The Search for Dark Photons at LHCb and Machine Learning in Particle Physics

by

Constantin Niko Weisser

MPhys, University of Manchester (2015)

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Physics, Statistics, and Data Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author . . . .
Department of Physics
April 2, 2021

Certified by . . . .
Mike Williams
Associate Professor
Thesis Supervisor

Accepted by . . . .
Deepto Chakrabarty
Associate Department Head

The Search for Dark Photons at LHCb and Machine Learning in Particle Physics

by

Constantin Niko Weisser

Submitted to the Department of Physics
on April 2, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Physics, Statistics, and Data Science

Abstract

Investigating hypothetical particles called dark photons helps shed light on the nature of dark matter, which is one of the biggest open questions in particle physics. This thesis presents world-leading limits in searches for prompt-like and long-lived dark photons decaying into two muons, as well as other dimuon resonances, produced in proton-proton collisions and collected by the LHCb experiment at the Large Hadron Collider at CERN.

In addition, this thesis proposes various machine and deep learning techniques and their applications to particle physics: classifier bias on a continuous feature can be controlled more flexibly with a novel moment decomposition loss function than with simple decorrelation, which can enhance bump hunt sensitivity; the first high precision generative model approach to high energy physics simulation has potential to help close the gap between pledged and required resources; we developed a simple, powerful, and novel deep learning approach to vertexing, a technique to determine the location of vertices of sprays of particles, given particle tracks; the statistics chapter is concluded by a pedagogical study of using machine learning classifiers for multivariate goodness-of-fit and two-sample tests.

Thesis Supervisor: Mike Williams
Title: Associate Professor

Acknowledgments

Every PhD student thanks their advisor. Prof Mike Williams, however, deserves this thanks more than most. From when I first stumbled into his office to the day I submit this thesis, he has always been more than supportive. Even when he had IAIFI proposals to write and IAP classes to teach, his door was always open. He listened to my ideas and patiently corrected mistakes in my thinking. Most importantly, he has shown flexibility regarding my projects, work location, and the internship opportunity I grew from. If I become even half as creative an analyst, proficient a writer, or captivating presenter as Mike, I will have succeeded.

I'm lucky to have found such a nurturing environment at LHCb and MIT LNS. Thanks to Tom Boettcher for letting me bounce ideas off him at CERN and MIT. To Yunjie Yang for countless conversations about detectors, our futures and the state of the world. To John Hardin for guidance and losing stick fights. To Yimin Wang for navigating professional life after graduation. To Phil Ilten and Dan Craik for discussing LHCb software-related issues. To William Holmkvist, Sharada Mohanty and Adam Dendek for keeping Geneva fun. To the LHCb collaboration for jokes at 4am in the control room. And to all the penthouse for being delightfully distracting.

I'm grateful for all the friends during grad school that have made life colorful. Thanks to all of my classmates who welcomed me at MIT including Dahlia Klein, Sameer Abraham, and Cyprian Lewandowski. To Pancho Eliott, Georgia Smits and Jordan Daniels for being my rock over the years. To the triathlon team that pushed me to my Ironman limits. To the MIT ballroom dance team for a positive flow of energy when research was slow, and my dance partners Miray Omurtak and Charlotte Emily Ryan for keeping me on my toes. Lastly, to the Alpacas for touring sights, game nights and socially-distanced snowball fights.

To my girlfriend, Xilian Sansoucy, for putting my life into perspective and a smile on my face every single day. Thanks for putting up with me (writing this thesis).

Thanks to my dad for the fun calls and inspiration and my mum who encouraged my curiosity from an early age and has never ever left my corner. This is for you all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

“The love of complexity without reductionism makes art; the love of complexity with reductionism makes science.”

—— E. O. Wilson [1]

To understand complex problems, humans often break them into smaller blocks to tackle them separately. The hope is that once the sub-problems are solved, an understanding of the whole emerges.

Particle physics is the modern approach to the ancient pursuit to understand the universe and make predictions by investigating the properties and interactions of the smallest building blocks of matter and radiation. The culmination of this century-long quest is summarized by the celebrated Standard Model of particle physics (see Chapter 2.1). However, there are some glaring holes in this model. It does not, for example, explain the nature of dark matter – a substance that constitutes 80% of the matter in our universe. One step towards solving this riddle is searching for the hypothetical particle called the dark photon (see Chapter 2.3).

To tackle this problem, my work has focused on the data science process of exploring the generative statistical model of nature. The LHCb experiment at the Large Hadron Collider at CERN that I am a part of does this by colliding particles and observing probabilistic outcomes with camera-like detectors. (see Chapter 3).

The main contributions of this thesis include:

- Development of various Machine Learning algorithms that can facilitate and speed up progress in particle physics, and other fields. (see Chapter 4)
- Searches for resonances in the spectrum of two oppositely charged muons at LHCb to test dark photon and other models. (see Chapter 5)

Chapter 2

Theory and Motivation

To set the scene for the rest of this thesis, this chapter will cover the necessary theory. First, a brief overview of the Standard Model of particle physics is given. Then Dark Matter will be covered, which the Standard Model does not explain.

2.1 The Standard Model

The Standard Model of particle physics (SM) unites all known fundamental particles and three of the four known fundamental forces in one model (see figure 2-1). It is formulated using quantum field theory (QFT), a theoretical framework that incorporates special relativity, quantum mechanics and classical field theory. Various quantum fields permeate all of space, while particles are excitations of these fields. All known matter and radiation is built up from these building blocks. For an in-depth treatment see Refs. [2] and [3].

The three forces the Standard Model tackles are electromagnetism, weak and strong nuclear forces [4]. Excitations in the corresponding fields create the gauge bosons: There are electrically charged massive W^+ and W^- bosons that act only on left-handed particles and right-handed antiparticles, a massive Z boson and one massless photon. Together these are associated with the unification of electromagnetism and the weak force, the electroweak interaction. Electromagnetism by itself can be described by quantum electro dynamics (QED). In addition, there are eight massless

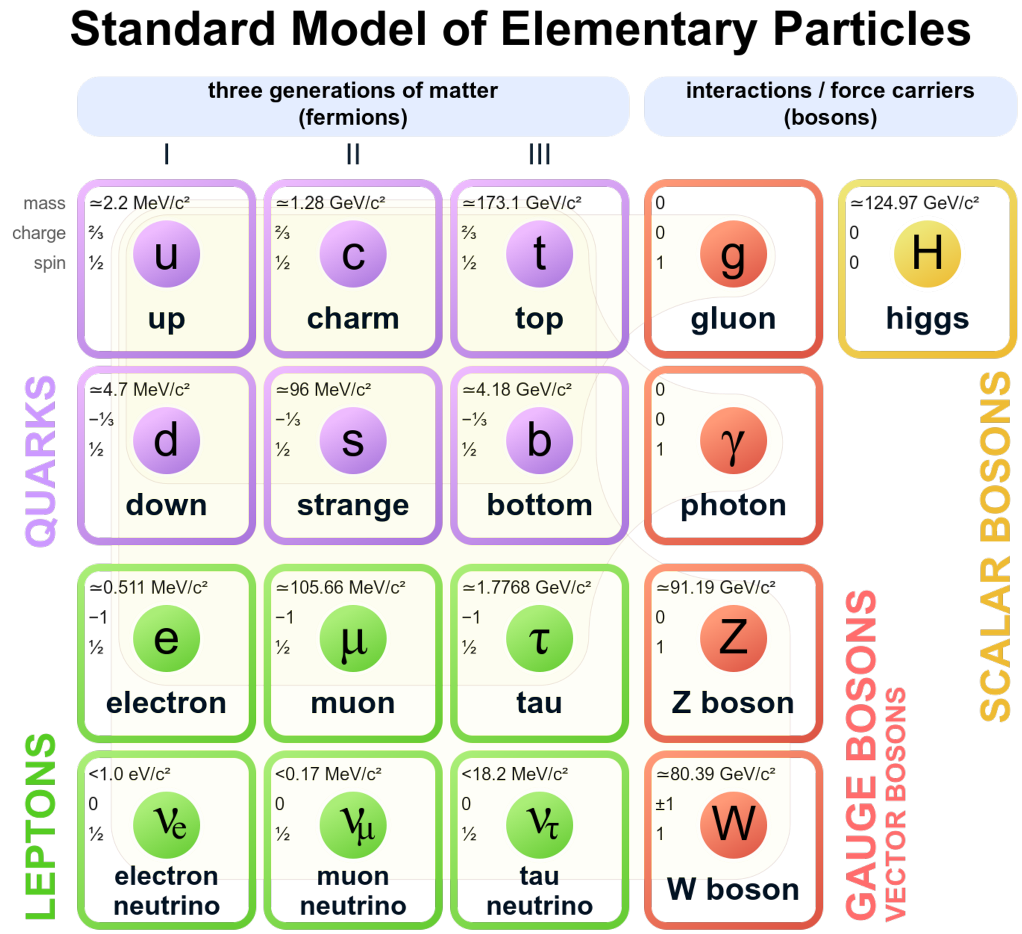


Figure 2-1: Quarks, leptons and bosons are the fundamental particles that make up the Standard Model of particle physics (Credit: Wikimedia Commons).

gluons associated with the strong interaction and modelled by quantum chromo dynamics (QCD). Gauge bosons carry a spin of 1, which means that they do not follow the Pauli exclusion principle and there is no limit on their spatial density. These particles are the messengers that mediate the forces, e.g. if two electrons interact electromagnetically they exchange photons, the force carriers of QED.

Not every force acts on all fundamental particles. Instead, we say a particle is charged under a certain interaction, or couples to it, if it is affected by that force, with the charge value proportional to the strength of the interaction. In fact, the Standard Model classifies particles by their charges under the various forces as seen in figure 2-1.

The 12 elementary matter particles that carry spin 1/2 are known as fermions, and respect the Pauli exclusion principle. For each of these matter particles, an antiparticle of the same mass, but opposite charges exists and all of them are affected by the weak force. The fermions are grouped into three so-called generations. Particles of the same class (same rows in figure 2-1) have the same charges, but their mass grows with increasing generation as signified by moving rightwards in figure 2-1. By the conservation of energy in its rest frame, higher mass particles can decay into lower mass ones if permitted by other conservation laws, but the reverse cannot occur. As first generation particles have the lowest masses and cannot decay, they constitute the building blocks for ordinary matter. The atom, for example is built up from electrons, protons and neutrons, which themselves are built from up and down quarks. The neutrinos are the exception to this rule as explained below.

The fermions are further subdivided into six quarks [up (u), down (d), charm (c), strange (s), top (t), bottom (t)] and six leptons [electron (e), electron neutrino (ν_e), muon (μ), muon neutrino (ν_μ), tau (τ), tau neutrino (ν_τ)]. The quarks interact via QCD and QED and therefore hold color and electromagnetic charge. One particularity of the strong force is that its attractive force gets stronger at larger distances. Hence, color-charged particles have to be accompanied by other particles rather than be isolated, so that the thereby-formed groups, called hadrons, have no overall color charge. This phenomenon is called confinement. Types of hadrons include quark-antiquark pairs called mesons of which B^0, B^+ and B_c^+ are examples with quark contents of ($\mathrm{d\bar{b}}$), ($\mathrm{u\bar{b}}$) and ($\mathrm{c\bar{b}}$), respectively. Baryons consisting of three (anti)quarks also are hadrons and include the proton and Λ_b^0 with quark contents (uud) and (udb). The different species of quarks are referred to as flavors. Flavor physics concerns itself with using mesons and baryons to study phenomena of interest, such as asymmetries between matter and antimatter. As leptons, electrons and their higher generation counterparts are electrically, but not color charged, allowing me to write this thesis on a computer. Neutrinos, however, only interact via the weak force, which makes them notoriously difficult to detect. Each flavor state (e.g. ν_e) is a superposition of three mass states and neutrinos oscillate between different flavor states.

As described by the Higgs mechanism, the Higgs field is a scalar (spin 0) field that through spontaneous symmetry breaking gives mass to particles that interact with it [5, 6]. The excitation of this field, the Higgs Boson, was discovered by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) at CERN [7, 8]. This completed the Standard Model, which can be compactly mathematically summarized by the Lagrangian density below:

$$\begin{aligned} \mathcal{L} = &-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} && I \\ &+ i\bar{\psi}\not{D}\psi + h.c. && II \\ &+ \bar{\psi}_i y_{ij}\psi_j\phi + h.c. && III \\ &+ |D_\mu\phi|^2 && IV \\ &- V(\phi) && V \end{aligned}$$

Line I posits the existence of the fundamental forces and their corresponding gauge bosons. Line II describes the coupling of particles to force carriers. Non-neutrino fermions (gauge bosons) acquire their masses via interactions with the Higgs boson by the term on line III (IV). Lastly, line V denotes the Higgs field.

The Standard Model predicted the confirmed existence of the gluon, W, Z and Higgs bosons, and the top and charm quarks and many of their properties to high precision. However, it is an incomplete description of nature and is missing the following pieces amongst others:

1. a quantized theory of gravity
2. neutrino oscillation and mass generation mechanisms
3. a mechanism to generate baryon asymmetry
4. a dark energy and a dark matter candidate

2.2 Dark Matter

The universe is expanding at an ever-accelerating rate alongside space itself and is cooling in the process [9, 10]. This suggests that the universe was denser and hotter in the past.

[11]. This substance does not interact with the Standard Model strong and electromagnetic forces and has, thereby, evaded direct observation with telescopes. As the nature of this substance is not known, it is given a placeholder name: Dark Matter (DM). The determination of its nature is one of the most fundamental and sought after goals of particle physics.

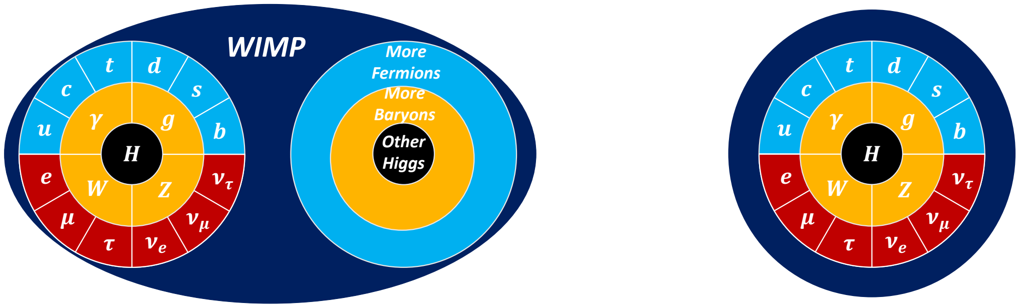
One possible objection to the proposal of dark matter is that the problem of galaxy dynamics could be solved by modifying the laws of gravity instead. However, the locality of these effects as demonstrated by gravitational lensing, evidence from the cosmic microwave background and galaxy formation amongst others make this an unfavored alternative [12–16].

The prevailing cosmological model of the observable universe is the flat ΛCDM Big Bang model. It estimates that dark matter makes up 85% of the matter in the universe [14]. The aim of this thesis is to help shed some light on its nature.

2.2.1 Paradigms

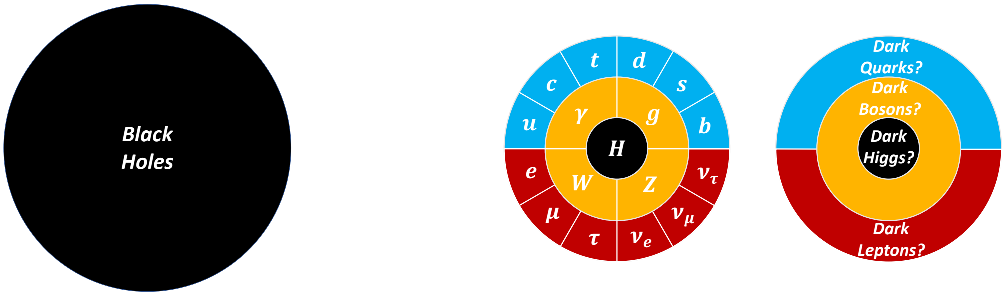
While the Standard Model lacks a dark matter candidate, there are many different theoretical models that provide them. These models can be classified into different paradigms. A larger symmetry group could encompass the SU(3) x SU(2) x U(1) structure of the Standard Model and extend it, yielding further particles, such as a dark matter candidate (see figure 2-2a). Even though this approach can solve many problems in particle physics at once, many attempts have been made to find evidence of examples of such theories like supersymmetry at the LHC and direct detection experiments, but no discovery has been made [17].

Alternatively, rather than requiring a new symmetry group, the SM could be extended by additional particles like the QCD axion as in figure 2-2b [18]. Another



(a) Superset symmetry groups

(b) Extensions of SM



(c) Black holes

(d) Dark sectors

Figure 2-2: There is a plethora of different dark matter paradigms as shown by these pictorial representations of dark matter paradigms (Figures adapted from Mike Williams).

idea portrayed in figure 2-2c would be that some black holes formed so early in the universe that density inhomogeneities rather than gravitational collapse formed them, which would allow them to have very small masses around $10^{-8} kg$.

What if there is a dark sector (DS) with a possible zoo of dark leptons and quarks, but there is no direct coupling of dark matter to the Standard Model forces? In that case as in figure 2-2d, dark matter could easily evade detection in particle-physics experiments pushing the energy frontier, even if the masses of the dark sector particles are within the kinematic reach of the experiment.

2.2.2 Dark Sector Portals

Portals are an appealing solution to the problem of observing DM in dark sector scenarios with a lack of coupling between the sectors: A new force can exist that is of similar structure as its SM counterpart and allows dark matter particles to interact with each other. This Self Interacting Dark Matter (SIDM) scenario has been used to explain the highly debated core-cusp problem [19]. Quantum mechanics could cause this new force to kinematically mix with its luminous counterpart, thus facilitating interaction between luminous and dark sector particles. This process is called a portal, which makes this model testable. For further information, consult Ref [17]. While the gauge boson of the dark force is not necessarily a viable dark matter candidate itself, confirming the existence of it would be a central step towards understanding dark matter.

2.3 Dark Photons

The photon portal is the most compelling option for thermal models of light DM. The other renormalizable options are the Higgs and neutrino portals [17]. When the dark photon is lighter than twice the mass of the lightest dark sector particle, there is no alternative, other than for it to decay to ordinary matter. These are called visible decays and is what I will concentrate on in this thesis.

If there exist fermions that are charged under both Standard Model and dark sector forces, they can form a loop and effectively couple the SM and DS mediators, even if these fermions are as heavy as the Plank scale. Alternatively, the SM and DS mediators can be coupled by two loops with fermions up to the GUT scale.

The dark photon couples to SM particles via the electromagnetic current and the kinematics of its production and decay are exactly equivalent to those for an off-shell photon of the same mass. This coupling is suppressed relative to that of the ordinary photon by a factor of ϵ, one of the two free parameters in the minimal dark photon model. The other parameter is the mass of the dark photon. The lifetime is related to these two quantities by the following equation:

$$\tau \propto \frac{1}{m\epsilon^2} \tag{2.1}$$

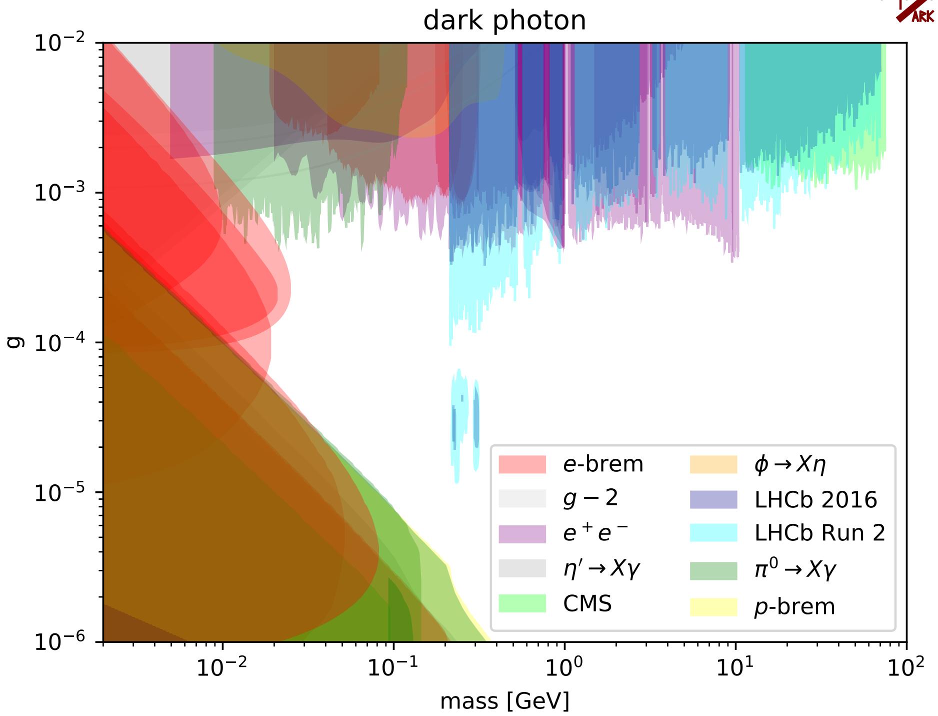


Figure 2-3: There are many still feasible values of the minimal dark photon model parameters as shown by unshaded regions in the exclusion plot. The LHCb full run 2 search described in chapter 5 and shown in cyan significantly improves on the earlier LHCb search with 2016 data only shown in blue (Figure adapted from Ref [20]).

The limits for the parameter space of the visible minimal dark photon model are presented in figure 2-3. The limits in the low mass, low ϵ lower left hand side of figure 2-3 are set by long decay length displaced vertex searches. The limits at high ϵ are set by hunting for a bump in the mass spectrum of final states of the dark photon. The BaBar results for example look for a leptonic final state in e^+e^- collisions.

Naively, the theoretically most interesting region is a box that is bounded by $10^{-5} < \epsilon < 10^{-3}$ due to the one- and two-loop regime and by $10^{-2}\,\text{GeV} < \epsilon < 0.4\,\text{GeV}$ by astrophysical considerations. However, any of the parameter space shown in figure 2-3 becomes interesting when assumptions are adjusted.

A collider of high energy is needed to cover the region in the top right, and with excellent vertex and invariant mass resolution some of the space in the centre of this plot can be covered. Furthermore, unique particle identification and trigger systems are needed. The LHCb experiment can meet all of these requirements.

The blue regions in figure 2-3 denote the limits from the 2016 LHCb search. This thesis is concerned with the cyan constraints obtained by searching for a dark photon decaying to two muons with LHCb data taken during the full run 2 (2015-2018). The luminosity in run 3 will improve this search, and LHCb will move towards a trigger-less readout making it highly efficient at selecting dark photon decays.

2.4 Dimuon Resonances

In the collision of two protons like at LHCb, there are various decays of particles that lead to two oppositely charged muons being produced at a single point in space and time. By measuring the energies and momenta of the decay products, in this case the muons, the mass of the decaying particle can be reconstructed. The number of occurrences (events) of these measured masses peaks at the rest mass of the decaying particle, forming a resonance in the so-called mass spectrum. This is a powerful tool to find evidence for the existence of particles and, if the resonance is sufficiently wide, its observed width can be used to estimate the particle's lifetime.

These resonances can be formed by the decays of known mesons such as J/ψ, η and ρ, but could also occur as a result of dark photon decays. Alternatively, there are other models of physics beyond the Standard Model that can cause these resonances in dimuon spectra. Amongst them are theories with any massive gauge boson with vector couplings to the Standard Model fermions, like a leptophobic B Boson, that couples to leptons via B–γ kinetic mixing and directly to Baryon number.

Other examples include models where a complex scalar singlet is added to the two-Higgs doublet potential and that often feature a light pseudoscalar boson that can decay into the dimuon final state (e.g. Ref. [21]) . References [22, 23], for example, considered the scenario where the pseudoscalar boson acquires all of its couplings to SM fermions through its mixing with the Higgs doublets; the corresponding X–H mixing angle is denoted as θ_H. CMS observed an potential excess in this final state [24]. Hidden-valley (HV) scenarios that exhibit confinement produce a high multiplicity of light hidden hadrons from showering processes [25]. These hidden hadrons typically have low p_{T} and decay displaced from the proton-proton collision resulting in observable dimuon resonances.

The experimental setup used in this thesis to discover novel dimuon resonances or constrain these theories will be described in the following chapter.

Chapter 3

The LHCb Experiment

In order to measure parameters of the Standard Model and probe new theories, we turn to experiments. As particle collisions are inherently probabilistic due to quantum mechanics, observing a large sample of their collision products acts like sampling from the generative model of nature that we want to investigate.

The LHCb experiment is a bowling-lane-long particle detector that reconstructs high energy hadron collisions at one of the four collision points at the Large Hadron Collider (LHC). Designed to excel at the study of heavy flavor physics, it is comprised of various detector components specialized to detect hadrons containing the heavy bottom (b) and charm (c) quarks. This specialization is also serendipitously advantageous to some non-flavor physics analyses and turn LHCb into a general purpose detector in the forward region. One of these analyses is the study of low mass dimuon resonances such as for the search of dark photons.

This chapter will first describe the machine to accelerate the hadrons in section 3.1 and motivate the design decisions of the LHCb experiment by considering the particularities of performing heavy flavor physics at a hadron collider in section 3.2. LHCb has showcased a successful physics program with over 500 publications to date due to its functioning sub-detector components (chapter 3.3) and data reconstruction and selection (chapter 3.4). LHCb is currently being upgraded alongside the LHC (see chapter 3.5) to provide increased luminosity and potential for novel discoveries ahead.

3.1 Large Hadron Collider

To probe new theories of physics, novel data on particle behaviour is needed. This data can be categorized by four major aspects: The particle types studied, the kinematics of these particles, the amount of data collected, and how much this data is polluted by other processes.

The Large Hadron Collider (LHC) at CERN, Geneva, is the largest particle accelerator in the world with the focus of providing the highest energies in the center-of-mass of the collision [26]. Let us see what design decisions this necessitates. Due to the process of boosting in special relativity, this aim necessitates that two beams of particles are directed at each other in a collider, rather than shooting a single beam at a fixed target.

Linear accelerators suffer from the fact that each particle passes each part of the accelerator exactly once, thereby requiring prohibitively long and expensive accelerators for a given energy. In circular accelerators particles can be accelerated by the same components repeatedly. The change of the direction of the beam induces energy losses due to synchrotron radiation, which are comparatively low for hadrons, but prohibitive for electrons at higher energy. Hence, circular hadron colliders are preferred to reach the highest energies.

Another accelerator design driver is how many data points (events) of rare processes can be generated in a certain time. Given the cross-section σ which measures how commonly this process occurs, instantaneous luminosity $\mathcal{L}$ is defined as the number of such rare events that occur per time per cross-section. If each beam arrives in bunches of N particles, the instantaneous luminosity is driven by accelerator parameters as below:

$$\mathcal{L} = \frac{N^2 f N_b}{4\pi\sigma_{beam}^2}, \tag{3.1}$$

where f is the revolution frequency, N_b the number of bunches in one beam, and σ_{beam} is the effective radius of the beam [27], which is assumed to be the same for the two beams. Antiproton production can become a bottleneck , which is why a proton proton collider can reach higher luminosity and was preferred to a proton antiproton

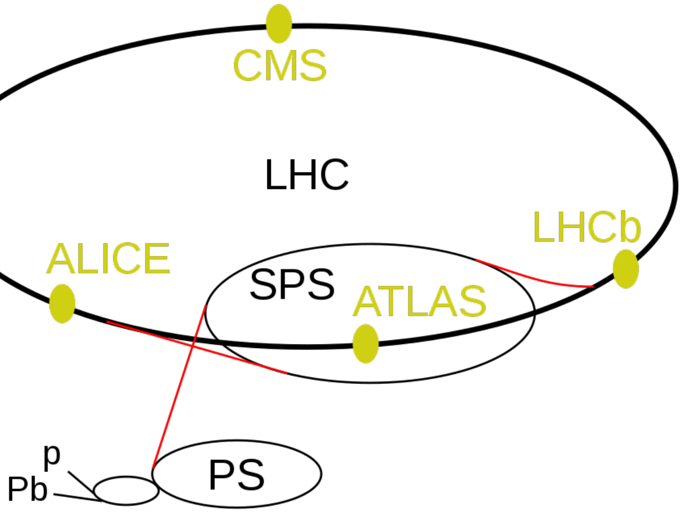


Figure 3-1: To achieve the design energy at the LHC, protons are preaccelerated by the Booster, the proton synchrotron, here labeled as PS, and the SPS as shown in this schematic (Credit: Arpad Horvath Wikimedia).

collider.

The LHC is a circular collider with fixed radius, called a synchrotron, which makes use of all of these insights. For the data collection run considered in this thesis, it collided two proton beams with up to 6.5 TeV of energy, each, to reach a center of mass energy of 13 TeV [28]. The LHC is filled with protons until the fill is dumped due to safety concerns or too many protons have been used up by the experiments (typically after 12 hours). Each such fill is separated into hour-long runs that contain many bunch crossings, also called events. It cost close to 5 billion USD to build and reaches a peak luminosity of $2.1 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ with 25 ns bunches [28, 29]. The LHC is housed in the tunnel of 27 km circumference that was previously used by the Large Electron Positron Collider (LEP) [26]. This ring lies at a depth ranging from 50 to 175 meters underground to reduce background pollution and simplify the civil engineering project in Geneva, Switzerland [30]. In order to get the protons up to speed to be injected into the LHC ring, a pre-acceleration system as shown in figure 3-1 is used [31]. After protons are extracted from hydrogen bottles they gain progressively more energy by passing through a linear accelerator, the Booster, the proton synchrotron, and super proton synchrotron (SPS).

Within the LHC, the beams are accelerated by 16 radio frequency (RF) cavities, steered by 1232 dipole magnets and focused by 392 quadrupole magnets [28]. Both types of magnets are constructed from copper-clad niobium-titanium and kept superconducting at a temperature of 1.9 K (-271°C), giving the LHC the title of largest cryogenic facility at liquid helium temperature in the world [26]. The strength of the magnets limits the energy of the beams for a fixed tunnel size.

The beams travel in parallel beam lines evacuated to 10^8 to 10^9 mbar and intersect at four collision points surrounded by a detector each. ATLAS and CMS are the two "general purpose" experiments that have a full symmetric coverage of the interaction region and cover a wide range of physics analyses. The ALICE experiment is a specialized detector that operates primarily when the LHC collides heavy ions instead of protons, which happens roughly one month out of a year. The LHCb experiment is located at LHC's point 8, which was previously occupied by the DELPHI experiment, and is further described in the next section. The LHCb detector performance decreases more rapidly with higher number of collisions per bunch crossing, commonly called "pile-up", than the increased data rate justifies. Hence, the luminosity is kept at its largest acceptable level (1.1 collisions every bunch crossing [29]) through luminosity leveling [32]: As, the number of particles per bunch, N of equation 3.1, decreases due to collision losses, the effective radius of the beam σ_{beam} is tuned by offsetting the beams around the LHCb collision point to control the overlap. Another advantage of luminosity leveling is lower radiation damage to detector components.

While first run of data taking at the LHC occurred between 2010 and 2013, the data that is relevant for this thesis was collected in Run 2 between 2015 and 2018 at a center-of-mass energy of 13 TeV and delivered luminosity to LHCb of roughly 7 fb^{-1} [29]. With Run 3 planned to start in 2021, at a center-of-mass energy of 14 TeV and five times increased instantaneous luminosity, many changes have been made to upgrade the LHCb detector for a new era of discovery.

3.2 LHCb Design Motivation

The LHCb experiment is motivated by the prospect of doing heavy flavor physics at the LHC by studying hadrons containing the bottom (b) or charm (c) quark. There is an abundance of matter over antimatter in the known universe, which is referred to as baryon asymmetry and for which a deviation from the CP symmetry is required. There are known mechanisms in the Standard Model weak interactions that violate the CP symmetry. These are, however, not powerful enough to account for the full extent of the discrepancy and CP violation mechanisms beyond the Standard Model are required, which can be studies with flavor physics [33].

Physics beyond the Standard Model often features new particles of high masses outside of the kinematic reach of colliders. However, these particles can temporarily come into existence in the decay of particles produced at the LHC. This behavior is referred to as a loop in a Feynman diagram, a pictorial representation of particle interactions [34]. Due to its appearance in a loop, new physics can alter the decay rates to certain final states and their relative frequency (branching fraction). Rare decays of B hadrons that are suppressed in the SM are often more affected fractionally than common decays, and are used to hunt for indirect evidence of new physics.

B factories like BaBar, Belle and Belle II explore these questions by observing asymmetric electron-positron collisions with a centre-of-mass energy of the $\Upsilon(4S)$ meson [35–37], which decays into a neural (charged) B meson and its antiparticle with a branching fraction of 49% (51%) [38]. While many discoveries like the observation of large CP violation in neural B mesons were made, the B factories cannot produce B mesons accompanied by heavy flavor quarks like $\mathrm{B}_\mathrm{c}^{\pm}$ or b baryons like $\mathit{\Lambda}_\mathrm{b}^{0}$. Furthermore, while the luminosity at B factories is significantly higher than at the LHC, the production cross sections are lower, resulting in lower production rates. Doing flavor physics at a hadron collider settles these issues, but complicates the analysis. The b quarks and antiquarks are produced in association with many other particles and have to be selected efficiently.

Heavy quark pairs are primarily produced in the forward direction boosted along the beamline [39, 40]. For the given budget, it is more advantageous to adopt a sophisticated detector system in one forward direction (a single arm), and to drop detector coverage in the central area and the other arm. LHCb can specialize in being a forward spectrometer with only a single arm.

Rather than being contained in a large solenoid magnet to measure particle momenta, a large dipole magnet that only affects a small region of the detector volume can be used. Particle tracks will traverse the various detector components in straight lines with only a deflection in the magnetic field region simplifying charged track reconstruction.

Detector elements can simply be pushed into and out of the beamline for assembly and disassembly, simplifying maintenance. Detector readout electronics can be permanently located outside of the acceptance range of the detector, improving the so-called material budget by reducing the uninstrumented matter than can inadvertently deflect particles.

Heavy flavor hadrons in the ground state cannot decay electromagnetically or strongly and therefore have to decay via the weak interaction resulting in long lifetimes. Due to the typically high energy at production, these hadrons are Lorentz boosted and traverse macroscopic distances in the detector from the collision point (primary vertex) before they decay. These "displaced vertex" signatures are crucial to a successful flavor program and demand high-granularity tracking close to the interaction region [33]. Vertex and momentum resolution drive the invariant mass resolution of a set of particles forming a vertex. This is crucial to reduce combinatorial background. Due to the variety of possible final states of B meson decays ($B^0 \to \mathrm{\pi}^+\mathrm{\pi}^-$, $\mathrm{B} \to \mathrm{DK}^{(*)}$, $B_s^0 \to \mathrm{D}_\mathrm{s}^\pm\mathrm{K}^\mp$), good particle identification systems are needed [33]. As some of the typical decay products are produced with little excess energy, particles with lower momenta have to be reconstructable compared to CMS and ATLAS, necessitating a flexible triggering scheme.

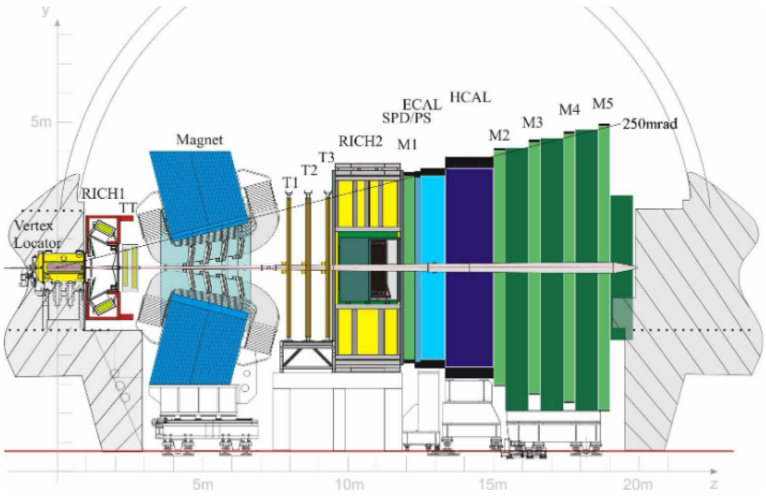


Figure 3-2: Schematic of the LHCb detector components. The coordinate system is right handed with the z axis along the beam and the y axis pointing upwards. Charged particles are deflected along the x axis by the magnet (Credit: Ref. [33]).

3.3 LHCb Detector

LHCb is a single-arm forward spectrometer that follows the insights of section 3.2 and is described fully in Refs. [33, 41, 42]. It covers the pseudorapidity range $2 < \eta < 5$, which corresponds to a range of roughly 1° to 15° from them beamline. Its layout is shown in figure 3-2.

The trajectories of charged particles are captured by the tracking system consisting of the Vertex Locator (VELO) and Tracker Turicensis (TT) upstream of the magnet and the forward tracking stations (T1, T2, T3) downstream of it. Energy deposits in the electromagnetic calorimeter (ECAL) and hadronic calorimeter (HCAL) are used alongside the Ring Imaging Cherenkov Detectors (RICH1 and RICH2) and five muon stations to identify particle type. The subdetectors are assembled in halves that can be retracted from the beamline.

One main components of LHCb is a warm dipole magnet consisting of saddle-shaped coils in a window-frame yoke with sloping poles with an integrated field of 4 Tm (also called bending power). The geometry was chosen so the field between the TT and the forward tracker could be maximized, while producing a field of no more than 2 mT in the region of RICH1.

3.3.1 Tracking Hardware

The principle of tracking detectors is as follows. A charged particle passes through a detector, leaving some kind of remnant behind (e.g. an electron-hole pair in a silicon detector), which is turned into an electrical signal and read out. Multiple channel hits in close proximity are grouped together with a cluster finding algorithm. These clusters, in turn, are associated with a matching algorithm, and a Kalman filter is run to extract track parameters.

