Doctoral Thesis

Searching for rare charm decays, performing alignment studies and improving the analysis ecosystem in HEP

Author:
Chris Burr

Supervisor:
Prof. Chris Parkes

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the

School of Physics and Astronomy
Faculty of Science and Engineering

2019

Blank page

Contents

Word count: 36,405

List of Figures

Blank page

List of Tables

Blank page

THE UNIVERSITY OF MANCHESTER

Abstract

Faculty of Science and Engineering
School of Physics and Astronomy

Doctor of Philosophy

title

by Chris Burr

A search for rare and forbidden decays of the form $D^+_{(s)} \to h^\pm l^+ l'^\mp$ has been performed, where h is a charged pion or kaon and l is an electron or muon, using $1.5\,\text{fb}^{-1}$ of data that was collected by the LHCb detector during 2016. No statistically significant deviations from the background only hypothesis are observed and upper limits are given for 25 final states, with 23 improving upon the previous world's best measurements. The majority of these limits are improved by more than an order of magnitude and some by up to a factor of 500.

Detector alignment studies are presented here that have been used to influence the development of the LHCb Upgrade VELO. This has included the alignment of several thousand testbeam datasets which were then analysed to provide results that guided the design of the new detector. Furthermore, a comprehensive study on the physics impact of thermally induced distortions of LHCb Upgrade VELO modules has been performed. This work resulted in changes to the manufacturing and quality assurance procedure and will help ensure optimal performance of the final detector.

This thesis also discusses various efforts that have been made to improve the analysis ecosystem in high energy physics in response to the challenges faced during the aforementioned studies. This includes improving the tools, procedures and software training that are available to analysts to improve productivity and encourage long term analysis preservation. Such improvements will be essential to effectively utilise modern high energy physics experiments which are unprecedented in both scale and duration.

Blank page

Declaration of Authorship

Candidate name: Chris Burr
Faculty: Faculty of Science and Engineering
Thesis title: Searching for rare charm decays, performing alignment studies and improving the analysis ecosystem in HEP

This work represents the combined efforts of the author and his colleagues in the LHCb collaboration. Some of the content has been published elsewhere and/or presented to several audiences. No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signed:

Date:

Blank page

Copyright Statement

[1] See `https://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420`

[2] See `https://www.library.manchester.ac.uk/about/regulations/`

[3] See `https://creativecommons.org/licenses/by/4.0/`

Blank page

Acknowledgements

In the declaration of authorship section it states that the work contained within this thesis is the combined effort of the entire LHCb collaboration. Of course this is true and is the justification by which our papers are signed by over 850 people but it is more than just that. These many hundreds of people manage to work towards their own goals and interests while still striving to help by offering support and guidance, volunteering sleepless nights, sacrificing their week of caffeination or by gallanting across borders to help broken down vehicles. I am thankful of how welcoming LHCb has been and how I can expect my daily interactions to be not good but excellent. There are many people whom could have been mentioned here both from within LHCb and other areas at CERN and, even if I do not name you, I am thankful for having met you all and wish you the best in all of your endeavours.

Ultimately it was a coin toss that saved me from a year of magnet trouble and instead led me to measure charm cross-sections with Alex Pearce, Dominik Müller, Patrick Spradlin, Sajan Easo and others. Each of you showed patience and compassion as you guided me through my first year in LHCb, I learnt so much and will always remember this time fondly.

Next I found myself in Manchester and here I must thank ~~Chris 1~~ Chris one for being the best PhD supervisor I could have wished for. You have given me the freedom to explore my own interests while also always being available to guide, mentor and advise me. I have learnt so much from you over the years that will continue to be useful regardless of what the future holds.

There are so many people whom I have met in Manchester and far too many to list here but you have all made it wonderful place to be over the last three and a half years. To name but a few of you: Thank you Marco Gersabeck for guiding me to Manchester. Thank you Silvia Borghi and Lucia Grillo for your support, especially during the tougher times of my PhD. Thank you to Igor Babuschkin for giving with me so much of the computing and statistics knowledge I use today. Gediminas Šarpis ir Judita Šarpienė, nuneš mane į kalnus.

To the organisers, teachers, helpers and participants of the Starterkit, I have thoroughly enjoyed being involved in all these activities with you. Most of all, a special thanks must go to Violaine Bellée. It was a pleasure to organise the 2017 Starterkit and 2018 Impactkit with you and I hope we can organise something else together one day.

Thank you Ben Couturier for your support and enabling my side projects. Chapter 7 would not have been the same without you.

I am also sincerely grateful to the European Union, AIDA-2020 and STFC for the funding which made this all possible.

Thank you to my examiners Rodger Jones and Terry Wyatt for taking the time to review this thesis and providing an enjoyable and thought provoking viva.

Looking beyond academia, thank you Mum, Dad and Matsie. Over the years you have supported me in everything I have wanted to do and formed me into the person I am today.

Last, but most certainly not least, mulțumesc frumos Cristina. Mulțumesc pentru prietenia ta și sprijinul acordat în ultimii trei ani. Am iubit timpul împreună micuța mea, fie în aventură sau ca niște cartofiori. Aștept cu nerăbdare călătoria noastră la Genève și mai departe ♥

Dedicated to Aunty Jean

Blank page

Preface

Modern high energy physics is an intensely computational field which has led to the advancement of many influential technologies such as the World Wide Web, grid computing and many detector technologies. The scale of the challenges faced has long since exceeded the capabilities of individual researchers. In response, large collaborations have been formed to enable monumental achievements such as detecting cosmic neutrinos, better understanding the strong force, discovering the Higgs boson and observing gravitational waves. With the looming upgrades to the Large Hadron Collider experiments, the challenges are greater than ever as datasets continue to grow. This will necessitate advancements to how research is performed and this has been a defining theme while the work documented here was performed. This thesis is structured as follows:

The first chapter gives a brief overview of the standard model of particle physics and concludes with some motivations for studying rare charm decays.

The second chapter introduces the LHCb experiment, one of the four main experiments at CERN's Large Hadron Collider in Geneva. The various sub detectors used during Run 1 and 2 of the LHC are described followed by a summary of some of the software tools which are most commonly used. Finally the LHCb Upgrade I programme is explained with a focus on the Vertex Locator.

The third chapter explains how the position of tracking detectors can be determined in a procedure known as *alignment* and concludes by explaining how these methods have been applied for testbeam data collected with the LHCb Timepix3 telescope.

The fourth chapter shows a study that was performed by the author to examine the potential physics impact of misalignment in the vertex locator of the LHCb Upgrade I programme.

The fifth and sixth chapters present a search for 28 decays of the form $D^+_{(s)} \to h^\pm l^+ l'^\mp$, where h is a kaon or pion and l is an electron or muon, for which the author was the primary analyst.

The seventh chapter is split into five sections which present additional work that has been performed by the author in response to some of the challenges that were found when performing the aforementioned work. This includes sections on analysis preservation, software training, software packaging, distributed computing and improvements to a statistical method known as the *energy test*.

Prior to this work, the author of this thesis first joined the LHCb collaboration in 2014 as a University of Southampton MPhys student working on the early measurements campaign of LHC Run 2. In this time he was a lead proponent of the $\sqrt{s} = 13\,\text{TeV}$ charm production measurements[1] and the analysis framework developed during this time was reused for charm production measurements at $\sqrt{s} = 5\,\text{TeV}$[2]. While both papers were published during my PhD, much of the work on the first paper was performed prior to joining Manchester. These analyses are also described in the theses [3, 4] and hence for brevity are not reported here.

Throughout this thesis the use of natural units ($c = \hbar = 1$) and the inclusion of charge conjugate processes is assumed unless otherwise stated.

Chapter 1

Theory

The standard model of particle physics has proven to be a tremendous achievement, providing predictions for a wide array of processes that have been subsequently shown to agree with experimental results with excellent precision. Despite this success, the standard model fails to provide justifications for many observations, such as the asymmetry between matter and antimatter in the universe, neutrino oscillations and the nature of dark matter and dark energy.[5] The desire to explain these gaps motivates physicists to look at processes that are rare, or even forbidden, in the standard model. Here observations can be dominated by new physics models that make negligible contributions in other more common processes where the standard model has already shown remarkable precision. In addition, measurements in flavour physics can probe regimes that are inaccessible to direct searches for new particles. Measurements of rare and forbidden decays can be sensitive to new physics, giving sensitivity at energy scales far beyond the centre of mass energy of the collider[6, 7] by the observation of contributions from virtual corrections.

1.1 The standard model of particle physics

In nature there are four known fundamental forces: *electromagnetic*, *gravitational*, *weak* and *strong*. Of these four, all except gravity can be described by the standard model of particle physics[5] as physical manifestations of local symmetries in a gauge theory that is built from the product of three special unitary groups,

$$\mathrm{SU}(3)_\mathrm{C} \otimes \mathrm{SU}(2)_\mathrm{L} \otimes \mathrm{U}(1)_\mathrm{Y}. \tag{1.1}$$

Each of these local symmetries are associated with a charge-like property of particles and gauge bosons that act as a force carrier. The electromagnetic force is mediated by the photon (γ) and couples to particles that carry electromagnetic charge. The strong force is mediated by eight gluons and couples to particles that carry colour charge. The weak force is mediated by the W and Z bosons and couples to particles that carry weak isospin.

In addition there are three generations of fermions with each containing an up-type quark, a down-type quark, a charged lepton and a neutral lepton. The up and down type quarks have electromagnetic charge of $+\frac{2}{3}$ and $-\frac{1}{3}$ respectively, relative to that of the

Particle			Spin	Electric charge	Mass		
u	c	t	$\frac{1}{2}$	$+\frac{2}{3}$	2.2 MeV	1.275 GeV	173 GeV
ū	c̄	t̄	$\frac{1}{2}$	$-\frac{2}{3}$	2.2 MeV	1.275 GeV	173 GeV
d	s	b	$\frac{1}{2}$	$-\frac{1}{3}$	4.7 MeV	95 MeV	4.18 GeV
d̄	s̄	b̄	$\frac{1}{2}$	$+\frac{1}{3}$	4.7 MeV	95 MeV	4.18 GeV
e^-	μ^-	τ^-	$\frac{1}{2}$	-1	0.511 MeV	105 MeV	1.78 GeV
e^+	μ^+	τ^+	$\frac{1}{2}$	$+1$	0.511 MeV	105 MeV	1.78 GeV
ν_e	ν_μ	ν_τ	$\frac{1}{2}$	0	<2 eV		
$\bar{\nu}_e$	$\bar{\nu}_\mu$	$\bar{\nu}_\tau$	$\frac{1}{2}$	0	<2 eV		
γ			1	0	0		
g			1	0	0		
$W^\pm$			1	± 1	80.4 GeV		
Z			1	0	91.2 GeV		
H			0	0	125 GeV		

Table 1.1: Summary of the properties of fundamental particles in the standard model with the upper half showing fermions and the lower half showing the force mediators (bosons). Values taken from Reference [8].

charged lepton. All known fermions interact via the weak force however only quarks carry colour charge. These properties are summarised in Table 1.1.

1.1.1 Quantum chromodynamics

The $\mathrm{SU}(3)_\mathrm{C}$ term of Equation 1.1 corresponds to the strong force, which is described by the theory of quantum chromodynamics (QCD), and particles that interact with the strong force are said to carry *colour charge*. There exists three types of colour charge denoted by *red*, *green* and *blue* as well as three anti-colours denoted by *anti-red*, *anti-green* and *anti-blue*. The strong force is mediated by gluons that couple to colour charged particles and each gluon carries one unit of colour and one unit of anti-colour resulting in nine possible combinations. However, one of these combinations would be a colour singlet state,

$$\frac{r\bar{r} + g\bar{g} + b\bar{b}}{\sqrt{3}}. \tag{1.2}$$

The existence of a colour singlet would result in the strong force having long range effects between hadrons and such interactions are not observed in nature. The remaining eight gluons form the *colour octet* for which one possible representation is

$$\begin{array}{cccc} \frac{r\bar{b}+b\bar{r}}{\sqrt{2}} & \frac{b\bar{g}+g\bar{b}}{\sqrt{2}} & \frac{-i(r\bar{b}-b\bar{r})}{\sqrt{2}} & \frac{-i(b\bar{g}-g\bar{b})}{\sqrt{2}} \\ \frac{r\bar{g}+g\bar{r}}{\sqrt{2}} & \frac{r\bar{r}-b\bar{b}}{\sqrt{2}} & \frac{-i(r\bar{g}-g\bar{r})}{\sqrt{2}} & \frac{(r\bar{r}+b\bar{b}-2g\bar{g})}{\sqrt{6}} \end{array}.$$

These cannot be combined to form the colour singlet state therefore preventing long distance strong interactions.

1.1.2 Electroweak interaction

The $\mathrm{SU}(2)_\mathrm{L} \otimes \mathrm{U}(1)_\mathrm{Y}$ term of Equation 1.1 represents the unification of the electromagnetic and weak forces; the electroweak force. The symmetries of the $\mathrm{SU}(2)_\mathrm{L}$ term give rise to three gauge bosons, W^1, W^2 and W^3, which interact with particles that carry weak isospin, T. The third component of this, T_3, is always conserved. The $\mathrm{U}(1)_\mathrm{Y}$ term gives rise to the B boson, which is associated with weak hypercharge, and is given by the Gell-Mann–Nishijima relation

$$Y = 2(Q - T_3), \tag{1.3}$$

where Q is the electric charge. At low energy scales the spontaneous symmetry breaking of the higgs mechanism where

$$\mathrm{SU}(3)_\mathrm{C} \otimes \mathrm{SU}(2)_\mathrm{L} \otimes \mathrm{U}(1)_\mathrm{Y} \rightarrow \mathrm{SU}(3)_\mathrm{C} \otimes \mathrm{U}(1)_\mathrm{em} \tag{1.4}$$

causes the observable electroweak force mediators to arise as linear combinations of the four gauge bosons. These are given by

$$W^{\pm} = \frac{1}{\sqrt{2}}(W^1 \pm W^2) \tag{1.5}$$

and

$$\begin{pmatrix} \gamma \\ Z \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} B \\ W^0 \end{pmatrix} \tag{1.6}$$

where θ_W is the weak mixing angle. These are the massive charged W bosons, the massive Z boson and the massless photon (γ). This mechanism also results in the existence of an additional field that gives rise the mass of the massive particles and also predicts the existence of a massive scalar boson, known as the Higgs[9–11]. The existence of this particle was confirmed in 2012 by the ATLAS[12] and CMS[13] collaborations. To date all measurements are in good agreement with standard model expectations for the Higgs boson, completing the standard model.

1.1.3 CP violation

Up until this point there has been no mechanism for quark flavour to be changed in the standard model. For strong and electromagnetic interactions flavour changing interactions do not occur however the weak force does not couple to the mass eigenstates. Instead the weak force couples to a linear combination of these and this superposition is described by the Cabibbo–Kobayashi–Maskawa (CKM) matrix[14] and is given by

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_\mathrm{CKM} \begin{pmatrix} d \\ s \\ b \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \tag{1.7}$$

where q' is the weak “counterpart” of the mass eigenstate, q, and $|V_{ij}|^2$ represents the probability that a freely propagating quark with flavour i decays to a quark with flavour

j. This leads to the constraint that the CKM matrix is unitary,

$$V_{\text{CKM}}^{\dagger} V_{\text{CKM}} = 1, \tag{1.8}$$

where $V_{\text{CKM}}^{\dagger}$ describes the interaction with anti-quarks. The existence of CP violation in weak decays, i.e. that quarks decay at a different rate to their anti-quark counterparts, is well established experimentally. It was first observed in the kaon system in the decays of K_L^0 mesons[15], and later in the b quark system by the Belle and Babar experiments[16, 17]. Very recently CP violation has been discovered in the charm system by the LHCb experiment[18]. These observations result in the constraint

$$V_{ij} \neq V_{ij}^{*} \tag{1.9}$$

and is facilitated by requiring each component of the matrix to be a complex number. From these constraints the CKM matrix can be reduced to three angles; θ_{12}, θ_{13} and θ_{23} as well as a phase term δ. The CKM matrix can then be constructed from three rotation matrices

$$V_{\text{CKM}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13}e^{-i\delta} \\ 0 & 1 & 0 \\ -s_{13}e^{i\delta} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.10}$$

$$= \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix}. \tag{1.11}$$

where c_{ij} is the cosine of angle θ_{ij} and s_{ij} is the sine of angle θ_{ij}.

A commonly used alternative representation of the CKM matrix is the Wolfenstein parameterisation[19] where $s_{12} = \lambda$, $s_{23} = A\lambda^2$ and $s_{13}e^{i\delta} = A\lambda^3(\rho - i\eta)$. This definition has the advantage that all four parameters (λ, A, ρ and η) are of order 1 therefore allowing the magnitude of each element of the CKM matrix to be easily determined. In this parametrisation the CKM matrix is given by

$$V_{\text{CKM}} = \begin{pmatrix} 1-\lambda^2/2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\lambda^2/2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4). \tag{1.12}$$

The current best estimate[8] of the magnitude of the CKM matrix elements are

$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} = \begin{pmatrix} 0.97446 \pm 0.00010 & 0.22452 \pm 0.00044 & 0.00365 \pm 0.00012 \\ 0.22438 \pm 0.0004 & 0.97359^{+0.00010}_{-0.00011} & 0.04214 \pm 0.00076 \\ 0.00896^{+0.00024}_{-0.00023} & 0.04133 \pm 0.0007 & 0.999105 \pm 0.000032 \end{pmatrix} \tag{1.13}$$

with the Wolfenstein parameters being estimated as $\lambda = 0.22465 \pm 0.00039$, $A = 0.832 \pm 0.009$, $\rho = 0.139 \pm 0.016$, $\eta = 0.346 \pm 0.010$. A graphical representation of the CKM matrix

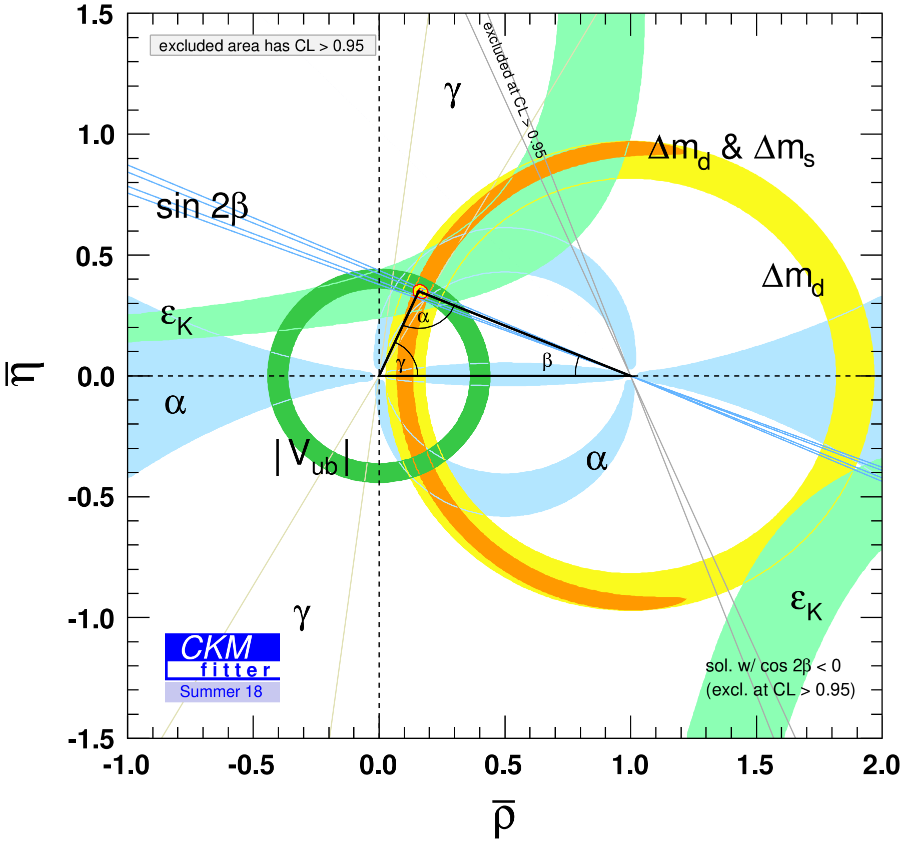


Figure 1.1: Unitary triangle in the $\bar{\rho}/\bar{\eta}$ plane with a global fit to experimental constraints superimposed.[20]

can be obtained using the unitary triangle described by the relation

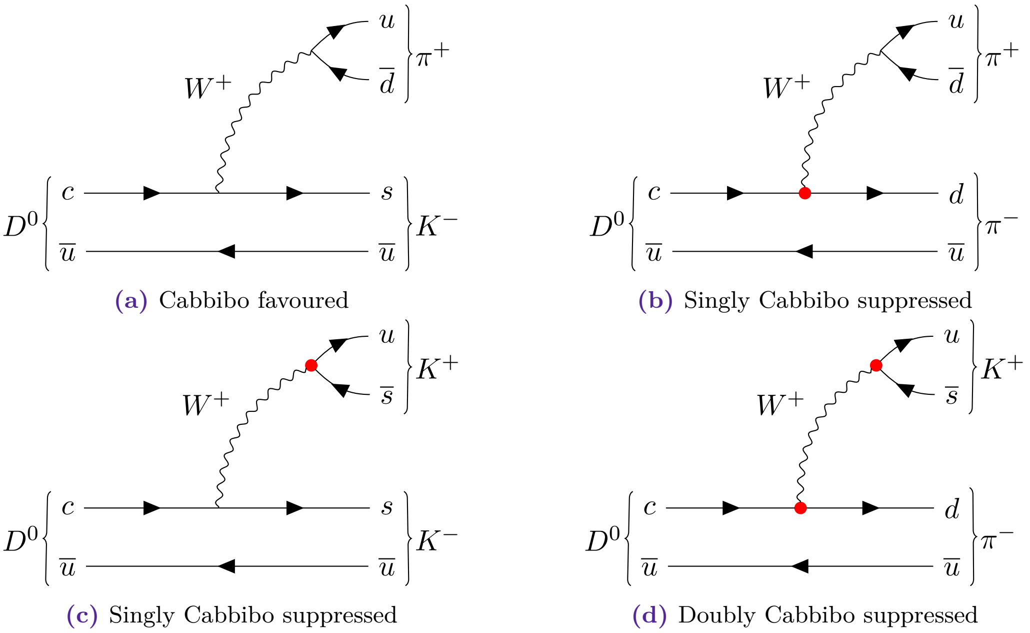
$$V_{ud}V_{ub}^{*} + V_{cd}V_{cb}^{*} + V_{td}V_{tb}^{*} = 0. \tag{1.14}$$

This triangle can be normalised to $V_{cd}V_{cb}^{*}$ to give one side unit length and one such representation is shown in Figure 1.1.

A key feature of the CKM matrix is that the diagonal elements are approximately one therefore the transitions corresponding to off-diagonal elements are heavily suppressed. Decays that use only these diagonal elements in their leading order Feynman diagrams are known as Cabbibo favoured decays. Decays that contain one off-diagonal vertex are said to be Cabbibo suppressed and those with two off-diagonal vertices are said to be doubly Cabbibo suppressed. Figure 1.2 shows an example of D^0 decays to two-body final states at various suppression levels.

1.1.4 Flavour changing neutral currents

Processes that change the generation of a quark without changing its charge are known as Flavour Changing Neutral Currents (FCNC). Such processes are forbidden at tree level. It is however possible for intermediate *loops* consisting of multiple weak interactions to



(a) Cabbibo favoured

(b) Singly Cabbibo suppressed

(c) Singly Cabbibo suppressed

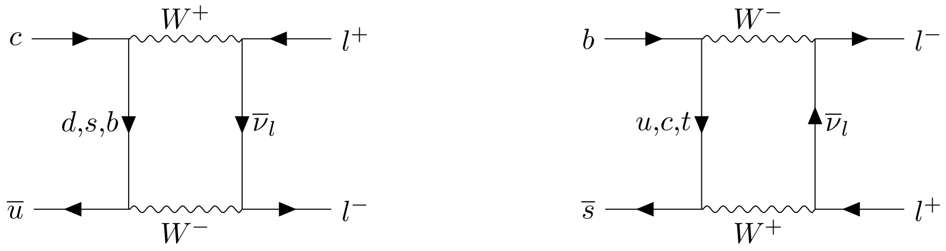
(d) Doubly Cabbibo suppressed

Figure 1.2: Four decays of a D^0 meson to two charged hadrons. The Cabbibo suppressed vertices shown with filled red circles.

result in FCNC via the exchange of a quark of the opposite type, such as those shown in Figure 1.3. In $c \to u$ and $b \to s$ transitions these processes are suppressed due to $V_{cd}V_{ud} \approx -V_{cs}V_{us}$ and $V_{cb}V_{cs} \approx -V_{tb}V_{ts}$ respectively.[1] This is known as the Glashow-Iliopoulos-Maiani (GIM) mechanism[21] and such decays are said to be GIM suppressed. In the standard model these decay rates are approximately proportional to the difference in mass between the quarks. In $b \to s$ transitions this cancellation is less perfect than in $c \to u$ due to the large mass of the top quark[22]. New physics models can increase the rate of FCNC processes by many orders of magnitude making them attractive tests of the standard model[23–25].

Historically FCNC decays have proven useful with the low rate of $K^0_L \to \mu^+\mu^-$[26–28] being used to predict the existence of the charm quark. In the standard model the branching fraction of $K^0_L \to \mu^+\mu^-$ is dominated by an intermediate real two photon state[29, 30] and can therefore only be used to obtain an upper bound on the corresponding FCNC decay. Another rare kaon decay, $K^+ \to \pi^+\nu\bar{\nu}$, is not affected by this QED background has been measured to be compatible with the standard model[31]. More recently $B^0_{(s)} \to \mu^+\mu^-$ decays have been seen as a promising area for new physics searches due to the high precision of the associated theoretical predictions. The more common $B^0_s \to \mu^+\mu^-$ decay has been observed in recent years by the LHCb and CMS experiments[32–34] though the rarer $B^0 \to \mu^+\mu^-$ decay remains elusive[32, 33]. To date all evidence and observations of FCNC decays have been consistent with the standard model and these measurements represent some of the strongest constraints on beyond the standard model contributions in flavour physics.

[1] The contribution from the additional generation of quarks is heavily suppressed from the smallness of $V_{cb}V_{ub}$ and $V_{ub}V_{us}$.



(a) $c \to ull$

(b) $b \to sll$

Figure 1.3: Examples of flavour changing neutral currents.

1.1.5 Rare charm decays

There are possible tensions in B decays between the standard model and experimental measurements in lepton flavour universality with the ratios

$$R_K = \frac{\mathcal{B}(B \to K\mu^+\mu^-)}{\mathcal{B}(B \to Ke^+e^-)} \text{ and } R_{K^*} = \frac{\mathcal{B}(B \to K^*\mu^+\mu^-)}{\mathcal{B}(B \to K^*e^+e^-)}. \tag{1.15}$$

In the standard model these are both predicted to be one with extremely high precision however the results of LHCb each show deviations of approximately 2.5σ[35, 36]. Additionally the P_5' angular observable in $B^0 \to K^*\mu^+\mu^-$ decays shows large local deviations at the 3.7σ and 4.3σ level[37]. A wide variety of potential explanations have been proposed from various new physics models[38–43]. Alternatively, suggestions have been made that the treatment of hadronic effects in standard model prediction may the cause of the apparent discrepancies[44].

The study of charm decays has the potential to offer insight to these discrepancies and such measurements are uniquely positioned to probe for new physics in the up quark sector. Despite the stringent constraints from measurements of kaon and B meson decays new physics contributions are also possible from models that have minimal interplay with down-type quarks[23]. Searches in charm are often hindered by long distance QCD contributions that screen the short distance contributions where new physics is most likely to contribute. These can cause intermediate *resonances* where $A \to BCD$ is predominantly observed as $A \to BX$ with $X \to CD$. The underlying physics process is fundamentally different and the possibility of new physics contributions is heavily constrained by existing measurements. Despite this there are still sizeable regions of phasespace where BSM physics could dominate over the long distance standard model background. In the literature many models have been considered that can generate $c \to ul^+l^-$ transitions[23, 24] including Minimal Supersymmetric Standard Model (MSSM)[45–48], two Higgs doublet models[47], warped extra dimensions[49, 50], little Higgs models[51, 52] and the existence of a up vector-like quark singlet[48].

For $D^+ \to \pi^+l^+l'^-$ decays most of these models predict negligible BSM contributions relative to the that of the intermediate dilepton resonances. Despite this there is room for contributions from MSSM processes at the level of the experimental constraints. After experimental searches reach the limit of the standard model background there remains

	$\mathcal{B}(D^+ \to \pi^+\mu^+\mu^-)$	$\mathcal{B}(D^0 \to \mu^+\mu^-)$	$\mathcal{B}(D^+ \to \pi^+e^\pm\mu^\mp)$	$\mathcal{B}(D^0 \to \mu^\pm e^\mp)$	$\mathcal{B}(D^+ \to \pi^+\nu\bar{\nu})$
i	SM-like	SM-like	$\lesssim 2 \times 10^{-13}$	$\lesssim 7 \times 10^{-15}$	$\lesssim 3 \times 10^{-13}$
ii.1	SM-like	$\lesssim 4 \times 10^{-13}$	0	0	$\lesssim 4 \times 10^{-12}$
ii.2	$\lesssim 7 \times 10^{-8}$	$\lesssim 3 \times 10^{-9}$	0	0	$\lesssim 8 \times 10^{-8}$
iii.1	SM-like	SM-like	$\lesssim 8 \times 10^{-15}$	$\lesssim 2 \times 10^{-16}$	$\lesssim 9 \times 10^{-15}$
iii.2	SM-like	SM-like	$\lesssim 2 \times 10^{-6}$	$\lesssim 4 \times 10^{-8}$	$\lesssim 2 \times 10^{-6}$

Table 1.2: Branching fractions integrated across the full dilepton invariant mass from various leptoquark models (see text). Cases where the standard model contribution dominates are denoted by *SM-like.*[24]

potential for null tests of the standard model to be made[24]. Significant CP violation could be induced near the ϕ resonance and in the high dilepton q^2 region, though these estimates are limited by the unknown strong phases of such decays. Furthermore, a wide range of leptoquark models could have significant contributions to $c \to ul^+l^{(\prime)-}$ where the leptoquark couples up-type quarks to leptons[23, 24]. Table 1.2 shows the potential branching fractions, integrated across dilepton q^2, that could arise from some of these models. Scenario *i* has suppressed couplings for lighter generations, scenario *ii* predominately couples to muons and scenario *iii* favours coupling within the same generation. For *ii* and *iii* the scenarios are split further into case *1* (*2*) where leptoquark is (is not) subject to constrains from experimental measurements of kaon decays. From this work all $c \to ue^+e^-$ transitions are expected to be dominated by standard model contributions and in these models decays to two leptons are often less sensitive than those with an additional hadron in the final state.

Experimentally, rare decays of D^+ and D_s^+ mesons have previously been searched for by the CLEO, BaBar, BESIII, D0, E653, E687, E791, Focus experiments[53–59]. In 2013 the LHCb Collaboration published a search for the 4 decays of the form $D_{(s)}^+ \to \pi^+l^+l^\mp$ that used $1\,\text{fb}^{-1}$ of data to improved upon previous results by approximately 2 orders of magnitude[60]. This result showed the potential LHCb has in this area and in Chapters 5 and 6 of this thesis a new search for the 28 rare charm decays of the form $D_{(s)}^+ \to h^\pm l^+l'^\mp$ is presented.

Chapter 2

LHCb and the LHC

The Large Hadron Collider (LHC) is a 27 km synchrotron at CERN in Geneva, Switzerland that produces proton-proton collisions for the fundamental physics research conducted by the ATLAS, CMS, LHCb and ALICE collaborations[61].

A series of older accelerators from the CERN accelerator complex (Figure 2.1) are used as pre-accelerators to the LHC. Initially, protons are produced using a duoplasmatron[62] and are subsequently accelerated by LINAC2 to 50 MeV. The Proton Synchrotron Booster (PBS) then accelerates the protons to 1.4 GeV before injecting them into the Proton Synchrotron (PS) where they are accelerated to 26 GeV. Finally the protons are accelerated to 450 GeV using the Super Proton Synchrotron (SPS) and injected into the Large Hadron Collider which has a design energy of 14 TeV[1]. The beam of the LHC is made up of 2,808 bunches each containing approximately 1.15×10^{11} protons which corresponds to a bunch spacing of 25 ns around the ring.[2] Space for around 800 bunches is left empty to provide regions in time that the kicker magnets can use to ramp to their full voltage when injecting/dumping the beam.

The Large Hadron Collider beauty (LHCb) experiment is situated approximately 100 m underground at point eight of the Large Hadron Collider and unlike the general purpose detectors, which aim for the highest integrated luminosities possible,[3] LHCb is designed to operate at lower luminosities to allow for precision measurements to be made. This is achieved by fixing the mean number of visible proton-proton interactions,[4] μ, in each bunch crossing through a feedback loop that varies the transverse distance between the beams in order to maintain a constant instantaneous luminosity as shown in Figure 2.2a.

[1] To date the center of mass collision energies used have been 7 TeV, 8 TeV and 13 TeV in 2011, 2012 and 2015 respectively.

[2] While the design bunch spacing is 25 ns the 2010-2012 data taking periods used reduced number of bunches and a 50 ns bunch spacing.

[3] Due to the better than expected performance of the LHC, ATLAS and CMS also decided to use a luminosity levelling scheme to reduce the detector occupancy at the beginning of each fill for part of Run 2.

[4] "Visible" is used to describe the collisions that result in hard interactions rather than elastic scattering and is generally taken to be 69.9 % of the total number of proton-proton interactions[64].

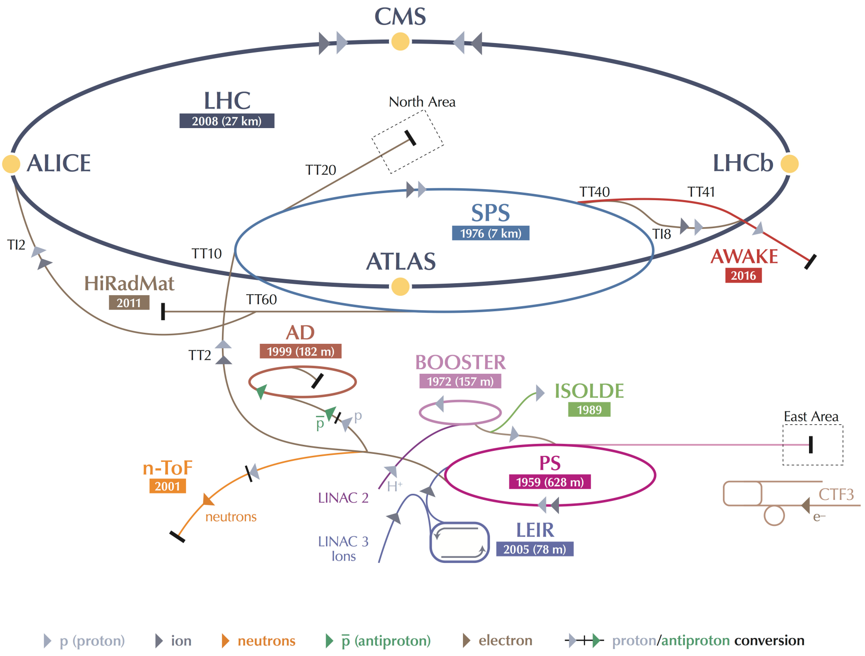


Figure 2.1: The layout of the CERN accelerator complex[63].

2.1 The LHCb detector

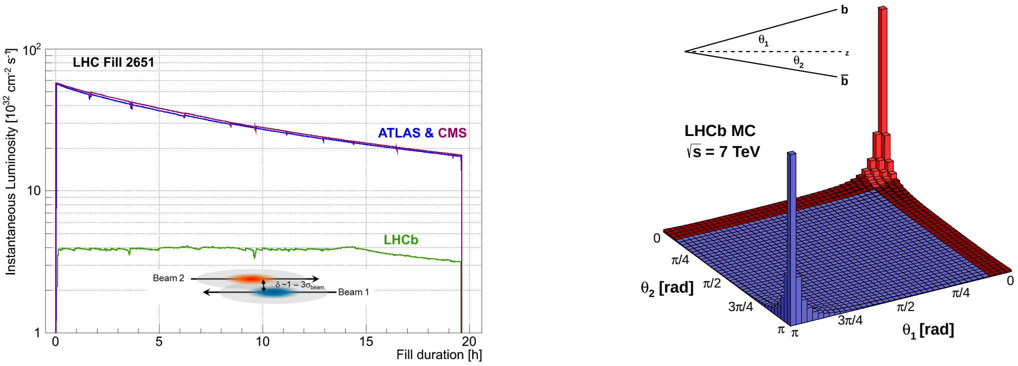
The LHCb detector is a single armed forward spectrometer that is designed to perform precision measurements of beauty and charm quarks. As these quarks are more favourably produced with a small angle to the beam pipe, as shown in Figure 2.2b, therefore the LHCb detector only covers the pseudorapidity[5] range $2 < \eta < 5$ (15 to 300 mrad)[68]. The forward design also allows the majority of the detector to be made in flat planes perpendicular to the beam pipe with most of the readout electronics and physical support structures being kept outside the acceptance of the detector. This reduces the amount of passive material that particles must pass though in order to be detected and helps allow LHCb to have excellent momentum resolution. In addition, LHCb can effectively perform flavour tagging of particles, thereby opening a wide range of opportunities, particularly in the measurement of CP violation.

The detector itself is made up of a vertex locator, five tracking stations, two ring imaging Cherenkov detectors, an electromagnetic calorimeter, a hadronic calorimeter and five muon stations arranged as shown in Figure 2.3.

2.2 Vertex Locator

The LHCb detector's vertex locator (VELO)[69] is unique at the LHC as it can be moved between a distance of 35 mm and 7 mm from the beam of the LHC. This movement is necessary to protect the VELO during the initial injection of protons when the beam is unstable and may deviate from its nominal path. After each fill of the LHC, the position

[5] Pseudorapidity is proportional to the angle between the particle's trajectory and the beam axis and defined as $\eta = \frac{1}{2}\log\left(\frac{|\mathbf{p}|+p_z}{|\mathbf{p}|-p_z}\right)$.



