting has since underpinned various impactful analyses. Recent studies on dijet phenomena have benefited significantly from this data acquisition technique [48–50]. A particularly notable achievement facilitated by HLT scouting was the observation of the elusive decay channel of the η meson into four muons in early 2023 [51].

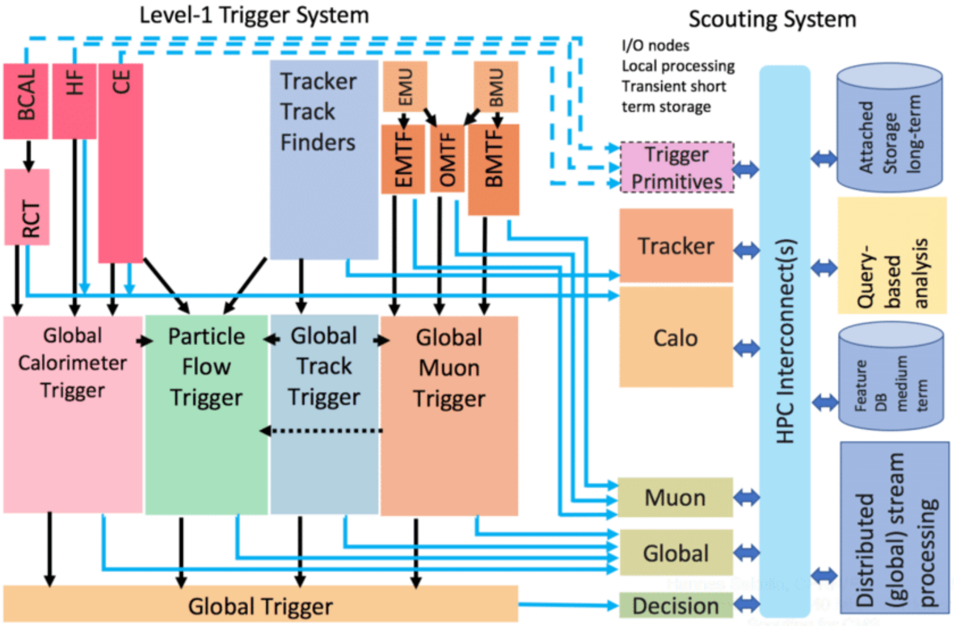


Figure 3.16: The CMS Level-1 trigger system for Phase-2 (left) and the proposed Scouting System (right). BCAL: Barrel Calorimeter; HF: Hadronic Calorimeter Forward; CE: Calorimeter Endcap (high granularity); RCT: Regional Calorimeter Trigger; EMU/BMU: Endcap/Barrel Muon System; EMTF/OMTF/BMTF: Endcap / Overlap / Barrel Muon Track Finder. Figure from [52].

3.4.2 Data scouting at the Level-1 Trigger

Scouting the HLT has historically played a significant role in facilitating specific physics analyses in the CMS experiment. However, with the technological strides associated with the High Luminosity and the subsequent Phase-2 upgrades enable scouting directly at the Level-1 trigger [52, 53], facilitating data capture at an unprecedented rate of 40 MHz. Unlike HLT scouting, which undergoes a more rigorous filtering process, L1T scouting offers a relatively unbiased view of the data, having undergone fewer steps in the trigger chain. This combination of high-frequency data capture and minimal filtering can amplify the sensitivity of new physics searches, providing a richer substrate for analysis.

The Level-1 scouting system is conceptualized to operate semi-independently, detached from the conventional trigger and data acquisition chain. This autonomy means that, in specific scenarios, the data sourced exclusively by the scouting system might be ample for deriving New Physics insights. Conversely, in other situations, it might offer preliminary indications that can guide the formulation of a specialized trigger algorithm for deeper exploration.

The envisaged architecture of the scouting system leverages the unused optical outputs of Level-1 trigger boards. It will receive trigger data in sync with the established 25 Gb/s serial interconnect technology, maintaining the link protocol intrinsic to the trigger. Dedicated FPGA boards mediate data acquisition. These boards bridge the synchronous trigger and the asynchronous scouting data acquisition domains and engage in preliminary processing like zero-suppression or recalibration. The immediate processing step in the scouting data landscape is executed in the I/O nodes, directly connected to the data acquisition boards. These nodes harness distributed algorithms for feature extraction while the data is momentarily buffered in short-term memory. These nodes might also be fortified with GPUs or other hardware accelerators. On detecting pertinent features, the associated events, or even condensed versions, known as "mini-events", are relayed over a high-performance computing network to a specialized processing farm. This processing farm utilizes distributed stream processing for feature

Table 3.1: Inputs to the Level-1 data scouting Run-3 demonstrator.

Input system	Number of 25 Gb/s links	Objects
uGMT	8 + duplicate 8	Up to 8 uGMT final muons 8 BMTF muon candidates
Calorimeter trigger	7 + 1 spare	e/γ, tau candidates, jets and energy sums including $E_{\mathrm{T}}^{\mathrm{miss}}$
BMTF	24	BMTF input super-primitives
uGT	18	Algorithm bits

reconstruction and extraction. Ultimately, its outputs are stored in a database designed for medium-term retention, facilitating analysis through queries. Only the results of these analyses are archived in permanent storage.

The scouting system, shown in Fig. 3.16, is designed for phased deployment. The foundational proposal entails sourcing data from the global trigger stages, specifically the Global Trigger's input (sGS) and output (sDS). Subsequent phases could integrate data from muon tracks, calorimeter objects (sLS), tracker tracks (sTS), and, in the final stage, the calorimeter trigger primitives themselves (sPS). For the Phase-2 scouting system, the L1 scouting project has pinpointed the DAQ800 board as the ideal hardware medium. This board, referenced in [54], is undergoing development specifically for the CMS Phase-2 upgrade. It has two robust Xilinx VU35P FPGAs, each chip connected to 6 × 4 FireFly connectors that are used to provide 24x 25 Gb/s input bandwidth and to 5 QSFP connectors that provide 5x 100 Gb/s output bandwidth. The DAQ800 read-out board aggregates and transmits data using a custom synchronous link protocol (SlinkRocket) to the central CMS DAQ system's receiver units. The sending module remains almost untouched for scouting purposes, but the receiving module is overhauled to handle data from the L1 trigger's asynchronous serial link protocol. Given that the board receiving bandwidth marginally surpasses its sending bandwidth, a zero suppression mechanism will be introduced before the sender module.

3.4.2.1 The Run-3 demonstrator

During LHC Run-3, a demonstrator system [52, 53, 55] has been set up to assess various concepts and understand system dynamics using real data. This demonstrator system draws data from the Phase-1 Global Trigger (uGT), the Global Muon Trigger (uGMT), the Calorimeter Trigger, and the Barrel Muon Track Finder (BMTF).

The Run-3 Level-1 data scouting demonstrator consists of a series of FPGA-based processing boards receiving data via optical links from the trigger system. From early 2023, the system receives data from the uGMT, the calorimeter trigger, the uGT and the BMTF, with details illustrated in Tab. 3.1. Afterward, the data is transferred to computing nodes (DSBU), where event construction and subsequent processing occur.

The Run-3 demonstrator manifests as a heterogeneous system comprised of three distinct receiver board types:

1. **Xilinx KCU1500:** This development kit hosts the KU115 FPGA, capable of handling eight optical links at 10 Gb/s each. It communicates with a host computer through PCIe and employs Direct Memory Access (DMA) for data transition to this host. Subsequently,

the data gets sent to the corresponding DSBU machine. The KCU1500 was initially applied in a smaller demonstrator at the close of Run-2, where it received inputs from the uGMT, as referenced in [56].

2. **Micron SB852:** This PCIe card hosts a Xilinx VU9P FPGA and is enhanced with the Micron Deep Learning Accelerator (MDLA). Functionally similar to the KCU1500, it also supports eight optical links at 10 Gbps and utilizes DMA to transfer data to the host computer.

3. **Xilinx VCU128:** This board hosts with its VU37P FPGA. It is designed to handle 24 input links at a capacity of 25 Gb/s and has four output channels rated at 100 Gb/s. These features mirror half the capabilities of a DAQ800 board. However, it is crucial to note that the input links are set to operate at 10 Gbps for the demonstrator, aligning with the transmission speed of the Phase-1 Level-1 trigger.

Data from the uGMT and the calorimeter trigger is transferred through eight 10 Gb/s optical links, respectively, to a pair of Xilinx KCU1500 boards, which decode the trigger link protocol, align the links with each other and performs firmware zero-suppression. This first stage of zero-suppression reduces the uGMT data rate by a factor of ∼10, discarding data from any bunch crossing where no muons have been found. A more fine-grained zero-suppression is instead performed on the host PC in software, thus reducing the data rate further. A duplicate set of GMT muons is sent to the Micron SB852 board, used to prototype on-the-fly muon histograms involved in luminosity measurements and neural network approaches for re-calibration and classification of L1 trigger objects. The BMTF super primitives and GT algorithm bits are sent over 24 and 18 links, respectively, to the Xilinx VCU128 boards, utilizing the High Bandwidth Memory of the VU37P chip to send data directly to a commercial PC.

We should note that L1 trigger objects are calibrated to achieve a specific efficiency at a given energy or transverse momentum threshold. For this reason, we cannot employ them for a direct physics analysis. Ongoing studies are exploiting Machine Learning, specifically neural networks, to re-calibrate the L1 information so that they can be used for semi-online analysis studies [55,57,58]. Although the L1 scouting system does not have to follow the strict latency requirements of the L1 trigger pipeline, it still needs to handle a large throughput of roughly 2 million muons per second. Therefore, the trained neural network must be capable of sustaining a high number of inferences per second. To accomplish this, the L1 data scouting system implements neural networks in the FPGA boards receiving the data from the L1 trigger system. While using Verilog or VHDL could be more efficient regarding resource utilization, an easier solution to implement a neural network model on FPGAs is exploiting alternative technologies. The Micron Deep Learning Accelerator (MDLA) [59] includes a software compiler that converts neural networks to hardware instructions for an FPGA processor. The models are trained in Tensorflow [60], then converted to Open Neural Network Exchange (ONNX) format and executed on hardware using the MDLA API. Another approach is implementing neural networks in the VU37P FPGA using the Python API and command-line tool HLS4ML [61] to translate trained neural networks to synthesizable FPGA firmware.

3.4.3 Data scouting at the DT Phase-2 front-ends

Despite CMS maintaining its two-stage trigger system, advancements in front-end (FE) electronics and trigger boards have enabled the reconstruction of high-quality physics objects at the Level-1 hardware trigger level. The CMS 40 MHz Level-1 trigger scouting initiative [52,53] intends to use spare optical outputs from Level-1 trigger boards. This data is then processed

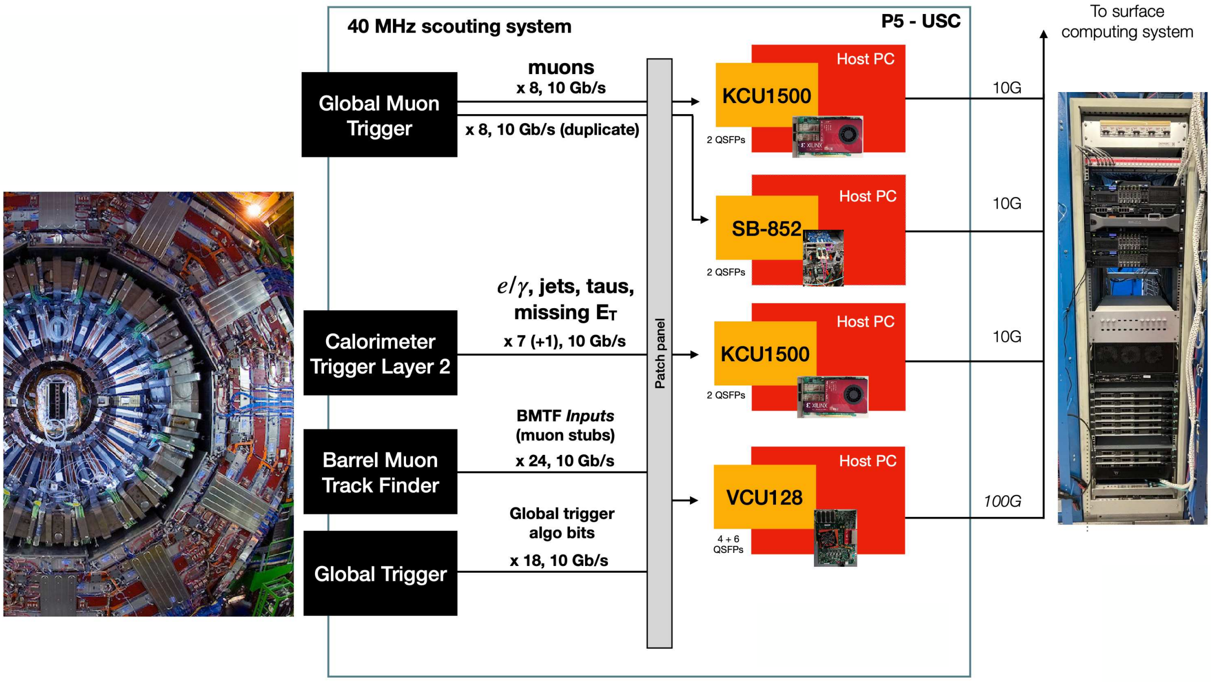
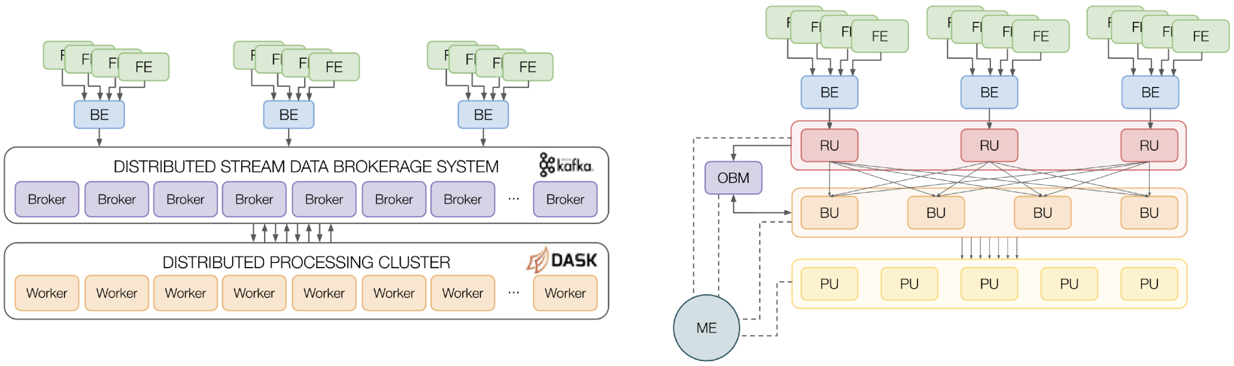


Figure 3.17: The L1 data scouting system demonstrator at the LHC Run-3. Figure from [57].

nearly online, leveraging dedicated computing resources without the latency constraints inherent to Level-1. This approach will aid numerous physics studies, especially those about processes with lower trigger efficiencies due to Level-1 trigger thresholds. Furthermore, analyses of rare processes, benefiting from increased available statistics, are also set to improve. The approach also supports non-standard reconstruction efforts, like searches for long-lived particles spanning multiple bunch crossings or analyzing appearing or disappearing tracks. With improvements to the front-end electronics, CMS can expand the scouting initiative, extracting data as close to the detector FEs as the throughput permits. The CMS is investigating the possibility of reconstructing data online at full resolution using 40 MHz data scouting at the detector's FE [62]. This effort involves collecting and processing data immediately after the detector's FE, even before any trigger intervention.

The inaugural application of the 40 MHz data scouting at the detector's FE was launched in CMS in 2022. The DT detector subsystem was elected for the initial deployment since four chambers (MB1 to MB4 of the DT sector 12 of wheel +2) incorporated the Phase-2 On-Board DT read-out boards (OBDT) [42, 43]. These boards manage Time-to-Digital Conversions of DT hit timings in FPGA with precision at the nanosecond scale. The DT Phase-2 Upgrade demonstrator is outfitted with 13 OBDT boards, encompassing 3120 distinct channels. The hit streams generated by the OBDTs are relayed through rapid optical links under the GBT [63] protocol to back-end (BE) apparatuses. Data is amassed and concurrently processed, divergent from the conventional CMS DAQ and trigger pathways. This enables a side-by-side evaluation of the legacy and the advanced demonstrator systems. BE devices utilize two Xilinx KCU1500 development boards, each fortified with a Kintex UltraScale XCKU115-2FLVB2104E FPGA. Every BE board accommodates two QSFP transceivers, handling up to 8 OBDT input connections. These BE boards are housed within a specific Dell PowerEdge R730 server, connected via PCIe Gen3 x8 interfaces on a bifurcated x16 edge connector. The FPGA firmware facilitates link deserialization under the GBTx-FPGA protocol [64] and consolidates all links into a unified data stream. Direct Memory Access (DMA) engines conduct data shifts to memory using the Advanced eXtensible Interface (AXI) stream protocol over the PCIe Gen 3 bus, alleviating the


FE
BE
DISTRIBUTED STREAM DATA BROKERAGE SYSTEM
kafka
Broker
...
DISTRIBUTED PROCESSING CLUSTER
DASK
Worker
FE
BE
RU
OBM
BU
PU
ME

(a) Distributed processing architecture

(b) Multistage architecture

Figure 3.18: Schematic representations of (a) the distributed processing data scouting architecture and (b) the multistage data scouting architecture. FE: front-end devices; BE: back-end devices; RU: read-out units; BU: builder units; PU: processing units; OBM: orbit manager; ME: master entity process. (Figure from [62]).

load on the server CPU. When monitoring LHC collisions, the CMS DT sector test produces physics hits with a throughput close to 40 MB/s, equating to an aggregate rate of 5 MHz for DT hits. The DMA's bandwidth over the PCIe Gen3 bus surpasses 1 GB/s, posing no restriction on the system's handling capacity. This data stream then transforms continuously to a computational farm for real-time processing. Two distinct architectural designs have been developed for data transfer and online processing.

3.4.3.1 Distributed processing and messaging systems

The architecture presented in Fig. 3.18a relies on using distributed computing frameworks that can scale horizontally. This framework attends to both the handling and processing of signals from the front-end. It utilizes Dask [65] as its primary engine for distributed processing, orchestrating data reconstruction tasks across multiple worker nodes. Apache Kafka [66] functions using the publisher-subscriber model to efficiently channel data from BE units to Dask's workers. Within each BE unit, there exists a persistent Kafka producer process that methodically scans server memory. This process broadcasts any newly acquired data to a remotely designated topic, which is, in turn, disseminated across several partitions spanning a collection of remote broker nodes. A cluster, composed of Dask worker nodes, plays the role of data consumers. They persistently draw from Kafka's partitions, assimilating raw data to form distributed data structures. This producer-consumer framework separates the tasks of writing and reading, enabling Dask worker nodes to retrieve and process data independently from the BE's data read-out from the FE. The inclusion of Kafka brokers functions as a buffering phase in this architecture.

This model is designed to work with the asynchronous 40 MHz data scouting read-out, reducing pressure on initial read-out stages. By leveraging the inherent data-analytic capabilities of the distributed framework, straightforward aggregated metrics—valuable for detector monitoring—are directly computed. An event-building phase segments data batches based on the LHC Orbit ID, which consists of 3564 intervals. Each interval is of 25 ns duration.

This decentralized processing setup has undergone practical testing using cloud computing. Both the Apache Kafka and Dask clusters were set up on virtual platforms steered by the Kubernetes orchestrator. The computing resources allocated for this endeavor consisted of five virtual machines, each endowed with four virtual CPUs and a memory of 16 GB.

3.4.3.2 Multistage architecture

As depicted in Fig. 3.18b, an advanced multi-stage aggregation framework has been devised as an alternative to the earlier described distributed cluster-centric system. Central to this design is synthesizing fragmented data accumulated from various BE devices, intending to curate holistic data structures—events—before dispatching them to processing modules. Given the intrinsic nature of the 40 MHz data scouting system, which amasses data from multiple FE devices uninterruptedly and asynchronously, discerning a traditional event structure is not immediately feasible before signal processing. Consequently, the LHC orbit identifier emerges as the primary event "key". Numerous Readout Units (RUs), seamlessly integrated within the servers that host the BE boards, serve as in-situ key-value stores, temporarily lodging the amassed DT hits within arrays delineated by the LHC Orbit ID. One distinct Orbit Manager process (OBM) oversees the cataloging of available LHC Orbits IDs spread across all RUs. This is achieved as RUs communicate a fresh message to the OBM whenever a new index is established within the RU cache. Concurrently, a separate suite of processes is tasked with amalgamating all hits corresponding to a specific LHC orbit, as cached across various RUs, culminating in a singular collection harmonized with a designated LHC Orbit ID. Every such entity, termed Builder Unit (BU), is a multi-threaded process that interacts directly with the roster of Orbit IDs managed by the OBM and the RUs to retrieve cached records. Following the event-construction phase, BUs channel the consolidated hits toward several Processing Units (PUs). These PUs stand at the forefront, analyzing these congruous sets of hits and elucidating attributes aligned with the muons' trajectory through the detector. To administer this ensemble of processes (RU, BU, PU, OBM), a Master Entity (ME) process is commissioned. The ME offers a computational interface, serving as the gateway for engagement with the entirety of the system's constituents, including initiating/terminating data acquisition and amassing metrics from all involved processes. Both the RU and BU processes are executed on the Dell PowerEdge R730 server, which houses the two BE boards. This server boasts dual Intel Xeon E5-2620 octa-core CPUs complemented with 64 GB of memory. Notably, the resource expenditure owing to the RU and BU processes remains confined to below 25% of the available capacity. PUs operate on distinct servers, and considering the data flux from the DT sector test, several PU processes could be co-located on the identical server earmarked for RU and BU processes.