Due to the magnet geometry, the track segments in each subdetector are straight lines, bent only in the region of magnetic field between the TT on one side and the T stations on the other. As particles of larger momentum-to-charge ratios are deflected less by a magnetic field, the momentum of charged particles can be measured from tracks. At LHCb, the momentum of charged particles can be measured with a relative uncertainty that varies from 0.5% at low momentum to 1.0% at 200 GeV [33].

This section first describes the VELO. It then covers the silicon tracker that makes up the TT and the inner section of the T stations that are called the Inner Tracker (IT). The Outer Tracker (OT) consists of straw tubes and completes the T stations.

VELO

The VErtex LOcator (VELO) is the first detector that particles produced in the collision can interact with. It provides high location precision to measure the position of primary and secondary vertices. These are used to determine decay lifetimes and the minimum distance of a track to a primary vertex, the impact parameter, which the VELO measures with a resolution of $(15 + [29\,\mathrm{GeV}]/p_{\mathrm{T}})\,\mu\mathrm{m}$ [43].

The radiation levels per year of running are equivalent to the bombardment with 1 MeV neutrons with a flux of $1.3 \cdot 10^{14}\ \mathrm{neq\,cm^{-2}}$. Due to requirements for precision and radiation hardness, silicon microstrip detectors made up of n-implants in n-bulk technology with p-spray strip isolation and a smallest strip pitch of 40 μm were chosen [43]. Modules are constructed from one layer measuring the radial coordinate (r) and another measuring the azimuthal coordinate (ϕ) (see figure 3-3). To provide

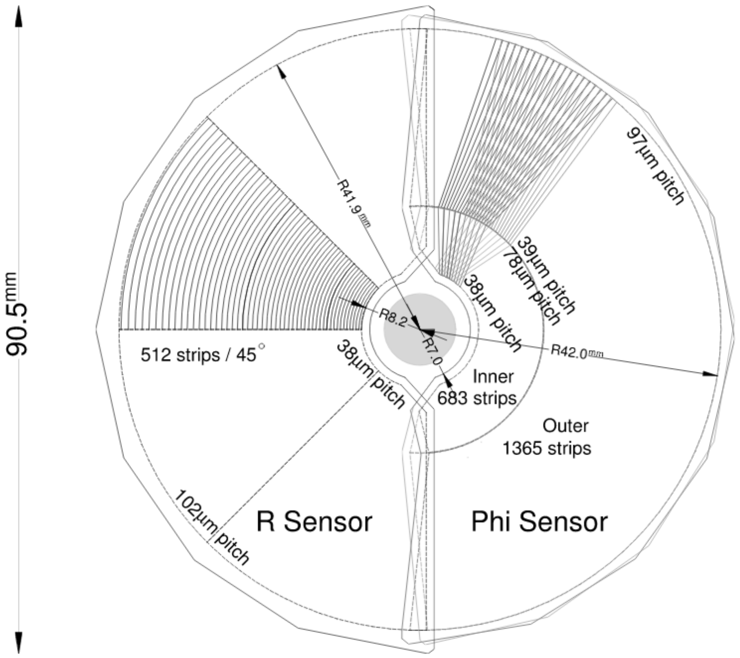


Figure 3-3: The geometry of VELO modules provides the necessary spacial resolution close to the beamline; only a portion of the strips are shown for clarity (Credit: Ref. [33]).

geometric coverage of decays anywhere in a 21.2 cm window along the beamline, 32 of these modules cover the central region. Ten more of the modules are placed further downstream of the interaction region to help extrapolate tracks to the RICH1 detector. Pile-up can be vetoed using two further VELO planes placed upstream of the interaction region.

In order to measure even small displacements form the primary vertex, the VELO is closer to the beam than the beampipe which leads to a novel operating procedure. When the beams start circulating, the VELO is retracted from the beamline. As the beams become stable, the VELO is pushed into its nominal operating position. For operation reasons, the beamline vacuum is separated from the VELO vacuum by a radiofrequency (RF) foil. Particles can scatter off of this foil and produce artificial secondary vertices that can act as background for some physics analyses such as low-mass displaced dimuon searches. This background can be filtered out by a dedicated veto algorithm described in chapter 5. The VELO has to be realigned every fill by a procedure whose success relies on the precision in the construction of the components

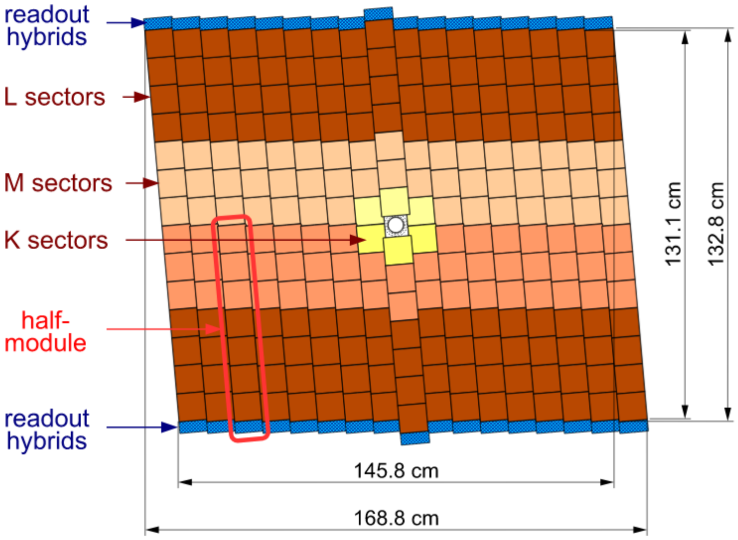


Figure 3-4: The third Tracker Turicensis layer is rotated by 5° to the vertical (Credit: Ref. [33]).

(mechanical tolerances), control of the movement of the components with precision of 10 µm, and software alignment, which processes the residual hits on reconstructed tracks with a non-iterative matrix inversion method [43].

Silicon Tracker

As the particle moves through the LHCb detector, it can interact with the silicon trackers. They are single sided p^+-on-n silicon microstrip detectors with a strip pitch of 200 µm, chosen to provide a single-hit resolution of 50 µm. This allows the momentum resolution to be dominated by multiple scattering for nearly all possible particle momenta. Four detection layers of strip sensors are arranged in a x-u-v-x configuration to form a station. In this configuration the first and last layer are arranged vertically, and the second and third layers are rotated around the beamline by -5° and 5°, respectively. As there are more particles passing through the middle of the detector than the edges, smaller readout strip lengths were chosen for the center, which keeps hit occupancies at a few percent. With this setup, a detection efficiency above 99.8% was achieved for signal-to-noise ratios above 10:1 [33].

The Tracker Turicensis (TT) or trigger tracker is the first silicon tracker. It is

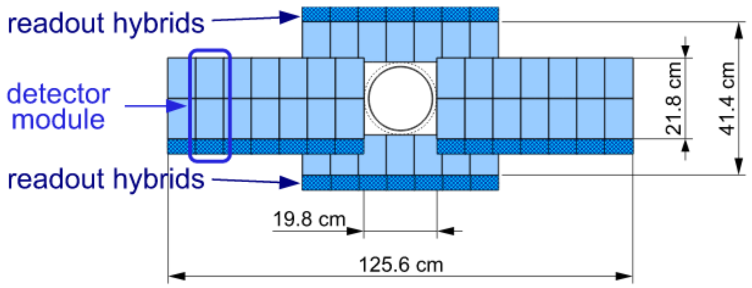


Figure 3-5: The Inner Tracker makes up the inner layer of the tracking stations (Credit: Ref. [33]).

150 cm wide and 130 cm high and covers the full acceptance of LHCb. The K sensors are the closest to the beam followed by the M and L sensors as shown in figure 3-4. At larger distance from the beam, more sensors are read out together to save money and material budget.

The Inner Tracker (IT) is the second silicon tracker and a 120 cm wide and 40 cm high cross shaped region in the middle of the three forward tracking stations (T1 - T3) downstream from the magnet. Each tracking station consists of four detector boxes as shown in 3-5. To ensure complete acceptance and aid the alignment process, adjacent modules are staggered by 4 mm in the beam directions and overlap by 3 mm horizontally.

Outer Tracker

The Outer Tracker surrounds the Inner Tracker and completes the T stations. It is a kapton and aluminum straw-tube drift-time detector with a total active area of $5971 \times 4850\,\mathrm{mm}^2$ [44]. Charged particles create ions in the gas which drift to the central wire of each tube to be read out. A gas mixture of 70% Argon and 30% CO_2 is chosen for its fast drift time under 50 ns and 200 µm coordinate resolution. The detection efficiency is larger than 99% in the centre of the straw.

Two monolayers of 64 drift-tubes each with 4.9 mm inner diameters are staggered to form a gas-tight module as shown in figure 3-6. Each of the T stations has 4 modules of the Outer Tracker arranged in the same x-u-v-x geometry as for the Inner Tracker.

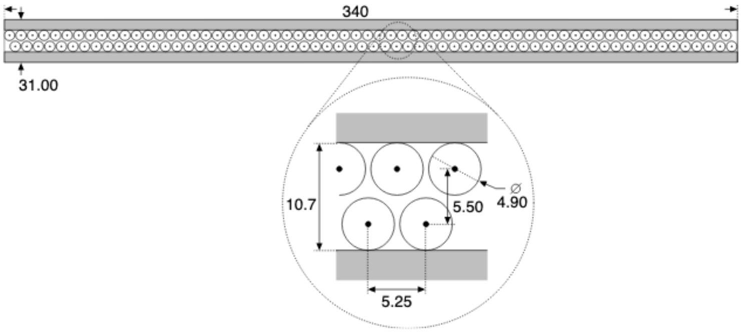


Figure 3-6: Each Outer Tracker module contains two monolayers (Credit: Ref. [33]).

3.3.2 Particle Identification Hardware

Particle Identification (PID) is the task of predicting the type of a particle, due to some of its characteristics like mass and EM and strong charge. LHCb uses the RICH system to measure a particle's velocity, the activity in its EM and hadronic calorimeters to identify electrons, photons and hadrons and its muon system interspersed with iron absorbers to select the highly penetrating muon particles.

Ring Imaging Cherenkov Detectors

When charged particles pass through a material faster than the speed of light in that medium, photons around it accumulate and form a Cherenkov light front like the sonic boom of a supersonic aircraft. The refractive index, n, and fraction of the speed of light the particle is travelling at, β, determine the emission angle of the light cone, θ, as follows:

$$\theta = \arccos\left(\frac{1}{n\beta}\right). \tag{3.2}$$

The LHCb Ring Imaging Cherenkov Detectors, RICH, contain a radiator gas mixture of known refractive index and reflect the Cherenkov photons with wavelengths between 200-600 nm on spherical and flat mirrors so they can be read out by Hybrid Photon Detectors (HPDs) (see figure 3-7). The measured speed of the particle together with its momentum information obtained from tracking specifies its rest mass and, therefore, particle type.

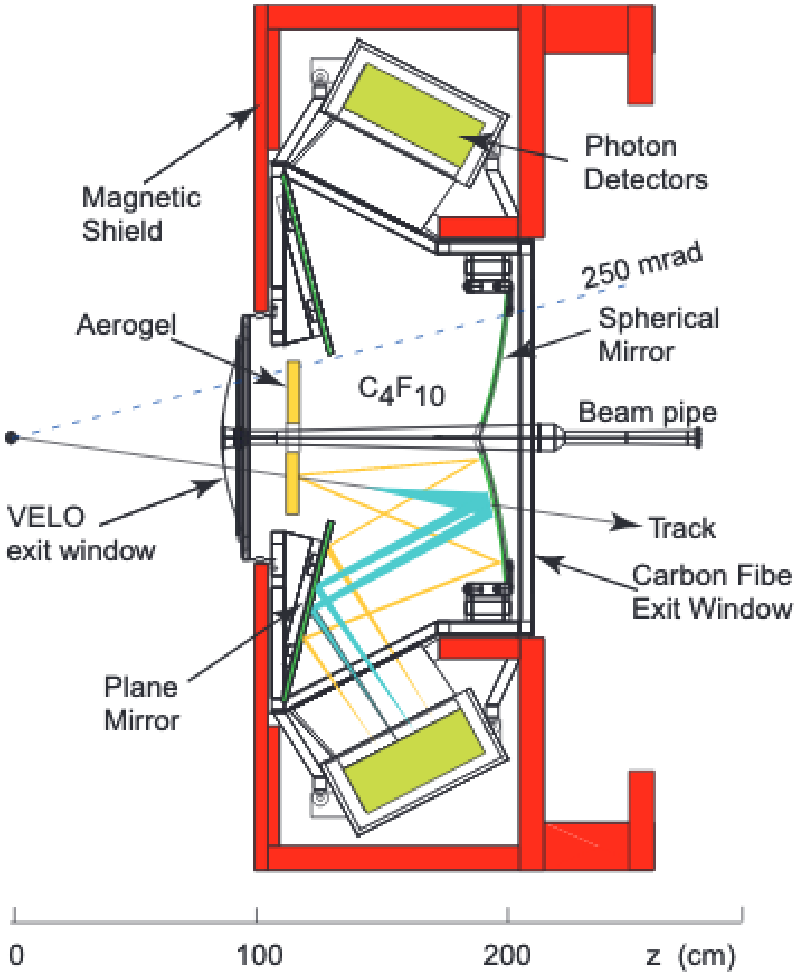


Figure 3-7: Cherenkov light of charged particles is reflected on mirrors to reach the photon detectors in RICH1 (Credit: Ref. [33]).

The photons are detected by hybrid photon detectors (HPD). Cherenkov photons convert in the photocathode of the detector and release a photoelectron that is accelerated by a high voltage of around 10 kV. This photoelectron hits a reverse-biased silicon pixel detector with 1024 pixels of $500\,\mu\text{m} \times 500\,\mu\text{m}$ size producing electron hole pairs that are read out. The HPD's dark count rate, the output in absence of incident photons, is kept at a minimum by limiting its thermionic electron emission at the photocathode and other measures. The HPDs are disturbed by magnetic fields and are, therefore, surrounded by external iron shields to operate in magnetic fields with a maximum of 50 mT.

LHCb needs to provide PID capabilities over a large range of momenta. As particle momenta tend to be higher at small polar angles, the RICH system is split into two detectors, RICH1 and RICH2. RICH1 is upstream of the magnet, is aligned horizonally and has full angular acceptance. It covers the low momentum particle range of $1 - 60$ GeV using fluorobutane C_4F_{10} gas radiators. RICH2 is downstream of the magnet, is aligned vertically and has limited angular acceptance of ± 15 mrad to ± 120 mrad horizontally and from 0 to ± 100 mrad vertically. It covers the high momentum range of 15 to ≥ 100 GeV using CF_4 gas. The performance of the radiators is shown in figure 3-8 and the refractive indices of the gases at 0°C and and 10^5 Pa are parametrized as [33]:

$$
\begin{aligned}
C_4F_{10} \qquad (n-1) \times 10^6 &= \frac{0.25324}{73.7^{-2} - \lambda^{-2}} \\
CF_4 \qquad (n-1) \times 10^6 &= \frac{0.12489}{61.8^{-2} - \lambda^{-2}}
\end{aligned}
\tag{3.3}
$$

where the photon wavelength λ is measured in nm. At a typical wavelength of 400 nm, this results in refractive indices n of 1.0014 and 1.0005 for C_4F_{10} and CF_4, respectively. With an effective radiator length of 95 cm and 180 cm for C_4F_{10} and CF_4, the photoelectron yield is 30 and 22, respectively, for charged particles near the speed of light. With all of these considerations considered in the design of RICH1 and RICH2, excellent $K - \pi$ separation is achieved in the momentum range from 2 to 100 GeV at LHCb.

Calorimeters

Calorimeters destructively measure a particle's energy. When a particle moves through it, a shower of particles is created, that produces scintillation light that is read out by wavelength-shifting fibres and Photo-Multipliers (PMTs). LHCb features an Electromagnetic Calorimeter (ECAL) for particles that primarily interact via electromagnetism and a Hadronic Calorimeter (HCAL) for particles that primarily interact via the strong force. Together with the SPD/PS calorimeter layers, these detectors provide hadron, electron and photon candidates of the highest transverse energy that can

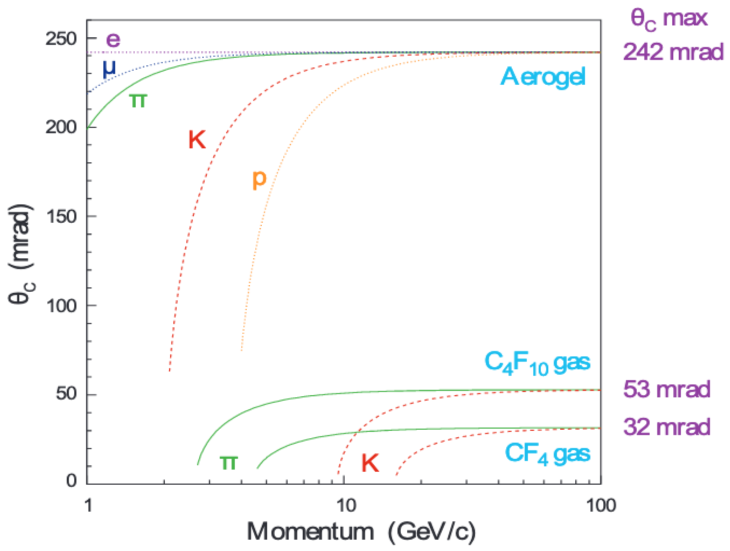


Figure 3-8: Cherenkov angle versus particle momentum for different particle types for the radiator materials using in the LHCb RICH system (Credit: Ref. [33]).

already be used by the hardware level trigger (L0). As the hit density varies by two orders of magnitude over the calorimeter coverage, all calorimeters adopt a variable segmentation transverse to the beam [45].

SPD/ PS To reject a background of neutral π^0, the scintillator pad detector (SPD), a rectangular scintillator pad of high granularity, is placed 12 meters from the interaction region. To reject a background of charged pions, longitudinal information of the electromagnetic shower is achieved by adding a slim preshower detector (PS), that is almost identical to the SPD in design, located between the SPD and the ECAL [33]. These two detectors are separated by a 15 mm lead converter that is 2.5 radiation lengths thick.

ECAL The LHCb ECAL is located 12.5 m from the interaction point and uses a shashlik calorimeter technology: Each module cell uses 66 alternating layers of 2 mm thick lead and 4 mm thick polystyrene scintillator tiles and is read out by plastic wavelength shifting (WLS) fibres [46]. The layers add up to a thickness of 42 cm which corresponds to 25 radiation lengths, which is necessary to fully contain

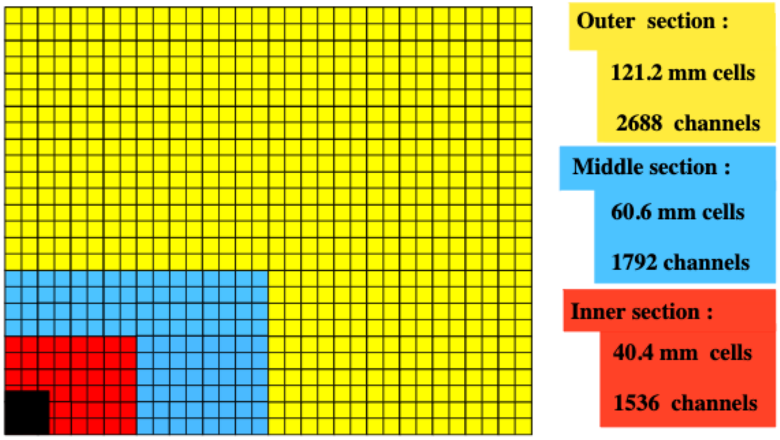


Figure 3-9: Lateral segmentation of the top right corner of the SPD/PS and ECAL provides more location information closer to the beamline. The black space is taken up by the beamline (Credit: Ref. [33]).

the EM shower for optimal energy resolution. The cell widths become progressively smaller the closer they are to the beamline to add location precision and deal with higher occupancies as shown in figure 3-9. This ECAL technology provides modest energy resolution of $\sigma_E/E = 10\%/\sqrt{E/\,\mathrm{GeV}} \bigoplus 1\%$, fast time response and acceptable radiation resistance.

HCAL The hadronic calorimeter (HCAL) follows a sampling calorimeter strategy with scintillating tiles interspersed with 1 cm thick iron slabs as absorbers. Unlike in many other calorimeters, the scintillating tiles run parallel to the beam and have a square shape of size 131.3 mm in the inner section and 262.6 mm in the outer section as shown in figure 3-10 [33]. Trigger requirements do not dictate a stringent hadronic shower containment and the thickness is set to 5.6 nuclear interaction lengths due to space limitations of the detector hall [47]. The scintillation light is transported through WLS fibres to photomultiplier tubes (PMTs) to be read out. The HCAL has an energy resolution of $\sigma_E/E = (69 \pm 5)\%/\sqrt{E/\,\mathrm{GeV}} \bigoplus (9 \pm 2)\%$ [33].

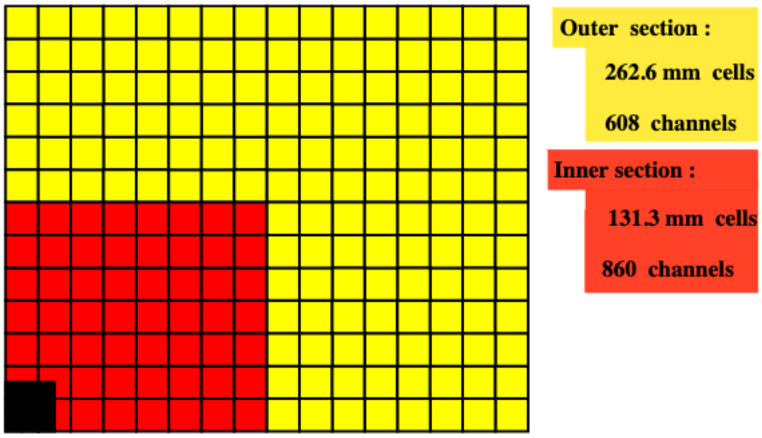


Figure 3-10: Lateral segmentation of the top right corner of the HCAL provides more location information closer to the beamline. The black space is taken up by the beamline (Credit: Ref. [33]).

Muon stations

High energy muons commonly do not lose enough energy in the ECAL to decay and do not interact hadronically in the HCAL. Because few other particle types are as highly penetrating, particles can be identified as muons by how far they traverse the so-called muon stations. This information is produced so fast that it is available in the hardware trigger. The muon stations are comprised of five rectangular shaped tracking stations (M1-M5) placed along the beam axis and contains 1380 chambers covering an area of $435m(A')$. While M1 resides upstream of the ECAL to improve the transverse momentum measurement in the trigger, M2 to M5 are downstream of the HCAL and are interleaved with 80 cm thick iron absorbers. M1 to M3 possess high spatial resolution along the x coordinate to estimate the transverse momentum to 20% accuracy and determine the track direction. M4 and M5 have limited spatial resolution and are used to measure penetration power. A minimum momentum of 6 GeV is needed for a muon to cross all five stations [48–50].

Multi-wire proportional chambers (MWPC) were chosen for all tracking except the most radiation-prone regions of M1. To achieve efficiency better than 95% in a 20 ns window at a gas gain of 10^5, a $\mathrm{Ar/CO_2/CF_4}$ $(40 : 55 : 5)$ gas mixture was adopted. Due to high radiation hardness requirements triple-GEM detectors were used for M1's

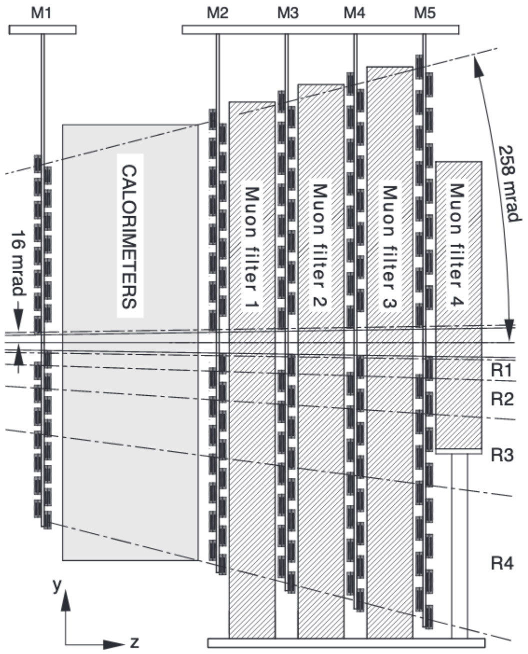


Figure 3-11: The muon chambers become larger with distance to the interaction region to retain constant angular acceptance (Credit: Ref. [33]).

central region. They consist of three gas electron multiplier (GEM) foils sandwiched between anode and cathode planes with an active area of 20 cm × 24 cm.

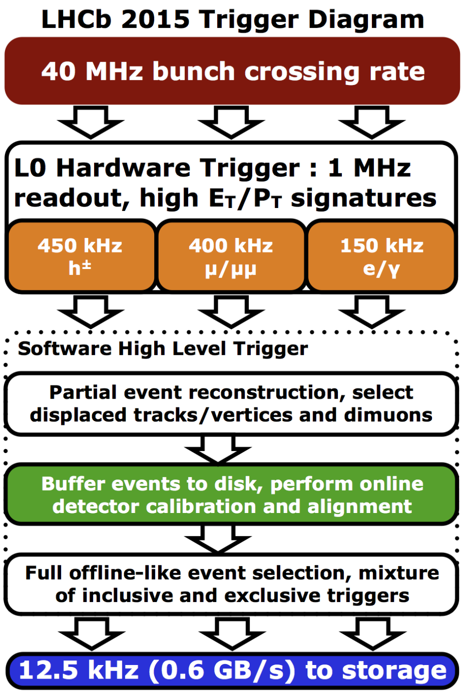


Figure 3-12: The LHCb Run 2 trigger reduces the event rate by more than three orders of magnitude (Credit: LHCb Speakers' bureau [52]).

3.4 Trigger and Reconstruction

Most of the observable physics processes that occur in proton-proton interactions at LHC energies have already been studied extensively. The goal of LHCb is to perform several physics analyses, each exploring an understudied process, in order to measure parameters of interest or to discover a deviation from the expected distribution, symbolising New Physics (NP). This requires sufficient data in this parameter space region also referred to as signal. Due to limited output and storage rates, the trigger application explained in detail in Refs. [32, 51] keeps data relevant for these analysis while discarding the rest which is referred to as background.

Data is recorded each time bunches of protons collide and each such bunch crossing is called an event. The trigger first takes a rough look at each event based on only certain limited information to decide whether to process it further or irreversibly discard it. At LHCb a hardware stage called L0 performs this task tackling a bunch crossing rate of 40 MHz (see figure 3-12). If an event passes this scrutiny, a second trigger armed with further information from reconstruction of other subdetectors selects the promising events. At LHCb, this is done in a software trigger called HLT1. Once more fine grained information is available after detector alignment and calibration, a final HLT2 software stage decides which of events passing HLT1 to save for offline use [53]. HLT1 and HLT2 are executed asynchronously on the Event Filter Farm (EFF), containing approximately 1700 nodes with 27000 physical cores. The output rate is 12.5 kHz, which is a reduction by a factor of 3200 and corresponds to 0.6 GB/s. Due to the identical performance of the online and offline reconstruction, physics analyses like those described in chapter 5 can be performed directly using trigger reconstructed candidates.

Each stage is organized into trigger lines that serve the different analyses. One of these lines firing means that the event is passed on to the next stage. Trigger lines can require specific conditions of trigger lines of earlier trigger stages. For example, if an HLT1 line requires a L0 line to be TIS of a signal decay, then it will only fire if the event was triggered independently of the signal decay. Requiring a line to be TOS of a signal decay selects events that triggered on this decay. Overall, there are around 20 HLT1 and 500 HLT2 trigger lines in total [32].

3.4.1 L0

At the LHC beam crossing rate of 40 MHz only information from the calorimeters, muon stations and VELO pile-up system are available. L0 is a hardware trigger implemented in field-programmable gate arrays (FPGAs) and its task is to reduce this rate to 1 MHz, at which the whole detector can be read out. Some characteristics of interesting events are the presence of muons and particles of high momenta transverse to the beam axis, p_T, and of high transverse energies E_T calculated like $E_\mathrm{T} = \sqrt{m^2 + {p_\mathrm{T}}^2}$, where m is the mass of the particle. L0 reconstructs the highest E_T hadron, electron and photon clusters in the calorimeters and the two highest p_T muon candidates in the muon chambers.

The muon reconstruction begins by connecting the hits in the muon system to form straight line tracks. The track momenta are estimated by assuming that the muon was produced in the interaction region encased by the VELO and deflected by one kick of the magnet. The lower the muon momentum, the more it is deflected. The L0Muon line passes any event with a muon candidate of sufficiently high p_{T}, while the L0DiMuon line cuts on the product of the transverse momenta of the two highest p_{T} muon candidates.

On many of the L0 lines a global event cut rejects events with a high number of primary pp interactions and total number of particles in the event as estimated by the VELO pile-up system and the number of SPD hits. Due to combinatorial issues these events would otherwise occupy a disproportionate fraction of bandwidth and processing power, and efficient signal-background discrimination would not possible.

3.4.2 HLT1

The events that passed the L0 selection are reconstructed further and examined by the HLT1 trigger. Clusters in the VELO are combined to create VELO tracks, which serve to reconstruct primary and secondary vertices (PVs and SVs). Vertices displaced from the beamline are a powerful indicator of interesting signatures such as heavy flavor physics and certain models of dark photons. These tracks are progressively propagated to the TT to form upstream tracks, to the T stations taking into account the curvature due to the magnet, and to the muon system. To speed up the reconstruction, tracks lower than a cutoff transverse momentum of 500 MeV are discarded. Momenta are estimated with a Kalman filter.

No muon identification can be performed for tracks with momenta below 3 GeV, as the penetrating power is too low to reach the muon stations downstream of the calorimeters. Below a momentum of 6 GeV, hits in the first two muon stations are required. In the regions of 6 to 10 GeV (>10 GeV), additional hits are needed in one (two) of the two remaining muon stations.

Events are selected inclusively based on one- or two-track signatures, displaced vertices, or dimuon candidates. In addition there are low multiplicity lines for central exclusive production scenarios and dimuon lines with no impact parameter cut as used in dark photon searches.

3.4.3 HLT2

Events that were selected by HLT1 are stored in a 10 PB buffer, which can set aside up to two weeks of consecutive data taking. This allows for the real time running of alignment for various subdetectors and calibration such as for the RICH gas refractive index and the ECAL, each of which has a dedicated HLT1 line. Buffering allows HLT2 to be processed during inter-fill periods when LHCb is not running, thereby alleviating output rate bottlenecks. HLT1 thresholds can also be tuned according to the size of the remaining disk buffer.

The HLT2 trigger stage runs a full offline level reconstruction on the events retrieved from the buffer. All tracks are reconstructed, this time with only a minimal p_{T} requirement, which is crucial for the study of charmed or strange hadrons. In HLT2 neutral particle candidates are constructed from calorimeter information and full particle identification is run. Fake tracks are rejected with a neural network and clone tracks are removed. This is the full offline level reconstruction that is only revised if bugs are found.

The traditional data-storage strategy (FULL) saves the whole 70 kB raw event to tape in addition to reconstruction. Analysts have to wait for stripping campaigns that are run several times a year and can extract information about any kind of particle in these events. An alternative (TURBO) data-storage strategy only saves the reconstructed objects and selected other information resulting in 5 kB size events that can be accessed by analysts immediately after data taking [53, 54]. Caution has to be exercised as the information needed to perform another reconstruction of the data is discarded. The smaller event size allows for a higher output rate, that benefit some analyses like the prompt-like dark photon analysis from chapter 5.

Inclusive topological and exclusive c-hadron trigger lines make up roughly 40% of the trigger output rate each, with the rest distributed among electroweak, dimuon, exotic and other exclusive trigger lines [32].

3.5 Upgrade

After an upgrade during the second long shutdown starting in 2018, the LHC is scheduled to resume operations in 2022 with Run 3, providing proton-proton collisions at a center of mass energy of 14 TeV [55]. To optimize the physics exploration potential, LHCb will also be upgraded with an overhaul of more than 90% of the active detector channels [56]. The main design driver is the goal to withstand the five times larger luminosity of $2 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ while maintaining or improving the current physics performance. Due to the timing requirements, LHCb will change all front-end electronics of all subdetectors. Further improvements will be made to optimize the physics program like a complete overhaul of the tracking system.

Tracking The upgrade VELO detector will remain similar in concept to Run 2, but replace the silicon microstrips with pixels. Furthermore, the distance to the beamline will be reduced from 8.2 mm in Runs 1 and 2 to 5.1 mm [57]. Together these adjustments improve vertex and impact parameter resolution. The TT will be replaced with the Upstream Tracker (UT) that covers a larger area of acceptance of LHCb. It is made up of four planes of high-granularity silicon micro-strips, each consisting of 16 or 18 vertical staves of 14 square sensors each [58]. Three Scintillating Fiber Tracker (SciFi) stations with four x-u-v-x detection layers each will replace the T stations downstream of the magnet. Each layer consists of 2.4 m long scintillating fibres with a diameter of 250 μm that are read out by silicon photomultipliers [58].

RICH While the structure of the Cherenkov detectors will remain the same, the photon detectors are replaced with commercial multianode photomultipliers (MaPMTs) with external readout electronics to adapt to the fast readout speed [59].

Calorimeters The modules of the ECAL and HCAL will continue to be used throughout Run 3, with a replacement of front and back-end electronics. The SPD and PS sub-detectors' main use cases were in the hardware trigger and they will be removed to reduce the material budget [59].

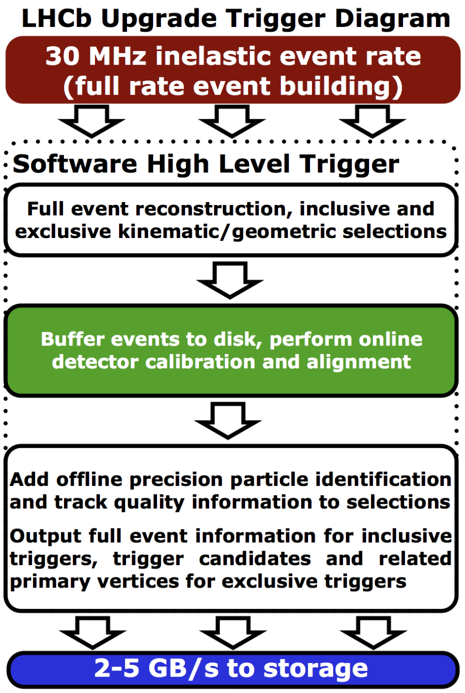


Figure 3-13: LHCb triggering scheme during run 3 (Credit: LHCb Speakers' bureau [52]).

Muons Due to shielding from detectors upstream of the muon stations and additional material introduced in front of M2, a redesign due to the increased luminosity was not necessary. The first muon station (M1) will be removed as its main use case was in the hardware trigger [59].

Trigger Requirements on the upgrade trigger are stringent. For example three quarters of the LHCb physics program has to be moved to the TURBO stream to stay within an acceptable output rate [56]. However, the main difficulty is reducing the trigger rate enough for HLT2 with the limited information available at the first trigger stage. The information available to the hardware trigger is insufficient to reduce the data rate enough without discarding significant fractions of the physics program. Hence, LHCb will implement a triggerless readout and all triggering will be done in software (see figure 3-13). The increase in luminosity leads to an input data rate of 5 TB/s, which is a five-fold increase over the current rate and a 100-fold increase to the rate processed by the software trigger [60, 61]. The first software trigger, HLT1, is implemented in C++ with CUDA extensions by the Allen project for use in GPUs [62]. The parallelism in the triggering process maps onto the parallelism of GPUs and allows for filtering to be performed in the Event Builder infrastructure, thereby reducing network costs. Rather than just meeting the triggering needs, Allen exceeds them and foreshadows the potential of a highly successful Run 3 for LHCb.

Chapter 4

Statistics, Machine Learning and Data Science

The goal of modern particle physics is to explore the generative statistical model of nature in particle interactions. It pursues this objective with a procedure rooted in statistics, machine learning (ML) and data science. This procedure processes the low level detector features and produces confidence intervals for measurementVs of fundamental properties, and discovery p-values or exclusion regions for searches for new particles.

Five topics will be highlighted here. To start with, a novel moment loss function (MoDe) is introduced that can train a classifier with the restriction that it has a certain order of dependence on a protected feature. Next, the first high precision generative model approach to high energy physics simulation is shown. Then, the use of natural language processing in particle physics is explored. In addition, a simple, powerful, and novel deep learning approach to vertexing will be discussed. Finally, a procedure for high dimensional goodness-of-fit and two-sample tests using ML classifiers is shown.

4.1 MoDe

4.1.1 Overview

When searching for new physics that appears as a resonance in a spectrum (often of a mass), sideband methods that assume a smooth background are routinely applied. Artificial structures induced by ML classifiers trained to reject backgrounds can lead to false claims of discovery and are impermissible. A plethora of existing tools allows constraining the classifier to be independent from the resonant feature. Such decorrelation is a sufficient, but not necessary condition to avoid localized structures. In Ref. [63], we proposed a novel moment loss function (Moment Decomposition or MoDe) that allows the bump hunter to constrain the classifier to allow a specified dependence on the resonant feature (like linear or monotonic quadratic), thereby improving classification accuracy [63]. This increases the discovery power without exacerbating the risk of false positives. [1]

4.1.2 Background

In particle collider physics, we often want to test a composite hypothesis: Does a hypothetical particle exist and, if so, with what mass? This is often determined by bump hunting: fitting an otherwise featureless spectrum with a smooth background and a localized signal pdf (e.g. a narrow Gaussian peak) [64]. Particle physics milestones from the discovery of the ρ meson [65] to the discovery of the Higgs boson [7, 8] were achieved in this way and this approach retains its prominence in nuclear and particle physics today [66–74].