(a) Variation of the instantaneous luminosity at the ATLAS, CMS and LHCb interaction points during fill 2651 in May 2012[65].

(b) Plot of the rapidity distribution of $b\bar{b}$ production. The red region represents the LHCb acceptance and shows that $b\bar{b}$ production is favourably produced at large rapidity[66].

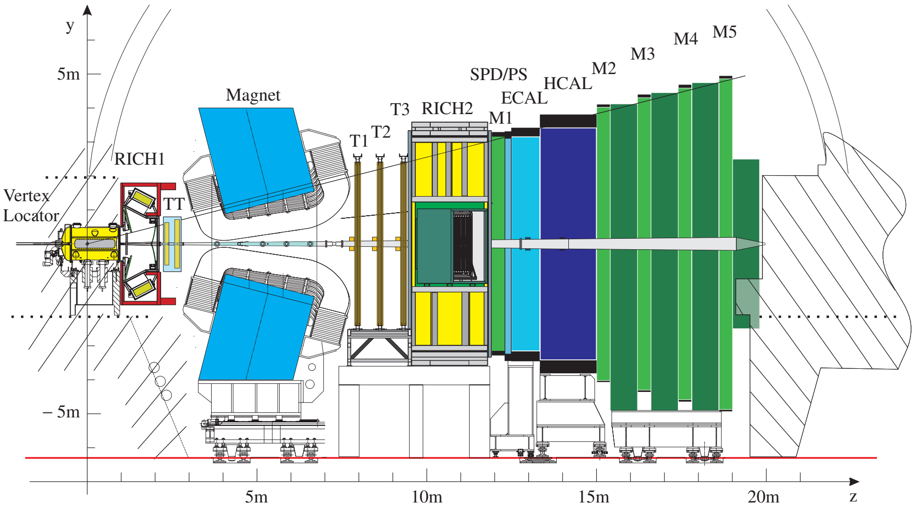


Figure 2.3: Cross-section of the LHCb detector[67].

of the beam is measured and the VELO is manoeuvred into position prior to the start of normal data taking.

The VELO itself is made up of 42 modules of silicon strip detectors with a pitch of 38 µm to 102 µm that varies linearly from the beam edge. Each module uses two sensors and provides a measurement of both the r and ϕ coordinates and are arranged as shown in Figure 2.4 to optimise the reconstruction performance. The close proximity of the VELO to the LHC beam allows for an excellent primary vertex (PV) resolution of 13 µm in the transverse plane and 71 µm in the axis parallel to the beam[70]. In addition, thanks to its close proximity to the interaction point, the VELO is also able to measure the flight distance of B-mesons, which have a typical lifetime of 1.5×10^{-12} s, opening up opportunities for precision lifetime measurements and the study of processes such as B_s^0 mixing[71].

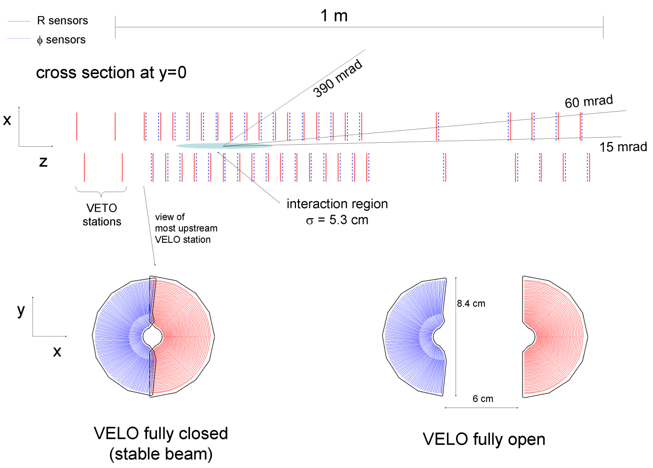


Figure 2.4: Top: Cross-section in the x-y plane of the LHCb VELO. Below: Example of two modules in the close and open position where the left, blue, half measures the angular component and the right, red, half measures the radial position.

2.3 Silicon Tracker

The Silicon Tracker is comprised of two parts; the Tracker Turicensis (TT) and the Inner Tracker (IT). Both parts are comprised of four vertical silicon microstrip detectors with a strip pitch of around 200 µm in a x-u-v-x layout where the inner two layers, u and v, are rotated by $-5°$ and $5°$ respectively. This configuration has a reduced precision when measuring the y position of particles, however this is less important for momentum measurements as the magnet predominantly bends tracks in the x plane. The TT is located upstream of the magnet and covers the full acceptance of the detector. In contrast to this the IT is placed downstream of the magnet and only covers the innermost region of the acceptance where the occupancy is greatest. The remainder of the acceptance is measured using the Outer Tracker as described in the following section. In all elements of the Silicon Trackers the length of the silicon strips is varied to minimise the expected occupancy in any given strip, without using an excessive number of readout channels in the lower occupancy regions of the detector.

2.4 Outer Tracker

The Outer Tracker (OT)[72] is a drift-time detector and is comprised of around 55 000 hollow tubes with an inner diameter of 4.9 mm containing a gas mixture and a thin wire in the centre. When a charged particle enters the tube the gas mixture is ionised resulting in the delocalisation of electrons that are then attracted towards the charged wire in the centre. As the electron drifts, a phenomenon known as Townsend discharge[73] occurs increasing the number of electrons to a level where they can be detected by electronics at the end of the wire. Rather than being limited to the spacing of the straws, the resolution can be improved by measuring the drift time of the electric charge relative to the interaction

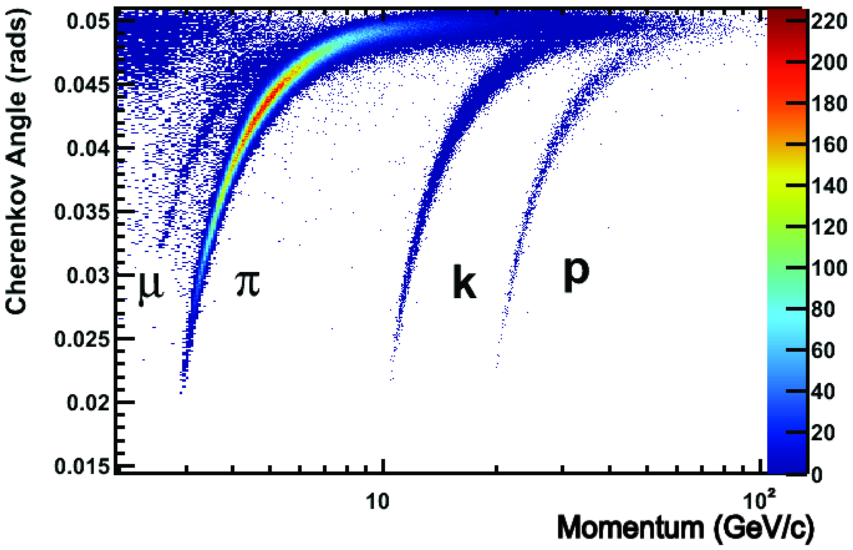


Figure 2.5: Plot of Cherenkov angle against momentum for 2 % of the real data taken in 2011 at $\sqrt{s} = 7\,\text{TeV}$ in the C_4F_{10} radiator of RICH 1[65].

time to establish the distance from the wire. This technique allows a precision of around 200 µm to be achieved. The gas mixture is chosen such that the maximum drift time is 50 ns to minimise spillover from other bunch crossings.[6]

2.5 Ring Imaging Cherenkov Detectors

RICH detectors contain a medium with refractive index (η) slightly greater than 1, therefore when charged particles pass through the detector, Cherenkov radiation is emitted at an angle, θ_c, given by

$$\cos{(\theta_c)} = \frac{c}{\eta v}$$

As this is related to the velocity of the particle, v, this can be combined with a momentum measurement from the tracking detectors to give a mass hypothesis. Plotting track momentum against Cherenkov angle, as in Figure 2.5, results in the formation of bands, each of which corresponds to a different species of particle.

In LHCb, two mirrors are used to reflect and focus this light onto Hybrid Photon Detectors (HPDs) that are placed outside of the LHCb acceptance. As pions are the most commonly produced particle at the LHC, a log-likelihood algorithm is used to perform a ratio likelihood test of whether a candidate is an electron, kaon, muon or proton against the likelihood that the candidate is a pion. Below threshold (9.3 GeV for kaons and 17.8 GeV for protons in RICH 1) the refractive indices of the RICH detectors are too small to produce Cherenkov radiation and the log-likelihood must instead be calculated in veto mode, i.e. the likelihood that this track is not a pion[74].

[6] For 25ns data taking the hits in the OT from the previous and following bunch crossing are used during reconstruction due to the drift time having some potential overlap.

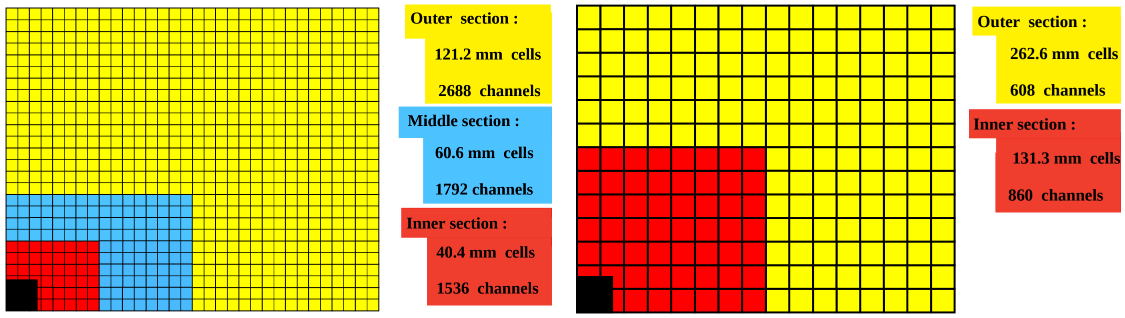


Figure 2.6: Construction of the upper right quadrant of ECAL (left) and HCAL (right). The black section in the innermost region represents the cut out for the LHC beampipe.[68]

2.6 Calorimeters

Calorimetry is used to measure the energy and position of electrons, photons and hadrons as well as providing signal for them at the lowest level of the LHCb trigger, prior to any tracking considerations. Most calorimeters follow the principle that scintillation light, that is, light emitted due to the presence of ionising radiation, is measured by photon detectors. In LHCb, wavelength shifting fibres are used to convert these photons into the spectral range of Multianode Photon Multiplier Tubes (MaPMTs). The first layer of the LHCb calorimeter is the Scintillating Pad Detector (SPD) that is used to distinguish between charged and neutral particles as they enter the calorimeter. This is then followed by the PreShower (PS) that distinguishes between electrons, photons and pions. These are both primarily used to provide a signal for the trigger. The Electromagnetic Calorimeter (ECAL) measures the transverse energy of electrons, photons and neutral pions and is used in the reconstruction of such particles, whereas, the Hadronic Calorimeter (HCAL) is mostly used to provide a transverse energy measurement for triggering purposes[75].

As shown in Figure 2.6, the electromagnetic calorimeter is split into three regions, with the granularity increasing as the detector gets closer to the beam pipe. This design ensures the occupancy remains acceptable, while reducing the cost of the large outer region. In this thesis the ECAL is used in Section 5 and 6 for reconstructing bremsstrahlung photons that are radiated by electrons as they travel through the detector material. Electrons are most significantly affected as they are the least massive charged particle and the power emitted is proportional to m^{-6}. The direction of the emitted photons is the same as the direction of electron's momentum at the time of emission. As the LHCb magnet contains effectively no material inside the detector acceptance this results in two possible regions in ECAL that could contain bremsstrahlung photons, as shown in Figure 2.7.

2.7 Muon system

The LHCb muon system[77] is made up of 1380 multi-wire proportional chambers (MWPCs) that are equally distributed between the 5 muon stations. Each MWPC is comprised of four (two in M1) 5 mm deep gaps containing a mixture of carbon dioxide, argon, and

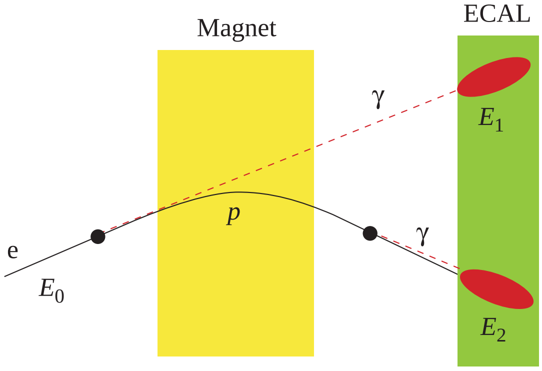


Figure 2.7: Depiction of an electron passing through LHCb, showing how bremsstrahlung is predominantly emitted in regions of the detector where there is effectively no magnetic field in the same direction as the electron's momentum[76].

CF_4. One side of each cell is sub-divided into regions that are used as cathodes. The anodes are formed of wires running down the centre of each gap with a spacing of 2 mm. The size of each MWPC is varied between muon stations and along rapidity to allow for approximately uniform occupancy. When charged particles pass through the muon stations, Townsend discharge occurs as in the straws of the outer tracker.

In the innermost region of M1, the occupancy is too high for multi-wire proportional chambers. Therefore, a triple gas electron multiplier (GEM) technology is used. GEMs use an insulating layer of Kapton foil that is clad with copper on both sides. Many small holes are then etched through the cladded foil. During operation a voltage is applied across the foil and if an electron enters one of the holes an avalanche of electrons is triggered. In LHCb three layers of these foils are placed in a gas pocket, similarly to the MWPCs, with a cathode placed after the third layer to detect the electrons released by the presence of a muon.

In order to allow the momentum of a track to be efficiently estimated in the Level 0 trigger, M2 to M5 are interleaved with 80 cm thick iron absorbers. The minimum momentum required to cross all five muon stations is around 6 GeV.

2.8 Particle Identification

Different species of particle leave signatures in different elements of the detector, as shown in Figure 2.8. In some cases these signatures are unambiguous. For example, if a energy deposit is found in the electromagnetic calorimeter with no corresponding track or HCAL energy deposit, it is almost certainly a photon. However, for most species of particle there is always some ambiguity. To resolve this, Particle Identification (PID) is used for distinguishing between long lived particles that have similar characteristics in the detector such as; protons, pions and kaons or neutral pions and photons. The main elements of LHCb for distinguishing these particles are the two ring-imaging Cherenkov detectors and the calorimeters.

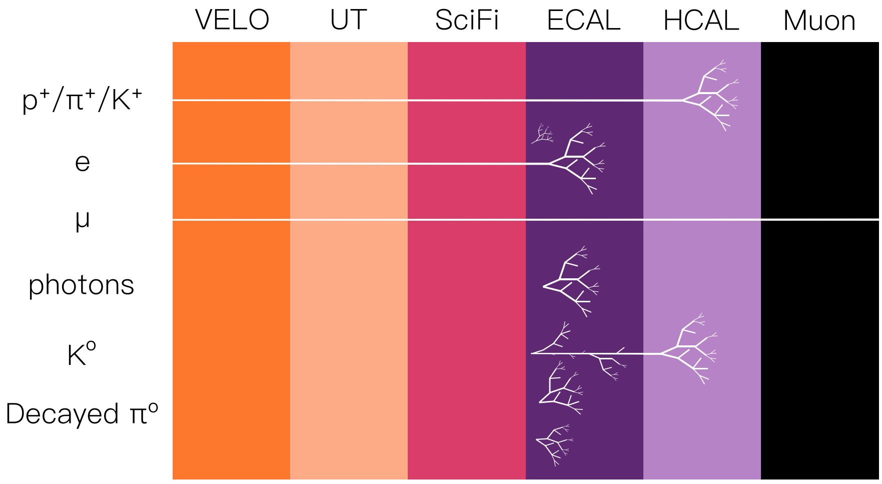


Figure 2.8: Signature left in LHCb for various species of particle. For electrons there can be multiple energy deposits in ECAL due to bremsstrahlung photons. Neutral pions are either detected directly or as two energy deposits, depending on if the π^0 decays to two photons before reaching ECAL.

In LHCb, each subdetector provides a likelihood for each particle species it is able to discriminate, such as the RICH detectors described in Section 2.5. To obtain better performance, the inputs from each element of the detector are combined to obtain global PID variables. One method computes a “combined delta log likelihood” (DLL) that is the sum of the log likelihoods from each detector, relative to the pion hypothesis. This method assumes all information can be presented as a likelihood and can result in unusual features, such as for particles whose momentum is below the pion identification threshold in the RICH.

To obtain better performance, a second class of global PID variables are computed, that use a multivariate classifier to simultaneously predict the true species of a particle that left a given signature in the detector. The most established technique for this in LHCb uses an artificial neural network to simultaneously compute a probability value for each species from a given detector signature and are known as `ProbNN` variables. Improved classifiers using popular tools such as `XGBoost` and `Keras` are under investigation, see Reference [78] for more details.

2.9 Simulation

Most analyses in high energy physics rely on Monte Carlo simulated events as an input to allow one to access aspects of the event that cannot be measured in real data. In LHCb simulated events are normally generated using `PYTHIA 8`[79] with a specific LHCb configuration[80]. The decays of hadronic states are then simulated using `EvtGen`[81] and final state radiation is modelled using `PHOTOS`[82]. Detector effects and interactions are implemented using `GEANT4`[83] as described in Reference [84].

2.10 Trigger

The trigger system of the LHCb detector is comprised of three stages[85, 86]. The first of these, known as L0, is implemented in hardware and used to reduce the rate from 40 MHz down to 1 MHz. The Level 0 trigger is subdivided into three classes:

- **Muons** The L0 Muon triggers search for combinations of hits in the muon chambers that are consistent with originating from the same track. The transverse momentum is then estimated using the slope of the track and a minimum requirement is made to reject muons originating from pion/kaon decays.
- **Charged hadrons** The L0 trigger for charged hadrons searches each region of four adjecent cells in the hadronic calorimeter to find the cluster with the most transverse energy, which is then required to be larger than a specified threshold.
- $\boldsymbol{\pi^0/\gamma/e^\pm}$ To identify neutral particles and electrons the electromagnetic calorimeter is used to identify deposits with large transverse energy in a similar way to the charged hadron trigger. Additionally the preshower is used to tag each cluster as either a π^0, photon or electron. The candidate with the most transverse energy is then read out for each of the three particles.

If an event is selected by L0, it is then processed by the first of two software level triggers, known as HLT1 and HLT2, where a partial track reconstruction is performed. If a well fitted track is found that also passes a transverse momentum requirement, the event is accepted and then processed by HLT2 where a full reconstruction is performed and compared to many analysis dependent selections, such as those described in Section 5.3.1.

Finally, if an event passes one or more HLT2 selections, the event is transferred away from the detector for permanent storage. This is then periodically processed with a higher quality reconstruction as part of centralised processing campaigns. The results of this reconstruction are then made available to physicists after passing an additional set of constraints as described in Section 2.11.

2.10.1 Turbo stream

As a result of work during long shutdown one (2013-2015) of the LHC, it was possible to unify the trigger and offline reconstruction software thanks to significant optimisations, and upgraded hardware that provides twice the computing capability of Run 1. In order to fully utilise this new capability the detector must be aligned prior to starting HLT2. To allow this to happen the alignment procedure of each part of the detector has been modified to start once enough data has been collected, with the resulting constants being applied automatically as required[87]. While the alignment is being performed events accepted by HLT1 are stored in a 5 PB buffer. This typically takes a few minutes however the buffer is able to store up to two weeks of data ensuring that data is not lost if problems occur with calibration procedure. This also has the added benefit of allowing events to be processed between fills of the LHC, further increasing the computing time that can be allocated to each event.

As a result of these changes the reconstruction performed by the trigger is now of equal quality to that previously obtained offline and can be used for analysis. To reflect this, a new output "stream" has been added known as the turbo stream.[88] This utilises this new capability by only storing information that is relevant for analysis and reduces the event size by a factor of up to 14. This has allowed the output rate of LHCb to be increased by 25 % while minimally increasing the bandwidth used. The disadvantage is that the full raw data is not kept and hence cannot be re-reconstructed in case of unforeseen issues. This system was able to produce two publications[1, 89] using data from the 2015 50 ns ramp very quickly after the data was collected. The author of this thesis was a lead proponent of one of these analyses. The preparations of this analysis were reported in his MPhys thesis[61] and completed early in the author's PhD. This analysis has already been described in two theses[3, 4] and hence for brevity is not covered here.

2.10.2 Computing resources

Modern high energy physics research is computationally intensive and most of the computing resources used are made available as part of the Worldwide LHC Computing Grid (WLCG)[90]. This allows resources to be shared between different experiments, resulting in more optimal utilisation of resources. The resources used to process LHCb data are split into four tiers:

- **Tier 0**: All LHC data passes through the CERN datacenter in Geneva, Switzerland and the Wigner Research Centre for Physics in Budapest, Hungary[7] to the Tier 1 centres.
- **Tier 1**: Seven large computing centres, each in a different country, have dedicated optical links to the Tier 0 sites.
- **Tier 2**: Smaller computing centres at universities and other scientific institutions that provide storage and processing resources. As of 2019 there are around 90 Tier 2 sites used by LHCb.
- **Tier 3**: Computing resources that are not pledged to WLCG, including clusters from university departments and physicist's personal computers.

Use of the Tier 0, 1 and 2 sites within LHCb is managed by a framework that is developed and maintained by the collaboration named `Dirac`[91]. This has also been adopted by other collaborations including the Belle 2 experiment[92], BES-III[93], the International Linear Collider community[94] and many others. `Dirac` manages compute resources and data storage and abstracts away differences in the infrastructure used by each computing site. Additionally, `Dirac` manages the scheduling of both central and individual jobs, assigning them to available resources while also ensuring data locality to minimise network congestion. The three tiers of WLCG site are not generally distinguished when processing data or running Monte Carlo simulations.

[7] This facility is currently being decommissioned.

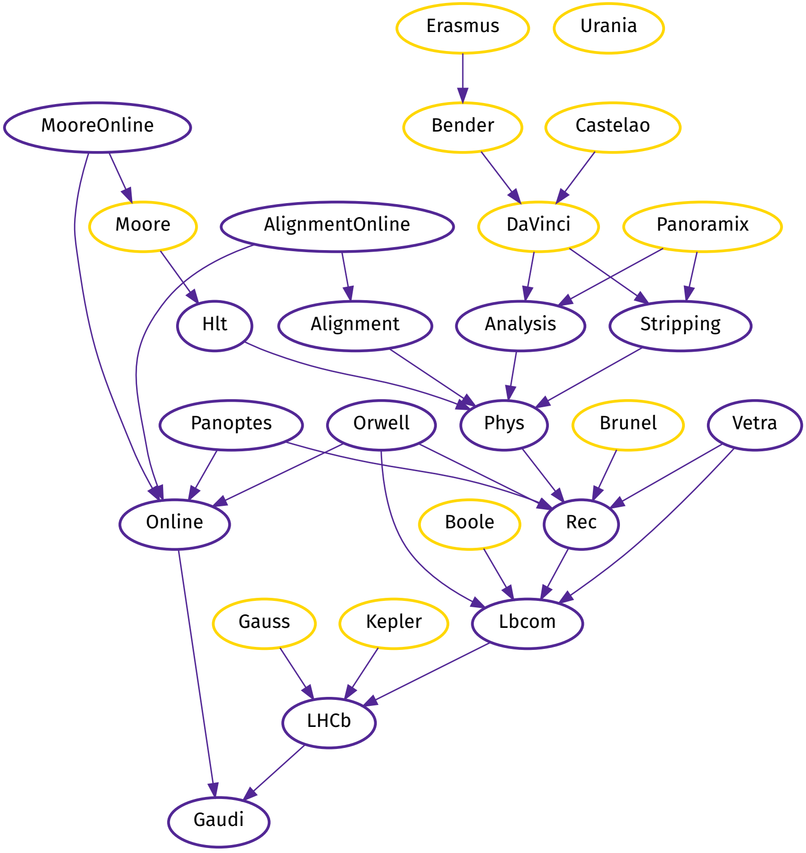


Figure 2.9: Dependency graph showing the main software packages maintained by the LHCb collaboration. High level applications that provide environment frequently used by physics analysts are circled in gold.

While user jobs can be submitted directly to Dirac, it provides no facility for the management and grouping of jobs created by users. An independent project, Ganga[95], provides users with an interactive IPython[96] prompt that can be used to manage groups of jobs. It supports a wide range of backends such as Dirac[91], LSF[97], Condor[98], ARC[99], CREAM[100] and the local machine, making it suitable for submitting to all WLCG and Tier 3 resources.[101]

2.11 Software stack

Software for the LHCb experiment is split into around 20 packages, each of which is stored in a separate Git repository on CERN's GitLab instance. The names and dependencies of most applications are shown in Figure 2.9.

The LHCb software framework is based upon the experiment independent Gaudi framework[102] that provides generic implementations of interfaces and services that are required processing events in HEP experiments. Two libraries, LHCb and Lbcom, build upon Gaudi to provide LHCb specific classes; such as those used for the detector geometry.

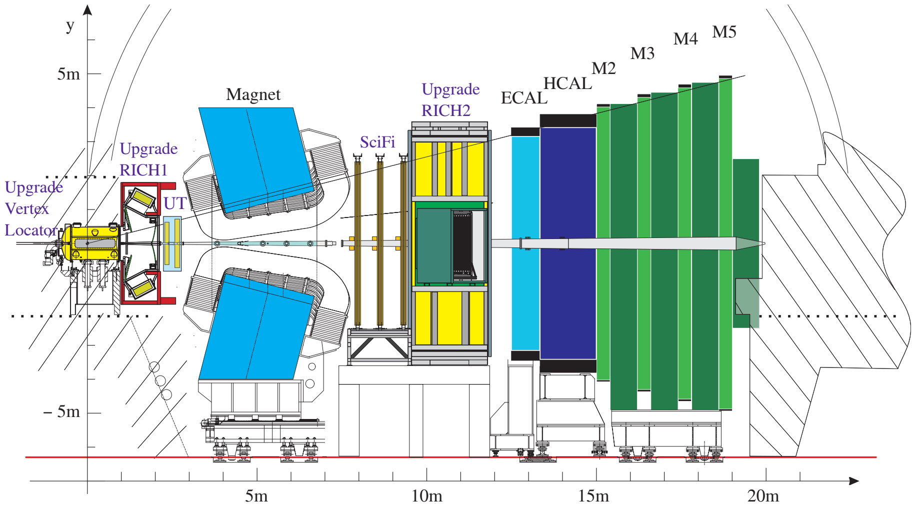


Figure 2.10: Cross-section of the LHCb Upgrade detector[67]. Detectors that are entirely replaced are written in purple.

A high level application, `Brunel`, is used to perform offline reconstruction as part of centralised processing campaigns on the Worldwide LHC Computing Grid.[90] Since the start of Run 2, an identical reconstruction is also performed in the high level trigger application, `Moore`. To allow for this the that code used for the reconstruction of LHCb events to be shared between the two applications it is developed as an independent library named `Rec`. Two additional libraries, `Phys` and `Analysis`, contain software used for physics analyses. These are most commonly used via the `Moore` and `DaVinci` applications.

Data is reconstructed using `Brunel` is filtered using `DaVinci` in centralised campaigns, in a process known internally as the `stripping`. The newer turbo stream output of `Moore` has a lightweight transformation applied using `Tesla` to produce files that are similar to those produced by the stripping. The output of these centralised campaigns is typically further filtered by individuals using `DaVinci` to produce `ROOT`[103] files that contain `TTrees` of signal candidates and their properties. It is possible to work with the stripping output directly, though this is only done by a minority of users. The `Bender` application can be used to make working interactively with these files easier however, this can also be done with any of the applications.

Monte Carlo simulations for LHCb are generated using a package named `Gauss` as described in Section 2.9. The output of `Gauss` undergoes a digitisation process, where the detector response is simulated using an application named `Boole`. The output of `Boole` can then be processed using the aforementioned software that is used for real data.

Additionally, there are many specialised applications that are part of the LHCb software stack but not described here.

2.12 The LHCb Upgrade

In a decade of running, the LHCb detector has collected $10\,\mathrm{fb}^{-1}$ of luminosity with an efficiency of over $90\,\%$ and, at the end of 2018, Run 2 of the LHC concluded and Long

Shutdown 2 (LS2) began. This marks the end of the useful lifespan of the current detector and during the shutdown most of the LHCb detector will be replaced[104] to allow for the detector to be operated at an order of magnitude higher instantaneous luminosity. The sensitive elements in all subdetectors, except the ECAL, HCAL and muon stations, will be replaced with higher resolution and more radiation tolerant designs that will allow the detector to operate effectively when the average bunch crossing contains 7 inelastic proton-proton collisions.

The VELO will be replaced with a hybrid pixel detector and is discussed in more detail in Section 2.12.1. An Upstream Tracker (UT)[105] made up of silicon strip detectors will replace the TT to provide tracking between the VELO and magnet. The tracking stations after the magnet (IT and OT) are to be replaced with a Scintillating Fibre tracker (SciFi)[105] that is made of three tracking stations. Each station contains four layers, in the same stereo angle configuration as is present in the current detector's tracking stations. For the RICH detectors[106], the HPDs are to be replaced with Multianode Photon Multiplier Tubes (MaPMTs) that have a higher resolution. The RICH2 mechanical structure is kept however RICH1 is entirely replaced. The mirrors used for focusing the Cherenkov light are also replaced with a further optimised arrangement. The gas mixture of RICH1 and RICH2 will remain the same. To make room for the larger SciFi detector, the PS, SPD and M1 detectors will be removed, allowing RICH2 to be placed closer to ECAL.

In addition to replacing the sensitive detector elements, the data acquisition (DAQ) system and readout electronics from all subdetectors are to be entirely replaced. This will allow for the entire detector to be read out at 40 MHz, i.e. for every LHC bunch crossing in LHCb. The L0 hardware trigger will also be removed, allowing for an all software trigger scheme. This results in significant improvements for areas of the LHCb physics programme such as charm physics, where the final state hadrons are often soft and the current L0 trigger is inefficient. The topology of the high level trigger will remain similar, albeit with entirely new hardware and software to facilitate the increased data rates.

2.12.1 VELO Upgrade

The LHCb Upgrade VELO[107] is comprised of 52 modules, each containing 12 256×256 pixel sensors with a pitch of $55\,\mu\text{m} \times 55\,\mu\text{m}$[8] that are bump bonded onto a custom Application Specific Integrated Circuit (ASIC) that is capable of 40 MHz readout[108]. The sensors are arranged as four, 3x1 sensor ladders with two being placed on each side of the module as shown in Figure 2.11. Each sensor is glued to the 500 µm thick silicon substrate.

Each module has to be able to dissipate 30 W of heat while keeping the sensors below $-20\,^{\circ}\text{C}$. This low temperature is required to ensure the sensors remain functional in the high radiation environment. As the module is operated in vacuum, the heat must be removed along the supports of the module. This is achieved with evaporative CO_2 cooling in microchannels that are etched inside the silicon substrate, as shown in Figure 2.12. Liquefied CO_2 is pumped into the channels at around 20 bar. As it passes under the heat

[8] The pixels are elongated at the edge of each sensor to improve the efficiency between each ASIC.

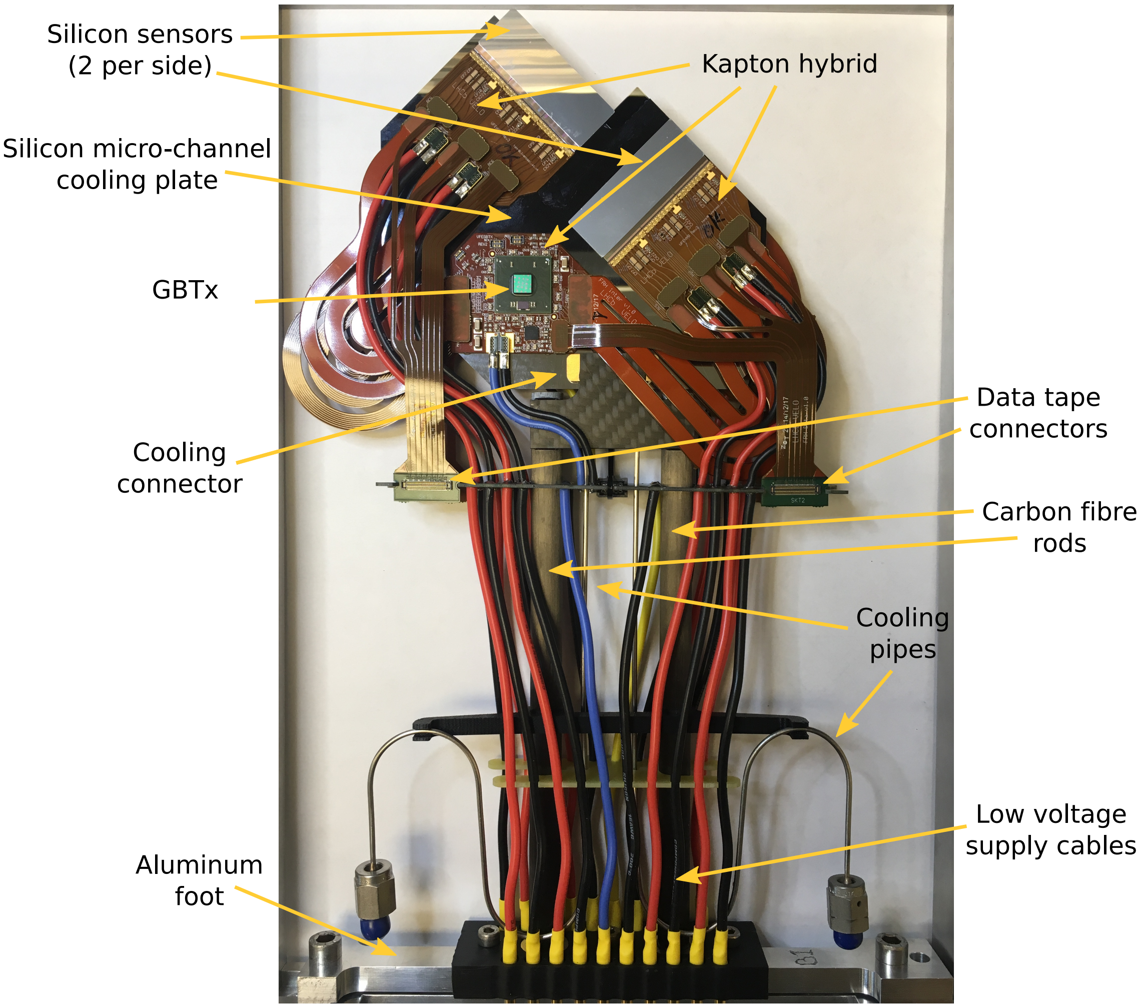


Figure 2.11: Diagram of a prototype upgrade VELO module that was manufactured at the University of Manchester with the main components labelled.

producing components of the module the CO_2 boils, absorbing heat from the component that is then removed as the CO_2 flows back out of the substrate.

To connect the substrate to the main detector assembly, an aluminium cooling block is soldered to the substrate. This also provides the connection to the cooling pipes that provide the high pressure CO_2. The cooling block is connected to carbon fibre supports that are mounted on an aluminium cooling block that provides a connection to the edge of the VELO's vacuum vessel. This design is susceptible to moving when cooled and the potential impact of these movements is described in Section 4.

2.12.2 Software improvements

Devloping an all software trigger for filtering LHC collisions is an extremely challenging undertaking. The 30 MHz of filled bunch crossings must all be processed in near realtime to avoid filling the available buffer space in the trigger. It is expected that the upgrade trigger will use around 1000 compute nodes and this means each node must be able to fully process an average event within 3.5 µs.

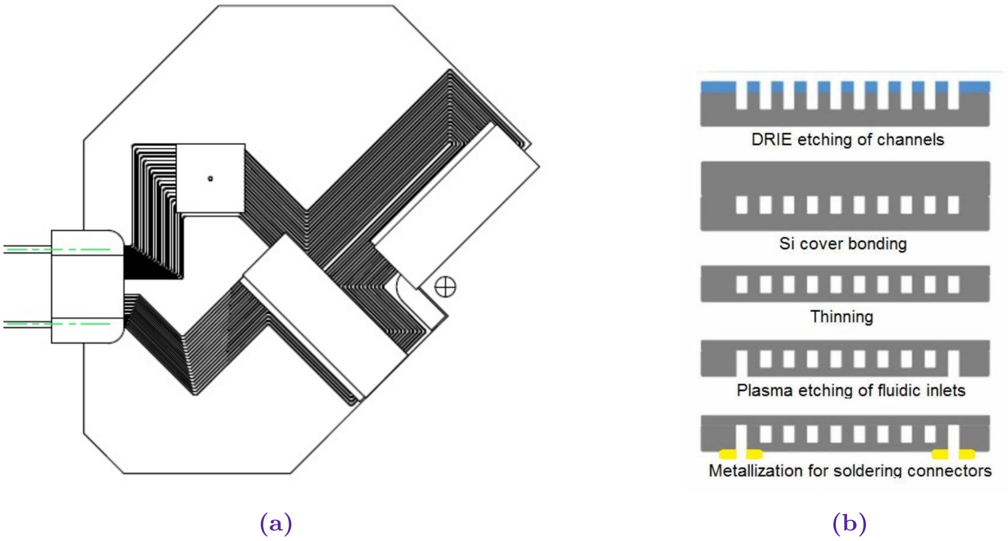

DRIE etching of channels
Si cover bonding
Thinning
Plasma etching of fluidic inlets
Metallization for soldering connectors
(a)
(b)

Figure 2.12: (a) Positions of the microchannels etched into the module's silicon substrate, chosen such that each channel has similar length and fluidic resistance while providing cooling to the GBTx and VeloPix ASICs. (b) Manufacturing process for the microchannel substrate. The silicon wafer is etched then bonded to a second wafer to embed the channels inside the substrate. This assembly is then thinned and holes are plasma etched and metallised to facilitate the connection of the cooling block.

As can be seen in Figure 2.13, the processor frequency and single threaded performance stagnated in the early 2000s. Fortunately the increase in the number of transistors has continued to follow Moore's Law[109] and grow exponentially, with the additional transistors mostly being used for new `SIMD` instructions as well as adding multiple processor units to a single CPU package.