Chapter 4

Online data quality monitoring as a demonstrator for New Physics searches with trigger-less muon data streams

In Chap. 2, a new technique for anomaly detection was discussed. It identifies statistical anomalies within the dataset without relying on predetermined assumptions about potential discrepancies from the reference distribution. Recent studies [3, 12, 67] have thoroughly evaluated the effectiveness of the New Physics Learning Machine (NPLM) in offline analyses. These evaluations mainly focused on how well the algorithm can detect New Physics signals in standard LHC analysis contexts, including the di-muon final state. The NPLM has been recognized for its sensitivity, resilience, and adaptability as a model-independent technique. It was compared to other analysis strategies, spanning model-dependent and independent methodologies. Data constraints inherently limit the effectiveness of data-driven approaches. In high-energy physics, data is selected, processed, and interpreted based on current comprehension of particle physics phenomenology, guided by the Standard Model.

In Chap. 3, we discussed the details of the CMS experiment and the significance of its trigger system. The trigger system is necessary, as it would be unfeasible to read-out all the data produced by collisions at the LHC. The Level-1 trigger applies a coarse, low-latency selection to identify interesting events from a large background rate. There is increasing awareness, however, that the trigger selection might be hiding possible signatures of New Physics. Although the trigger system allows room for searching for New Physics, these processes are extremely rare and often difficult to measure, even with specialized search methods. The CMS experiment requires upgrades to handle the upcoming High Luminosity phase of the LHC, which will significantly enhance New Physics investigations. The improvements include an enhanced detector and trigger system that ensures almost real-time resolution at Level-1. The currently deployed Run-3 demonstrator highlights physics use cases suitable for unbiased anomaly detection algorithms, such as the NPLM. A possible application is using cosmic muon events collected in specialized runs as a reference dataset for the NPLM. In this scenario, the algorithm can analyze orbit gaps in standard collision scouting datasets. These gaps are bunch crossings that are empty of collisions. Since cosmic muons constantly interact with the detector, they mainly occupy these gaps. Therefore, the NPLM allows comparing the reference cosmic muon distribution with the analyzed orbit gap data. Anomalies, potentially indicative of long-lived particles, might populate these gaps, which the NPLM algorithm would be required to detect.

This chapter focuses on applying the NPLM algorithm to unfiltered muon data streams collected by a set of Drift Tube (DT) chambers in the context of DT scouting, as introduced in Sec. 3.4.3. This configuration promises unparalleled sensitivity by utilizing raw detector data with minimal pre-processing. However, the early stages of electronic prototypes and computational infrastructure for real-time data processing at 40 MHz make using such advanced strategies for New Physics explorations difficult. The focus is on providing comprehensive feedback on the algorithm and prototypes while having better control over the experimental setup and the collected data stream. Implementing an online anomaly detection system on an unfiltered 40 MHz muon data stream presents a practical opportunity for data quality monitoring (DQM). At its core, DQM is similar to exploring New Physics. Anomalies in DQM indicate detector issues, while New Physics suggests unexpected phenomena. Integrating Phase-2 electronics and scouting enhancements into an unbiased anomaly detection system, similar to exploring New Physics, reveals numerous innovative processing and analysis opportunities during the High Luminosity era. It should be accentuated that the goal of this chapter, and this thesis in general, is not to provide a comprehensive exploration of New Physics. The focus is on how Phase-2 upgrades, including 40 MHz data scouting, real-time processing, and analysis, can enhance the investigative potential for revealing previously obscure signatures during the High Luminosity phase. This chapter will provide an overview of DQM at collider experiments and explain how NPLM can improve these techniques. We will then introduce the experimental setup for real-time anomaly detection on trigger-less muon data streams. The data acquisition and processing pipeline is integral to the setup as it enables the extraction and processing of data streams to fit the algorithm input requirements. Finally, we detail the anomaly detection strategy using the NPLM algorithm and show results regarding sensitivity and scalability to collider experiments with more challenging data throughput.

4.1 Monitoring collider experiments data quality

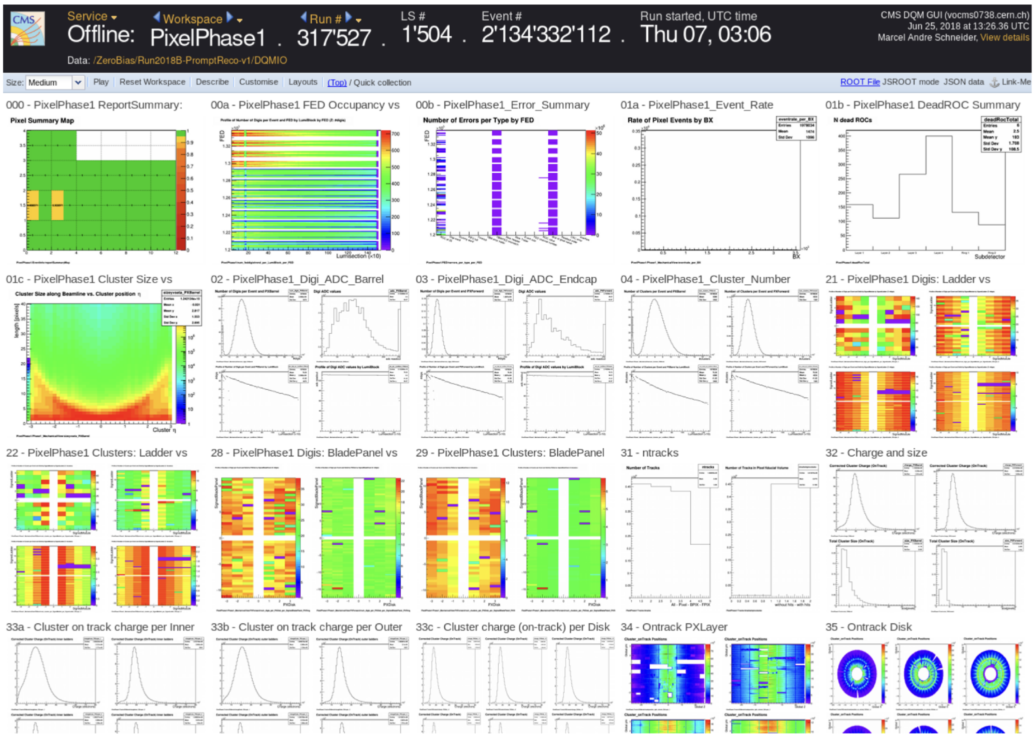
This study showcases the possibility of unbiased searches for New Physics at the High Luminosity LHC (HL-LHC) using CMS Phase-2 upgrades. The data quality monitoring (DQM) application was adopted as a test bench to enhance flexibility and testing control. However, we suggest potential improvements to existing state-of-the-art DQM systems and techniques using our proposed pipeline and analysis framework. This section will detail the current state of DQM and indicate areas that could benefit from our work.

4.1.1 Introduction to DQM

Modern high-energy physics experiments operating at particle colliders are extremely sophisticated and advanced systems. Millions of sensors are sampled every few nanoseconds, resulting in a vast amount of complex data. Different technologies are utilized to detect and measure the particles produced from collisions. These experiments operate under demanding environmental conditions, making consistently achieving the required performance metrics challenging.

Failures are inevitable within a comprehensive system like the CMS experiment. Ensuring that the CMS data is suitable for physics analysis is crucial because it underpins the reliability of all published results from the CMS Collaboration. Even though there are design redundancies within various subsystems, measurements can be affected by part malfunctions or by potential misinterpretations of spurious signals. To put it in perspective, around 7% of DT components have issues and 2% of the data is discarded [68]. These figures are not primarily due to overarching malfunctions of the detector but more often are related to localized problems. Meaningful physics analysis can still be performed using data from undamaged detector areas. As a result, the monitoring system must deliver a general status and accurately pinpoint localized issues.

The monitoring scope is not limited to overseeing operational parameters like power, electronic configuration, or temperature of hardware components. Continuous quality checks on data from all sources are essential to promptly identify and, if possible, diagnose the root



Service Offline: Workspace PixelPhase1 . Run # 317'527 . LS # 1'504 . Event # 2'134'332'112 . Run started, UTC time Thu 07, 03:06

CMS DQM GUI (vocms0738.cern.ch) Jun 25, 2018 at 13:26:36 UTC Marcel Andre Schneider, View details

Data: /ZeroBias/Run2018B-PromptReco-v1/DQMIO

Size: Medium | Play | Reset Workspace | Describe | Customise | Layouts | (Top) / Quick collection | ROOT File JSROOT mode JSON data Link-Me

000 - PixelPhase1 ReportSummary:
00a - PixelPhase1 FED Occupancy vs
00b - PixelPhase1_Error_Summary
01a - PixelPhase1_Event_Rate
01b - PixelPhase1 DeadROC Summary
01c - PixelPhase1 Cluster Size vs
02 - PixelPhase1_Digi_ADC_Barrel
03 - PixelPhase1_Digi_ADC_Endcap
04 - PixelPhase1_Cluster_Number
21 - PixelPhase1 Digis: Ladder vs
22 - PixelPhase1 Clusters: Ladder vs
28 - PixelPhase1 Digis: BladePanel vs
29 - PixelPhase1 Clusters: BladePanel
31 - ntracks
32 - Charge and size
33a - Cluster on track charge per Inner
33b - Cluster on track charge per Outer
33c - Cluster charge (on-track) per Disk
34 - Ontrack PXLayer
35 - Ontrack Disk

Figure 4.1: Screenshot of the CMS DQMGUI user interface. Figure from [73].

causes of any anomalies. Due to the high data rate and large number of sensors, automation is beneficial for the monitoring process. Machine learning (ML) techniques have become increasingly relevant for this purpose and are now used by several experimental collaborations, as referenced in [69–71]. These techniques complement traditional methods referenced in [72, 73]. In essence, data quality monitoring involves comparing data batches with reference samples collected under standard conditions. Any differences can then be analyzed to determine their source. This process must consider computational limits, which are influenced by data batch delivery frequency, volume, sensor grouping granularity, and desired statistical accuracy.

4.1.2 Traditional methods

Quality data is essential for physics analyses in the CMS Collaboration. Rigorous quality criteria are applied to ensure only the highest standard of datasets are used. During data acquisition, a selected fraction of data is processed in real time to produce a series of histograms encapsulating critical data metrics. An example is provided in Fig. 4.1. These histograms are cross-referenced with a predetermined reference set—typifying the detector's response under standardized operational conditions. Based on this comparative analysis, expert operators, or "shifters", address any discrepancies, taking decisions as significant as halting data acquisition based on the severity and nature of the detected anomaly. The historical context of LHC operations and any prior detected issues are invaluable tools in this decision-making process.

The monitoring approach used in CMS can be divided into two main categories:

- **Online monitoring:** Provides real-time feedback on data quality during acquisition, allowing for quick resolution of issues.
- **Offline monitoring:** Validates data integrity post-acquisition by transforming raw detector data into a coherent list of detected particle events through centralized processing.

These domains can be distinguished based on the following:

- **Latency:** While online monitoring necessitates near-instantaneous data assessment to ensure immediate interventions, offline evaluations typically span several days.
- **Data access:** The online mechanism processes data at 100 Hz, equivalent to roughly 10% of the total data archived. In contrast, offline mechanisms handle the entirety of events sanctioned by the trigger system, approximating 1 kHz of data.
- **Granularity:** Offline monitoring is more holistic, concentrating on the general status of various sub-detectors. In contrast, the online domain delves deeper, identifying specific malfunctioning elements within the sub-detectors.

Both monitoring strategies share a methodological foundation, scrutinizing predefined sets of histograms and statistical tests to detect known failure patterns. Detector specialists compare each data distribution to its reference, identifying anomalies such as noise, inactive detector zones, or calibration issues. A detailed explanation of CMS data quality monitoring infrastructure can be found in [73].

4.1.2.1 Online data monitoring

Online DQM prioritizes data sampled during the High-Level Trigger (HLT) processing. The objective is to assess sub-detector components in real time with the current LHC beam conditions. Given its emphasis on immediacy, histograms are updated with minimal latency, offering a dynamic visualization of the detector's operational performance. This live feedback mechanism is invaluable. It provides experts and operators with insight into the detector's current state. These insights guide decision-making, incorporating past anomalies and LHC operational metrics. Upon the conclusion of each data run, shifters are entrusted with the task of annotating a quality flag for every sub-detector, marking problematic subsystems as "bad", unless overruled by the shift leader or an expert in the specific subsystem.

4.1.2.2 Offline data certification

Once data has been collected, it goes through a thorough vetting process known as Data Certification (DC). This ensures its suitability for comprehensive physics analysis. DC ensures reconstructed events meet stringent requirements for optimal detector functionality. Through rigorous training, specialists identify and pinpoint anomalies arising from hardware discrepancies or software glitches by analyzing histograms that depict crucial data metrics. The definitive certification flag is determined by comparing these findings with a predetermined reference that encapsulates the standard detector response under normal conditions. Certification shifters use their knowledge of past issues to make decisions. Given the diverse origin of monitoring data from various CMS sub-detectors, comprehensive data quality depends on the collective performance of these individual components. This requires the expertise of specialists who are familiar with the specific behaviors of sub-detectors. As a result, it is a collaborative effort that involves about seventy experts. The complexity of decision-making and pressure to certify quickly can lead to errors in quality labeling, making a human-centric approach vulnerable to minor inconsistencies.

4.1.3 Machine learning approaches

The complexity of monitoring collider experiment data quality is amplified by the evolution of detector technology and the transition to High Luminosity configurations. Due to the diverse range of LHC operational conditions, the detectors generate a large amount of monitoring data. This increase in data, especially the histogram output that must be evaluated, is further compounded by the inclusion of newly identified failure modes. While this enriched data is pivotal for anomaly identification and mitigation, it also results in delays due to the extensive volume requiring analysis. The extensive monitoring requires a significant amount of personnel for both dedicated shifts and updating references. Given these considerations, an inevitable trend towards automation, fortified by machine learning, becomes imperative. Embracing automated mechanisms, specifically ML-enhanced anomaly detection, in anticipation of future LHC operational phases will not only optimize current workflows but also cultivate robust expertise within the collaboration regarding advanced ML frameworks.

Implementing machine learning techniques marks a significant change for the CMS DQM methodology. Using extensive monitoring data, coupled with expert annotations, presents an opportunity to train and refine algorithms. This dataset provides a reliable platform for training algorithms to identify complex patterns and deliver detailed insights. Envisioning a ML-driven DQM system entails algorithmic data pre-processing, with expert review reserved for complex or ambiguous scenarios, as suggested in [74].

Methodologically, monitoring metrics can be systematically categorized based on characteristic patterns. Monitoring at the detector level involves hit and occupancy maps that show spatial distributions of key parameters. These visualizations help identify and address inefficient regions within detectors. As discussed in [68], image classification techniques offer promising diagnostic tools for such data structures, whether supervised for known anomalies or semi-supervised for unfamiliar ones. Concurrently, physics objects—hadrons, leptons, and photons—are evaluated for deviations in their statistical distributions during the data certification phase. A major challenge in this field is the complex nature of the data, which often requires a comprehensive understanding of interdependent detector components. However, emerging representation learning methods appear to offer practical solutions to these challenges.

Nonetheless, moving towards a monitoring approach focusing on machine learning poses unique challenges. The high dimensionality inherent in collider data often precludes simplistic parametric modeling. The scarcity of labeled datasets for online monitoring and the intricacies of label pollution in offline scenarios increase these challenges. Due to the dynamic nature of potential failure scenarios and evolving operational parameters, continuous model recalibration is necessary. Therefore, while the prospective benefits of such a transition are considerable, it necessitates meticulous strategic planning and iterative refinements.

4.1.4 The NPLM contribution to DQM

Traditional DQM methods mostly relied on examining a multitude of one-dimensional distributions. The NPLM algorithm represents a significant shift in how data quality monitoring could be approached. The NPLM offers a comprehensive examination of the entire phase space, not just predefined input variables, making it capable of discerning intricate variable correlations. Furthermore, the algorithm can ingest lower-level data quantities that necessitate minimal preprocessing. For the DQM framework, this translates into the ability to assimilate near-raw data extracted directly from the detectors' front-end electronics. This reduction in data manipulation simplifies the process and reduces the risk of biases that may mask underlying data anomalies.

The NPLM's model-independent approach offers versatility across various applications, avoiding restrictions from labeled datasets. This feature distinguishes it from conventional supervised learning algorithms limited by the specifics of their training data. As a result,

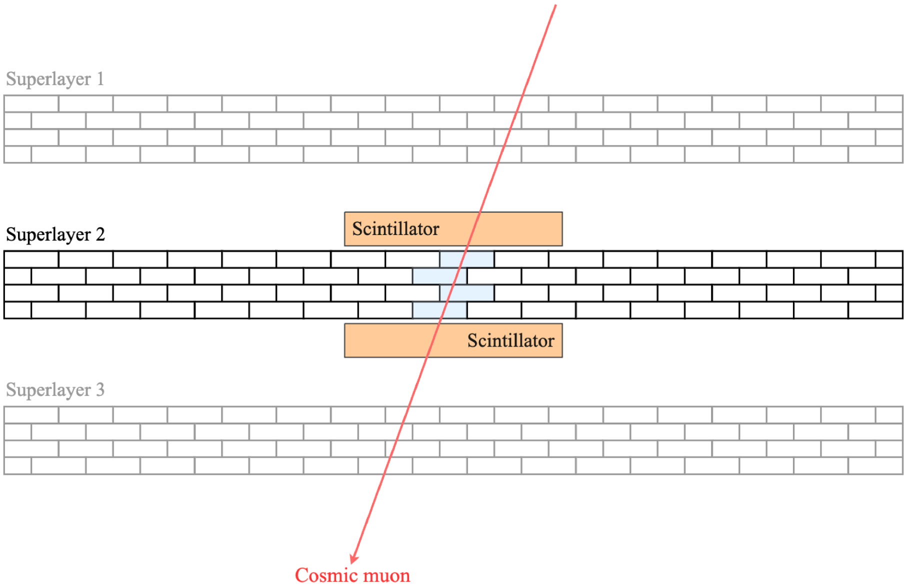


Figure 4.2: Schematic representation of the muon telescope setup at the Legnaro National Laboratories. Only the central superlayer has been used for the work presented in this chapter.

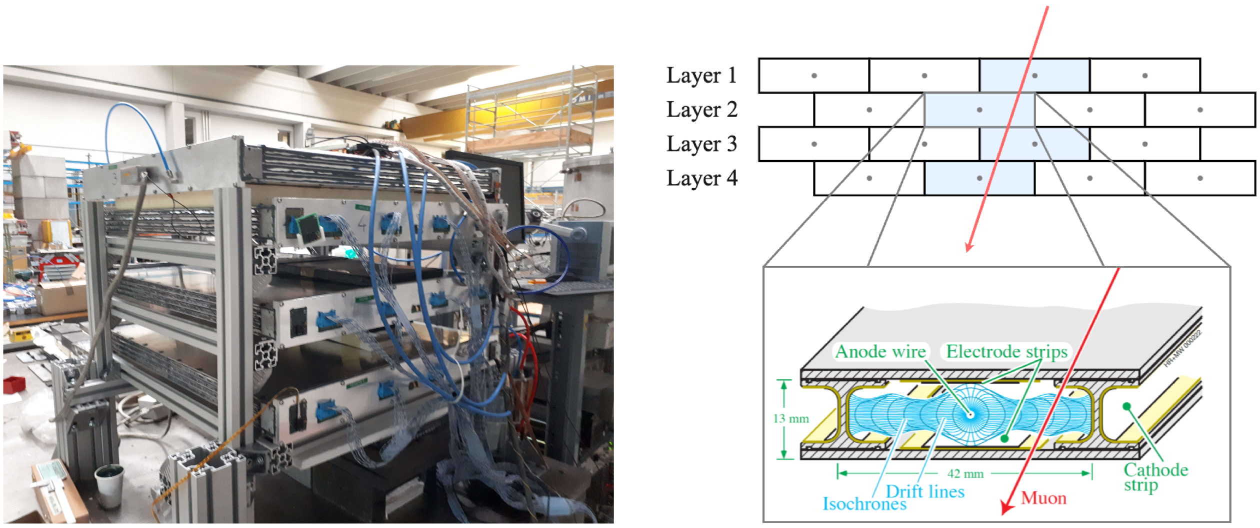
the NPLM improves its ability to understand data quality nuances. Furthermore, the NPLM anomaly detection algorithm provides a more comprehensive perspective than many existing machine learning DQM algorithms. Most anomaly detection algorithms detect occupancy-based anomalies but may not identify all data inconsistencies. The NPLM seamlessly integrates occupancy-based anomalies within its test statistic (Eq. 2.22), emphasizing the analysis of raw detector observables to detect a broader range of anomalies. This approach improves its ability to identify significant and minor deviations from the standard working conditions.