Many of these searches rely on estimating the background under a signal candidate by parametrically or non-parametrically [75] interpolating or extrapolating the spectrum from a region where no signal is expected: the sidebands. A central assumption of these methods is the smoothness of the background.

Often the strength of the statistical test is improved by selecting a specific area

---

[1] This work is adapted from the paper in Ref. [63] of which I am a co-author.

of phase space of auxiliary variables (e.g. decay kinematics) to perform the test in, namely the event selection. Often the selection criteria are optimized with a proxy for the power of the test such as the value of $S/\sqrt{S+B}$, where S and B are the number of Signal and Background events, respectively. While simple cut-based selection is still prominent, a modern way of selecting an area of phasespace for the test is with deep machine learning algorithms. Borrowing from advances in machine learning, higher level correlations can be harnessed to deliver superior signal sensitivity [76–80].

W jet tagging, the process of distinguishing jets originating from a Lorentz-boosted W boson from other sources, is a main ingredient in searches for low-mass dark matter and others. Both ATLAS [81] and CMS [82] have used machine learning classifiers for this task.

Consider a setting where a classifier is trained on a sample with the resonant feature / protected attribute (e.g. mass) peaked at a value m_S and a background uniform in that feature. If the resonant feature were fed into the classifier, the algorithm would predict a signal only for values around m_S. Even running the machine learning algorithm on only the background will result in a peak. This "sculpts the background", violating the smoothness assumption discussed above. Some kind of regularization is needed.

Training on signal samples with different mass hypotheses only allows for nonzero signal estimation close to the masses of the signal samples. Another option is to remove the mass feature from training. However, a sufficiently powerful classifier can learn the correlations and recover the mass variable.

Another promising procedure bakes the regularisation into the loss function. Decorrelating the classifier score from the mass (protected) variable and, thereby, producing independence prevents the problem of a sculpted background. A variety of techniques have been developed for decorrelation [83–100] and harnessed in bump hunts [101–119]. In the machine learning literature, similar methods have been used under the names of fairness or domain adaptation [120–123].

Independence of a resonant feature is a sufficient, but not necessary condition to prevent violating the smoothness assumption of the background. A low order polynomial response cannot induce a strong localized response, while still being able to capture correlations between the resonant feature and the classification target. Relaxing the decorrelation restriction can result in superior rejection performance.

The work presented here demonstrates a method capitalizing on these insights: *Moment Decomposition* (MoDe) is a flexible, novel loss function that allows bump hunters to control their desired level of correlation from linear to arbitrarily high polynomials while retaining control on parameters such as the slope of the linear dependence.

In the remainder of this section, existing decorrelation methods are discussed and MoDe defined in sub-section 4.1.3. Performance in a toy and a physics example is shown in sub-section 4.1.4. A discussion of the work follows in sub-section 4.1.5.

4.1.3 Methods

Existing Decorrelation Methods

The relevant statistical problem for decorrelation is binary classification, where the data consists of a feature vector $X \in \mathcal{X}$, the target labels $Y \in \mathcal{Y} := \{0, 1\}$ and the resonant feature $M \in \mathcal{M}$. X can be independent of M, which would mean that no decorrelation is needed, but in general it can have any dependence on M. While M could be any feature on which the bump hunt is performed, it is assumed to be mass for the rest of this work.

The decorrelation task is to devise a mapping $f : \mathcal{X} \to \mathcal{S}$ where $s \in \mathcal{S}$ are classifier scores with the condition that

$$p(f(X) = s|M = m, Y = y) = p(f(X) = s|Y = y) \; \forall \; m \in \mathcal{M} \text{ and } \forall \; s \in \mathcal{S}, \tag{4.1}$$

for certain values of y. While Eq. (4.1) can be required to hold for background, signal or both, the standard is to require independence only for background. A simple cutoff on classifier scores produces predictions $\hat{y} \in \mathcal{Y}$.

Decorrelation methods that are currently in use can have a loss function of the following form:

$$\begin{aligned}\mathcal{L}[f(x)] = \sum_{i\in \mathrm{S}} L_{\text{class}}(f(x_i),1) + \sum_{i\in \mathrm{B}} w(m_i)\, L_{\text{class}}(f(x_i),0) \\ +\lambda \sum_{i\in \mathrm{B}} L_{\text{decor}}(f(x_i),m_i)\,,\end{aligned} \tag{4.2}$$

where $S = \{i|\, y_i = 1\}$ and $B = \{i|\, y_i = 0\}$ correspond to signal and background, respectively, L_{class} is the usual classification loss such as the binary cross entropy $L_{\text{BCE}}(f(x), y) = y\log(f(x)) + (1-y)\log(1-f(x))$, w is a weighting function, λ is a hyperparameter, and L_{decor} generically denotes some form of decorrelation loss.

Traditional classification without decorrelation sets $w(m) = 1$ and $\lambda = 0$. Types of decorrelation methods are:

- Planing [98, 124]: The total importance of the signal and background can be set to unity, by reweighting signal event with $w(m_i) \approx p_S(m)/p_B(m)$. In this case, there is no decorrelation loss function and $\lambda = 0$.

- Flatness [94]: A cumulative decorrelation loss $L_{\text{decor}} = \sum_m b_m \int |F_m(s) - F(s)|^2\, \mathrm{d}s$ is added with $\lambda > 0$, where the squared difference between the empirical CDF in a mass bin and the empirical CDF over the whole spectrum at a classifier score s is summed over all mass bins, with a prefactor of the fraction of events in that bin, b_m. No reweighting of the signal events is performed and $w(m) = 1$.

- Distance Correlation (DisCo) [90, 97]: where $\lambda > 0$. The decorrelation loss function is specified by the *distance correlation* [125–128] between $f(x)$ and m for the background and has to be computed on a batch of data. There is no reweighting of signal events and $w(m) = 1$.

- Adversaries [83, 87, 92, 99]: $w(m) = 1$, $\lambda < 0$, the correlation loss is the loss of a second, adversarial network trained to predict some properties of m from the classifier output $f(x)$.

Moment Decorrelation

In this section the new simple, fast and robust decorrelation procedure that achieves state-of-the-art performance is motivated. However, MoDe's main merit is its extension beyond simple decorrelation.

The insight for this work and the flatness loss [94] is that the uniformity constraint in Eq. (4.1) has an equivalent expression for the CDFs of scores:

$$F(f(X)=s|Y=y)=F(f(X)=s|M=m,Y=y)\ \forall\ m\in\mathcal{M}\text{ and }\forall\ s\in\mathcal{S}. \tag{4.3}$$

We are going to focus on a setting where only the background is made independent of s and the conditional CDF is discretized by binning in mass, which brings with it a more compact notation

$$F(f(X)=s|M=m,Y=y)\rightarrow F_m(s), \tag{4.4}$$

It is assumed that $\mathcal{M}$ lies in $[-1,1]$. This can be accomplished by a linear transformation or a more sophisticated choice.

If the following structure of a loss function is accepted, the constraint Eq. (4.3) is satisfied.

$$L_{\text{decor}}\rightarrow L^0_{\text{MoDe}}\equiv\sum_m\int|F_m(s)-F^0_m(s)|^2\mathrm{d}s. \tag{4.5}$$

F^0_m is the 0^{th} Legendre moment of $F_m(s)$ in a mass bin as defined in

$$F^0_m(s)=c_0(s)P_0(\tilde{m})=\frac{1}{2}\int_{-1}^{+1}P_0(m')F(s|m')\mathrm{d}m'\approx\frac{1}{2}\sum_{m'}\Delta_{m'}F_{m'}(s), \tag{4.6}$$

where c_0, and polynomial, $P_0(x)=1$. Note that the Legedre polynomial can be replaced by any other orthogonal polynomial. The bin m has width Δ_m and central value $\tilde{m}$.

This loss in Eq. (4.5) is minimized for

$$F_m(s) = F_m^0(s) = c_0(s)\ \forall\ m, \tag{4.7}$$

In this case Eq. (4.3) holds. Hence, $f(X)$ is decorrelated and independent of M. In the situation where all bins are of the same width and occupancy, the loss in Eq. (4.5) reduces to the flatness loss of Ref. [94]. With bins of varying width and occupancy these solutions diverge with MoDe outperforming Ref. [94].

If decorrelation is more important for one region of the spectrum, or a score threshold, then nonuniform weighting of the form $\mathrm{d}s \to w(s)\mathrm{d}s$ can be introduced, as long as the function derivative of the loss is still calculabale as shown in Sec. 4.1.3.

Beyond decorrelation: Moment decomposition

We will generalize moment decorrelation by loosening the independence restriction and allowing a controlled measure of correlation. This is accomplished by letting the bump hunter choose the orders of Legendre modes that the classifier score dependence can take. Choosing orders up to 0 restores the decorrelation loss. If orders up to ℓ were chosen, the generalized loss of a classifier $f(X)$ becomes:

$$\mathcal{L}[f] = L_{\text{class}} + \lambda L^{\ell}_{\text{MoDe,}} \tag{4.8}$$

where

$$L^{\ell}_{\text{MoDe}} \equiv \sum_m \int |F_m(s) - F_m^{\ell}(s)|^2 \mathrm{d}s. \tag{4.9}$$

Compared to (4.5), a linear combination of Legendre modes up to ℓ replaces the 0^{th} order Legendre mode F_m^0.

$$F_m^{\ell}(s) = \sum_{l=0}^{\ell} c_l(s) P_l(\tilde{m}), \tag{4.10}$$

with Legendre moments described by

$$c_l(s) = \left[\frac{2l+1}{2}\right] \int_{-1}^{1} P_l(m')F(s|m')\mathrm{d}m' \approx \left[\frac{2l+1}{2}\right] \sum_{m'} \Delta_{m'} P_l(\tilde{m}')F_{m'}(s). \tag{4.11}$$

When the mass dependence is perfectly described by orders up to ℓ, then $F_m(s) = F_m^\ell(s)\ \forall\ m, s$, leading to a optimal MoDe loss in Eq. (4.9). Setting $\ell = 0$ results in an classifier of zero MoDe loss being independent of mass. Increasing ℓ to 1 and 2 leads to linear and quadratic dependence. Furthermore, the maximum magnitude of the linear slope can be defined by making the replacement in Eq. (4.10):

$$c_1(s) \to c_1^{\max} c_0(s) \tanh\left(\frac{c_1(s)}{c_1^{\max} c_0(s)}\right) \tag{4.12}$$

The analyst can control $c_1^{\max}$, which bounds the quotient of the linear slope magnitude and $c_0(s) > 0$. The hyperbolic tangent function is chosen due to its range monotonicity, differentiability, and oddness, though it can be replaced by similar alternatives.

When $\ell = 2$ the derivative of $F_m^\ell(s)$ is nonzero on $(-1, 1)$ if $3|c_2(s)| \leq |c_1(s)|$ is the case. This allows us to specify monotonic mass dependence by making the following replacement in Eq. (4.10).

$$c_2(s) \to \frac{c_1(s)}{3} \tanh\left(\frac{3c_2(s)}{c_1(s)}\right) \tag{4.13}$$

Both these options are hyperparameters in MoDe which can be set independently of each other. Equivalently, higher order relations can be derived.

Computational Details

For an implementation of MoDe, an efficient calculation of its loss function and gradient is needed that can be achieved with the following approximations. For every batch of n samples that the training data is split into, the empirical CDF in bin m

with n_m samples is calculated as

$$F_m(s) \approx \frac{1}{n_m} \sum_{i=1}^{n} \Theta(s - s_i)\delta_{m,m_i}, \tag{4.14}$$

where $s_i \equiv f(x_i)$ is the score of sample i, and Θ is the Heaviside step function defined as $\Theta(x) = 1$ if $x > 0$ and $\Theta(x) = 0$ otherwise.

To propagate gradients, the functional derivative of L^{ℓ}_{MoDe} with respect to the classifier scores is

$$\delta L^{\ell}_{\mathrm{MoDe}} = \delta s_i \sum_m \int 2 \left[F_m(s) - F^{\ell}_m(s)\right] \left[\frac{\partial F_m}{\partial s_i} - \frac{\partial F^{\ell}_m}{\partial s_i}\right] \mathrm{d}s, \tag{4.15}$$

where the approximation from Eq. (4.14) leads to

$$\frac{\partial F_m}{\partial s_i} \approx -\frac{1}{n_{m_i}} \delta(s - s_i)\delta_{m,m_i}. \tag{4.16}$$

The derivatives of the Legendre CDFs are acquired from this equation, Eqs. (4.10) and (4.11) to be

$$\frac{\partial F^0_m}{\partial s_i} \approx \frac{1}{2}\Delta_{m_i} \frac{\partial F_{m_i}}{\partial s_i} = -\frac{\Delta_{m_i}}{2n_{m_i}} \delta(s - s_i), \tag{4.17}$$

$$\frac{\partial F^1_m}{\partial s_i} \approx \frac{\partial F^0_m}{\partial s_i} + \frac{3}{2}\tilde{m} \cdot \tilde{m}_i \Delta_{m_i} \frac{\partial F_{m_i}}{\partial s_i} = -\frac{\Delta_{m_i}}{2n_{m_i}} \delta(s - s_i)\left[1 + 3\tilde{m} \cdot \tilde{m}_i\right], \tag{4.18}$$

$$\vdots$$

Note that as each derivative only affects one mass bin, the sum over mass bins can be dropped. The integrals over s can be discarded due to the factors of $\delta(s - s_i)$, which generate high-precision gradients as no integral approximation is needed.

This means that the gradient for $\ell = 1$ can be summarised as.

$$\delta L^{1}_{\mathrm{MoDe}} \approx -\delta s_i \sum_m \frac{1}{n_{m_i}} \left[F_m(s) - F^{1}_m(s)\right] \left[2\delta_{m,m_i} - \Delta_{m_i}(1 + 3\tilde{m} \cdot \tilde{m}_i)\right]. \tag{4.19}$$

Gradients for other values of ℓ take similar forms.

For weighted samples with per-sample weights w_i, the CDFs in Eq. (4.14) become

$$F_m(s) \approx \frac{1}{w_m} \sum_i^n w_i \Theta(s - s_i) \delta_{m,m_i}, \tag{4.20}$$

where the sum of weights in a bin m is computed as

$$w_m = \sum_i^n w_i \delta_{m,m_i} \tag{4.21}$$

Equation (4.16) changes to

$$\frac{\partial F_m}{\partial s_i} \approx -\frac{w_i}{w_{m_i}} \delta(s - s_i) \delta_{m,m_i}, \tag{4.22}$$

The MoDe method and all computation above is valid for small batch size. However, unsurprisingly due to its global approach, MoDe can demonstrate slow convergence due to a large discrepancy between mini-batch and full-batch statistics. Performance is more optimal for larger batches.

In the evaluation direction of the neural network, MoDe scales linearly with the number of steps in the integral approximation in s, n_s and the batch size n, which is a main contributor to cost. With dynamic binning sorts and reindexing, the time complexity of MoDe is $\mathcal{O}(n_s \times n + n \log n)$. In the current implementation an additional $\mathcal{O}(n \times n_m)$ component is added due to caching $F_m(s_i) - \tilde{F}_m(s_i)$ for the backwards pass. We leave it to further work to reduce this component (e.g. through approximating the CDFs at s_i by nearest neighbor interpolation). For the space complexity MoDe adds a $\mathcal{O}(n \times n_m)$ component over pure data storage due to storing the CDF for each input for each bin when calculating gradients. An implementation in PyTorch [129] and Tensorflow [130] was performed and is available at `https://github.com/okitouni/MoDe`.

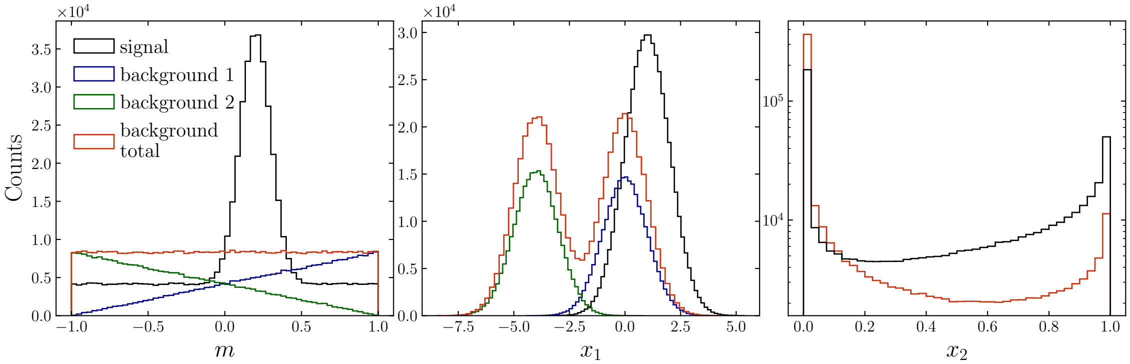


Figure 4-1: Mass, x_1 and x_2 distributions for the simple model problem (Credit: Ref. [63]).

4.1.4 Results

This section contains MoDe's performance on a toy model and the W-jet tagging setting tackled by the decorrelation work in Refs. [89, 90].

Simple Model

First, a binary classification model with two background sources and the following two features is considered:

$$x_1 \sim \begin{cases} \mathcal{N}\big(1,1\big) & \text{when } Y = 1, \\ \mathcal{N}\big(0,1\big) & \text{when } Y = 0 \text{ for background type 1,} \\ \mathcal{N}\big(-4,1\big) & \text{when } Y = 0 \text{ for background type 2,} \end{cases} \tag{4.23}$$

$$x_2 = \exp\left[-\frac{(m-0.2)^2}{2\cdot 0.1^2}\right] \quad \text{when } Y = 0 \text{ or } Y = 1, \tag{4.24}$$

where the normal distribution is written as $\mathcal{N}$. For the signal process, when $Y = 1$, the mass is sampled from $\mathcal{N}(0.2, 0.1)$ and $\mathcal{U}(-1, 1)$ with equal probabilities. The background masses are produced by sampling $U \sim \mathcal{U}(0, 1)$ and setting $m = 1 - 2\sqrt{U}$ and $m = -1+2\sqrt{U}$ for types types 1 and 2, respectively. The distributions generated by this procedure are displayed in Fig. 4-1.

In the absence of any dependence constraints, a classifier can use x_2 to recover the mass feature in this situation if its capacity is high enough. The unconstrained classifier can then discriminate between signal and background using the mass feature,

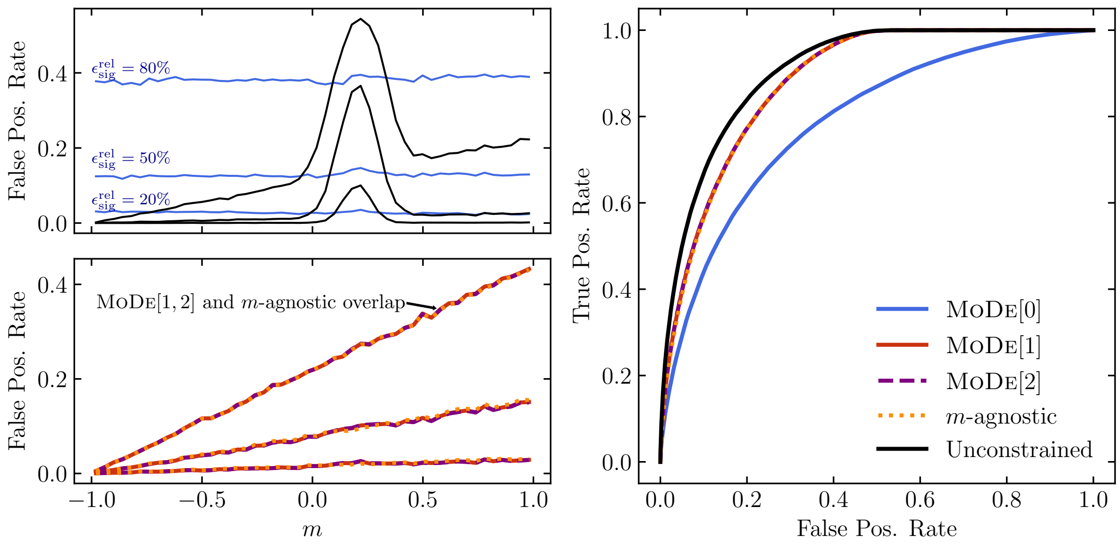


Figure 4-2: **Left:** The false positive rate versus mass. Different classifiers have lines colored as shown in the legend on the right with lines of the same color corresponding to the signal efficiencies $\epsilon_{\mathrm{sig}}^{\mathrm{rel}} = 80, 50, 20\%$. The bottom panel shows that MoDe[1], MoDe[2] and the m-agnostic model resulted in the same classifier dependence on mass. This is expected due to the optimality of linear dependence for an agnostic classifier. **Right:** ROC curves for different classifiers. The unconstrained classifier reached superior ROC scores at the price of background sculpting, while MoDe[1], MoDe[2] and the m-agnostic model improved on the uncorrelated MoDe[0] model, while preventing background sculpting. Finding an equivalent of the m-agnostic model in practice is often impossible (Credit: Ref. [63]).

as shown in practice in Figure 4-2. This classifier violates the smoothness assumption of the background and sculpts the background making it unusable in a real-world scenario. This is exactly what we are tying to avoid. In contrast, the MoDe[0] classifier produces scores that are independent of mass as advertised, making it a feasible real-world solution.

In some situations it is possible to remove all explicit mass information while exploiting all other information to separate signal from background, resulting in an optimal, analytical, m-agnostic solution. In the simple example here, this is possible and amounts to ignoring x_2. The resulting m-agnostic model is linearly correlated with mass due to the distribution of x_1 and the background mass distribution. The constraints on a decorrelating classifier, however, prevent even this helpful linear correlation, resulting in a sub-optimal performance. Without this unnecessary restriction, MoDe[1] can recover the linear dependence and the classification performance of

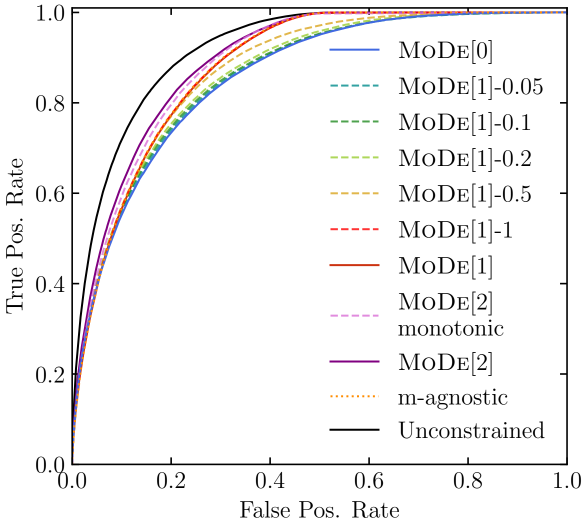


Figure 4-3: ROC curves for different classifiers as in Fig. 4-2 for the modified simple problem. A monotonic version of MoDe[2] and constrained maximum slope versions of MoDe[1] are added (Credit: Ref. [63]).

the m-agnostic solution. Despite MoDe[2]'s freedom to choose quadratic dependence it does not fail to find the optimal linear mass dependence as shown in Fig. 4-2.

The analyst has even more fine-grained control in MoDe like the maximum linear slope and monotonicity of the quadratic dependence as explained in Sec. 4.1.3. To show this potential, consider a situation in which the second feature functional form is changed to

$$x_2 \to \exp(m) + 2m. \tag{4.25}$$

In this situation, MoDe[2] has a higher ROC AUC than MoDe[1] as seen in Fig 4-3. Fig 4-4 demonstrates that MoDe[2] does indeed have a quadratic and nonlinear dependence in this problem. Control over the slope monotonicity and slope maxima is displayed in Figs 4-4 and 4-5, respectively. Any constraint can lead to a loss in performance (see Fig. 4-3), which an analyst has to weigh up against the increase in background simplicity in their particular problem.

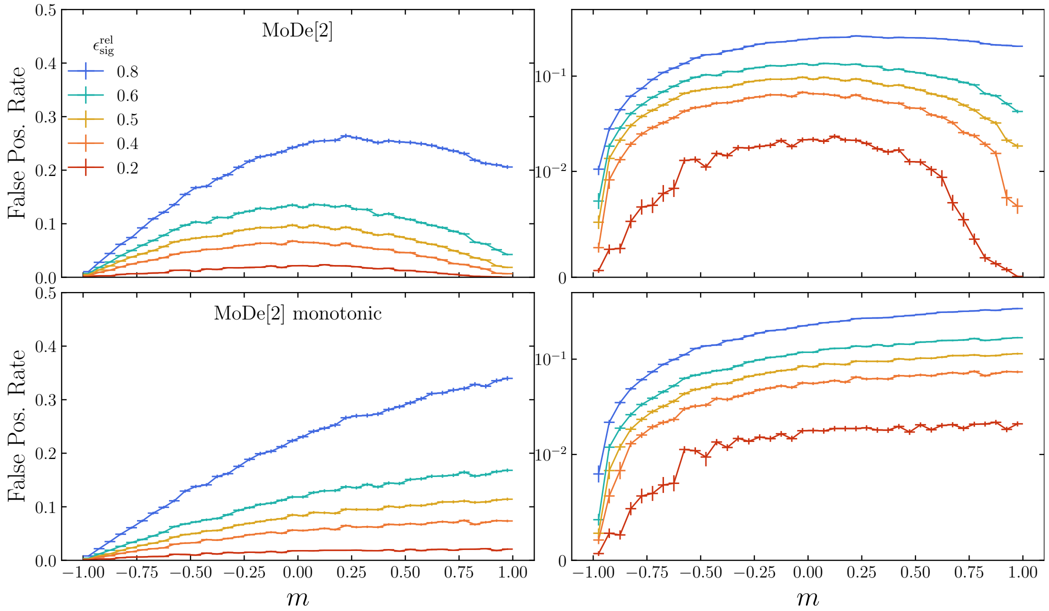


Figure 4-4: Monotonicity demonstration: With the monotonicity option disabled in the top panel, MoDe[2] for the modified simple example in Eq. (4.25) has a nonmonotonic false positive rate as a function of mass. If the option is enabled, the response is monotonically increasing as shown in the bottom panel. The right panels show a version of the left plots on logarithmic axes (Credit: Ref. [63]).

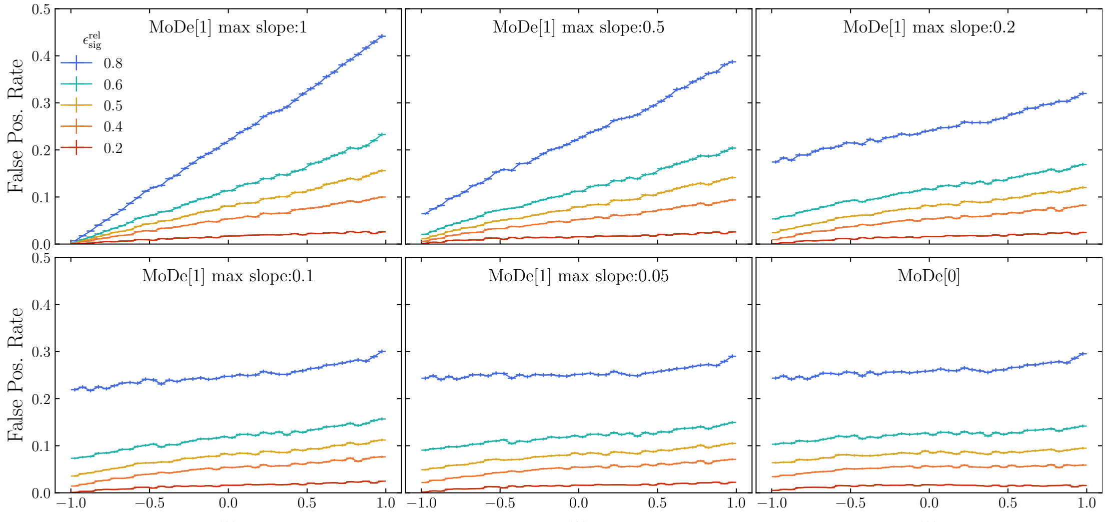


Figure 4-5: Slope cap demonstration: Varying the maximum slope of MoDe[1] for the modified toy example results in corresponding false positive rate to mass relations (Credit: Ref. [63]).

Boosted Hadronic W Tagging

Distinguishing jets originating from highly Lorentz-boosted W bosons from other jets is a common task in searches for physics beyond the Standard Model as discussed in Sec. 4.1.2. Spectra of the mass of the W candidate jet, or a dijet system including the W candidate are usually bump hunted. A variety of informative features exist for this task, deep classifiers are beneficial because of correlations in the data, and violating the smooth background assumption is unacceptable due to biased signal estimators. For these reasons boosted W jet tagging is a benchmark system to study decorrelation models.

The benchmark simulation data is taken from Ref. [90], which itself followed the procedure from Ref. [89]. The data is available from Zenodo at Ref. [131] and further details can be seen in these references.

Classifier Details: MoDe and DisCo Both the DisCo and MoDe classifiers use a 3-layer neural network of the structure discussed in Ref. [90]. For a significant performance improvement over the original studies in Refs. [90] and [89], Swish activation [132] was applied after each of the fully connected layers, with batch normalization applied after the first layer. The output layer is comprised of a single node with a sigmoid activation. For MoDe and DisCo, the ADAM optimizer [133] was used with a 1 cycle learning rate policy [134], a starting learning rate of 10^{-3} and a maximum learning rate (lr) of 10^{-2}. This was achieved using a cosine annealing strategy [135] and the learning rate decayed to 10^{-5} during the last few iterations. Momentum is cycled in the inverse direction from 0.95 to a minimum of 0.85. A learning rate range test was the basis of hyperparameter testing. Large batches of 10–20% of the training data were chosen for training. This does not necessarily complicate training due to the 1 cycle learning policy as discussed in Ref. [136].

Classifier Details: Adversarial Decorrelation An adversarial setting as in Refs. [89] and [90] was used to train two networks: the same classifier as used by MoDe and DisCo and a Gaussian Mixture Network (GMN) [137] that outputs means, covariances and mixing coefficients of 20 normal distributions. The mixture network used a single 64 node hidden layer with ReLu activation connected to 60 output nodes to model the posterior probability density function $p_{\theta_{\text{adv}}}(m|f(X) = s)$, where θ_{adv} are the GMN parameters. The loss function of the adversary is given by the data likelihood:

$$L_{\text{adv}} = \mathbb{E}_{s\sim f(X)}\mathbb{E}_{m\sim M|s}\left[-\log p_{\theta_{\text{adv}}}(m|s)\right] . \tag{4.26}$$

The classifier was trained without the adversary for 20 epochs with a learning rate of $lr = 10^{-4}$, followed by training of the adversary for 20 epochs with $lr = 5 \cdot 10^{-3}$. Lastly, both networks were trained jointly by optimizing the following loss function

$$\arg\min_{\theta_{\text{class}}} \max_{\theta_{\text{adv}}} \left[L_{\text{class}(\theta_{\text{class}})} - \lambda L_{\text{adv}}(\theta_{\text{class}}, \theta_{\text{adv}})\right] , \tag{4.27}$$

where θ_{class} and θ_{adv} are the parameters of the classifier and the adversary, respectively. Controlling the λ parameter between 1 and 100 allows choosing an operating point for the classification-decorrelation trade-off. As common in adversarial training, the non-convexity of the loss complicates convergence, and leads to a reliance on carefully selected hyperparameters.

Decorrelation This subsection shows that MoDe[0] outperforms existing techniques in the decorrelation task. We adopt the classification and decorrelation metrics of Ref. [90]: R50, the background rejection power (inverse false positive rate) at 50% signal efficiency; and 1/JSD, where the Jensen-Shannon divergence (JSD) is a symmetrized version of the Kullback-Leibler divergence. JSD measures the relative entropy between background mass distributions that pass and fail the classifier selections, respectively.

As expected and shown in Fig. 4-6, an unconstrained classifier selects background with a mass close to the W boson, sculpting a fake peak in the background. For a sufficiently large λ, MoDe[0] yields a selection uncorrelated to mass as advertised and shown in Fig. 4-6. A more fine-grained differentiation between different decorrelation methods is provided by the metrics from Ref. [90] as shown in Fig. 4-7. While the

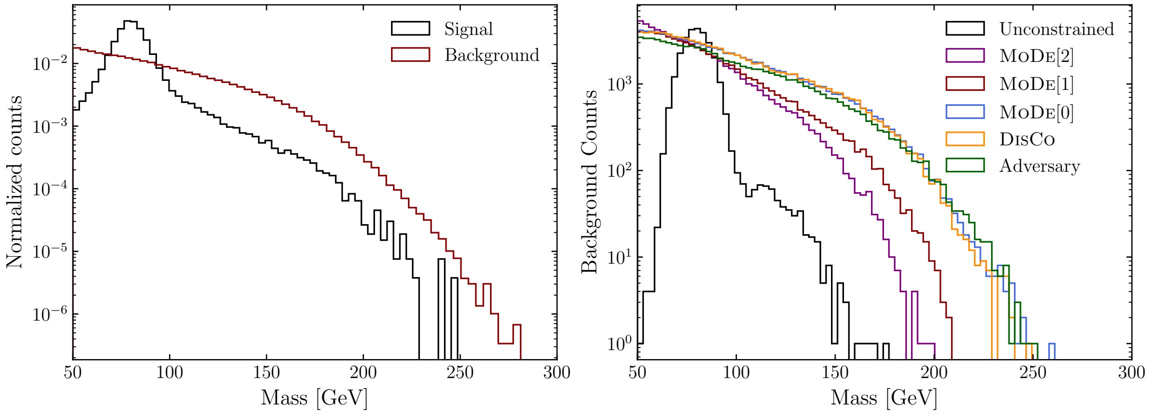


Figure 4-6: **Left:** Signal and background distributions without selection. **Right:** Background distributions at 50% signal efficiency for different classifiers. Only the unconstrained classifier sculpts a fake peak (Credit: Ref. [63]).

adversarial methods achieve slightly more rejection power for every value of JSD, they demonstrate unstable behavior and are notoriously hard to tune and train due to the saddle point loss landscape. In contrast, DisCo and MoDe[0] introduce only a single additional hyperparameter and posses a convex loss function leading to robust and stable training that achieves nearly identical rejection performance. Given that MoDe[0] is more resource efficient as shown in Fig. 4-8 and has a slightly better performance, it can become the practitioner's decorrelation method of choice.

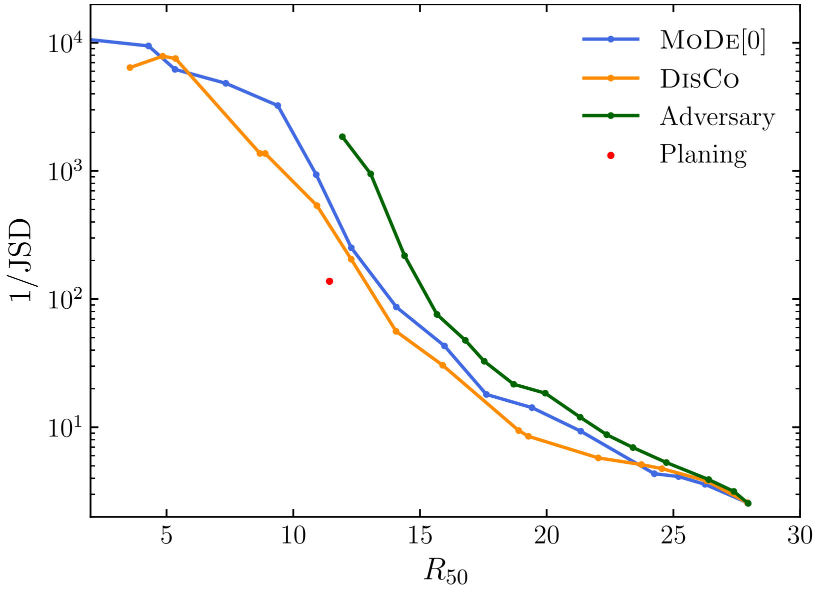


Figure 4-7: Decorrelation vs background-rejection power using the metrics of Ref. [90], where points in the top right are optimal. All methods achieve similar results (Credit: Ref. [63]).

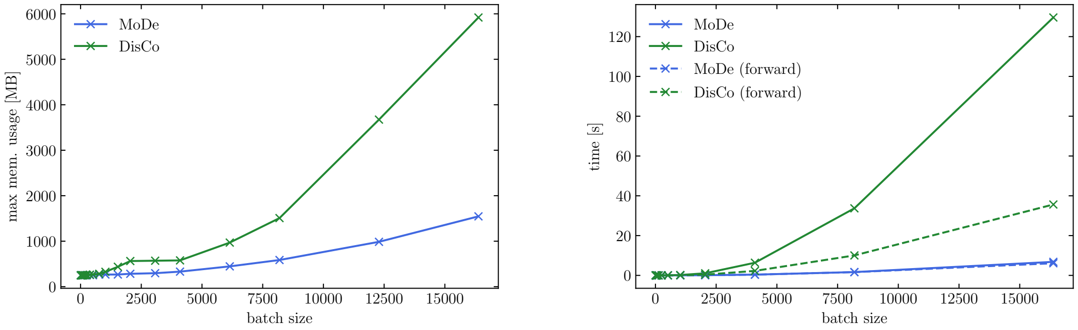


Figure 4-8: Maximum memory usage and CPU time versus batch size for DisCo and MoDe (Credit: Ref. [63]).