Single Instruction Multiple Data (`SIMD`) is a class of computation that is analogous to matrix operations, where a single operation is applied to multiple values at the same time. In a modern CPU, the number of data points that can simultaneously be operated varies from 4 to 16 and this can correspond to a equivalent speed up provided a suitable algorithm is available. GPUs and FPGAs are an expansion of this paradigm, with them ideally operating on many thousands of data points simultaneously. Contrastingly, multi core processors are equivalent to attaching several single core processors together. This allows them to each perform completely independent operations, at the cost of the processor being more complex and expensive to construct. There are two options when writing software to make use of multi core processors:

- **Multi process**: Equivalent to running multiple independent programs at the same time. Each process is only able to perform one operation at a time and has its own pool of memory.
- **Multi thread**: A single process uses multiple threads that are able to perform operations synchronously. A single pool of memory is shared between the threads.

The current LHCb software framework utilises multi core CPUs by using multiple independent `Gaudi` processes. This has the advantage of being able to use the existing

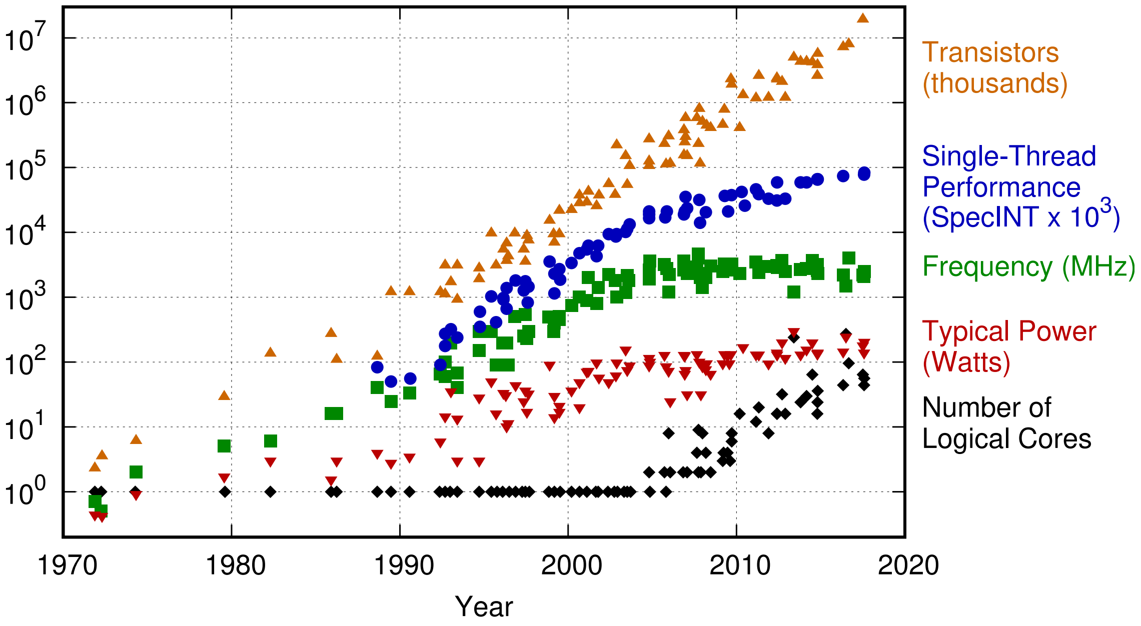


Figure 2.13: Specification of commercially available computer processors as a function of their release date. Taken from Reference [111].

framework that was designed before mainstream multi core processors were announced. As the number of available cores in each node increases, so too must the available memory to allow for each process to have its own memory pool. A significant fraction of this data is duplicated between each process, such as the detector geometry.[9] Additionally, the high memory usage can result in worse performance by causing data to be evicted from the `L1`/`L2`/`L3` caches of the CPU, therefore reducing its effectiveness. To avoid this issue, `Gaudi` is moving to a new functional framework where all algorithms must explicitly declare their data dependencies instead of using memory that is globally available. This allows a single `Gaudi` process to efficiently analyse multiple events simultaneously. Investigations are also ongoing to assess to suitability in LHCb of alternative microarchitectures, such as `arm64`, or special purpose hardware, such as GPUs or FPGAs.[101, 110]

[9] While it is also possible to make use of shared regions of memory between multiple processes, there are many limitations when using this model.

Chapter 3

Alignment

Many experiments in high energy physics depend on being able to accurately measure the trajectories of particles as they traverse through detectors. These measurements are then used to calculate a wide range of physical quantities such as momenta, lifetimes and angles. In order to allow these quantities to be reconstructed with the greatest precision, the position of the detector elements must be precisely known. The process of determining these positions is known as *alignment*. This chapter will start by introducing how track reconstruction is performed, before giving an overview of some of the available methods that can be used for aligning tracking detectors. Finally, it will be shown how these alignment methods have been applied to data collected during LHCb testbeam campaigns.

3.1 Tracking methods

Most experiments in HEP detect charged particles using sensitive elements that give point measurements of the particle's position. Several of these measurements can be combined to reconstruct a *track* that approximates the true path of the particle.

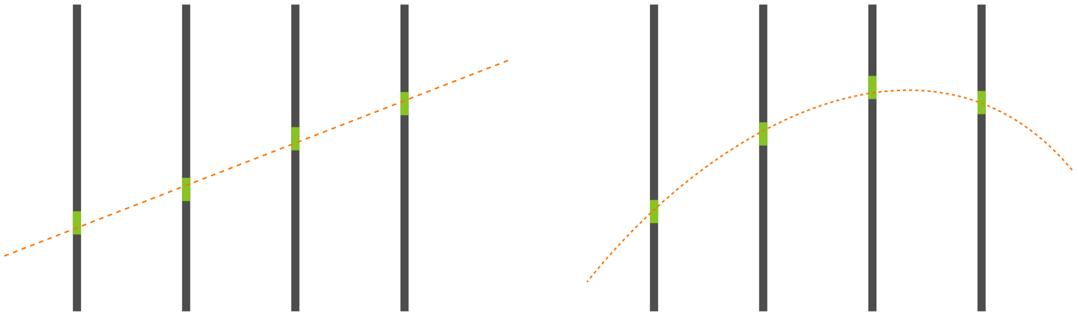
Figure 3.1a shows the simplest scenarios where there is no magnetic field present and the track can be approximated by a straight line. In this situation a least squares fit can be used to find the parameters that describe the track. To assess the quality of a track, χ^2_{track} can be computed using

$$\chi^2_{\text{track}} = \sum_{i}^{\text{clusters}} \left[\frac{(x_i - (t_{x0} + t_x z))^2}{\sigma^2_{x_i}} + \frac{(y_i - (t_{y0} + t_y z))^2}{\sigma^2_{y_i}} \right] \tag{3.1}$$

where x_i/y_i is the position of the measured cluster, $\sigma_{x_i}/\sigma_{y_i}$ is the uncertainty of the measured position, t_x/t_y is the slope of the track and t_{x0}/t_{y0} is the track intercept.

In order to measure the momentum of a particle, many detectors are placed in or near a strong magnetic field. This causes the trajectories of charged particles to become curved, as shown in Figure 3.1b, due to the *Lorentz force* given by

$$\overrightarrow{F} = q\overrightarrow{v} \times \overrightarrow{B} \tag{3.2}$$



(a) Without magnetic field

(b) With magnetic field

Figure 3.1: Example detector response (green) from a charged particle (orange) traversing through.

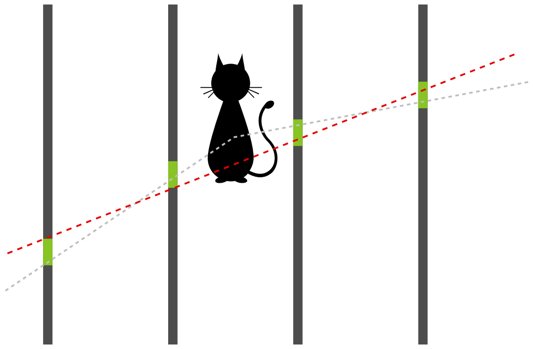


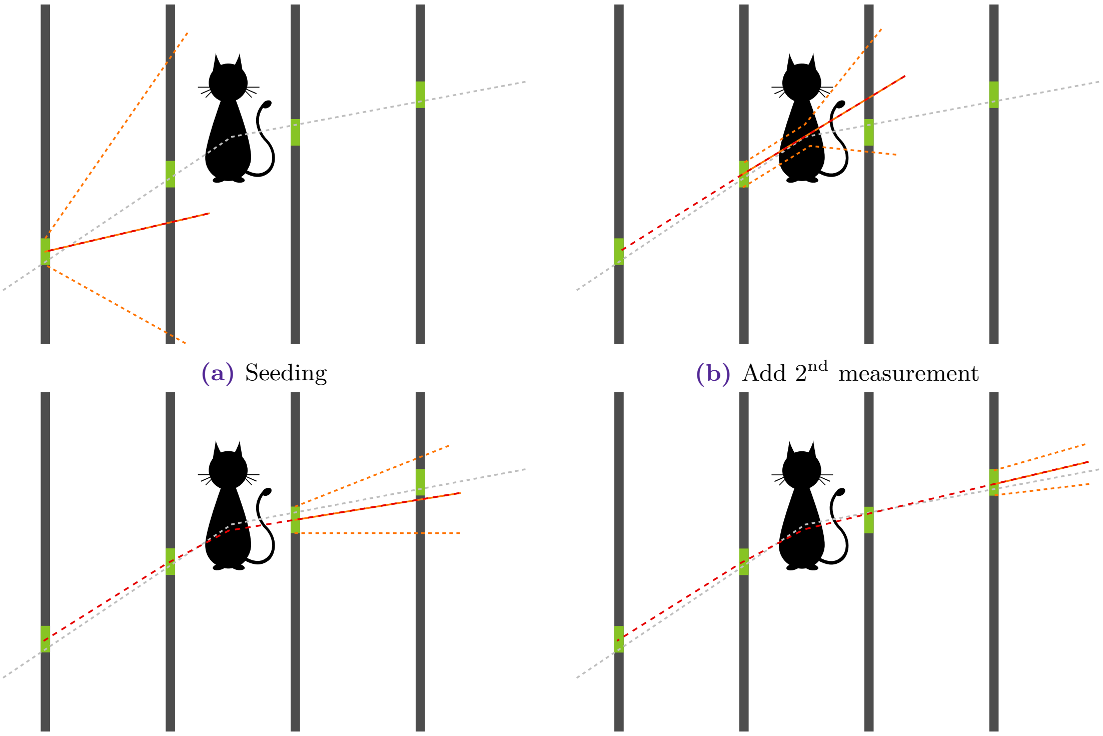
Figure 3.2: Demonstration of how a straight line track fit (red) can fail as a result of material interactions introducing a kink in the true particle's path (grey).

where $\overrightarrow{F}$ is the force on the particle, $\overrightarrow{B}$ is the magnetic field vector, q and $\overrightarrow{v}$ are the particle's charge and velocity. If the field is homogeneous, this can be accounted for by replacing the straight line function with a helix when performing track fits. In practice however, this is rarely used as particles can also experience energy loss and multiple scattering as shown in Figure 3.2. Instead, most experiments choose instead to fit tracks using a *Kalman filter*.

3.1.1 Kalman tracking

A Kalman filter[112] is a Bayesian technique for combining multiple measurements to produce a best estimate of parameters in the presence of statistical noise. When used for tracking the procedure is made up of four steps:

1. **Seeding:** Make an initial estimate of the track trajectory and the associated covariance matrix at the point of the first state.
2. **Prediction:** Predict the position of the next measurement, again with an associated covariance matrix.
3. **Projection:** Correct the prediction using the new measurement.



(a) Seeding

(b) Add 2$^{\text{nd}}$ measurement

(c) Add 3$^{\text{rd}}$ measurement

(d) Add 4$^{\text{th}}$ measurement

Figure 3.3: Steps involved in fitting a track using a Kalman filter when material interactions may occur between the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ measurements. The grey dashed line shows the true particle's path, the green squares show the detector response, the red line shows the reconstructed track and the orange dashed line shows the current prediction for the next measurement.

4. **Smoothing:** Propagate the new information iteratively to update the positions of the previous states.

Figure 3.3 shows this pictorially, with Figure 3.3a showing the initial prediction and Figure 3.3b, 3.3c and 3.3b showing the effect of adding additional measurements. To make predictions of future measurements a *transport function* must be defined, see Reference [113] for an overview of the transport functions used in LHCb.

3.2 Alignment

If the true position of tracking detectors differs from the position that is used when tracks are reconstructed the quality of the tracks is degraded. This increases the value of χ^2_{track} (Equation 3.1) and can have a secondary effect of increasing the rates of

- **Ghosts:** Reconstructed tracks that are made up of hits that were not caused by the same underlying particle.
- **Clones:** Hits from a particle are assigned to different reconstructed tracks resulting in two tracks from a single particle.

These effects are particularly important when the occupancy of the detector is high and many particles are present relative to the detector resolution. The χ^2_{track} is not a good measure of the overall quality of the detector alignment due to the fact that a track only contains $\mathcal{O}(10)$ measurements and does not have a statistically significant number of measurements in each detector element. To account for this a global χ^2, defined by

$$\chi^2 = \sum_{\text{track}}^{\text{tracks}} \chi^2_{\text{track}}, \tag{3.3}$$

can be computed and, given a large enough sample of tracks, will be at a minimum in the case of a well aligned detector.

There can also be misalignments that have very little effect on the global χ^2, these are known as *weak modes* and correspond to geometric effects to which the track residual is independent or insensitive. The simplest example of a weak mode would be a global translation and/or rotation of the entire system (Figures 3.4a and 3.4b). By definition this can have no effect on reconstructed quantities, however, it does prevent optimisation algorithms from being able to minimise the χ^2 due to the infinite number of minima. This is avoided by introducing *constraints*, the simplest of these would be keeping the position of one or more elements fixed. Alternatively, it can be more useful to allow all parameters to float and instead require that an equation be satisfied. This can be achieved using the Lagrange multiplier method that introduces an additional term in the χ^2,

$$\chi^2 = \sum_{\text{track}}^{\text{tracks}} \chi^2_{\text{track}} - \lambda \cdot g(\vec{\theta}) \tag{3.4}$$

where λ is an additional parameter and $g(\vec{\theta})$ is a function of the alignment parameters that has a minimum when the constraint is met.

Depending on the detector geometry there can also exist other weak modes. In the case of telescope-like detectors[1], such as the LHCb VELO, transformations known as *scaling* and *shearing* can be present. In the limit that all tracks are parallel to each other, these weak modes have no affect on the χ^2 and instead change the angle of the reconstructed tracks as shown in Figures 3.4c and 3.4d. These can be constrained with a careful choice of track selection and by adding Lagrange constraints on quantities such as vertices and invariant masses.[114] See Reference [115] for an overview of how these additional constraints are used in LHCb.

3.2.1 Survey methods

One method of aligning detector is with the use of survey techniques where the positions of the detector elements are measured directly. These methods are powerful however, modern detectors often demand higher precision than survey methods can reasonably provide, especially when movements due to operating temperature or humidity are taken into account. Additionally, detectors are often used for long periods of time and in the vicinity of large magnetic fields that can result in the position of the detectors evolving over time. As

[1] A telescope-like detector is one made up of approximately parallel planes of sensitive elements.

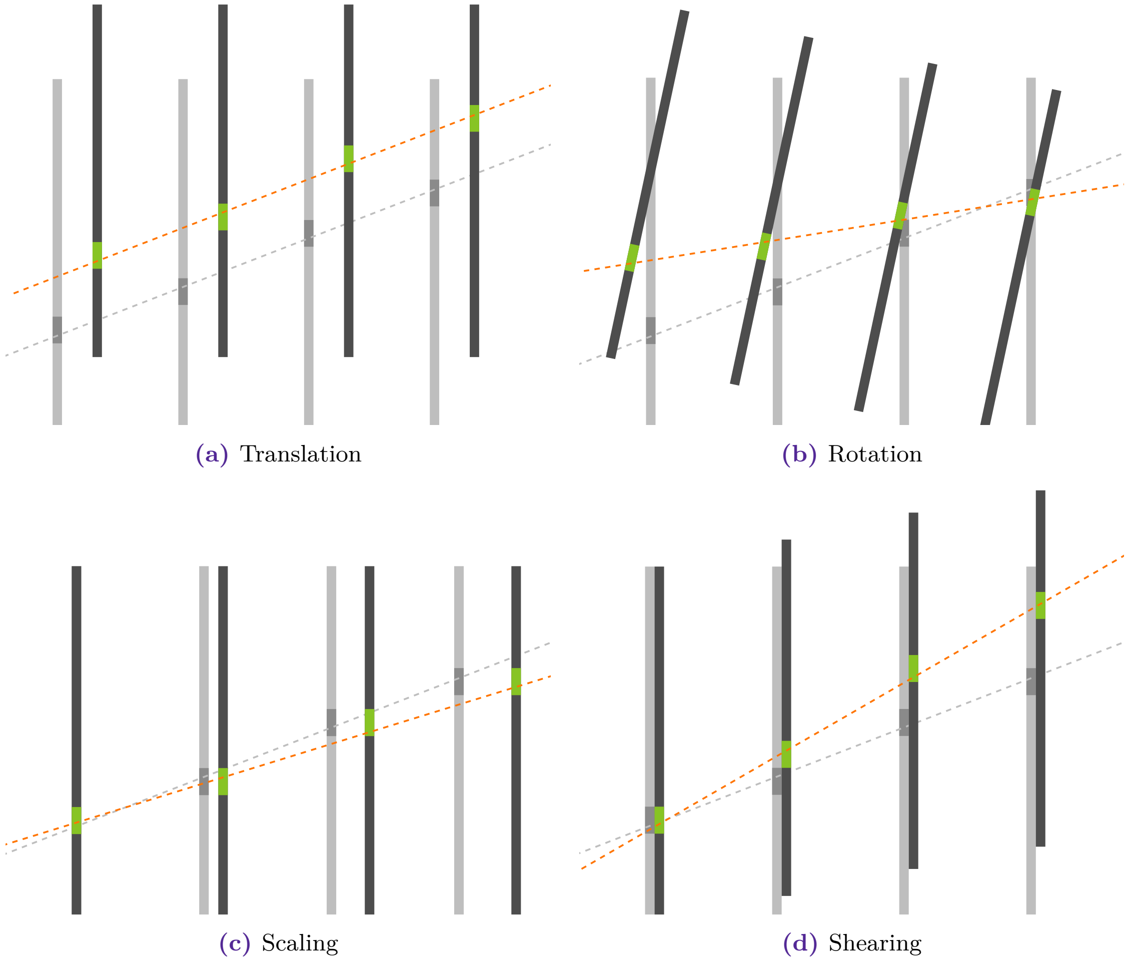


Figure 3.4: Pictorial demonstrations of the weak modes of the alignment of a telescope-like detector including the resulting effect on the reconstructed tracks.

a result, survey techniques are typically used only to obtain an approximate *starting position* of the elements, with data driven methods being used to further refine the position. The survey results can be included in other alignment procedures by introducing Lagrange constraints that ensure the aligned position is compatible with the survey uncertainties.

3.2.2 Histogram

One of the simplest data driven techniques for aligning detectors is with the use of histogram based methods. These work by producing histograms of each component (x/y) of the residual. The peaks of these distributions can then be used as translational alignment constants that shift the peak to be centred at zero. See Figure 3.5 for an example. Histogram based alignment is a simple yet powerful technique for obtaining an initial alignment, particularly if the starting position is not well known. It is however limited to only correcting misalignments that arise from translations.

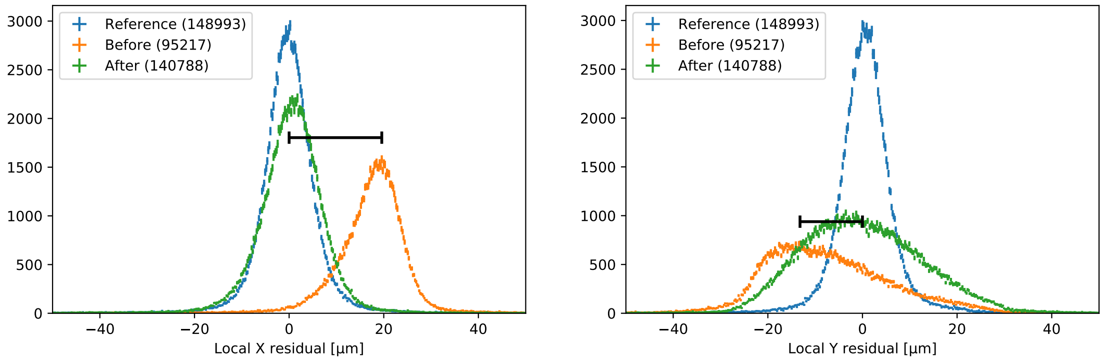


Figure 3.5: Residual before and after using a histogram method to align data taken with the LHCb VELO Timepix3 telescope with a known good alignment overlaid. The black line shows the correction that was made and the legend shows the number of reconstructed tracks in each scenario.

3.2.3 Iterative minimisation

If the initial alignment is sufficiently good to form tracks, track based methods can be used. The simplest implementation of this is to use a general purpose optimisation algorithm such as `MIGRAD`[116], `BFGS`[117] or simulated annealing[118] to find the alignment parameters that minimise the global χ^2 subject to any constraints. It is most common to model each detector element with six parameters: three translations (Δ_x, Δ_y, Δ_z) and three rotations (r_x, r_y, r_z).[2] For a system of n detector elements this results in $6n$ alignment parameters. Additional free parameters are introduced due to the track parameters. In the simplest case of straight tracks (Equation 3.1) this introduces 4 parameters for each track. For a simple telescope consisting of 8 detector planes, such as that described in Section 3.3, the number of tracks used for alignment $\mathcal{O}\left(10\,000\right)$ tracks will typically be used to align 8 detector elements. This results in a minimisation problem with $\mathcal{O}\left(40\,000\right)$ free parameters, of which only 48 are the alignment parameters.

To counteract the large number of free parameters the problem is normally simplified by converting this into an *iterative* method. This is achieved by minimising the global χ^2 for each detector element's parameters individually. The track parameters are held fixed during the minimisation and are refitted between each iteration. To account for possible biases in the original track parameters this method must typically be repeated several times. Even with these repeated iterations it is often impossible to find the optimal alignment conditions due to correlations between the parameters. Instead it is common for the iterations to oscillate between local minima near the true global minimum of the χ^2.

An example of an iterative alignment algorithm that uses a general purpose optimiser is the `NUMERIC` method in the `Bach` alignment toolkit.[119]

[2] Other alignment parameters can be needed to model effects such as the non-linearity introduced from large elements bending. These methods can be extended to accommodate such effects however care must be taken to ensure the minimisation problem remains well constrained.

3.2.4 Millepede

To avoid the limitations iterative methods, it is desirable to instead use global methods that can fully account for the correlations between the track and alignment parameters. One such algorithm is `Millepede`[120, 121] and has been widely used in the high energy physics community including the H1[122], ZEUS[123], ATLAS[124], CMS[125] and LHCb[126, 127] experiments. This algorithm involves constructing a matrix using the partial derivatives of the global χ^2 with respect to the track and alignment parameters. Using the property that the global χ^2 is at a minimum when its derivatives are zero, solving for the alignment parameters becomes a matrix inversion problem, i.e.

$$\mathbf{C}\vec{d} = -\vec{g} \tag{3.5}$$

where $\vec{d}$ is a vector of the track and alignment parameters, $\mathbf{C}$ is an invertible matrix and $\vec{g}$ is a vector that is dependent on the track parameters. The alignment parameters can then be extracted using

$$\vec{d} = -\mathbf{C}^{-1}\vec{g}. \tag{3.6}$$

Inverting large matrices is computationally intensive, with a computational complexity of $\mathcal{O}\left(n^3\right)$ when using the row reduction method. Fortunately, it is possible to avoid inverting the entire matrix by exploiting the fact that only the alignment parameters are of interest. This is the basis of the `Millepede` method.

By transforming Equation 3.5 $\mathbf{C}$ can be restructured such that the track and alignment parameters are independent and this simplification makes it possible to use `Millepede` to compute thousands of alignment parameters on conventional computing hardware. See Reference [121, 128] for further detail.

3.2.5 Kalman

A method based around the Kalman track fit described in Section 3.1.1 can be used for computing the alignment parameters. As with the `Millepede` method, this involves taking the partial derivatives of the global χ^2 with respect to the alignment parameters and solving this system of equations to find the minimum. This method has the benefit that the χ^2 that is used for the alignment is the same as the one used for the track fit. In particular it allows the uncertainty of the measurements and of the transport function to be fully incorporated. A global alignment procedure based on the Kalman method is used to align the LHCb's tracking detectors and more detail is available in Reference [115].

3.3 The Timepix3 Telescope

Developing new detectors for experiments in high energy physics is normally a multi-year research project in itself. In almost all cases there is a desire to push the boundary of what is possible and create elements that are more efficient, radiation tolerant, have higher resolution and consist of less material. During this process, lab based testing of devices is

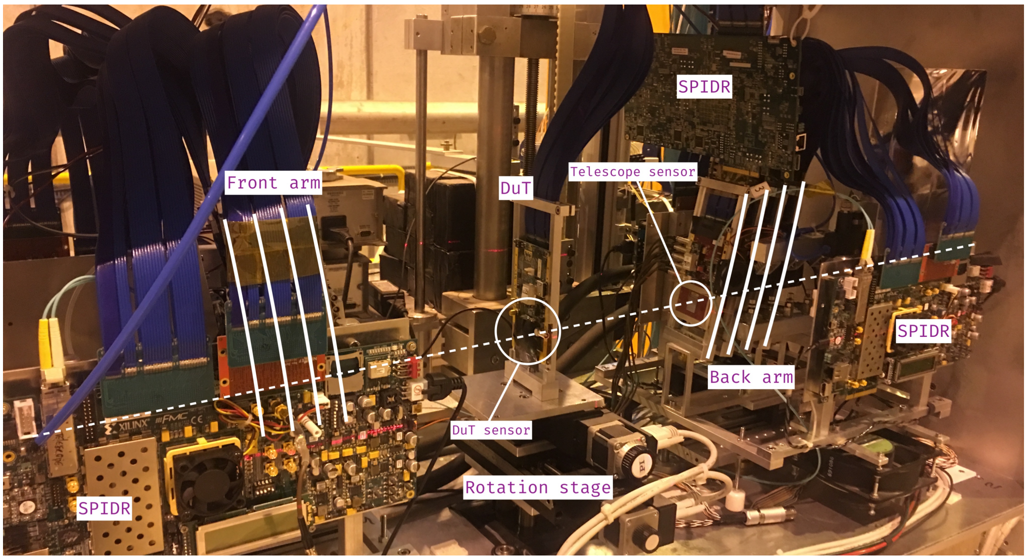


Figure 3.6: Annotated photograph of the LHCb VELO Timepix3 telescope taken during the June 2017 testbeam. The beam position is shown with a white dashed line and the solid white lines show the position of the telescope sensors. Two additional `SPIDR` boards are placed on the opposite side of the assembly and are not visible.

essential and can involve electrical testing or the use of radioactive sources to study the performance of devices. Sometimes however it is impossible to adequately simulate the running conditions of the final experiment in a laboratory environment and in these cases testbeams can be used.

Testbeam facilities typically provide a small area where experiments can be placed into a particle beam for the purposes of testing and validation. The beam is often configurable, allowing users to set the particle species, intensity and energy.

CERN's north area is often used by LHCb and is able to provide up to 400 GeV beams consisting of protons, electrons, muons or a mixture of hadrons that originate from the Super Proton Synchrotron. The beam is not continuous and consists of *spills* during which up to 1×10^7 particles[3] are provided in 5 s bursts.[129] While beam monitoring is provided to indicate the beam's intensity, profile and make up, it is often useful to be able to reconstruct tracks to precisely measure the trajectory of each particle. The LHCb VELO group normally uses a device known as the LHCb VELO Timepix3 telescope[130], which is comprised of 8 planes each with a sensor consisting of 256x256 55x55 µm pixels connected to a Timepix3 ASIC.[131] As shown in Figure 3.6, these devices are arranged in two *arms* that are each angled at 7° in both x and y. As the beam is approximately collimated, this angle improves the resolution of the devices.

The spacing between the arms can be varied depending on the size of the device that is being tested. Often the Device under Test (DuT) needs to be biased with high voltages and cooled to low temperatures. These requirements result in the sensor needing to be kept under vacuum to prevent sparking and condensation. Rather than operate the entire telescope assembly in vacuum, the telescope arms can be adjusted to allow an aluminium

[3] Can be increased to 1×10^8 in special cases and is limited to 1×10^5 for electron beams.

Period	Number of runs	Period	Number of runs
July 2014	34	May 2016	1192
Oct 2014	719	Aug 2016	297
Nov 2014	296	Nov 2016	751
Dec 2014	240	June 2017	340
May 2015	1577	July 2017	324
July 2015	2100	July 2018	634
Sep 2015	705	Oct 2018	1021
Nov 2015	339		

Table 3.1: Number of runs collected during each testbeam campaign using the LHCb VELO Timepix3 telescope. These numbers include bad runs and those used for set up and debugging.

vacuum enclosure to be inserted. To minimise the material that the beam is exposed to, and therefore minimise the chance of multiple scattering, the sides contain foil covered “windows”. Inside the vacuum box, the DuT is attached to a Peltier module that moves heat to a reservoir that is connected to an external chiller. This system allows the DuT to be biased up to 1000 V and cooled to around −25 °C.

Data from the sensors is read out using the `SPIDR`[132] readout system. Four boards are used for the telescope, one for each pair of planes and an additional board is used for the DuT. Each `SPIDR` is connected to a dedicated computer that acts as the primary store of the data. This data is later copied to `EOS`[133] to make it available to analysts.

Data is collected in 1-2 week *testbeam campaigns*, also know as *periods*. By convention the LHCb VELO group refers to these by the month and year when the campaign started. Data is collected in *runs* that are manually started between spills and stopped once a sufficient quantity of data has been collected. This typically involves 2 or 3 spills but can require many more depending on the beam conditions and reason for collecting the data.

The LHCb VELO Timepix3 telescope can be controlled remotely, this includes varying the high voltage power supplies and translating the entire assembly in x and y. The middle assembly for the DuT can be rotated around y and translated in x and y. This capability is extremely useful as testing a device typically involves scanning through parameters, such as angle or voltage, running for only a few minutes in each configuration. As shown in Table 3.1 this can result in many hundreds of runs being collected during a testbeam campaign over the course of 1-2 weeks. Due to the necessary safety procedures, entering the experimental area is relatively slow, taking at least 10 minutes each time and collecting data in a large number of configurations would be impractical without the ability to remotely control the assembly.

For processing data collected during testbeam campaigns, a software package called `Kepler`[134] is used. This is based on the `Gaudi`[102] framework and is independent of the LHCb software stack. Unlike other LHCb applications, users of `Kepler` typically compile the entire application from source and modifications are made to the source code directly instead of configuring algorithms.

3.4 Alignment of the Timepix3 telescope

While it is possible to perform some studies without knowledge of the telescope and DuT position, many require the telescope and DuT to be well aligned. The changing conditions necessitate a generic alignment procedure that can be applied to each run and, due to the weak modes of telescope geometries, it is necessary to find a good alignment to use as a starting position. This is found by manually inspecting the monitoring plots produced by `Kepler`. Using this information to adjust the starting position and parameters of the alignment algorithms until a good alignment is found.

Once an acceptable starting position has been found for a given testbeam campaign, an automated procedure can be used. This is split into three steps: calibration, alignment and validation.

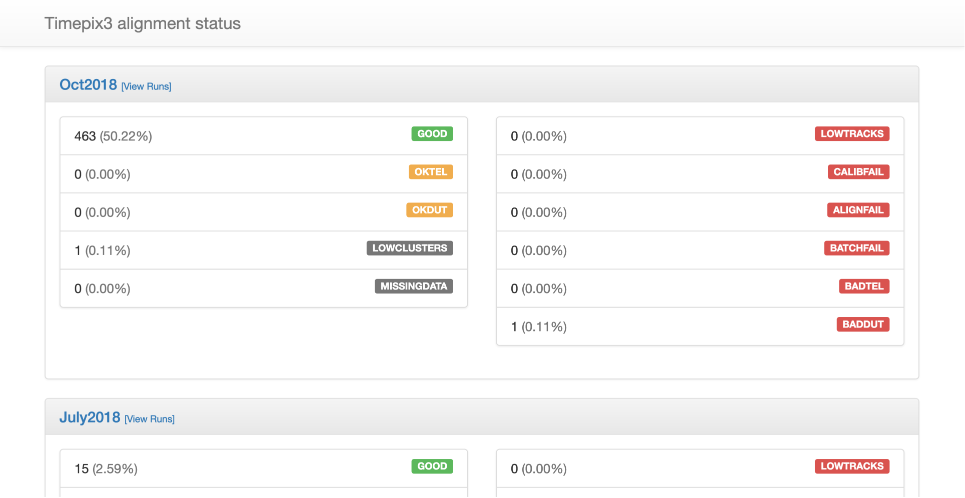
In the calibration step an algorithm is applied to mask noisy pixels and use a histogram based alignment method to calculate a time alignment from the time residuals of the tracks. The correction in the time alignment is typically small due to the `SPIDR` readout boards being synchronised. However, it is occasionally necessary to make corrections up to the 100 ns level. As the occupancy of the telescope is low relative to the 1.5625 ns time resolution of the Timepix3 ASIC,[131] the selection used for track reconstruction can be relaxed making this procedure approximately independent of the detector alignment.

After calibrating the run, it is aligned using a multi-step procedure. First the `Millepede` method[120] is applied multiple times with the DuT removed from the alignment and tracking algorithms. To make the procedure more stable, earlier instantiations of `Millepede` only correct for degrees of freedom to which the global χ^2 is most sensitive; i.e. translations in x and y and rotations around z. Once the telescope is aligned, it is assumed the tracks are sufficiently well understood that an iterative method can be used to find the alignment parameters that minimise the global χ^2 using the `MIGRAD` optimiser in `MINUIT`.[116] Again, this is applied multiple times with only the most sensitive degrees of freedom being allowed to vary in the earlier steps.

To validate the results of the alignment, a separate invocation of `Kepler` is used to produce the unbiased residual distributions. For the telescope planes this involves refitting the tracks with the hits from the current plane excluded to remove the bias that is introduced from including the current hit in the track fit.[4] These residuals are fitted using a normal distribution, with the mean and width being used to determine if the alignment of the telescope and DuT is “good”. Additionally, the x residual is plotted as a function of the x position in the local frame of the sensor. Ideally this should be uncorrelated and any slope implies the presence of a rotational misalignment in the system. This distribution is fitted using a straight line and the slope of this line is also used to determine if the alignment can be flagged as “good”.

To present this information in an easily accessible form, a web application was developed in `Python` using the `Flask`[135] microframework. This website allows users of

[4] Some bias remains due to the tracking algorithm using the hits from all telescope sensors. This bias is generally assumed to be small and neglected. The DuT residuals are not affected as these hits are not used by the tracking algorithm.