Incorporating machine learning into the DQM framework, the introduction of NPLM is a significant advancement. The NPLM takes a holistic approach to data analysis, which integrates occupancy-based with a broader spectrum of anomalies, unlike other machine learning algorithms relying solely on image classification and other supervised learning mechanisms. These distinguishing attributes underscore its promise and reinforce its positioning as an invaluable asset in the sophisticated and ever-evolving domain of DQM.

4.2 Experimental setup of the demonstrator

In high-energy physics, monitoring data quality is essential. So, we created a test-bench setup to validate the New Physics Learning Machine's capabilities for online data quality monitoring. This setup emulates CMS Phase-2 DT muon chambers with upcoming front-end electronics (see Sec. 3.4.3). As discussed, the On-Detector Boards for Drift Tubes (OBDT) [42, 43] allow direct data stream extraction, providing an unfiltered 40 MHz muon data stream. The setup operates as a cosmic muon telescope, exposing it to much lower throughput than LHC collisions. Therefore, although the CMS setup is mirrored closely, this setup provides greater testing capacity, increased control in the experimental setup, and an overall more manageable data collection environment.

The system has two main purposes. Firstly, it aims to develop advanced online DQM methodologies for managing up to 40 MHz raw data streams with minimal preprocessing. Secondly, it serves as a demonstrator for New Physics searches during the High Luminosity phase. It utilizes NPLM's anomaly detection abilities to search for potential New Physics signals in unfiltered, unbiased, *scouting* data streams.



(a) (b)

Figure 4.3: Left (a): a picture of the muon telescope setup at the INFN Legnaro National Laboratories. Right (b): a schematic representation of the staggered layers configuration (top) and a schematic view of the drift cell (bottom).

4.2.1 The Muon Telescope at the Legnaro National Laboratories

To evaluate the potential of the NPLM algorithm for sensitivity and scalability in the CMS Phase-2 experiment at the HL-LHC, we employed a specialized experimental apparatus for DQM tasks. This apparatus, located at the Legnaro INFN National Laboratory in Padova, Italy, consists of three drift tube chambers (see the schematic representation of the setup in Fig. 4.2 and the picture of the setup in Fig. 4.3a) modeled after the barrel muon chambers of the CMS experiment at the LHC, detailed in Sec. 3.2.1. This compact replica was conceived in alignment with the design principles of the original CMS chambers. Their operational efficacy was first validated in the 2019 test-beam campaigns under the LEMMA project [75].

Chamber Construction and Geometry Each chamber (also called *superlayer*) consists of four layers, with 16 tubes or cells in each layer. The cells have a transverse dimension of $42 \times 13\,\mathrm{mm}^2$. To increase tracking precision, cells in adjacent layers are staggered, offset by half the width of one cell.

Gas Composition The cells are filled with an Ar-CO_2 gas mixture in an $85\% - 15\%$ ratio and maintained at atmospheric pressure. When a charged particle, like a muon, passes through this medium, it causes ionization events within the gas.

Electrode Configuration At the core of ionization detection is an anodic wire with a voltage of $V_{\mathrm{wire}} = +3600\,\mathrm{kV}$ placed at the center of the cell. The cell sidewalls, or cathodes, have a potential of $V_{\mathrm{cathode}} = -1800\,\mathrm{kV}$. Additionally, the chamber's top and bottom walls are equipped with electrodes (strips), both set at $V_{\mathrm{strip}} = -1200\,\mathrm{kV}$. This arrangement guarantees a balanced electric field.

Upon ionization, primary electrons are attracted to the anodic wire. As they draw closer, secondary ionization augments the original signal, typically by a magnitude of 10^5. The anodic wire then seizes the resultant electron cloud. The multi-electrode configuration ensures a consistent drift speed of roughly $54\,\mu\mathrm{m/ns}$ within the cell.

Signal Generation and Processing An external temporal framework is supplied by plastic scintillators nestled between the DT chambers, which captures the passage of the muons. When a muon passes through a chamber, it can produce up to four hits, exhibiting temporal coherence. Each hit captured by the wire undergoes a series of processes: amplification, shaping, and discrimination against a set threshold, guided by an Application-Specific Integrated Circuit (ASIC) [76] situated within the chamber's gaseous environment.

From this procedure, we can determine the hit timestamp t, denoted in TDC units, which represents the time of arrival of the ionized electron cloud. This timestamp on its own does not convey information about the muon's timing. To determine this, it must be referenced to a time pedestal t_0 that marks the moment a muon passes the detector. The difference $t - t_0$ provides the drift time t_{drift}, which is the elapsed time taken by the electrons to drift towards the wire following the muon's passage.

This drift time directly ties into the crossing distance x to the wire:

$$x_{\pm} = \pm \left[v_{\text{drift}} \times t_{\text{drift}} \right] \tag{4.1}$$

where v_{drift} is the drift velocity obtained through calibration. Based solely on the drift time, we can determine the distance of the muon from the wire with a left-right ambiguity, making track reconstruction more intricate. Particularly, we will need to solve a combinatorial challenge determining the optimal left-right hit combination that relates to the crossing of a muon.

Data Acquisition The chambers continuously intercept cosmic muons, registering hits at a typical rate of about 1 per minute per cm^2 at sea level. The 40 MHz data acquisition system operates autonomously without any trigger logic, simplifying the data collection process for further analysis. More details will be given in Sec. 4.3.1.

4.2.2 Production of real anomalies

Utilizing small-scale CMS DT replicas at Legnaro National Laboratories gives us complete control over the detector, electronics, and data acquisition and processing. This control is crucial for our work as a demonstrator. Therefore, we prefer conducting a mock-up experiment instead of using the CMS sector already equipped with OBDT prototypes extracting data at 40 MHz at the LHC. Our setup allows us to modify detector parameters to induce genuine anomalies in our data. This contrasts with most research on machine learning's role in DQM for CMS subdetectors, which rely on synthesized anomalies and software-engineered malfunctions crafted to mirror real-case scenarios.

Typical malfunctions within DTs encompass a range of issues: a broken cell, which could be attributed to the anodic wire or issues spanning the charge collection to signal acquisition chain; a hyperactive cell registering hits at an abnormally high frequency irrespective of muon presence; gas leakages or inconsistent gas pressure which consequently disrupt the linear correlation between drift time and the track position as elucidated in Eq. 4.1. Additionally, irregularities in the voltage of the electrodes can also be problematic.

We aimed to cover as many detector malfunctions as possible with minimal interventions to the hardware of the experiment. To achieve this, we modified the voltage of the cathodic strips and the front-end thresholds. Altering the cathodic strips perturbs the electric field's configuration while adjusting the front-end thresholds simulates the abrupt interference of noise sources. Notably, our approach was more nuanced rather than a binary type of fault wherein a component either functions perfectly or fails outright. We reduced the cathodic strips' voltage and the front-end thresholds to 75%, 50%, and 25% of their nominal values, corresponding to −900 V, −600 V, and −300 V for the strips and 75 mV, 50 mV, and 25 mV for the thresholds.

Such a gradient of failure, instead of an absolute 100% to 0% switch, provides a richer landscape of malfunction scenarios. For instance, issues like gas leakage or inconsistent pressure might mimic the effects of reducing the strips' voltage by a certain percentage.

Finally, data collected under these intentionally perturbed conditions undergo the same processing pipeline as “normal” data. This methodology will be elaborated upon in the subsequent section. In short, these datasets are collected under different anomalous conditions and used as test samples, $\mathcal{D}$. They will be compared to a reference sample, $\mathcal{R}$, collected under nominal detector working conditions, using the New Physics Learning Machine to test the algorithm's sensitivity, specificity, and time performance.

4.3 Online processing and computing infrastructure

The need for real-time processing is increasingly evident in modern high-energy physics experiments. The unique design of our instrumentation ensures a seamless transition from data acquisition, through preprocessing and processing, to the final stages of analysis, all taking place online. For a more nuanced comprehension of this data flow, it is essential to recognize its adherence to the LHC timing standards [77]. The data stream is structured with precise timing parameters, including an orbit counter, a bunch crossing (BX) counter, and the TDC timestamp for each orbit and BX. Within one orbit, there are 3564 bunch crossings. With a BX rate of 40 MHz, each BX spans 25 ns and is further segmented into 30 distinct bins via the TDC timestamps. Together with timing data, channel information indicates which detector cell has been triggered. Integrating this temporal and spatial data and leveraging our detailed understanding of the detector's geometry, we can translate this local spatial data into a global laboratory frame of reference. This enables the computation of the distance to the wire, x, using the drift time, as detailed in Eq. 4.1, laying the groundwork for global track reconstruction.

Our system, operating “trigger-less” at a rate of 40 MHz, can process vast quantities of data. Such an acquisition rate necessitates a meticulous approach to data management, from its initial collection to its final representation, ensuring rapidity, accuracy and efficiency. In progressing from initial data acquisition to advanced data processing, we utilize big-data tools and hardware acceleration, particularly GPUs and algorithmic parallelization. These technologies allow for the efficient calculation of the distance from the wire for each detected hit based on the drift time. The central part of our data processing is track reconstruction, which primarily focuses on quadruplets—four distinct hits, one in each chamber layer, representing the trajectory of a muon through the entire chamber. These quadruplets offer four specific points, indicating the potential paths of the muon within the detector. Subsequent data manipulations ensure the data is in an appropriate format, complete with necessary features for further analysis.

The main objective of this section is to outline the process of our real-time data flow, from acquisition to processing. While the later sections detail the analysis techniques, it is essential first to establish a foundational understanding of the design and operation of our present system.

4.3.1 Trigger-less readout

Our setup employs two Xilinx VC707 evaluation boards, equipped with Virtex-7 XC7VX 485T FPGAs, to function as front-end (FE) read-out boards. Each board performs the time-to-digital conversion (TDC) of the LVDS signals, ensuring precise tagging of the hits' arrival time relative to a reference clock. This TDC operation is executed in firmware through the standard IOSERDES, which operates at a speed of 1.2 GSps. An external oscillator distributes a 120 MHz clock across both VC707 boards. Emulating the typical clock distribution system

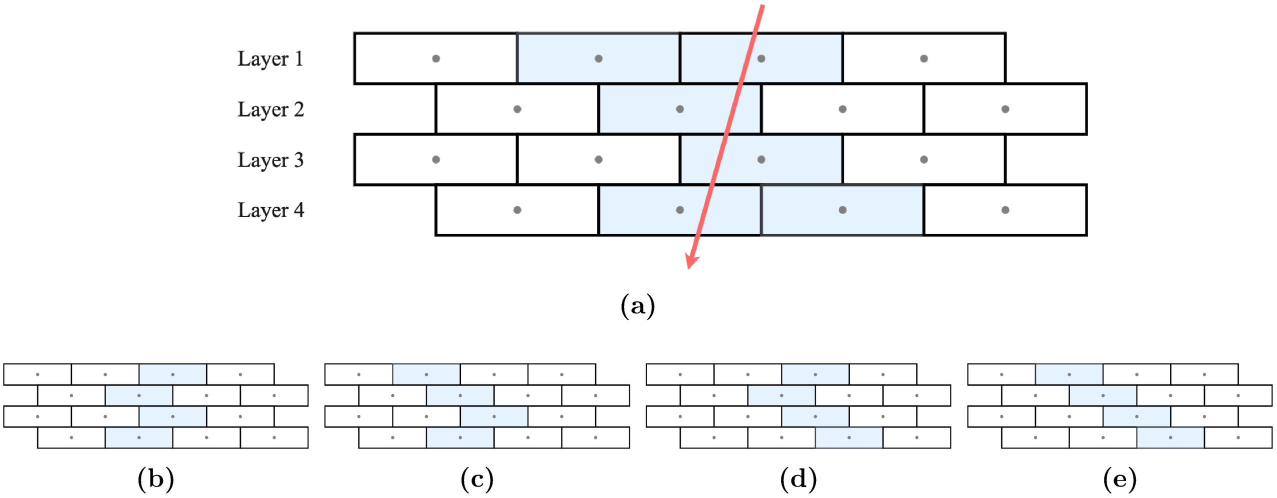


Figure 4.4: Top (4.4a): Schematic representation of the case in which there are more than four hits (six in this example represented as blue cells) within the orbit associated with the crossing of a muon. Bottom (4.4b to 4.4e): the four quadruplets that can be formed using the six hits in the detector.

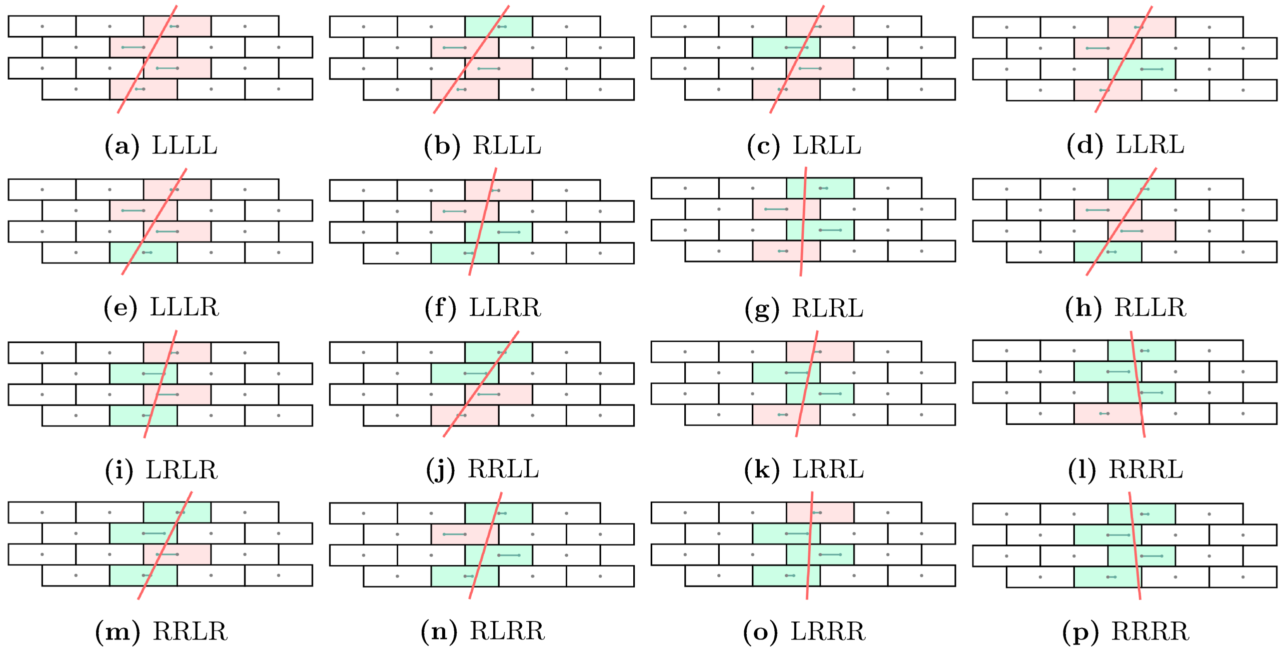
of LHC experiments [78], an internal 40 MHz clock is generated within each VC707, derived by down-scaling the external reference. With this 40 MHz clock, the TDC ascertains the time measurement of the signal's rising edge.

Each VC707 can accumulate data from 138 channels, signifying its ability to collect hits from two chambers of 64 channels each. The FE boards perform this without applying filters or selections to the signals, ensuring the comprehensive data stream of TDC hits is serialized with the GBTx-FPGA protocol [64]. This serialized data is relayed to SFP+ transceivers via 5 GSps optical links, directed toward a back-end (BE) board. The process is distinguished by its absence of a trigger signal, guaranteeing uninterrupted and asynchronous data flow. As the FE boards register signals, they are immediately digitized and transmitted to a BE board.

The BE board, Xilinx KCU1500 evaluation board, is mounted in a Dell PowerEdge server and equipped with a Kintex UltraScale FPGA. Here, the GBTx-FPGA protocol proceeds to deserialize data from each optical link. The FPGA's fast transceivers retrieve the clock directly from the received data through the Clock Data Recovery (CDR) mechanism. The data streams from these links are individually processed by an FPGA-based algorithm [79], which undertakes the task of reconstructing trajectories consistent with muons' passage through each chamber—though this particular method is not deployed in this study, as our aim is a novel approach. This processed data is then dispatched to a Direct Memory Access (DMA) engine by leveraging the Advanced eXtensible Interface (AXI) stream protocol, occurring over a PCIe Gen 3 bus. To optimize the DMA's throughput, a FIFO is implemented, determining the optimal size for DMA data transfers. Utilizing the Xilinx AXI-DMA engine, the data stream is seamlessly transmitted to the server memory. Given that the achieved throughput in DMA transfers surpasses that recorded in the subsequent online data processing, it does not pose any limitations to the data acquisition design described herein.

4.3.2 Online processing pipeline

After the trigger-less read-out, which accumulates data at a 40 MHz rate, this data is first directed into the server memory using DMA and then temporarily stored in the server's ramdisk. The NVIDIA A100 GPU, connected to the Dell PowerEdge R750 server, reads the data from the ramdisk. This system is then subjected to a thorough processing regimen, maximizing the GPU's computational capacities to ensure peak throughput. While we currently leverage standard DMA to the server memory, we are actively exploring the potential of incorporating



(a) LLLL (b) RLLL (c) LRLL (d) LLRL

(e) LLLR (f) LLRR (g) RLRL (h) RLLR

(i) LRLR (j) RRLL (k) LRRL (l) RRRL

(m) RRLR (n) RLRR (o) LRRR (p) RRRR

Figure 4.5: The sixteen combinations following from the left-right ambiguity of the quadrupled. Red cells indicate the choice of the left (L) possibility, while green cells indicate the choice of the right (R) possibility. The red line crossing the four layers represents the fit of the left-right combinations of distance x from the wire. In this particular example, 4.5i would then be chosen as muon track as it returns the best-fit red-line among the other combinations.

Remote DMA (RDMA) [80] directly into the GPU memory. Once seamlessly integrated into the current trigger-less read-out and processing pipeline, this enhancement promises significant advancements in efficiency.

Our processing efforts primarily center on the A100 GPU, where the data is organized into distinct streams. Each of these streams is systematically processed batch-wise, with each batch corresponding to one or more orbits, and is then subjected to the track reconstruction process.

The primary data at our disposal consists of the TDC information of the hits and the scintillator readings. If an orbit does not have a scintillator signal, it is immediately discarded, and the system moves on to the next orbit. When a scintillator signal is present in an orbit, the drift time for each hit is calculated to estimate the distance from the muon wire using the known drift velocity. Furthermore, exploiting the precise knowledge of the detector geometry, we map the x-coordinate of the hits into a global frame of reference. As mentioned earlier, the estimation of distance from the wire is challenged by the left-right (LR) ambiguity.

The individual hits are systematically grouped into sets of quadruplets, with the constraint that each hit in a set must come from a distinct layer. This grouping process is complex because external factors like noise interference and incidental passage of electrons can result in additional hits beyond what would be expected from just the muon's passage. To counteract this, when a muon signal (as evidenced by the scintillator's signal) is detected within an orbit, we collate all available hits in that orbit. We generate all possible quadruplets from this collection, ensuring that each hit within a set is sourced from a unique layer. A graphical representation of this scenario is detailed in Fig. 4.4, where the upper figure 4.4a visually depicts a crossing muon and six activated drift cells. Although only four of these are directly activated by the muon's crossing, the track reconstruction process mandates the division of all hits into quadruplets as illustrated in Figs. 4.4b, 4.4c, 4.4d, and 4.4e.

After forming these quadruplets, we address the combinatorial challenge introduced by the LR ambiguity. This results in $2^4 = 16$ potential combinations for each quadruplet, an aspect

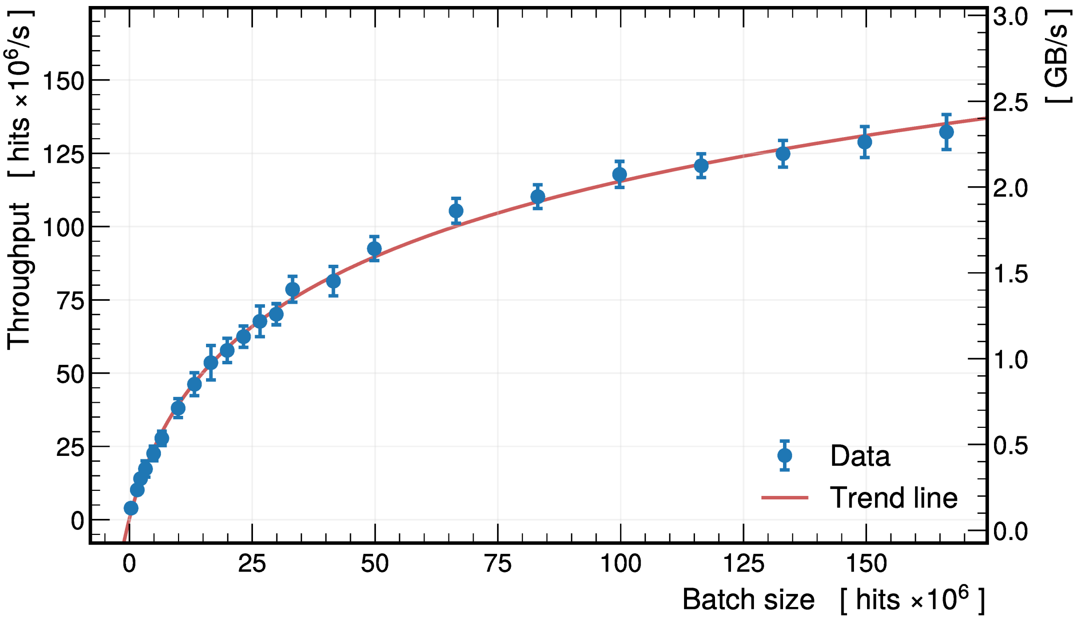


Figure 4.6: Throughput performance of the preprocessing algorithm, executed entirely on the GPU, plotted against batch sizes measured in millions of hits. The throughput values are expressed in processed hits (left y-axis) and gigabytes (right y-axis) per second of processing time. The blue data points represent measured throughput values, and the solid red line signifies a trend fit to the data.

visually depicted in Fig. 4.5. A fitting process is initiated for each combination to minimize the χ^2. The fits are then evaluated for accuracy, and the most precise one is selected as the potential muon track candidate.