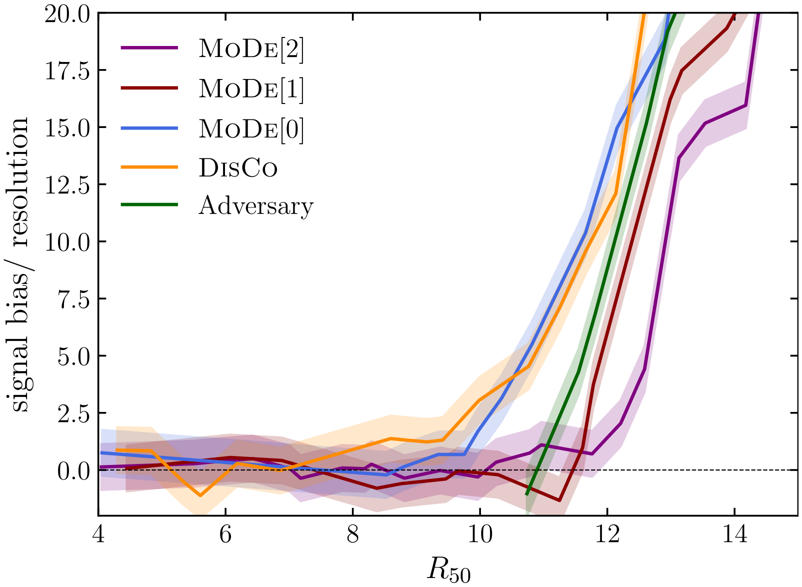


Figure 4-9: Signal bias normalized by its uncertainty versus background-rejection power. Widening the range of permissible classifiers, MoDe[1] and MoDe[2] achieve larger background-rejection power for unbiased estimators (Credit: Ref. [63]).

Beyond Decorrelation While MoDe[0] was shown to be a state-of-the art decorrelation method, the main focus of this work is on loosening the restriction of decorrelation, where the 1/JSD metric becomes irrelevant. Indeed, MoDe[1] and MoDe[2] do not sculpt a peak as shown in Fig.4-6 and become important in this section.

We replace this metric by the signal bias induced by the classifier selection as measured by the a polynomial fit to the background-only samples. This is exactly what matters to analysts in the bump hunting scenario.

For Figure 4-9, the signal bias is divided by their uncertainties, so that of values of roughly unity are consistent with no bias, and values substantially larger than unity demonstrate substantial bias, which could result in a false claim of discovery. This measure is often well approximated by the square root of the number of background events.

Figure 4-9 demonstrates that for small background rejection power all methods are consistent with no bias. Specifically DisCo and MoDe[0] retain unbiased results up to R50 $\lesssim 9$, which occurs at 1/JSD $\gtrsim 1000$ (see Fig. 4-7).

While larger values of 1/JSD are possible, this does not translate to real advantages to the bump-hunter. Loosening the restriction of decorrelation allows MoDe[1] and MoDe[2] to find unbiased signal estimators at larger rejection power, leading to tangible gains in sensitivity in a real-world analysis.

4.1.5 Conclusions and Outlook

In summary, a key problem in bump hunting is to find a selection method that provides both unbiased signal estimators and the largest possible background rejection power for a given signal efficiency. While unconstrained classifiers achieve the latter and fail at the former, the plethora of existing decorrelation methods prevent inducing localized structures, they do so at a relatively low classification power.

This work proposes a novel moment loss function (Moment Decomposition or MoDe) as a flexible method to achieve the best of both worlds. It does so by abandoning the restriction of decorrelation and instead allows the analyst to introduce controllable polynomial dependence on the resonant feature (often a mass). Analysts can choose the order of the dependence (linear, quadratic,...), as well as the maximum slope of the linear dependence, and the monotonicity of the higher-order dependence. This flexible approach opens the door for physics analysis with a higher sensitivity to new physics without compromising the fidelity of the background estimation.

4.2 Generative Models for Simulation and Compression

With the proposed increase in data rates, the CPU requirements of simulation samples far exceed the pledged resources of LHCb. While effort has been spent to incrementally improve existing methods that simulate the full microphysics of detector interactions and then reconstruct the resulting signals, a fundamental change in approach is needed to make up for the resource gap. Furthermore, data rates are limited. Compressing particle information before events are read out leads to smaller event sizes, which in turn allows for more events to be collected.

We developed novel, purpose-built neural architectures to address both issues. These modified autoencoder and Camér generative adversarial network architectures generate realistic and high precision samples of particle identification (PID) features. This approach was validated to 2% precision in the context of a high impact modern physics analysis.[2]

4.2.1 Generative Models

In statistics, generative models estimate the joint distribution of target and observation features. Alternatively, discriminative approaches model the conditional distribution of the target given the observation. While discriminative models often deliver superior results in classification and regression tasks because of their specialization, they cannot be used to generate samples. This is the realm for generative models, of which variational autoencoders (VAEs) and generative adversarial networks (GANs) are prominent examples.

Autoencoder An autoencoder is a neural network trained in an unsupervised manner with only examples, but no labels [139]. It consists of an encoder and a decoder. The encoder takes a list of input features and returns an intermediate output of a hidden (latent) dimension. The decoder takes this intermediate output and returns a

---

[2] This section was adapted from Ref. [138] of which I am a co-author.

list of features. Training of the whole structure has the aim of increasing the similarity between the output and the input.

Typically, the latent dimension is lower than the input dimension. If the similarity loss between the input and output is optimal, the decoder can be used to reconstruct the input from the latent dimension, e.g., a perfect autoencoder turns a cat image into the same cat image. Hence, the latent data is a lossless lower-dimensional representation of the input, which can be used in other tasks. Achieving dimensionality reduction with the use of a bottleneck latent dimension is an example of representation learning. If all inputs are independent of each other, any lower dimensional representation loses information needed to reconstruct the input. In practice, there is some trade-off between the loss of the latent representation and the benefit gained from the compression which has to be evaluated on a case-by-case basis.

Variational autoencoders (VAE) are probabilistic graphical models whose posterior is approximated by a neural network [140]. They retain the encoder-decoder structure of traditional autoencoders, but the distributions in the latent dimension are forced to be standard Gaussians. If samples of standard Gaussians are decoded by the network, they approximate samples from the original joint distribution, turning the VAE into a generative model.

Generative Adversarial Networks A GAN is a training regime, where networks compete with each other in a zero-sum two-player minimax game. A generator learns to produce output data similar to the training data. A discriminator / critic decreases its loss by successfully differentiating between the training data and the data produced by the generator [141]. In practice, neural networks are often used for both the generator and discriminator functions. The exact training schedule for the models has to be chosen carefully to prevent vanishing gradients: If the discriminator becomes too good too quickly, the generator will never produce samples that defeat the discriminator and training effectively stops being useful. Even with a good training schedule, traditional GANs exhibit the problem of mode collapse: Sometimes the same outputs are produced repetitively, which leads to coverage issues and an insufficient reproduction of the joint probability.

The Wasserstein GAN targets vanishing gradients and mode collapse by using a Wasserstein-1-distance-based loss function. The activation function of the output of the discriminative network is changed from a sigmoid to a linear activation function and the RMSProp optimizer is used to aid convergence [142]. Together, these improvements yield increased training stability and more varied samples. The sample gradients, however, are biased. The Camér GAN [143] used in this work produces unbiased gradients by using the Camér metric.

4.2.2 Data Compression

The data collected by experiments like LHCb is rich in a variety of features describing reconstructed particles. Some features require high precision and any bias in them can lead to knock-on effects (e.g. mass and momentum of the particle). Others are only approximate. Miscalculating them does not bias an analysis, but only makes it sub-optimal. One such example are particle identification (PID) features. Low level information is summarised by high-level PID features like the muon vs pion classification score.

One particular analysis might, for example, look for specific physics processes that produce kaons. However, there are other very similar physical interactions, producing pions, that are background and could potentially be misreconstructed as being signal. The analysis might choose a pion vs kaon classifier to remove this background. A kaon vs "everything else" classifier might be sub-optimal by unnecessarily rejecting kaons with ghost-track-like signatures, even though ghosts are not a relevant background in the analysis. Hence, a list of highly correlated, but yet separately important PID features is retained in practice. These should be recoverable in any acceptable solution.

At LHCb 40% of features are PID features. If these could be compressed by half, a 20% bandwidth reduction is possible, potentially enabling the implementation of entirely new analyses. This proposal will investigate the feasibility of compressing PID features.

4.2.3 Simulation

A necessary step in making a physics discovery is to optimize the parameter space in which to perform the search for new physics. For example, in the hypothesis test for the existence of a new particle, the null hypothesis is represented by the Standard Model prediction (background), while the alternative hypothesis is taken to be the background probability density function (PDF) plus a PDF for a new particle (signal). Combinatorial backgrounds can also be estimated using simulation. To increase test power, the selection can be optimized to contain as much signal and as little background as possible.

Both the signal and the background for selection optimization are often estimated by sophisticated and precise simulations with the procedure shown in figure 4-11: Firstly, the particles originating from the collision are generated and their interactions with the particle tracking stations modelled [144]. Secondly, the microscopic interactions of the particle with specialized PID detector elements are simulated, a series of pattern-recognition algorithms are run and high level PID features inferred. These two processes are independent and factorise up to a small number of special cases to be dealt with separately.

With the upgrade of detectors like LHCb, the number of simulated data samples required will increase significantly, outpacing the available computational resources. If the required amount of simulated data is not reached, systematic uncertainties in analyses will further dominate other types uncertainties. While several attempts have been made to speed up computation [145], none have made significant progress to achieve the goal to reduce this systematic uncertainty enough for it not to be the limiting factor of analyses. A fundamentally different strategy of simulation is required. If the second stage of the simulations process were made negligibly fast, the CPU cost could be reduced by an order of magnitude and 400 TeraFLOPs would be saved.

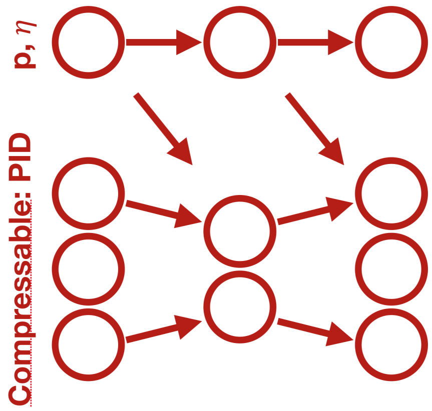


Figure 4-10: The proposed autoencoder architecture allows the encoder and decoder to be informed by non-PID features like the momentum (p) and pseudorapidity (η).

4.2.4 Proposal

This section proposes using generative models on PID features for data compression and simulation at LHCb and beyond. For compression in an autoencoder setup, PID features are compressed by an encoder for each particle separately and decompressed by a decoder (see Fig. 4-10). In this framework, the intermediate layer can be stored, reducing file size and throughput. A custom architecture was created so that non-PID features are used as a pass-through, informing the compression and decompression of PID features at any layer. One alternative idea is to compress the PID features and perform analyses from them directly. This would break backward-compatibility and complicate revisiting old analyses with the proven and tested original features. Roughly 1000 collaboration members have relied on these and produced over 500 papers to date with them.

Simulation samples can be generated by sampling standard Gaussian distributed numbers and processing them with a VAE decoder. These numbers can be stored and combined with conventionally generated non-PID features. This plan would skip the microphysics and reconstruction steps in the traditional PID simulation procedure as shown in figure 4-11, thereby saving on the computation time of all PID features, which constitute 40% of all LHCb features. The performance of the simulation can be evaluated by comparing the outputs of the conventional and the proposed simulations.

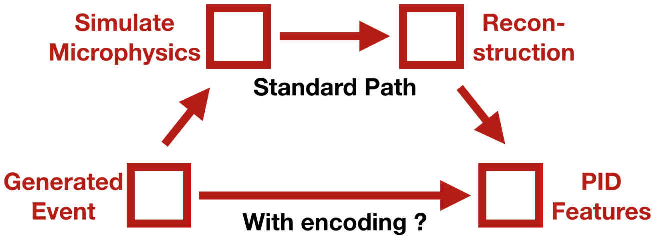


Figure 4-11: Generative models can help avoid the computationally expensive process of simulating particle interactions with PID subdetectors. Interactions of particles with other detector systems remains unchanged.

For the simulation studies, we replace the VAEs with Camér GANs. This gives up on the task of data compression, but results in better generative performance.

Trying to tune a traditional simulation to properly model the real-world detector is time-consuming and in some situations unsolved. The retraining of the generative model is expected to be significantly easier and might produce more accurate simulations.

One potential issue with the generative approach proposed here is the assumption of factorization. As every particle's PID information is handled separately, correlations between them could lead to invalid samples. However, it has been shown that the number of particles that are affected by these concerns is small and can be dealt with separately.

4.2.5 Compression Studies

To evaluate the potential of autoencoders to compress LHCb PID information, features of equal numbers (≈ 550k) of different charged particle types (e, μ, π, K, ghost) were collected from LHCb minimum bias Monte Carlo simulations. Ghost tracks are fake tracks that can occur when reconstruction algorithms mistakenly match unrelated combinations of hits. This data was published at `zenodo.org/record/1230552#.WuIbVlMvw6h` to enable collaboration with computer scientists without a background in physics. This study concentrates on 2.5 million examples where general tracking, ECAL and RICH information was available. The data contained 18 features

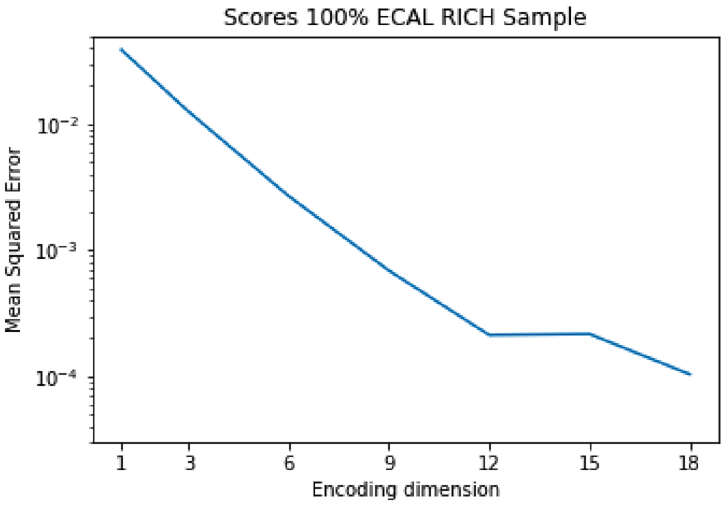
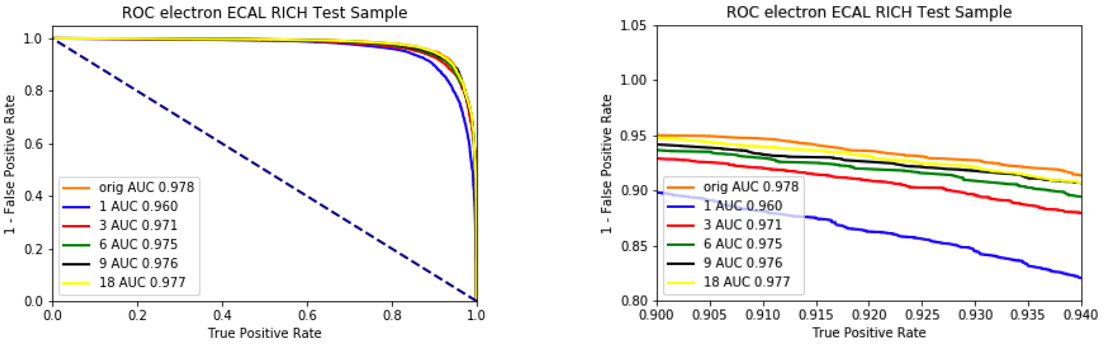


Figure 4-12: The autoencoder MSE loss is monotonically decreasing for increasing encoding dimension.

to compress and 8 additional auxiliary features that are present at both compression and decompression phases.

An autoencoder with 5 hidden fully connected layers was trained with auxiliary features concatenated after each layer. The training, validation, and test sets were chosen to correspond to 70%, 20%, and 10% of the total data set. The mean squared error (MSE) loss between the input of the encoder and output of the decoder was minimized. Figure 4-12 shows the trade-off between low reconstruction error and encoding dimension.

However, it is difficult to know what value of the loss is acceptable in practice. A relevant metric for practitioners is the PID classification power. For this reason, a cross-check is performed in which boosted decision tree (BDT) classifiers are trained on the uncompressed and decompressed features and the receiver operating curve (ROC) for electron selection compared. The dotted line in figure 4-13a represents random selection, with an area under the curve (AUC) of 0.5. Ideal separation would be achieved by a value of zero for false positive rates for all true positive rates, resulting in an AUC of 1. The PID classifiers lie between these two extremes. The AUC is reduced by only 0.07 when the 18 features are compressed to 3 in figure 4-13, which is promising. Yet, it is unclear whether this is acceptable for real analyses.



(a) Receiver Operating Curve (ROC)

(b) Zoomed in version of figure 4-13a

Figure 4-13: Compressing 18 PID features with an autoencoder (AE) to 3 dimensions reduces the electron selection AUC from 0.978 to 0.971.

Importantly, the compression achieved with an autoencoder is independent of the traditional lossless LZ4 compression common in particle physics. LZ4 only uses a dictionary-matching stage, and unlike other common compression algorithms does not combine it with an entropy coding stage. Hence, both forms of compression don't compete, as demonstrated by the linear relation between file size and encoding dimension in figure 4-14.

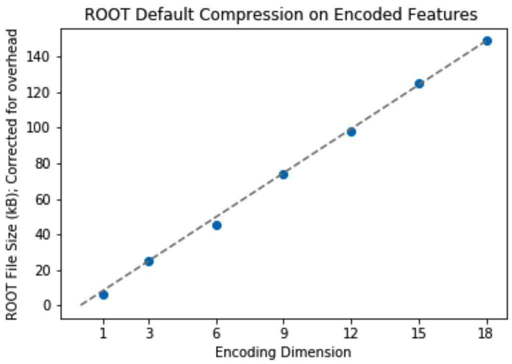


Figure 4-14: Encoding the PID data with an autoencoder does not reduce the fractional compression with the LZ4 algorithm common in particle physics.

4.2.6 Simulation Generation Studies

The goal of improving the generation of simulation is to maximise the discovery potential for a variety of analyses without introducing bias. The GAN is trained to reproduce traditionally simulated data from one set of physics interactions ($B \to K^\star \mu^+ \mu^-$ and $B \to K^\star e^+ e^-$) and uses the squared difference between the traditional (GEANT) and GAN-generated simulated PID features as the figure of merit. To test the robustness of GAN-generated simulation samples, the selection efficiency in a high-profile and high-precision measurement using different decays (Lepton Flavour Universality in $B^+ \to K^- l^+ l^-$ decays [146]) is computed with GAN-based simulation data.

The success of this project will be evaluated on whether the systematic effects induced by using the fast GAN-based simulation are small compared to the precision of the measurement otherwise. In that case there will be no negative impact on the physics from using the GAN-based simulation and CPU and disk space can be saved by its adoption. Alternatively, the same CPU and disk budget could be used to generate more simulation, improving the discovery potential for a few analyses. In practice, the systematic effects are captured by the selection efficiencies. If the selection efficiency between the two types of simulation samples differs by less than 5% in each bin of features such as the square of the four momentum transferred, Q^2, this approach will be deemed feasible for further study.

The observed ratio of GAN to traditional (GEANT) selection efficiencies as a function of Q^2 as shown in figure 4-15 does not deviate from unity by more than 2%. The ratio appears to be dependent on the occupancy in the detector as shown by 4-16, which remains a problem to be studied. To mitigate this issue, separate generative models could be trained for different occupancy bins.

Overall, the study shows that GAN-based simulation has the potential to replace the traditional methods, thereby maximizing the scientific output of the existing detectors.

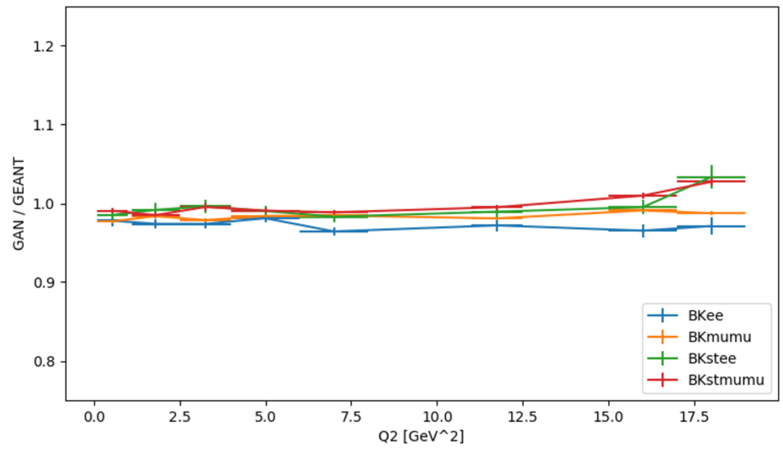


Figure 4-15: The ratio of selection efficiencies for GAN and GEANT simulations for the lepton flavour universality analysis in Ref [146] hovers around 1. This implies that this analysis would likely not suffer from having GAN simulation samples.

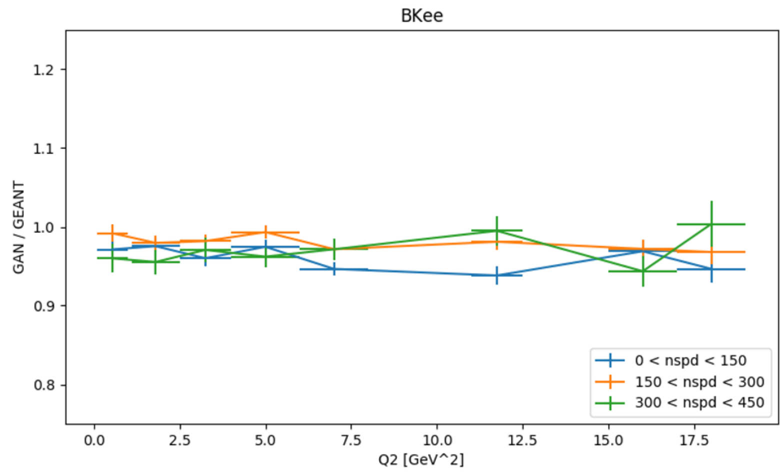


Figure 4-16: The ratio from figure 4-15 is dependent on the occupancy of events.

4.2.7 Summary

Overall, the benefits from utilizing generative models for PID features at experiments like LHCb demand attention. They maximise the data collection and simulation production capacities in a finite computational resource scenario. It is stressed that this procedure is not specific to LHCb and PID features, but can be trivially extended for other experiments and use cases.

4.3 Natural Language Processing

Natural Language Processing (NLP) is the study of automatic manipulation of natural language. As such it lies at the intersection of linguistics, artificial intelligence, and computer science. Some of its tasks include turning a text given in one language into text in another (machine translation), determining the syntactic structure of a text (parsing), and synthesizing the most important information of a text into a shorter text (summarization) [147].

Modern NLP often makes use of neural architectures. In many applications, words are first represented through a vector embedding like Word2Vec, GloVe, and contextualized word embeddings [148–150]. Then, the text is represented with a one dimensional sequence of tokens for which the order matters. Until 2019, the state of the art models were recurrent neural networks (RNNs) like LSTM architectures that tackle variable input lengths by weight sharing and maintaining a memory [151]. Since then, attention based transformer architectures [152] often based on the BERT paradigm [153] have been achieving the best results in many tasks.

Particle physics problems with a similar sequential structure can also be tackled by NLP techniques. One example is the problem of jet tagging that is explored by Ref [154] and section 4.1. Ref [155] increased the background rejection for 50% signal efficiency of top tagging by more than a factor of 2 by replacing fully connected network architectures by LSTMs [156]. Ref [157] further improved jet tagging performance with the use of an attention based architecture. The order of the jet constituents is determined by their momenta or other criteria. As the awareness of NLP based method increases, algorithms such as those used for text summarization could be applied to physics problems of sequential nature.

4.4 A Hybrid Deep Learning Approach to Vertexing

4.4.1 Background

The LHCb experiment will be upgraded substantially as part of the transition to Run 3 in 2022 featuring $\sqrt{s} = 14\,\text{TeV}$ proton-proton collisions as discussed in section 3.5. As part of this effort, the number of expected visible Primary Vertices (PVs) will increase from 1.1. to 5.6 due to a five-fold increase in the design instantaneous luminosity. As a response and to ensure maximum possible physics output, LHCb will adopt Allen, a software trigger developed for graphics processing units (GPUs) running at 30 MHz [158, 159]. This brings an increased interest in GPU friendly algorithms for tracking and reconstruction.

Vertexing is one such reconstruction task that determines the location of vertices of jets of particles, given particle tracks. Existing procedures evaluate each vertex candidate independently, ignoring the high level correlations between candidates. Due to its simplicity of implementation, a simple PV finding GPU algorithm was the baseline solution for LHCb [160]. However, factorizing this problem can lead to sub-optimal vertex location prediction accuracy.

My collaborators and I devised a simple, powerful, and novel deep learning prototype algorithm that takes into account their correlation as an ambitious alternative vertexing algorithm.[3] Its procedure is outlined in figure 4-17 with an implementation

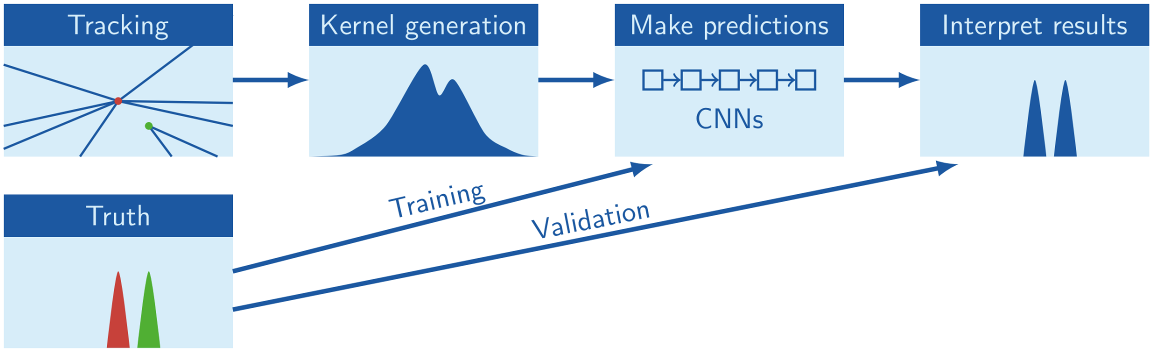


Figure 4-17: Schematic workflow of the proposed hybrid deep learning algorithm for vertex finding. (Credit : Ref [161]).

---

[3] This work is adapted from papers in Ref [161–163] of which I am a co-author.

on gitlab [162, 164]. It first computes the value of a custom kernel for each point in space by summing up the probabilities of the various particles having passed through that point. Because peaks in the discretized kernel projected into one dimension are highly correlated to the positions of the vertices, we then processed this 1D kernel with a Convolutional Neural Network (CNN) to predict the probability of a vertex being present at every point.

Our method is the first of its kind to use CNNs for this task, demonstrates a distinct advantage over traditional non-ML methods and shows promise to tackle operating conditions in Run 3 at LHCb. This is why it is considered for incorporation into Allen with the main complication of realizing ML inference in Allen. While custom-made for LHCb, this project can inspire similar algorithms for other experiments. The rest of this section will show how an initial model (ACAT-2019) [162] achieved better than 94% efficiency on the vertexing task with no more than 0.25 False Positives (FPs) per event and how an improved architecture and improved target histograms [161] reduce the number of false positives per event by a factor of two.

4.4.2 Kernel Generation

Simulated data is produced by generating particle tracks originating from Primary Vertices (PVs) and Secondary Vertices (SVs) with PYTHIA [165, 166], propagating the tracks and intersecting them with the dedicated toy model of the LHCb VELO detector [167] to produce hits in the 26 detection planes of both VELO halves. Scattering in the RF foil and other detector parts is taken into account.

The process of creating tracks from raw hits is called tracking and is handled by Allen. In the toy data setup and for the presented results, "prototracking" is used instead: Hits are radially sorted, grouped into triplets, and the invalid triplets filtered out with a $\chi^2 < 10$ cutoff. The list of these triplets is considered input to the kernel. In parallel to using prototracking, this algorithm has been implemented in the LHCb software stack and been validated using complete LHCbRun 3 MC data with complete VELO tracking.

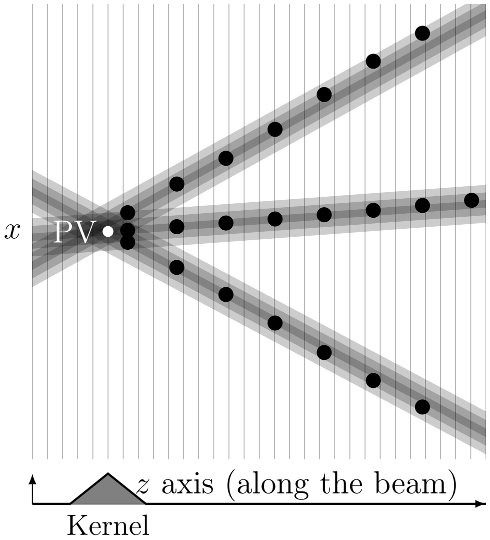


Figure 4-18: The 3D kernel is computed by adding Gaussian contributions of each track to each voxel. The maximum values of voxels in x and y for each z constitute the 1D kernel (Credit : Ref [162]).

Kernel generation produces a dense 1D histogram from spare 3D tracks. Each one of the 4000 $100\mu m$ wide bins in the active VELO area along the beamline (z axis) is filled with the maximum kernel in x and y directions, with the kernel defined as:

$$\mathcal{K}(x,y,z) = \frac{\sum_{\text{tracks}} \mathcal{G}(\Delta_z x, \Delta_z y)^2}{\sum_{\text{tracks}} \mathcal{G}(\Delta_z x, \Delta_z y)} - \sum_{\text{tracks}} \mathcal{G}(\Delta_z x, \Delta_z y) \ , \tag{4.28}$$

where $\mathcal{G}(\Delta_z x, \Delta_z y)$ is the 2D standard normal Gaussian pdf, $\Delta_z x$ is the difference between the x position at which the kernel is evaluated and the x position of a track at a given z divided by its uncertainty. $\Delta_z y$ is defined as the equivalent in the y direction.

The target 4000 bin histogram for the CNN training is generated from PYTHIA's truth information. Vertices with less than 5 detectable tracks within the detector are ignored for training and testing. For every true PV, a normalized Gaussian is added to the target histogram. In the original model (ACAT-2019) the width was taken to be fixed to $100\mu m$. An improvement was made to the target histograms by taking PV resolution due to track multiplicity into account, with resulting efficiency changes shown in figure 4-19.

4.4.3 Network Design

The kernel is processed by a Convolutional Neural Network (CNN) and a list of PV candidate locations is created by running a simple peak finding algorithm on the resulting output probabilities.

The original network (ACAT-2019) was comprised of four convolutional layers with leaky ReLu activations for the hidden layers and softplus activation for the output layer due to this being an easier input for the peakfinding algorithm to handle. It used the Adam optimizer, mini-batch gradient descent, and dropout regularization. All models are trained and tested using PyTorch [168] and in the LHCb software stack inference is run with TorchScript.

A loss function inspired by cross entropy was devised that penalizes fractionally over- or underestimates of the true value by utilizing r_i, the ratio of predicted label $\hat{y}_i$ to real label y_i.

$$r_i \equiv \frac{\hat{y}_i + \epsilon}{y_i + \epsilon}, \tag{4.29}$$

$$z_i \equiv \frac{2r_i}{r_i + r_i^{-1}}, \tag{4.30}$$

$$z_i' \equiv z_i \left(1 + ae^{-r_i}\right), \tag{4.31}$$

$$\text{loss} \equiv -\sum_{\text{bins}} \ln z_i'. \tag{4.32}$$

Small values of $\epsilon = 10^{-5}$ were added to both the numerator and denominator of r_i for mathematically tractability. The asymmetry term a was shown to control the false positive and efficiency tradeoff.

The last convolutional layer and the rest of the architecture are trained in stages. First the last layer is replaced by a fully connected layer and the main architecture trained. Then, the main layers are fixed and the last convolutional layer updated. Lastly, all components are finetuned together. This approach stabilized the training.

An improved architecture was created compared to the ACAT-2019 model by adding a transverse module. The additional information of x and y positions of the kernel maximum at each z was fed into a separate 3 layer CNN. The responses of the

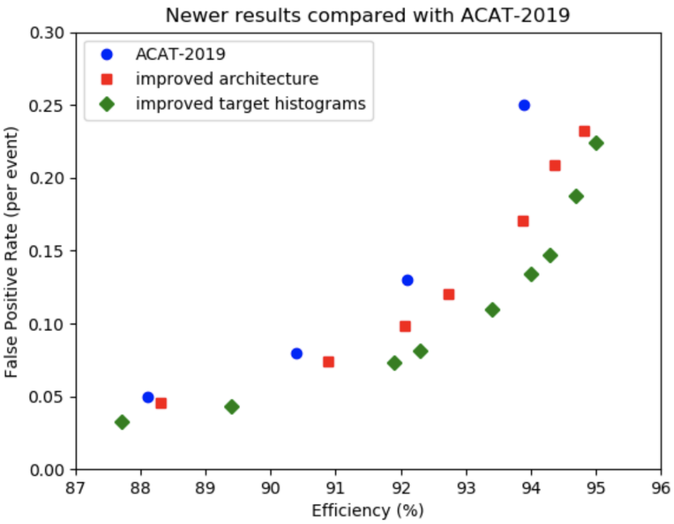


Figure 4-19: The Hybrid deep learning approach to vertexing presented in this work [161] achieves a promising efficiency for a fixed value of false positives. Changes to the architecture and target histograms further increase this efficiency (Credit : Ref [161]).

original model and the transverse module were multiplied. This reflects the expert prior that x and y information should be able to veto vertex candidates due to high x and y gradients that often lead to false positives. However, the transverse module should not singlehandedly set the prediction to true. After retuning, this improved architecture has a softmax rather than a softplus activation for the output layer, and six rather than four convolutional layers.

4.4.4 Results

The two metrics of interest in the PV finding task are the fraction of true vertices discovered, also called efficiency, and the number of locations without a vertex that were predicted to have a vertex per event, the false positives. On the prototracking dataset, these metrics are displayed in figure 4-19.

The original architecture achieves 94% efficiency for under 0.25 false positives per event, which is promising for LHCb. With a modified target histogram, the addition of a transverse module and retuning of the network, the number of false positives per event is reduced by a factor of two. Alternatively, one could say that for 0.14 false positives, the improvements increased the efficiency from 92% to 94%.

For a one-to-one comparison with LHCb's baseline algorithm, a performance study on official LHCb simulation is to follow, but initial studies show comparable performance between the datasets. Due to the official kernels being more pronounced, retraining on the official simulation is likely to bring further efficiency boosts.

4.4.5 Discussion

We developed the first particle PV finding tool that harnessed higher level correlations between different vertices using a convolutional neural network. While preliminary results are promising and a version of the algorithm has been privately deployed into the LHCb CPU software stack, an implementation with LHCb's Allen GPU framework requires further work.

4.5 Goodness-of-Fit and Two-Sample Tests

4.5.1 Setting

Many nuclear and particle physics analyses require high dimensional goodness-of-fit or two-sample tests. One example is when a probability density function (PDF) is fit to data using the unbinned maximum likelihood procedure. As the value of the maximized likelihood cannot be used to determine how well the PDF fit the data, a separate goodness-of-fit procedure is required.

In another relevant setting, the question is whether independently collected data were drawn from the same underlying distribution, regardless of the functional form of such a distribution. Exploring the difference between matter and antimatter decay distributions in some high dimensional parameter space can help shed light on matter antimatter asymmetries and CP violation [169].

In this pedagogical work, we discuss using ML classifiers to project data into a single dimension, so traditional low-dimensional tests can be performed. This paper aims to translate these tests for use in physics, shows a software implementation, and encourages adoption. The inclusion of systematic uncertainties is discussed.[4]

4.5.2 Formalism

Let us denote a d-dimensional vector of features as $\vec{x}$ and the corresponding PDF as $f(\vec{x})$. In this work, the null hypothesis holds when two distributions are the same, while the alternate hypothesis holds when they are different:

$$\begin{aligned} H_0 &: f_a(\vec{x}) = f_b(\vec{x}) \\ H_1 &: f_a(\vec{x}) \neq f_b(\vec{x}) \end{aligned} \tag{4.33}$$

A goodness-of-fit task tests whether the empirical distribution $f_a(\vec{x})$ is different from the fit distribution $f_b(\vec{x})$, possibly retrieved from an unbinned maximum likelihood fit. For a two-sample test, $f_a(\vec{x})$ and $f_b(\vec{x})$ are the empirical PDFs of the two samples.

---

[4] This work is taken from a previously presented paper in reference [170] of which I am a co-author.

The frequentist p-value of these tests is defined as the probability of observing a more extreme value of the test statistic than the actual observed value under the null hypothesis. It should be noted that a goodness-of-fit test can be recast into a two-sample test by sampling from the fit PDF, which is assumed for the rest of this work.

4.5.3 Approaches

χ^2 and Univariate Methods

The go-to solutions for practitioners in particle physics is the χ^2 test [171], even though other techniques are relevant (e.g. [172, 173]). Like a variety of other tests, the χ^2 test suffers from the curse of dimensionality [174], because due to the increasing data sparsity it rapidly loses power as the dimensionality grows.

While tests like Kolmogorov-Smirnov (KS) [175, 176], Anderson Darling (AD) [177], and Cramér-van-Mises (CvM) [178, 179] perform well in a single dimension, the extension to higher dimensions is not trivial. This is due to a lack of a canonical ordering of dimensions and difficulties with extensions of concepts like ranks [180].

Unsupervised Techniques

One could imagine turning to unsupervised machine learning techniques to solve this task. For example, one could compress the high dimensional data using Principal Component Analysis or Autoencoders followed by a traditional 1D goodness-of-fit test.