Timepix3 alignment status

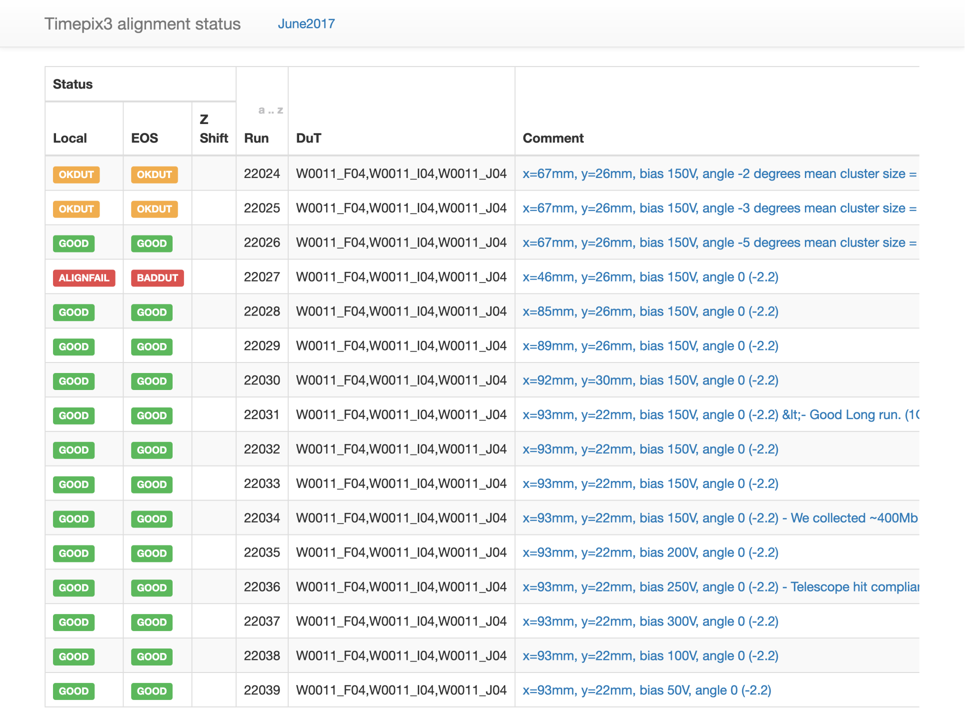
Oct2018 [View Runs]

463 (50.22%)	GOOD	0 (0.00%)	LOWTRACKS
0 (0.00%)	OKTEL	0 (0.00%)	CALIBFAIL
0 (0.00%)	OKDUT	0 (0.00%)	ALIGNFAIL
1 (0.11%)	LOWCLUSTERS	0 (0.00%)	BATCHFAIL
0 (0.00%)	MISSINGDATA	0 (0.00%)	BADTEL
		1 (0.11%)	BADDUT

July2018 [View Runs]

15 (2.59%)	GOOD	0 (0.00%)	LOWTRACKS

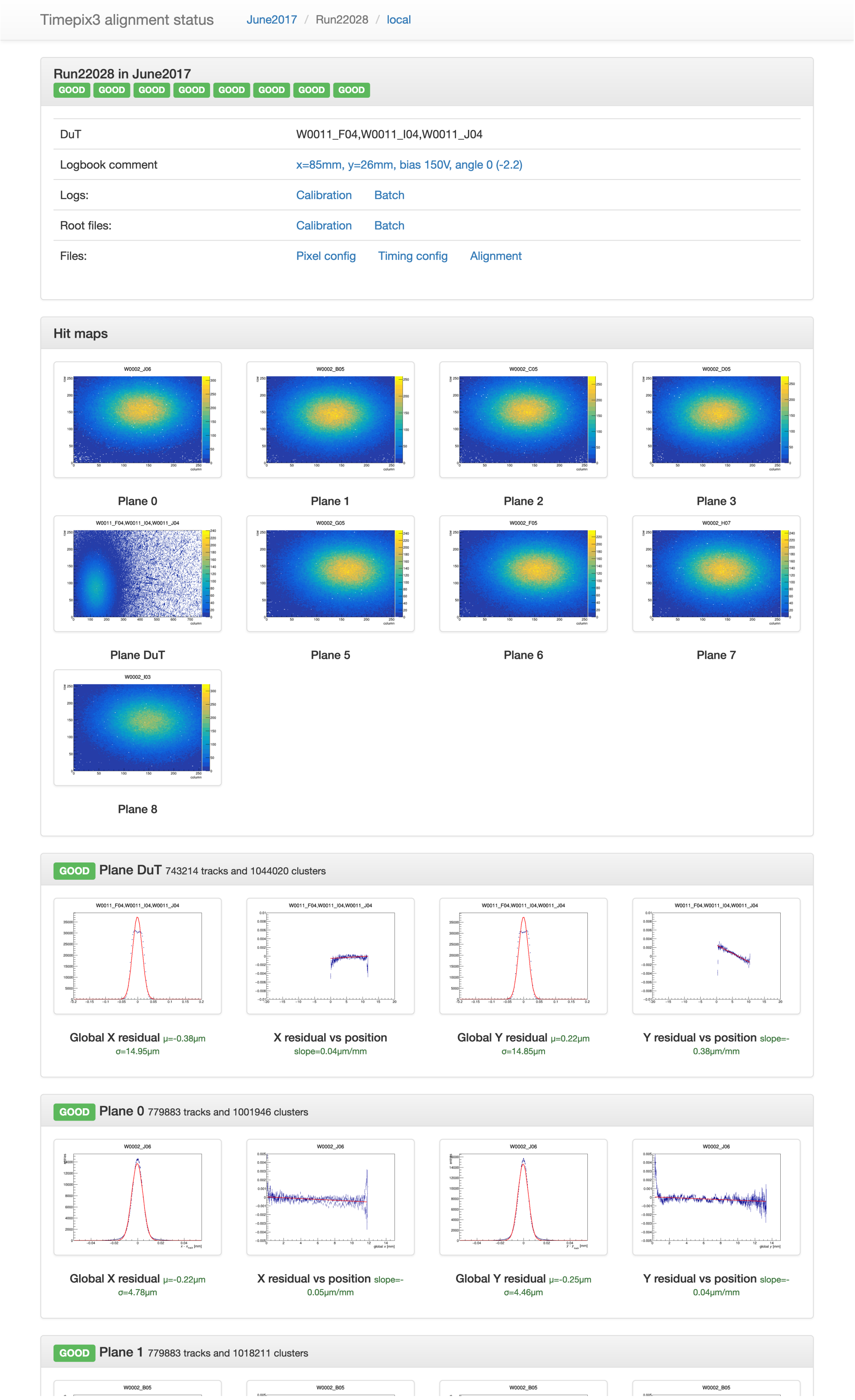
Figure 3.7: Home package of the web application used to see an overview of the alignment status of data collected with the LHCb VELO Timepix3 telescope. Most of the runs without a good alignment correspond to runs in which useful data was not collected.



Timepix3 alignment status June2017

Status			a .. z		
Local	EOS	Z Shift	Run	DuT	Comment
OKDUT	OKDUT		22024	W0011_F04,W0011_I04,W0011_J04	x=67mm, y=26mm, bias 150V, angle -2 degrees mean cluster size =
OKDUT	OKDUT		22025	W0011_F04,W0011_I04,W0011_J04	x=67mm, y=26mm, bias 150V, angle -3 degrees mean cluster size =
GOOD	GOOD		22026	W0011_F04,W0011_I04,W0011_J04	x=67mm, y=26mm, bias 150V, angle -5 degrees mean cluster size =
ALIGNFAIL	BADDUT		22027	W0011_F04,W0011_I04,W0011_J04	x=46mm, y=26mm, bias 150V, angle 0 (-2.2)
GOOD	GOOD		22028	W0011_F04,W0011_I04,W0011_J04	x=85mm, y=26mm, bias 150V, angle 0 (-2.2)
GOOD	GOOD		22029	W0011_F04,W0011_I04,W0011_J04	x=89mm, y=26mm, bias 150V, angle 0 (-2.2)
GOOD	GOOD		22030	W0011_F04,W0011_I04,W0011_J04	x=92mm, y=30mm, bias 150V, angle 0 (-2.2)
GOOD	GOOD		22031	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 150V, angle 0 (-2.2) &lt;- Good Long run. (1
GOOD	GOOD		22032	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 150V, angle 0 (-2.2)
GOOD	GOOD		22033	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 150V, angle 0 (-2.2)
GOOD	GOOD		22034	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 150V, angle 0 (-2.2) - We collected ~400Mb
GOOD	GOOD		22035	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 200V, angle 0 (-2.2)
GOOD	GOOD		22036	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 250V, angle 0 (-2.2) - Telescope hit complia
GOOD	GOOD		22037	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 300V, angle 0 (-2.2)
GOOD	GOOD		22038	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 100V, angle 0 (-2.2)
GOOD	GOOD		22039	W0011_F04,W0011_I04,W0011_J04	x=93mm, y=22mm, bias 50V, angle 0 (-2.2)

Figure 3.8: Summary page for checking the status of each run in a data taking period.



Timepix3 alignment status June2017 / Run22028 / local

Run22028 in June2017

GOOD GOOD GOOD GOOD GOOD GOOD GOOD GOOD

DuT	W0011_F04,W0011_I04,W0011_J04
Logbook comment	x=85mm, y=26mm, bias 150V, angle 0 (-2.2)
Logs:	Calibration Batch
Root files:	Calibration Batch
Files:	Pixel config Timing config Alignment

Hit maps

Plane 0

Plane 1

Plane 2

Plane 3

Plane DuT

Plane 5

Plane 6

Plane 7

Plane 8

GOOD Plane DuT 743214 tracks and 1044020 clusters

Global X residual μ=-0.38μm σ=14.95μm

X residual vs position slope=0.04μm/mm

Global Y residual μ=0.22μm σ=14.85μm

Y residual vs position slope=-0.38μm/mm

GOOD Plane 0 779883 tracks and 1001946 clusters

Global X residual μ=-0.22μm σ=4.78μm

X residual vs position slope=-0.05μm/mm

Global Y residual μ=-0.25μm σ=4.46μm

Y residual vs position slope=-0.04μm/mm

GOOD Plane 1 779883 tracks and 1018211 clusters

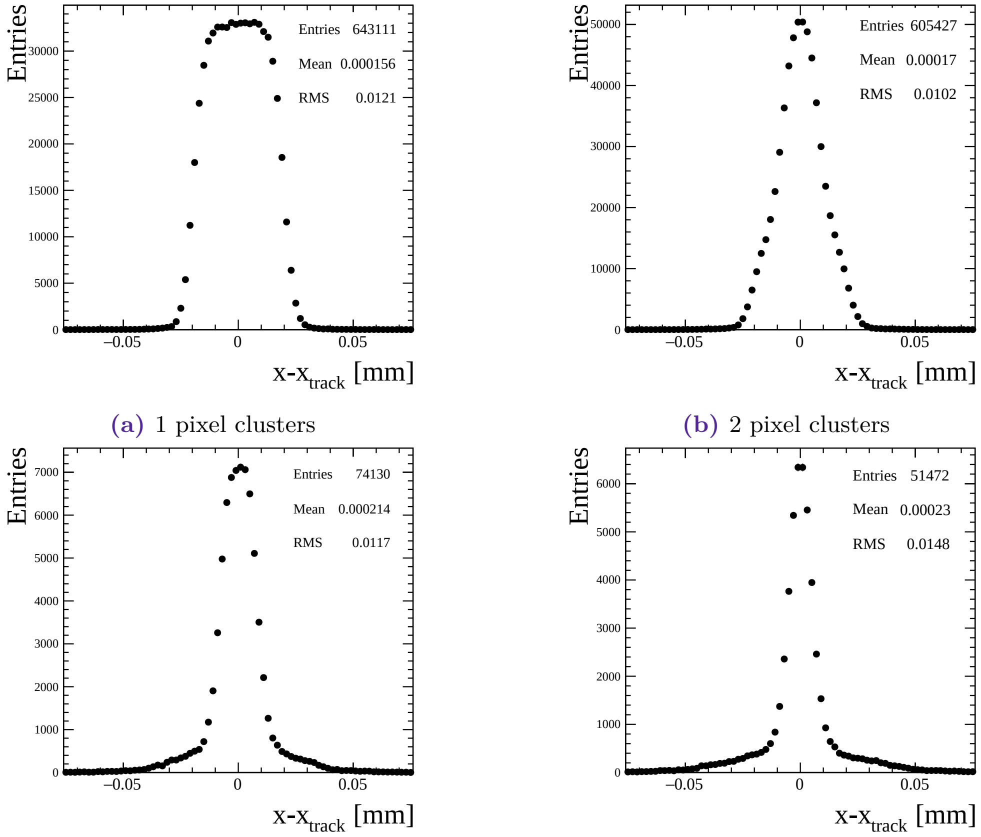
Figure 3.9: Summary page of the alignment status for a single run collected during the October 2018 test beam campaign. The page is split into three sections: a summary, hit maps showing the occupancy of each sensor and track, and a detailed view of the alignment quality of each sensitive element.

testbeam data see an overview of the alignment status of each testbeam campaign (Figure 3.7). From this page users can navigate to a page containing a summary of each run in a given testbeam period (Figure 3.8) including: the DuT that was present, the quality of the alignment and a link to the `elog`[136] entry for that run. Detailed information is then available for a specific run including: hit maps of clusters, the fitted residuals from the validation step, the number of hits/clusters and output logs of `Kepler` for each step. An example of this page is shown in Figure 3.9.

3.5 Example results from the Timepix3 telescope

The work described in Section 3.4 has facilitated measurements of the sensor performance and the validation of the design. One such study, shown in Figure 3.10, shows the resolution of VELO sensors as a function of the cluster size and similar results are available[137] showing these effects as a function of irradiation level, track angle and operating voltage. These measurements are useful for understanding the tracking performance and condition of the detector. Furthermore, they could potentially be used to further improve the alignment procedures used for the LHCb VELO.

This work has culminated in three prototype modules, constructed in Manchetser and Nikhef, being used for a testbeam campaign at CERN in October 2018 using the setup shown in Figure 3.11. These modules have essentially the final design and use a mixture of production and final prototype components. Preliminary analysis shows the modules are functioning as expected and work to further validate their performance is ongoing.



(a) 1 pixel clusters

(b) 2 pixel clusters

(c) 3 pixel clusters

(d) 4 pixel clusters

Figure 3.10: Track residuals as a function of cluster size for a prototype VELO sensor.[137]

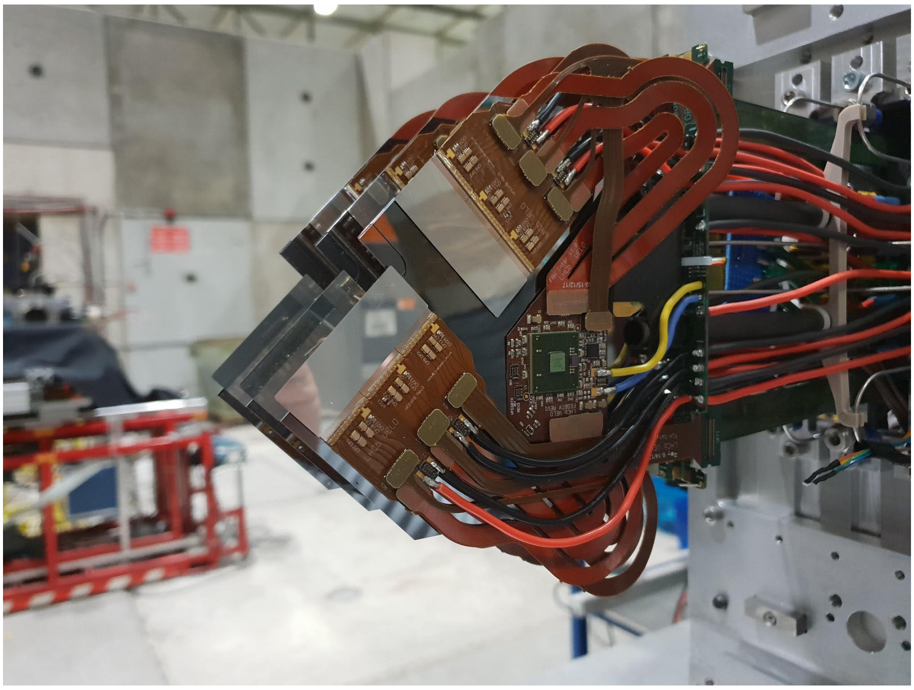


Figure 3.11: Three prototype upgrade velo modules being tested at the SPS north area in October 2018.

Chapter 4

Effect of cooling induced distortions in the LHCb VELO Upgrade

During the development of the LHCb VELO Upgrade modules, described in Section 2.12.1, the initial prototypes were found to distort when cooled from room temperature to their operating temperature (approximately −30 °C). The prototypes tested were based on two designs:

- **Plan A:** Uses a silicon substrate that will have microchannels etched inside to allow bi-phase CO_2 to be circulated.[138] Onto this substrate the readout hybrid and silicon sensors and readout chips are mounted. The initial prototypes used a silicon plate as microchannel plates were not available.
- **Plan B:** Uses a ceramic, Shapal, substrate with embedded stainless steel pipes.

Measurements of these prototypes performed at both Nikhef and the University of Manchester suggested that the tip of the module (corresponding to D3 in Figure 4.1) may move significantly more than had been anticipated.[139] The magnitude of this distortion varied between prototypes, up to a maximum of around 500 µm from the cooling block to the tip of the module, and corresponds to a rotation at the mid-plate around y in the global LHCb frame. For Plan B it was envisaged to include a constraint system to grip the module on either side at 60 mm away in y from the tip of the module. In this case the distortion would be much reduced and would likely have opposite sign at the tip and base.

This chapter describes studies that were performed using the full LHCb simulation and reconstruction to assess the potential physics impact of these distortions. This information was used to guide the choice of module substrate at the VELO upgrade module mechanical module engineering design review.[140, 141]

4.1 Methodology

In order to study the potential impact of the laboratory measurements, full Monte Carlo simulations were generated using the default, perfectly aligned, geometry. These samples

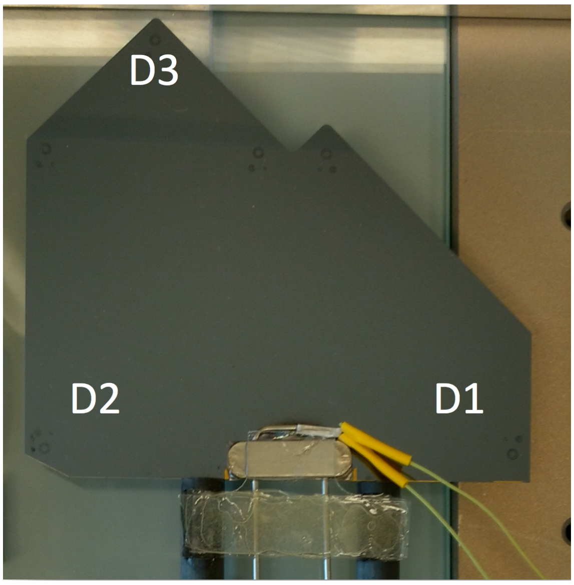


Figure 4.1: Image of a prototype bare module, based on a silicon substrate, with the positions of the displacement sensors marked. The rotations observed are around the mid-plate that is seen at the bottom of the figure between the two carbon fibre support rods.

were then passed through a full event reconstruction, using a detector model with misaligned modules, under a range of different scenarios. Reconstruction and analysis level quantities were then determined for each scenario, including the perfectly aligned case.

For these studies a Monte Carlo sample of $D^{*+} \to (D^+ \to K^+K^-)\,\pi^+$ decays was used. This was produced with modules being perfectly aligned and parallel to the x/y plane in the global LHCb frame. To simulate the effect of the distortions, a modified detector geometry database[142] was created for each scenario and this modified geometry was used during the event reconstruction. The method is similar to that used for previous VELO alignment studies.[143] This approach, whereby the true module positions are perfectly aligned, but the reconstruction assumes a misaligned geometry, is in fact the converse of the scenario of interest (where the detector is misaligned in reality, but the reconstruction assumes perfect alignment). As the deflection angles are small, these two cases differ only in the sign of the z--displacement of the clusters. To ensure the results were not sensitive to this, module distortions in both directions are included in this study. Additionally, this method fails in extreme scenarios as the cluster position in the local frame of the module is held constant, whereas in reality large distortions would cause different pixels, if any, to detect each particle.

All rotations considered were applied around the y axis at the position of the cooling block, based on measurements of prototype modules. These are then described by the z displacement (δ_z) of a point 10 cm away, corresponding to the approximate position of the tip, as shown in Figure 4.2. Three different classes of scenario were considered as the variations in the distortions of modules will only be known once a significant part of the production has been completed:

- **All same:** Here all modules are rotated exactly the same magnitude in the same

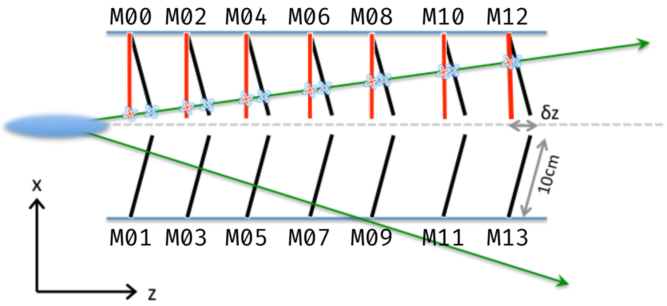


Figure 4.2: Example difference between the nominal (red) and the misaligned (black) VELO module position in the global x/y plane of LHCb. Note that if the modules misalign in a uniform way then the two sides of the VELO will give a delta z with the same sign, as illustrated.

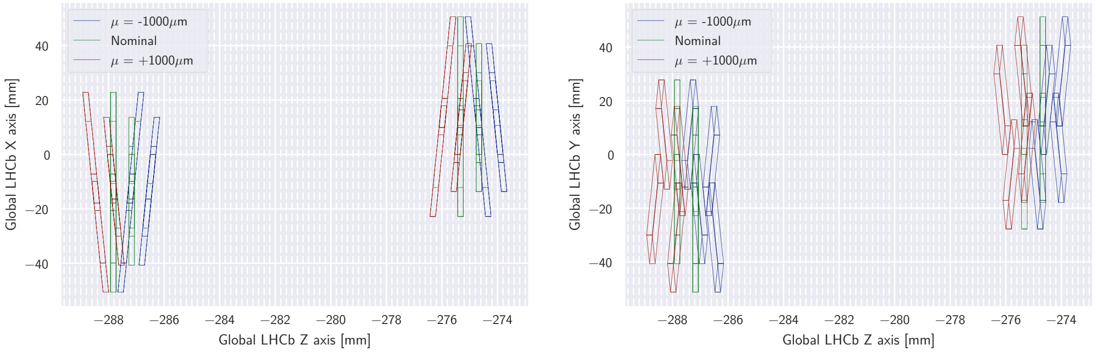
direction for ±100 µm, ±250 µm, ±500 µm and ±1000 µm. This corresponds to all modules distorting the same way with cooling and is a best case, as will be seen below.

- **Random:** Here all modules are rotated by an magnitude taken from a normal distribution with σ equal to half the mean, for means of ±100 µm, ±250 µm, ±500 µm and ±1000 µm. This corresponds to module distortions being correlated, but having significant variability and may be considered the most realistic scenario. Note that for each scenario only one misaligned detector model was randomly generated making this sensitive to the specific values that are randomly generated. In particular, quantities that are sensitive to individual modules may vary significantly, see Figure 4.15 for an example of this.
- **Alternating:** Here only alternate stations are displaced, e.g. `M02`, `M03`, `M06 M07` etc. This corresponds to a worst case scenario maximising the effects seen and is performed for $\delta_z = \pm 25\,\mu\text{m}, \pm 50\,\mu\text{m}, \pm 100\,\mu\text{m}, \pm 125\,\mu\text{m}, \pm 150\,\mu\text{m}, \pm 200\,\mu\text{m}$ and ±1000 µm.

Each scenario is implemented by editing the module alignment constants in the LHCb conditions database[142] before performing the full track reconstruction using `Brunel`.[144] The resulting data files were then processed in the `Panoramix`[145] software environment.

4.2 Cross-checks

In order to validate that the distortions are applied correctly, the x-z and y-z projections were plotted for both the entire VELO and a single module pair. In all cases the observed module positions match the desired misalignments included in the conditions database, as can be seen in Figures 4.3-4.5.



(a) x-z projection

(b) y-z projection

Figure 4.3: Birds-eye and side-on views of the ASICs on the first two VELO modules shown to validate that the distortions are applied correctly to the LHCb geometry. The plotted lines correspond to the true alignment and the *all same* scenario with tip displacements of ±1000 µm, as described in Section 4.1.

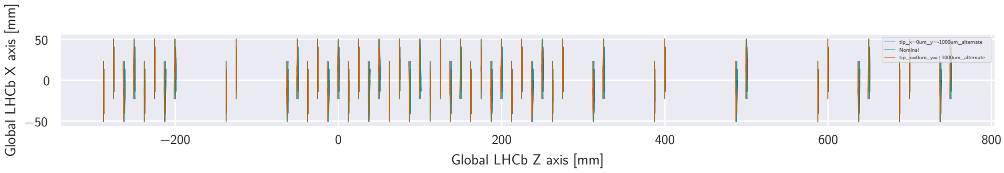


Figure 4.4: Birds-eye (z/x) view of the ASIC positions for the entire VELO with the true alignment and *alternating* scenario with a maximum tip displacement of ±1000 µm, as described in Section 4.1.

4.3 Track level results

In this section the effect of the simulated VELO upgrade module distortions on the: impact parameter resolution, primary vertex resolution, residuals between reconstructed track and cluster positions, tracking performance and momentum resolution is studied. These quantities were chosen as these all depend upon the position of the VELO clusters and/or the extrapolation of VELO tracks to the rest of the LHCb detector.

4.3.1 Impact parameter resolution

The Impact Parameter (IP) for tracks produced in a primary p-p interaction is a measure of the accuracy and precision with which tracks are reconstructed. In this study it is defined as the distance between the track and the reconstructed primary vertex, in the z plane of the reconstructed primary vertex as shown in Figure 4.6.

In the ideal case all components should be zero, assuming the track originates from the primary vertex, and is therefore an important variable in the suppression of prompt background in most flavour physics analyses. The IP is dependent on both the extrapolation distance from the first VELO hit and multiple scattering effects in detector material.

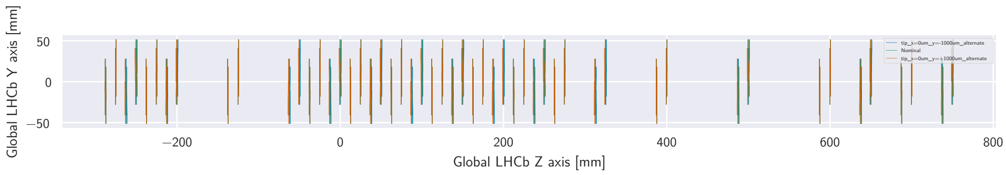


Figure 4.5: Side-on (z/y) view of the ASIC positions for the entire VELO with the true alignment and *alternating* scenario with a maximum tip displacement of ±1000 µm, as described in Section 4.1.

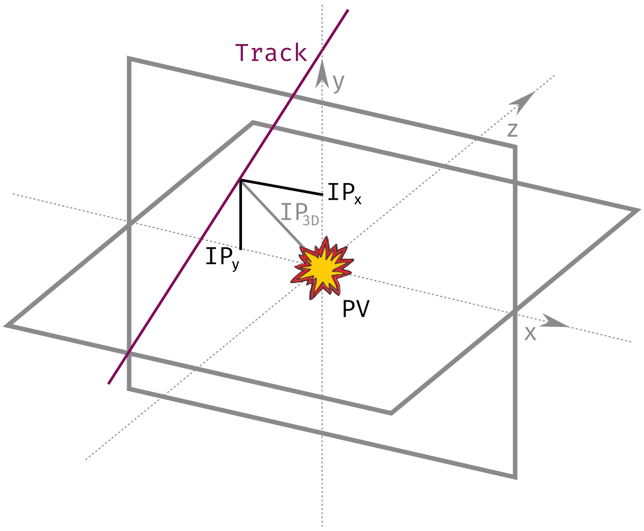
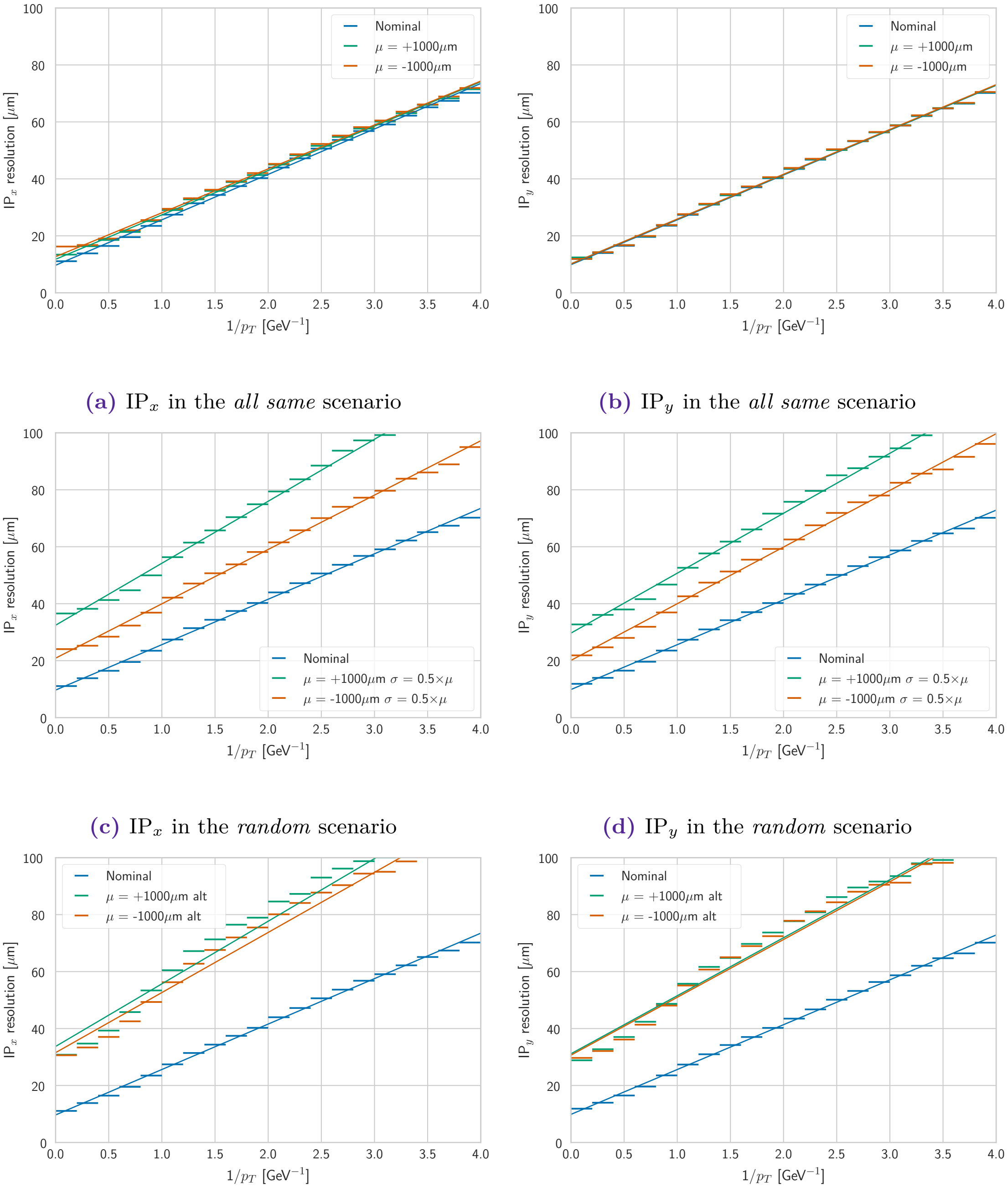


Figure 4.6: Pictorial representation showing the impact parameter is defined by the minimum distance between a track and primary vertex, in the z plane of primary vertex. All values correspond to reconstructed quantities.

These effects introduce a strong momentum dependence and, as a result, the impact parameter resolution is plotted as a function of inverse transverse momentum. Figure 4.7a and 4.7b show the IP plotted against inverse p_T for the correctly aligned detector alongside the scenario where all modules are displaced by ±1000 µm. Even in this extreme scenario the effect is small with high momentum track performance being minimally degraded, likely due to the scenario being approximately equivalent to a global translation of the VELO in z.[128, 146] As discussed in Section 3.2, this is a known *weak mode* of the LHCb VELO geometry. The small degradation at high momentum likely results from extrapolation effects to the other tracking detectors.

When the magnitude of the distortion is allowed to vary between modules it can no longer be approximated as a global translation and the effect becomes more severe, as can be seen in Figures 4.7c, 4.7d, 4.7e and 4.7f.

In order to find the level where these distortions start to dominate over other detector effects the y intercept, corresponding to the best case resolution with high p_T tracks, was examined as a function of the various scenarios. Figure 4.8 shows the effect is small in all *all same* scenarios. In the *random* scenarios, the resolution starts to significantly degrade after 250 µm, corresponding to a 125 µm standard deviation of the variation. For extreme



(a) IP$_x$ in the *all same* scenario

(b) IP$_y$ in the *all same* scenario

(c) IP$_x$ in the *random* scenario

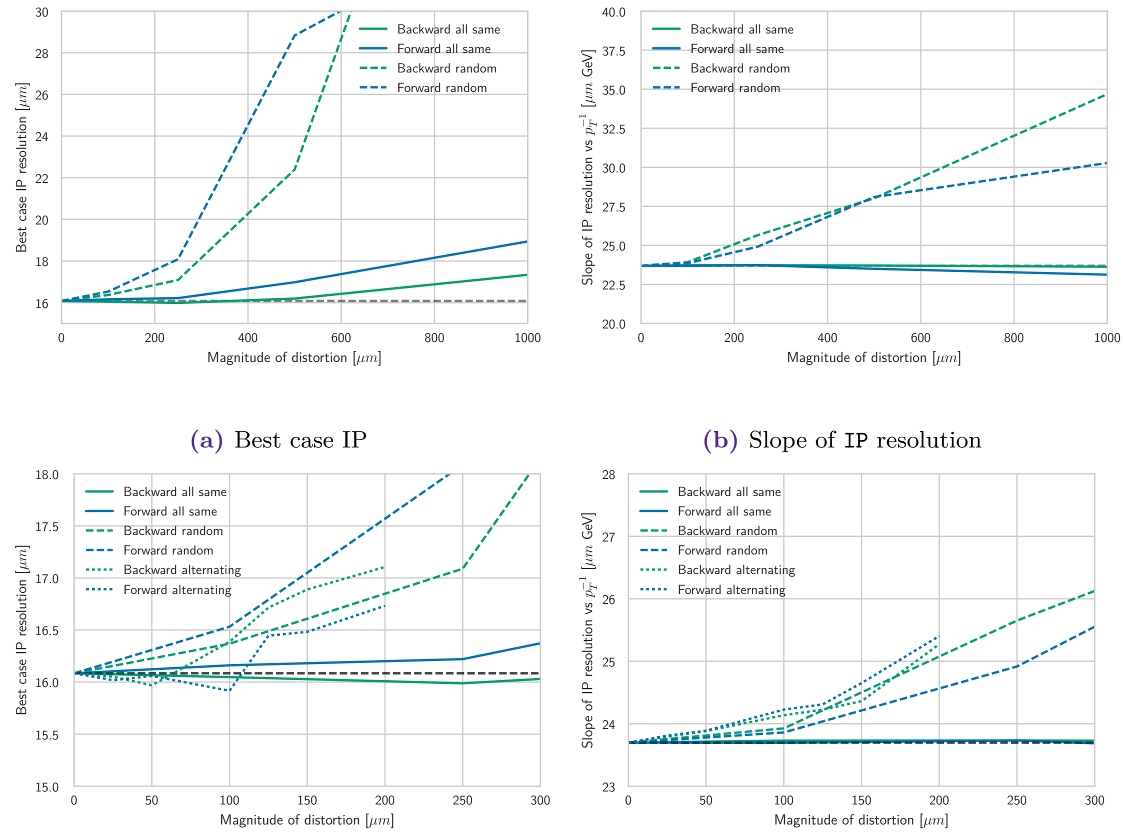
(d) IP$_y$ in the *random* scenario

(e) IP$_x$ in the *alternating* scenario

(f) IP$_y$ in the *alternating* scenario

Figure 4.7: Components of the impact parameter between the reconstructed track and primary vertex in the plane of the reconstructed PV. All three scenarios are shown with a displacement magnitude of ±1000 µm.

cases the positive and negative z displacements show different results, this is believed to be caused by the single misalignment scenario used for this study and likely represents the result's sensitivity to the misalignment of a small number of modules.



(a) Best case IP

(b) Slope of IP resolution

(c) Best case IP

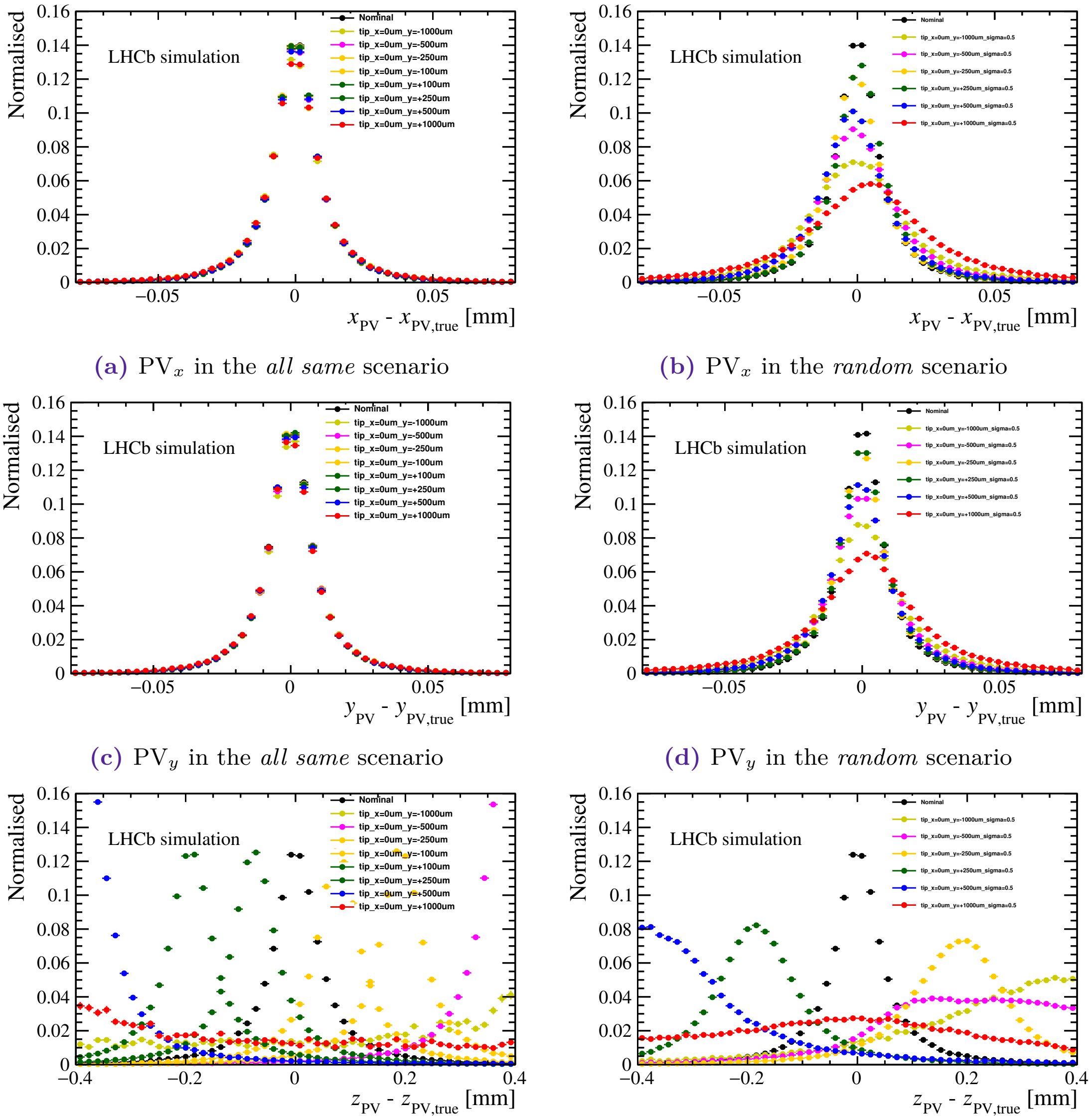
(d) Slope of IP resolution

Figure 4.8: Combined results for the different scenarios showing the evolution of the resolution of the impact parameter resolution as a function of the misalignment scenario. The lower plots show a smaller range of misalignment magnitudes with the *alternate* scenario included.

4.3.2 PV resolution

Primary vertices are built in LHCb using tracks formed of only VELO hits and are therefore exclusively dependent on the performance of the VELO. In order to assess the primary vertex resolution, plots were produced to compare the true and reconstructed position of the primary vertex. This method relies on Monte Carlo truth information and therefore a global shift in the PV position is visible. For the purposes of this section, only effects on the resolution is of interest. As can be seen in Figure 4.9 the results for PV$_x$ and PV$_y$ are consistent with those in Section 4.3.1 with only varied distortions showing any appreciable degradation in performance.