To optimize our operations, we use the Rapids library [81]. It handles GPU-based dataframes proficiently via the cuDF library [82], and for GPU-optimized arithmetic operations, we deploy CuPy [83]. These libraries allow efficient array arithmetic and column-wise vectorized operations on dataframes in a GPU-parallelized environment. Such tools prove indispensable for preliminary tasks, such as the drift time and wire distance computations, predominantly grouping by orbit and isolating only those orbits manifesting a scintillator signal.
Special optimization techniques are instead needed to address the track fitting combinatorial challenge, particularly with the variable number of hits per orbit, quadruplets, and LR ambiguity. To meet this requirement, we use Numba [84] with custom CUDA kernels to design a specialized C++ kernel [85] that performs a linear regression on the quadruplet combination.

Each quadruplet has its own set of 16 possible combinations at this stage. For even greater efficiency, we flatten the potential combinations of all quadruplets and parallelize the fitting calculations across all viable combinations.

The other possibilities are discarded once the best combination is found, acting as a filter on the original dataframe of hits. This step purges the dataframe of noise hits, preserving only the combination of hits attributable to the muon. The conserved hits are then combined with the track's slope, which is valuable because it is linked to the angle at which the muon crosses.

To assess the efficiency and robustness of our processing pipeline, we performed a series of benchmarking tests by modulating the input batch size. Specifically, in our trigger-less model, we initiated the processing only after a predefined number of hits were streamed into the server ramdisk. Given our computational resources, this procedure allowed us to measure the throughput efficiency and calibrate the pipeline to maximize throughput. Figure 4.6 showcases our findings. The blue data points indicate observed throughput, quantified as the processed batch size in terms of 'number of hits' (left y-axis) and 'gigabytes' (right y-axis) relative to the processing time in seconds. The red line represents a fitted curve illustrating the general trend in our data. The trend suggests a logarithmic growth in throughput, with most batch sizes being processed within a second. Notably, our system exhibits potential for further optimization as we have not yet reached the throughput plateau for the tested batch sizes.

It is worth noting that the potential for multiple muons within a single orbit is theoretically feasible. Nonetheless, due to the concise $\sim$90 µs span of an orbit, it is rare for more than one muon to be present. For most orbits, no muons are detected. Instead, when developing applications specific to LHC dynamics, it is important to explore different approaches. One such approach uses a mean timer to determine the time pedestal, eliminating the need for an external time reference usually provided by the scintillators. This methodology is detailed further in [79] and has been proposed as a foundation for the anticipated CMS Phase-2 upgrades. For the scope of this study, we have chosen not to adopt this method to maintain simplicity.

4.3.3 Dataset extraction

Upon completing track reconstruction and selecting the optimal track representing the muon, we obtain a dataframe consisting of muon hits. By grouping by orbit, we can meticulously isolate the four hits associated with the muon. Our dataframe is transformed from hit-based to event-based, with data encapsulating an event's four drift times, crossing angle, and the number of hits recorded in a one-second time window centered around the muon crossing time. These six features will be subsequently analyzed and monitored.

The rationale behind selecting these particular features is rooted in their sensitivities to various physical phenomena. Specifically:

1. **Drift times and Crossing Angle (Slope):** Both these metrics are inherently sensitive to non-homogeneities in the electric field and variations in gas pressure, primarily because they are intricately linked to drift velocity. What distinguishes the slope from drift times is its dependency on track reconstruction. Changes in drift velocity or irregularities in the electric field can influence the calculation of the wire's distance, potentially leading to inaccurate track reconstructions.

2. **Number of Hits in the Orbit (n_{hits}):** This metric represents the number of hits recorded in a time window of approximately one second centered around the muon's passage. Beyond genuine muon hits, many spurious hits are registered. The noise rate is contingent upon environmental conditions: the LHC's noise levels are several orders of magnitude higher than our laboratory in Legnaro. However, the recorded rate of these spurious hits can also be influenced by the operational conditions of the detector.

While the monitoring methodology primarily uses the four drift times to provide granularity at the single-layer level, further enhancements in granularity can be pursued. Specifically, the chambers, each constituted of 16 cells in a layer, can be subdivided into macrocells, each comprising four cells per layer. By initiating multiple concurrent monitoring processes, granularity can be achieved at the level of individual layers within each macrocell, amplifying the granularity by a factor of four. The attributes of the specific detector under consideration intrinsically influence the granularity selection. In deploying this approach to the CMS DTs at the LHC, meticulous adjustments would be imperative to align with the experimental rate and the computational resources at hand.

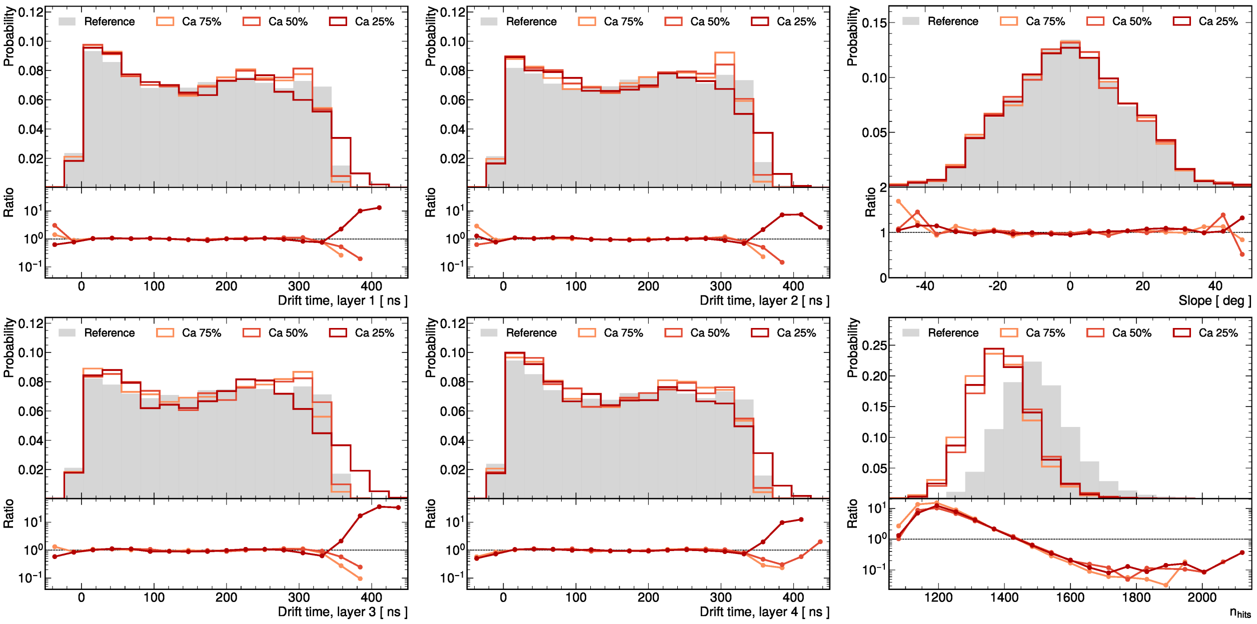


Figure 4.7: The distribution of the dataset features in the reference and in three anomalous working conditions of the cathodes voltages

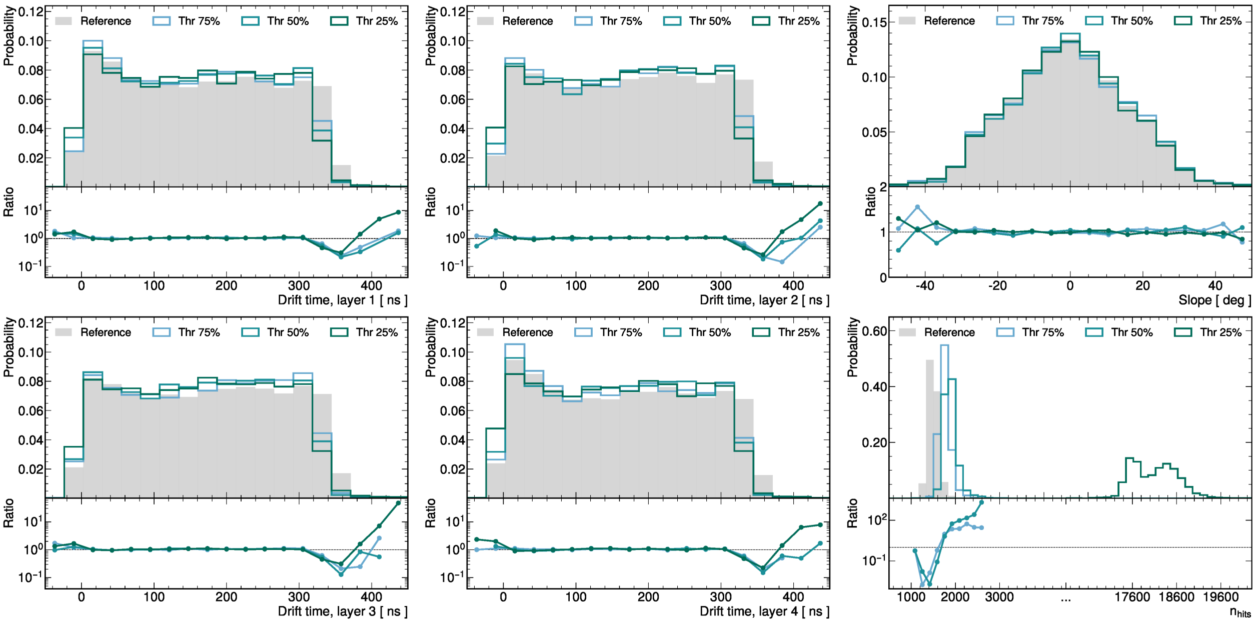


Figure 4.8: The distribution of the dataset features in the reference and in three anomalous working conditions of the front-end thresholds.

In our mock-up configuration, the preprocessing and dataset extraction phases reduce the number of rows by a factor of $\sim 10^3$, where the filter that drops all the orbits without the scintillator signal contributes with a factor of $\sim 10^2$ and the track reconstruction within the orbit contributes with the remaining factor of 10. For instance, to extract 2000 muons for the analysis, we typically start with more than 2 million hits.

We conducted a dedicated data acquisition campaign in these six anomalous configurations anticipated in Sec. 4.2.2, collecting around 10^4 events (meaning muons or equivalently scintillator signals) for each configuration. We also collected around 3×10^5 data points in the normal (or Reference) working conditions of the apparatus.[3] The plots in Figures 4.7 (*cathode* anomalies) and 4.8 (*threshold* anomalies) provide a comprehensive visualization of the dataset, including the reference dataset collected under standard working conditions of the detector and the datasets collected during the anomaly injection campaign.

4.4 Anomaly detection methodology

In our quest to monitor the quality of the unbiased data stream as a demonstrator for future New Physics searches, we employ the New Physics Learning Machine (NPLM) algorithm, detailed in Chap. 2. At its core, NPLM constructs a log-likelihood ratio test, yet it does so without prescribing a specific alternative hypothesis. Instead, it learns directly from the data. This learning process leverages universal approximators, either neural networks or kernel methods, resulting in a model-independent anomaly detection algorithm rooted in machine learning principles. We opt for the kernel-based implementation of NPLM, powered by the Falkon library [7]. This library is specifically optimized for training on GPUs, aligning with our infrastructure wherein the muon data stream is processed on the A100 GPU. As a result, the data already exists in GPU memory, facilitating an efficient transition from preprocessing to analysis. In essence, the analysis can be viewed as a continuation of the preprocessing, executed with similar computational efficiency. We would like to note that the parts of this section are published in [86].

4.4.1 Adapting the kernel-based NPLM algorithm to DQM

In the setup described in the previous section, we are interested in assessing the quality of individual batches of data collected by the apparatus, each of which is denoted as $\mathcal{D} = \{x_i\}_{i=1}^{\mathcal{N}_\mathcal{D}}$. Namely, we ask whether the statistical distribution of the data points in $\mathcal{D}$ coincides with the one expected under reference working conditions, $P(x\,|\,\mathrm{R})$. We thus aim to perform what is known in statistics as a *goodness-of-fit test.*[4] Refer to [87] for a detailed study on the NPLM algorithm in the context of classifier-based goodness-of-fit tests.

The reference distribution $P(x\,|\,\mathrm{R})$ is not available in closed form. What is available is instead a second dataset $\mathcal{R} = \{x_i\}_{i=1}^{\mathcal{N}_\mathcal{R}}$ collected by the same apparatus when operated in the reference working conditions, such that the data in $\mathcal{R}$ do follow the $P(x\,|\,\mathrm{R})$ distribution. Our goodness-of-fit test is thus carried out by comparing the two datasets $\mathcal{D}$ and $\mathcal{R}$, asking whether they are sampled from the same statistical distribution. The problem can then be formulated as a *two-sample test*, in which $\mathcal{D}$ and $\mathcal{R}$ play asymmetric roles.

The data batch $\mathcal{D}$ is what needs to be tested. Therefore its composition and size, $\mathcal{N}_\mathcal{D}$, are among the specification requirements of the DQM methodology we are developing. Typically, $\mathcal{N}_\mathcal{D} \sim 1000$ is in the ballpark of what is considered by DQM applications deployed at CMS.

The reference dataset $\mathcal{R}$ is instead created within the methodology design, with mild or no limitation on its size, $\mathcal{N}_\mathcal{R}$. A larger $\mathcal{R}$ dataset offers a more faithful representation of the underlying reference statistical distribution and, therefore, a more accurate test. Furthermore, taking $\mathcal{N}_\mathcal{R}$ larger than $\mathcal{N}_\mathcal{D}$ reduces the effect of the $\mathcal{R}$ dataset statistical fluctuation on the test outcome, leaving only those inherently due to the fluctuations of $\mathcal{D}$. This makes the outcome for a given data batch $\mathcal{D}$ nearly independent of the specific instance of the set $\mathcal{R}$ employed for the test, making the result more robust. In what follows, we will thus preferentially consider an unbalanced setup for the two datasets, with $\mathcal{N}_\mathcal{R} > \mathcal{N}_\mathcal{D}$. We will further exploit the availability

[3] Dataset available at https://doi.org/10.5281/zenodo.7128223.

[4] See Cousins' note and cited references at physics.ucla.edu/∼cousins/stats/ongoodness6march2016.pdf

of a relatively large volume of data collected under the reference working conditions for calibrating the test statistics variable and for selecting the hyperparameters, as discussed in the following. The availability of accurately labeled large datasets collected under reference detector conditions deserves further discussion. Such datasets are routinely available, especially in high-energy physics experiments, and are used for designing and calibrating traditional DQM methods [72, 73]. As discussed in Sec. 4.1.2, data is validated by a careful offline inspection, which requires human intervention. This validation process is too demanding and slow to employ as a DQM algorithm. Our purpose is to monitor the data quality online, i.e., while they are being collected. The offline validation is instead straightforwardly capable of producing labeled reference data samples that are way larger than individual data batches.

As already introduced, we employ the New Physics Learning Machine method, proposed and developed to address a similar problem in the different context of searches for new physical laws at collider experiments. The search for New Physics is performed by comparing the measured data with a reference dataset whose statistical distribution is predicted by a standard set of physical laws that supposedly describe the experimental setup. The purpose of the comparison is not to assess the data quality like in DQM. Still, the quality of the distribution prediction and, in turn, to check whether the standard laws are adequate or new physical laws are needed to model the experimental setup. However, this conceptual difference does not have practical consequences. The NPLM setup of $\mathcal{D}$ versus $\mathcal{R}$ data comparison is straightforwardly portable to DQM problems.

Using the ideas and equations laid down in Chap. 2, the design of the NPLM method for DQM works as follows. We first pick up a model for $f_{\mathbf{w}}(x)$ and select its hyperparameters. The hyperparameters selection strategy is described in the next section for the kernel-based implementation of NPLM. Next, we need to calibrate the test statistics variable,

$$t(\mathcal{D}) = 2\sum_{x\in\mathcal{D}} \log\frac{P(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}})}{P(x\,|\,\mathrm{R})} = 2\sum_{x\in\mathcal{D}} f_{\hat{\mathbf{w}}}(x)\,, \tag{4.2}$$

to be able to associate its value $t(\mathcal{D})$ to a probability $p[\,t(\mathcal{D})\,]$, the p-value. This probability will be the output of the DQM algorithm. Based on its value, the analyzer will eventually judge the quality of each data batch $\mathcal{D}$. For instance, the analyzer might define a probability threshold, below which the data batch is discarded or set apart for further analyses. Above the threshold, the batch could be retained as a good batch.

It should be noted that the selected hyperparameters and the p-value depend on the DQM problem's detailed setup under consideration. However, once these elements are made available for a given setup, they can be used to evaluate the quality of all the $\mathcal{D}$ batches in that setup. The only operation that the DQM algorithm has to perform at run-time is one single training of $\mathcal{D}$ against $\mathcal{R}$, out of which $t(\mathcal{D})$ is obtained and, in turn, $p[\,t(\mathcal{D})\,]$.

Notice that the test statistic definition in Eq. 2.22 and Eq. 4.2 do not match. Unlike in NPLM applications to new physics searches, the total number of data points in $\mathcal{D}$ is not a random variable but instead fixed to the data batch size. Therefore, in DQM applications we employ the regular likelihood rather than the extended likelihood.

Calibration is performed as follows. The test statistics are preferentially large and positive if the best-fit alternative distribution $P(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}})$ accommodates the data better than the reference distribution $P(x\,|\,\mathrm{R})$ does, signaling that the data batch is likely not thrown from $P(x\,|\,\mathrm{R})$.

Large $t(\mathcal{D})$ should thus correspond to a small probability. The precise correspondence is established by comparison with the typical values that t attains when the data batch is good. We thus compute the distribution, $P(t\,|\,\mathrm{R})$, that the t variable possesses when the data follow the reference statistical distribution and the p-value is defined as

$$p\,[\,t(\mathcal{D})\,] = \int_{t(\mathcal{D})}^{\infty} P(t'\,|\,\mathcal{R})\,\mathrm{d}t'\,. \tag{4.3}$$

The physical meaning of $p\,[\,t(\mathcal{D})\,]$ is the probability that a good data batch gives a value of t that is more unlikely (i.e., larger) than the value $t(\mathcal{D})$ produced by the batch $\mathcal{D}$. If a threshold is set on p, this threshold measures the frequency at which good data batches are not recognized as such by the algorithm. The $P(t\,|\,\mathrm{R})$ distribution is straightforwardly estimated empirically, thanks to the availability of reference-distributed labeled data points. We create several artificial data batches—called *toy* datasets—of the same size $\mathcal{N}_{\mathcal{D}}$ as the true batches. We run the training and compute t on each of them. Each toy dataset should be statistically independent and also independent from the reference dataset $\mathcal{R}$ employed for training. An extensive sample of reference-distributed data is thus used in order to produce both the toy batches and the reference dataset. By histogramming the values of t computed on the toys, we could easily obtain an estimate of $P(t\,|\,\mathrm{R})$ and hence of $p\,[\,t(\mathcal{D})\,]$.

A different procedure is adopted here, exploiting the empirical observation that $P(t\,|\,\mathrm{R})$ is well approximated by a chi-squared (χ^2) distribution, as discussed in Sec. 2.2.3. The number of degrees of freedom of the χ^2 depends on the setup but can be determined by fitting the empirical distribution of the t values computed on the toys. The survival function (one minus the cumulative) of the corresponding χ^2 distribution will be used as an estimate of $p\,[\,t(\mathcal{D})\,]$. By proceeding in this way, we will be formally able to compute very small p-values that correspond to highly discrepant data batches with very large $t(\mathcal{D})$. However, the agreement of $P(t\,|\,\mathrm{R})$ with the χ^2 cannot be verified in the high-t region, which the toys do not populate, and there is no theoretical reason to expect that this agreement will persist in that region. Our quantification of the p-value is thus only accurate in the region that the toys statistically populate. For instance, if 300 toys are thrown, only p-values larger than around $1/300$ are accurately computed. Suppose $t(\mathcal{D})$ falls in a region where our determination of p is much smaller than that. In that case, ours should be regarded as a reasonable estimate that is particularly useful to compare the level of discrepancy of different batches, but it cannot be directly validated. However, in those cases, we can ensure that $p\,[\,t(\mathcal{D})\,] \lesssim 1/300$ by directly comparing with the t values on the toys.