While these tests would provide valid p-values, the power in the Neaman-Pearson paradigm would be suboptimal. This is, because the dimension reduction was performed without the specific task of separating two specific data samples.

Measure Transportation

Motivated by measure transportation, Ref. [180] proposes a non-parametric distribution-free two-sample goodness-of-fit test that uses the Halton sequence, empirical range map, and energy statistic. Their method reduces to the Cramér-von Mises statistic for one dimension. While their method is promising, it was not available at the writing of this pedagogical paper and would be a good candidate for further exploration.

Supervised Techniques

An alternative is to reduce dimensionality by minimizing a loss function based on the differences in samples. This can be achieved through supervised machine learning as first introduced by Friedman [181] and studied by us [170]. Here, we use a training set to train a classifier to separate two samples. The classifier scores on examples from unseen test sets of both samples are computed, and their compatibility tested with a traditional 1D test.

The response of some ML classifiers is discrete which violates assumptions of some tests like KS. In these cases, a permutation test has to be applied to achieve correct p-values [182].

When the difference between the distributions is practically meaningful, can be described analytically, and a powerful, specialized two-sample test statistic exists, then this specialized should be used. It will often have larger power than the classification accuracy, which is especially important if the data is not abundant. However, in the general setting studied here, the difference between the distributions is not practically meaningful. According to Kim et. al., who study theoretical guarantees for a specialized problem [183], the flexible approach of supervised techniques is advisable in such cases.

Most of the particle physics community seems to be unaware of the possibility of the use case of supervised learning for high dimensional goodness-of-fit and two-sample tests. However, their general ease of use, and available software are strong arguments for adopting a classification-based approach to high-dimensional two-sample tests in high-energy physics.

4.5.4 Systematic Uncertainties

Systematic uncertainties can be incorporated into a supervised ML two-sample test by binning the classifier response and choosing the χ^2 test to obtain the p-value. This way, systematic uncertainties can by introduced bin-by-bin in any of the ways common in particle physics.

One procedure is to assign a random weighting to the input data points to account for systematic effects, and to repeat the supervised ML testing steps to produce a one-dimensional response function. The shifts in the response function yield in each bin is then assigned as the systematic uncertainty.

Another procedure to include systematic uncertainties is to generate samples with systematic effects incorporated, to repeat the supervised ML testing steps, and to incorporate systematic uncertainties bin-by-bin as above.

For the remainder of this work we assume a systematic uncertainty of 1% in all bins to demonstrate how to include these uncertainties. As the number of events are equal for the two samples, the χ^2 statistic is calculated as follows to account for these uncertainties:

$$\chi^2 = \sum_{k=1}^{n_{bins}} \frac{(a_k - b_k)^2}{a_k + b_k + (0.01 \cdot a_k)^2 + (0.01 \cdot b_k)^2}, \tag{4.34}$$

where a_k (b_k) denotes the contents of the kth bin of sample a (b).

Gaussian smearing is used to implement the systematic uncertainties. Practically, this means that for both the multivariate χ^2 tests and the univariate χ^2 tests on the classifier response, the values of a_k and b_k inserted into the statistic above are sampled from a Gaussian. The mean was the generated count in the corresponding bin and the standard deviation 1% of that value. For real-wold analyses, existing systematic uncertainties should be accounted for, but no new ones introduced as explained here.

4.5.5 Experimental Setup

To experimentally compare the power of different tests, a χ^2 test, and ML tests as explained in section 4.5.3 were run on example problems. Our implementation of the χ^2 test uses adaptive binning defined such that the expected number of points is equal for each bin in the infinite sample limit under the null hypothesis [184]. The number of bins was treated as a hyperparameter and its optimal value was in all cases lower than naively expected.

As ease of use and availability is important to particle physics practitioners, the ML algorithms below were selected.

1. BDT: boosted decision trees with a tunable number of trees using XGBoost [185]

2. ANN: shallow, fully connected neural networks with zero or one hidden layer and a tunable number of neurons per layer

3. SVM: support vector machines with a the radial basis function kernel and variable C and γ parameters

Hyperparameters were optimized on a validation data set using Bayesian optimization as implemented in the SPEARMINT package [186]. While other example settings were studied, here a Gaussian setting is shown because of its pedagogical value, followed by a more application-focused setting regarding a two-body particle decay.

For each setting, problems were generated for correct ($f_a = f_1$ and $f_b = f_1$) and incorrect ($f_a = f_1$ and $f_b = f_2$) null hypothesis for a number of dimensions ranging from 2 to 10. In problems with a correct null hypothesis, it was confirmed that the p-values were distributed uniformly between 0 and 1. Problems with an incorrect null hypothesis were used to test the power of the tests by counting the fraction of tests rejecting the null hypothesis at 95% confidence level. In each problem, 100 data sets with ten thousand examples were sampled.

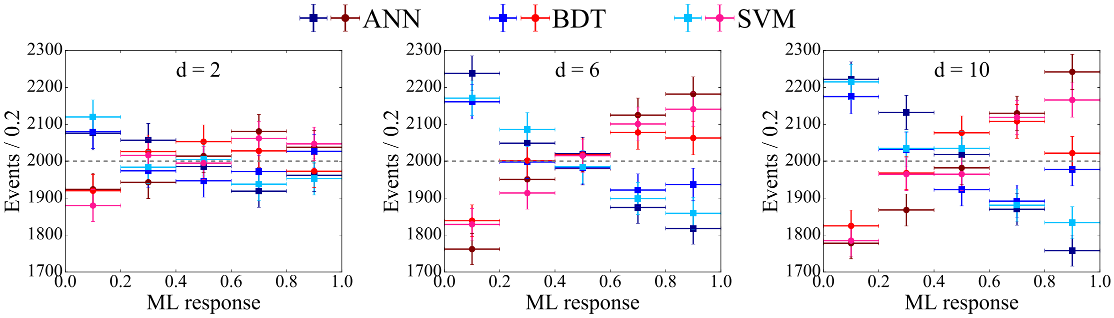


Figure 4-20: Gaussian setting: ML response distributions for example data sets sampled from f_1 and f_2 for (left)d=2, (middle)d=6, and (right)d=10 (Credit: Ref. [170]).

4.5.6 Gaussian Setting

In this setting, data are sampled iid from a d-dimensional Gaussian distribution with a zero vector mean and diagonal covariance matrix as follows:

$$f_b(\vec{x}) = \prod_{k=1}^{d} \mathcal{G}(x_k|\mu_k = 0, \sigma_k), \tag{4.35}$$

where $\sigma_k = 1$ for f_1 and $\sigma_k = 0.95$ for f_2.

Valid p values were observed under the null hypothesis, as shown in the right of figure 4-21. For the setting in which the alternate hypothesis is correct, samples were generated and some example ML-response distributions displayed in Fig 4-20.

In this Gaussian setting, the ANN and SVM demonstrated a larger power than the BDT algorithms (see the left of figure 4-21). This was not surprising as the PDFs were hyperspherical, which does not lend itself to the one-dimensional (rectangular) splitting based approach that boosted decision trees take.

The main take-away of this setting is that, as expected, the curse of dimensionality affects the χ^2 test more heavily than the ML-based tests. Were the underlying model known, one could design a powerful test by handcrafting a feature. In this case, running a 1D test on the d-dimensional Euclidean distance from the origin would result in close to optimal power. In fact, such handcrafted features are common in particle physics and are encouraged. However, when it is difficult to find such features, the ML-based approach is more flexible and powerful than traditional methods.

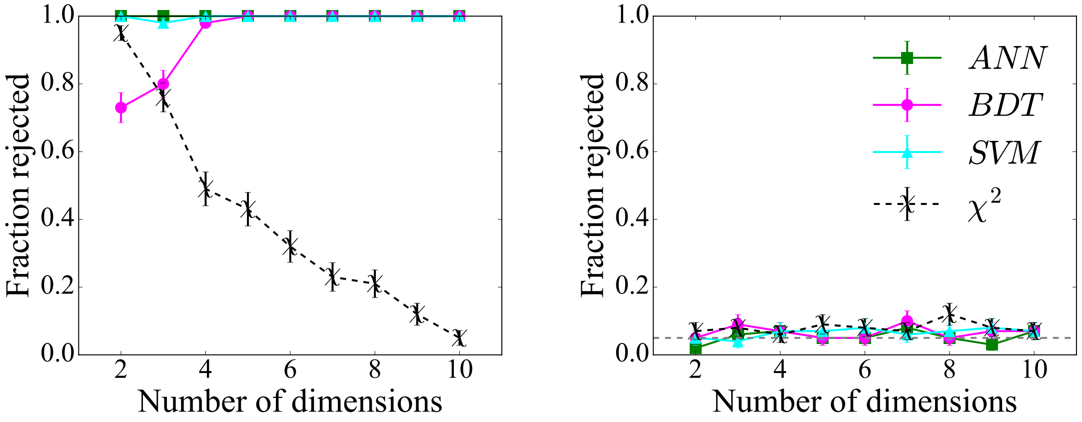


Figure 4-21: Gaussian example problem: Power of each method versus dimensionality d (Credit: Ref. [170]).

4.5.7 Two-Body Decay Setting

This setting represents a simple, real-world physics use case of discovering a new particle in a two body decay. The energy and momenta of two particles α and β, with masses $m(\alpha)$ and $m(\beta)$, are measured producing an 8-dimensional feature space.

The background process f_1 consists of processes with an invariant mass uniformly distributed on $[0.5m(X)$,$1.5m(X)]$ decaying into particles α and β. In the signal process $X \rightarrow \alpha\beta$, a new, to-be-discovered particle X of mass $m(X)$ and effective resolution $m(X)/10$ decays into two particles of lighter equal mass. The generative process for f_2 consists of binary sampling with probability of success of p_{signal}. In case of success, a data point is generated using the signal process, and the background process, otherwise.

In a real search, the collected sample is to be compared to a sample that is known to be generated by the background process f_1. This sample could be generated using simulation, collected from same-sign candidates in data, or other means. Under the null hypothesis, both samples a and b originate from the background process and uniform p-values were observed.

The ML-based and χ^2 tests were run on the 8-dimensional problem, as p_{signal} was varied from zero to 10% under the alternate hypothesis with results displayed in figure 4-22. As expected, the classification tests surpass the χ^2 in power.

However, from special relativity it is known that the invariant mass is an effective

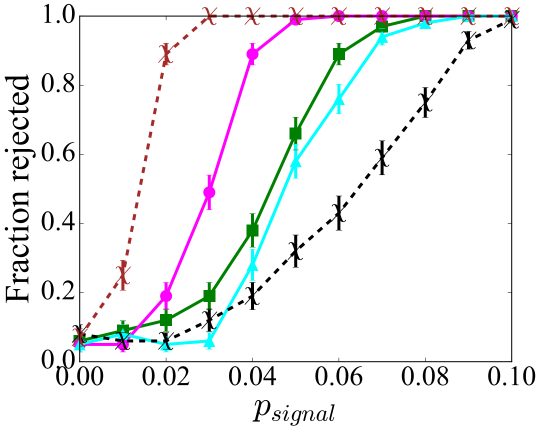


Figure 4-22: Two Body Decay Setting: Power of each method versus signal contribution fraction. The markers are the same as in figure 4-21 with the additional red χ^2 marker for the one dimensional χ^2 test on the invariant mass feature (Credit: Ref. [170]).

discriminating feature and can be constructed from the 8 features. Indeed, a χ^2 test trained on this feature alone outperforms all other tests shown here. While more powerful ML algorithms should be capable of bridging this performance gap [187], this example setting is a clear reminder that specialized information should be incorporated into tests if available. In their absence, classification-based tests are a flexible alternative.

4.5.8 Summary

While a variety of powerful methods for multivariate goodness-of-fit and two-sample tests exist for low dimensional feature spaces, such tests rapidly lose power as the dimensionality increases and the data inevitably become sparse. We show that flexible, easy to use, and powerful non-parametric two-sample and goodness-of-fit tests can be performed using ML classifiers to project data into one dimension, so traditional low-dimensional tests can be run [170]. This work aims to translate these tests for use in physics, demonstrates ease of use, and encourages adoption.

Chapter 5

Dark Photon Searches

Searches are performed for prompt-like and long-lived dark photons, A', decaying into two muons, $A' \to \mu^+\mu^-$, and other low-mass dimuon resonances, X, produced in proton-proton collisions at a center-of-mass energy of 13 TeV. The data samples for A' and X searches were collected with the LHCb detector and correspond to an integrated luminosity of 5.1 fb^{-1}. The prompt-like A' search explores the mass region from near the dimuon threshold up to 70 GeV, and places the most stringent constraints to date on dark photons with $214 < m(A') \lesssim 740\,\text{MeV}$ and $10.6 < m(A') \lesssim 30\,\text{GeV}$. The search for long-lived $A' \to \mu^+\mu^-$ decays places world-leading constraints on low-mass dark photons with lifetimes $\mathcal{O}(1)$ ps. The $X \to \mu^+\mu^-$ decays can be either prompt-like or displaced from the proton-proton collision, where in both cases the requirements placed on the event and the assumptions made about the production mechanisms are kept as minimal as possible. The prompt-like X searches explore the mass range from near the dimuon threshold up to 60 GeV, with nonnegligible X widths considered above 20 GeV. The searches for displaced $X \to \mu^+\mu^-$ decays consider masses up to 3 GeV. None of the searches finds evidence for a signal and 90% confidence-level exclusion limits are placed on the γ–A' kinetic-mixing strength for A' searches and on the $X \to \mu^+\mu^-$ cross sections for X searches. In addition, these results are used to place world-leading constraints on two-Higgs-doublet and hidden-valley scenarios.[1]

---

[1] This work is taken from two previously published papers [188, 189] of which I am a proponent.

5.1 Overview

Substantial effort has been dedicated recently [190–192] to searching for the dark photon, A', a hypothetical massive vector boson that could mediate the interactions of dark matter particles[193], similar to how the ordinary photon, γ, mediates the electromagnetic (EM) interactions of charged Standard Model (SM) particles. The dark photon as described in section 2.3 does not couple directly to SM particles; however, it can obtain a small coupling to the EM current due to kinetic mixing between the SM hypercharge and A' field strength tensors [194–199]. This coupling, which is suppressed relative to that of the photon by a factor labeled ε, would provide a portal through which dark photons can be produced in the laboratory, and also via which they can decay into visible SM final states. If the kinetic mixing arises due to processes described by one- or two-loop diagrams containing high-mass particles, possibly even at the Planck scale, then $10^{-12} \lesssim \varepsilon^2 \lesssim 10^{-4}$ is expected [191]. Exploring this *few-loop* ε region is one of the most important near-term goals of dark-sector physics.

Constraints have been placed on visible A' decays by previous beam-dump [199–215], fixed-target [71, 216–218], collider [219–224], and rare-meson-decay [225–234] experiments. These experiments ruled out the few-loop region for dark-photon masses $m(A') \lesssim 10\,\mathrm{MeV}$ ($c = 1$ throughout this chapter); however, most of the few-loop region at higher masses remains unexplored. Decays into invisible dark-sector particles are expected to be dominant if $m(A')$ is large enough to allow such decays. Constraints on invisible A' decays can be found in Refs. [235–247]; only the visible scenario is considered here.

Many proposals have been put forward to further explore the $[m(A'), \varepsilon^2]$ parameter space [248–268]. For example, Ref. [262] proposed an inclusive search for $A' \to \mu^+\mu^-$ decays with the LHCb experiment, and showed that such a search will provide sensitivity to large regions of otherwise inaccessible parameter space with the data collected by the end of Run 3 in 2023. The LHCb collaboration first performed this search using Run 2 data corresponding to $1.6\,\mathrm{fb}^{-1}$ collected in 2016 [269]. The constraints placed on prompt-like dark photons, where the dark-photon lifetime is small compared to the detector resolution, were the most stringent to date for $10.6 < m(A') < 70\,\mathrm{GeV}$ and comparable to the best existing limits for $m(A') < 0.5\,\mathrm{GeV}$. The search for long-lived dark photons was the first to achieve sensitivity using a displaced-vertex signature, though only small regions of $[m(A'), \varepsilon^2]$ parameter space were excluded.

The minimal A' model is not the only viable dark-sector scenario. The strongest connection to the dark sector may not arise via kinetic mixing, and the dark sector itself could be populated by additional particles that have phenomenological implications. Searches for dark photons can provide serendipitous discovery potential for other types of particles, especially vector particles that share the same production mechanisms as the minimal dark photon [20], yet many well-motivated models would have avoided detection in all previous experimental searches [270, 271]. For example, hidden-valley (HV) scenarios that exhibit confinement produce a high multiplicity of light hidden hadrons from showering processes [25]. These hidden hadrons would typically decay displaced from the proton-proton collision, thus failing the criteria employed in Refs. [269, 272] to suppress backgrounds due to heavy-flavor quarks [258, 262]. Furthermore, the sensitivity to various model scenarios can be improved by exploiting additional signatures (e.g. the presence of a b-quark jet produced in association with the X boson [21]). Therefore, it is desirable to perform searches that are less model dependent, including some that explore additional signatures in the event.

This chapter presents searches for both prompt-like and long-lived dark photons produced in proton-proton, pp, collisions at a center-of-mass energy of 13 TeV, looking for $A' \to \mu^+\mu^-$ decays using a data sample corresponding to an integrated luminosity of $5.1\,\mathrm{fb}^{-1}$ collected with the LHCb detector in 2016–2018. The LHCb detector is a single-arm forward spectrometer covering the pseudorapidity range $2 < \eta < 5$, described in detail in chapter 3. The strategies employed in these searches are the same as in Ref. [269], though the three-fold increase in integrated luminosity, improved trigger efficiency during 2017–2018 data taking, and improvements in the analysis provide much better sensitivity to dark photons. The prompt-like A' search is performed from near the dimuon threshold up to 70 GeV, achieving a factor of 5 (2) better sensitivity to ε^2 at low (high) masses than Ref. [269]. It is based on a data sample that employs the novel TURBO data-storage strategy explained in chapter 3.4, greatly reducing the event size. The long-lived A' search is restricted to the mass range $214 < m(A') < 350\,\text{MeV}$, where the data sample potentially has sensitivity, and provides access to much larger regions of $[m(A'), \varepsilon^2]$ parameter space.

Furthermore, this chapter also presents other searches for low-mass dimuon resonances in the same proton-proton collisions with an integrated luminosity of $5.1\,\text{fb}^{-1}$. The $X \to \mu^+\mu^-$ decays can be either prompt-like, *i.e.* consistent with being prompt, or displaced from the proton-proton collision. In both cases, the requirements placed on the event and the assumptions made about the production mechanisms are kept as minimal as possible. Two variations of the prompt-like X search are performed: an inclusive version, and an $X + b$ search, where the X boson is required to be produced in association with a beauty quark. Two variations are also considered of the search for displaced $X \to \mu^+\mu^-$ decays: an inclusive version, and one where the X boson is required to be produced promptly in the proton-proton collision. The prompt-like X searches explore the mass range from near the dimuon threshold up to $60\,\text{GeV}$, with nonnegligible widths, $\Gamma(X)$, considered above $20\,\text{GeV}$. The searches for displaced $X \to \mu^+\mu^-$ decays consider masses up to $3\,\text{GeV}$.

The online event selection is performed by a trigger [273] consisting of a hardware stage using information from the calorimeter and muon systems, followed by a software stage that performs a full event reconstruction. At the hardware stage, A' events are required to have a muon with momentum transverse to the beam direction $p_{\mathrm{T}}(\mu) \gtrsim 1.8\,\text{GeV}$, or a dimuon pair with $p_{\mathrm{T}}(\mu^+)p_{\mathrm{T}}(\mu^-) \gtrsim (1.5\,\text{GeV})^2$, while non-minimal A' events are required to have a dimuon pair with $p_{\mathrm{T}}(\mu^+)p_{\mathrm{T}}(\mu^-) \gtrsim (1.5\,\text{GeV})^2$. In both settings, events with more than 900 hits in the scintillating-pad detector are discarded, which prevents high-occupancy events from dominating the processing time in the software trigger stages. The long-lived A' search also uses events selected at the hardware stage independently of the $A' \to \mu^+\mu^-$ candidate.

In the software stage, where the p_{T} resolution is substantially improved *cf.* the hardware stage, $A' \to \mu^+\mu^-$ and $X \to \mu^+\mu^-$ candidates are built from two oppositely charged tracks that form a good-quality vertex and satisfy stringent muon-identification criteria, though these criteria were loosened considerably in the low-mass region during 2017–2018 data taking. All searches require $p_{\mathrm{T}}(A') > 1\,\mathrm{GeV}$ and $2 < \eta(\mu) < 4.5$. The prompt-like A' and $X \to \mu^+\mu^-$ searches use muons that are consistent with originating from the PV, with $p_{\mathrm{T}}(\mu) > 1.0\,\mathrm{GeV}$ and momentum $p(\mu) > 20\,\mathrm{GeV}$ in 2016, and $p_{\mathrm{T}}(\mu) > 0.5\,\mathrm{GeV}$, $p(\mu) > 10\,\mathrm{GeV}$, and $p_{\mathrm{T}}(\mu^+)p_{\mathrm{T}}(\mu^-) > (1.0\,\mathrm{GeV})^2$ in 2017–2018. The searches for displaced A' and $X \to \mu^+\mu^-$ decays use muons with $p_{\mathrm{T}}(\mu) > 0.5\,\mathrm{GeV}$ and $p(\mu) > 10\,\mathrm{GeV}$ that are inconsistent with originating from any PV, and require $2 < \eta(A') < 4.5$. In addition, the searches for a long-lived A' and a long-lived promptly produced X boson requires a decay topology consistent with a dimuon resonance originating from a PV.

Simulation is required to model the effects of the detector acceptance and its response to A' and $X \to \mu^+\mu^-$ decays. In the simulation, pp collisions are generated using PYTHIA 8 [274, 275] with a specific LHCb configuration [276]. Decays of unstable particles are described by EVTGEN [277], in which final-state radiation is generated using PHOTOS [278]. The interaction of the generated particles with the detector, and its response, are implemented using the GEANT4 toolkit [279, 280] as described in Ref. [281]. Simulation is also used to place constraints on specific models. Prompt limits for light-pseudoscalar models are set with next-to-next-to-leading order cross-sections from HIGLU [282, 283] using the NNPDF3.0 PDF set [284], branching fractions from HDECAY [285, 286], and fiducial acceptances from PYTHIA 8 [287]. Displaced limits for HV models are set with PYTHIA 8 [287] using a running α_{HV} scheme [288], and couplings from DARKCAST [20].

5.2 Minimal A'

Both the production and decay kinematics of the $A' \to \mu^+\mu^-$ and $\gamma^* \to \mu^+\mu^-$ processes are identical, since dark photons produced in *pp* collisions via γ–A' mixing inherit the production mechanisms of off-shell photons with $m(\gamma^*) = m(A')$. Furthermore, the expected $A' \to \mu^+\mu^-$ signal yield is related to the observed prompt $\gamma^* \to \mu^+\mu^-$ yield in a small $\pm\Delta m$ window around $m(A')$, $n_{\rm ob}^{\gamma^*}[m(A')]$, by [262]

$$n_{\rm ex}^{A'}[m(A'), \varepsilon^2] = \varepsilon^2 \left[\frac{n_{\rm ob}^{\gamma^*}[m(A')]}{2\Delta m}\right] \mathcal{F}[m(A')]\, \epsilon_{\gamma^*}^{A'}[m(A'), \tau(A')], \tag{5.1}$$

where the dark-photon lifetime, $\tau(A')$, is a known function of $m(A')$ and ε^2, $\mathcal{F}$ is a known $m(A')$-dependent function, and $\epsilon_{\gamma^*}^{A'}[m(A'), \tau(A')]$ is the $\tau(A')$-dependent ratio of the $A' \to \mu^+\mu^-$ and $\gamma^* \to \mu^+\mu^-$ detection efficiencies. For prompt-like dark photons, $A' \to \mu^+\mu^-$ decays are experimentally indistinguishable from prompt $\gamma^* \to \mu^+\mu^-$ decays, resulting in $\epsilon_{\gamma^*}^{A'}[m(A'), \tau(A')] \approx 1$. This facilitates a fully data-driven search where most experimental systematic effects cancel, since the observed $A' \to \mu^+\mu^-$ yields, $n_{\rm ob}^{A'}[m(A')]$, can be normalized to $n_{\rm ex}^{A'}[m(A'), \varepsilon^2]$ to obtain constraints on ε^2 without any knowledge of the detector efficiency or luminosity. When $\tau(A')$ is larger than the detector decay-time resolution, $A' \to \mu^+\mu^-$ decays can potentially be reconstructed as displaced from the primary *pp* vertex (PV) resulting in $\epsilon_{\gamma^*}^{A'}[m(A'), \tau(A')] \neq 1$; however, only the $\tau(A')$ dependence of the detection efficiency is required to use Eq. (5.1). Finally, Eq. (5.1) is altered for large $m(A')$ to account for additional kinetic mixing with the Z boson [289, 290].

5.2.1 Prompt

The prompt-like A' sample is contaminated by prompt $\gamma^* \to \mu^+\mu^-$ production, various resonant decays to $\mu^+\mu^-$, whose mass-peak regions are avoided in the search, and by the following types of misreconstruction: (hh) two prompt hadrons misidentified as muons; ($h\mu_Q$) a misidentified prompt hadron combined with a muon produced in the decay of a heavy-flavor quark, Q, that is misidentified as prompt; and ($\mu_Q\mu_Q$)

two muons produced in Q-hadron decays that are both misidentified as prompt. The impact of the $\gamma^* \to \mu^+\mu^-$ background is reduced, *cf.* Ref. [269], by constraining the muons to originate from the PV when determining $m(\mu^+\mu^-)$, which improves the resolution, $\sigma[m(\mu^+\mu^-)]$, by about a factor of 2 for small $m(A')$. The misreconstructed backgrounds are highly suppressed by the stringent muon-identification and prompt-like requirements applied in the trigger; however, substantial contributions remain for $m(A') \gtrsim 1.1\,\text{GeV}$. In this mass region, dark photons are expected to be predominantly produced in Drell–Yan processes, from which they would inherit the well-known signature of dimuon pairs that are largely isolated. Therefore, the signal sensitivity is enhanced by applying the anti-k_{T}-based [291–293] isolation requirement described in Refs. [269, 294] for $m(A') > 1.1\,\text{GeV}$.

The observed prompt-like $A' \to \mu^+\mu^-$ yields, which are determined from fits to the $m(\mu^+\mu^-)$ spectrum, are normalized using Eq. (5.1) to obtain constraints on ε^2. The $n_{\text{ob}}^{\gamma^*}[m(A')]$ values in Eq. (5.1) are obtained from binned extended maximum likelihood fits to the $\min[\chi^2_{\text{IP}}(\mu^\pm)]$ distributions, where $\chi^2_{\text{IP}}(\mu)$ is defined as the difference in the vertex-fit χ^2 when the PV is reconstructed with and without the muon. The $\min[\chi^2_{\text{IP}}(\mu^\pm)]$ distribution provides excellent discrimination between prompt muons and the displaced muons that constitute the $\mu_Q\mu_Q$ background. Since $\chi^2_{\text{IP}}(\mu)$ approximately follows a χ^2 probability density function (PDF), with two degrees of freedom, the $\min[\chi^2_{\text{IP}}(\mu^\pm)]$ distributions have minimal mass dependence for each source of dimuon candidates. The prompt-dimuon PDFs are taken directly from data at $m(J/\psi)$ and $m(Z)$, where prompt resonances are dominant. Small corrections are applied to obtain these PDFs at all other $m(A')$, which are validated near threshold, at $m(\phi)$, and at $m[\Upsilon(1S)]$, where the data predominantly consist of prompt dimuon pairs. Based on these validation studies, a small shape uncertainty is applied in each $\min[\chi^2_{\text{IP}}(\mu^\pm)]$ bin. Same-sign $\mu^\pm\mu^\pm$ candidates provide estimates for the PDF and yield of the sum of the hh and $h\mu_Q$ contributions, where each involves misidentified prompt hadrons. The $\mu^\pm\mu^\pm$ yields are corrected to account for the difference in the production rates of $\pi^+\pi^-$ and $\pi^\pm\pi^\pm$, since the hh background largely consists of $\pi^+\pi^-$ pairs where both pions are misidentified. The uncertainty due to the finite size of the

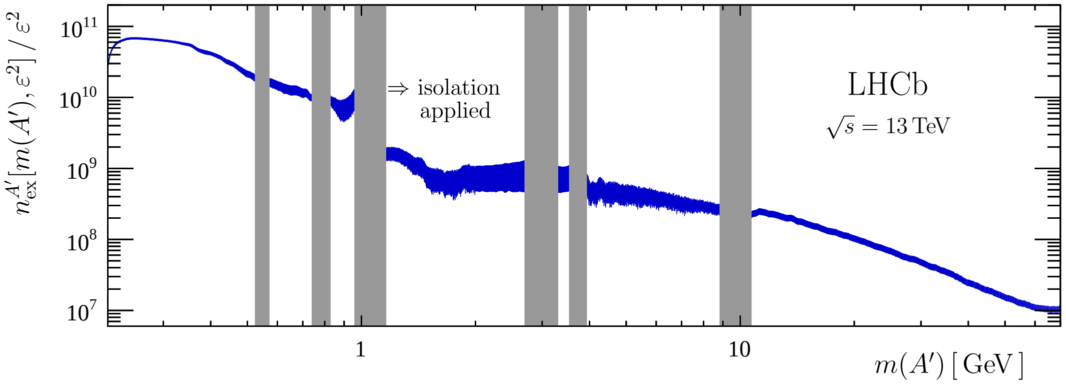


Figure 5-1: Expected reconstructed and selected prompt-like $A' \to \mu^+\mu^-$ yield divided by ε^2, where the displayed uncertainties include the systematic contributions. The gray boxes cover the regions with large SM resonance contributions, where no search for dark photons is performed. The anti-$k_{\rm T}$-based isolation requirement is applied for $m(A') > 1.1\,\text{GeV}$ (Credit: Ref. [188]).

$\mu^\pm\mu^\pm$ sample in each bin is included in the likelihood. Simulated Q-hadron decays are used to obtain the $\mu_Q\mu_Q$ PDFs, where the dominant uncertainties are from the relative importance of the various Q-hadron decay contributions at each mass. Example $\min[\chi^2_{\rm IP}(\mu^\pm)]$ fits, and the resulting prompt-like candidate categorization versus $m(\mu^+\mu^-)$, are provided in Ref. [294]. Finally, the $n_{\rm ob}^{\gamma^*}[m(A')]$ yields are corrected for bin migration due to bremsstrahlung, which is negligible except near the low-mass tails of the J/ψ and $\Upsilon(1S)$, and the small expected Bethe–Heitler contribution is subtracted [262], resulting in the $n_{\rm ex}^{A'}[m(A'),\varepsilon^2]$ values shown in Fig. 5-1.

The prompt-like mass spectrum is scanned in steps of $\sigma[m(\mu^+\mu^-)]/2$ searching for $A' \to \mu^+\mu^-$ contributions [294], using the strategy from Ref. [269]. At each mass, a binned extended maximum likelihood fit is performed in a $\pm 12.5\,\sigma[m(\mu^+\mu^-)]$ window around $m(A')$. The profile likelihood is used to determine the p-value and the upper limit at 90% confidence level (CL) on $n_{\rm ob}^{A'}[m(A')]$. The signal mass resolution is determined with 10% precision using a combination of simulated $A' \to \mu^+\mu^-$ decays and the observed $p_{\rm T}$-dependent widths of the large resonance peaks in the data. The method of Ref. [295] selects the background model from a large set of potential components, which includes all Legendre modes up to tenth order and dedicated terms for known resonances, by performing a data-driven process whose uncertainty

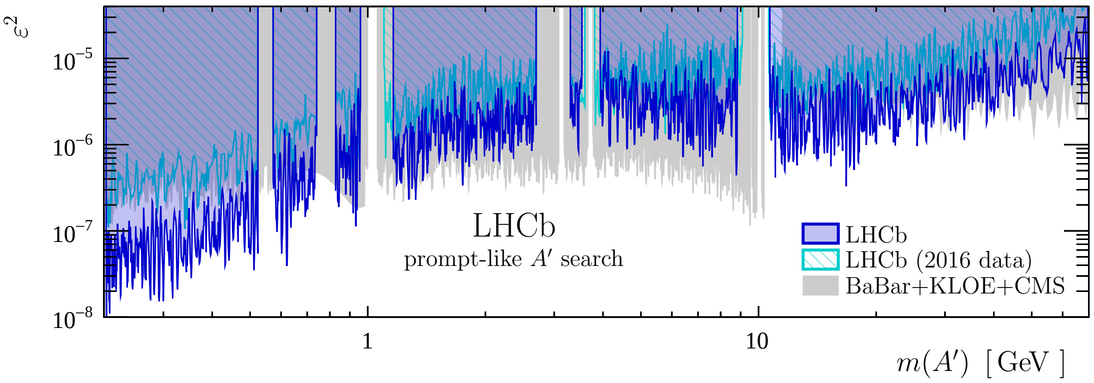


Figure 5-2: Regions of the $[m(A'), \varepsilon^2]$ parameter space excluded at 90% CL by the prompt-like A' search compared to the best published [221, 224, 269] and preliminary [297] limits (Credit: Ref. [188]).

is included in the profile likelihood following Ref. [296]. No significant excess is found in the prompt-like $m(A')$ spectrum, after accounting for the trials factor due to the number of signal hypotheses.

Dark photons are excluded at 90% CL where the upper limit on $n_{\rm ob}^{A'}[m(A')]$ is less than $n_{\rm ex}^{A'}[m(A'), \varepsilon^2]$. Figure 5-2 shows that the constraints placed on prompt-like dark photons are the most stringent for $214 < m(A') \lesssim 740\,{\rm MeV}$ and $10.6 < m(A') \lesssim 30\,{\rm GeV}$. The low-mass constraints are the strongest placed by a prompt-like A' search at any $m(A')$. These results are corrected for inefficiency that arises due to $\tau(A')$ no longer being negligible at such small values of ϵ^2. The high-mass constraints are adjusted to account for additional kinetic mixing with the Z boson [289, 290], which alters Eq. (5.1). Since the LHCb detector response is independent of which Drell–Yan process produces the dark photon above 10 GeV, it is straightforward to recast the results in Fig. 5-2 for other models [20].

5.2.2 Displaced

For the long-lived A' search, contamination from prompt particles is negligible due to the stringent criteria applied in the trigger. Therefore, the dominant background contributions are: photons that convert into $\mu^+\mu^-$ in the silicon-strip vertex detector that surrounds the pp interaction region, known as the VELO [298]; b-hadron decay chains that produce two muons; and the low-mass tail from $K^0_\mathrm{S} \to \pi^+\pi^-$ decays, where both pions are misidentified as muons. A p-value is assigned to the photon-conversion hypothesis for each long-lived $A' \to \mu^+\mu^-$ candidate using properties of the decay vertex and muon tracks, along with a high-precision three-dimensional material map produced from a data sample of secondary hadronic interactions [299]. A $m(A')$-dependent requirement is applied to these p-values to reduce conversions to a negligible level. The remaining backgrounds are highly suppressed by the decay topology requirement applied in the trigger. Furthermore, since muons produced in b-hadron decays are often accompanied by additional displaced tracks, events are rejected if they are selected by the inclusive heavy-flavor software trigger [300, 301] independently of the presence of the $A' \to \mu^+\mu^-$ candidate. In addition, boosted decision tree classifiers are used to reject events containing tracks consistent with originating from the same b-hadron decay as the signal muon candidates [302].

The long-lived A' search is also normalized using Eq. (5.1); however, in this case $\epsilon^{A'}_{\gamma^*}[m(A'),\tau(A')]$ is not unity, in part because the efficiency depends on the decay time, t. The kinematics are identical for $A' \to \mu^+\mu^-$ and prompt $\gamma^* \to \mu^+\mu^-$ decays for $m(A') = m(\gamma^*)$; therefore, the t dependence of $\epsilon^{A'}_{\gamma^*}[m(A'),\tau(A')]$ is obtained by resampling prompt $\gamma^* \to \mu^+\mu^-$ candidates as long-lived $A' \to \mu^+\mu^-$ decays, where all t-dependent properties, e.g. $\min[\chi^2_\mathrm{IP}(\mu^\pm)]$, are recalculated based on the resampled decay-vertex locations (the impact of background contamination in the prompt $\gamma^* \to \mu^+\mu^-$ sample is negligible). This approach is validated using simulation, where prompt $A' \to \mu^+\mu^-$ decays are used to predict the properties of long-lived $A' \to \mu^+\mu^-$ decays. The relative uncertainty on $\epsilon^{A'}_{\gamma^*}[m(A'),\tau(A')]$ is estimated to be 5%, which arises largely due to limited knowledge of how radiation damage affects the performance of the VELO as a function of the distance from the pp interaction region. The looser kinematic, muon-identification, and hardware-trigger requirements applied to long-lived $A' \to \mu^+\mu^-$ candidates, *cf.* prompt-like candidates, also increase the efficiency. This t-independent increase in efficiency is determined using a control data sample of dimuon candidates consistent with originating from the PV, but otherwise satisfying the long-lived criteria. The $n^{A'}_\mathrm{ex}[m(A'),\varepsilon^2]$ values obtained using these data-

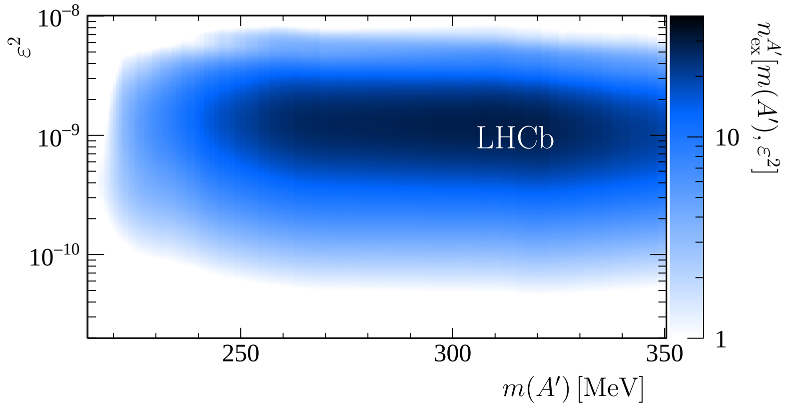


Figure 5-3: Expected reconstructed and selected long-lived $A' \to \mu^+\mu^-$ yield (Credit: Ref. [188]).

driven $\epsilon_{\gamma^*}^{A'}[m(A'),\tau(A')]$ values, along with the expected prompt-like $A'\to\mu^+\mu^-$ yields in Fig. 5-1, are shown in Fig. 5-3.