For PV$_z$ a clear effect is seen with the primary vertex being biased by around 80 % of the tip displacement. If the bias is disregarded the resolution is again only affected for varied distortions, although the smearing is more significant than that seen for PV$_x$ and PV$_y$.



(a) PV$_x$ in the *all same* scenario

(b) PV$_x$ in the *random* scenario

(c) PV$_y$ in the *all same* scenario

(d) PV$_y$ in the *random* scenario

(e) PV$_z$ in the *all same* scenario

(f) PV$_z$ in the *random* scenario

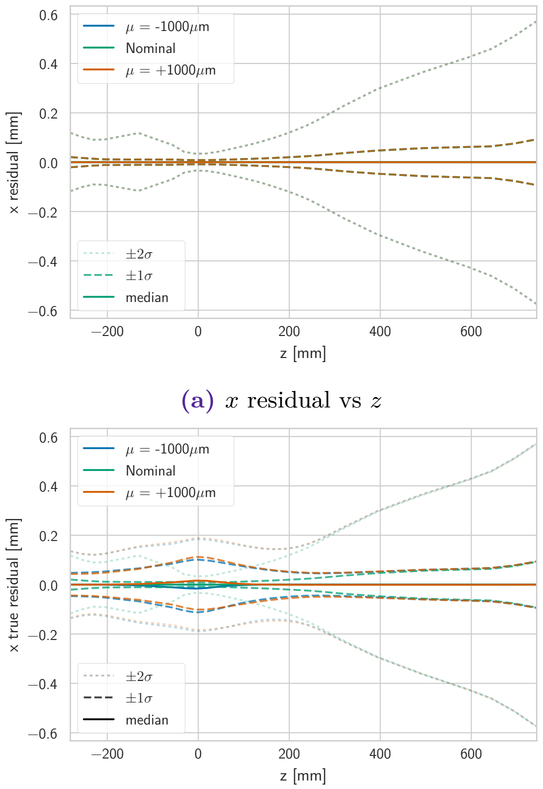
Figure 4.9: Primary vertex resolution for the *all same* and *random* scenarios.

4.3.3 Track residuals

In order to assess the effect on the reconstructed tracks two different definitions of the track residuals were considered. Firstly, the biased residual between the track and the misaligned cluster position was calculated as a function of z (Figure 4.10a) and this shows a degradation of performance within 200 mm of the PV and a slight improvement further upstream. The origin of the apparent improvement in the forward VELO region may be caused by a loss of tracking efficiency for already poorly reconstructed tracks.

The second residual considered is again a *partially biased residual*[1] corresponding to the minimum distance between the reconstructed track and cluster, however, in this case

[1] This is described as a partially biased residual as the track is biased to the cluster position using the misaligned geometry.



(a) x residual vs z

(b) "True" x residual vs z

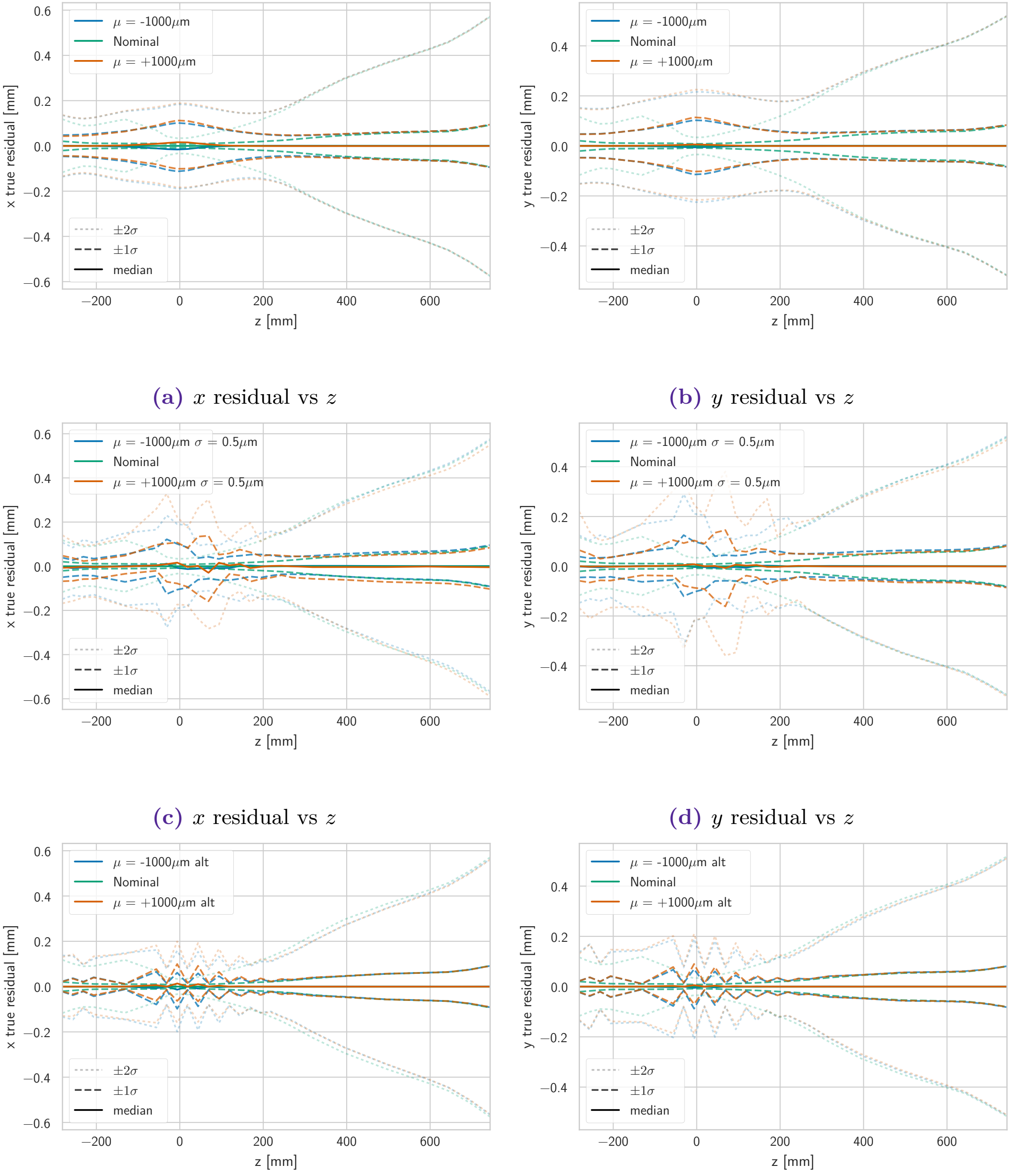
Figure 4.10: Comparison between the residuals obtained using the true and reconstructed cluster positions for the scenario where the tip of alternate planes have been displaced by ±1000 µm.

the cluster position is extracted using perfect detector alignment (Figure 4.10b) and is referred to as the *true residual*. All further plots shown correspond the *true residual*.

Figure 4.11 shows the true residuals for each of the three classes of distortion with a mean displacement of ±1000 µm. All show a degradation of performance near the interaction point. In the forward VELO region the effect is hidden due to the nominal resolution being worse. This is caused by the angular distribution of the tracks tending to intercept the forward sensors at large angles. In the *all same* scenarios the residual smoothly varies with z whereas in the *random* case a discontinuous distribution is obtained with the exact displacement of each sensor directly affecting the residual at each point. This can be best seen in the *alternating* (Figure 4.11e and 4.11f) scenario where the residual approximately varies between that seen in the *all same* and *random* scenarios and depending on if the sensor was held fixed.

4.3.4 Tracking performance

The tracking performance can be characterised by using many quantities, such as the clone rate, ghost rate and track finding efficiency. In addition, these can be calculated on different groupings of tracks, like the track type or by using some physical property such as the track momentum or the quark content of the corresponding particle. These options



(a) x residual vs z

(b) y residual vs z

(c) x residual vs z

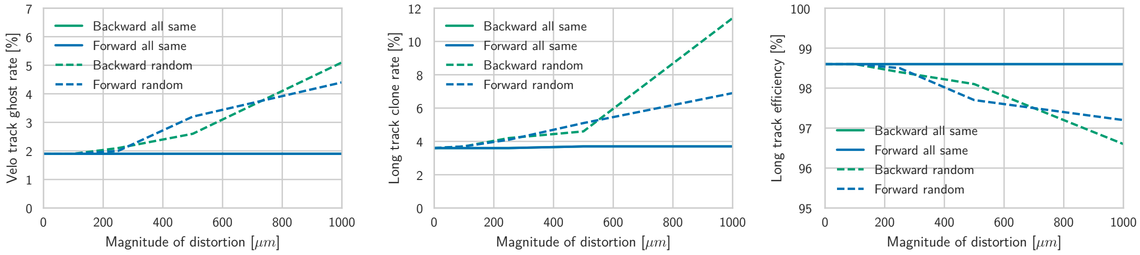
(d) y residual vs z

(e) x residual vs z

(f) y residual vs z

Figure 4.11: Top down (a, c, e) and side on (b, d, f) plots of the residual between the true cluster position and the reconstructed track for the *all same* (a, b), *random* (c, d) and *alternating* (e, f) scenarios. All misalignments are shown with a displacement magnitude of ±1000 µm.

have varying sensitivity to different effects and for this study the ghost rate of VELO tracks and the clone rate and track finding efficiency of long tracks were considered in detail as these are the most important track types for primary vertex fitting and general analyses. The tracking performance was assessed using $\mathcal{O}(250{,}000)$ tracks and the results are shown in Figure 4.12 as a function of the distortion. The *clone* and *ghost* rate are defined in Section 3.2 and the track types shown are defined as follows:



(a) VELO track ghost rate **(b)** Long track clone rate **(c)** Long track efficiency

Figure 4.12: Tracking ghost rate, clone rate and efficiency as a function of scenario.

- **VELO tracks:** Tracks created using only information from the vertex locator.
- **Long tracks:** Tracks that include hits in both the vertex locator and the post magnet tracking stations.

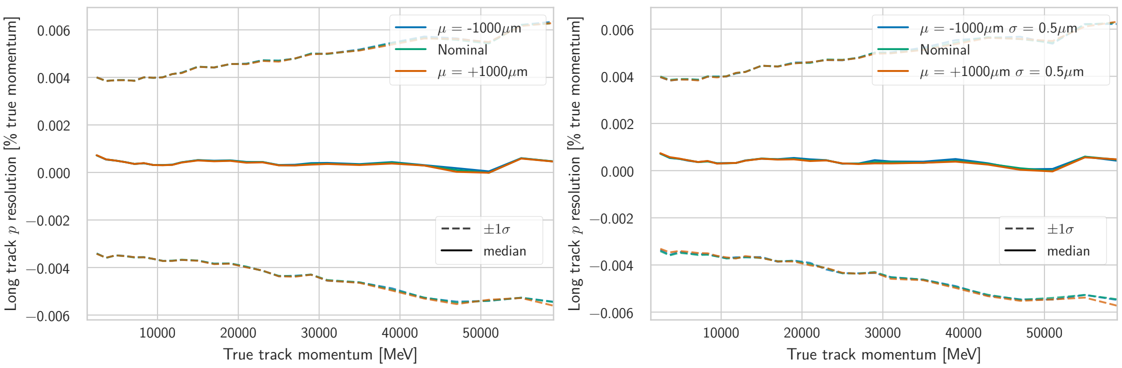
The three quantities studied are all affected in similar ways and consistent with the results found in the other sections of this note, with the tracking performance degrading significantly for the *random* scenario with a variation with standard deviation greater than 125 µm.

4.3.5 Momentum resolution

In LHCb the VELO cannot be used directly to make momentum measurements of particles as the LHCb magnet is placed around 4 m away, causing there to be a minimal magnetic field at the position of the VELO modules. Instead, momentum measurements are made as a result of particles being curved by passing through the magnetic field between the VELO and the later tracking detectors. As such, it is conceivable that a weak mode in the VELO could result in larger effects when tracks are propagated to the remainder of the detector, as a result of a misalignment-driven-shift between the reconstructed z positions of the VELO and the downstream tracking detectors. This would result in the measured momentum of tracks being smeared or biased and was tested by plotting the reconstructed momentum resolution of long tracks as a function of true momentum. Figure 4.13 shows the effect on the measured momentum is negligible for all situations considered.

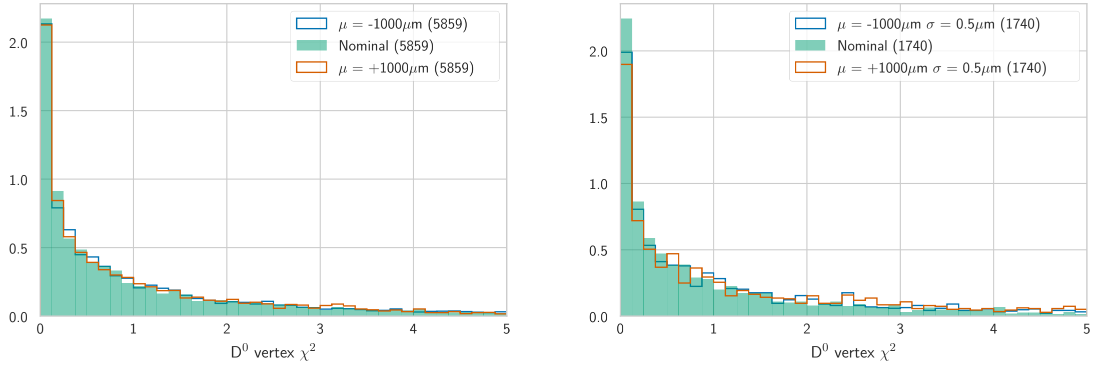
4.4 D^0 lifetime measurement

In this section a model analysis is performed to measure the D^0 lifetime using simulated data under each misaligned scenario. The analysis is performed using `Panoramix` and reconstructed candidates are formed using truth matching with the standard reconstruction. To minimise the effect of statistical fluctuations on the results show in this section, each plot corresponds to exactly the same underlying simulated particles and hence the results are highly correlated. If a signal event is not reconstructed for one or more of the scenarios, it is rejected from all datasets used in the plot.



(a) All same **(b)** Varied randomly

Figure 4.13: Median and ±1σ intervals of the momentum resolution for all long tracks as a function of track momentum.

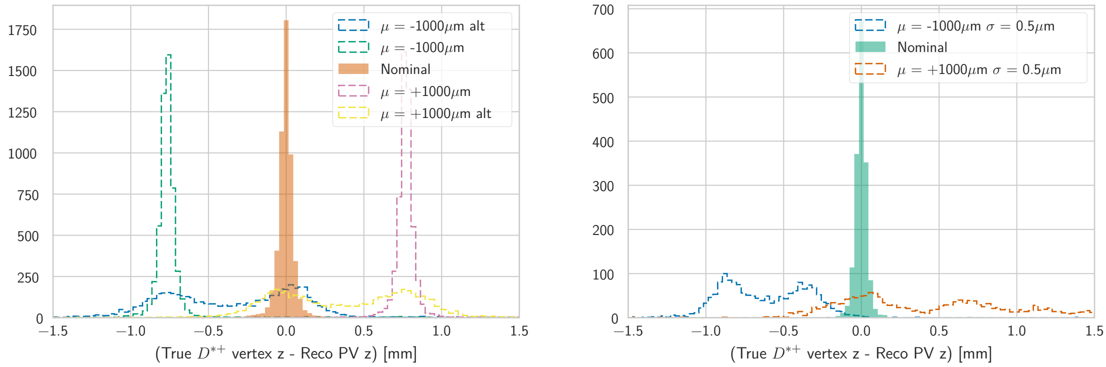


(a) *All same* scenario **(b)** *Random* scenario

Figure 4.14: The D^0 vertex fit quality obtained when forming a vertex from the two reconstructed and truth-matched kaon tracks, using simulation of the *all same* and *random* scenarios outlined in Section 4.1. The values in the legend surrounded by parentheses gives the number of candidates in each sample.

4.4.1 D^0 decay vertex

In order to measure the reconstructed lifetime, the decay vertex of the D^0 meson must first be found and fitted. This was implemented by iterating over all well reconstructed tracks, using truth matching to find those corresponding to a single real $D^{*+} \to \left(D^0 \to K^+K^-\right)\pi^+$ decay. To ensure that all of the reconstructed D^0 candidates originated from the primary vertex a requirement was made such that parent particles to the D^0 must have lifetimes of less than 1×10^{-7} ns. Additionally, a veto was made on D^0 candidates that came from a primary vertex more than 15 µm away in x or y from the true D^{*+} decay vertex to remove candidates that are incorrectly associated to a primary vertex. The D^0 decay vertex was then fitted from the two kaons and the vertex fit quality is shown in Figure 4.14. As in Section 4.3 the effect was minimal in the *all same* scenarios, however *random* scenarios resulted in a degradation in the quality of the reconstructed vertex.



(a) *All same* and *alternating* scenarios

(b) *Random* scenario

Figure 4.15: Difference between the z position of the true D^{*+} vertex and the associated primary vertex for the *all same* and *alternating* scenarios (4.15a) and the *random* scenario (4.15b).

4.4.2 D^{*+} decay vertex

The D^{*+} decay vertex was then chosen to be the reconstructed primary vertex closest to the slow pion, where the slow pion was found using truth matching, as described in Section 4.4.1.

Figure 4.15a shows the difference between the true D^{*+} decay vertex and the primary vertex associated using this method. In the *all same* scenario a shift in the mean of the distribution is seen, with the resolution remaining unaffected. This is expected as this is a weak mode of the vertex locator geometry and all fully reconstructed quantities are expected to be insensitive. As expected from the results in Section 4.3, the *alternating* scenario shows a large effect with the primary vertex resolution being significantly degraded. Despite this a peaking structure can be found that peaks like a mixture of the two corresponding *all same* scenarios with a significant smearing to the resolution. The peaks are likely caused by primary vertices being dominated by tracks fitted from clusters that are on planes that all are in the same position as an *all same* scenario. If their vertex is instead formed using an equal mixture of planes from each scenario the reconstructed position ends up in between the two peaks. The *random* scenario is the generalised case, where each plane is from a different scenario with a complex structure being formed (Figure 4.15a). This distribution is very sensitive the combination of misalignments generated for each plane, weighted by its relative importance in the primary vertex reconstruction.

4.4.3 D^0 lifetime

Once both the D^0 and D^{*+} decay vertices are found the flight distance can be calculated using the distance between them. This is then combined with the reconstructed D^0 momentum to calculate the proper lifetime using

$$t_{\text{proper}} = \frac{\text{flight distance} \cdot \gamma}{c} \tag{4.1}$$

where

$$\gamma = \frac{1}{\sqrt{1 + \frac{p}{m_{D^0}}^2}}. \tag{4.2}$$

The lifetime for each scenario is then shown in Figure 4.16 with the resolution plot comparing the reconstructed lifetime with the Monte Carlo truth information. An estimate of the measured lifetime is shown in the legend and obtained by fitting the lifetime with an exponential function for candidates with lifetimes less than 3 ps. From these plots it is clear that the lifetime is sensitive to varying displacement of modules. Figure 4.17 shows the difference between the calculated lifetime for the nominal and the misaligned scenarios as a function of the displacement magnitude. Similarly Figure 4.18 shows the bias for alternating scenarios below 250 µm. The uncertainty on the lifetime bias for each point is obtained using bootstrapping to account for the correlation that arises from using the same underlying D^0 mesons.

4.5 Comparison with Δm_s prospects

In this section the results of Section 4.4 are compared with the most recent published measurement of Δm_s.[147] This measurement is chosen as uncertainties derived from misalignment of the VELO are already known to be a dominant systematic uncertainty and could be larger than the statistical uncertainty with the full upgrade dataset. While measuring Δm_s is unlikely to be of significant importance, as the current LHCb measurement already exceeds the precision of the currently available theoretical predictions, it remains a useful indicator of potential issues that could affect other time dependent measurements. To make this comparison the following assumptions are made:

1. The upgrade LHCb VELO will be used to collect 50 fb^{-1} and the efficiency for the reconstruction and selection of $B_s^0 \to D_s^- \pi^+$ decays will remain similar to that used for [147].
2. The absolute error from these misalignments is independent of the meson lifetime, i.e. the resulting relative uncertainty can be scaled by the ratio of the D^0 lifetime to the B_s^0 lifetime.
3. The relative systematic error on Δm_s is the same as the relative error on the measurement of the B_s^0 lifetime.[148]
4. The bias on the lifetime is linearly correlated with the magnitude of the misalignment.[149]

The maximum allowed systematic on Δm_s that can arise from the misalignments studied here is calculated using Assumption 1 to scale the statistical uncertainty on the current best measurement (17.768 ± 0.023 (stat) ± 0.006 (syst) ps^{-1}). This gives a maximum target uncertainty of:

$$0.023\,\text{ps}^{-1} \times \sqrt{\frac{1\,\text{fb}^{-1}}{50\,\text{fb}^{-1}}} = 3.3 \times 10^{-3}\,\text{ps}^{-1}\,(0.018\,\%)$$

Assumption 2 and 3 can then be used to scale the maximum allowed systematic on Δm_s to a maximum allowed bias that can be compared with the results of Section 4.4. Using the mean lifetime of D^0 and B_s^0 mesons taken from [150] the maximum allowed bias on the measured D^0 lifetime is calculated to be 0.27 ps or 0.067 %. Each of the individual results shown in Figure 4.18 are not sensitive enough to measure systematics at the required scale however using Assumption 4 the data points can be fitted to improve the sensitivity as shown in Figure 4.19. Ideally any misalignment should not exceed the $+1\,\sigma$ line to allow the limit of this systematic to be conservatively known. This gives a maximum displacement, at the tip of the module from a rotation around the cooling block, of 28 µm.[2]

In future it may be possible to further improve this systematic uncertainty, or tolerate larger misalignments, by applying track based alignment techniques or by better replicating the Δm_s analysis instead of relying upon Assumption 3.

4.6 Summary and conclusion

The prototype LHCb VELO upgrade modules have been observed to distort when cooled to their operating temperature, rotating in LHCb global y around their mid-plates. The effect of this distortion has been assessed through a simulation study. A primary conclusion from the study is that if all modules distort in the same manner the VELO upgrade physics performance is insensitive to these distortions, even if the scale of distortion is at the level of hundreds of microns. However, in practice this will only be known once the production is highly advanced, unless the modules positions are mechanically constrained to each other.

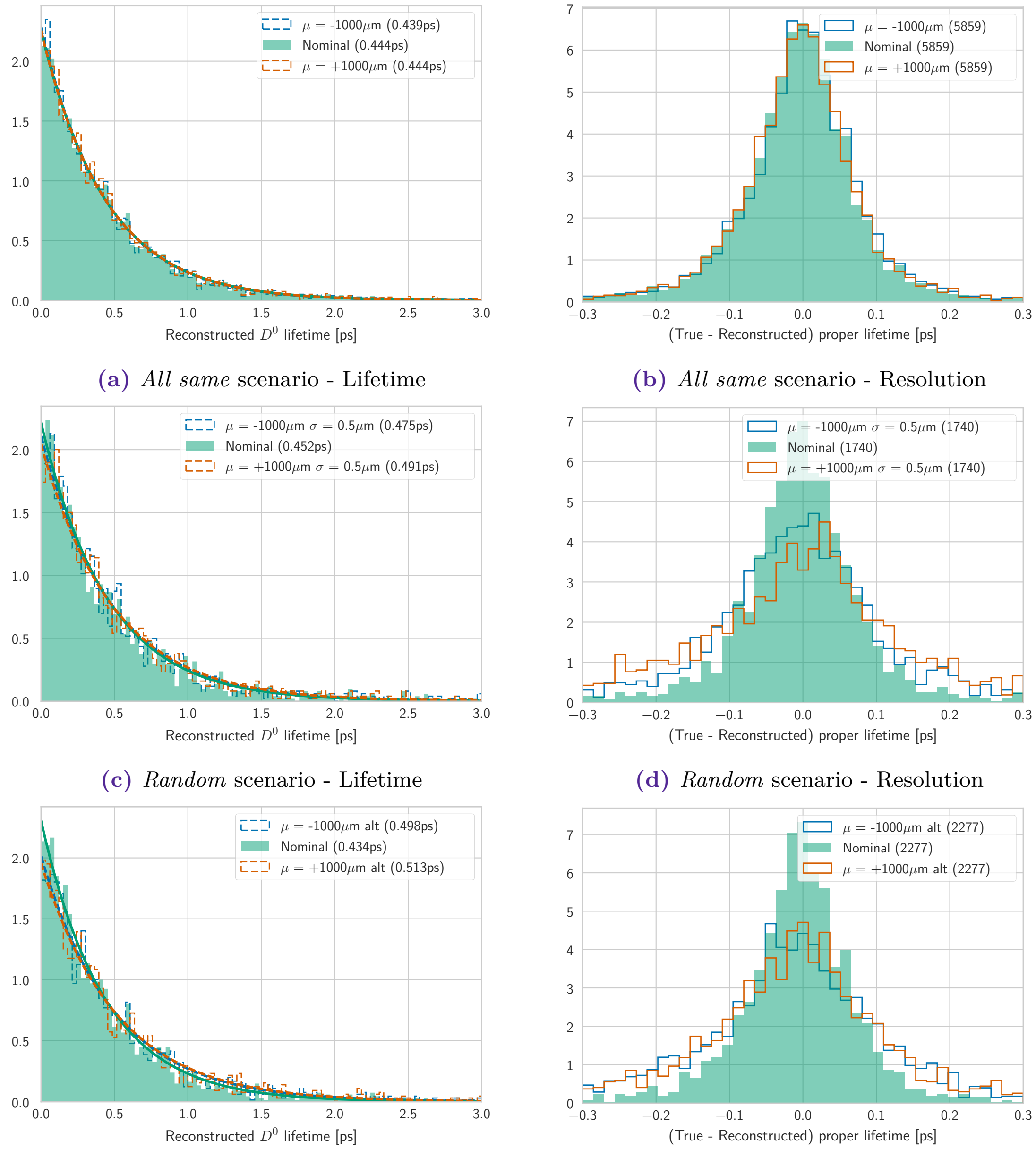
As Plan B module designs contain additional material with respect to Plan A they exhibit worse performance in the absence of other differences. Comparisons between the results in this note and studies into the impact of the material change[151] show that the impact parameter resolution and tracking performance of perfectly aligned Plan B modules is comparable to that of Plan A modules with up to 150 µm misalignments in the z distance between each pair of modules.

This study shows the most stringent constraints on the permitted cooling induced module distortions is from the precise measurement of lifetime-like quantities (Section 4.5). Extrapolating the existing measurement of Δm_s to the full upgrade dataset, and requiring that final statistical uncertainty is no smaller than the systematic uncertainty from the movement of modules due to cooling, conservatively gives a maximum uncertainty on the module tip position between modules of 28 µm. In this study the distortion has not been determined by an alignment procedure, so it is possible that the effect could be partially mitigated but this is currently not known. Additionally, future developments of this study could be useful in better estimating the systematics that arise from detector misalignments.

[2] Similarly the mean and $-1\,\sigma$ uncertainty lines give 35 µm and 45 µm respectively.

It may also be possible to use develop new techniques to correct for these effects and one such method has been studied by the author while supervising a CERN summer student.[152] As particles pass through the detector material there is a small probability of inelastic interactions occurring resulting multiple charged tracks originating from a secondary vertex. These can be reconstructed by the VELO and an example of this technique with simulated data can be seen in Figure 4.20. Misalignment along the beam axis results in the resolution of this image being degraded as shown in Figure 4.21.

After this study was performed the Plan A module substrate was chosen and at the time of writing final preparations are being made to start the module production. This study demonstrated the potential physics effects of relatively modest distortions of the module. Design studies with prototypes have shown how distortions can be reduced and refinements have been made to the design and manufacturing process to minimise potential deformations. The distortions of the tip in the final design and after full population of components are found smaller in the final prototypes, hence allaying fears from the bare module studies. Furthermore, during manufacture, measurements will now be made of each module's deformation as a function of temperature, with the view that this information can be used an input for the detector alignment procedure.



(a) *All same* scenario - Lifetime

(b) *All same* scenario - Resolution

(c) *Random* scenario - Lifetime

(d) *Random* scenario - Resolution

(e) *Alternating* scenario - Lifetime

(f) *Alternating* scenario - Resolution

Figure 4.16: Reconstructed proper lifetime for ±1000 µm under the three different scenario classes

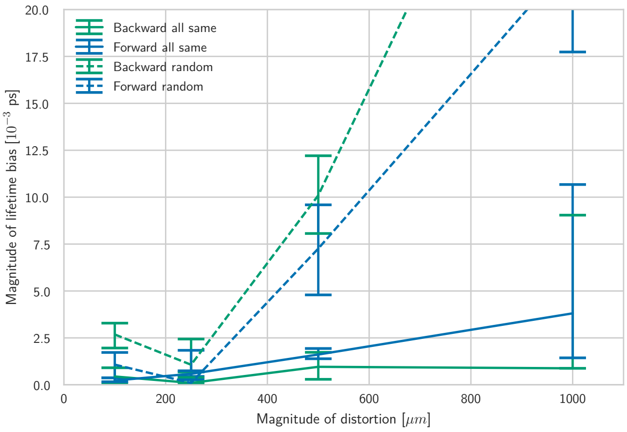


Figure 4.17: Variation of the lifetime bias as a function of the magnitude of the module displacement for *all same* and *random* scenarios with magnitudes up to 1000 µm.

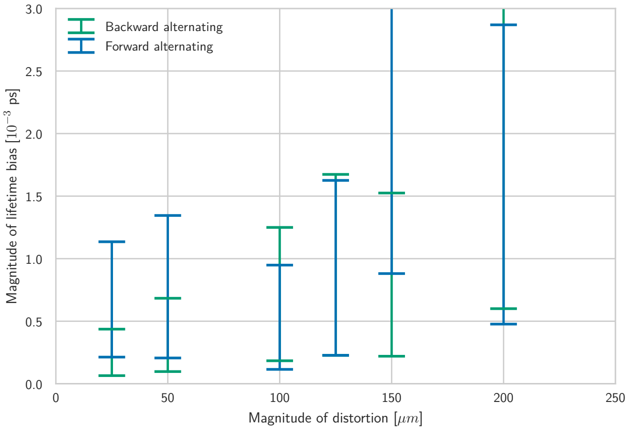
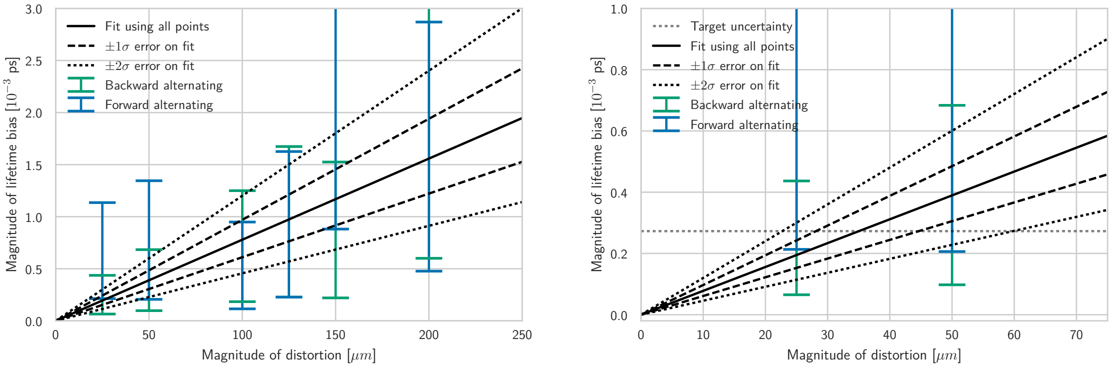


Figure 4.18: Variation of the lifetime bias as a function of the magnitude of the module displacement for *alternating* scenarios with magnitudes up to 200 µm.



(a) Full fit range

(b) Region of interest

Figure 4.19: Figure 4.18 with the addition of a fit using Assumption 4 to better evaluate the maximum acceptable variation. All uncertainties are calculated using bootstrapping to account for the correlation between the underlying D^0 mesons as with Section 4.4.3.

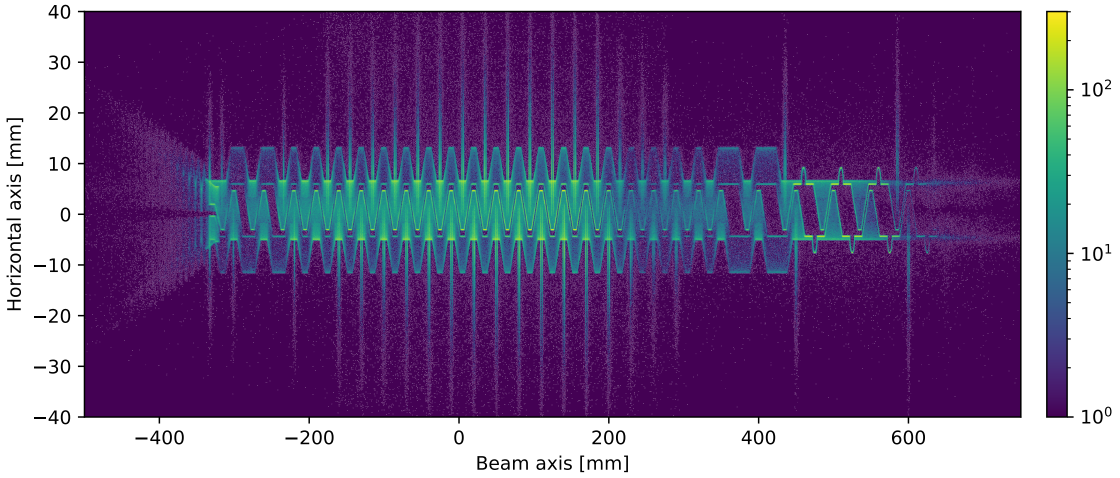
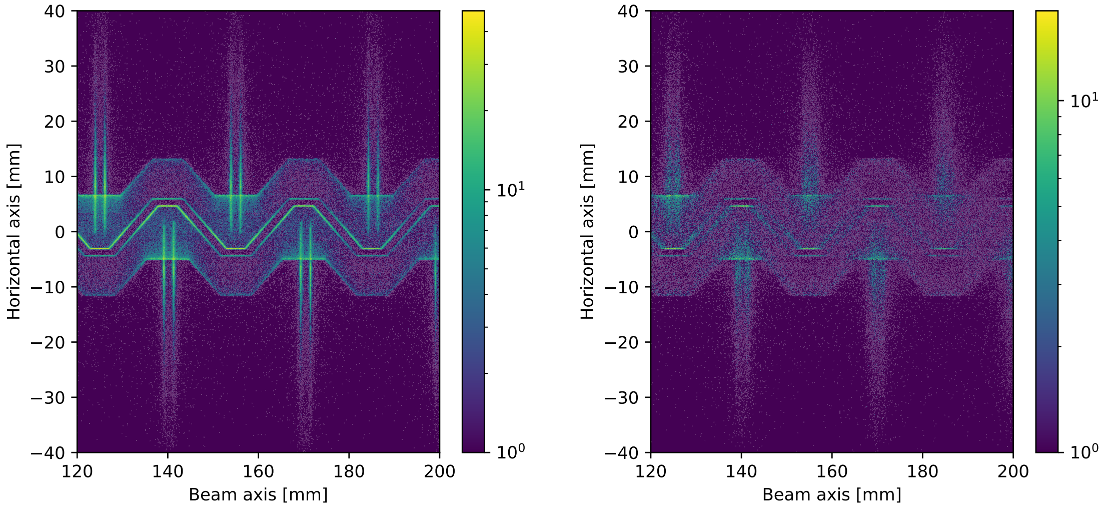


Figure 4.20: Reconstructed vertices that originate from outside the luminous region in simulated beam-gas data for a perfectly aligned detector. The “zig-zag” pattern is caused by the aluminium RF foil and the vertical lines are the VELO modules.



(a) Correct alignment

(b) 10 mm scaling along z

Figure 4.21: Reconstructed vertices that originate from outside the luminous region in simulated beam-gas data with and without misalignment applied.

Blank page

Chapter 5

Analysis methodology of $D^+_{(s)} \to h^\pm l^+ l'^\mp$

This chapter describes the method that is used in a search for $D^+_{(s)} \to h^\pm l^+ l'^\mp$ using LHCb data, where h is a charged kaon or pion and l is a electron or muon. This analysis is performed using $1.5\,\text{fb}^{-1}$ of data that was collected in 2016 and all components were primarily performed by the author of this thesis. Throughout this section the use of natural units ($c = \hbar = 1$) and the inclusion of charge conjugate processes is assumed unless otherwise stated. None of the decays covered by this analysis have previously been observed without proceeding via intermediate resonances and, while the 28 signal channels studied in this analysis all have the same decay topology, the processes by which they can occur varies. Eight of the decays are allowed within the standard model and can occur through a Weak Annihilation (WA) diagram as shown in Figure 5.1. Additionally, four of these eight decays can also occur via Flavour Changing Neutral Currents (FCNC) via the diagrams shown in Figure 5.2 with the light quark acting as a spectator. These decays do not occur at tree level in the standard model and are suppressed at the loop level by the GIM mechanism discussed in Section 1.1.4. In principle the effect of the FCNC diagrams can be disentangled from the WA diagram using the four decays that do not occur via a FCNC however, in practice, these decays will likely be dominated by resonant contributions as discussed in Section 5.1.

While the remaining 20 decays studied here are Lepton Flavour Violating (LFV), eight are technically permitted within the standard model and can occur via an oscillating

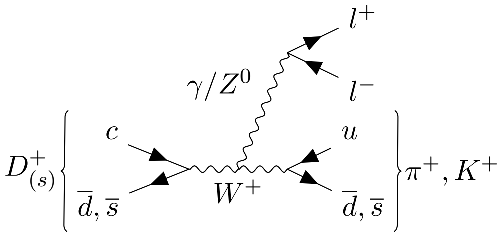
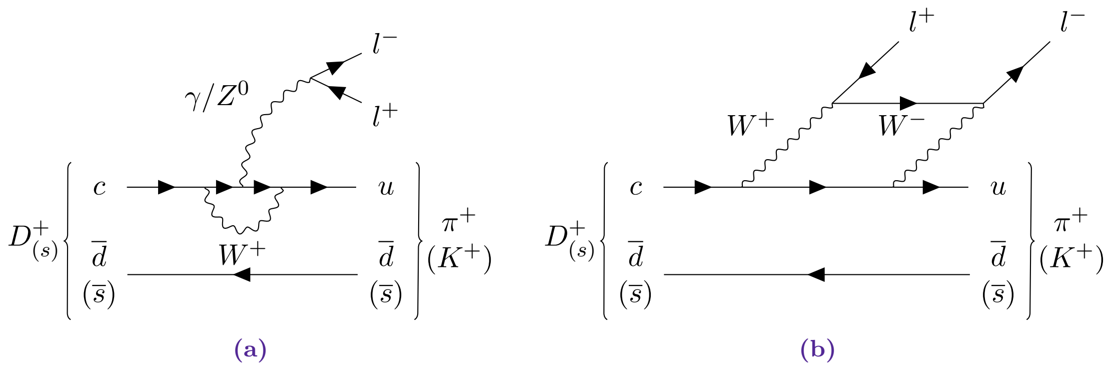


Figure 5.1: Feynman diagram for the eight weak annihilation decays of a $D^+_{(s)}$ meson that are searched for in this analysis.