Another feature of the NPLM approach is the possibility of exploiting the function $f_{\hat{\mathbf{w}}}$ learned during the training task to characterize anomalous batches of data. The function $f_{\hat{\mathbf{w}}}$ represents the log-ratio between $P(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}})$ and $P(x\,|\,\mathrm{R})$ and, hence, can be used to deform and adapt the reference distribution to the data by reweighting it according to

$$P(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}}) = e^{f_{\hat{\mathbf{w}}}(x)} P(x\,|\,\mathrm{R})\,. \tag{4.4}$$

The exponential $\exp\left(f_{\hat{\mathbf{w}}}(x)\right)$ will be close to 1 if the data are well-described by the reference distribution, while it will depart from it otherwise. One should therefore be able to gain additional information about the anomalous batch by inspecting this quantity as a function of the input variables or any combination of them, even when not explicitly provided as an input feature for the training. Having access to this kind of information is a valuable element in the context of the search for New Physics [3, 12, 67], since the physics-motivated variables that one might want to inspect to explain a potential anomalous score could be some type of nontrivial combination of the input features with a clear physical meaning, such as the invariant mass of a many-body final state. For DQM applications, this analysis is less relevant since a direct visual inspection of the ratio between the binned data and reference marginal distributions is already quite informative. The user may not be interested in exploring specific high-level features in the first place. On the other hand, one can still exploit the possibility of reconstructing the data distribution using $f_{\hat{\mathbf{w}}}$ as a debugging tool to check whether the learning model correctly recognizes if the data deviates from the reference and how. Moreover, aside from this thesis's primary goal, the NPLM-DQM application's output could be exploited to study data batches that display significant deviations from the reference and, depending on the characteristics of the departures, classify them into different anomalous categories. In this respect, further investigations on a possible application extension are left for future work.

Applying NPLM to the DQM problem is more straightforward than using it for New Physics searches. For New Physics searches, one must worry about imperfections in the reference data that stem from mis-modeling the reference distribution based on the underlying standard physical laws. Including these effects in NPLM is possible but requires dedicated work and domain-specific expertise [3]. Mis-modeling is not a concern in DQM problems because no modeling is required at all: the reference-distributed data are merely collected from the same experimental apparatus and not simulated. NPLM algorithms for DQM can thus be designed more efficiently and systematically without needing extremely specialized domain knowledge. DQM applications are, however, much more computationally demanding than New Physics searches. For New Physics searches, there is typically only one dataset $\mathcal{D}$ to be analyzed. For DQM, a large flow of data batches needs to be analyzed online. Our DQM algorithm must respond on a competitive timescale to apply to that problem. The original implementation of NPLM based on neural networks is incompatible with this requirement. On the other hand, the one based on kernel methods is much faster to train on problems of comparable scale [12]. It could thus match the specification requirements for applications to LHC detectors.

The performance of the kernel-based version of NPLM stems from those of the Falkon [7] library, the core algorithm powering our implementation. The fundamental theoretical and algorithmic ideas implemented in Falkon, developed in Ref. [9, 88, 89], have already been discussed in Sec. 2.3. In short, with kernel methods, one learns functions of the form:

$$f_{\mathbf{w}}(x) = \sum_{i=1}^{\mathcal{N}} w_i \, k_\sigma(x,\, x_i)\,, \tag{4.5}$$

with $\mathcal{N} = \mathcal{N}_{\mathcal{D}} + \mathcal{N}_{\mathcal{R}}$ the total size of the training set. Here, $k_\sigma(x,\, x_i)$ is the kernel function and σ some hyperparameter. We consider Gaussian kernels defined as

$$k_\sigma(x,\, x') = e^{-||x-x'||^2\,/\,2\sigma^2}\,, \tag{4.6}$$

so that $f_{\mathbf{w}}$ is a linear combination of Gaussians of fixed width σ, centered at the training data points. The optimization of the model parameters $\mathbf{w}$ is achieved by minimising the empirical risk $L(f_{\mathbf{w}})$, plus a regularization term

$$L_\lambda = L(f_{\mathbf{w}}) + \lambda\, R(f_{\mathbf{w}})\,. \tag{4.7}$$

The empirical risk, in our case, is the one associated with the logistic loss

$$L(f_{\mathbf{w}}) = \sum_{i=1}^{\mathcal{N}} l(y_i,\, f_{\mathbf{w}}(x_i))\,, \tag{4.8}$$

where the (weighted) logistic loss is

$$l(y,\, f_{\mathbf{w}}(x)) = (1-y)\,(1+\mathcal{N}_{\mathcal{D}}/\mathcal{N}_{\mathcal{R}})\log\left(1+e^{+f_{\mathbf{w}}(x)}\right) + y\,(1+\mathcal{N}_{\mathcal{R}}/\mathcal{N}_{\mathcal{D}})\log\left(1+e^{-f_{\mathbf{w}}(x)}\right)\,. \tag{4.9}$$

We have already seen that to train the NPLM with the logistic loss, we can exploit a classical result of statistical learning: a continuous-output classifier trained to tell apart two datasets approximates—possibly up to a given monotonic transformation—the log ratio between the probability distribution of the two training sets. Thus, by assigning label $y = 0$ to the data in $\mathcal{R}$ and $y = 1$ to those in $\mathcal{D}$ the model $f_{\hat{\mathbf{w}}}$ trained with the logistic loss approaches the logarithm of $P(x \,|\, \mathrm{H}_{\hat{\mathbf{w}}}) \,/\, P(x \,|\, \mathrm{R})$. The regularization term is instead given by

$$R(f_{\mathbf{w}}) = \sum_{i,j} w_i w_j k_\sigma(x_i,\, x_j)\,, \tag{4.10}$$

and its relative importance in the optimization target in Eq. 4.7 is controlled by the hyperparameter λ.

We remark that kernel methods are non-parametric approaches because the number of parameters $\mathbf{w}$ in Eq. 4.5 increases automatically with the total number of data points. Gaussian kernel methods are universal, meaning they can recover any continuous function in the large sample limit [90, 91]. However, optimizing the function in Eq. 4.5 with the target in Eq. 4.7 requires handling an $N \times N$ matrix—the kernel matrix—with entries $k_\sigma(x_i,\, x_j)$. The computational complexity of the optimization thus scales cubically in time and quadratically space with respect to the number of training points N. These costs prevent the application to large-scale settings, and some approximation is needed. Within the Falkon library, the problem of minimizing Eq. 4.7 is formulated in terms of an approximate Newton method (see Algorithm 2 of [7]). The algorithm is based on the Nyström approximation, which is used twice. First, to reduce the size of the problem, we consider solutions of the form

$$f_{\mathbf{w}}(x) = \sum_{i=1}^{M} w_i\, k_\sigma(x,\, \tilde{x}_i)\,, \tag{4.11}$$

where $\{\tilde{x}_1,\, \ldots,\, \tilde{x}_M\} \subset \{x_1,\, \ldots,\, x_{\mathcal{N}}\}$ are called Nyström centers and are sampled uniformly at random from the input data. The number of centers $M \leq \mathcal{N}$ is a hyperparameter to be chosen. Then, Nyström approximation is again used to derive an approximate Hessian matrix

$$\tilde{\mathbf{H}} = \frac{1}{M} T \tilde{D} T^{\mathrm{T}} + \lambda\, I\,. \tag{4.12}$$

Here, T is such that $T^{\mathrm{T}} T = \tilde{K}$ (Cholesky decomposition), with $\tilde{K} \in \mathbb{R}^{M \times M}$ the kernel matrix subsampled with respect to both rows and columns. Eq. 4.12 is used as a preconditioner for conjugate gradient descent. With this strategy, the overall computational cost to achieve optimal statistical bounds is $\mathcal{O}(\mathcal{N})$ in memory and, of particular importance for our scope, $\mathcal{O}(\mathcal{N}\sqrt{\mathcal{N}} \log \mathcal{N})$ in time. It is known in the literature [8, 92] that the effect of the projection in the subspace determined by the centers is a form of regularization. On the other hand, the stochasticity of the projection can potentially lead to a subspace that does not guarantee stability. From this point of view, the inclusion of a further explicit penalty term can be used to ensure stability as needed. Indeed, the regularization level is determined by both the penalty parameter and the number of centers. These ideas are formalized and made quantitative in [8].

4.4.2 Hyperparameter selection

The selection of the three Falkon hyperparameters M, σ and λ follows the prescriptions discussed extensively in Sec. 2.3.3. In short, the hyperparameters selection employs data collected under the reference working condition and proceeds as follows.

1. The model's expressive power is controlled by the number of centers M, so it should be set as high as possible to maintain sensitivity to anomalous distributions with intricate shapes. It must also be at least as large as $\sqrt{\mathcal{N}}$ to achieve statistically optimal bounds of the training convergence. At the same time, training is faster if M is smaller.

2. The Gaussian width σ is selected as the 90th percentile of the pairwise distance between reference-distributed data points.

3. The regularization parameter λ is kept as small as possible while keeping training stable, i.e., avoiding large training times or non-numerical outputs.

Several reference-distributed toy data batches are employed for this study, each trained against the reference sample $\mathcal{R}$. The experiments performed in this work employ relatively smaller data batches (e.g., between $\mathcal{N}_{\mathcal{D}} = 250$ and $\mathcal{N}_{\mathcal{D}} = 1000$) than those considered in New Physics applications [12]. In these new conditions, we observe that the compatibility of the test statistic distribution with a χ^2 is violated for very small λ. In these cases, we raise λ until the agreement with the χ^2 is restored.

It should be emphasized that the hyperparameters selection problem for NPLM is somewhat different than for regular applications of Falkon or other types of classifiers. The hyperparameters for regular classifiers can be optimized based on the performances in the specific classification task under examination. NPLM aims instead at goodness-of-fit, namely at attaining good sensitivity to a broad class of anomalous data distributions that are unknown or specified a priori. Hence, a prior reasonable choice of the hyperparameters must be performed and cannot be re-optimized a posteriori. In particular, no re-optimization can be or has been performed to enhance the sensitivity to the specific types of anomalies considered in this work to demonstrate the method's sensitivity.

4.4.3 Alternative approaches

Goodness-of-fit and two-sample test problems are of interest in several domains of science. Many approaches exist, and developing new strategies is an active area of research. One heuristic reason to choose NPLM for DQM, among the many different options, is that it has been developed in the challenging context of New Physics searches. Prior experimental and theoretical knowledge suggests that New Physics is elusive. The target for New Physics searches is thus to spot minor departures of the actual data from the reference distribution. These departures could emerge as minor corrections to the distribution shape or as relatively large corrections like sharp peaks, which only account for a small fraction of the experimental data. Detecting such minor effects requires precisely comparing the reference distribution with large datasets, which NPLM is designed to perform. Using NPLM for DQM could thus enable more accurate data monitoring, offering sensitivity to more subtle failures of the apparatus. The number of input features in the data that are typically relevant for New Physics searches ranges from few to tens, which is an adequate number for monitoring individual detectors and detector systems fully exploiting the correlations among the variables. For comparison, methods to assess the quality of generated images target instead order thousand-dimensional input data. They could be less performant for DQM as they are designed to address a radically different problem. These heuristic considerations suggest that NPLM is a reasonable starting point for developing novel DQM algorithms based on advanced multivariate goodness-of-fit or two-sample test methods, which we advocate in this paper. On the other hand, no comprehensive comparative study of the NPLM performances is currently available. Such a comparison is beyond the scope of this paper. However, the DQM problems and datasets we study will be valuable benchmarks for

Table 4.1: Configuration of the FALKON-based NPLM hyperparameters for the five-dimensional (5D) and six-dimensional (6D) experiments. The first column details the dataset configuration. Subsequent columns signify the sizes of the reference sample and data batch fed into the NPLM algorithm. Columns three through five delineate the chosen FALKON hyperparameters. The concluding column demonstrates the best-fitting χ^2 degrees of freedom (d.o.f.) to the empirical $P(t \,|\, \mathrm{R})$ distribution.

	$\mathcal{N}_{\mathcal{R}}$	$\mathcal{N}_{\mathcal{D}}$	M	σ	λ	d.o.f.
5D	2000	250	2000	4.5	10^{-6}	40
5D	2000	500	2000	4.5	10^{-7}	83
5D	2000	1000	2000	4.5	10^{-8}	171
6D	2000	250	2000	4.5	10^{-6}	58
6D	2000	500	2000	4.5	10^{-6}	78
6D	2000	1000	2000	4.5	10^{-6}	109

future work in this direction. Recent work has initiated [87, 93] to compare NPLM with a particular class of "classifier-based" methods. The classifier-based approaches [94] are all those that entail training a classifier to tell apart $\mathcal{D}$ from $\mathcal{R}$ and using the trained classifier to construct a test statistic for the hypothesis test. A simple implementation [95] employs classification accuracy as test statistics. Following the standard pipeline for classifiers, the model is trained on a subset of the $\mathcal{D}$ and $\mathcal{R}$ datasets (the training set). Instead, the accuracy is evaluated on the remaining data (the test set). The idea is that while the accuracy will be poor (around random guess) if $\mathcal{D}$ and $\mathcal{R}$ follow the same distribution, it will be higher if their distributions differ.

NPLM is technically a classifier-based method. Its major peculiarities are the choice of the likelihood ratio test statistic and the fact that the entire datasets are employed for training and evaluating the test statistics. None of these choices is motivated by the viewpoint of the theory of classification. At the same time, they are both natural or required from the perspective of the theory of hypothesis testing that underlies the NPLM approach. Performance studies in [87] show these choices benefit sensitivity. These results partly contradict Ref. [93], which, however, employs different classification models and criteria for hyperparameters selection and uses permutation tests to estimate the sensitivity rather than computing it empirically as in NPLM. These differences are responsible for the different findings, and more work is needed for a conclusive assessment.

4.5 Results and scalability

In Sec. 4.3.3, we introduced and depicted in Fig. 4.7 and Fig. 4.8 the input data comprising six features: the four of drift times, the muon's incident angle relative to the vertical axis, and the number of hits. Notably, as illustrated in the bottom-right plots of Fig. 4.7 and Fig. 4.8, the latter feature is a discriminating factor for the anomalies scrutinized in this study. This is especially relevant for anomalies affecting thresholds; a lower threshold corresponds to heightened noise. However, at the LHC, the varying luminosity delivered to the experiment within a single run could influence this feature's value. Since it does not always correlate with detector issues, we also contemplate a scenario wherein only the initial five features guide the algorithm. This modification aids in gauging the proficiency of the NPLM DQM methodology in harnessing correlations between variables to identify anomalies, especially when their manifestations are unanticipated and subtly concealed. Indeed, pinpointing what we term 'threshold anomalies'

Table 4.2: Reported median p-values for the 5D dataset configuration across varying anomaly intensities (cathode voltages and front-end thresholds) and data batch sizes.

Cathodes voltage			**Front-end thresholds**		
Anomaly	$\mathcal{N}_{\mathcal{D}}$	median p-value	Anomaly	$\mathcal{N}_{\mathcal{D}}$	median p-value
75%	250	1.4×10^{-1}	75%	250	2.8×10^{-7}
50%	250	2.9×10^{-2}	50%	250	$< 10^{-7}$
25%	250	3.4×10^{-3}	25%	250	$< 10^{-7}$
75%	500	1.9×10^{-3}	75%	500	$< 10^{-7}$
50%	500	3.4×10^{-4}	50%	500	$< 10^{-7}$
25%	500	1.1×10^{-6}	25%	500	$< 10^{-7}$
75%	1000	$< 10^{-7}$	75%	1000	$< 10^{-7}$
50%	1000	$< 10^{-7}$	50%	1000	$< 10^{-7}$
25%	1000	$< 10^{-7}$	25%	1000	$< 10^{-7}$

by solely observing the drift times and reconstructed track's slope is challenging. A surge in noise hits could inadvertently influence track reconstruction, generating aberrations in the distributions under observation. Consequently, we adopt the notation "6D" to represent the configuration where all six input features are monitored and "5D" when the hit count is excluded.

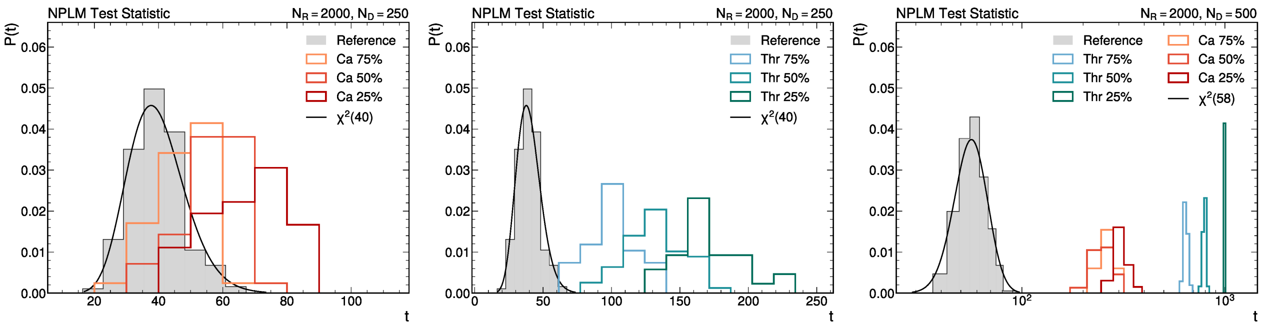
We utilize a fixed-size reference dataset of $\mathcal{N}_{\mathcal{R}} = 2000$ for both configurations, drawn from the pool of $\sim 3 \times 10^5$ muons captured under standard detector conditions. We then monitor batches of varying sizes: $\mathcal{N}_{\mathcal{D}} = 250$, $\mathcal{N}_{\mathcal{D}} = 500$, and $\mathcal{N}_{\mathcal{D}} = 1000$. This is to dissect the sensitivity of the analytical procedure to batch size fluctuations. In alignment with each of the six possible input configurations, we meticulously calibrate the Falkon hyperparameters, adhering to the blueprint presented in Sec. 4.4.2. An exhaustive summary of these configurations, including the resultant hyperparameter settings, is cataloged in Tab. 4.1.

4.5.1 Anomaly detection performance

Here, we will assess how well the NPLM can detect anomalies and determine its efficiency. After calibrating the test statistic in Eq. 4.2 using toy datasets from the standard detector conditions, we sample a few datasets (without replacement) from each anomaly category and fill a histogram with the output $t(\mathcal{D})$ values. Then, we compute the p-value of the median of the test statistic distribution using Eq. 4.3 and a fitted χ^2 distribution approximating $P(t \,|\, \mathrm{R})$.

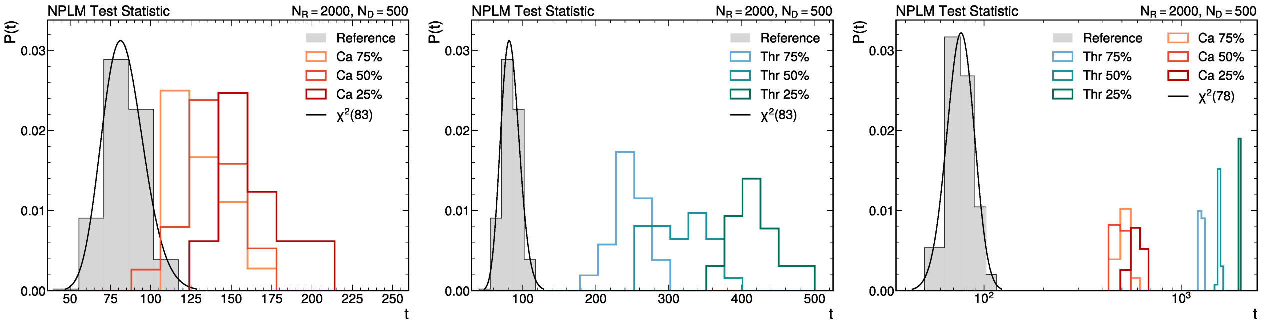
NPLM test statistic distributions Figures 4.9a, 4.10a, and 4.11a display the NPLM test statistic distribution for the cathodes anomaly class, given a 5D configuration with data batch sizes $\mathcal{N}_{\mathcal{D}}$ of 250, 500, and 1000, respectively. Likewise, for the thresholds anomaly class, the distributions are depicted in Figures 4.9b, 4.10b, and 4.11b. Furthermore, Figures 4.9c, 4.10c, and 4.11c illustrate the $t(\mathcal{D})$ distribution of both anomaly classes in the 6D configuration.

The grey histograms in the above figures represent the empirical test statistic distribution under the reference working conditions, $P(t \,|\, \mathrm{R})$. They are derived by empirically processing batches of toy data and are subsequently fit to a χ^2 distribution, which allows for the extraction of the asymptotic degrees of freedom.