The long-lived $m(A')$ spectrum is also scanned in discrete steps of $\sigma[m(\mu^+\mu^-)]/2$ looking for $A'\to\mu^+\mu^-$ contributions [294]; however, discrete steps in $\tau(A')$ are also considered here. Binned extended maximum likelihood fits are performed to the three-dimensional feature space of $m(\mu^+\mu^-)$, t, and the consistency of the decay topology as quantified in the decay-fit $\chi^2_{\rm DF}$, which has three degrees of freedom. The photon-conversion contribution is derived in each $[m(\mu^+\mu^-), t, \chi^2_{\rm DF}]$ bin from the number of dimuon candidates that are rejected by the conversion criterion. Both the b-hadron and $K^0_{\rm S}$ contributions are modeled in each $[t, \chi^2_{\rm DF}]$ bin by second-order polynomials of the energy released in the decay, $\sqrt{m(\mu^+\mu^-)^2 - 4m(\mu)^2}$. These contributions are validated using the following large control data samples: candidates that fail the b-hadron suppression requirements; and candidates that fail, but nearly satisfy, the stringent muon-identification requirements. The profile likelihood is used to obtain the p-values and confidence intervals on $n_{\rm ob}^{A'}[m(A'),\tau(A')]$. No significant excess is observed in the long-lived $A'\to\mu^+\mu^-$ search (the three-dimensional data distribution and the background-only pull distributions are provided in Ref. [294]).

Since the relationship between $\tau(A')$ and ε^2 is known at each mass [262], the upper limits on $n_{\rm ob}^{A'}[m(A'),\tau(A')]$ are easily translated into limits on $n_{\rm ob}^{A'}[m(A'),\varepsilon^2]$. Regions of the $[m(A'),\varepsilon^2]$ parameter space where the upper limit on $n_{\rm ob}^{A'}[m(A'),\varepsilon^2]$ is less

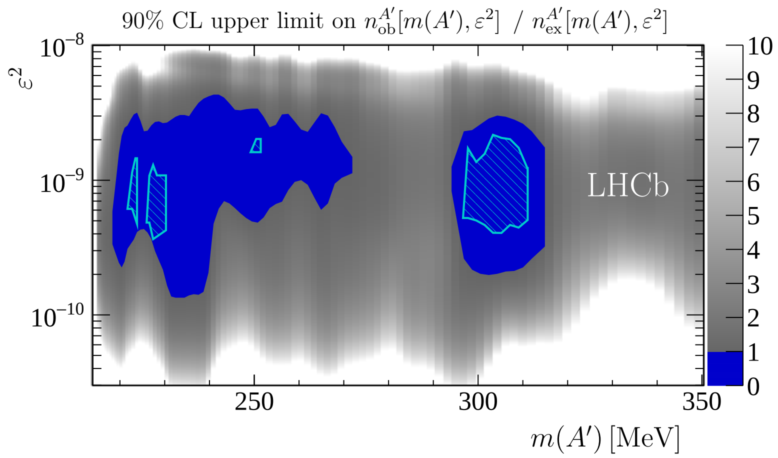


Figure 5-4: Ratio of the observed upper limit on $n_{\mathrm{ob}}^{A'}[m(A'),\varepsilon^2]$ at 90% CL to the expected dark-photon yield, $n_{\mathrm{ex}}^{A'}[m(A'),\varepsilon^2]$, where regions less than unity are excluded. The only constraints in this region are from (hashed) the previous LHCb search [269] (Credit: Ref. [188]).

than $n_{\mathrm{ex}}^{A'}[m(A'),\varepsilon^2]$ are excluded at 90% CL. Figure 5-4 shows that sizable regions of $[m(A'),\varepsilon^2]$ parameter space are excluded, which are much larger than those excluded by LHCb in Ref. [269]. Furthermore, most of the parameter space shown in Fig. 5-4 would have been accessible if the data sample was roughly three times larger. The expected number of recorded $A'\to\mu^+\mu^-$ decays should increase by a factor $\mathcal{O}(100)$ in the data sample to be collected in Run 3 by the upgraded LHCb detector.

5.3 Non-Minimal A'

Searches for dark photons can provide serendipitous discovery potential for other types of particles, especially vector particles that share the same production mechanisms as the minimal dark photon [20]. In order to cover a wide variety of these other models, the searches have to be modified. To ensure that the detector response is sufficiently model independent in the kinematic regions where results are reported, the fiducial regions used for each non-minimal A' searches are defined in Table 5.1. The requirements placed on the momenta, p, and transverse momenta, p_{T}, of the muons make them sufficiently energetic to be selected by the trigger, but not so energetic that their charges cannot be determined. The main motivation for defining the maximum charged-particle multiplicity is the requirement for 900 hits in the scintillating-pad detector in the hardware trigger stage. Only events with at least one reconstructed proton-proton primary vertex (PV) are used in the analysis, which requires that at least five charged prompt-like particles, including the muons of the X decay if this is prompt-like, are produced in the same collision as the X boson. An upper limit is also placed on the number of charged particles produced in the collision, since the detector response depends on the charged-particle multiplicity. The dimuon opening angle is required to be $\alpha(\mu^+\mu^-) > 1\,(3)\,\mathrm{mrad}$ in the searches for prompt-like (displaced) $X \to \mu^+\mu^-$ decays to ensure that the reconstruction efficiency factorizes into the product of the two individual muon efficiencies, which subsequently leads to an upper limit on $p_{\mathrm{T}}(X)$ to remove regions where the $\alpha(\mu^+\mu^-)$ requirement is rarely satisfied. The $X + b$ analysis is performed using jets clustered with the anti-k_{T} algorithm [291] using a distance parameter $R = 0.5$. The jets are required to have $20 < p_{\mathrm{T}}(\mathrm{jet}) < 100\,\mathrm{GeV}$ and a pseudorapidity in the range $2.2 < \eta(\mathrm{jet}) < 4.2$ so that the b-tagging efficiency is nearly uniform within the fiducial region. Finally, the displaced $X \to \mu^+\mu^-$ secondary vertex (SV) is required to be transversely displaced from the PV in the range $12 < \rho_{\mathrm{T}} < 30\,\mathrm{mm}$, which results in minimal dependence on the SV location distribution. For example, this requirement leads to the efficiency being nearly independent of the X lifetime, $\tau(X)$; however, the probability that the

Table 5.1: Fiducial regions of the searches for prompt-like and displaced $X \to \mu^+\mu^-$ decays.

All searches	$p_\mathrm{T}(\mu) > 0.5\,\mathrm{GeV}$ $10 < p(\mu) < 1000\,\mathrm{GeV}$ $2 < \eta(\mu) < 4.5$ $\sqrt{p_\mathrm{T}(\mu^+)p_\mathrm{T}(\mu^-)} > 1\,\mathrm{GeV}$ $5 \leq n_\mathrm{charged}(2 < \eta < 4.5, p > 5\,\mathrm{GeV}) < 100$ (from same PV as X)
Prompt-like $X \to \mu^+\mu^-$ decays	$1 < p_\mathrm{T}(X) < 50\,\mathrm{GeV}$ X decay time $< 0.1\,\mathrm{ps}$ $\alpha(\mu^+\mu^-) > 1\,\mathrm{mrad}$ $20 < p_\mathrm{T}(b\text{-jet}) < 100\,\mathrm{GeV},\ 2.2 < \eta(b\text{-jet}) < 4.2$ ($X + b$ only)
Displaced $X \to \mu^+\mu^-$ decays	$2 < p_\mathrm{T}(X) < 10\,\mathrm{GeV}$ $2 < \eta(X) < 4.5$ $\alpha(\mu^+\mu^-) > 3\,\mathrm{mrad}$ $12 < \rho_\mathrm{T}(X) < 30\,\mathrm{mm}$

X boson decays in this region is strongly dependent on $\tau(X)$.

5.3.1 Selection

The selection criteria are largely applied online in the trigger and most are the same as those used in the minimal dark-photon search [272]. The prompt-like dimuon sample selected by the trigger described in Sec. 5.1 predominantly consists of genuine prompt dimuon pairs. The only selection criteria applied offline for the inclusive prompt-like $X \to \mu^+\mu^-$ search, $p_\mathrm{T}(X) < 50\,\mathrm{GeV}$ and $\alpha(\mu^+\mu^-) > 1\,\mathrm{mrad}$, are included in the definition of the fiducial region. In addition to these, the search for a prompt-like X boson produced in association with a beauty quark requires at least one b-tagged jet with $p_\mathrm{T}(\mathrm{jet}) > 20\,\mathrm{GeV}$ and $2.2 < \eta(\mathrm{jet}) < 4.2$. The jets are formed by clustering charged and neutral particle-flow candidates [293] using the anti-k_T clustering algorithm as implemented in FastJet [292]. The b-tagging requires an SV in the jet that satisfies the criteria given in Ref. [303]. Figure 5-5 shows the $m(\mu^+\mu^-)$ distributions of both prompt-like data samples in bins of width $\sigma[m(\mu^+\mu^-)]/2$, where $\sigma[m(\mu^+\mu^-)]$ denotes the dimuon invariant-mass resolution which varies from $0.6\,\mathrm{MeV}$ near threshold to $0.6\,\mathrm{GeV}$ at $m(\mu^+\mu^-) = 60\,\mathrm{GeV}$.

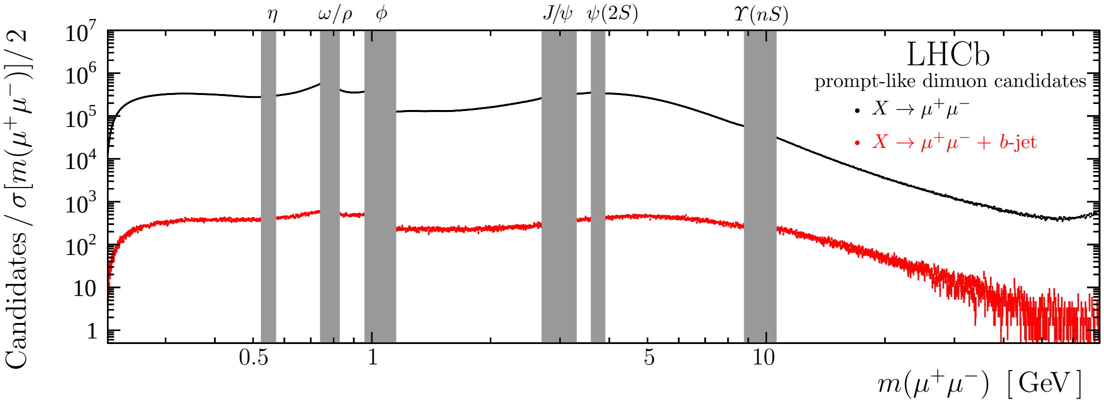


Figure 5-5: Prompt-like dimuon mass spectra showing the (black) inclusive and (red) $X + b$ samples. The grey boxes show the regions vetoed due to large contributions from QCD resonances (Credit: Ref. [189]).

In the searches for displaced $X \to \mu^+\mu^-$ decays, contamination from prompt particles is negligible due to a stringent criterion applied in the trigger that requires muons to be inconsistent with originating from any PV. Furthermore, the fiducial region requires a transverse displacement from the PV of $12 < \rho_T < 30\,\text{mm}$, which is applied offline in both searches for displaced $X \to \mu^+\mu^-$ decays and highly suppresses the background from b-hadron decay chains that produce two muons. Therefore, the dominant background contributions are due to material interactions in the VELO, e.g. photons that convert into $\mu^+\mu^-$ pairs, and from $K^0_S \to \pi^+\pi^-$ decays, where both pions are misidentified as muons, which is the dominant background in the search for $K^0_S \to \mu^+\mu^-$ decays [304]. A p-value is assigned to the material-interaction hypothesis for each displaced $X \to \mu^+\mu^-$ candidate using properties of the SV and muon tracks, along with a high-precision three-dimensional material map produced from a data sample of secondary hadronic interactions [299]. The same mass-dependent requirement used in Ref. [272] is applied to the p-values in this analysis, which highly suppresses the material-interaction background. Figure 5-6 shows the $m(\mu^+\mu^-)$ distributions of both displaced-dimuon data samples.

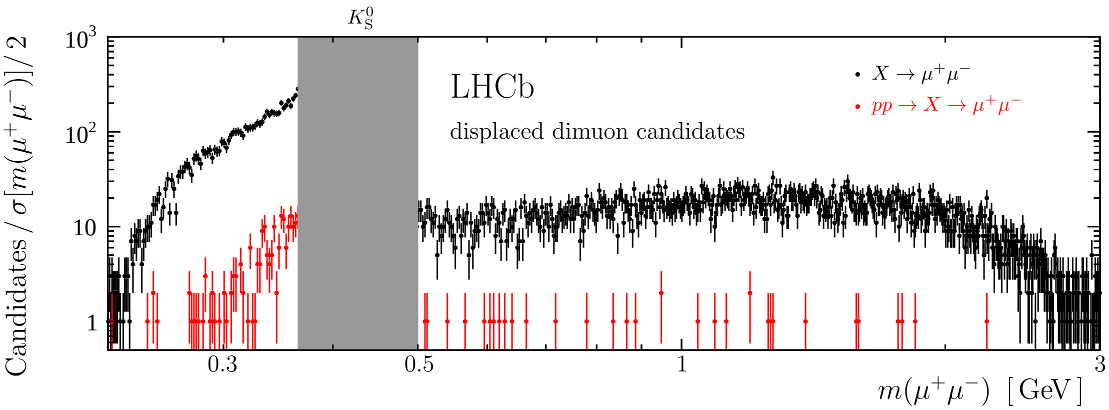


Figure 5-6: Displaced dimuon mass spectra showing the (black) inclusive and (red) promptly produced samples. The grey box shows the region vetoed due to the large doubly misidentified K^0_S background, whose low-mass tail extends into the search region (Credit: Ref. [189]).

5.3.2 Signal searches

The signal-search strategies and methods employed are similar to those used in Ref. [272]. The dimuon mass spectra are scanned in around 6000 steps of about $\sigma[m(\mu^+\mu^-)]/2$ searching for $X \to \mu^+\mu^-$ contributions. For $m(X) < 20\,\text{GeV}$, the data are binned in $p_T(X)$ and each p_T bin is searched independently for each $m(X)$ hypothesis; whereas at higher masses, p_T bins are not necessary since both the resolution and efficiency are nearly independent of $p_T(X)$. All searches use the profile likelihood method to determine the local p-values and the confidence intervals on the signal yields. The trial factors are obtained using pseudoexperiments in each search. The confidence intervals are defined using the *bounded likelihood* approach [305], which involves taking $\Delta \log \mathcal{L}$ relative to zero signal, rather than the best-fit value, if the best-fit signal value is negative. This approach enforces that only physical (nonnegative) upper limits are placed on the signal yields, and prevents defining exclusion regions that are much better than the experimental sensitivity in cases where a large deficit in the background yield is observed. The signal $m(\mu^+\mu^-)$ distributions are well modeled by a Gaussian function, whose resolution is determined with 10% precision using a combination of simulated $X \to \mu^+\mu^-$ decays and the observed p_T-dependent widths of the large known resonance peaks present in the data. The mass-resolution uncertainty is included in the profile likelihood.

The fit strategy used in the prompt-like $X \to \mu^+\mu^-$ searches below 20 GeV, which is the same as in Refs. [269, 272], was first introduced in Ref. [295]. At each $m(X)$ hypothesis, a binned extended maximum-likelihood fit is performed in a $\pm 12.5\,\sigma[m(\mu^+\mu^-)]$ window around the $m(X)$ value. Near the dimuon threshold, the energy released in the decay, $Q = \sqrt{m(\mu^+\mu^-)^2 - 4m(\mu)^2}$, is used instead of the mass because it is easier to model. The background model for each fit window takes as input a large set of potential components, then the data-driven model-selection process of Ref. [295] is performed, whose uncertainty is included in the profile likelihood following Ref. [296]. Specifically, the method labeled *aic-o* in Ref. [295] is used, where the log-likelihood of each background model is penalized for its complexity (number of parameters). The confidence intervals are obtained from the profile likelihoods, including the penalty terms, where the model index is treated as a discrete nuisance parameter, as originally proposed in Ref. [296]. In the $X + b$ search there are not many candidates near the dimuon threshold. Therefore, just in this region, the counting-experiment-based method of Ref. [306] is used, which is also used in the searches for displaced $X \to \mu^+\mu^-$ decays and described in detail below.

In this analysis, the set of possible background components is the same as in Ref. [272] and includes all Legendre modes up to tenth order at every $m(X)$. Additionally, dedicated background components are included for sizable narrow SM resonance contributions. The use of 11 Legendre modes adequately describes every doubly misidentified peaking background that contributes at a significant level; therefore, these do not require dedicated background components. In mass regions where such complexity is not required, the data-driven model-selection procedure reduces the complexity, which increases the sensitivity to a potential signal contribution. Therefore, the impact of the background-model uncertainty on the size of the confidence intervals is mass dependent, though on average it is about 30%. As in Ref. [295], all fit regions are transformed onto the interval $[-1, 1]$, where the $m(X)$ value is mapped to zero. After such a transformation, the signal model is (approximately) an even function; therefore, odd Legendre modes are orthogonal to the signal component, which means that the presence of odd modes has minimal impact on the variance of the observed signal yield. In the prompt-like fits, all odd Legendre modes up to ninth order are included in every background model, while even modes must be selected for inclusion in each fit by the data-driven method of Ref. [295].

Regions in the mass spectrum with large SM resonance contributions are vetoed in the searches for prompt-like $X \to \mu^+\mu^-$ decays. Furthermore, the region near the η' meson is treated uniquely. Since it is not possible to distinguish between $X \to \mu^+\mu^-$ and possible $\eta' \to \mu^+\mu^-$ contributions at $m(\eta')$, the p-values near this mass are ignored. The small observed excess at $m(\eta')$ is simply absorbed into the signal yield when setting the limits, which is conservative in that the $\eta' \to \mu^+\mu^-$ contribution weakens the constraints on $X \to \mu^+\mu^-$ decays.

Figure 5-7 shows the signed local significances for all $m(X)$ below 20 GeV for both prompt-like $X \to \mu^+\mu^-$ searches. The largest local excess in the inclusive search in this mass region is 3.7σ at 349 MeV in the $3 < p_{\mathrm{T}}(X) < 5\,\mathrm{GeV}$ bin; however, its neighboring p_{T} bin at this mass has a small deficit and the global significance is only $\approx 1\sigma$. Similarly, the largest local excess in the $X + b$ search below 20 GeV is 3.1σ at 2424 MeV in the $10 < p_{\mathrm{T}}(X) < 20\,\mathrm{GeV}$ bin, though again, the neighboring p_{T} bins both have deficits at the same mass, and the global significance is below 1σ. Therefore, no significant excess is found in either prompt-like spectrum for $m(X) < 20\,\mathrm{GeV}$.

In the $20 < m(X) < 60\,\mathrm{GeV}$ region, the background is nearly monotonic, which permits the use of a simplified fit strategy. The entire $12 < m(\mu^+\mu^-) < 80\,\mathrm{GeV}$ region is fitted when considering all $m(X)$ values above 20 GeV. The background model is comprised of three falling power-law terms and an eighth-order polynomial that collectively describe the Drell–Yan, heavy-flavor, and misidentified-background contributions, along with a rising power-law term to describe the low-mass tail of the Z boson, where all parameters are free to vary. This background model is validated by studying simulated Drell–Yan dimuon production, same-sign dimuon data which predominantly consists of heavy-flavor and misidentification backgrounds, and candidates in the data sample itself above the search region. Unlike at lower masses, nonnegligible widths are considered. At each $m(X)$, a scan is performed covering the

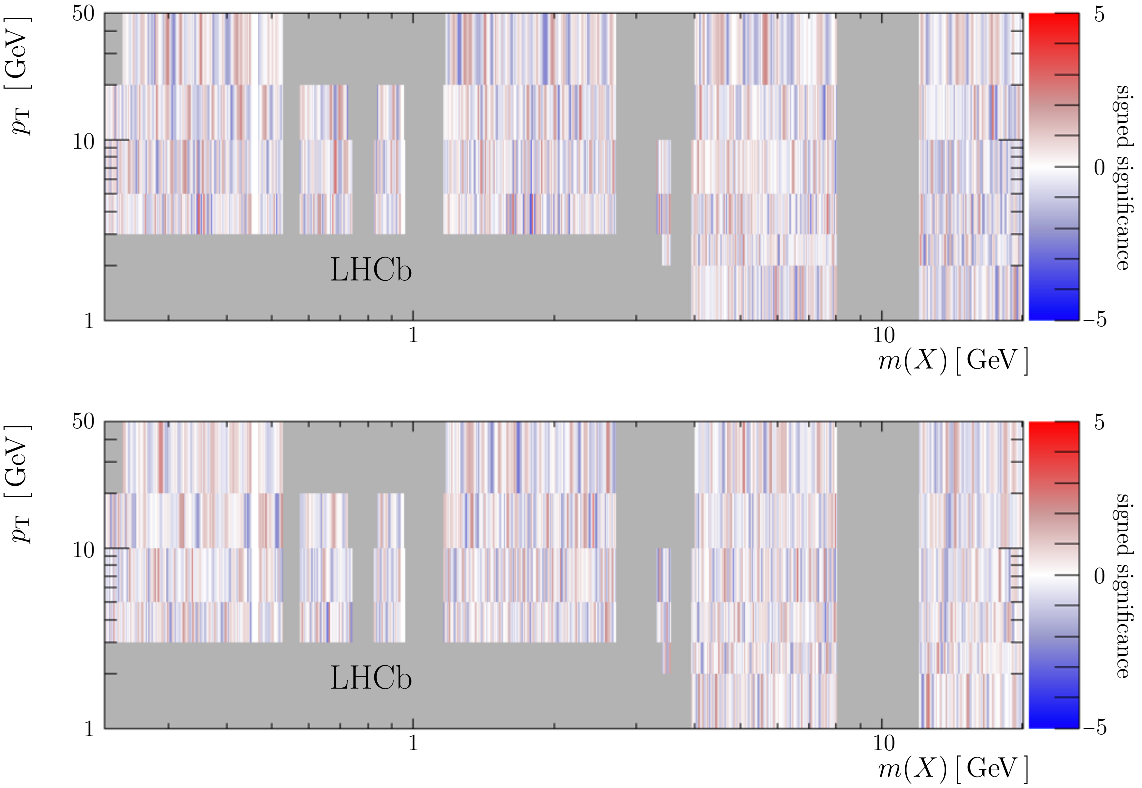


Figure 5-7: Signed local significances in the $m(X) < 20\,\mathrm{GeV}$ region for the (top) inclusive and (bottom) associated beauty prompt-like $X \to \mu^+\mu^-$ searches. If the best-fit signal-yield estimator is negative, the signed significance is negative and *vice versa*. The grey regions are excluded either due to a nearby large QCD resonance contribution, or because the overlap of the bin with the fiducial region in Table 5.1 is small (Credit: Ref. [189]).

range $0 \leq \Gamma(X) \leq 3\,\mathrm{GeV}$. The signals are modelled by a Gaussian resolution function convolved with the modulus of a Breit–Wigner function.

Figure 5-8 shows the signed local significances for the $m(X) > 20\,\mathrm{GeV}$ region for both prompt-like $X \to \mu^+\mu^-$ searches. The largest local excess in the inclusive search in this mass region is 3.2σ at $m(X) = 36\,\mathrm{GeV}$ for $\Gamma(X) = 1.5\,\mathrm{GeV}$, which corresponds to a global p-value of about 11% (considering only the $m(X) > 20\,\mathrm{GeV}$ mass region). In the $X + b$ search, no local significance exceeds $\approx 2\sigma$ in this mass region. Therefore, no significant excess is found in either prompt-like spectrum for $m(X) > 20\,\mathrm{GeV}$.

Motivated by the possible excess seen by CMS [24] in $X + b\bar{b}$ events, a dedicated search for a resonance with $27 < m(X) < 30\,\mathrm{GeV}$ and $0.5 < \Gamma(X) < 3.0\,\mathrm{GeV}$ is performed in the subset of the $X + b$ data sample that contains at least two b-tagged

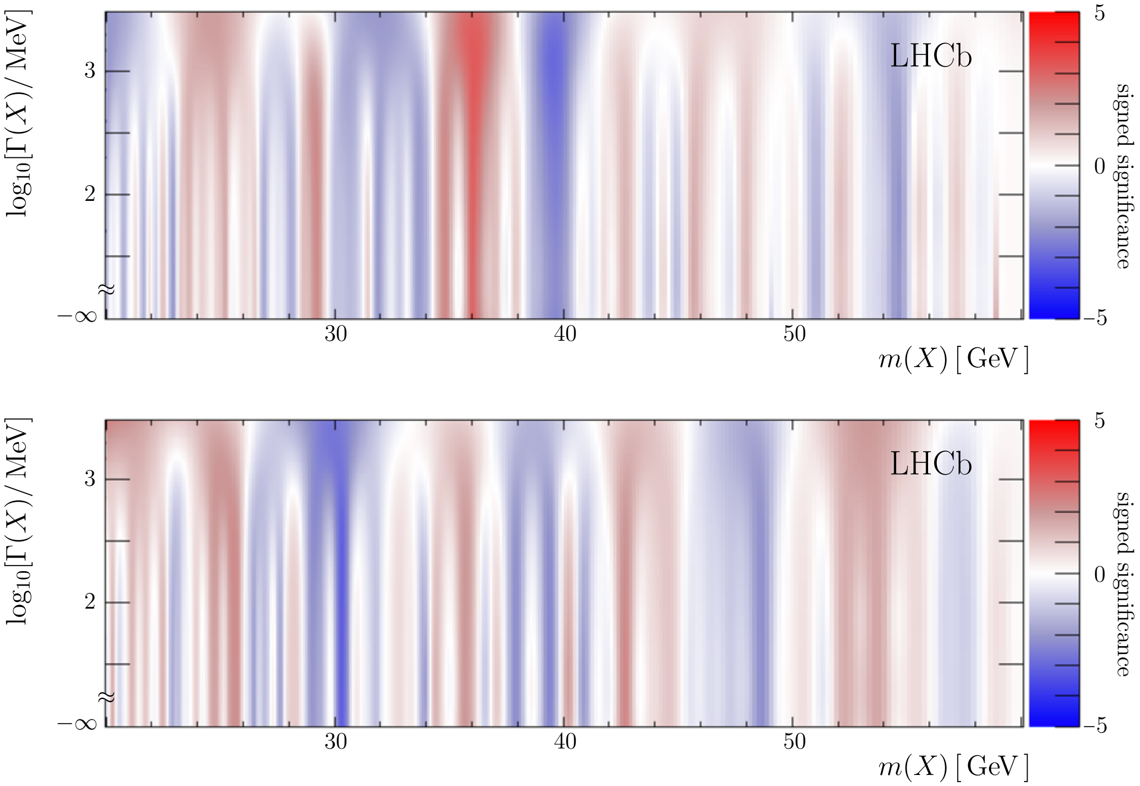


Figure 5-8: Signed local significances in the $m(X) > 20\,\mathrm{GeV}$ region for the (top) inclusive and (bottom) associated beauty prompt-like $X \to \mu^+\mu^-$ searches. The lower limit on the vertical axis of $\log_{10}[\Gamma(X)/\,\mathrm{MeV}] = -\infty$ corresponds to $\Gamma(X) = 0$ (Credit: Ref. [189]).

jets. The mass spectrum in the range 20–40 GeV is fitted using a model consisting of a second-order polynomial background and a signal whose mass and width are free to vary within the $m(X)$ and $\Gamma(X)$ ranges specified above. Figure 5-9 shows the result of this fit. The best-fit signal yield is negative in the region considered; therefore, no evidence for a signal is observed. Using the efficiency and luminosity from Sec. 5.3.3, and their associated uncertainties, the upper limits on the $X(\mu^+\mu^-) + b\bar{b}$ cross section in the $m(X)$ and $\Gamma(X)$ regions considered are no larger than $15\,\mathrm{fb} \times \sqrt{\Gamma(X)/\,\mathrm{GeV}}$.

The fit strategy used in the searches for displaced $X \to \mu^+\mu^-$ decays below the K^0_S mass is also the same as in Refs. [269, 272]. Binned extended maximum-likelihood fits are performed to the Q spectrum in each p_T bin. The region near the K^0_S mass is vetoed to avoid the sizable background from doubly misidentified $K^0_\mathrm{S} \to \pi^+\pi^-$ decays. The expected photon-conversion contribution is derived from a sample of

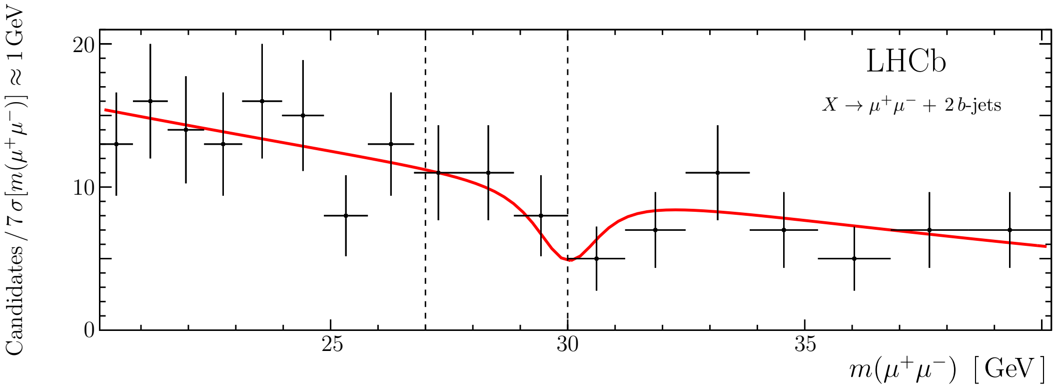


Figure 5-9: Fit to the $m(\mu^+\mu^-)$ spectrum in events with at least two b-tagged jets. The $27 < m(X) < 30\,\text{GeV}$ search region is marked by the vertical dashed lines (Credit: Ref. [189]).

candidates that are consistent with a photon originating from a PV. Two large control samples are used to develop and validate the modeling of the K_S^0 and remaining material-interaction contributions: dimuon candidates that fail, but nearly satisfy, the stringent muon-identification criteria; and a sample of dimuon candidates that is rejected by the material-interaction criterion. Both contributions are well modeled by second-order polynomials in Q below the K_S^0 veto region. The material-interaction contribution, apart from the dedicated photon-conversion component, is not needed in the search that requires a decay topology consistent with an X boson originating from a PV.

The fit strategy used in the searches for displaced $X \to \mu^+\mu^-$ decays above the K_S^0 veto region, specifically, in the $0.5 < m(X) < 3.0\,\text{GeV}$ mass range, is the same as used in the LHCb search for hidden-sector bosons produced in $B^0 \to K^{(*)}X(\mu^+\mu^-)$ decays [307, 308]. This strategy was first introduced in Ref. [306]. Since no sharp features are expected in the background in this region, and due to the small bin occupancies, the background is estimated by interpolating the yields in the sidebands starting at $\pm 3\sigma[m(\mu^+\mu^-)]$ from $m(X)$. The statistical test at each mass is based on the profile likelihood ratio of Poisson-process hypotheses with and without a signal contribution. The uncertainty on the background interpolation is modeled by a Gaussian term in the likelihood.

Figure 5-10 shows the signed local significances for both searches for displaced

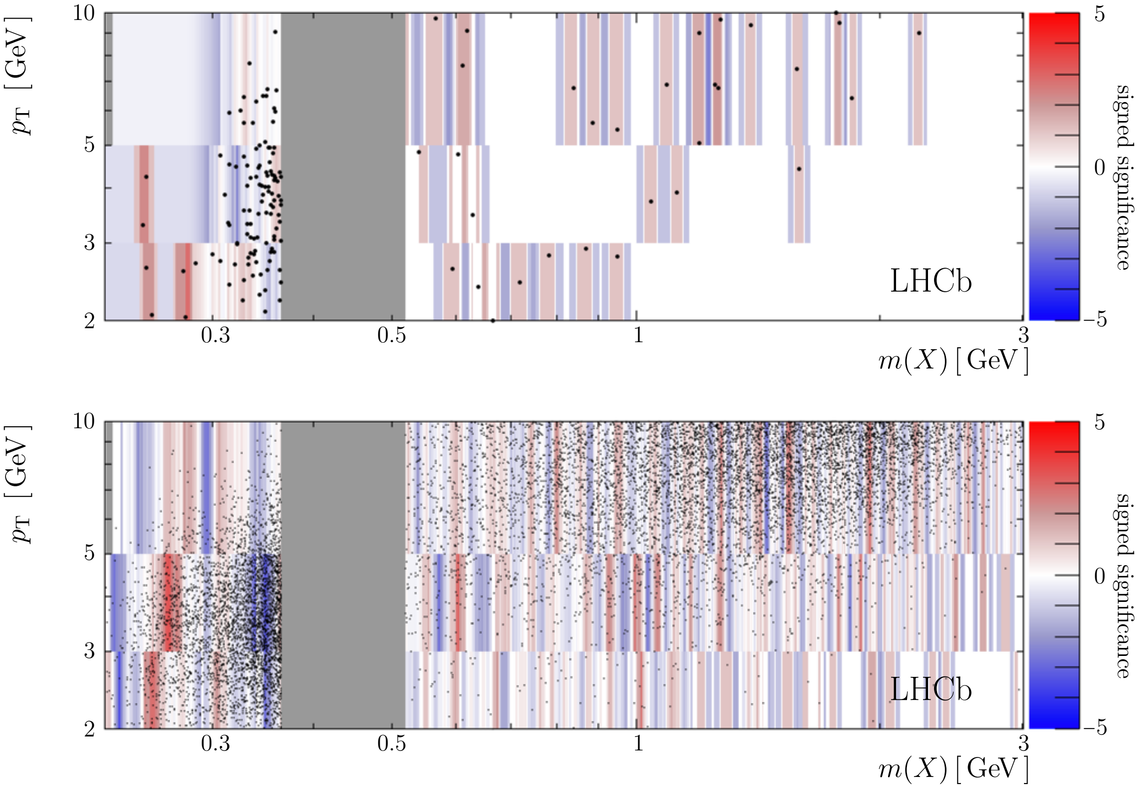


Figure 5-10: Signed local significances for the (top) promptly produced and (bottom) inclusive searches for displaced $X \to \mu^+\mu^-$ decays. The black points show the individual candidates (Credit: Ref. [189]).

$X \to \mu^+\mu^-$ decays. The largest local excess in the search for a promptly produced long-lived X boson is 2.8σ, which occurs at 280 MeV in the $2 < p_T(X) < 3$ GeV bin. The largest local excess in the inclusive search for displaced $X \to \mu^+\mu^-$ decays is 3.1σ at 604 MeV in the $3 < p_T(X) < 5$ GeV bin. Both of these correspond to global excesses below 1σ; therefore, no significant excess is found in either search for displaced $X \to \mu^+\mu^-$ decays.

5.3.3 Efficiency and luminosity

The $X \to \mu^+\mu^-$ yields are corrected for detection efficiency, which is determined as the product of the trigger, reconstruction, and selection efficiencies. The trigger efficiency is measured as a function of $\sqrt{p_T(\mu^+)p_T(\mu^-)}$ using a displaced J/ψ calibration sample. Events selected by the hardware trigger independently of the J/ψ candidate, e.g. due to the presence of a high-p_{T} hadron, are used to determine the trigger efficiency directly from the data. The muon reconstruction efficiency is obtained from simulation in bins of $[p(\mu), \eta(\mu)]$. Scale factors that correct for discrepancies between the data and simulation are determined using a data-driven tag-and-probe approach on an independent sample of $J/\psi \to \mu^+\mu^-$ decays [309]. The contribution to the selection efficiency from the muon-identification performance is measured in bins of $[p_{\mathrm{T}}(\mu), \eta(\mu)]$ using a highly pure calibration sample of $J/\psi \to \mu^+\mu^-$ decays. Finally, the contributions from the vertex-quality and prompt-like muon criteria are determined from simulation, and validated using a calibration sample of prompt QCD resonance decays to the $\mu^+\mu^-$ final state.

The uncertainty due to the methods used to determine each of these components of the total efficiency is assessed by repeating the data-based efficiency studies on simulated events, where the difference between the true and efficiency-corrected yields in kinematic bins is used to determine the systematic uncertainty. These uncertainties are in the 2–5% range, depending on X-boson kinematics. Additional uncertainties arise due to the unknown production mechanisms of the X bosons. The muon reconstruction and identification efficiencies depend on the charged-particle multiplicity. The corresponding systematic uncertainty is determined to be 5%, which covers both minimal and maximal charged-particle multiplicities defined in Table 5.1 at the 2σ level. The unknown kinematic distributions in both p_{T} and η within the wide p_{T} bins used in the analysis lead to sizable uncertainties. The variation in the efficiencies across the kinematic regions allowed in each bin are used to determine bin-dependent uncertainties that vary from 10 to 30%.