(a) (b)

Figure 5.2: Feynman diagrams for the four flavour changing neutral current decays of a $D^+_{(s)}$ meson that are searched for in this analysis.

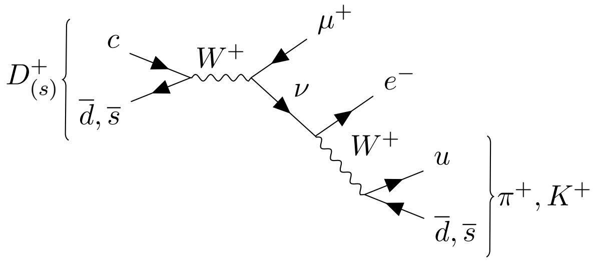


Figure 5.3: Feynman diagram showing how lepton flavour violating decays of a $D^+_{(s)}$ meson can occur within the standard model via an oscillating neutrino. The small branching fraction of such processes would render them inaccessible to any conceivable experimental search.

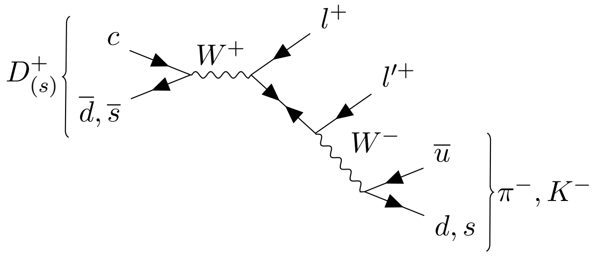


Figure 5.4: Feynman diagram showing how decays of a $D^+_{(s)}$ meson via a Majorana neutrino can result in lepton number and lepton flavour violation.

Channel	Type	Current Limit	Experiment	Resonance
$D^+ \to \pi^+e^+e^-$	FCNC	$< 1.1 \times 10^{-6}$	BABAR [53]	$\phi, 1.7 \times 10^{-6}$
$D^+ \to \pi^+\mu^+\mu^-$	FCNC	$< 7.3 \times 10^{-8}$	LHCb [155]	$\phi, 1.8 \times 10^{-6}$
$D^+ \to K^+e^+e^-$	Weak annihilation DCS	$< 1.0 \times 10^{-6}$	BABAR [53]	Not Seen
$D^+ \to K^+\mu^+\mu^-$	Weak annihilation DCS	$< 4.3 \times 10^{-6}$	BABAR [53]	Not Seen
$D^+ \to \pi^+e^+\mu^-$	LFV	$< 2.9 \times 10^{-6}$	BABAR [53]	-
$D^+ \to \pi^+e^-\mu^+$	LFV	$< 3.6 \times 10^{-6}$	BABAR [53]	-
$D^+ \to K^+e^+\mu^-$	LFV	$< 1.2 \times 10^{-6}$	BABAR [53]	-
$D^+ \to K^+e^-\mu^+$	LFV	$< 2.8 \times 10^{-6}$	BABAR [53]	-
$D^+ \to \pi^-e^+e^+$	LNV LFV	$< 1.1 \times 10^{-6}$	CLEO [54]	-
$D^+ \to \pi^-\mu^+\mu^+$	LNV LFV	$< 2.2 \times 10^{-8}$	LHCb [155]	-
$D^+ \to K^-e^+e^+$	LNV LFV	$< 0.9 \times 10^{-6}$	BABAR [53]	-
$D^+ \to K^-\mu^+\mu^+$	LNV LFV	$< 10 \times 10^{-6}$	BABAR [53]	-
$D^+ \to \pi^-e^+\mu^+$	LNV LFV	$< 2.0 \times 10^{-6}$	BABAR [53]	-
$D^+ \to K^-e^+\mu^+$	LNV LFV	$< 1.9 \times 10^{-6}$	BABAR [53]	-

Table 5.1: D^+ decay channels studied in this analysis. The primary process through which the decay proceeds or the Standard Model (SM) conservation laws that it violates is listed. The current world's best limits and the experiment that provided these results are given. In the case of the FCNC and weak annihilation processes the limits are for the non-resonant contribution only. The final column notes if a resonant contribution has been measured, and in the case that it has been observed specifies the resonance meson and the central value of the BF.

neutrino as shown in Figure 5.3. In practice the branching fraction of such decays would be beyond the reach of any conceivable experimental study and an observation of any of these 20 modes would be conclusive evidence of Beyond the Standard Model (BSM) physics. Interestingly, given the flavour anomalies that are being investigated in FCNC B meson decay[35–37, 153], leptoquark models have been shown to have potential contributions to the modes considered here at the level of the current experimental constraints[24].

The 12 of these decays that are both Lepton Number Violating (LNV) and LFV are forbidden within the standard model however extensions have been predicted that could permit such decays. One example would be via a Majorana neutrino as shown in Figure 5.4.

The world's best limits on these processes prior to the measurement described here are given in Tables 5.1 and 5.2.[1] Only four of these 28 decays had previously been searched for by LHCb[155] and these results improved the previous limits by around a factor of fifty showing the potential LHCb has in this area.

5.1 Resonant contributions

Some of the signal channels studied in this analysis have resonant contributions with the same final state. These resonances occur in the dilepton q^2 distribution and are of the form $D^+_{(s)} \to h^+X$, where X represents the resonances $(\rho, \omega, \phi, \eta)$ that can decay into e^+e^- or $\mu^+\mu^-$. The analysis described here both benefits and suffers from the presence of these

[1] BESIII has recently shown new results for $D^+ \to h^\pm e^+e^\mp$ at ICHEP2018[154] however, at the time of writing, the corresponding paper has not been released.

Channel	Type	Current Limit	Experiment	Resonance
$D_s^+ \to \pi^+ e^+ e^-$	Weak annihilation CF	$< 13 \times 10^{-6}$	CLEO	$\phi, 6 \times 10^{-6}$
$D_s^+ \to \pi^+ \mu^+ \mu^-$	Weak annihilation CF	$< 4.1 \times 10^{-7}$	LHCb [155]	Not seen
$D_s^+ \to K^+ e^+ e^-$	FCNC	$< 3.7 \times 10^{-6}$	BABAR [53]	Not seen
$D_s^+ \to K^+ \mu^+ \mu^-$	FCNC	$< 21 \times 10^{-6}$	BABAR [53]	Not seen
$D_s^+ \to \pi^+ e^+ \mu^-$	LFV	$< 12 \times 10^{-6}$	BABAR [53]	-
$D_s^+ \to \pi^+ e^- \mu^+$	LFV	$< 20 \times 10^{-6}$	BABAR [53]	-
$D_s^+ \to K^+ e^+ \mu^-$	LFV	$< 14 \times 10^{-6}$	BABAR [53]	-
$D_s^+ \to K^+ e^- \mu^+$	LFV	$< 9.7 \times 10^{-6}$	BABAR [53]	-
$D_s^+ \to \pi^- e^+ e^+$	LNV LFV	$< 4.1 \times 10^{-6}$	CLEO [54]	-
$D_s^+ \to \pi^- \mu^+ \mu^+$	LNV LFV	$< 1.2 \times 10^{-7}$	LHCb [155]	-
$D_s^+ \to K^- e^+ e^+$	LNV LFV	$< 5.2 \times 10^{-6}$	BABAR [53]	-
$D_s^+ \to K^- \mu^+ \mu^+$	LNV LFV	$< 1.3 \times 10^{-5}$	BABAR [53]	-
$D_s^+ \to \pi^- e^+ \mu^+$	LNV LFV	$< 8.4 \times 10^{-6}$	BABAR [53]	-
$D_s^+ \to K^- e^+ \mu^+$	LNV LFV	$< 6.1 \times 10^{-6}$	BABAR [53]	-

Table 5.2: D_s^+ decay channels studied in this analysis. The primary process through which the decay proceeds or the Standard Model (SM) conservation laws that it violates is listed. The current world's best limits and the experiment that provided these results are given. In the case of the FCNC and weak annihilation processes the limits are for the non-resonant contribution only. The final column notes if a resonant contribution has been measured, and in the case that it has been observed specifies the resonance meson and the central value of the BF.

decays. They are highly beneficial as they provide excellent control channels to which the signal channels are normalised. However, the resonant regions must be vetoed in the signal searches and the tails of these resonant decays are also expected to be the dominant standard model contribution to the signal q^2 regions of these channels.

In these decay channels the search for the signal decays is performed in the dilepton invariant mass squared q^2 away from the resonances, see Section 5.2.2. The dominant resonant contributions in the tails of the q^2 range above the ϕ peak are due to the ϕ and ρ. The regions sensitive to the signal are taken to be the same as in the previous LHCb publication [155] as they have been adopted for theoretical predictions in the literature [23, 24]. Figure 5.5 shows a comparison of this analysis's q^2 binning with the standard model prediction. The expected sensitivities, from Section 6.4 and 6.5, are between 1×10^{-8} and 4×10^{-6} depending on the parent hadron and final state.

The resonant regions are also separated into three q^2 bins, corresponding to the η, ρ/ω and ϕ regions. The ϕ dominated regions are used for signal normalisation. The branching fractions of these decays are calculated using the world average values[150].

$$\mathcal{B}(D^+ \to \phi\pi^+, \phi \to K^+K^-) = (2.77 \pm 0.10) \times 10^{-3},$$

$$\mathcal{B}(D_s^+ \to \phi\pi^+, \phi \to K^+K^-) = (2.27 \pm 0.08) \times 10^{-2},$$

$$\mathcal{B}(D_s^+ \to \phi K^+, \phi \to K^+K^-) = (8.9 \pm 2.0) \times 10^{-5},$$

$$\mathcal{B}(\phi \to K^+K^-) = (4.89 \pm 0.05) \times 10^{-1},$$

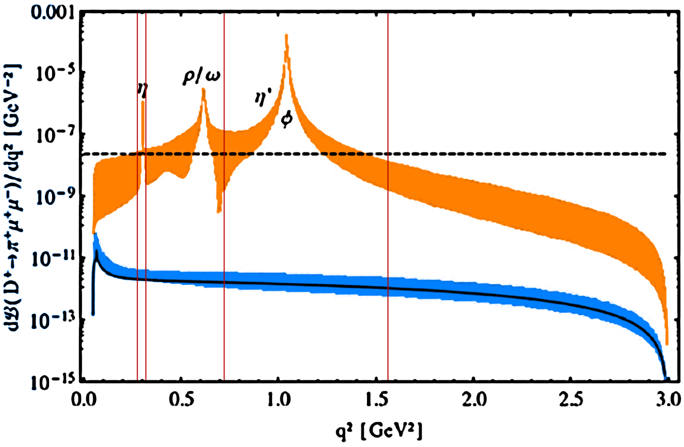


Figure 5.5: Theory predictions from Reference [24] for $\mathrm{d}\mathcal{B}\left(D^+\to\pi^+\mu^+\mu^-\right)/\mathrm{d}q^2\left(\mu^+\mu^-\right)$ with this analysis's q^2 binning overlaid. Three resonant region bins are selected, the bin containing the η, the bin containing the ρ and ω, and the bin containing the ϕ. The solid blue curve is the non-resonant prediction, the orange band is the pure resonant contribution and the dashed black line shows the previous LHCb 90 % CL limit[155] of 7.3×10^{-8}.

$$\mathcal{B}(\phi\to e^+e^-) = (2.954\pm0.030)\times10^{-4},$$

$$\mathcal{B}(\phi\to \mu^+\mu^-) = (2.87\pm0.19)\times10^{-4}.$$

The corresponding branching fractions proceeding through a ϕ resonance to the following final states are:

$$\mathcal{B}(D^+\to\pi^+e^+e^-) = (1.67\pm0.07)\times10^{-6},$$

$$\mathcal{B}(D^+\to\pi^+\mu^+\mu^-) = (1.63\pm0.12)\times10^{-6},$$

$$\mathcal{B}(D_s^+\to\pi^+e^+e^-) = (1.37\pm0.05)\times10^{-5},$$

$$\mathcal{B}(D_s^+\to\pi^+\mu^+\mu^-) = (1.33\pm0.10)\times10^{-5},$$

$$\mathcal{B}(D_s^+\to K^+e^+e^-) = (5.38\pm1.21)\times10^{-8},$$

$$\mathcal{B}(D_s^+\to K^+\mu^+\mu^-) = (5.22\pm1.22)\times10^{-8}.$$

For the 20 decays in which there is no SM contribution no q^2 cuts are required. Should new physics signals be observed in other channels the q^2 distributions in dileptons and opposite sign hadron-lepton would be examined for any indications of peaking contributions that could shed light on the nature of the new physics contribution.

5.2 Experimental overview

Experimentally the decay channels that are searched for all have the same topology. The three final state tracks are required to come from a common vertex. Hence, the processes searched for are not optimised for contributions from long-lived new physics particles, such as a long-lived Majorana neutrino in LNV decays.

The reconstruction of the momentum of muons is significantly more accurate than that of electrons at LHCb. Hence, an experimentally useful classification is into the eight decays with **two electrons** in the final state (four for D^+ and four for D^+_s), the twelve decays with **one electron and one muon**, and the eight decays with **two muons**. This classification is useful for fitting the expected signal mass distributions, and is used in Section 5.4.1.

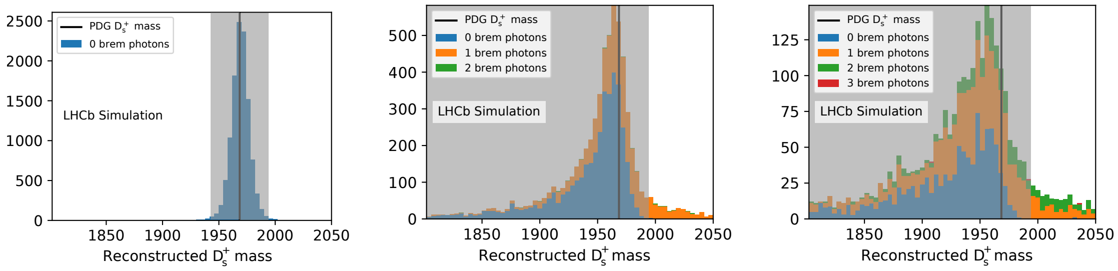
A further experimentally useful classification is into groups of channels where similar background processes are expected to contribute. This classification is by hadron type (pion or kaon) and whether the leptons have the same charge and is discussed further in Section 5.4.2.

This analysis is performed using data that has been processed as part of the central LHCb stripping campaigns described in Section 2.11. The data is then further selected using a multivariate classifier and Particle IDentification (PID).

Signal channel efficiencies are extracted from simulation generated using a minimum bias generation with generator level selections applied to ensure the decay products are within the LHCb acceptance. The agreement of the simulation with data is improved by applying a multivariate reweighting technique using `hep_ml`.[156] These corrections are applied using a number of kinematic distributions, track multiplicity and reconstruction parameters. Four different corrections are computed using the data from each of the control channels: $D^+_{(s)} \to (\phi \to \mu^- \mu^+)\ \pi^+$ and $D^+_{(s)} \to (\phi \to e^- e^+)\ \pi^+$. In addition, the PID variables in simulated data are sampled from kernel density estimates of real data. This is described in Section 5.5.

The $D^+_{(s)} \to (\phi \to \mu^- \mu^+)\ \pi^+$ decay channel is used as the normalization channel for all signal decay modes using the known branching fraction for this mode. It is assumed that, after the simulation reweighting is applied, the simulation correctly models the selection efficiency for muons. A systematic for this assumption is applied for mixed final states, see Section 6.1.4. Under this assumption, the second pair of control channels, $D^+_{(s)} \to (\phi \to e^- e^+)\ \pi^+$, are used to correct the simulation efficiency for electrons. Having applied this correction the branching fraction of all channels is determined in comparison with $D^+_{(s)} \to (\phi \to \mu^- \mu^+)\ \pi^+$.

Signal yields are extracted using one dimensional maximum likelihood fits to the invariant mass of the $D^+_{(s)}$. These likelihoods are then used to obtain 90 % confidence limits using the CL_s method, described in Section 6.2, with the systematic uncertainties included as nuisance parameters. Systematic uncertainties are included for the background model, finite simulation statistics, the normalisation channel branching fractions, the normalisation channel yield, the track reconstruction efficiency and the probability density function used in the fit to model the signal component. An additional systematic is applied for channels containing an electron and a muon where there is no direct control channel.



(a) $D_s^+ \to K^-\mu^+\mu^+$ **(b)** $D_s^+ \to K^-\mu^+e^+$ **(c)** $D_s^+ \to K^-e^+e^+$

Figure 5.6: Reconstructed mass plots using truth matched 2016 simulation for three example decay channels studied in this analysis illustrating the differing resolutions for final states containing two muons (a) one muon and one electron (b) and two electrons (c).

5.2.1 Blinding

A common concern when performing searches in high energy physics analyses is that the results might be biased by statistical fluctuations. Nominally these effects should be quantified by the uncertainties on the final result however, unconscious biases can occur when developing an analysis. A example of this would be to choose a selection that happens to result in a signal peak or choosing a background parametrisation that suppresses an unexpected signal. To counteract this, many analyses in high energy physics are performed without knowledge of the final result. This is known as performing a *blind analysis*[157] and can be achieved in a multitude of ways. The simplest method of blinding is to hide part or all of the data while developing an analysis making it impossible to be influenced by statistical fluctuations or unexpected results. Once the analysis method has been fully finalised it can be applied to the real data to obtain the final result. Ideally this should then be published regardless of how unexpected it might be, ensuring that pre-existing biases do not influence the result. Depending on the situation more complex strategies can be used such as, assigning false labels to the data or injecting a unknown quantity of simulated data. See Reference [157] for an overview of these methods.

The analysis described in this chapter was performed with the signal regions blinded. In the case of final states containing two muons this was simply achieved by removing candidates from two mass regions, i.e those contained by the intervals:

- D^+: 1844 MeV $< M <$ 1896 MeV
- D_s^+: 1942 MeV $< M <$ 1994 MeV

In the case of channels containing electrons this strategy was unsuitable as there exists no region in $M_{D^+_{(s)}}$ where signal events cannot be found: either due to the failure to reconstruct bremsstrahlung photons causing a long low mass tail, or the incorrect addition of bremsstrahlung photons producing a high mass tail. In order to work around this it was decided to blind all candidates with electrons with bremsstrahlung photons recovery. Additionally, all events below the upper sideband ($M_{D^+_{(s)}} <$ 1994 MeV) were also blinded, as illustrated in Figure 5.6.

Bin name	Lower edge	Upper edge
`low m`	250 MeV	525 MeV
η	525 MeV	565 MeV
ρ/ω	565 MeV	850 MeV
ϕ	850 MeV	1250 MeV
`high m`	1250 MeV	2000 MeV

Table 5.3: Dilepton invariant mass binning used for the allowed channels in this analysis.

This choice of blinding strategy results in the dataset being unsuitable for evaluating the backgrounds present in final states containing electrons, to work around this limitations it was decided to develop the analysis using the dimuon final states under the assumption that the backgrounds should be similar when a muon is substituted for an electron.

To validate this assumption a second dataset was used, corresponding to 300 pb^{-1} of data collected by LHCb during 2015. This dataset was not used for the final result and was processed using inputs that are only valid when used with 2016 data, such as the simulated data samples. This makes the precise values of the results obtained with this dataset unreliable, the data set is purely used to check the analysis strategy and provide cross-checks. Despite this, using the 2015 dataset allows this analysis to be performed while keeping the 2016 analysis blind. This resulted in the analysis being split into three stages:

1. First, develop the analysis keeping both the 2015 and 2016 datasets blind.
2. Once the main steps of the analysis are understood, in particular the signal and background model, the 2015 dataset can be unblinded to allow the analysis to be validated. At this stage there is the potential to observe signals that occur at a lower rate than the previous world's best limit.
3. After the analysis framework has been fully developed and the procedure has undergone the collaboration's internal review procedure, the 2016 dataset is unblinded. In practice, at this stage further refinements were needed to the fit model to account for unexpectedly large background contributions and this was achieved by allowing some parameters to float in the final fit. These parameters were originally fixed due to the 2015 dataset not providing enough data to constrain their values. The text here describes the final fit model applied.

5.2.2 q^2 binning

Binning is used to remove possible resonances in the channels with SM contributions ($D^+_{(s)} \to \pi^+\mu^+\mu^-$, $D^+_{(s)} \to K^+\mu^+\mu^-$, $D^+_{(s)} \to \pi^+e^+e^-$ and $D^+_{(s)} \to K^+e^+e^-$) using the bins defined in Table 5.3. These bins are shown overlaid in Figure 5.5. The `low m` and `high m` bins are combined when calculating the fitted branching fraction by correcting the efficiency under the assumption the phase space of the decay is uniform.

5.3 Selection

When signal candidates are created by naively fitting a vertex using three reconstructed tracks a huge fraction of the candidates originate from various forms of background. These candidates are often split into two categories:

- **Combinatorial background** is formed of random tracks that do not originate from the same underlying decay process. For most LHCb analyses this is the main background component.
- **Physical backgrounds** arise from true decays from processes other than the signal that is being studied.

To remove these backgrounds *selections* are used to filter candidates by making requirements on their measured properties. This section will describe the selection process for the candidates used in this analysis.

5.3.1 Trigger

More than 15 % of primary vertices in LHCb contain over 20 charged particles[158] and many events contain multiple primary vertices. This results in a huge number of potential three track combinations. As described in Section 2.10, a trigger system is necessary to reduce these combinations to an level that is acceptable for permanent storage.

The determination of the best triggers lines to require for each channel was performed using simulated data by considering the efficiency of all trigger lines relative to the stripping selection (discussed in Section 5.3.2). This was done separately for all available `L0`, HLT1 and `HLT2` triggers and the conditions are applied by requiring that at least one of the particles in the $D^+_{(s)}$ candidate has been flagged by at least one of the triggers.

Level 0

For the hardware trigger only a small number of trigger lines are available[68] and most analyses do not develop a dedicated line to select the relevant signal. From the aforementioned studies it was decided to choose the triggers for each signal channel according to the logical `OR` of the following criteria:

- **All channels:** `L0Hadron` requires that there is a energy deposit in the hadronic calorimeter with greater than 3.7 GeV of transverse energy.
- **Has 1 or more muons:** `L0Muon` requires that a collection of hits exists in the muon stations with more than 1.8 GeV of transverse momentum. The transverse momentum is estimated from the track slope assuming the particle came from the primary vertex. This approximation has a momentum resolution of approximately 25 %.
- **Has 1 or more electrons:** `L0Electron` is similar to the `L0Hadron` trigger, except the deposit is searched for in the electromagnetic calorimeter and is required to have transverse energy greater than 2.4 GeV.

- **Has 2 muons:** `L0DiMuon` uses the same procedure as the `L0Muon` trigger except the two highest momentum tracks are used. The product of these two momentums is required to be greater than $2.25\,\mathrm{GeV}^2$.

The thresholds vary between years and those listed above are valid for the majority of 2016 data taking.[159] Additionally, most `L0` triggers are found to be inefficient for high multiplicity events. As these events are the most time consuming to reconstruct in the later software stages of the trigger a maximum detector occupancy cut is also made, based on the number of hits found in the SPD detector.

HLT1

For the first stage of the software level trigger (`HLT1`) it is found that the use of specialised triggers does not benefit the analysis. As a result two inclusive `HLT1` trigger lines are used. These triggers are designed to select charged tracks that do not originate from the primary vertex and were improved for LHC Run 2 with the addition of multivariate classification techniques[159, 160] and are defined as follows:

- `Hlt1TrackMVA` This trigger selects a single track by making a hyperbolic requirement in a 2D plane formed of the track's transverse momentum and displacement from the primary vertex.
- `Hlt1TwoTrackMVA` This trigger selects a pair of tracks that appear to originate from the same displaced vertex. A MatrixNet classifier[161] is used to select candidates based on the: vertex fit quality and displacement, the scalar sum of the two tracks' transverse and the primary vertex displacement of each track.

HLT2

For the second stage of the software trigger (`HLT2`) dedicated trigger lines are implemented for all channels. This corresponds to 14 trigger lines with each trigger line covering both the D^+ and D^+_s decay. At the beginning of 2016 six additional lines were added to select "unphysical" channels where all final state particles have the same charge, i.e. $D^+_{(s)} \to h^{+}l^{+}l'^{+}$. These unphysical combinations are useful for studying backgrounds. This analysis's `HLT2` triggers each use similar selection criteria, most notably:

- $\frac{\chi^2_{\mathrm{track}}}{N_{\mathrm{dof}}} < 3$: The fit quality of each of the decay product tracks.
- Track$P_T > 300$MeV: The reconstructed transverse momentum of each of the final state tracks.
- Track$P > 2000$MeV: The reconstructed magnitude of the three-momentum of each of the final state tracks.
- $\mathrm{IP}_{\mathrm{min}}\frac{\chi^2}{N_{\mathrm{dof}}} > 5$: The Impact Parameter (IP) is the distance of closest approach between the reconstructed decay product and the associated primary vertex. This requirement ensures all tracks are consistent with originating from a displaced vertex.

- DIRA > 0.9999: The cosine of the angle between a line drawn from the primary vertex to the decay vertex of the $D^+_{(s)}$ candidate and the sum of the 4-momentum of its decay products. If the decaying particle originates from the primary vertex this angle should be consistent with zero.
- $\text{DOCA}_{\text{max}} < 0.15\,\text{mm}$: The maximum distance of closest approach between all pairs of tracks. If the tracks originate from the same $D^+_{(s)}$ meson this quantity should be small.
- Vertex displacement $\chi^2 > 20.0$: The reconstructed decay vertex of the $D^+_{(s)}$ meson must be displaced from the associated primary vertex. This quantity is used in units of the vertex fit quality.

Additionally, a requirement was made that all final state tracks are associated with the same PV. This requirement contained a bug for the first half of Run 2[2] data taking and effectively required that there only be a single primary vertex in the event.[162] The main effect on this analysis is a reduction in signal efficiency though it also affects the reweighting procedure as described in Section 5.5.1.

5.3.2 Offline selection

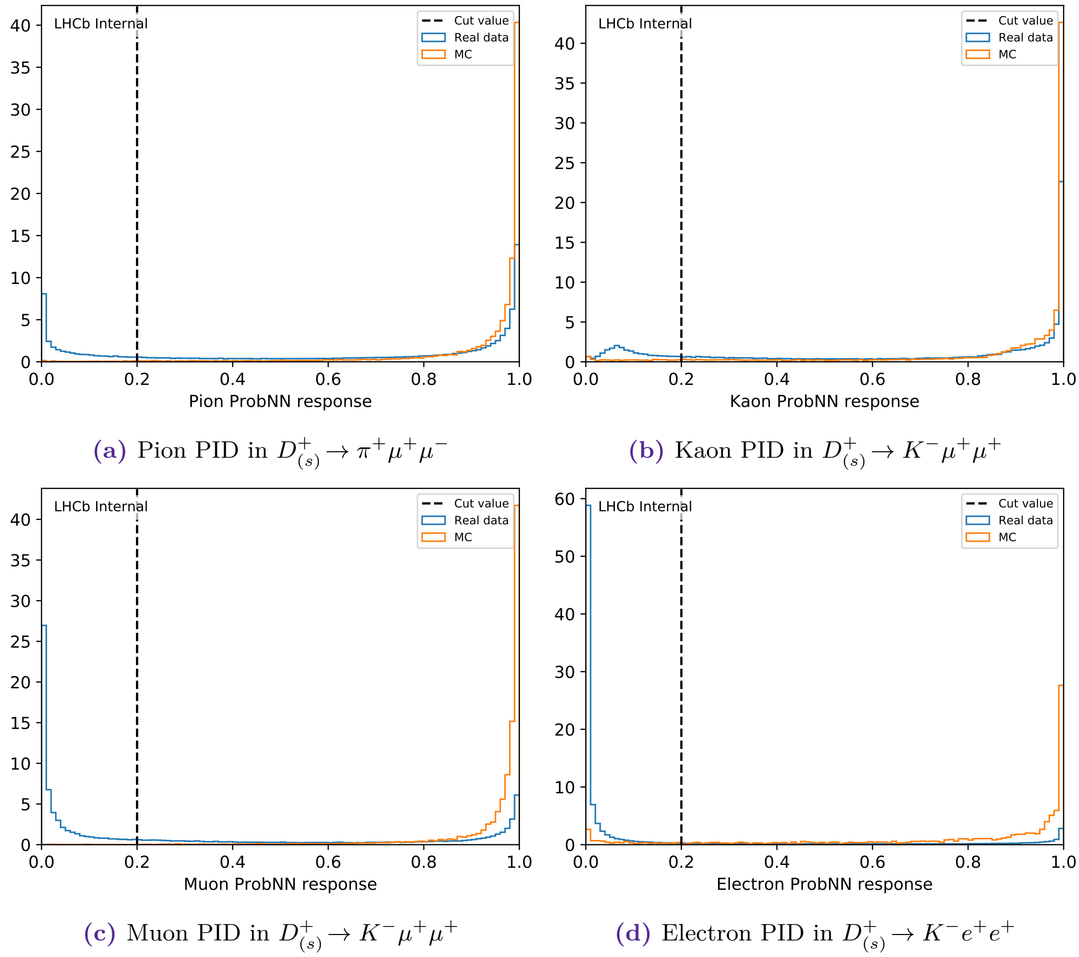
The input data to this analysis is processed as part of the centralised LHCb stripping campaigns that were introduced in Section 2.11.[163] For each final state in this analysis, a dedicated stripping line was used to build $D^+_{(s)}$ meson candidates from reconstructed particles built using the following minimal selection requirements on the reconstructed tracks:

- $\text{IP}\frac{\delta\chi^2}{N_{\text{dof}}} > 5$ The change in primary vertex fit quality between including and excluding each final state track. This requirement ensures all tracks are consistent with originating from a displaced vertex.
- `PIDK-PIDpi` > -1.0 The likelihood that the track was a kaon relative to the likelihood the track is a pion, computed using information from the RICH detectors as described in Section 2.8. This requirement is only used for kaons.

From these containers of $D^+_{(s)}$ candidates a three track vertex is created. The following selection criteria are then made:

- $1763\,\text{MeV} < M < (M_{\text{PDG}}(D^+) + 200\,\text{MeV})$: The invariant mass of the three track combination must be consistent with the PDG[164] value. This is known as the *mass window*. For this analysis it is defined relative to the D^+ mass, however the window is wide enough to also include the D^+_s peak.
- $M(l^+ l'^{\mp}) > 250.0\,\text{MeV}$: The invariant mass between the final state leptons. This requirement is approximately equal to kinematically allowed lower bound.
- $D^+_{(s)}$ decay vertex$\frac{\chi^2}{N_{\text{dof}}} < 5$: The fit quality of the $D^+_{(s)}$ vertex.

[2] From 2015 until mid 2017.



(a) Pion PID in $D^+_{(s)} \to \pi^+\mu^+\mu^-$

(b) Kaon PID in $D^+_{(s)} \to K^-\mu^+\mu^+$

(c) Muon PID in $D^+_{(s)} \to K^-\mu^+\mu^+$

(d) Electron PID in $D^+_{(s)} \to K^-e^+e^+$

Figure 5.7: Normalised distributions of PID variables in background and signal for each species of particle in data and simulation for 2016 data taking conditions.

- $D^+_{(s)}$ IP$\delta\chi^2 < 25$: The change in primary vertex fit quality between including and excluding the $D^+_{(s)}$ candidate. This requirement suppresses contributions from combinatorial background and secondary decays.

Some additional selection criteria were made prior to performing any further processing. A PID requirement was made on all final state particles to shrink the dataset and speed up processing times and required the track probability (`MC15TuneV1_ProbNN`) to be greater than 0.2 for each species of particle. This is effectively an additional stripping requirement, however this was performed offline to allow the *inverse PID* samples to be created for background studies, these samples are discussed in Section 5.4.2. As shown in Figure 5.7 this has negligible impact on the selection efficiency. A further selection is also made on each flavour of lepton:

- Muon tracks are required to have p_T >800 MeV in order to ensure the PID efficiency can be correctly determined as described in Section 5.5.2.
- Electron tracks are required to have $1.9 < \eta < 4$ to avoid regions where the agreement between simulated and real data is poor.[165] A data driven method is used to

compute this efficiency, as described in Section 5.5.3. The efficiency in data is low outside this range and hence the effect on the signal is expected to be small.

5.3.3 Normalisation channel selections

All results in this analysis are obtained as branching ratio limits or measurements relative to $D^+_{(s)} \to (\phi \!\to \mu^-\mu^+)\ \pi^+$. In order to minimise the selection-related systematics introduced by substituting one or more muons for electrons, or the pion for a kaon, the normalisation channel selection was intentionally kept loose. A dimuon q^2 requirement was made, requiring it to be within 20 MeV of the ϕ mass from the PDG (1019.445 MeV).[164] Only loose particle identification selections were applied.[3]

The $D^+_{(s)} \to (\phi \!\to e^-e^+)\ \pi^+$ sample is used for the branching fraction cross-check described in Section 6.3 and for comparing data simulation differences in electrons. Using the equivalent loose selection to that used for $D^+_{(s)} \to (\phi \!\to \mu^-\mu^+)\ \pi^+$ is not suitable as the background contribution is too large. To allow the offline selection to still be validated a simple cut based selection is used of $D^+_{(s)}$ $\texttt{IP}\chi^2 < 5$ and $D^+_{(s)}$ decay vertex $\frac{\chi^2}{N_{\text{dof}}} < 6$ in addition to the aforementioned particle identification requirements. A ±20 MeV dielectron invariant mass cut was found to be unsuitable as it significantly shaped the resulting $D^+_{(s)}$ invariant mass. Instead, a $^{+40\,\text{MeV}}_{-100\,\text{MeV}}$ dielectron invariant mass cut was made around the PDG ϕ mass.

5.3.4 Classifier training

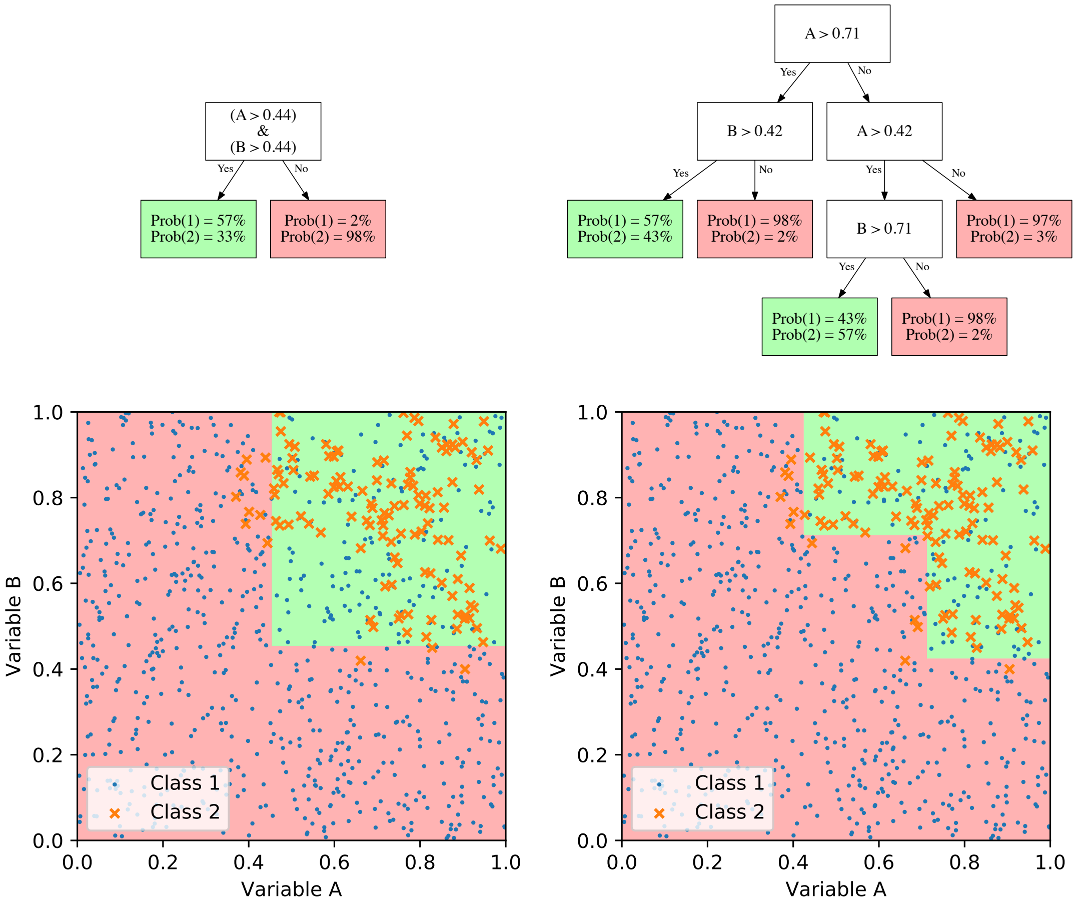
Up until this point in the analysis, the selection criteria have been applied by specifying minimum or maximum values for each quantity.[4] This method of selecting data is often referred to as *rectangular cuts*. A limitation of this strategy is that there can exist correlations between variables that can be exploited to better select signal from background. This is known as a *classification* problem in machine learning where labels (or classes) are predicted for a given data point in a multidimensional space. Many algorithms exist for making these predictions and after evaluation of a wide range of these it was decided to use a boosted ensemble of binary decision trees for this analysis.