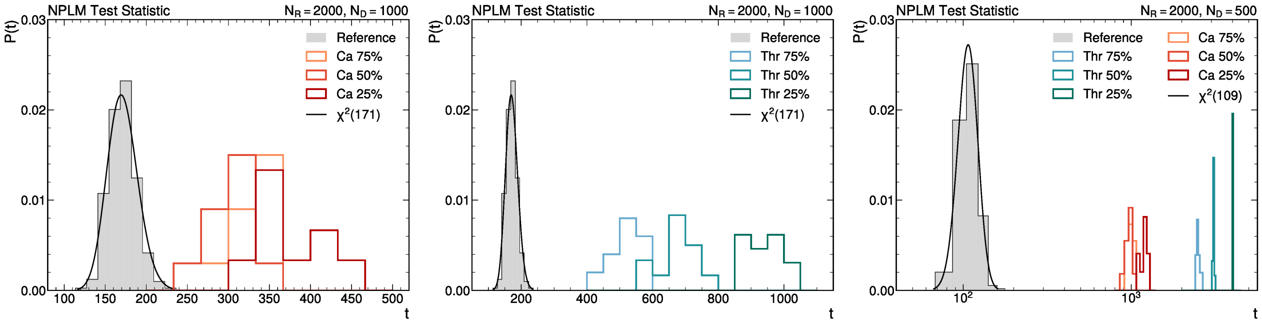
(a) 5D - cathodes **(b)** 5D - thresholds **(c)** 6D - cathodes and thresholds

Figure 4.9: Distribution of the test statistic for a data batch size of $\mathcal{N}_{\mathcal{D}} = 250$.



(a) 5D - cathodes **(b)** 5D - thresholds **(c)** 6D - cathodes and thresholds

Figure 4.10: Distribution of the test statistic for a data batch size of $\mathcal{N}_{\mathcal{D}} = 500$.



(a) 5D - cathodes **(b)** 5D - thresholds **(c)** 6D - cathodes and thresholds

Figure 4.11: Distribution of the test statistic for a data batch size of $\mathcal{N}_{\mathcal{D}} = 1000$.

From these histograms, two key observations can be made:

1. The test statistics for anomalous batches, highlighted in varying shades of red and blue for the cathodes and thresholds anomaly classes, respectively, are noticeably distinct from the reference distribution. This distinction indicates the monitoring algorithm's capability to identify anomalies.

2. The 6D dataset configuration, represented in the right panels, demonstrates enhanced anomaly detection when supplemented with the n_{hits} variable compared to the 5D configuration portrayed in the left and center panels. Consequently, subsequent discussions will primarily focus on the 5D configuration for a more comprehensive assessment of the monitoring algorithm's sensitivity.

The divergence from the reference conditions is numerically gauged using the median p-value. Specifically, this value is computed from the median of the $t(\mathcal{D})$ distribution, leveraging the asymptotic χ^2 formula as a reference point. As expected, larger batch sizes $\mathcal{N}_\mathcal{D}$ strengthen the sensitivity to anomalies. This is evident both visually, by comparing Figures 4.9, 4.10, and 4.11, and quantitatively, with smaller p-values in Tab. 4.2 for increasing $\mathcal{N}_\mathcal{D}$.

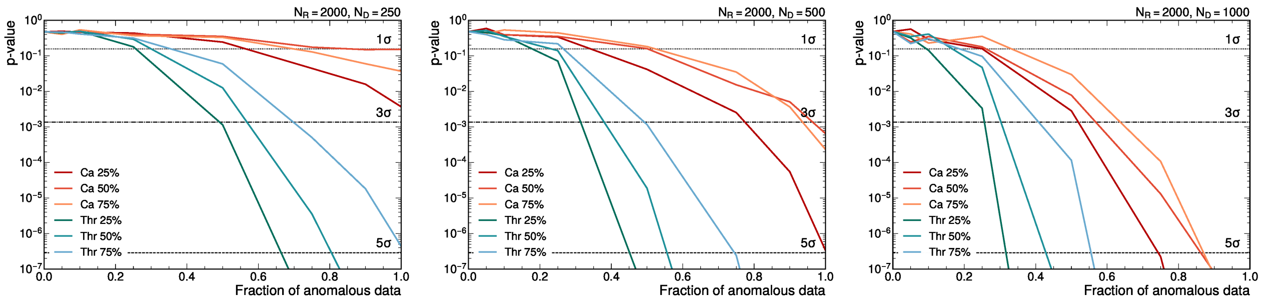
Furthermore, it is evident that as the severity of the detector failure amplifies, so does the sensitivity to anomalies. This is visually confirmed by comparing the $t(\mathcal{D})$ distributions corresponding to the 25% and 75% configurations. The former consistently displays larger test statistic values, diverging further from the reference t distribution, while the latter remains relatively closer to $P(t\,|\,\mathrm{R})$. Tab. 4.2 quantitatively reinforces this observation, as the p-values diminish progressively from 75% of the standard conditions to 25%.

Anomaly injection into reference-distributed batches We conducted a series of tests in which we purposefully manipulated data batches in order to evaluate the algorithm's ability to detect and respond to anomalous data. The initial results were based on data batches exclusively containing anomalous data, though it is not known a priori what points in the data deviate from the Reference distribution. Therefore, we restructured the data batches to gain a clearer insight into the algorithm's monitoring capabilities. We started with data from the reference working conditions and then introduced a fraction of points gathered under anomalous circumstances. This approach ensured that the majority of the dataset adhered to the reference distribution $P(x\,|\,\mathrm{R})$, with only a portion collected under anomalous conditions. This technique of integrating anomalous data into a primarily reference-based dataset is referred to as 'anomaly injection'. This procedure was executed for three batch sizes, $\mathcal{N}_\mathcal{D}$: 250, 500, and 1000. The subsequent median p-values of the distributions, as functions of the fractions of anomalous data injected, are illustrated in Figures 4.12a, 4.12b, and 4.12c, respectively.

Consistent with our prior discussions, we focus primarily on the 5D dataset configuration. It is curious to note that anomalies related to thresholds are much more noticeable. This is demonstrated by consistently reaching a significance level of 5σ in almost every configuration. On the other hand, the detectability of cathode anomalies depends on the proportion of anomalous data. Discerning these anomalies becomes challenging for scenarios where this fraction is not substantial, even for the NPLM monitoring algorithm. Reiterating a prior observation, augmenting the monitored batch size inherently facilitates anomaly detection. In the context of our 'anomaly injection' analysis, this is manifested in the rapid attainment of higher significances even when confronted with reduced fractions of anomalous data. This reaffirms our initial assertion. Noteworthily, the algorithm consistently identifies anomalies at reliable significances, especially when the anomaly's magnitude is consequential enough to be deemed problematic.

Comparison with the two-sample Kolmogorov-Smirnov test For a comparative performance assessment, we applied the Kolmogorov–Smirnov (KS) test and the NPLM model to the same dataset. Our choice to compare the NPLM's results with the KS test is based on their shared non-parametric and unbinned nature. While widely used to compare empirical distributions, the KS test has a significant limitation in multivariate scenarios. It operates on a feature-by-feature basis, ignoring potential correlations between variables. This univariate approach can miss nuances and sometimes lead to misleading results. In contrast, the NPLM algorithm is inherently multivariate and can identify discrepancies across features, offering a more comprehensive analysis.

For different data batch sizes, $\mathcal{N}_\mathcal{D}$, we recorded the KS test's median p-values for each feature. This data is presented alongside the global, five-dimensional median p-value from the NPLM test. These results can be found in Tabs. 4.3, 4.4, and 4.5 for batch sizes of $\mathcal{N}_\mathcal{D} = 250$,



(a) $\mathcal{N}_\mathcal{D} = 250$

(b) $\mathcal{N}_\mathcal{D} = 500$

(c) $\mathcal{N}_\mathcal{D} = 1000$

Figure 4.12: Median p-value plotted against the fraction of anomalous data injected into synthetic datasets. Sub-figures correspond to different batch sizes, $\mathcal{N}_\mathcal{D}$: (a) 250, (b) 500, and (c) 1000. The figures elucidate the sensitivity of the monitoring algorithm to varying concentrations of anomalies within the datasets.

500, and 1000, respectively. A clear trend emerges: as the batch size $\mathcal{N}_\mathcal{D}$ increases, the KS test's p-values decrease, suggesting that larger data batches make anomalies more detectable. Moreover, the KS test is more adept at pinpointing threshold anomalies than cathode anomalies, a trait mirrored by the NPLM. However, the KS test's p-values are noticeably larger than those of the NPLM, indicating its lesser capability to distinguish anomalies. For a visual representation, we have included figures displaying the distributions for the NPLM and KS test statistics across different anomalies and batch sizes. We remember that the NPLM test statistic distributions are displayed in Figures 4.9a, 4.10a, and 4.11a for the cathodes anomaly class with data batch sizes $\mathcal{N}_\mathcal{D}$ of 250, 500, and 1000, respectively, while in Figures 4.9b, 4.10b, and 4.11b for the thresholds anomaly class with data batch sizes $\mathcal{N}_\mathcal{D}$ of 250, 500, and 1000, respectively. Instead, the KS test statistic distributions are presented in Figures 4.13, 4.15 and 4.17 for the cathodes anomaly class with data batch sizes $\mathcal{N}_\mathcal{D}$ of 250, 500, and 1000, respectively, while in Figures 4.14, 4.16, and 4.18 for the thresholds anomaly class with data batch sizes $\mathcal{N}_\mathcal{D}$ of 250, 500, and 1000, respectively.

In conclusion, while the simpler KS test offers some insights, it is not comprehensive enough for our specific application. The NPLM, tailored to detect even slight deviations from reference distributions, proves to be the more effective tool. Lastly, every anomaly becomes starkly evident when we factor in n_{hits}. All resulting p-values drop below 10^{-7} and are hence omitted from our tables. This means that just this variable in the NPLM DQM test, or a standard KS test, can effectively pinpoint the anomalies, echoing our earlier observations.

Table 4.3: Reported NPLM median p-values for the 5D dataset configuration with $\mathcal{N}_{\mathcal{D}} = 250$ across varying anomaly intensities (cathode voltages and front-end thresholds) and KS median p-values for each of the five features, n_{hits} excluded.

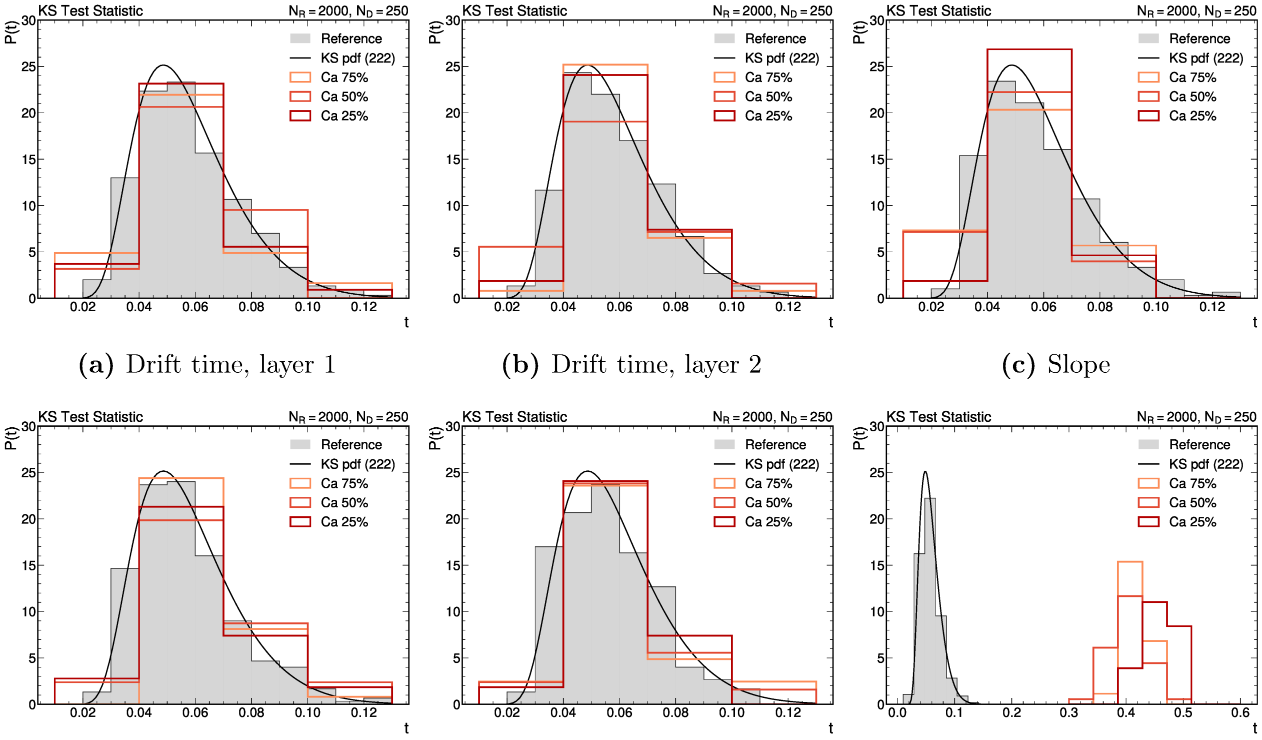
	Median p-value					
Anomaly	NPLM (5D)	KS (t^1_{drift})	KS (t^2_{drift})	KS (t^3_{drift})	KS (t^4_{drift})	KS (ϕ)
Cathode 75%	1.4×10^{-1}	0.796	0.790	0.433	0.801	0.834
Cathode 50%	2.9×10^{-2}	0.813	0.630	0.813	0.730	0.766
Cathode 25%	3.4×10^{-3}	0.834	0.766	0.784	0.760	0.778
Threshold 75%	2.8×10^{-7}	0.580	0.450	0.472	0.439	0.824
Threshold 50%	$< 10^{-7}$	0.356	0.376	0.284	0.496	0.778
Threshold 25%	$< 10^{-7}$	0.396	0.301	0.230	0.396	0.926

Table 4.4: Reported NPLM median p-values for the 5D dataset configuration with $\mathcal{N}_{\mathcal{D}} = 500$ across varying anomaly intensities (cathode voltages and front-end thresholds) and KS median p-values for each of the five features, n_{hits} excluded.

	Median p-value					
Anomaly	NPLM (5D)	KS (t^1_{drift})	KS (t^2_{drift})	KS (t^3_{drift})	KS (t^4_{drift})	KS (ϕ)
Cathode 75%	1.9×10^{-3}	0.446	0.438	0.133	0.453	0.499
Cathode 50%	3.4×10^{-4}	0.468	0.272	0.468	0.368	0.409
Cathode 25%	1.1×10^{-6}	0.499	0.409	0.431	0.402	0.424
Threshold 75%	$< 10^{-7}$	0.231	0.143	0.157	0.137	0.483
Threshold 50%	$< 10^{-7}$	0.093	0.103	0.062	0.171	0.424
Threshold 25%	$< 10^{-7}$	0.113	0.069	0.042	0.113	0.664

Table 4.5: Reported NPLM median p-values for the 5D dataset configuration with $\mathcal{N}_{\mathcal{D}} = 1000$ across varying anomaly intensities (cathode voltages and front-end thresholds) and KS median p-values for each of the five features, n_{hits} excluded.

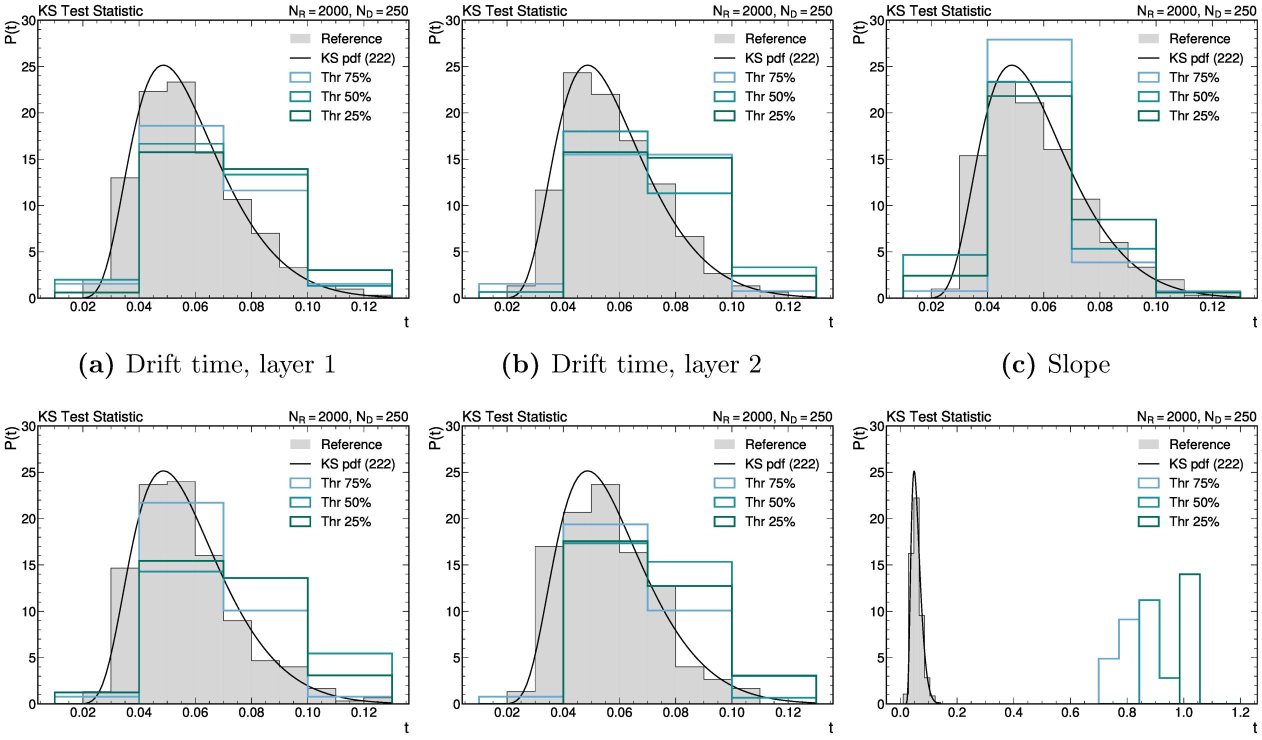
	Median p-value					
Anomaly	NPLM (5D)	KS (t^1_{drift})	KS (t^2_{drift})	KS (t^3_{drift})	KS (t^4_{drift})	KS (ϕ)
Cathode 75%	$< 10^{-7}$	0.170	0.165	0.023	0.175	0.207
Cathode 50%	$< 10^{-7}$	0.185	0.074	0.185	0.123	0.147
Cathode 25%	$< 10^{-7}$	0.207	0.147	0.161	0.143	0.156
Threshold 75%	$< 10^{-7}$	0.056	0.025	0.030	0.024	0.196
Threshold 50%	$< 10^{-7}$	0.012	0.015	0.006	0.034	0.156
Threshold 25%	$< 10^{-7}$	0.017	0.008	0.003	0.017	0.346



(a) Drift time, layer 1 **(b)** Drift time, layer 2 **(c)** Slope

(d) Drift time, layer 3 **(e)** Drift time, layer 4 **(f)** n_{hits}

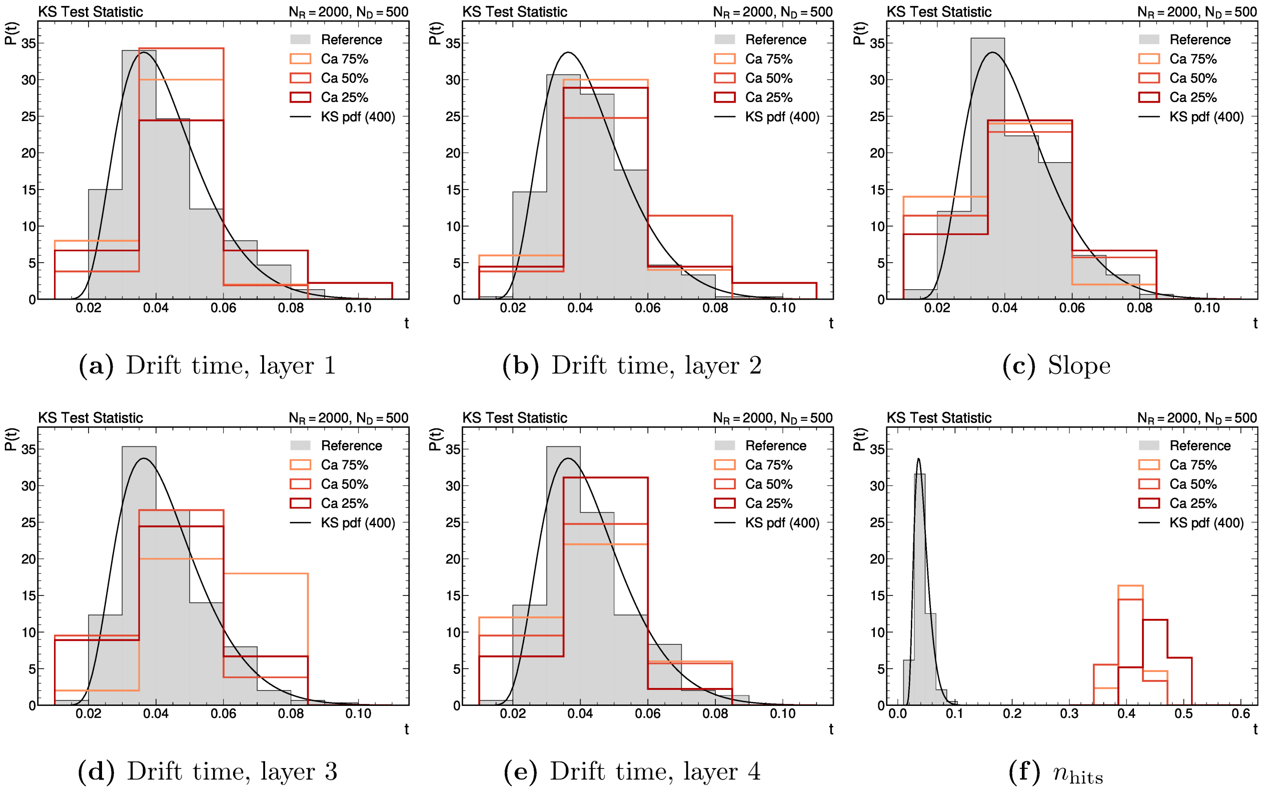
Figure 4.13: Distributions of the KS test statistics for a data batch size of $\mathcal{N_D} = 250$ with cathode anomalies (in shades of red) and the reference t distribution in gray.