The $X + b$ analysis uses the SV-based b-tagging method described in detail in Ref. [303], though without placing any criteria on the boosted decision tree algorithms. The b-tagging efficiency is estimated to be $(65 \pm 7)\%$, where the uncertainty covers both the variation of the b-tagging efficiency across the b-jet fiducial region and possible data-simulation discrepancies. An additional uncertainty arises since the efficiency for a b-tagged jet in the fiducial region to be reconstructed with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ depends on the unknown underlying jet p_{T} spectrum. The detector response to jets is studied using the p_{T}-balance distribution of $p_{\mathrm{T}}(\mathrm{jet})/p_{\mathrm{T}}(Z)$ in nearly back-to-back Z-boson+jet events using the same data-driven technique as in Ref. [293]. Based on this study, and considering jet p_{T} spectra as soft as QCD di-b-jet production and hard enough to result in negligible inefficiency, this efficiency is estimated to be $(90 \pm 5)\%$.

The searches for displaced $X \to \mu^+\mu^-$ decays must also account for effects that arise due to the displacement of the SV from the PV. The relative efficiency of displaced compared to prompt-like dimuon production is obtained as a function of $m(X)$ and $p_{\mathrm{T}}(X)$ by resampling prompt $X \to \mu^+\mu^-$ candidates as displaced $X \to \mu^+\mu^-$ decays, where all displacement-dependent properties are recalculated based on the resampled SV locations. The high-precision material map produced in Ref. [299] forms the basis of the material-interaction criterion applied in the selection. This map is used to determine where each muon would hit active sensors, and thus, have recorded hits in the VELO. The resolution on the vertex location and other displacement-dependent properties varies strongly with the location of the first VELO hit on each muon track, though this dependence is largely geometric, making rescaling the resolution of prompt tracks straightforward. This approach is validated using simulation, where prompt $X \to \mu^+\mu^-$ decays are used to predict the properties of long-lived $X \to \mu^+\mu^-$ decays; these predictions are found to agree within 2% with the actual values. The efficiencies at both short and long distances, which are driven by the muon displacement criterion and the minimum number of VELO hits required to form a track, respectively, are well described. The dominant uncertainty, which arises due to limited knowledge of how radiation damage has affected the VELO performance, is estimated to be 5% by rerunning the resampling method under different radiation-damage hypotheses.

The efficiency of the material-interaction criterion is validated separately using two control samples. The predicted efficiency for an X boson with the same mass and lifetime as the K^0_{S} meson is compared to the efficiency observed in a control sample of K^0_{S} decays. The predicted and observed efficiencies agree to 1%. Additionally, in Ref. [299] the expected performance of the material-interaction criterion was shown to agree with the performance observed in a control sample of photon conversions to the $\mathcal{O}(10^{-4})$ level. Finally, the distribution of the SV locations is unknown, which leads to a 10% uncertainty in the efficiency determined by comparing the efficiency of an X boson that rarely survives long enough to enter the decay fiducial region to an extremely long-lived X boson.

Most of the data used in this analysis is from data-taking periods that do not yet have fully calibrated luminosities. Therefore, the efficiency-corrected yield of $Z/\gamma^* \to \mu^+\mu^-$ decays observed in the data sample—and the corresponding high-precision LHCb cross-section measurement made using 2015 data [310]—are used to infer the luminosity. A small correction factor is obtained from Pythia 8 to account for the different fiducial regions. This luminosity determination is validated by also determining the $\Upsilon(1S)$ differential cross section from this data sample and comparing the results to those published by LHCb using the 2015 data sample [311]. The different fiducial region is again corrected for using a scale factor obtained from Pythia 8. The results are found to agree to $\approx$ 5% in each p_{T} bin, which is assigned as a systematic uncertainty and combined with the 4% luminosity uncertainty from Ref. [310] to obtain the total uncertainty on the luminosity of this data sample. Based on both of these studies, the luminosity is determined to be $5.1 \pm 0.3\,\mathrm{fb}^{-1}$. The minimal dark-photon search [272], which used the same data sample but did not require knowledge of the luminosity, quotes an uncalibrated luminosity value that is 7% larger. The efficiency corrections used to infer the luminosity are highly correlated to those used to correct the observed $X \to \mu^+\mu^-$ yields, which is accounted for when determining the total normalization uncertainties.

5.3.4 Cross-section results

The upper limits on the signal yields obtained in Sec. 5.3.2 are normalized using the efficiencies and luminosity described in Sec. 5.3.3. The systematic uncertainties on the signal yield, efficiency, and luminosity are included in the profile likelihood when determining the cross-section upper limits. These uncertainties are described in detail in Secs. 5.3.2 and 5.3.3, and summarized in Table 5.2. The resulting upper limits at 90% confidence level on $\sigma(X \to \mu^+\mu^-)$ for all searches are shown in Figs. 5-11–5-13,

Table 5.2: Summary of systematic uncertainties. The luminosity and efficiency uncertainties are highly correlated, which is accounted for when obtaining the total uncertainties.

Source	Relative uncertainty
Signal model	5%
Background model	data driven, see Sec. 5.3.2
Trigger, reconstruction, selection	2–5% (bin dependent)
Charged-particle multiplicity	5%
X kinematics	10–30% (bin dependent)
b-jet selection	11% ($X+b$ only)
SV selection	5% (SV-based only)
X SV distribution	10% (SV-based only)
Luminosity	6%
Total	11–30% (bin dependent)

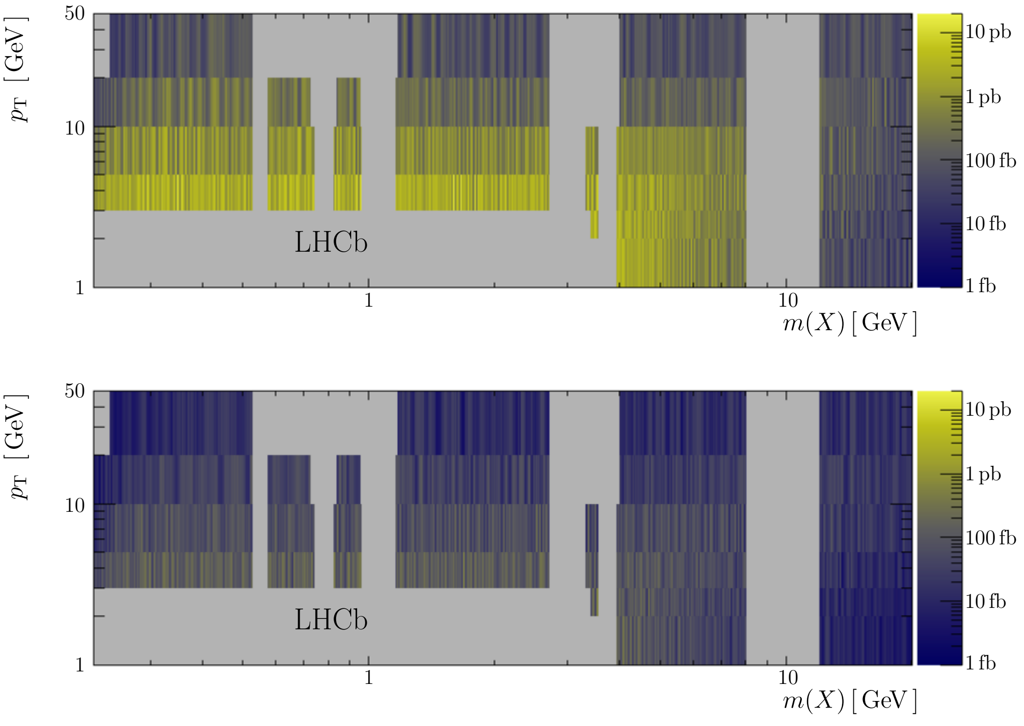


Figure 5-11: Upper limits at 90% confidence level on the cross section $\sigma(X \to \mu^+\mu^-)$ in the $m(X) < 20\,\text{GeV}$ region for the (top) inclusive and (bottom) associated beauty prompt-like $X \to \mu^+\mu^-$ searches (Credit: Ref. [189]).

and provided numerically in Ref. [294].

The model-independent limits in Figs. 5-11–5-12 can be used to place constraints

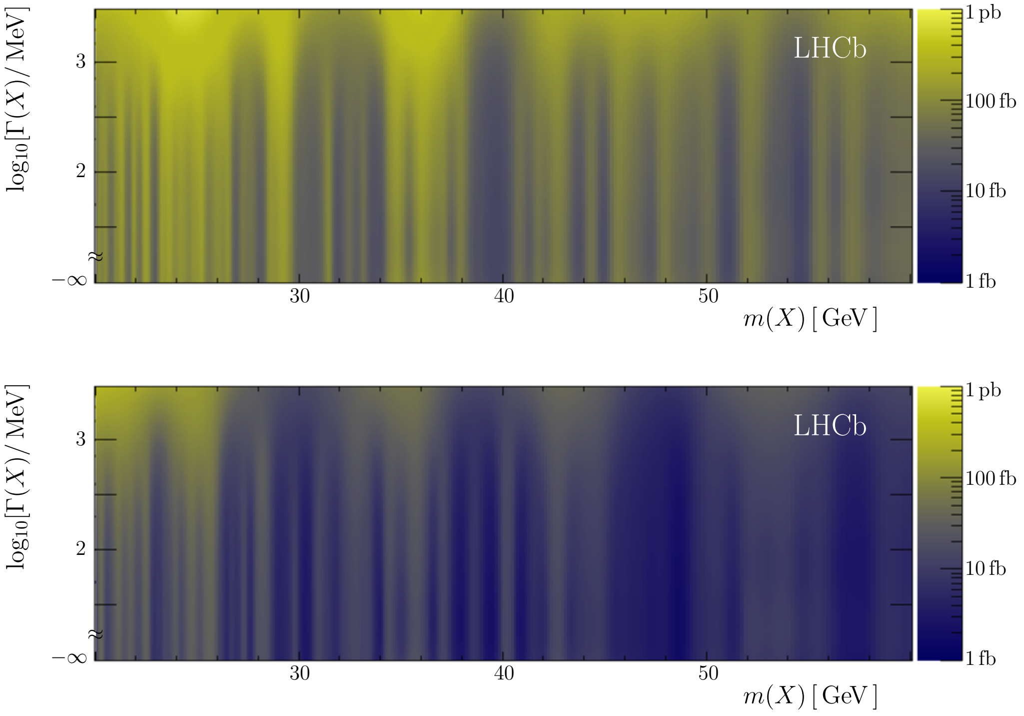


Figure 5-12: Upper limits at 90% confidence level on the cross section $\sigma(X \to \mu^+\mu^-)$ in the $m(X) > 20\,\text{GeV}$ region for the (top) inclusive and (bottom) associated beauty prompt-like $X \to \mu^+\mu^-$ searches (Credit: Ref. [189]).

on any model that would produce a prompt-like low-mass dimuon resonance within the fiducial region of Table 5.1. For example, models where a complex scalar singlet is added to the two-Higgs doublet (2HDM) potential often feature a light pseudoscalar boson that can decay into the dimuon final state; see, e.g. Ref. [21]. References [22, 23] considered the scenario where the pseudoscalar boson acquires all of its couplings to SM fermions through its mixing with the Higgs doublets; the corresponding X–H mixing angle is denoted as θ_H. Figure 5-14 shows that world-leading constraints are placed on θ_H by the prompt-like $\sigma(X \to \mu^+\mu^-)$ limits shown in Figs. 5-11–5-12. Furthermore, assuming the $X + b\bar{b}$ topology produced by this type of model permits direct comparison with the excess seen by CMS in this final state [24]. For this scenario, the $X + b$ limits from Fig. 5-12 are about 20 times lower than the excess observed by CMS.

The limits on displaced $X \to \mu^+\mu^-$ decays in Fig. 5-13 can also be used to place constraints on specific models. One example is HV scenarios that exhibit confine-

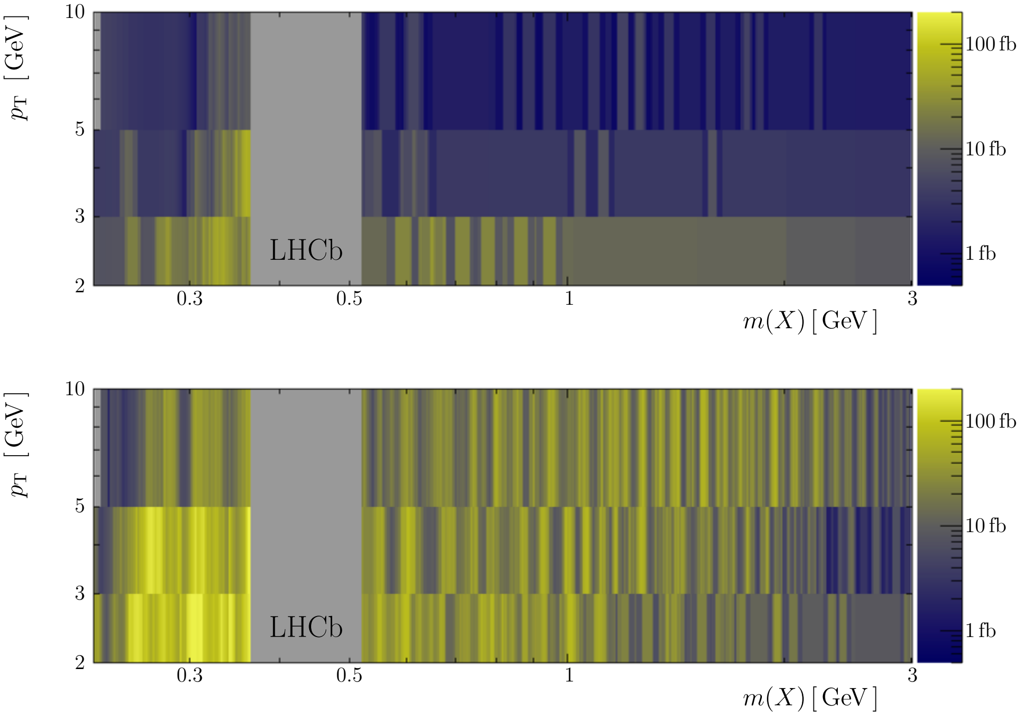


Figure 5-13: Upper limits at 90% confidence level on the cross section $\sigma(X \to \mu^+\mu^-)$ for the (top) promptly produced and (bottom) inclusive searches for displaced $X \to \mu^+\mu^-$ decays (Credit: Ref. [189]).

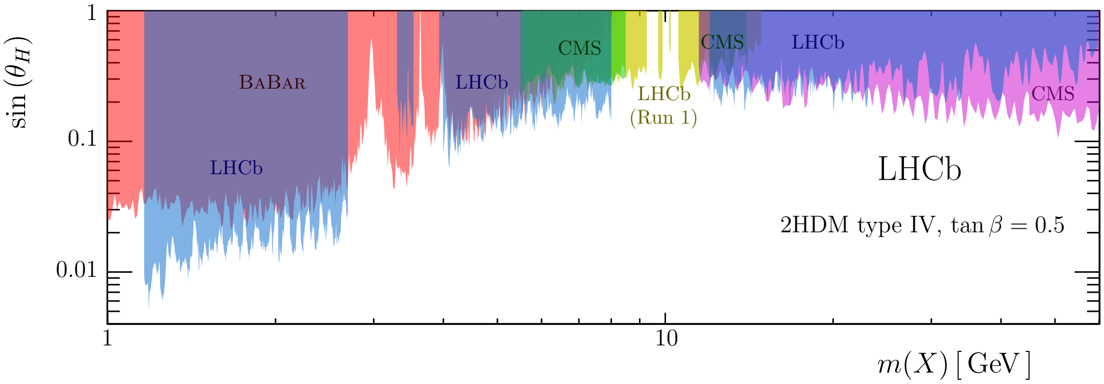


Figure 5-14: Upper limits at 90% confidence level on the X–H mixing angle, θ_H, for the 2HDM scenario discussed in the text (blue) from this analysis compared with existing limits from (red) BaBar [312], (green) CMS Run 1 [313], (magenta) CMS Run 2 [314] and (yellow) LHCb Run 1 [315] (Credit: Ref. [189]).

ment, which result in a large multiplicity of light hidden hadrons from showering processes [25]. These hidden hadrons typically have low p_T and decay displaced from the proton-proton collision. Figure 5-15 shows the limits placed on this type of HV

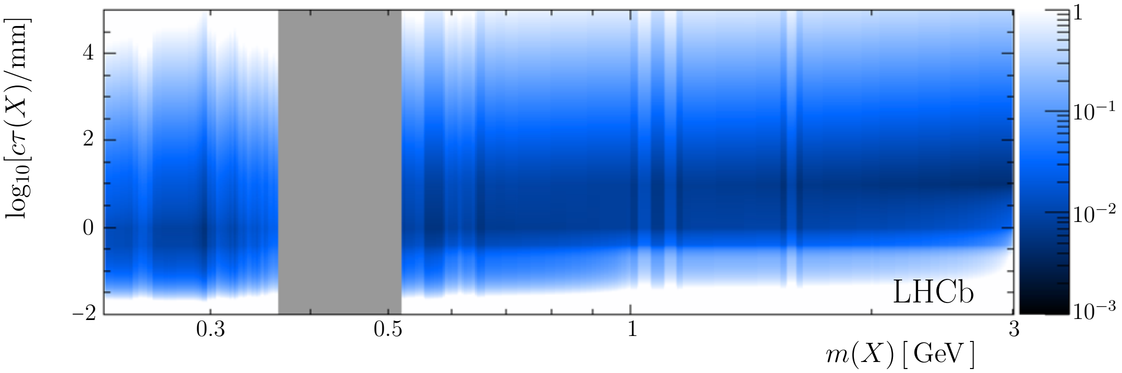


Figure 5-15: Upper limits at 90% confidence level on the γ–Z_{HV} kinetic mixing strength for the HV scenario discussed in the text (Credit: Ref. [189]).

scenario by the search for displaced $X \to \mu^+\mu^-$ decays. These are the most stringent constraints to date. Specifically, constraints are placed on the kinetic-mixing strength between the photon and a heavy HV boson, Z_{HV}, with photon-like couplings. The kinematics of the hidden hadrons depend upon the average HV hadron multiplicity, $\langle N_{\mathrm{HV}}\rangle$, and are largely independent of the model parameter space. In Fig. 5-15 $\langle N_{\mathrm{HV}}\rangle$ is fixed at ≈ 10 for all hidden hadron masses. These are the first results that constrain the kinetic-mixing strength to be less than unity in this mass region.

5.4 Search Summary

Searches are performed for prompt-like and long-lived dark photons, A', decaying into two muons, $A' \to \mu^+\mu^-$, and other low-mass dimuon resonances, X, produced in proton-proton collisions at a center-of-mass energy of 13 TeV. The data samples for A' and X searches were collected with the LHCb detector and correspond to an integrated luminosity of $5.1\,\text{fb}^{-1}$. The three-fold increase in integrated luminosity, improved trigger efficiency during 2017–2018 data taking, and improvements in the analysis result in the searches presented in this chapter achieving much better sensitivity to dark photons than the previous LHCb results [269]. The prompt-like A' search achieves a factor of 5 (2) better sensitivity to ε^2 at low (high) masses than Ref. [269], while the long-lived A' search provides access to much larger regions of $[m(A'), \varepsilon^2]$ parameter space. No evidence for a signal is found in either search, and 90% CL exclusion regions are set on the γ–A' kinetic-mixing strength. The prompt-like A' search is performed from near the dimuon threshold up to 70 GeV, and produces the most stringent constraints on dark photons with $214 < m(A') \lesssim 740\,\text{MeV}$ and $10.6 < m(A') \lesssim 30\,\text{GeV}$. The long-lived A' search is restricted to the mass range $214 < m(A') < 350\,\text{MeV}$, where the data sample potentially has sensitivity, and places world-leading constraints on low-mass dark photons with lifetimes $\mathcal{O}(1)\,\text{ps}$.

The $X \to \mu^+\mu^-$ decays can be either prompt-like or displaced from the proton-proton collision, where in both cases the requirements placed on the event and the assumptions made about the production mechanisms are kept as minimal as possible. Two variations of the prompt-like X search are performed: an inclusive version, and one where the X boson is required to be produced in association with a beauty quark. Two variations are also considered of the search for displaced $X \to \mu^+\mu^-$ decays: an inclusive version, and one where the X boson is required to be produced promptly in the proton-proton collision. The prompt-like X searches explore the mass range from near the dimuon threshold up to 60 GeV, with nonnegligible X widths considered above 20 GeV. The searches for displaced $X \to \mu^+\mu^-$ decays consider masses up to 3 GeV. None of the searches finds evidence for a signal, and 90% confidence-level exclusion limits are placed on the $X \to \mu^+\mu^-$ cross sections, each with minimal model dependence. These results demonstrate the unique sensitivity of the LHCb experiment to low-mass dimuon resonances, even using a data sample collected with a hardware-trigger stage that is highly inefficient for low-mass $A' \to \mu^+\mu^-$ decays. The removal of this hardware-trigger stage in Run 3, along with the planned increase in luminosity, should greatly increase the potential yield of $A' \to \mu^+\mu^-$ and $X \to \mu^+\mu^-$ decays in the low-mass region compared to the 2016–2018 data sample, and therefore, greatly increase the dark-photon discovery potential of the LHCb experiment.

Chapter 6

Conclusion

Searches were performed for prompt-like and long-lived dark photons, A', decaying into two muons and other low-mass dimuon resonances, X. For this task, proton-proton collisions produced at the LHC at a center-of-mass energy of 13 TeV were studied with the LHCb detector. No evidence for a signal is found in any of these searches, and 90% confidence level exclusion regions are set. The prompt-like A' search achieves a factor of 5 (2) better sensitivity to ε^2 at low (high) masses than in previous studies and sets the world's most stringent constraints on dark photons with $214 < m(A') \lesssim 740\,\mathrm{MeV}$ and $10.6 < m(A') \lesssim 30\,\mathrm{GeV}$. The long-lived A' search places world-leading on low-mass dark photons with lifetimes of $\mathcal{O}(1)$ ps. Furthermore, exclusion limits are placed on prompt-like and long-lived low-mass dimuon resonances with minimal assumptions about production mechanisms. These limits can be translated to constraints on various theories like two-Higgs-doublet and hidden-valley models. By excluding previously viable models, theorists and experimentalists can concentrate on remaining theories, thereby making progress on the question of the nature of dark matter.

The luminosity upgrade of the LHC accelerator and the improvements made to the LHCb detector, including the removal of its hardware trigger, will result in more powerful searches for dark photons decaying into two muons once there is data collected from 2021 to 2024. Searches for $A' \to e^+e^-$ inclusive decays and those produced via the $\mathrm{D}^*(2007)^0 \to \mathrm{D}^0 A'$ process will further restrict the space of viable models and will be able to probe dark photon masses as low as twice the electron mass.

In addition, this thesis proposed various machine and deep learning techniques. They can help increase the sensitivity of bump hunts, enhance the number and accuracy of Monte Carlo simulations, improve vertex reconstruction, and explain the use of machine learning for goodness of fit. Together these tools can help pave the way to a new, more powerful generation of analyses.

Bibliography

1. Wilson, E. O. *Consilience: The Unity of Knowledge* (2014).
2. Thomson, M. *Modern particle physics* ISBN: 978-1-107-03426-6 (Cambridge University Press, New York, 2013).
3. Griffiths, D. *Introduction to elementary particles* ISBN: 978-3-527-40601-2 (2008).
4. Gaillard, M. K., Grannis, P. D. & Sciulli, F. J. The Standard model of particle physics. *Rev. Mod. Phys.* **71,** S96–S111 (1999).
5. Higgs, P. W. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.* **13,** 508–509 (16 1964).
6. Englert, F. & Brout, R. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.* **13,** 321–323 (9 1964).
7. CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B* **716,** 30 –61. ISSN: 0370-2693 (2012).
8. ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B* **716,** 1 –29. ISSN: 0370-2693 (2012).
9. Perlmutter *et al.*, S. Measurements of Ω and Λ from 42 high redshift supernovae. *Astrophys. J.* **517,** 565–586 (1999).
10. Riess *et al.*, A. G. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116,** 1009–1038 (1998).
11. Rubin, V., Ford, W. & Thonnard, N. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 /R = 4kpc/ to UGC 2885 /R = 122 kpc/. *The Astrophysical Journal* **238,** 471–487 (May 1980).
12. Natarajan, P. *et al.* Mapping substructure in the HST Frontier Fields cluster lenses and in cosmological simulations. *Mon. Not. Roy. Astron. Soc.* **468,** 1962–1980 (2017).
13. Ade, P. A. R. *et al.* Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594,** A13 (2016).
14. Jarosik *et al.*, N. Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors, and Basic Results. **192,** 14 (Feb. 2011).

15. Markevitch, M. Chandra observation of the most interesting cluster in the universe. *ESA Spec. Publ.* **604,** 723 (2006).

16. Clowe *et al.*, D. A direct empirical proof of the existence of dark matter. *Astrophys. J. Lett.* **648,** L109–L113 (2006).

17. Alexander *et al.*, J. *Dark Sectors 2016 Workshop: Community Report* in (Aug. 2016). arXiv: `1608.08632 [hep-ph]`.

18. Peccei, R. & Quinn, H. CP Conservation in the Presence of Pseudoparticles. *Physical Review Letters - PHYS REV LETT* **38,** 1440–1443 (June 1977).

19. Moore, B. Evidence against dissipation-less dark matter from observations of galaxy haloes. **370,** 629–631 (Aug. 1994).

20. Ilten, P., Soreq, Y., Williams, M. & Xue, W. Serendipity in dark photon searches. *JHEP* **06,** 004 (2018).

21. Branco *et al.*, G. Theory and phenomenology of two-Higgs-doublet models. *Phys. Rept.* **516,** 1–102 (2012).

22. Haisch, U. & Kamenik, J. F. Searching for new spin-0 resonances at LHCb. *Phys. Rev.* **D93,** 055047 (2016).

23. Haisch, U., Kamenik, J. F., Malinauskas, A. & Spira, M. Collider constraints on light pseudoscalars. *JHEP* **03,** 178 (2018).

24. Sirunyan *et al.*, A. M. Search for resonances in the mass spectrum of muon pairs produced in association with b quark jets in proton-proton collisions at $\sqrt{s} = 8$ and 13 TeV. *JHEP* **11,** 161 (2018).

25. Pierce *et al.*, A. Searching for confining hidden valleys at LHCb, ATLAS, and CMS. *Phys. Rev.* **D97,** 095033 (2018).

26. Bruning *et al.*, O. S. LHC Design Report Vol.1: The LHC Main Ring. doi:`10.5170/CERN-2004-003-V-1` (June 2004).

27. Herr, W. & Muratori, B. *Concept of luminosity* in (2006). doi:`10.5170/CERN-2006-002.361`. `http://cds.cern.ch/record/941318`.

28. CERN public relations. *Facts and figures about the LHC* `https://home.cern/resources/faqs/facts-and-figures-about-lhc` (2021).

29. Wenninger, J. Operation and Configuration of the LHC in Run 2. `https://cds.cern.ch/record/2668326` (2019).

30. Buning *et al.*, O. LHC Design Report. 2. The LHC infrastructure and general services. doi:`10.5170/CERN-2004-003-V-2` (Nov. 2004).

31. Benedikt *et al.*, M. LHC Design Report. 3. The LHC injector chain. doi:`10.5170/CERN-2004-003-V-3` (Dec. 2004).

32. Aaij *et al.*, R. Design and performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC. *JINST* **14,** P04013 (2019).

33. The LHCb Collaboration. The LHCb Detector at the LHC. *Journal of Instrumentation* **3,** S08005–S08005 (2008).

34. Feynman, R. P. The Theory of Positrons. *Physical Review* **76,** 749–759 (Sept. 1949).

35. Aubert, B. *et al.* The BaBar detector. *Nucl. Instrum. Meth. A* **479,** 1–116 (2002).

36. Abashian, A. *et al.* The Belle Detector. *Nucl. Instrum. Meth. A* **479,** 117–232 (2002).

37. Abe, T. *et al.* Belle II Technical Design Report. arXiv: `1011.0352 [physics.ins-det]` (Nov. 2010).

38. Tanabashi *et al.*, M. Review of Particle Physics. *Phys. Rev. D* **98,** 030001 (3 2018).

39. Aaij, R. *et al.* Measurement of $\sigma(pp \to \mathrm{b}\overline{\mathrm{b}}X$ at $\sqrt{s}$ =7 TeV in the forward region. *Phys. Lett.* **B694,** 209 (2010).

40. Elsässer, C. *$\bar{b}b$ production angle plots* `{\href{https://lhcb.web.cern.ch/lhcb/speakersbureau/html/bb\_ProductionAngles.html}{https://lhcb.web.cern.ch/lhcb/speakersbureau/html/bb\_ProductionAngles.html}}`.

41. Aaij, R. *et al.* The LHCb detector at the LHC. *JINST* **3,** S08005 (2008).

42. Aaij *et al.*, R. LHCb detector performance. *Int. J. Mod. Phys.* **A30,** 1530022 (2015).

43. Barbosa-Marinho *et al.*, P. R. *LHCb VELO (VErtex LOcator): Technical Design Report* `https://cds.cern.ch/record/504321` (CERN, Geneva, 2001).

44. Barbosa-Marinho *et al.*, P. R. *LHCb outer tracker: Technical Design Report* `http://cds.cern.ch/record/519146` (CERN, Geneva, 2001).

45. Djeliadine, R, Iouchtchenko, O & Obraztsov, V. F. *LHCb hadron trigger and Hcal cell size and length optimization* tech. rep. LHCb-99-035 (CERN, Geneva, 1999). `http://cds.cern.ch/record/691688`.

46. Barsuk *et al.*, S. *Design and construction of electromagnetic calorimeter for LHCb experiment* tech. rep. LHCb-2000-043 (CERN, Geneva, 2000). `http://cds.cern.ch/record/691508`.

47. Dzhelyadin, R. The LHCb hadron calorimeter. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **494,** 332 –339. ISSN: 0168-9002 (2002).

48. Barbosa-Marinho *et al.*, P. R. *LHCb muon system: Technical Design Report* `http://cds.cern.ch/record/504326` (CERN, Geneva, 2001).

49. LHCb Collaboration. *LHCb muon system: addendum to the Technical Design Report* `http://cds.cern.ch/record/600536` (CERN, Geneva, 2003).

50. LHCb Collaboration. *LHCb muon system: second addendum to the Technical Design Report* Submitted on 9 Apr 2005. `http://cds.cern.ch/record/831955` (CERN, Geneva, 2005).

51. LHCb collaboration. LHCb Trigger and Online Technical Design Report (2014).

52. LHCb collaboration. *Trigger Schemes* `http://lhcb.web.cern.ch/lhcb/speakersbureau/html/TriggerScheme.html`.

53. Dujany, G. & Storaci, B. Real-time alignment and calibration of the LHCb Detector in Run II. *J. Phys. Conf. Ser.* **664,** 082010 (2015).

54. Aaij, R. *et al.* Tesla: An application for real-time data analysis in high energy physics. *Comput. Phys. Commun.* **208,** 35–42 (2016).

55. LHCb Collaboration. *Letter of Intent for the LHCb Upgrade* tech. rep. CERN-LHCC-2011-001. LHCC-I-018 (Geneva, 2011). `https://cds.cern.ch/record/1333091`.

56. LHCb Collaboration. *Computing Model of the Upgrade LHCb experiment* tech. rep. CERN-LHCC-2018-014. LHCB-TDR-018 (CERN, Geneva, 2018). `https://cds.cern.ch/record/2319756`.

57. LHCb Collaboration. *LHCb VELO Upgrade Technical Design Report* tech. rep. CERN-LHCC-2013-021. LHCB-TDR-013 (2013). `https://cds.cern.ch/record/1624070`.

58. LHCb Collaboration. *LHCb Tracker Upgrade Technical Design Report* tech. rep. CERN-LHCC-2014-001. LHCB-TDR-015 (2014). `https://cds.cern.ch/record/1647400`.

59. LHCb Collaboration. *LHCb PID Upgrade Technical Design Report* tech. rep. CERN-LHCC-2013-022. LHCB-TDR-014 (2013). `https://cds.cern.ch/record/1624074`.

60. LHCb collaboration. Physics case for an LHCb Upgrade II — Opportunities in flavour physics, and beyond, in the HL-LHC era. arXiv: `1808.08865 [hep-ex]` (2018).

61. LHCb collaboration. Expression of Interest for a Phase-II LHCb Upgrade: Opportunities in flavour physics, and beyond, in the HL-LHC era (2017).

62. LHCb Collaboration. *LHCb Upgrade GPU High Level Trigger Technical Design Report* tech. rep. CERN-LHCC-2020-006. LHCB-TDR-021 (CERN, Geneva, 2020). `https://cds.cern.ch/record/2717938`.

63. Kitouni, O., Nachman, B., Weisser, C. & Williams, M. *Enhancing searches for resonances with machine learning and moment decomposition* 2020. arXiv: `2010.09745 [hep-ph]`.

64. Williams, M. A novel approach to the bias-variance problem in bump hunting. *Journal of Instrumentation* **12,** P09034âĂŞP09034. ISSN: 1748-0221 (2017).

65. Button *et al.*, J. Pion-Pion Interaction in the Reaction $\bar{p}+p \to 2\pi^{+}+2\pi^{-}+n\pi^{0}$. *Phys. Rev.* **126,** 1858–1863 (1962).

66. Sirunyan *et al.*, A. M. Search for high mass dijet resonances with a new background prediction method in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP* **05,** 033 (2020).

67. Aad *et al.*, G. Search for new resonances in mass distributions of jet pairs using 139 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP* **03,** 145 (2020).

68. Aaij *et al.*, R. Searches for low-mass dimuon resonances. arXiv: `2007.03923 [hep-ex]` (July 2020).

69. Adam *et al.*, J. Pair invariant mass to isolate background in the search for the chiral magnetic effect in Au+Au collisions at $\sqrt{s_{\mathrm{NN}}}$= 200 GeV. arXiv: `2006.05035 [nucl-ex]` (June 2020).

70. Acharya *et al.*, S. J/ψ elliptic and triangular flow in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 5.02 TeV. arXiv: `2005.14518 [nucl-ex]` (May 2020).

71. Adrian *et al.*, P. Search for a dark photon in electroproduced e^+e^- pairs with the Heavy Photon Search experiment at JLab. *Phys. Rev. D* **98,** 091101 (2018).

72. McCracken *et al.*, M. Search for baryon-number and lepton-number violating decays of Λ hyperons using the CLAS detector at Jefferson Laboratory. *Phys. Rev. D* **92,** 072002 (2015).

73. Ablikim *et al.*, M. Observation of the leptonic decay $D^+ \to \tau^+\nu_\tau$. *Phys. Rev. Lett.* **123,** 211802 (2019).

74. Belle-II. Search for Axion-Like Particles produced in e^+e^- collisions at Belle II. arXiv: `2007.13071 [hep-ex]` (July 2020).

75. Frate *et al.*, M. Modeling Smooth Backgrounds and Generic Localized Signals with Gaussian Processes. arXiv: `1709.05681 [physics.data-an]` (2017).

76. Larkoski, A. J., Moult, I. & Nachman, B. Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning. *Phys. Rept.* **841,** 1–63 (2020).

77. Guest, D., Cranmer, K. & Whiteson, D. Deep Learning and its Application to LHC Physics. arXiv: `1806.11484 [hep-ex]` (2018).

78. Albertsson *et al.*, K. Machine Learning in High Energy Physics Community White Paper. arXiv: `1807.02876 [physics.comp-ph]` (2018).

79. Radovic *et al.*, A. Machine learning at the energy and intensity frontiers of particle physics. *Nature* **560,** 41–48 (2018).

80. Bourilkov, D. Machine and Deep Learning Applications in Particle Physics. *Int. J. Mod. Phys. A* **34,** 1930019 (2020).

81. Aaboud *et al.*, M. Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC. *Eur. Phys. J. C* **79,** 375 (2019).

82. Sirunyan *et al.*, A. M. Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques. *JINST* **15,** P06005 (2020).

83. Louppe, G., Kagan, M. & Cranmer, K. in *Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al.*) 981–990 (Curran Associates, Inc., 2017). eprint: `1611.01046`. `http://papers.nips.cc/paper/6699-learning-to-pivot-with-adversarial-networks.pdf`.

84. Dolen *et al.*, J. Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure. *JHEP* **05,** 156 (2016).
85. Moult, I., Nachman, B. & Neill, D. Convolved Substructure: Analytically Decorrelating Jet Substructure Observables. *JHEP* **05,** 002 (2018).
86. Stevens, J. & Williams, M. uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers. *JINST* **8,** P12013 (2013).
87. Shimmin *et al.*, C. Decorrelated Jet Substructure Tagging using Adversarial Neural Networks. arXiv: `1703.03507 [hep-ex]` (2017).
88. Bradshaw *et al.*, L. Mass Agnostic Jet Taggers. *SciPost Phys.* **8,** 011 (2020).
89. ATLAS collaboration. *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS* tech. rep. ATL-PHYS-PUB-2018-014 (CERN, Geneva, 2018). `https://cds.cern.ch/record/2630973`.
90. Kasieczka, G. & Shih, D. DisCo Fever: Robust Networks Through Distance Correlation. arXiv: `2001.05310 [hep-ph]` (2020).
91. Xia, L.-G. QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics. *Nucl. Instrum. Meth.* **A930,** 15–26 (2019).
92. Englert *et al.*, C. Machine Learning Uncertainties with Adversarial Neural Networks. *Eur. Phys. J.* **C79,** 4 (2019).
93. Wunsch *et al.*, S. Reducing the dependence of the neural network function to systematic uncertainties in the input space. arXiv: `1907.11674 [physics.data-an]` (2019).
94. Rogozhnikov *et al.*, A. New approaches for boosting to uniformity. *JINST* **10,** T03002 (2015).
95. CMS Collaboration. A deep neural network to search for new long-lived particles decaying to jets. *Machine Learning: Science and Technology.* doi:`10.1088/2632-2153/ab9023`. eprint: `1912.12238` (2020).
96. Clavijo, J. M., Glaysher, P. & Katzy, J. M. Adversarial domain adaptation to reduce sample bias of a high energy physics classifier. arXiv: `2005.00568 [stat.ML]` (2020).
97. Kasieczka *et al.*, G. ABCDisCo: Automating the ABCD Method with Machine Learning. arXiv: `2007.14400 [hep-ph]` (July 2020).
98. Chang, S., Cohen, T. & Ostdiek, B. What is the Machine Learning? *Phys. Rev. D* **97,** 056009 (2018).
99. Clavijo, J. M., Glaysher, P. & Katzy, J. M. Adversarial domain adaptation to reduce sample bias of a high energy physics classifier. arXiv: `2005.00568 [stat.ML]` (May 2020).
100. Sirunyan *et al.*, A. M. A deep neural network to search for new long-lived particles decaying to jets. arXiv: `1912.12238 [hep-ex]` (Dec. 2019).