A *Binary Decision Tree* (BDT) is a method of making decisions. At each level in the tree a rectangular cut is applied and, instead of using this cut to directly classify the data point, additional cuts can be applied to allow arbitrary complex shapes to be made in the phase space of the problem. This is shown pictorially in Figure 5.8. In practice, the *depth* of the tree often has to be limited to prevent the trees exploiting statistical artefacts of the training data which is known as *over fitting*.

In the late 1980s a question was posed, can a set of weak learners create a single strong learner?[166] When applied to this specific problem this means: can multiple BDTs be combined to be give better classification performance that any individual BDT can provide? The answer is yes and the process of combining weak learners is know as *boosting*.

[3] `MC15TuneV1_ProbNNx` > 0.2 where `x` is the species of each final state particle

[4] Technically the `HLT1` trigger lines also use multivariate selection criteria.



(a) Rectangular selection

(b) Tree-based selection

Figure 5.8: Comparison between using a rectangular and tree-based selection to separate between two classes of data using two features. In this case the simple tree-based selection results in a 7 % higher significance.

While it is relatively simple to combine the weighted output multiple classifiers to create a pseudo-continuous variable, the process of choosing the best set of classifiers is a complex statistical problem and an active field of research.[167–173] It is a type of *supervised machine learning* where a sample of *training data* is given to the algorithm with associated labels stating the true classification for the given data point. The algorithm then attempts to find the an ensemble of learners that provides an output variable with good separation between the classes.

For this analysis it was decided to use the `XGBoost`[170] algorithm for training the BDT. For the training dataset a mixture of simulated D^+ and D_s^+ signal data was used, with the weights from Section 5.5.1 applied to improve the agreement with data. For the training data representing background, a sample of unphysical candidates[5] containing all three final state particle of the same charge was used. This data can act as a proxy for combinatorial background while ensuring no signal will be present. Several strategies were tried for training the BDT, most notably:

[5] The selection of this sample was discussed in Section 5.3.1.

Channel	$N_{\text{background}}$	N_{D^+}	$N_{D_s^+}$	$N_{\text{background}}$	N_{D^+}	$N_{D_s^+}$
$D^+_{(s)} \to K^+e^+e^-$	382124	1664	1455	381821	1543	1409
$D^+_{(s)} \to K^-e^+e^+$	382124	1849	1355	381821	1919	1327
$D^+_{(s)} \to K^+e^+\mu^-$	352878	4803	3164	352582	4777	3173
$D^+_{(s)} \to K^+\mu^+e^-$	352321	4923	2995	353139	5005	2999
$D^+_{(s)} \to K^-\mu^+e^+$	352321	4944	3120	353139	4964	3161
$D^+_{(s)} \to K^+\mu^+\mu^-$	152495	7675	5661	152640	7691	5532
$D^+_{(s)} \to K^-\mu^+\mu^+$	152495	8039	5722	152640	8024	5750
$D^+_{(s)} \to \pi^+e^+e^-$	721086	2180	1700	724199	2122	1738
$D^+_{(s)} \to \pi^-e^+e^+$	721086	2106	1581	724199	2115	1613
$D^+_{(s)} \to \pi^+e^+\mu^-$	795960	5375	3649	797963	5347	3654
$D^+_{(s)} \to \pi^+\mu^+e^-$	795972	5634	3629	797951	5719	3646
$D^+_{(s)} \to \pi^-\mu^+e^+$	795972	5746	3675	797951	5737	3655
$D^+_{(s)} \to \pi^+\mu^+\mu^-$	329643	9341	6583	329143	9391	6650
$D^+_{(s)} \to \pi^-\mu^+\mu^+$	329643	9774	6918	329143	10055	6918

Table 5.4: Number of events in each sample that is used when training the BDT that is used for 2016 data. The two columns are for the two different k-folds.

- **Single BDT for all channels:** Initially it was hoped that generating full simulation for all channels could be avoided by training a single BDT using a representative mixture of the final states. This resulted in notably worse performance for channels that were not used for training.
- **One BDT per channel, 3 output classes (background, D^+ and D_s^+):** It was studied whether the difference in lifetime between the D^+ and D_s^+ would provide discrimination that helps improve the limit for channels with electrons where the D^+ and D_s^+ overlap. This extra information turned out not to be very useful for the limit.
- **One BDT per channel, 2 output classes (background, $D^+_{(s)}$):** When neglecting imperfections in the training algorithm, this should be equivalent to combining the D^+ and D_s^+ output variables from the aforementioned 3 class BDT and this was found to be the case.

From these studies it was decided to train a separate classifier for each final state and year combination but based on common variables. As the analysis is studying 28 channels it is important to keep the procedure as similar as possible in all decays so that the studies can be highly automated. In each case three samples are included in the training: D^+ signal simulation, D_s^+ signal simulation and background data from the unphysical combination of tracks that all have the same sign. The two signal samples are mixed and given equal weighting during the training. Over-training is avoided using k-folding with $k = 2$, i.e. two classifiers are trained, each using half of the available dataset. Each classifier is then used to get a response of the opposite half of the data, ensuring the BDT response is never taken from a classifier training with that data point. Table 5.4 shows the number of events of each type that are used when training the BDTs for 2016 data. The following variables are included in the training:

- Primary vertex fit quality
- $D^+_{(s)}$ decay vertex fit quality per degree of freedom
- Pseudorapidity of the $D^+_{(s)}$ meson
- Flight distance of the $D^+_{(s)}$
- Impact parameter χ^2 of the final state particles
- Magnitude of each of the final state particle's three-momenta
- Impact parameter between the reconstructed $D^+_{(s)}$ and the primary vertex
- p_T asymmetry in a 2 radian cone around the $D^+_{(s)}$ candidate, see Reference [174] for details
- Maximum distance of closest approach between all final state particles
- Angle between the reconstructed $D^+_{(s)}$ momentum and the line between the particle's origin and decay vertices
- Reconstructed proper lifetime of the $D^+_{(s)}$

As with most machine learning techniques, `XGBoost` provides hyperparameters that can be adjusted to improve the classifier performance. For this analysis these were chosen empirically to avoid overtraining as further tuning was found to give no appreciable improvement in performance. The values used are shown in Table 5.5.

Parameter	Value
`max_depth`	4
`learning_rate`	0.1
`min_child_weight`	10
`colsample_bytree`	0.5
`gamma`	0.1
`scale_pos_weight`	$\frac{N_{\text{bkg}}}{N_{\text{Signal MC}}}$
`n_estimators`	150

Table 5.5: Parameters used when training the multivariate classifier. See Reference [175] for the definition of each parameter.

For evaluating the performance of the classifier three main plots were used:

- **Training progress** The logloss[176] performance of each classifier/sample after each tree is added to the classifier. This is the metric used for training the classifier and is shown for both the training and testing datasets. Any bias between these curves is evidence of over-training however this would not impact the final result due to the use of k-folding. It is also shown for each k-fold and differences between these curves indicates more training data is required or a larger value of k.

Channel	Sensitivity target	
	2015	2016
$D^+ \to \pi^+\mu^+\mu^-$	1×10^{-7}	4×10^{-8}
$D^+ \to \pi^-\mu^+\mu^+$	1×10^{-7}	4×10^{-8}
$D^+ \to K^+\mu^+\mu^-$	1×10^{-7}	4×10^{-8}
$D^+ \to K^-\mu^+\mu^+$	1×10^{-7}	4×10^{-8}
$D_s^+ \to K^+e^+\mu^-$	6×10^{-7}	3×10^{-7}
$D_s^+ \to K^+\mu^+e^-$	6×10^{-7}	3×10^{-7}
$D_s^+ \to K^-\mu^+e^+$	6×10^{-7}	3×10^{-7}
$D_s^+ \to \pi^+e^+\mu^-$	6×10^{-7}	3×10^{-7}
$D_s^+ \to \pi^+\mu^+e^-$	6×10^{-7}	3×10^{-7}
$D_s^+ \to \pi^-\mu^+e^+$	6×10^{-7}	3×10^{-7}
$D_s^+ \to \pi^+e^+e^-$	2×10^{-6}	1×10^{-6}
$D_s^+ \to \pi^-e^+e^+$	2×10^{-6}	1×10^{-6}
$D_s^+ \to K^+e^+e^-$	2×10^{-6}	1×10^{-6}
$D_s^+ \to K^-e^+e^+$	2×10^{-6}	1×10^{-6}

Table 5.6: Target sensitivities in each decay channel used for selection optimisation. 2016 sensitivities relative to 2015 are scaled by the root of the approximate ratio of collected luminosity, $\sqrt{5}$.

- **Classifier response** Response of the classifier to sideband data and signal simulation. This should generally be smoothly varying and distinct features are likely to be an artefact of discretised input variables or unphysical differences between the difference classes of the training data.
- **ROC curve** Receiver Operating Characteristic (ROC) curve, shows the true positive rate against the false positive rate for the blinded real data (background) and signal simulation. This curve is independent of the ratio between the signal and background and better performing classifiers tend to maximise the area under this curve.

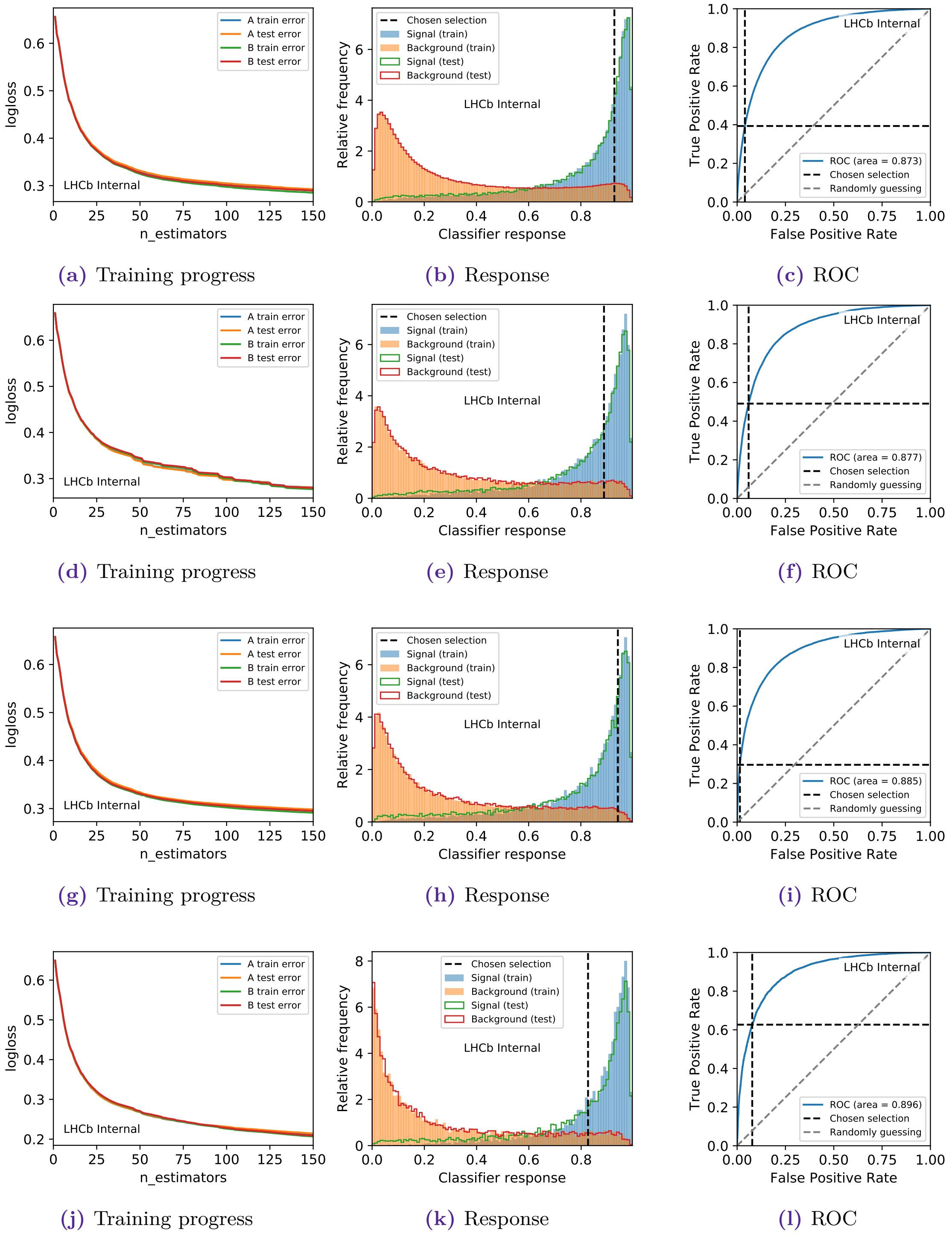
A representative sample of these performance plots are show in Figure 5.9.

5.3.5 Selection optimisation

The post-stripping selection is comprised of three PID cuts, one for each final state particle, and a cut on the output variable of the aforementioned classifier. This makes the optimisation of the selection a four dimensional problem. In the case of both leptons having the same flavour, the problem was reduced to being three dimensional by only cutting on the minimum of the two leptons PID. This is equivalent to cutting on both leptons' PID identically.

Other LHCb analyses have found that applying multiple cuts to divide the sample into more and less pure samples and then combining limits has improved analysis performance [177, 178]. This approach was studied here but not found to result in better performance.

To optimise the working point in the three/four variables, it is necessary to define a Figure of Merit (FoM) that can be maximised. Various options exist for this and one



(a) Training progress **(b)** Response **(c)** ROC

(d) Training progress **(e)** Response **(f)** ROC

(g) Training progress **(h)** Response **(i)** ROC

(j) Training progress **(k)** Response **(l)** ROC

Figure 5.9: Classifier performance plots for 2016 data. From top to bottom the classifiers are those used for the $D^+_{(s)} \to \pi^+\mu^+\mu^-$, $D^+_{(s)} \to \pi^-\mu^+e^+$, $D^+_{(s)} \to K^+e^+\mu^-$ and $D^+_{(s)} \to K^-e^+e^+$ datasets.

of the most popular is the Punzi figure of merit[179]. For a search with an appreciable background component, as is the case for this analysis, this can be simplified to $\frac{S}{\sqrt{S+B}}$. For the optimisation, this FoM is measured at 10 different working points in each variable. This gives 1000 bins in the 3D case and 10 000 bins in the 4D case. The bin boundaries are chosen for each bin to contain the same number of signal simulation events, making them uniformly spaced in the efficiency of the final selection.

Channel	BDT cut	PID(h)	PID(l1)	PID(l2)
$D^+_{(s)} \to K^+e^+e^-$	0.887	0.911	0.200	0.200
$D^+_{(s)} \to K^-e^+e^+$	0.880	0.907	0.200	0.200
$D^+_{(s)} \to K^+e^+\mu^-$	0.934	0.895	0.201	0.896
$D^+_{(s)} \to K^+\mu^+e^-$	0.943	0.770	0.912	0.201
$D^+_{(s)} \to K^-\mu^+e^+$	0.943	0.897	0.907	0.522
$D^+_{(s)} \to K^+\mu^+\mu^-$	0.938	0.200	0.925	0.925
$D^+_{(s)} \to K^-\mu^+\mu^+$	0.937	0.201	0.950	0.950
$D^+_{(s)} \to \pi^+e^+e^-$	0.867	0.752	0.369	0.369
$D^+_{(s)} \to \pi^-e^+e^+$	0.867	0.779	0.359	0.359
$D^+_{(s)} \to \pi^+e^+\mu^-$	0.920	0.802	0.529	0.899
$D^+_{(s)} \to \pi^+\mu^+e^-$	0.908	0.780	0.907	0.530
$D^+_{(s)} \to \pi^-\mu^+e^+$	0.908	0.801	0.902	0.503
$D^+_{(s)} \to \pi^+\mu^+\mu^-$	0.960	0.770	0.926	0.926
$D^+_{(s)} \to \pi^-\mu^+\mu^+$	0.959	0.204	0.930	0.930

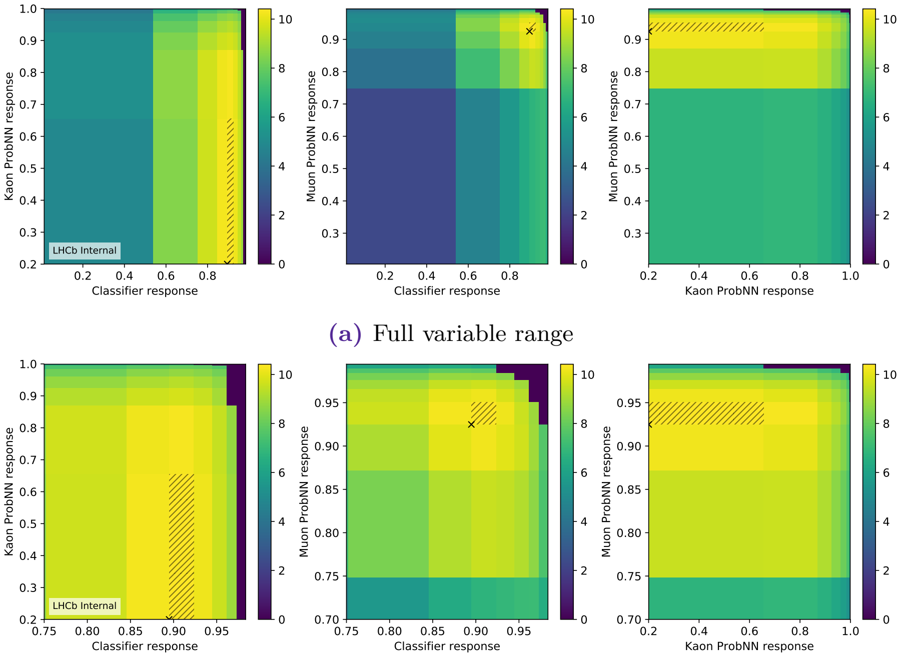
(a) Cut values used for 2015 data.

Channel	BDT cut	PID(h)	PID(l1)	PID(l2)
$D^+_{(s)} \to K^+e^+e^-$	0.826	0.763	0.203	0.203
$D^+_{(s)} \to K^-e^+e^+$	0.812	0.782	0.201	0.201
$D^+_{(s)} \to K^+e^+\mu^-$	0.942	0.752	0.203	0.908
$D^+_{(s)} \to K^+\mu^+e^-$	0.902	0.712	0.903	0.201
$D^+_{(s)} \to K^-\mu^+e^+$	0.893	0.758	0.919	0.204
$D^+_{(s)} \to K^+\mu^+\mu^-$	0.891	0.201	0.930	0.930
$D^+_{(s)} \to K^-\mu^+\mu^+$	0.895	0.200	0.925	0.925
$D^+_{(s)} \to \pi^+e^+e^-$	0.864	0.772	0.202	0.202
$D^+_{(s)} \to \pi^-e^+e^+$	0.854	0.769	0.201	0.201
$D^+_{(s)} \to \pi^+e^+\mu^-$	0.927	0.817	0.201	0.917
$D^+_{(s)} \to \pi^+\mu^+e^-$	0.890	0.795	0.918	0.201
$D^+_{(s)} \to \pi^-\mu^+e^+$	0.888	0.784	0.916	0.200
$D^+_{(s)} \to \pi^+\mu^+\mu^-$	0.928	0.201	0.931	0.931
$D^+_{(s)} \to \pi^-\mu^+\mu^+$	0.949	0.206	0.931	0.931

(b) Cut values used for 2016 data.

Table 5.7: Results of the selection optimisation procedure.

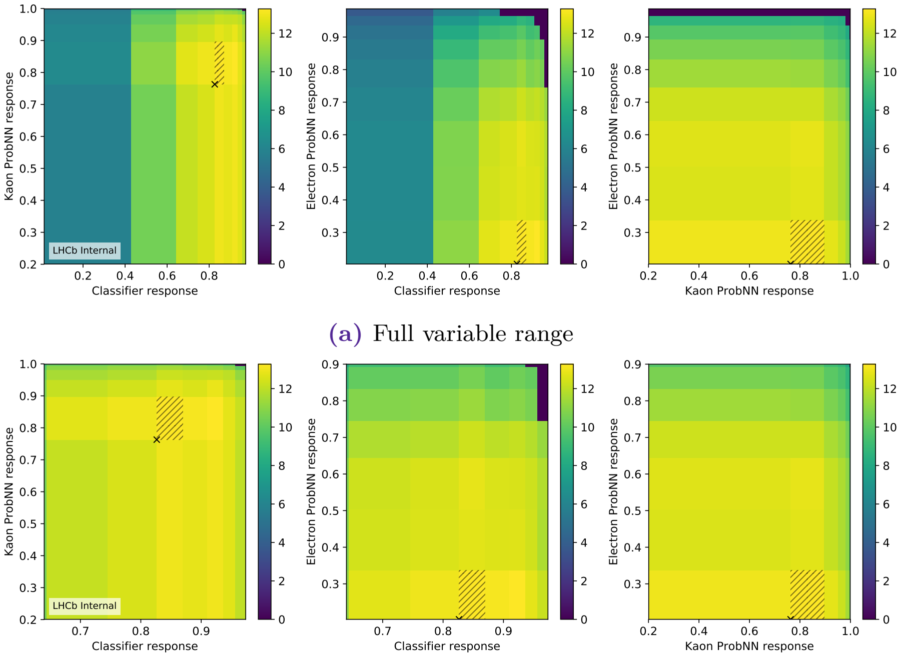
The signal yield S is obtained using the normalisation channel with the fitting procedure described in Section 5.4.4. The number of selected normalisation sample events is corrected by the efficiency of the selection obtained from simulation. The known branching ratio is then used to give the estimated number of $D^+_{(s)}$ mesons produced. This is then



(a) Full variable range

(b) Near optimum

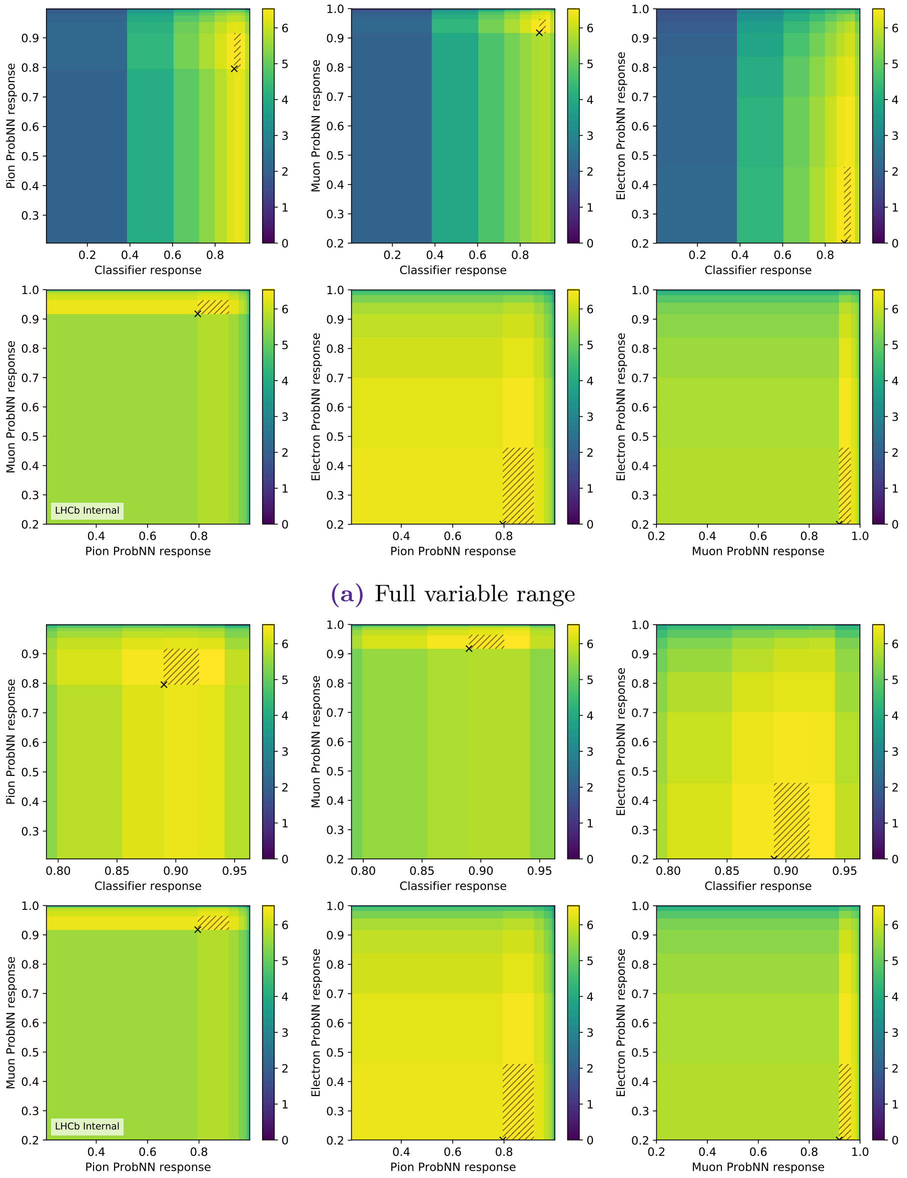
Figure 5.10: $\frac{S}{\sqrt{S+B}}$ vs selection cuts for $D^+ \to K^-\mu^+\mu^+$ in 2016 using $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+$. Each graphic shows each 2D slice (of the 3D grid) at the optimum of the other selection variables. The maximal bin is shown by the shaded region and the small x corresponds to the chosen selection.



(a) Full variable range

(b) Near optimum

Figure 5.11: $\frac{S}{\sqrt{S+B}}$ vs selection cuts for $D^+ \to K^+e^+e^-$ in 2016 using $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+$. Each graphic shows each 2D slice (of the 3D grid) at the optimum of the other selection variables. The maximal bin is shown by the shaded region and the small x corresponds to the chosen selection.



(a) Full variable range

(b) Near optimum

Figure 5.12: $\frac{S}{\sqrt{S+B}}$ vs selection cuts for $D^+ \to \pi^+\mu^+e^-$ in 2016 using $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\ \pi^+$. Each graphic shows each 2D slice (of the 4D grid) at the optimum of the other selection variables. The maximal bin is shown by the shaded region and the small `x` corresponds to the chosen selection.

scaled by the signal efficiency in simulation and approximate branching fractions to which this analysis is expected to be sensitive. The estimated sensitivities are given in Table 5.6 and were estimated based on the previous LHCb analysis[155] of $D^+_{(s)} \to \pi^\pm\mu^+\mu^\mp$, scaled by an efficiency estimate. This estimate is sufficient for the optimisation as the final results of this analysis were later found to not have a strong dependence on the values chosen. The background yield B is estimated from the all-same-sign background samples.

When tight selections cuts are applied this metric becomes sensitive to statistical fluctuations. To avoid optimising for these, the metric is set to zero when there are fewer than 100 events in the simulation sample or there are fewer than 30 events in the all-same-sign sample. To check the validity of these selections 2D slices are plotted for each variable, at the optimal value of the other variables. Some example plots are shown in Figures 5.10–5.12. The final working points used for each year are shown in Table 5.7. While it appears very different cuts are chosen for similar channels, the difference between a PID cut of 0.2 and 0.8 often has very little effect on the signal efficiency and background rejection of the cut. The difference between these cuts only represents one bin in the optimisation procedure due to the `ProbNN` variables being sharply peaked at 1, as is shown in Figure 5.7.

5.4 Yield extraction

Ideally when performing an analysis the only data that remains after applying the selection corresponds to the signal being studied. For a search this would mean the presence of any data would be definitive evidence of the given process. In practice it is rare to make a measurement with truly no background contributions, therefore it is necessary to separate each component of the data. In high energy physics this is commonly achieved by modelling the data as a sum of Probability Density Functions (PDFs).[180] Each component in the fit is denoted by

$$f_i(\vec{x}|\vec{\theta_i}), \tag{5.1}$$

where i is the number of the component, $\vec{x}$ is a vector of observables and $\vec{\theta_i}$ is a vector of parameters that affect the shape of the distribution. A set of m PDFs can be combined using $m-1$ fractions, α_i, that are each bounded by the interval $[0,1]$ and combined to ensure the combined PDF, f, remains normalised. For example in the two-component case $f(\vec{x}|\vec{\theta})$ is given by

$$f(\vec{x}|\vec{\theta_0}\vec{\theta_1}\alpha_0\alpha_1) = \alpha_0 \cdot f_0(\vec{x}|\vec{\theta_0}) + (1-\alpha_0) \cdot f_1(\vec{x}|\vec{\theta_1}) \tag{5.2}$$

and in the three-component case $f(\vec{x}|\vec{\theta})$ is given by

$$f(\vec{x}|\vec{\theta_0}\vec{\theta_1}\vec{\theta_2}\alpha_0\alpha_1\alpha_2) = \alpha_0 \cdot f_0(\vec{x}|\vec{\theta_0}) + (1-\alpha_0)\alpha_1 \cdot f_1(\vec{x}|\vec{\theta_1}) + (1-\alpha_0)(1-\alpha_1) \cdot f_2(\vec{x}|\vec{\theta_2}). \tag{5.3}$$

In order for this statistical model to be useful it is necessary to be able to extract the best values of the parameters, $\vec{\theta}$, given a dataset consisting of n values of x. A frequentist method for achieving this is to compute the product of the probabilities of observing each data point, x_i, to obtain a likelihood,

$$\mathcal{L} = \mathcal{L}(\vec{\theta}|\vec{x}) = \prod_j^n f(x_j|\vec{\theta}). \tag{5.4}$$

The value of $\vec{\theta}$ that maximises the likelihood corresponds to the best fit. This can only be maximised analytically in only the most trivial of cases making it necessary to use general purpose optimisation algorithms such as `BFGS`[117] and `MIGRAD`[116]. For most datasets the likelihood is too small to be represented using the `IEEE 754` standard[181] for floating point numbers making it instead useful to maximise the log-likelihood,

$$\log \mathcal{L} = \sum_{j}^{n} \log f(x_j|\vec{\theta}). \tag{5.5}$$

Furthermore, the convention for optimisation algorithms is to minimise the given function making it instead useful to minimise the negative log-likelihood, $\mathcal{N}$, given by

$$\mathcal{N} = -\sum_{j}^{n} \log f(x_j|\vec{\theta}). \tag{5.6}$$

This procedure for finding the optimal set of parameters is known as a *maximum likelihood fit*.

It is rare for the recursively defined fractions, α_i, to be physical meaningful. Instead it is more common for the model to be comprised of m Poisson processes that each contribute some number of entries to the dataset. This results in the total PDF being given by

$$f(x|\vec{\theta}) = \sum_{i}^{m} N_i \cdot f_i(x|\vec{\theta_i}) \tag{5.7}$$

where N_i is the number of events from the ith component. It is not possible to maximise this function as it is under-constrained. As shown in Reference [180], this can be accounted for in what is known as an *extended maximum likelihood fit* and has the effect of introducing a second term in the log-likelihood,

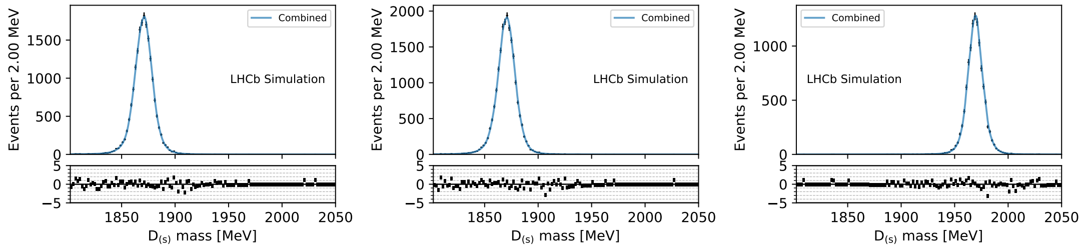
$$\log \mathcal{L} = \sum_{j}^{n} \log f(x_j|\vec{\theta}) - \sum_{i} N_i, \tag{5.8}$$

that accounts for the underlying Poisson process by which each data point is generated.

For this analysis, one dimensional fits to the reconstructed invariant mass of the $D^+_{(s)}$ meson are used. In some cases a *simultaneous* fit is performed which means the likelihoods from multiple datasets are summed and this combined value is maximised. Additionally, many of the PDFs used in this analysis are *Kernel Density Estimations* (KDEs) from simulated data. This is a statistical technique for approximating an unknown distribution from which a finite dataset has been sampled. One method of achieving this is known as `RooKeysPDF` and uses N Gaussian distributions to describe a reference dataset consisting of N data points. The widths (σ) of the Gaussian distributions are adjusted according to the local density of events[182].

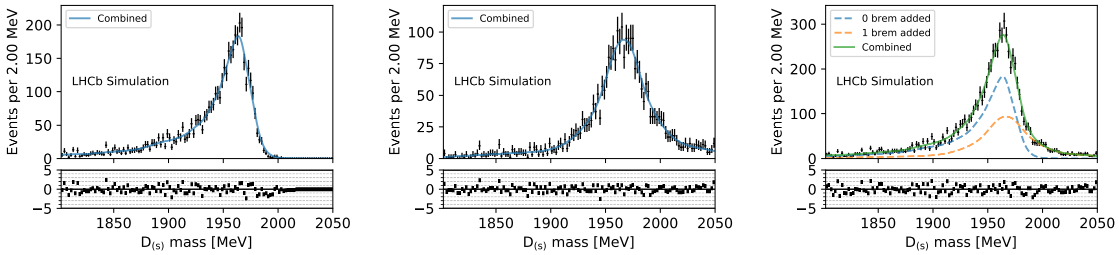
5.4.1 Signal

As most of the decays studied are expected to be forbidden in the standard model it is not possible to use real data to find an parametrisation for the signal mass shape, therefore,



(a) $D^+ \to \pi^+\mu^+\mu^-$ **(b)** $D^+ \to \pi^-\mu^+\mu^+$ **(c)** $D_s^+ \to K^+\mu^+\mu^-$

Figure 5.13: Reconstructed D^+ and D_s^+ mass distributions in simulated events for three example decays channels. The curve shows a fit to the distribution obtained with a kernel density estimation technique[182]. The lower axis shows the *pull* between the data and KDE, in units of statistical significance.



(a) 0 photons **(b)** 1+ photons added **(c)** Combined

Figure 5.14: Signal templates in reconstructed invariant mass obtained with a kernel density estimation technique[182] for 2016 $D_s^+ \to \pi^+e^+\mu^-$ simulation, split by the number of bremsstrahlung photons that have been added to the candidates.

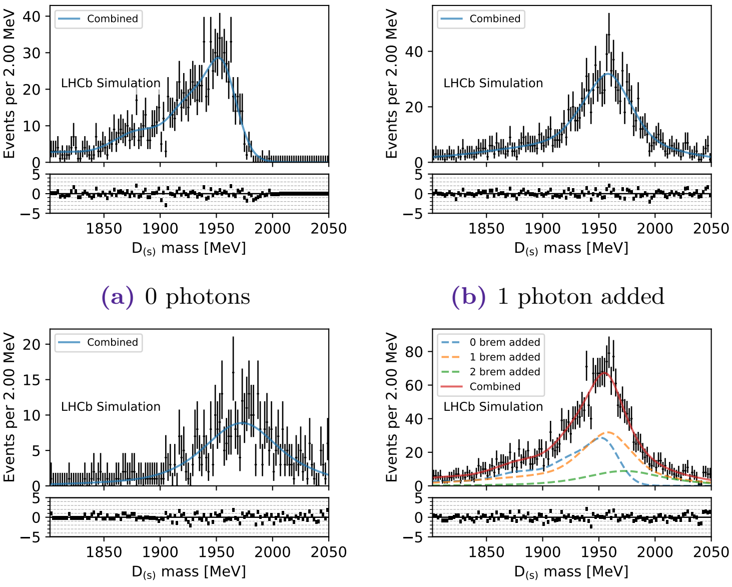
truth matched simulation is used with `RooKeysPDF`[182] (with `rho = 2`)[6] to obtain a kernel density estimation for the shape. These shapes are validated against data as discussed in Section 6.1.7.

The shape of the signal is expected to vary significantly depending on the number of bremsstrahlung photons involved in the decay, with the resolution worsening for more photons. Consequently, the shapes are split into $N + 1$ categories, where N is the number of electrons in the final state. Each category corresponds to the number of bremsstrahlung photons added and the last category includes any number of reconstructed photons. When performing the fit for the final result the relative yields are fixed from simulated data. Examples of the signal fit shapes are given in Figure 5.13, 5.14, and 5.15.

5.4.2 Peaking backgrounds

In addition to combinatorial background, a number of backgrounds are expected from other decays of $D^+_{(s)}$ mesons where particles have been missed or misidentified. Initial studies to find the specific backgrounds expected in each final state were performed using a simplified phase space simulation[183]. This allows the effect of reconstructing real decays under the wrong mass hypothesis to be studied. These results are then scaled by

[6] The bandwidth of the KDE, larger values promote smoothness over preserving detail.



(a) 0 photons

(b) 1 photon added

(c) 2+ photons added

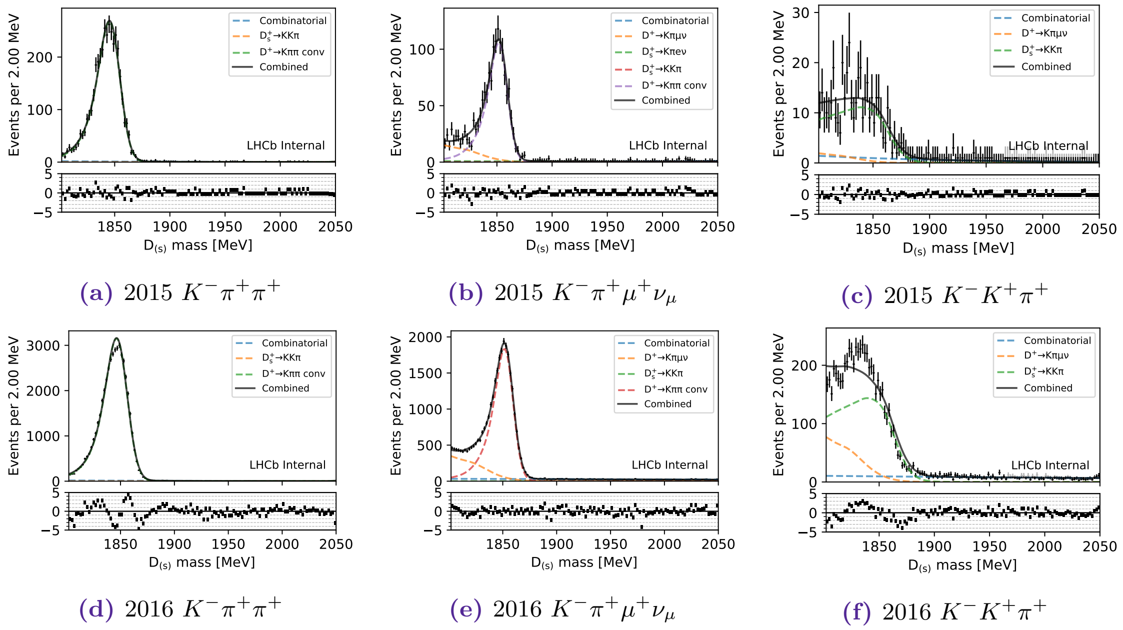
(d) Combined

Figure 5.15: Signal templates in reconstructed invariant mass obtained with a kernel density estimation technique[182] for 2016 $D_s^+ \to K^+e^+e^-$ simulation, split by the number of bremsstrahlung photons that have been added to the candidates.

the measured LHCb charm cross section [184] and the PDG 2016 [150] branching fractions. The studies are made for all possible three charged-body decays of D^+ and D_s^+ meson to pions, kaons, muons and electrons, with a neutrino included where appropriate.