(a) Drift time, layer 1 **(b)** Drift time, layer 2 **(c)** Slope

(d) Drift time, layer 3 **(e)** Drift time, layer 4 **(f)** n_{hits}

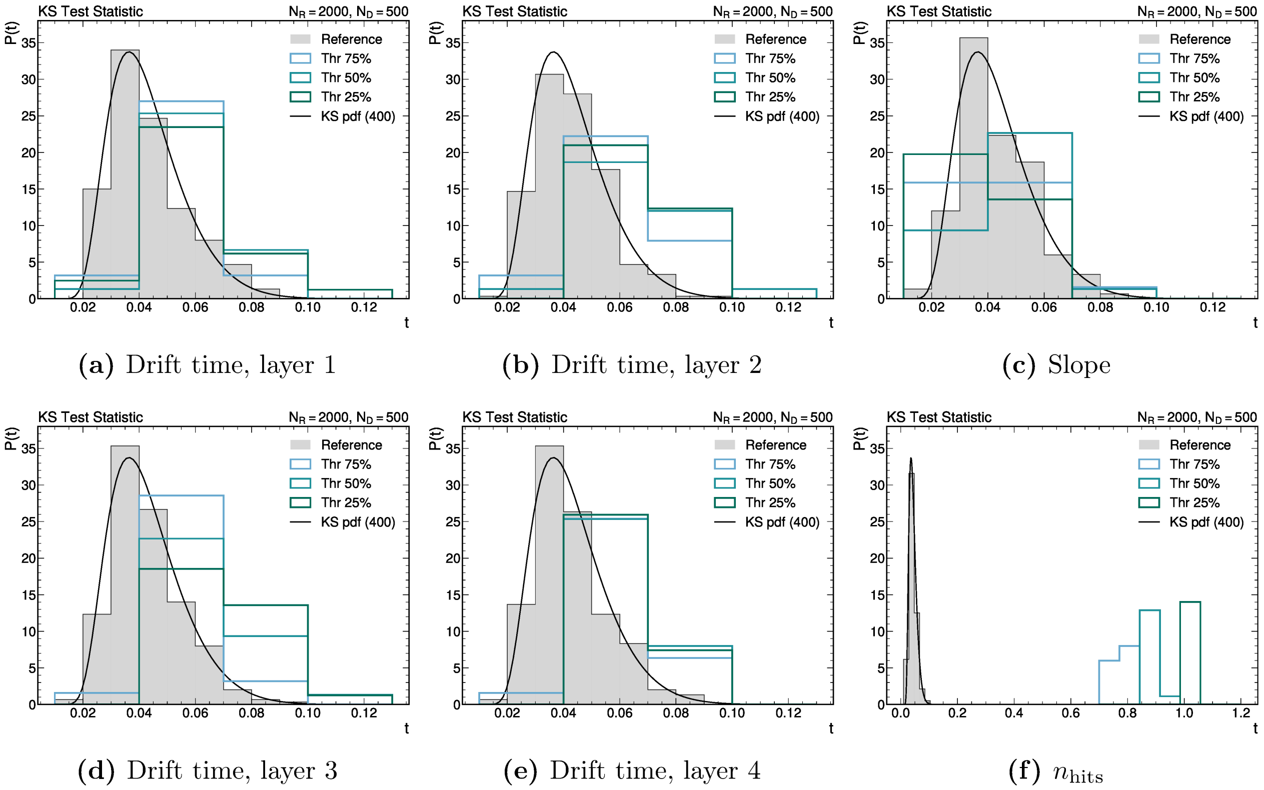
Figure 4.14: Distributions of the KS test statistics for a data batch size of $\mathcal{N_D} = 250$ with threshold anomalies (in shades of blue) and the reference t distribution in gray.



(a) Drift time, layer 1 **(b)** Drift time, layer 2 **(c)** Slope

(d) Drift time, layer 3 **(e)** Drift time, layer 4 **(f)** n_{hits}

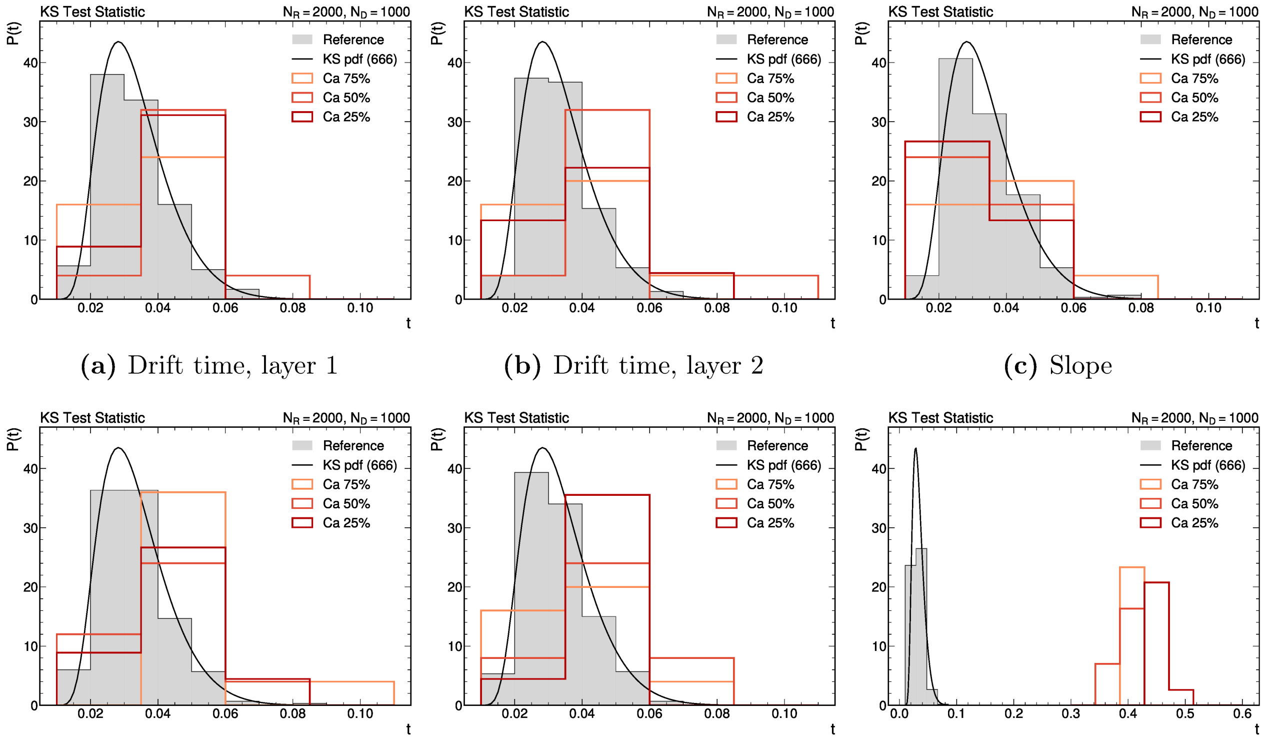
Figure 4.15: Distributions of the KS test statistics for a data batch size of $\mathcal{N}_\mathcal{D} = 500$ with cathode anomalies (in shades of red) and the reference t distribution in gray.



(a) Drift time, layer 1 **(b)** Drift time, layer 2 **(c)** Slope

(d) Drift time, layer 3 **(e)** Drift time, layer 4 **(f)** n_{hits}

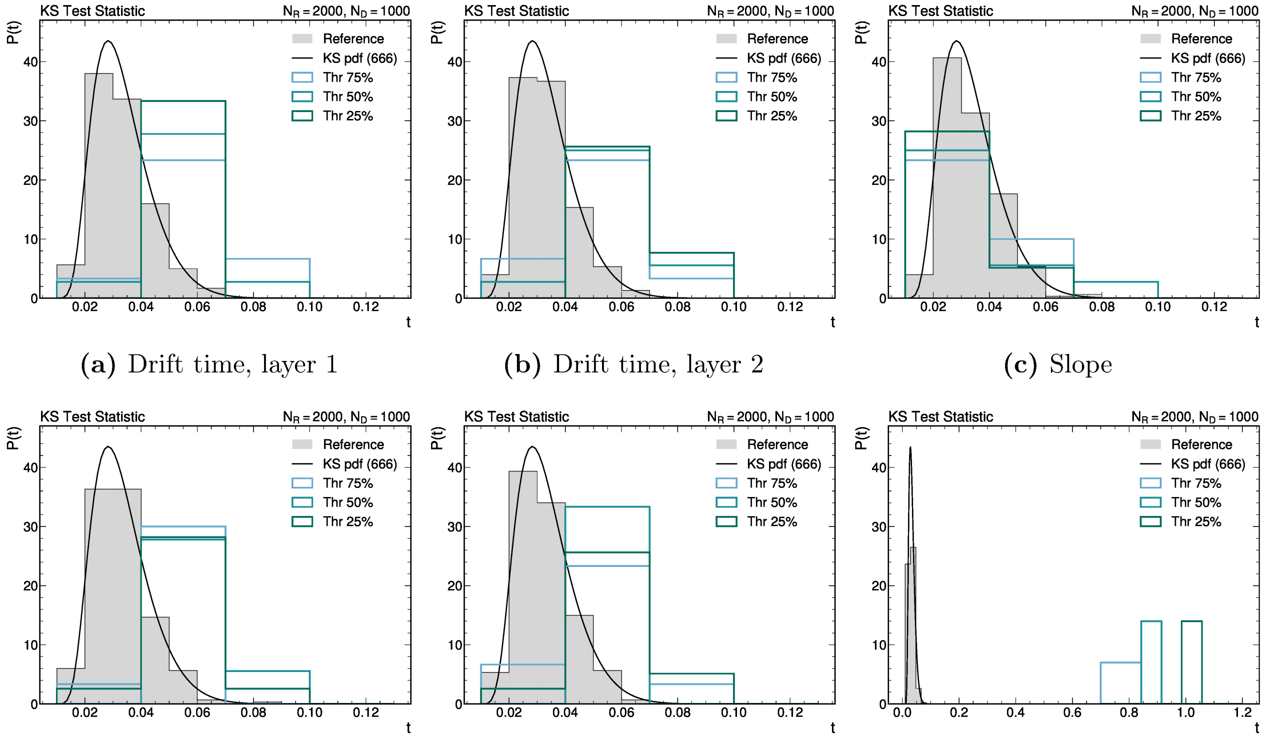
Figure 4.16: Distributions of the KS test statistics for a data batch size of $\mathcal{N}_\mathcal{D} = 500$ with threshold anomalies (in shades of blue) and the reference t distribution in gray.



(a) Drift time, layer 1 **(b)** Drift time, layer 2 **(c)** Slope

(d) Drift time, layer 3 **(e)** Drift time, layer 4 **(f)** n_{hits}

Figure 4.17: Distributions of the KS test statistics for a data batch size of $\mathcal{N}_{\mathcal{D}} = 1000$ with cathode anomalies (in shades of red) and the reference t distribution in gray.



(a) Drift time, layer 1 **(b)** Drift time, layer 2 **(c)** Slope

(d) Drift time, layer 3 **(e)** Drift time, layer 4 **(f)** n_{hits}

Figure 4.18: Distributions of the KS test statistics for a data batch size of $\mathcal{N}_{\mathcal{D}} = 1000$ with threshold anomalies (in shades of blue) and the reference t distribution in gray.

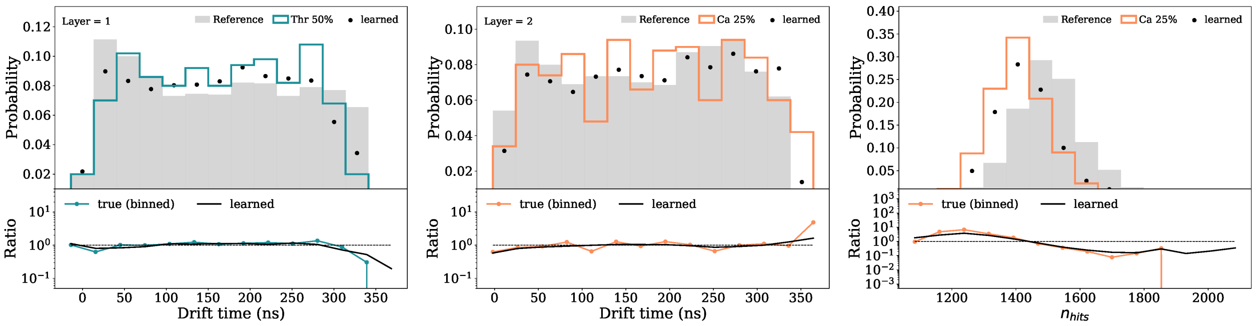


Figure 4.19: Examples of input data and respective learned likelihood ratios. The data configuration is $\mathcal{N}_\mathcal{R} = 2000$ and $\mathcal{N}_\mathcal{D} = 500$. The left panel displays the drift time distribution in the first layer of the chamber with the threshold 50% anomaly. The middle panel displays the drift time distribution in the second layer of the chamber with the cathode 25% anomaly. The right panel displays the n_{hits} distribution with cathode 25% anomaly.

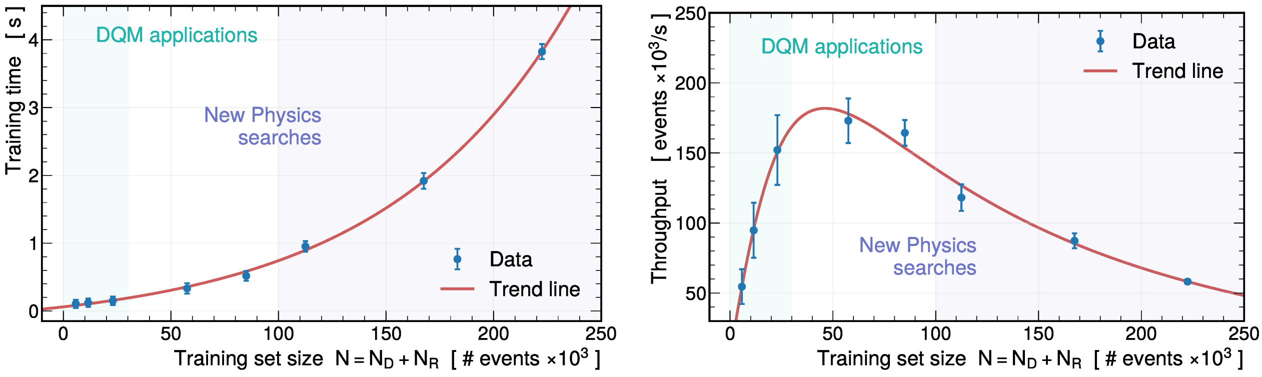
Reconstructing the learned data distribution In wrapping up this section, we focus on the data's marginal distribution as reconstructed by our model. Though such reconstructions hold immense value in searches for New Physics, their utility in DQM applications can be considered secondary. This is primarily because a direct visual comparison between the binned data and reference marginal distributions is usually insightful enough for DQM purposes. Nevertheless, the ability to reconstruct the data distribution using $f_{\hat{\mathbf{w}}}$ serves as a valuable diagnostic tool. It enables us to assess the model's ability to detect differences between the data and the reference sample.

Fig. 4.19 showcases this concept. The three plots presented are derived by reweighting each event from the reference sample used during training, using the exponential factor $\exp\left(f_{\hat{\mathbf{w}}}(x)\right)$. After binning both the reweighted reference and the actual data samples, we then calculate their ratios relative to the unaltered reference sample, displayed in the bottom panels of the figure. A side-by-side comparison of the actual data-to-reference ratio (labeled as "true") with the model's reconstructed version (labeled as "learned") sheds light on the model's accuracy. This parallel allows us to gauge the model's effectiveness in discerning anomalies, reinforcing our confidence in the machine learning task's outcomes.

4.5.2 Scalability of the algorithm to collider experiments

Having established the efficacy of the NPLM algorithm, our attention pivots to its time performance. This subsection scrutinizes the time performance as a function of the training set size, denoted as $\mathcal{N} = \mathcal{N}_\mathcal{D} + \mathcal{N}_\mathcal{R}$. This training set size largely dictates the performance intricacies of the non-parametric kernel-based NPLM. The overarching aim is real-time data quality monitoring. Thus, the algorithm should be primed to process one batch of data before the next is available. Considering its application at the LHC, where data rates surpass our current setup, adaptability becomes crucial. Continuous monitoring, down to every single hit, becomes impractical given these immense rates. Periodic monitoring, perhaps in intervals of tens of seconds or minutes, could be more feasible. During the High Luminosity phase of CMS Phase-2, a refined strategy is essential for online, trigger-less New Physics searches using DT scouting data. This may include selectively preprocessing to judiciously trim the input rate.

Delving into our evaluation metrics, we examined the 5D data configuration, methodically adjusting both $\mathcal{N}_\mathcal{R}$ and $\mathcal{N}_\mathcal{D}$. Our objective was to enlarge the total event count in the training set $\mathcal{N}$ by mainly increasing $\mathcal{N}_\mathcal{R}$, all the while ensuring the ratio $N_\mathcal{D} \,/\, \mathcal{N}_\mathcal{R}$ remains relatively sta-



(a) (b)

Figure 4.20: (a) Left panel: Training time as a function of the training set size $\mathcal{N}$ for the NPLM algorithm. (b) Right panel: Throughput, calculated as the number of muon events divided by the average training time, plotted against the training set size $\mathcal{N}$. Light green and light violet shaded areas represent regions typical for DQM applications and New Physics searches using NPLM, respectively.

ble. Then, for every unique configuration, which spanned from $\mathcal{N} \approx 2 \times 10^3$ to $\mathcal{N} \approx 230 \times 10^3$, individualized hyperparameter tuning was performed, aligning with methods elucidated in earlier sections. In practice, the NPLM algorithm underwent approximately 1000 iterations for each configuration. We captured the mean training times as our core metric and incorporated the standard deviation to provide insight into the associated uncertainty. This resulted in a comprehensive representation of training time as a function of the training set size $\mathcal{N}$, illustrated in Fig. 4.20a. Moreover, we calculated the throughput—derived by dividing the number of events (in this case, muons) by the average training time. This throughput plotted as a function of $\mathcal{N}$ is displayed in Fig. 4.20b. To further refine our visualization, we have shaded two specific ranges within the plots: a light green zone represents $\mathcal{N}$ values typical for DQM applications. In contrast, a light violet patch demarcates values commonly associated with New Physics searches using NPLM. Such distinctions visually emphasize the disparities between these regions in terms of both training time and throughput.

Upon close observation, the training time exhibits an evident exponential trend. Nonetheless, the actual durations appear promisingly short, bolstering its viability for real-time applications. The throughput analysis reveals an optimal performance point between the DQM and New Physics domains. Here, the GPU's computational power is utilized to its maximum potential without overwhelming it. Nonetheless, the performance of the kernel-NPLM appears promising. In conclusion, it's imperative to note that the available computational resources significantly influence the performance results. Additionally, real experimental conditions have a profound impact on performance requirements. While this chapter presents our demonstrator, which validates the possibility of conducting online or near-online analyses on scouting data processed in real-time, actual integration into the CMS experiment will necessitate extensive research and several years of development. This would encompass hardware upgrades, board firmware adjustments, refining the computing infrastructure, and optimizing the processing and analysis software to ensure cohesive functionality.

Chapter 5

Conclusions

Particle physics aims to reveal the building blocks of nature and understand the fundamental constituents of the universe. Particle colliders serve as sophisticated instruments in this mission, testing the Standard Model with remarkable precision. These experiments leverage state-of-the-art detector technologies and advanced statistical analyses to probe deeper into the quantum realm. This thesis identifies a pervasive challenge arising from the data collected from these experiments and the currently adopted analysis strategies.

At the heart of our exploration of particle physics is the underlying quest to unveil New Physics. To date, despite our meticulous efforts, these phenomena have remained intriguingly concealed. This thesis delves deeply into potential barriers obstructing our discovery. More than just highlighting these impediments, it contemplates their roots—pondering whether entrenched biases in our understanding or analysis techniques might be blinding us to these phenomena. While our state-of-the-art instruments, such as the particle detectors in the LHC, consistently capture intricate particle interactions, there is an emerging realization that our current frameworks might be inadvertently veiling the very signals of New Physics we seek.