101. CMS Collaboration. Search for Low Mass Vector Resonances Decaying to Quark-Antiquark Pairs in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.* **119,** 111802 (2017).
102. CMS Collaboration. Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair. *Phys. Rev. Lett.* **120,** 071802 (2018).
103. Sirunyan *et al.*, A. M. Search for low mass vector resonances decaying into quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP* **01,** 097 (2018).
104. ATLAS Collaboration. Search for light resonances decaying to boosted quark pairs and produced in association with a photon or a jet in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Lett. B* **788,** 316 (2019).
105. CMS Collaboration. Search for low-mass resonances decaying into bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D* **99,** 012005 (2019).
106. Sirunyan *et al.*, A. M. Search for dark matter produced in association with a Higgs boson decaying to a pair of bottom quarks in proton–proton collisions at $\sqrt{s} = 13\,$TeV. *Eur. Phys. J. C* **79,** 280 (2019).
107. Sirunyan *et al.*, A. M. Measurement and interpretation of differential cross sections for Higgs boson production at $\sqrt{s} = 13$ TeV. *Phys. Lett. B* **792,** 369–396 (2019).
108. CMS Collaboration. Search for Low-Mass Quark-Antiquark Resonances Produced in Association with a Photon at $\sqrt{s}$ =13 TeV. *Phys. Rev. Lett.* **123,** 231803 (2019).
109. CMS Collaboration. A multi-dimensional search for new heavy resonances decaying to boosted WW, WZ, or ZZ boson pairs in the dijet final state at 13 TeV. *Eur. Phys. J. C* **80,** 237 (2020).
110. CMS Collaboration. Search for low mass vector resonances decaying into quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D* **100,** 112007 (2019).
111. Aad *et al.*, G. Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector. *Phys. Rev. Lett.* **125,** 131801 (2020).
112. Sirunyan *et al.*, A. M. Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV. arXiv: `2006.13251 [hep-ex]` (June 2020).
113. Aaij *et al.*, R. Amplitude analysis of the $B^+ \to D^+D^-K^+$ decay. arXiv: `2009.00026 [hep-ex]` (Aug. 2020).
114. Aaij *et al.*, R. A model-independent study of resonant structure in $B^+ \to D^+D^-K^+$ decays. arXiv: `2009.00025 [hep-ex]` (Aug. 2020).

115. Aaij *et al.*, R. Measurement of the CP-violating phase ϕ_s from $B_s^0 \to J/\psi\pi^+\pi^-$ decays in 13 TeV pp collisions. *Phys. Lett. B* **797,** 134789 (2019).

116. Aaij *et al.*, R. Search for a dimuon resonance in the Υ mass region. *JHEP* **09,** 147 (2018).

117. Aaij *et al.*, R. Search for hidden-sector bosons in $B^0 \to K^{*0}\mu^+\mu^-$ decays. *Phys. Rev. Lett.* **115,** 161802 (2015).

118. Aaij *et al.*, R. First observation of forward $Z \to b\bar{b}$ production in pp collisions at $\sqrt{s} = 8$ TeV. *Phys. Lett. B* **776,** 430–439 (2018).

119. Aaij *et al.*, R. Measurement of forward $t\bar{t}$, $W + b\bar{b}$ and $W + c\bar{c}$ production in pp collisions at $\sqrt{s} = 8$ TeV. *Phys. Lett. B* **767,** 110–120 (2017).

120. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A Survey on Bias and Fairness in Machine Learning. arXiv: `1908.09635 [cs.LG]` (2019).

121. Chouldechova, A. & Roth, A. The Frontiers of Fairness in Machine Learning. arXiv: `1810.08810 [cs.LG]` (2018).

122. Edwards, H. & Storkey, A. J. *Censoring Representations with an Adversary* in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2016). eprint: `1511.05897`.

123. Ganin *et al.*, Y. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* **17,** 1–35 (2016).

124. De Oliveira, L., Kagan, M., Mackey, L., Nachman, B. & Schwartzman, A. Jet-Images – Deep Learning Edition. *JHEP* **07,** 069 (2016).

125. SzÃĺkely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35,** 2769–2794 (2007).

126. SzÃĺkely, G. J. & Rizzo, M. L. Brownian distance covariance. *Ann. Appl. Stat.* **3,** 1236–1265 (2009).

127. Székely, G. J. & Rizzo, M. L. The Distance Correlation T-test of Independence in High Dimension. *J. Multivar. Anal.* **117,** 193–213. ISSN: 0047-259X (2013).

128. Székely, G. J. & Rizzo, M. L. Partial distance correlation with methods for dissimilarities. *Ann. Statist.* **42,** 2382–2412 (2014).

129. Paszke *et al.*, A. in *Advances in Neural Information Processing Systems 32* (eds Wallach, H. *et al.*) 8024–8035 (Curran Associates, Inc., 2019). `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

130. et.al., M. A. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* Software available from tensorflow.org. 2015. `http://tensorflow.org/`.

131. Kasieczka, G. & Shih, D. *Datasets for Boosted W Tagging* version v1. Jan. 2020. doi:`10.5281/zenodo.3606767`. `https://doi.org/10.5281/zenodo.3606767`.

132. Ramachandran, P., Zoph, B. & Le, Q. V. Searching for Activation Functions. arXiv: `1710.05941 [cs.NE]` (2017).

133. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2015). `http://arxiv.org/abs/1412.6980`.

134. Smith, L. N. & Topin, N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. arXiv: `1708.07120 [cs.LG]` (2018).

135. Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. arXiv: `1608.03983 [cs.LG]` (2017).

136. Smith, L. N. & Topin, N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. arXiv: `1708.07120 [cs.LG]` (2017).

137. Bishop, C. M. *Mixture density networks* Technical Report (Birmingham, 1994). `http://publications.aston.ac.uk/id/eprint/373/`.

138. Weisser *et al.*, C. *Autoencoders for Compression and Simulation in Particle Physics* in (). Poster presented at ICLR FSAI 2020.

139. Ballard, D. H. *Modular Learning in Neural Networks* in *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1* (AAAI Press, Seattle, Washington, 1987), 279âĂŞ284. ISBN: 0934613427.

140. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *CoRR* **abs/1312.6114** (2014).

141. Goodfellow *et al.*, I. *Generative Adversarial Nets* in *Advances in Neural Information Processing Systems* **27** (Curran Associates, Inc., 2014), 2672–2680.

142. Arjovsky, M., Chintala, S. & Bottou, L. *Wasserstein GAN* 2017. arXiv: `1701.07875 [stat.ML]`.

143. Bellemare *et al.*, M. G. The Cramer Distance as a Solution to Biased Wasserstein Gradients. *CoRR* **abs/1705.10743.** arXiv: `1705.10743`. `http://arxiv.org/abs/1705.10743` (2017).

144. Agostinelli, S. *et al.* GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.* **A506,** 250–303 (2003).

145. Müller, D., Clemencic, M., Corti, G. & Gersabeck, M. ReDecay: A novel approach to speed up the simulation at LHCb. *Eur. Phys. J.* **C78,** 1009 (2018).

146. Aaij, R. *et al.* Search for lepton-universality violation in $\mathrm{B}^+ \to \mathrm{K}^+\ell^+\ell^-$ decays. *Phys. Rev. Lett.* **122,** 191801 (2019).

147. *et al.*, T. Y. Recent Trends in Deep Learning Based Natural Language Processing. *CoRR* **abs/1708.02709.** arXiv: `1708.02709`. `http://arxiv.org/abs/1708.02709` (2017).

148. Pennington, J., Socher, R. & Manning, C. D. *GloVe: Global Vectors for Word Representation* in *Empirical Methods in Natural Language Processing (EMNLP)* (2014), 1532–1543. `http://www.aclweb.org/anthology/D14-1162`.

149. Mikolov, T., Chen, K., Corrado, G. S. & Dean, J. *Efficient Estimation of Word Representations in Vector Space* 2013. `http://arxiv.org/abs/1301.3781`.

150. *et al.*, M. E. P. Deep contextualized word representations. *CoRR* **abs/1802.05365.** arXiv: `1802.05365`. `http://arxiv.org/abs/1802.05365` (2018).

151. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9,** 1735âĂŞ1780. ISSN: 0899-7667 (Nov. 1997).

152. *et al.*, A. V. *Attention is All You Need* in (2017). `https://arxiv.org/pdf/1706.03762.pdf`.

153. Devlin, J. t. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019), 4171–4186. doi:`10.18653/v1/N19-1423`. `https://www.aclweb.org/anthology/N19-1423`.

154. ATLAS Collaboration. *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment* tech. rep. ATL-PHYS-PUB-2017-003 (CERN, Geneva, 2017). `https://cds.cern.ch/record/2255226`.

155. Egan *et al.*, S. *Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC* 2017. arXiv: `1711.09059 [hep-ex]`.

156. Pearkes *et al.*, J. Jet Constituents for Deep Neural Network Based Top Quark Tagging. arXiv: `1704.02124 [hep-ex]` (Apr. 2017).

157. Li, J. & Sun, H. An Attention Based Neural Network for Jet Tagging. arXiv: `2009.00170 [hep-ph]` (Aug. 2020).

158. Aaij *et al.*, R. *Comput Softw Big Sci 4, 7 (2020).* doi:`10.1007/s41781-020-00039-7`. arXiv: `1912.09161 [physics.ins-det]` (Dec. 19, 2019).

159. LHCb Collaboration. *LHCb Upgrade GPU High Level Trigger Technical Design Report* tech. rep. CERN-LHCC-2020-006. LHCB-TDR-021 (Geneva, 2020). `https://cds.cern.ch/record/2717938`.

160. Reiss, F. Fast parallel Primary Vertex reconstruction for the LHCb Upgrade. `https://cds.cern.ch/record/2717609` (2020).

161. Akar *et al.*, S. An updated hybrid deep learning algorithm for identifying and locating primary vertices. arXiv: `2007.01023 [physics.ins-det]` (July 2020).

162. Fang, R., Schreiner, H. F., Sokoloff, M. D., Weisser, C. & Williams, M. A hybrid deep learning approach to vertexing. *Journal of Physics: Conference Series* **1525,** 012079. ISSN: 1742-6596 (2020).

163. Akar, S. *et al.* Progress in developing a hybrid deep learning algorithm for identifying and locating primary vertices. arXiv: `2103.04962 [hep-ex]` (Mar. 2021).

164. Schreiner *et al.*, H. *pv-finder repository* `https://gitlab.cern.ch/LHCb-Reco-Dev/pv-finder`. 2020.

165. Sjöstrand, T., Mrenna, S. & Skands, P. PYTHIA 6.4 physics and manual. *Journal of High Energy Physics* **2006,** 026âĂŞ026. ISSN: 1029-8479 (2006).

166. Sjöstrand, T., Mrenna, S. & Skands, P. A brief introduction to PYTHIA 8.1. *Computer Physics Communications* **178,** 852âĂŞ867. ISSN: 0010-4655 (2008).

167. LHCb Collaboration. *LHCb VELO Upgrade Technical Design Report* tech. rep. CERN-LHCC-2013-021. LHCB-TDR-013 (2013). `https://cds.cern.ch/record/1624070`.

168. Paszke *et al.*, A. Automatic differentiation in PyTorch (2017).

169. Williams, M. Observing CP violation in many-body decays. *Phys. Rev. D* **84,** 054015 (5 2011).

170. Weisser, C. & Williams, M. Machine learning and multivariate goodness of fit. arXiv: `1612.07186` (2016).

171. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series* **5 50 (302,** 157âĂŞ175 (1900).

172. D'Agostino, R. B. & Stephens, M. A. *Goodness-of-Fit Techniques* ISBN: 0824774876 (Marcel Dekker, Inc., USA, 1986).

173. Williams, M. How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics. *Journal of Instrumentation* **5,** P09004–P09004. ISSN: 1748-0221 (2010).

174. Bellman, R. *Adaptive control processes: a guided tour* (Princeton University Press, 1961).

175. Kolmogorov, A. Sulla Determinazione Empirica di um Legge di Distribuizione. *Giornale Dell'Instituto Italiano Degli Attuari* **4,** 83–91 (1933).

176. Smirnov, N. Table for Estimating the Goodness of Fit of Empirical Distributions. *Ann. Math. Statist.* **19,** 279–281 (June 1948).

177. Anderson, T. W. & Darling, D. A. Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes. *Ann. Math. Statist.* **23,** 193–212 (June 1952).

178. CramÃĺr, H. On the composition of elementary errors. *Scandinavian Actuarial Journal* **1928,** 13–74 (1928).

179. Von Mises, R. *Wahrscheinlichkeit, Statistik und Wahrheit* (Julius Springer, 1928).

180. Deb, N. & Sen, B. Multivariate Rank-based Distribution-free Nonparametric Testing using Measure Transportation. *arXiv.* arXiv: `1909.08733`. `http://arxiv.org/abs/1909.08733` (2019).

181. Friedman, J. H. On multivariate goodness of fit and two sample testing. *eConf* **C030908,** 311–313 (2003).

182. Fisher, R. A. The Design of Experiments. *Nature* **137.** doi:`10.1038/137252a0` (1936).

183. Ramdas, A., Singh, A. & Wasserman, L. Classification Accuracy as a Proxy for Two Sample Testing. arXiv: `1602.02210`. `http://arxiv.org/abs/1602.02210` (2016).

184. Roederer, M. e. a. Probability binning comparison: a metric for quantitating univariate distribution differences. *Cytometry* **45, 37** (2001).

185. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *CoRR* **abs/1603.02754.** arXiv: `1603.02754`. `http://arxiv.org/abs/1603.02754` (2016).

186. Snoek, J. *SPEARMINT* 2016. `https://github.com/HIPS/Spearmint`.

187. Baldi, P., Sadowski, P. & Whiteson, D. Enhanced Higgs Boson toÏĎ+ÏĎâĹŠSearch with Deep Learning. *Physical Review Letters* **114.** ISSN: 1079-7114. doi:`10.1103/physrevlett.114.111801`. `http://dx.doi.org/10.1103/PhysRevLett.114.111801` (2015).

188. LHCb Collaboration. Search for $A^{'\to\mu^+\mu^-}$ Decays. *Physical Review Letters* **124,** 41801. ISSN: 10797114 (2020).

189. LHCb Collaboration. Searches for low-mass dimuon resonances. *Journal of High Energy Physics* **2020,** 1–26. ISSN: 10298479 (2020).

190. Essig, R. *et al. Working Group Report: New light weakly coupled particles* in *Proceedings, 2013 Community Summer Study on the Future of U.S. Particle Physics: Snowmass on the Mississippi (CSS2013): Minneapolis, MN, USA, July 29-August 6, 2013* (2013). doi:`notadoi`. arXiv: `1311.0029 [hep-ph]`. `https://inspirehep.net/record/1263039/files/arXiv:1311.0029.pdf`.

191. Alexander, J. *et al. Dark Sectors 2016 Workshop: Community Report* in (2016). doi:`notadoi`. arXiv: `1608.08632 [hep-ph]`. `http://inspirehep.net/record/1484628/files/arXiv:1608.08632.pdf`.

192. Battaglieri, M. *et al.* US Cosmic Visions: New Ideas in Dark Matter 2017: Community Report. arXiv: `1707.04591 [hep-ph]` (2017).

193. Tulin, S. & Yu, H.-B. Dark matter self-interactions and small scale structure. *Phys. Rept.* **730,** 1–57 (2018).

194. Okun, L. B. Limits of electrodynamics: Paraphotons? *Sov. Phys. JETP* **56.** [Zh. Eksp. Teor. Fiz. **83** (1982) 892], 502 (1982).

195. Galison, P. & Manohar, A. Two Z's or not two Z's? *Phys. Lett.* **B136,** 279 (1984).

196. Holdom, B. Two $U(1)$'s and ϵ charge shifts. *Phys. Lett.* **B166,** 196 (1986).

197. Pospelov, M., Ritz, A. & Voloshin, M. B. Secluded WIMP dark matter. *Phys. Lett.* **B662,** 53–61 (2008).

198. Arkani-Hamed, N., Finkbeiner, D. P., Slatyer, T. R. & Weiner, N. A theory of dark matter. *Phys. Rev.* **D79,** 015014 (2009).

199. Bjorken, J. D., Essig, R., Schuster, P. & Toro, N. New fixed-target experiments to search for dark gauge forces. *Phys. Rev.* **D80,** 075018 (2009).

200. Bergsma, F. *et al.* A search for decays of heavy neutrinos in the mass range 0.5 GeV to 2.8 GeV. *Phys. Lett.* **B166,** 473 (1986).

201. Konaka, A. *et al.* Search for neutral particles in electron-beam-dump experiment. *Phys. Rev. Lett.* **57,** 659 (1986).

202. Riordan, E. M. *et al.* Search for short-lived axions in an electron-beam-dump experiment. *Phys. Rev. Lett.* **59,** 755 (1987).

203. Bjorken, J. D. *et al.* Search for neutral metastable penetrating particles produced in the SLAC beam dump. *Phys. Rev.* **D38,** 3375 (1988).

204. Bross, A. *et al.* A search for short-lived particles produced in an electron beam dump. *Phys. Rev. Lett.* **67,** 2942–2945 (1991).

205. Davier, M. & Nguyen Ngoc, H. An unambiguous search for a light Higgs boson. *Phys. Lett.* **B229,** 150 (1989).

206. Athanassopoulos, C. *et al.* Evidence for $\nu_\mu \to \nu_e$ oscillations from pion decay in flight neutrinos. *Phys. Rev.* **C58,** 2489–2511 (1998).

207. Astier, P. *et al.* Search for heavy neutrinos mixing with tau neutrinos. *Phys. Lett.* **B506,** 27–38 (2001).

208. Essig, R., Harnik, R., Kaplan, J. & Toro, N. Discovering new light states at neutrino experiments. *Phys. Rev.* **D82,** 113008 (2010).

209. Williams, M., Burgess, C. P., Maharana, A. & Quevedo, F. New constraints (and motivations) for abelian gauge bosons in the MeV-TeV mass range. *JHEP* **08,** 106 (2011).

210. Blümlein, J. & Brunner, J. New exclusion limits for dark gauge forces from beam-dump data. *Phys. Lett.* **B701,** 155–159 (2011).

211. Gninenko, S. Constraints on sub-GeV hidden sector gauge bosons from a search for heavy neutrino decays. *Phys. Lett.* **B713,** 244–248 (2012).

212. Blümlein, J. & Brunner, J. New exclusion limits on dark gauge forces from proton bremsstrahlung in beam-dump data. *Phys. Lett.* **B731,** 320–326 (2014).

213. Banerjee, D. *et al.* Search for a hypothetical 16.7 MeV gauge boson and dark photons in the NA64 experiment at CERN. *Phys. Rev. Lett.* **120,** 231802 (2018).

214. Andreas, S., Niebuhr, C. & Ringwald, A. New limits on hidden photons from past electron beam dumps. *Phys. Rev.* **D86,** 095019 (2012).

215. Bergsma, F. *et al.* Search for axion like particle production in 400-GeV proton-copper interactions. *Phys. Lett.* **157B,** 458–462 (1985).

216. Abrahamyan, S. *et al.* Search for a new gauge boson in electron-nucleus fixed-target scattering by the APEX experiment. *Phys. Rev. Lett.* **107,** 191804 (2011).

217. Merkel, H. *et al.* Search at the Mainz Microtron for light massive gauge bosons relevant for the muon g-2 anomaly. *Phys. Rev. Lett.* **112,** 221802 (2014).

218. Merkel, H. *et al.* Search for light gauge bosons of the dark sector at the Mainz Microtron. *Phys. Rev. Lett.* **106,** 251802 (2011).

219. Aubert, B. *et al.* Search for dimuon decays of a light scalar boson in radiative transitions $\Upsilon \to \gamma A^0$. *Phys. Rev. Lett.* **103,** 081803 (2009).

220. Curtin, D. *et al.* Exotic decays of the 125 GeV Higgs boson. *Phys. Rev.* **D90,** 075004 (2014).

221. Lees, J. P. *et al.* Search for a dark photon in e^+e^- collisions at BaBar. *Phys. Rev. Lett.* **113,** 201801 (2014).

222. Ablikim, M. *et al.* Dark photon search in the mass range between 1.5 and 3.4 GeV/c^2. *Phys. Lett.* **B774,** 252–257 (2017).

223. Anastasi, A. *et al.* Limit on the production of a low-mass vector boson in $e^+e^- \to U\gamma$, $U \to e^+e^-$ with the KLOE experiment. *Phys. Lett.* **B750,** 633–637 (2015).

224. Anastasi, A. *et al.* Combined limit on the production of a light gauge boson decaying into $\mu^+\mu^-$ and $\pi^+\pi^-$. *Phys. Lett.* **B784,** 336–341 (2018).

225. Bernardi, G. *et al.* Search for neutrino decay. *Phys. Lett.* **B166,** 479 (1986).

226. Meijer Drees, R. *et al.* Search for weakly interacting neutral bosons produced in π^-p interactions at rest and decaying into e^+e^- pairs. *Phys. Rev. Lett.* **68,** 3845–3848 (1992).

227. Archilli, F. *et al.* Search for a vector gauge boson in ϕ meson decays with the KLOE detector. *Phys. Lett.* **B706,** 251–255 (2012).

228. Gninenko, S. Stringent limits on the $\pi^0 \to \gamma X, X \to e^+e^-$ decay from neutrino experiments and constraints on new light gauge bosons. *Phys. Rev.* **D85,** 055027 (2012).

229. Babusci, D. *et al.* Limit on the production of a light vector gauge boson in ϕ meson decays with the KLOE detector. *Phys. Lett.* **B720,** 111–115 (2013).

230. Adlarson, P. *et al.* Search for a dark photon in the $\pi^0 \to e^+e^-\gamma$ decay. *Phys. Lett.* **B726,** 187–193 (2013).

231. Agakishiev, G. *et al.* Searching a dark photon with HADES. *Phys. Lett.* **B731,** 265–271 (2014).

232. Adare, A. *et al.* Search for dark photons from neutral meson decays in pp and dAu collisions at $\sqrt{s_{NN}} = 200$ GeV. *Phys. Rev.* **C91,** 031901 (2015).

233. Batley, J. R. *et al.* Search for the dark photon in π^0 decays. *Phys. Lett.* **B746,** 178–185 (2015).

234. Anastasi, A. *et al.* Limit on the production of a new vector boson in $e^+e^- \to U\gamma$, U$\to \pi^+\pi^-$ with the KLOE experiment. *Phys. Lett.* **B757,** 356–361 (2016).

235. Essig, R., Mardon, J., Papucci, M., Volansky, T. & Zhong, Y.-M. Constraining light dark matter with low-energy e^+e^- colliders. *JHEP* **11,** 167 (2013).

236. Davoudiasl, H., Lee, H.-S. & Marciano, W. J. Muon $g-2$, rare kaon decays, and parity violation from dark bosons. *Phys. Rev.* **D89,** 095006 (2014).

237. Banerjee, D. *et al.* Search for invisible decays of sub-GeV dark photons in missing-energy events at the CERN SPS. *Phys. Rev. Lett.* **118,** 011802 (2017).

238. Lees, J. P. *et al.* Search for invisible decays of a dark photon produced in e^+e^- collisions at BaBar. *Phys. Rev. Lett.* **119,** 131804 (2017).

239. Adler, S. *et al.* Further evidence for the decay $K^+ \to \pi^+\nu\bar{\nu}$. *Phys. Rev. Lett.* **88,** 041803 (2002).

240. Adler, S. *et al.* Further search for the decay $K^+ \to \pi^+\nu\bar{\nu}$ in the momentum region $p < 195$ MeV/c. *Phys. Rev.* **D70,** 037102 (2004).

241. Artamonov, A. V. *et al.* Study of the decay $K^+ \to \pi^+\nu\bar{\nu}$ in the momentum region $140 < P_\pi < 199$ MeV/c. *Phys. Rev.* **D79,** 092004 (2009).

242. Fayet, P. Constraints on light dark matter and U bosons, from ψ, Υ, K^+, π^0, η and η' decays. *Phys. Rev.* **D74,** 054034 (2006).

243. Fayet, P. U-boson production in e^+e^- annihilations, ψ and Υ decays, and light dark matter. *Phys. Rev.* **D75,** 115017 (2007).

244. Fox, P. J., Harnik, R., Kopp, J. & Tsai, Y. LEP shines light on dark matter. *Phys. Rev.* **D84,** 014028 (2011).

245. Cortina Gil, E. *et al.* Search for production of an invisible dark photon in π^0 decays. *JHEP* **05,** 182 (2019).

246. Abdallah, J. *et al.* Photon events with missing energy in e^+e^- collisions at $\sqrt{s} =$ 130-GeV to 209-GeV. *Eur. Phys. J.* **C38,** 395–411 (2005).

247. Abdallah, J. *et al.* Search for one large extra dimension with the DELPHI detector at LEP. *Eur. Phys. J.* **C60,** 17–23 (2009).

248. Essig, R., Schuster, P., Toro, N. & Wojtsekhowski, B. An electron fixed target experiment to search for a new vector boson A' decaying to e^+e^-. *JHEP* **02,** 009 (2011).

249. Freytsis, M., Ovanesyan, G. & Thaler, J. Dark force detection in low energy *ep* collisions. *JHEP* **01,** 111 (2010).

250. Balewski, J. *et al. DarkLight: A search for dark forces at the Jefferson Laboratory Free-Electron Laser Facility* in *Proceedings, 2013 Community Summer Study on the Future of U.S. Particle Physics: Snowmass on the Mississippi (CSS2013): Minneapolis, MN, USA, July 29-August 6, 2013* (2013). doi:`notadoi`. arXiv: `1307.4432 [physics.ins-det]`. `https://misportal.jlab.org/ul/publications/view_pub.cfm?pub_id=12467`.

251. Wojtsekhowski, B., Nikolenko, D. & Rachek, I. Searching for a new force at VEPP-3. arXiv: `1207.5089 [hep-ex]` (2012).

252. Beranek, T., Merkel, H. & Vanderhaeghen, M. Theoretical framework to analyze searches for hidden light gauge bosons in electron scattering fixed target experiments. *Phys. Rev.* **D88,** 015032 (2013).

253. Echenard, B., Essig, R. & Zhong, Y.-M. Projections for dark photon searches at Mu3e. *JHEP* **01,** 113 (2015).

254. Battaglieri, M. *et al.* The Heavy Photon Search test detector. *Nucl. Instrum. Meth.* **A777,** 91–101 (2015).

255. Raggi, M. & Kozhuharov, V. Proposal to search for a dark photon in positron on target collisions at DAΦNE Linac. *Adv. High Energy Phys.* **2014,** 959802 (2014).

256. Alekhin, S. *et al.* A facility to Search for Hidden Particles at the CERN SPS: the SHiP physics case. *Rept. Prog. Phys.* **79,** 124201 (2016).

257. Gardner, S., Holt, R. J. & Tadepalli, A. S. New prospects in fixed target searches for dark forces with the SeaQuest experiment at Fermilab. *Phys. Rev.* **D93,** 115015 (2016).

258. Ilten, P., Thaler, J., Williams, M. & Xue, W. Dark photons from charm mesons at LHCb. *Phys. Rev.* **D92,** 115017 (2015).

259. Curtin, D., Essig, R., Gori, S. & Shelton, J. Illuminating dark photons with high-energy colliders. *JHEP* **02,** 157 (2015).

260. He, M., He, X.-G. & Huang, C.-K. Dark photon search at a circular e^+e^- collider. *Int. J. Mod. Phys.* **A32,** 1750138 (2017).

261. Kozaczuk, J. Dark photons from nuclear transitions. *Phys. Rev.* **D97,** 015014 (2018).

262. Ilten, P., Soreq, Y., Thaler, J., Williams, M. & Xue, W. Proposed inclusive dark photon search at LHCb. *Phys. Rev. Lett.* **116,** 251803 (2016).

263. Alexander, J. MMAPS: Missing-Mass A-Prime Search. *EPJ Web Conf.* **142,** 01001 (2017).

264. He, M., He, X.-G., Huang, C.-K. & Li, G. Search for a heavy dark photon at future e^+e^- colliders. *JHEP* **03,** 139 (2018).

265. Feng, J. L., Galon, I., Kling, F. & Trojanowski, S. ForwArd Search ExpeRiment at the LHC. *Phys. Rev.* **D97,** 035001 (2018).

266. Nardi, E., Carvajal, C. D. R., Ghoshal, A., Meloni, D. & Raggi, M. Resonant production of dark photons in positron beam dump experiments. *Phys. Rev.* **D97,** 095004 (2018).

267. D'Onofrio, M., Fischer, O. & Wang, Z. S. Searching for dark photons at the LHeC and FCC-he. arXiv: `1909.02312 [hep-ph]` (2019).

268. Tsai, Y.-D., deNiverville, P. & Liu, M. X. The high-energy frontier of the intensity frontier: Closing the dark photon, inelastic dark matter, and muon g-2 windows. arXiv: `1908.07525 [hep-ph]` (2019).

269. Aaij, R. *et al.* Search for dark photons produced in 13 TeVpp collisions. *Phys. Rev. Lett.* **120,** 061801 (2018).

270. Fayet, P. Extra U(1)'s and new forces. *Nucl. Phys.* **B347,** 743–768 (1990).

271. Berlin, A., Blinov, N., Gori, S., Schuster, P. & Toro, N. Cosmology and accelerator tests of strongly interacting dark matter. *Phys. Rev.* **D97,** 055033 (2018).

272. Aaij, R. *et al.* Search for $A' \to \mu^+\mu^-$ decays. *Phys. Rev. Lett.* **124,** 041801 (2020).

273. Aaij *et al.*, R. The LHCb trigger and its performance in 2011. *JINST* **8,** P04022 (2013).

274. Sjöstrand, T., Mrenna, S. & Skands, P. A brief introduction to PYTHIA 8.1. *Comput. Phys. Commun.* **178,** 852–867 (2008).

275. Sjöstrand, T., Mrenna, S. & Skands, P. Z. PYTHIA 6.4 Physics and Manual. *JHEP* **05,** 026 (2006).

276. Belyaev, I. *et al.* Handling of the generation of primary events in Gauss, the LHCb simulation framework. *J. Phys. Conf. Ser.* **331,** 032047 (2011).

277. Lange, D. J. The EvtGen particle decay simulation package. *Nucl. Instrum. Meth.* **A462,** 152–155 (2001).

278. Golonka, P. & Was, Z. PHOTOS Monte Carlo: A precision tool for QED corrections in Z and W decays. *Eur. Phys. J.* **C45,** 97–107 (2006).

279. Allison, J., Amako, K., Apostolakis, J., Araujo, H., Dubois, P., *et al.* Geant4 developments and applications. *IEEE Trans.Nucl.Sci.* **53,** 270 (2006).

280. Agostinelli, S. *et al.* Geant4: A simulation toolkit. *Nucl. Instrum. Meth.* **A506,** 250 (2003).

281. Clemencic, M *et al.* The LHCb simulation application, Gauss: Design, evolution and experience. *J. Phys. Conf. Ser.* **331,** 032023 (2011).

282. Spira, M., Djouadi, A., Graudenz, D. & Zerwas, P. Higgs boson production at the LHC. *Nucl. Phys. B* **453,** 17–82 (1995).

283. Spira, M. HIGLU: A program for the calculation of the total Higgs production cross-section at hadron colliders via gluon fusion including QCD corrections. arXiv: `hep-ph/9510347` (Oct. 1995).

284. Ball, R. D. *et al.* Parton distributions for the LHC Run II. *JHEP* **04,** 040 (2015).

285. Djouadi, A., Kalinowski, J. & Spira, M. HDECAY: A Program for Higgs boson decays in the standard model and its supersymmetric extension. *Comput. Phys. Commun.* **108,** 56–74 (1998).

286. Djouadi, A., Kalinowski, J., Muehlleitner, M. & Spira, M. HDECAY: Twenty$_{++}$ years after. *Comput. Phys. Commun.* **238,** 214–231 (2019).

287. Sjöstrand *et al.*, T. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.* **191,** 159–177 (2015).

288. Schwaller, P., Stolarski, D. & Weiler, A. Emerging Jets. *JHEP* **05,** 059 (2015).

289. Cassel, S., Ghilencea, D. M. & Ross, G. G. Electroweak and dark matter constraints on a Z' in models with a hidden valley. *Nucl. Phys.* **B827,** 256–280 (2010).

290. Cline, J. M., Dupuis, G., Liu, Z. & Xue, W. The windows for kinetically mixed Z'-mediated dark matter and the galactic center gamma ray excess. *JHEP* **08,** 131 (2014).

291. Cacciari, M., Salam, G. P. & Soyez, G. The anti-k$_T$ jet clustering algorithm. *JHEP* **0804,** 063 (2008).

292. Cacciari, M., Salam, G. P. & Soyez, G. FastJet user manual. *Eur.Phys.J.* **C72,** 1896 (2012).

293. Aaij, R. *et al.* Study of forward Z +jet production in pp collisions at $\sqrt{s}$ =7 TeV. *JHEP* **01,** 033 (2014).

294. See the Supplemental Material to this Letter for additional fit details and figures.

295. Williams, M. A novel approach to the bias-variance problem in bump hunting. *JINST* **12,** P09034 (2017).

296. Dauncey, P. D., Kenzie, M., Wardle, N. & Davies, G. J. Handling uncertainties in background shapes. *JINST* **10,** P04015 (2015).

297. *Search for a narrow resonance decaying to a pair of muons in proton-proton collisions at 13 TeV* tech. rep. CMS-PAS-EXO-19-018 (CERN, Geneva, 2019). `https://cds.cern.ch/record/2684861`.

298. Aaij *et al.*, R. Performance of the LHCb Vertex Locator. *JINST* **9,** P09007 (2014).

299. Alexander *et al.*, M. Mapping the material in the LHCb vertex locator using secondary hadronic interactions. *JINST* **13,** P06008 (2018).

300. Likhomanenko, T. *et al.* LHCb topological trigger reoptimization. *J. Phys. Conf. Ser.* **664,** 082025 (2015).

301. Gligorov, V. V. & Williams, M. Efficient, reliable and fast high-level triggering using a bonsai boosted decision tree. *JINST* **8,** P02013 (2013).

302. Aaij, R. *et al.* Measurement of the $B_s^0 \to \mu^+\mu^-$ branching fraction and effective lifetime and search for $B^0 \to \mu^+\mu^-$ decays. *Phys. Rev. Lett.* **118,** 191801 (2017).

303. Aaij, R. *et al.* Identification of beauty and charm quark jets at LHCb. *JINST* **10,** P06013 (2015).

304. Aaij, R. *et al.* Improved limit on the branching fraction of the rare decay $K^0_{\mathrm{S}} \to \mu\mu$. *Eur. Phys. J.* **C77,** 678 (2017).

305. Rolke, W. A., Lopez, A. M. & Conrad, J. Limits and confidence intervals in the presence of nuisance parameters. *Nucl. Instrum. Meth.* **A551,** 493–503 (2005).

306. Williams, M. Searching for a particle of unknown mass and lifetime in the presence of an unknown non-monotonic background. *JINST* **10,** P06002 (2015).

307. Aaij, R. *et al.* Search for hidden-sector bosons in $\mathrm{B}^0 \to \mathrm{K}^{*0}\mu^+\mu^-$ decays. *Phys. Rev. Lett.* **115,** 161802 (2015).

308. Aaij, R. *et al.* Search for long-lived scalar particles in $\mathrm{B}^+ \to \mathrm{K}^+\chi(\mu^+\mu^-)$ decays. *Phys. Rev.* **D95,** 071101 (2017).

309. Aaij *et al.*, R. Measurement of the track reconstruction efficiency at LHCb. *JINST* **10,** P02007 (2015).

310. Aaij, R. *et al.* Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s}$ =13 TeV. *JHEP* **09,** 136 (2016).

311. Aaij, R. *et al.* Measurement of Υ production cross-section in pp collisions at $\sqrt{s}$ =13 TeV. *JHEP* **07,** 134 (2018).

312. Lees, J. *et al.* Search for di-muon decays of a low-mass Higgs boson in radiative decays of the Υ(1S). *Phys. Rev. D* **87.** [Erratum: Phys.Rev.D 87, 059903 (2013)], 031102 (2013).

313. Chatrchyan, S. *et al.* Search for a light pseudoscalar Higgs boson in the dimuon decay channel in *pp* collisions at $\sqrt{s}$ = 7 TeV. *Phys. Rev. Lett.* **109,** 121801 (2012).

314. Sirunyan, A. M. *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,** 131802 (2020).

315. Aaij, R. *et al.* Search for a dimuon resonance in the Υ mass region. *JHEP* **09,** 147 (2018).