In the search for New Physics, there are distinct challenges that may obstruct our discoveries. Two central issues are biases in data collection and biases in our analysis methods. The first arises from the trigger systems used in experiments. While these systems are necessary to filter and manage massive amounts of data from particle interactions, they are based on our current knowledge. This means they could unintentionally exclude valuable data pointing to New Physics. The second issue stems from the statistical tests we employ to analyze the data. Created with our current understanding of physics, these tests might not be adequately equipped to identify new, unexpected phenomena. The indispensable nature of the trigger systems in managing the deluge of data from colliders is well-understood. While these systems efficiently curate and filter data based on our current knowledge of physics, they inadvertently introduce biases that might mask potential New Physics phenomena. Recognizing this challenge, the CMS collaboration has pioneered the data scouting approach. Unlike traditional methods, data scouting allows for extracting data directly from different levels of the trigger chain as it is being collected. By capturing and analyzing online low-resolution information in this manner, the method minimizes the influence of 'physics-motivated' selections and algorithms, thereby presenting a more unadulterated view of the data. This strategy, essentially, brings us closer to the raw, unfiltered stream of particle interactions, opening doors to previously unseen events. Traditional analysis techniques, instead rooted in established statistical methodologies, often operate within the framework of a predetermined alternative hypothesis. This implies that we approach the data with preconceived notions of what New Physics might look like, thereby injecting potential biases into our analyses. A central theme of this thesis, the New Physics Learning Machine (NPLM), offers a paradigm shift. Rather than constraining our search with specific expectations, the NPLM leverages machine learning to autonomously detect anomalies in data. By refraining from specifying the alternative hypothesis and allowing the data to dictate the narrative, the NPLM heralds a new era of model-independent anomaly detection. This innovative approach ensures that our search for New Physics remains as unbiased and comprehensive as possible. The most remarkable strides in advancing our search for New Physics come when we combine these two tools, mitigating biases at every stage. This thesis combines machine learning-based model-independent anomaly detection with unfiltered scouting data. Specifically, it demonstrates the synergy between the kernel-based NPLM and real-time analysis of an unfiltered muon data stream. This seamless integration was facilitated by the CMS Drift Tubes (DT) muon chambers front-end electronics prototypes foreseen for the High Luminosity phase of the LHC, which allow for direct scouting of DT hits from the front-ends. In a controlled setting at the Legnaro National Laboratories, we replicated the CMS drift tubes on a smaller scale to simulate a cosmic muon telescope. With the assistance of OBDT prototypes, an unfiltered muon stream was scouted directly from the front-ends. The research then expanded to develop a demonstrator that provides a blueprint for computing infrastructure, efficiently using heterogeneous computing elements like FPGAs and GPUs throughout the data acquisition, processing, and analysis phases. Our chosen use-case, data quality monitoring (DQM), essentially assessed the integrity of the acquired data stream for anomalies in a model-independent manner. Notably, our approach has underscored improvements in the DQM domain, highlighting how the multivariate characteristics of the NPLM amplify anomaly detection sensitivity compared to more conventional methods. Beyond this immediate application, the core significance lies in establishing a viable, online processing anomaly detection framework. By harnessing the full, unfiltered data stream, analyzing it instantaneously, and ensuring there are no predetermined biases, this work has the potential to substantially broaden the scientific horizons of the CMS experiment.

With the impending upgrades to detectors as the High Luminosity LHC era approaches, particle physics could make groundbreaking discoveries in the future. By combining 40 MHz data scouting, real-time data processing, and advanced statistical techniques like NPLM, we embark on a deeper exploration of particle physics. It is essential to highlight, however, that the demonstrator explored in this thesis serves as a precursor to the broader deployment expected in the CMS experiment during the High Luminosity phase. The forthcoming years will be dedicated to translating this preliminary work into a tangible framework for New Physics searches at CMS. To achieve this translation, it will be necessary to adapt to the processing pipeline and computing infrastructure of the CMS experiment. While the findings of this thesis provide a robust foundation, demonstrating scalability to meet the demands of the CMS experiment, nuances in the experiment's requirements and its existing infrastructure mandate these modifications. Furthermore, significant effort will be concentrated on DT scouting in the near future. As OBDT prototypes are integrated into the CMS front-ends, the stage is being set to scout data directly from these front-ends. This direct access will pave the way for the envisioned processing and analysis demonstrator, culminating in an even more comprehensive exploration of particle physics.

Our pursuit of particle physics is driven by cutting-edge technology and refined statistical methods. Each new thesis and research paper brings us closer to our goals. Every demonstrator, test and incremental improvement helps advance our knowledge and brings us closer to groundbreaking discoveries. While the quest for New Physics remains enigmatic, what is becoming increasingly clear is the paradigm shift in how we approach particle physics research. Whether this transformative approach will indeed unveil new facets of the universe remains a question. Nevertheless, we edge closer to that horizon with every innovative tool and methodology, eager to see what lies beyond, even if its true nature will remain unknown to humankind.

Bibliography

[1] G. Grosso, *Searching for Unexpected New Physics at the LHC with Machine Learning*, Ph.D. thesis, Padua U., 1, 2023.

[2] J. Neyman and E. S. Pearson, *On the problem of the most efficient tests of statistical hypotheses*, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231** (1933) 289.

[3] R. T. d'Agnolo, G. Grosso, M. Pierini, A. Wulzer and M. Zanetti, *Learning new physics from an imperfect machine*, *Eur. Phys. J. C* **82** (2022) 275 [2111.13633].

[4] F. Scarselli and A. Chung Tsoi, *Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results*, *Neural Networks* **11** (1998) 15.

[5] T. Hofmann, B. Schölkopf and A. J. Smola, *Kernel methods in machine learning*, *The Annals of Statistics* **36** (2008) .

[6] R. Barlow, *Extended maximum likelihood*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **297** (1990) 496.

[7] G. Meanti, L. Carratino, L. Rosasco and A. Rudi, *Kernel methods through the roof: Handling billions of points efficiently*, in *Advances in Neural Information Processing Systems*, vol. 33, pp. 14410–14422, Curran Associates, Inc., 2020, 2006.10350, DOI.

[8] A. Rudi, R. Camoriano and L. Rosasco, *Less is more: Nyström computational regularization*, 2016.

[9] U. Marteau-Ferey, F. Bach and A. Rudi, *Globally convergent newton methods for ill-conditioned generalized self-concordant losses*, in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[10] D. Calandriello and L. Rosasco, *Statistical and computational trade-offs in kernel k-means*, in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[11] Z. Li, J.-F. Ton, D. Oglic and D. Sejdinovic, *Towards a unified analysis of random Fourier features*, in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 3905–3914, PMLR, 09–15 Jun, 2019.

[12] M. Letizia, G. Losapio, M. Rando, G. Grosso, A. Wulzer, M. Pierini et al., *Learning new physics efficiently with nonparametric methods*, *Eur. Phys. J. C* **82** (2022) 879 [2204.02317].

[13] E. Mobs, *The CERN accelerator complex. Complexe des accélérateurs du CERN*. 2016.

[14] O. S. Brüning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole et al., *LHC Design Report*, CERN Yellow Reports: Monographs. CERN, Geneva, 2004, 10.5170/CERN-2004-003-V-1.

[15] O. S. Brüning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole et al., *LHC Design Report*, CERN Yellow Reports: Monographs. CERN, Geneva, 2004, 10.5170/CERN-2004-003-V-2.

[16] M. Benedikt, P. Collier, V. Mertens, J. Poole and K. Schindl, *LHC Design Report*, CERN Yellow Reports: Monographs. CERN, Geneva, 2004, 10.5170/CERN-2004-003-V-3.

[17] L. Evans, *The large hadron collider*, *New Journal of Physics* **9** (2007) 335.

[18] The ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *Journal of Instrumentation* **3** (2008) S08003.

[19] The CMS Collaboration, *The CMS experiment at the CERN LHC*, *Journal of Instrumentation* **3** (2008) S08004.

[20] ALICE collaboration, K. Aamodt, A. A. Quintana, R. Achenbach, S. Acounis, D. Adamová, C. Adler et al., *The ALICE experiment at the CERN LHC*, *Journal of Instrumentation* **3** (2008) S08002.

[21] LHCB collaboration, A. A. Alves, L. M. Andrade, F. Barbosa-Ademarlaudo, I. Bediaga, G. Cernicchiaro, G. Guerrer et al., *The LHCb Detector at the LHC*, *JINST* **3** (2008) S08005.

[22] The HL-LHC Project, *LHC/ HL-LHC Plan*, 2022.

[23] I. Zurbano Fernandez et al., *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*, vol. 10/2020. 12, 2020, 10.23731/CYRM-2020-0010.

[24] CMS collaboration, V. Karimäki, M. Mannelli, P. Siegrist, H. Breuker, A. Caner, R. Castaldi et al., *The CMS tracker system project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[25] CMS collaboration, *The CMS electromagnetic calorimeter project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[26] CMS collaboration, *The CMS hadron calorimeter project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[27] CMS collaboration, *The CMS magnet project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997, 10.17181/CERN.6ZU0.V4T9.

[28] CMS collaboration, J. G. Layter, *The CMS muon project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[29] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, E. Asilar, T. Bergauer et al., *Performance of the cms muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$=13 tev*, *Journal of Instrumentation* **13** (2018) P06015.

[30] J. Hauser, *Cathode strip chambers for the cms endcap muon system*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **384** (1996) 207.

[31] M. Abbrescia, A. Colaleo, G. Iaselli, F. Loddo, M. Maggi, B. Marangelli et al., *The rpc system for the cms experiment at the lhc*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **508** (2003) 137.

[32] M. Abbrescia, G. Bruno, A. Colaleo, G. Iaselli, G. Lamanna, F. Loddo et al., *Beam test results on double-gap resistive plate chambers proposed for cms experiment*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **414** (1998) 135.

[33] T. C. collaboration, *The performance of the cms muon detector in proton-proton collisions at $\sqrt{s}$ = 7 tev at the lhc*, *Journal of Instrumentation* **8** (2013) P11002.

[34] CMS collaboration, G. L. Bayatyan, N. Grigorian, V. G. Khachatrian, A. T. Margarian, A. M. Sirunyan, S. Stepanian et al., *CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems*, Technical design report. CMS.

[35] CMS collaboration, S. Cittolin, A. Rácz and P. Sphicas, *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project*, Technical design report. CMS. CERN, Geneva, 2002.

[36] D. Contardo, M. Klute, J. Mans, L. Silvestris and J. Butler, *Technical Proposal for the Phase-II Upgrade of the CMS Detector*, tech. rep., Geneva, 2015. 10.17181/CERN.VU8I.D59J.

[37] C. CMS, *A MIP Timing Detector for the CMS Phase-2 Upgrade*, tech. rep., CERN, Geneva, 2019.

[38] CMS collaboration, *The Phase-2 Upgrade of the CMS Endcap Calorimeter*, tech. rep., CERN, Geneva, 2017. 10.17181/CERN.IV8M.1JY2.

[39] CMS collaboration, *The Phase-2 Upgrade of the CMS Muon Detectors*, tech. rep., CERN, Geneva, 2017.

[40] CMS collaboration, *The Phase-2 Upgrade of the CMS Barrel Calorimeters*, tech. rep., CERN, Geneva, 2017.

[41] CMS collaboration, *The Phase-2 Upgrade of the CMS Tracker*, tech. rep., CERN, Geneva, 2017. 10.17181/CERN.QZ28.FLHW.

[42] A. Triossi, A. Navarro Tobar, D. Redondo, C. Fernández Bedoya, J. Puerta-Pelayo, I. Redondo et al., *Electronics Developments for Phase-2 Upgrade of CMS Drift Tubes*, in *Proceedings of Topical Workshop on Electronics for Particle Physics — PoS(TWEPP2018)*, vol. 343, p. 035, 2019, DOI.

[43] J. Javier Sastre-Alvaro, A. Triossi, A. Bergnoli, A. Griggio and D. Redondo Ferrero, *The OBDT board: A prototype for the Phase 2 Drift Tubes on detector electronics*, in *Proceedings of Topical Workshop on Electronics for Particle Physics — PoS(TWEPP2019)*, vol. 370, p. 115, 2020, DOI.

[44] CMS collaboration, *The Phase-2 Upgrade of the CMS Level-1 Trigger*, tech. rep., CERN, Geneva, 2020.

[45] J. Duarte, *Fast Reconstruction and Data Scouting*, in *4th International Workshop Connecting The Dots 2018*, 8, 2018, 1808.00902.

[46] CMS collaboration, S. Mukherjee, *Data Scouting and Data Parking with the CMS High level Trigger*, vol. EPS-HEP2019. 2020, 10.22323/1.364.0139.

[47] CMS collaboration, D. Anderson, *Data Scouting in CMS*, *PoS* **ICHEP2016** (2016) 190.

[48] CMS collaboration, *Search for Narrow Resonances using the Dijet Mass Spectrum in pp Collisions at sqrt s of 7 TeV*, tech. rep., CERN, Geneva, 2012.

[49] CMS Collaboration collaboration, A. M. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, T. Bergauer, M. Dragicevic et al., *Search for a narrow resonance lighter than 200 gev decaying to a pair of muons in proton-proton collisions at* $\sqrt{s} = 13$ TeV, *Phys. Rev. Lett.* **124** (2020) 131802.

[50] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, T. Bergauer, M. Dragicevic et al., *Search for dijet resonances using events with three jets in proton-proton collisions at s=13tev*, *Physics Letters B* **805** (2020) 135448.

[51] CMS collaboration, A. Hayrapetyan et al., *Observation of the rare decay of the* η *meson to four muons.* 5, 2023, [2305.04904].

[52] Badaro, Gilbert, Behrens, Ulf, Branson, James, Brummer, Philipp, Cittolin, Sergio, Da Silva-Gomes, Diego et al., *40 mhz level-1 trigger scouting for cms*, *EPJ Web Conf.* **245** (2020) 01032.

[53] R. Ardino, C. Deldicque, M. Dobson, S. Giorgetti, G. Grosso, T. James et al., *A 40 mhz level-1 trigger scouting system for the cms phase-2 upgrade*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1047** (2023) 167805.

[54] C. Collaboration, *The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger*, tech. rep., CERN, Geneva, 2021.

[55] CMS collaboration, T. O. James, *The Level 1 Scouting system of the CMS experiment*, tech. rep., CERN, Geneva, 2023.

[56] Badaro, Gilbert, Behrens, Ulf, Branson, James, Brummer, Philipp, Cittolin, Sergio, Da Silva-Gomes, Diego et al., *40 mhz level-1 trigger scouting for cms*, *EPJ Web Conf.* **245** (2020) 01032.

[57] T. O. James, "The Level 1 Scouting system of the CMS experiment." 2022.

[58] T. O. James, "Real-time deep learning inference and FPGA based processing for level 1 trigger scouting at CMS." 2023.

[59] Micron Technology Inc., *Micron deep learning accelerator software development kit*, 2023.

[60] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro et al., *Tensorflow: Large-scale machine learning on heterogeneous systems*, 2015.

[61] J. Duarte, S. Han, P. Harris, S. Jindariani, E. Kreinar, B. Kreis et al., *Fast inference of deep neural networks in FPGAs for particle physics*, *Journal of Instrumentation* **13** (2018) P07027.

[62] M. Migliorini, J. Pazzini, A. Triossi, M. Zanetti, A. Zucchetta and on behalf of the CMS collaboration, *Trigger-less readout and unbiased data quality monitoring of the cms drift tubes muon detector*, *Journal of Instrumentation* **18** (2023) C01003.

[63] P. Moreira, R. Ballabriga, S. Baron, S. Bonacini, O. Cobanoglu, F. Faccio et al., *The GBT Project*. 2009, 10.5170/CERN-2009-006.342.

[64] S. Baron, J. P. Cachemiche, F. Marin, P. Moreira and C. Soos, *Implementing the GBT data transmission protocol in FPGAs*. 2009, 10.5170/CERN-2009-006.631.

[65] M. Rocklin, *Dask: Parallel computation with blocked algorithms and task scheduling*, in *Proceedings of the 14th python in science conference*, no. 130-136, Citeseer, 2015.

[66] M. J. Sax, *Apache Kafka*, pp. 1–8. Springer International Publishing, Cham, 2018. 10.1007/978-3-319-63962-8-196-1.

[67] R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer and M. Zanetti, *Learning multivariate new physics*, *Eur. Phys. J. C* **81** (2021) 89 [1912.12155].

[68] A. A. Pol, *Machine Learning Anomaly Detection Applications to Compact Muon Solenoid Data Quality Monitoring*, Ph.D. thesis, LRI, Paris 11, 2020.

[69] A. A. Pol, G. Cerminara, C. Germain, M. Pierini and A. Seth, *Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider*, *Comput. Softw. Big Sci.* **3** (2019) 3 [1808.00911].

[70] CMS collaboration, A. A. Pol, V. Azzolini, G. Cerminara, F. De Guio, G. Franzoni, C. Germain et al., *Deep learning for certification of the quality of the data acquired by the CMS Experiment*, *J. Phys. Conf. Ser.* **1525** (2020) 012045.

[71] Azzolin, Virginia, Andrews, Michael, Cerminara, Gianluca, Dev, Nabarun, Jessop, Colin, Marinelli, Nancy et al., *Improving data quality monitoring via a partnership of technologies and resources between the cms experiment at cern and industry*, *EPJ Web Conf.* **214** (2019) 01007.

[72] M. Rovere and on behalf of the CMS Collaboration, *The data quality monitoring software for the cms experiment at the lhc*, *Journal of Physics: Conference Series* **664** (2015) 072039.

[73] Azzolini, Virginia, Broen van, Besien, Bugelskis, Dmitrijus, Hreus, Tomas, Maeshima, Kaori, Javier Fernandez, Menendez et al., *The data quality monitoring software for the cms experiment at the lhc: past, present and future*, *EPJ Web Conf.* **214** (2019) 02003.

[74] M. Borisyak, F. Ratnikov, D. Derkach and A. Ustyuzhanin, *Towards automation of data quality system for cern cms experiment*, *Journal of Physics: Conference Series* **898** (2017) 092041.

[75] N. Amapane, M. Antonelli, F. Anulli, G. Ballerini, L. Bandiera, N. Bartosik et al., *Study of muon pair production from positron annihilation at threshold energy*, *Journal of Instrumentation* **15** (2020) P01036.

[76] F. Gonella and M. Pegoraro, *The MAD, a Full Custom ASIC for the CMS Barrel Muon Chambers Front End Electronics*. 2001, 10.5170/CERN-2001-005.204.

[77] RD12 collaboration, B. G. Taylor, *Timing distribution at the LHC*. 2002, 10.5170/CERN-2002-003.63.

[78] J. Hegeman, J.-M. André, U. Behrens, J. Branson, O. Chaze, S. Cittolin et al., *The cms timing and control distribution system*, in *2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–3, 2015, DOI.

[79] M. Migliorini, J. Pazzini, A. Triossi, M. Zanetti and A. Zucchetta, *Muon trigger with fast neural networks on fpga, a demonstrator*, *Journal of Physics: Conference Series* **2374** (2022) 012099.

[80] R. Mittal, A. Shpiner, A. Panda, E. Zahavi, A. Krishnamurthy, S. Ratnasamy et al., *Revisiting network support for rdma*, `1806.08159`.

[81] "RAPIDS: GPU Accelerated Data Science." `https://rapids.ai/`.

[82] "cuDF." `https://docs.rapids.ai/api/cudf/stable/`.

[83] "CuPy." `https://cupy.dev/`.

[84] "Numba." `https://numba.pydata.org/`.

[85] "CUDA." `https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html`.

[86] G. Grosso, N. Lai, M. Letizia, J. Pazzini, M. Rando, L. Rosasco et al., *Fast kernel methods for data quality monitoring as a goodness-of-fit test*, *Machine Learning: Science and Technology* **4** (2023) 035029.

[87] G. Grosso, M. Letizia, M. Pierini and A. Wulzer, *Goodness of fit by Neyman-Pearson testing*, `2305.14137`.

[88] A. Rudi, L. Carratino and L. Rosasco, *Falkon: An optimal large scale kernel method*, in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017, `1705.10958`, DOI.

[89] U. Marteau-Ferey, D. Ostrovskii, F. Bach and A. Rudi, *Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance*, in *Proceedings of the Thirty-Second Conference on Learning Theory*, vol. 99 of *Proceedings of Machine Learning Research*, pp. 2294–2340, PMLR, 25–28 Jun, 2019, https://proceedings.mlr.press/v99/marteau-ferey19a.html.

[90] C. A. Micchelli, Y. Xu and H. Zhang, *Universal kernels*, *Journal of Machine Learning Research* **7** (2006) 2651.

[91] A. Christmann and I. Steinwart, *Support vector machines*. Springer, 2008.

[92] S. Manzhos and M. Ihara, *Rectangularization of gaussian process regression for optimization of hyperparameters*, 2022.

[93] P. Chakravarti, M. Kuusela, J. Lei and L. Wasserman, *Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests*, 2102.07679.

[94] J. H. Friedman, *On multivariate goodness of fit and two sample testing*, *eConf* **C030908** (2003) THPD002.

[95] D. Lopez-Paz and M. Oquab, *Revisiting classifier two-sample tests*, in *International Conference on Learning Representations*, (Toulon, France), Apr., 2017, https://inria.hal.science/hal-01862834.