The presence and position of backgrounds in the mass distributions resulting from these studies agree with the observed sidebands of selected $D_{(s)}^+ \to h^\pm\mu^+\mu^\mp$ in data. However, for other final states, the dominant backgrounds are hidden in the blinded regions. To work around this, use is made of *inverse PID* cuts, to select samples over the full mass window without unblinding. Each misidentified lepton is required to have a high probability of being misidentified while also passing a PID requirement under the true hypothesis for the background channel. Some additional cuts are applied to suppress other backgrounds, especially pions from $D^+ \to K^-\pi^+\pi^+$ decays. This allowed useful samples of the specific backgrounds to be studied. From these studies it was clear that the signal final states could best be grouped by hadron flavour and lepton charge to express commonality between the expected backgrounds.

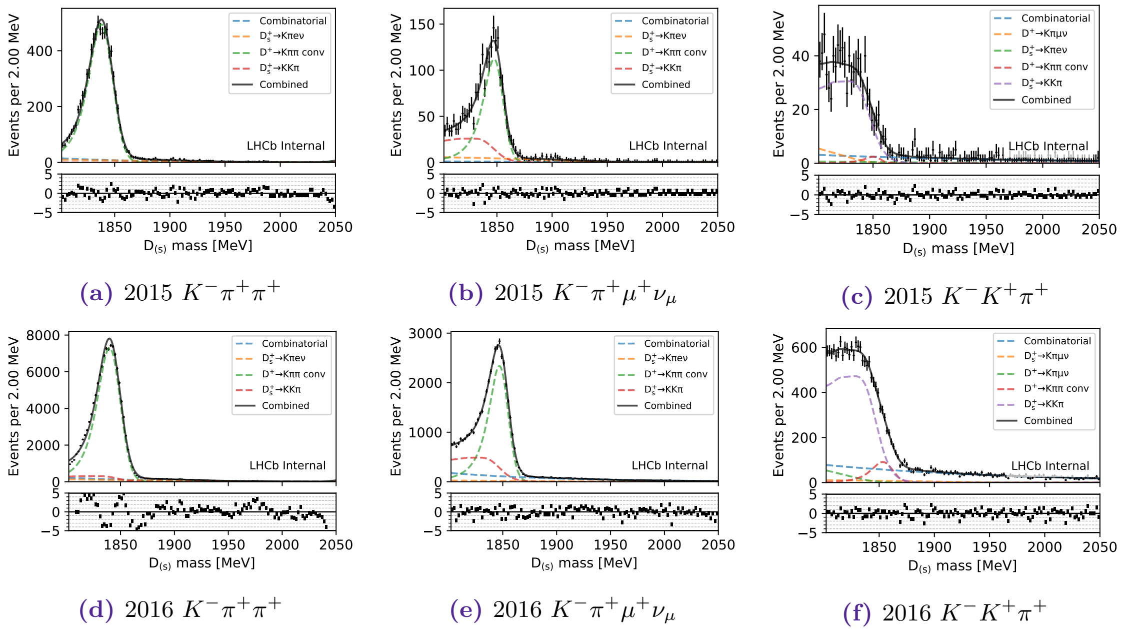
A description of the mass distributions shapes for these specific backgrounds is required for the measurement. An attempt to fit the inverse-cut samples using analytic PDFs was made but the shapes were found to be difficult to describe without introducing fit instabilities. Additionally, the shapes developed using the 2015 dataset did not fit well to the larger 2016 sample. Instead the background PDFs are obtained using RapidSim[185] which generates simulated samples using the phase space technique described in Reference [183]. The kinematics of the initial hadrons are corrected for LHC conditions using FONLL[186, 187] predictions and smearing is applied to emulate the LHCb detector resolution. These samples were then fitted using `RooKeysPDF` to obtain probability density functions.



(a) 2015 $K^-\pi^+\pi^+$ **(b)** 2015 $K^-\pi^+\mu^+\nu_\mu$ **(c)** 2015 $K^-K^+\pi^+$

(d) 2016 $K^-\pi^+\pi^+$ **(e)** 2016 $K^-\pi^+\mu^+\nu_\mu$ **(f)** 2016 $K^-K^+\pi^+$

Figure 5.16: Reconstructed mass distributions for backgrounds to $D^+ \to K^-\mu^+\mu^+$ enhanced using PID cuts for fitted using `RooKeysPDF`. The model used for (c) provides a poor description however this has negligible impact on the analysis results as this background source is found not to contribute at an appreciable level.



(a) 2015 $K^-\pi^+\pi^+$ **(b)** 2015 $K^-\pi^+\mu^+\nu_\mu$ **(c)** 2015 $K^-K^+\pi^+$

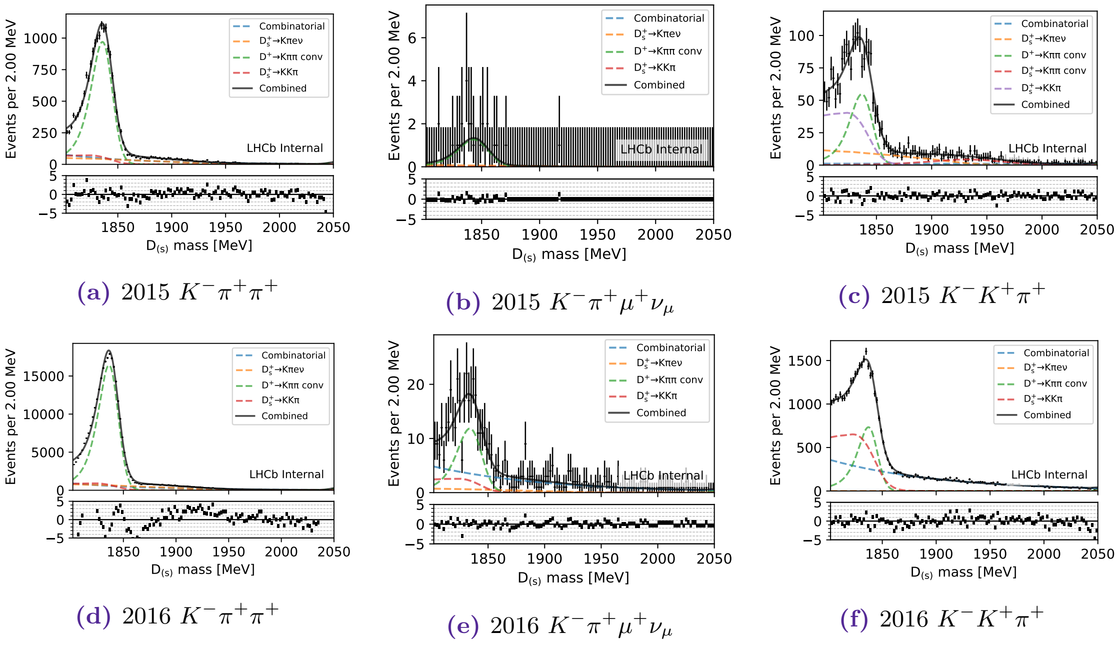
(d) 2016 $K^-\pi^+\pi^+$ **(e)** 2016 $K^-\pi^+\mu^+\nu_\mu$ **(f)** 2016 $K^-K^+\pi^+$

Figure 5.17: Reconstructed mass distributions for backgrounds to $D^+ \to K^-\mu^+e^+$ enhanced using PID cuts for fitted using `RooKeysPDF`.

To validate these shapes, the inverse PID samples were then fitted using a two step procedure:

- All backgrounds with candidates in the analysis mass window were given equal starting yields.[7] An extended maximum likelihood fit was then performed.

[7] $\frac{N_{\text{events}}}{N_{\text{components}}}$, where $N_{\text{components}}$ is the number of components included in the fit.



(a) 2015 $K^-\pi^+\pi^+$

(b) 2015 $K^-\pi^+\mu^+\nu_\mu$

(c) 2015 $K^-K^+\pi^+$

(d) 2016 $K^-\pi^+\pi^+$

(e) 2016 $K^-\pi^+\mu^+\nu_\mu$

(f) 2016 $K^-K^+\pi^+$

Figure 5.18: Reconstructed mass distributions for backgrounds to $D^+ \to K^-e^+e^+$ enhanced using PID cuts for fitted using RooKeysPDF.

- Components with fitted yields of less than 1 % of the dataset size were removed from the fit. The yields were reset to again be all equal given the new number of background components. A second extended maximum likelihood fit was then performed.

The results of this procedure can be found in Figures 5.16-5.29.

This procedure was found to describe the reconstructed mass distribution of many of the inverse PID data samples with high precision using the fixed shapes that were extracted from the simplified simulation.

The fit quality for backgrounds arising from the misidentification of a non-muon as a muon particle candidate are found to be worse than others, as seen in Figure 5.16 (vs Figure 5.18) It is hypothesised that this may be an artefact of the contradictory selection applied, since the candidates are required to be muons according to the muon chambers (ISMUON) but also required to have a low probability to be a muon from the inverse PID selection.

Furthermore, it is expected that this method will not perfectly describe the true background PDF in data as the mass resolution for peaking backgrounds in the simplified simulation can be expected to be better than that seen in real data. For almost all cases the contribution of these backgrounds is either very small, or potentially not present at all, so this has negligible effect on the analysis. The main exception to this is $D^+ \to K^-\pi^+\pi^+$ that has a significant yield. Consequently a third step is added for some channels with the RapidSim PDF smeared by a Gaussian convolution. This method is able to describe all of the observed discrepancies to within an acceptable degree of precision, as is illustrated in Figures 5.16-5.29, and a systematic uncertainty is assigned for the choice of model in Section 6.1.10.

Channel	Backgrounds		
$D^+_{(s)} \to K^- \mu^+ \mu^+$	$D^+ \to K^- \pi^+ \mu^+ \nu$	$D^+ \to K^- \pi^+ \pi^+$	$D^+_s \to K^- K^+ \pi^+$
$D^+_{(s)} \to K^- \mu^+ e^+$	$D^+ \to K^- \pi^+ \mu^+ \nu$	$D^+ \to K^- \pi^+ \pi^+$	$D^+_s \to K^- K^+ \pi^+$
$D^+_{(s)} \to K^- e^+ e^+$	$D^+ \to K^- \pi^+ \mu^+ \nu$	$D^+ \to K^- \pi^+ \pi^+$	$D^+_s \to K^- K^+ \pi^+$
$D^+_{(s)} \to \pi^- \mu^+ \mu^+$	$D^+ \to \pi^+ \pi^+ \pi^-$	$D^+_s \to \pi^+ \pi^+ \pi^-$	
$D^+_{(s)} \to \pi^- \mu^+ e^+$	$D^+ \to \pi^+ \pi^+ \pi^-$	$D^+_s \to \pi^+ \pi^+ \pi^-$	
$D^+_{(s)} \to \pi^- e^+ e^+$	$D^+ \to \pi^+ \pi^+ \pi^-$	$D^+_s \to \pi^+ \pi^+ \pi^-$	
$D^+_{(s)} \to \pi^+ \mu^+ \mu^-$	$D^+ \to \pi^+ \pi^+ \pi^-$	$D^+_s \to \pi^+ \pi^+ \pi^-$	
$D^+_{(s)} \to \pi^+ \mu^+ e^-$	$D^+ \to \pi^+ \pi^+ \pi^-$	$D^+_s \to \pi^+ \pi^+ \pi^-$	
$D^+_{(s)} \to \pi^+ e^+ \mu^-$	$D^+ \to \pi^+ \pi^+ \pi^-$	$D^+_s \to \pi^+ \pi^+ \pi^-$	
$D^+_{(s)} \to \pi^+ e^+ e^-$	$D^+ \to \pi^+ \pi^+ \pi^-$	$D^+_s \to \pi^+ \pi^+ \pi^-$	
$D^+_{(s)} \to K^+ \mu^+ \mu^-$	$D^+ \to K^+ \pi^+ \pi^-$	$D^+_s \to K^+ \pi^+ \pi^-$	
$D^+_{(s)} \to K^+ \mu^+ e^-$	$D^+ \to K^+ \pi^+ \pi^-$	$D^+_s \to K^+ \pi^+ \pi^-$	
$D^+_{(s)} \to K^+ e^+ \mu^-$	$D^+ \to K^+ \pi^+ \pi^-$	$D^+_s \to K^+ \pi^+ \pi^-$	
$D^+_{(s)} \to K^+ e^+ e^-$	$D^+ \to K^+ \pi^+ \pi^-$	$D^+_s \to K^+ \pi^+ \pi^-$	

Table 5.8: Specific background components that are included in the analyses when fitting the signal samples in the limit setting procedure. In addition to these a combinatorial background component is included for all samples, as described in Section 5.4.3.

In summary, shapes are obtained for the specific backgrounds from fits to simplified simulation (RapidSim) samples. These are validated on anti-cut samples in data and additional Gaussian smearing is applied in some cases. Table 5.8 shows the backgrounds that are included in the fits to the signal samples when computing the final result. The aforementioned convolutions are applied to the $D^+ \to K^- \pi^+ \pi^+$ and $D^+_{(s)} \to \pi^+ \pi^+ \pi^-$ contributions.[8]

5.4.3 Other backgrounds

In addition to the peaking backgrounds, a smooth background component is expected across the invariant mass distribution. This contribution arises from a mixture of processes, such as random combinations of tracks or incorrectly reconstructed decays. For the studies in the previous section this was described using an exponential distribution and, while this is sufficient for validating the peaking background shapes, a more advanced description is required for the signal datasets. Figure 5.30 shows that there is a clear difference in this component when electrons are present in the final state therefore, an exponential distribution is used for the four dimuon channels and a cubic is used remaining ten final states containing electrons. This choice is also motivated by shape in the background enhanced samples, as can be see in Figures 5.23, 5.24 and 5.25. In all cases, the change in fit quality between the exponential and cubic models is negligible for dimuon final states.

Channel	Year			Fitted yields		
		N_{D^+}	$N_{D_s^+}$	$N_{D^+\to\pi^+\pi^+\pi^-}$	$N_{D_s^+\to\pi^+\pi^+\pi^-}$	N_{comb}
$D^+\to\pi^+\,(\phi\to e^+e^-)$	2015	290±52	518±50	17±15	79±23	289±77
$D^+\to\pi^+\,(\phi\to\mu^+\mu^-)$	2015	1817±109	3994±99	76±68	143±78	3241±140
$D^+\to\pi^+\,(\phi\to e^+e^-)$	2016	2156±175	5323±178	49±55	235±72	5789±273
$D^+\to\pi^+\,(\phi\to\mu^+\mu^-)$	2016	18105±341	42042±378	460±250	1542±287	49008±486

Table 5.9: Summary of yields for each fit performed to normalisation channel data.

5.4.4 Normalisation channels

All results obtained for this analysis are obtained by normalising the yield of the signal channels to $D^+_{(s)}\to(\phi\!\to\mu^-\mu^+)\,\pi^+$. In order to extract the most precise signal yield for the dimuon normalisation channel, a triple Gaussian is fitted to both the D^+ and D_s^+ signal peak. In each peak, all three normal distributions are required to have the same mean value, with varying widths and relative normalisations that are each fitted. The only expected significant physics background is from $D^+_{(s)}\to\pi^+\pi^+\pi^-$ and the shape for this is obtained using the method described in Section 5.4.2. Additionally, the combinatorial background is described using an exponential distribution. The results of these fits are given in Figure 5.33.

In the case of the dielectron channel, $D^+_{(s)}\to(\phi\!\to e^-e^+)\,\pi^+$ used in Sections 5.5.1, 5.5.3 and 6.1, the signal peak can be clearly seen but cannot be trivially separated from the combinatorial background due to the presence of tails. The upper tail is caused by photons being incorrectly associated to an electron. Conversely, the lower tail is caused by failure to reconstruct all bremsstrahlung radiation emitted by the electron.

To model these shapes the kernel density estimations from simulation are used to describe the signal shape as described in Section 5.4.1. The shapes have no free parameters so only the normalisations are fitted. A simultaneous fit is performed in three categories; no bremsstrahlung added, one photon added and two or more photons added. The number of photons is defined to be the sum of the number of photons added to each electron in the final state. In each category an exponential distribution is used to describe the background, with the slope of the exponential being independent between each category. The results of these fits are shown in Figure 5.31 and 5.32 and integrated projections are shown in Figure 5.34. The fit yields are summarised in Table 5.9.

[8] Due to the smaller dataset, the convolutions are not used for the $D^+_{(s)}\to\pi^+\pi^+\pi^-$ contributions in the cross-check with 2015 data described in Section 6.3.

5.5 Efficiency corrections

For this analysis the signal channel efficiency needs to be known relative to the normalisation channel, given as

$$\epsilon_{\text{ratio}} = \frac{\epsilon\left(D^+_{(s)} \to h^\pm l^+ l'^\mp\right)}{\epsilon\left(D^+_{(s)} \to (\phi \to \mu^+\mu^-)\,\pi^+\right)}.$$

The selection of the normalisation has been made with a loose selection as described in Section 5.3.2 to simplify extrapolations to different particle species and to allow for a data driven cross-check against the normalisation channel (Section 6.3). The same parent particle (D^+ or D^+_s) as the signal channel is always used for setting the branching fraction.

The resonant structure in the signal channels will depend upon the new physics model that is unknown. Consequently, we choose to produce results by assuming a uniform distribution of events in their Dalitz plane phase-space.

5.5.1 Correcting data/MC agreement

It is known that the agreement between real and simulated data is imperfect. Consequently reweighting is applied using the four control channels as a reference. The results of the analysis directly depends on the relative efficiencies of the signal channel to $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+$ being the same in data as in simulation. For channels containing electrons, an efficiency correction is computed as described in Section 5.5.3.

The reweighting of the simulation to describe the control channels does not use traditional binned distributions but instead is achieved using a multivariate reweighting technique. The input variables to the reweighting technique are the same as those used in the selection classifier, listed in Section 5.3.4, with the addition of the $D^+_{(s)}$ meson P_T, $D^+_{(s)}$ meson η and the event multiplicity, N_{tracks}.

The multivariate classifier, here a gradient boosted ensemble of decision trees, is trained to distinguish between real and simulated data. The results are used to generate weights that can negate the differences found[188]. The implementation of this algorithm is taken from `hep_ml`[156]. This technique is simple to implement and accounts for correlations between the variables. It also avoids the issues in a traditional 1D approach of selecting the binning or the order of reweighting of distributions. Four reweighting classifiers are trained for this analysis for each data taking period (2015 and 2016). These are one for each of the four control channels $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+$ and $D^+_{(s)} \to (\phi \to e^-e^+)\,\pi^+$.

In order to train the reweighting classifier it is necessary to use a sample of real data as a reference. As the normalisation channel data contains a mixture of signal and background processes, it is necessary to separate these contributions so the distributions of each variable used in the training are known independently. This is known as *unfolding*.

One of the simplest methods of unfolding is *sideband subtraction*, which allows two components to be separated by assigning negative weights to a region of the invariant mass spectrum that contains no signal contribution. When a histogram of another variable is made these negative weights have the effect of subtracting the background contribution, under the assumption that the variable is uncorrelated with the invariant mass and the samples are statistically large. This method can be generalised to use the result of a maximum likelihood fit to allow arbitrarily complex mixtures to be unfolded, provided the assumptions remain true. This is known as the *sPlot*[189] method and is used by this analysis to unfold the D^+, D_s^+ and combinatorial background contributions in the normalisation channel data.

The weights from these classifiers are then applied to the simulated data before any further analysis is performed. This means the classifier training, $\epsilon_{\text{Selection}}$ and $\epsilon_{\text{electron}}$ all use weighted simulation. The classifier used always corresponds to the same parent meson (D^+ or D_s^+). In the case of the dielectron channels the $\phi \to e^- e^+$ classifier is used and for dimuon channels the $\phi \to \mu^- \mu^+$ classifier is used. For the mixed channels electron and muon channels there is of course no possibility to train an appropriate reweighting classifier. For these channels the $\phi \to \mu^- \mu^+$ derived reweighting is used and then in the systematic uncertainty (see Section 6.1.4) the change in the results from applying the $\phi \to e^- e^+$ reweighting of the simulation is considered.

Correcting event multiplicities

While training the reweighting classifiers discussed above, it was found that the total number of tracks in each event in the data is badly described in the simulation. In the data the number of reconstructed primary vertices is restricted to one (due to the trigger bug described in Section 5.3.1). The poor agreement with the simulation when restricting to a single PV is shown in Figure 5.35.

The simulation data with a single PV no longer provides a good coverage of the full variable range in the data. Consequently the reweighting classifier is unable to correct for the data/simulation difference in multiplicity.

Instead, it was decided to use the simulation without applying the single reconstructed PV requirement. This then increases the multiplicity in the simulation, as shown in Figure 5.35, allowing the reweighting of the simulation to data to be performed. This is implemented by not applying the standard HLT2 on the simulated data and instead applying an equivalent selection. The efficiency correction for the `nPVs == 1` cut is then provided by the normalisation channel, and it cancels perfectly in the efficiency ratio used to correct the obtained signal yield.

5.5.2 PID Efficiencies

It is known that the PID variable response is not well described in simulated data and to account for this various tools are available for obtaining the PID efficiency using data-driven methods. The `PIDGen`[190] tool uses `Meerkat`[191] to fit a Kernel Density Estimation (KDE) of a given particle identification variable using real data. This is parametrised as a function of three variables the track p_T and η, along with the total number of reconstructed

Channel	MagDown ϵ_{Gen} (%)	MagUp ϵ_{Gen} (%)	ϵ_{Gen} (%)	$\epsilon_{\text{Reco+Strip}}$ (%)	$\epsilon_{\text{Trigger}}$ (%)	$\epsilon_{\text{Selection}}$ (%)	Total
$D_s^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	19.30±0.04	19.28±0.04	38.64±0.05	2.44±0.02	58.34±0.36	93.10±0.25	(5.12±0.05)x10^{-3}
$D_s^+ \to \pi^-\mu^+\mu^+$	19.78±0.06	19.79±0.06	39.56±0.05	2.46±0.02	58.68±0.32	16.07±0.32	(9.19±0.20)x10^{-4}
$D_s^+ \to \pi^+\mu^+\mu^-$	19.82±0.06	19.82±0.06	39.53±0.05	2.38±0.02	58.70±0.33	12.42±0.29	(6.87±0.17)x10^{-4}
$D_s^+ \to K^-\mu^+\mu^+$	20.67±0.06	20.74±0.06	41.43±0.05	2.06±0.01	58.61±0.35	27.69±0.43	(1.39±0.03)x10^{-3}
$D_s^+ \to K^+\mu^+\mu^-$	20.66±0.06	20.67±0.06	41.41±0.05	2.02±0.01	57.96±0.36	8.70±0.27	(4.22±0.14)x10^{-4}
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	19.38±0.04	19.47±0.04	38.80±0.05	3.48±0.02	54.72±0.35	92.89±0.25	(6.87±0.07)x10^{-3}
$D^+ \to \pi^-\mu^+\mu^+$	19.87±0.07	19.91±0.07	39.69±0.05	3.48±0.02	55.74±0.26	24.65±0.31	(1.90±0.03)x10^{-3}
$D^+ \to \pi^+\mu^+\mu^-$	19.80±0.07	19.83±0.07	39.66±0.05	3.38±0.02	55.37±0.27	14.81±0.26	(1.10±0.02)x10^{-3}
$D^+ \to K^-\mu^+\mu^+$	20.82±0.07	20.88±0.07	41.67±0.05	2.96±0.02	54.74±0.29	35.30±0.39	(2.38±0.03)x10^{-3}
$D^+ \to K^+\mu^+\mu^-$	20.84±0.07	20.98±0.07	41.64±0.05	2.85±0.02	55.22±0.30	7.84±0.22	(5.13±0.15)x10^{-4}
$D_s^+ \to \pi^-\mu^+e^+$	19.54±0.06	19.42±0.06	39.01±0.05	1.91±0.01	41.55±0.37	28.53±0.53	(8.84±0.19)x10^{-4}
$D_s^+ \to \pi^+\mu^+e^-$	19.48±0.06	19.48±0.06	39.00±0.05	1.90±0.01	41.32±0.37	26.86±0.52	(8.22±0.19)x10^{-4}
$D_s^+ \to \pi^+e^+\mu^-$	19.54±0.06	19.52±0.06	38.98±0.05	1.90±0.01	41.26±0.37	18.96±0.46	(5.79±0.16)x10^{-4}
$D_s^+ \to K^-\mu^+e^+$	20.32±0.06	20.29±0.06	40.70±0.05	1.65±0.01	40.46±0.39	27.50±0.57	(7.45±0.18)x10^{-4}
$D_s^+ \to K^+\mu^+e^-$	20.35±0.06	20.35±0.06	40.71±0.05	1.64±0.01	40.74±0.41	26.39±0.57	(7.19±0.18)x10^{-4}
$D_s^+ \to K^+e^+\mu^-$	20.38±0.06	20.44±0.06	40.68±0.05	1.65±0.01	40.51±0.39	17.74±0.49	(4.81±0.15)x10^{-4}
$D^+ \to \pi^-\mu^+e^+$	19.54±0.06	19.46±0.06	39.11±0.05	2.73±0.02	37.82±0.28	36.88±0.46	(1.49±0.02)x10^{-3}
$D^+ \to \pi^+\mu^+e^-$	19.49±0.06	19.52±0.06	39.12±0.05	2.69±0.02	38.44±0.28	37.81±0.46	(1.53±0.02)x10^{-3}
$D^+ \to \pi^+e^+\mu^-$	19.62±0.06	19.46±0.06	39.11±0.05	2.70±0.02	38.44±0.29	27.36±0.44	(1.11±0.02)x10^{-3}
$D^+ \to K^-\mu^+e^+$	20.41±0.07	20.33±0.06	40.87±0.05	2.41±0.01	37.37±0.30	35.07±0.49	(1.29±0.02)x10^{-3}
$D^+ \to K^+\mu^+e^-$	20.38±0.06	20.30±0.06	40.86±0.05	2.39±0.01	37.98±0.30	34.46±0.49	(1.28±0.02)x10^{-3}
$D^+ \to K^+e^+\mu^-$	20.47±0.07	20.41±0.06	40.85±0.05	2.39±0.01	37.52±0.30	24.54±0.45	(8.99±0.19)x10^{-4}
$D_s^+ \to \pi^+ (\phi \to e^+e^-)$	18.75±0.04	18.79±0.04	37.51±0.05	1.67±0.01	23.82±0.38	52.98±0.95	(7.90±0.21)x10^{-4}
$D_s^+ \to \pi^-e^+e^+$	19.20±0.06	19.15±0.07	38.50±0.05	1.53±0.01	24.43±0.38	33.64±0.86	(4.82±0.15)x10^{-4}
$D_s^+ \to \pi^+e^+e^-$	19.28±0.07	19.12±0.07	38.44±0.05	1.50±0.01	25.21±0.37	16.80±0.66	(2.45±0.10)x10^{-4}
$D_s^+ \to K^-e^+e^+$	20.00±0.07	20.04±0.07	40.05±0.05	1.34±0.01	22.73±0.39	39.89±0.98	(4.85±0.15)x10^{-4}
$D_s^+ \to K^+e^+e^-$	20.06±0.07	20.08±0.07	40.04±0.05	1.32±0.01	24.33±0.40	10.42±0.59	(1.34±0.08)x10^{-4}
$D^+ \to \pi^+ (\phi \to e^+e^-)$	18.88±0.06	18.80±0.06	37.64±0.05	2.34±0.02	20.36±0.31	48.80±0.97	(8.75±0.24)x10^{-4}
$D^+ \to \pi^-e^+e^+$	19.28±0.07	19.33±0.07	38.51±0.05	2.14±0.02	21.94±0.30	42.39±0.84	(7.66±0.20)x10^{-4}
$D^+ \to \pi^+e^+e^-$	19.30±0.07	19.19±0.07	38.50±0.05	2.13±0.02	22.57±0.30	19.29±0.67	(3.57±0.14)x10^{-4}
$D^+ \to K^-e^+e^+$	20.13±0.07	20.16±0.07	40.11±0.05	1.99±0.01	20.27±0.29	44.50±0.90	(7.20±0.20)x10^{-4}
$D^+ \to K^+e^+e^-$	20.12±0.07	20.13±0.07	40.16±0.05	1.97±0.02	19.89±0.31	9.58±0.58	(1.50±0.10)x10^{-4}

Table 5.10: Overview of efficiencies obtained from simulation samples that represent the 2016 data sample. The value of ϵ_{Gen} is doubled due to the event flipping described in Section 5.5.3.

tracks (`nTracks`). Once fitted, the KDE can be sampled from to produce a realistic simulation of the true PID variables. The efficiency can then be obtained alongside the rest of the offline selection efficiency. To ensure the validity of the response obtained from `PIDGen`, muon tracks are required to have p_T >800 MeV as the reference sample contains this kinematic requirement.

5.5.3 Efficiency factorisation

Calculation of the efficiency is factorised as follows:

$$\epsilon = \epsilon_{\text{Gen}} \cdot \epsilon_{\text{Reco+Strip}} \cdot \epsilon_{\text{Trigger}} \cdot \epsilon_{\text{Selection}} \cdot \epsilon_{\text{electron}}$$

where each efficiency is relative to the previous step. The values obtained for these efficiencies using 2016 data and simulation are given in Table 5.10. Additionally, Table 5.11 shows the offline selection efficiency used for 2015 data. This is computed using simulated data generated under 2016 conditions that is then reweighted against the 2015 normalisation data. All other efficiencies are taken to be the same as in the 2016 simulation samples.

ϵ_{Gen}

When simulating data with Monte Carlo methods most of the generated data is outside of LHCb's acceptance. Simulating the detector response to these events is extremely time consuming[192] and not useful to most analyses therefore *generator level cuts* are applied to

Channel	$\epsilon_{\text{Selection}}$ (%)	Total
$D_s^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	93.28±0.26	(5.13±0.05)x10^{-3}
$D_s^+ \to \pi^-\mu^+\mu^+$	16.32±0.34	(9.33±0.21)x10^{-4}
$D_s^+ \to \pi^+\mu^+\mu^-$	7.80±0.25	(4.31±0.14)x10^{-4}
$D_s^+ \to K^-\mu^+\mu^+$	19.33±0.39	(9.69±0.22)x10^{-4}
$D_s^+ \to K^+\mu^+\mu^-$	6.62±0.25	(3.21±0.12)x10^{-4}
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	93.08±0.26	(6.88±0.07)x10^{-3}
$D^+ \to \pi^-\mu^+\mu^+$	23.75±0.32	(1.83±0.03)x10^{-3}
$D^+ \to \pi^+\mu^+\mu^-$	10.26±0.24	(7.62±0.19)x10^{-4}
$D^+ \to K^-\mu^+\mu^+$	25.14±0.36	(1.70±0.03)x10^{-3}
$D^+ \to K^+\mu^+\mu^-$	6.46±0.21	(4.23±0.14)x10^{-4}
$D_s^+ \to \pi^-\mu^+e^+$	24.53±0.53	(7.60±0.19)x10^{-4}
$D_s^+ \to \pi^+\mu^+e^-$	24.27±0.53	(7.43±0.19)x10^{-4}
$D_s^+ \to \pi^+e^+\mu^-$	23.87±0.53	(7.29±0.18)x10^{-4}
$D_s^+ \to K^-\mu^+e^+$	15.44±0.48	(4.18±0.14)x10^{-4}
$D_s^+ \to K^+\mu^+e^-$	18.13±0.52	(4.94±0.16)x10^{-4}
$D_s^+ \to K^+e^+\mu^-$	21.69±0.54	(5.88±0.17)x10^{-4}
$D^+ \to \pi^-\mu^+e^+$	32.42±0.46	(1.31±0.02)x10^{-3}
$D^+ \to \pi^+\mu^+e^-$	32.63±0.46	(1.32±0.02)x10^{-3}
$D^+ \to \pi^+e^+\mu^-$	29.98±0.47	(1.22±0.02)x10^{-3}
$D^+ \to K^-\mu^+e^+$	21.57±0.43	(7.94±0.18)x10^{-4}
$D^+ \to K^+\mu^+e^-$	25.10±0.46	(9.31±0.20)x10^{-4}
$D^+ \to K^+e^+\mu^-$	27.64±0.48	(1.01±0.02)x10^{-3}
$D_s^+ \to \pi^+ (\phi \to e^+e^-)$	48.47±1.02	(7.23±0.21)x10^{-4}
$D_s^+ \to \pi^-e^+e^+$	23.02±0.84	(3.30±0.14)x10^{-4}
$D_s^+ \to \pi^+e^+e^-$	10.80±0.60	(1.57±0.09)x10^{-4}
$D_s^+ \to K^-e^+e^+$	23.41±0.92	(2.85±0.13)x10^{-4}
$D_s^+ \to K^+e^+e^-$	5.55±0.48	(7.13±0.64)x10^{-5}
$D^+ \to \pi^+ (\phi \to e^+e^-)$	37.40±1.17	(6.71±0.26)x10^{-4}
$D^+ \to \pi^-e^+e^+$	24.38±0.89	(4.41±0.19)x10^{-4}
$D^+ \to \pi^+e^+e^-$	11.19±0.66	(2.07±0.13)x10^{-4}
$D^+ \to K^-e^+e^+$	24.26±0.97	(3.92±0.18)x10^{-4}
$D^+ \to K^+e^+e^-$	4.83±0.53	(7.59±0.86)x10^{-5}

Table 5.11: Overview of efficiencies used for 2015 data that are obtained using simulated data generated in 2016 conditions with reweighting applied to match the 2015 normalisation data. The efficiency varies significantly even between channels that seem similar, due to the different backgrounds changing the chosen working point.

reject events that are not of interest. While these cuts can be arbitrarily complex, the most common selection criteria is that all final state particles must be within a cone representing the acceptance of the LHCb detector. This cut is known internally as `DaughtersInLHCb`. If the signal meson is travelling away from the detector, the transformation $z \to -z$ is applied as an additional optimisation to *flip* the event into the LHCb acceptance.[9]

The efficiency of the generator level cuts vary slightly between the channels due to the slightly different kinematics of the various final states. To correct for these private simulation productions were run for every event type, with the `GEANT4` step disabled and the generator level cut removed. The data was then selected with the q^2 cuts as is used for the analysis and the `DaughtersInLHCb` cut was applied to obtain an efficiency. Only the magnet down polarity is generated as the `DaughtersInLHCb` generator level cut (Gen) is independent of the magnet polarity, as can be seen in Table 5.10. Note the values used for the analysis are doubled with respect to the official tables due to the event flipping[10] performed in `Gauss`. This efficiency is typically between 19 %-21 % and independent of the

[9] This assumes the collisions are forward-backward symmetric and is disabled when generating some types of simulated data such as beam-gas or proton-ion collisions.

[10] By default, if the hadron generated by Pythia has $p_z < 0$ the entire event is flipped by `Gauss`.

charge of the final state particles or if the parent particle is a D^+ or D^+_s. It is larger for final states containing kaons/muons than it is for final states containing pions/electrons.

$\epsilon_{\text{Reco+Strip}}$

The reconstruction and stripping efficiency is taken directly from simulated data. An additional correction for the tracking efficiency was considered however these were found to be negligible when taken as a ratio to the normalisation channel, as shown in Table 6.1, therefore these corrections are handled with the systematic uncertainties. Further considerations of the tracking efficiency are required for electrons and these are covered by the correction factor discussed in Section 5.5.3. This efficiency is typically between 1 %-4 % and independent of the charge of the final state particles. It is larger for decays of D^+ than D^+_s and for final states containing pions/muons than it is for final states containing kaons/electrons.

$\epsilon_{\text{Trigger}}$

The trigger efficiency is taken from simulation relative to the reconstruction and stripping efficiency. For HLT2 the selection is simulated using identical cuts for the reasons discussed in Section 5.5.1. This efficiency is typically between 20 %-60 % and behaves similarly to $\epsilon_{\text{Reco+Strip}}$ except the efficiency for D^+_s is larger than D^+.

$\epsilon_{\text{Selection}}$

The final selection efficiency is also taken from simulated data using the cuts obtained. This efficiency also includes the effects of the reduced mass window used in the fit (1802 MeV to 2050 MeV) as well as any kinematic regions that were excluded to remove resonances as described in Section 5.2.2. This efficiency varies significantly between channels due to the different backgrounds and dilepton q^2 requirements causing the optimal working point of the PID and BDT to change between channels.

$\epsilon_{\text{electron}}$

The precision to which electrons are described in simulation is known to be less precise than other final state particles due to their more complex interactions with the material in the detector. As a result it is expected that the correction applied to achieve good data/simulation agreement when calculating the absolute efficiency will differ for final states with electrons and final states with muons.

To probe this effect the two different decays of the ϕ meson can be used to obtain an efficiency correction from the double ratio of efficiency corrected yields in $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+$ and $D^+_{(s)} \to (\phi \to e^-e^+)\,\pi^+$ in real and simulated data. It is assumed that the effect on the efficiency correction for each electron ($\epsilon_{\text{electron}}$) is independent, therefore the per electron efficiency is the square root of the ratio, i.e.

$$\epsilon^2_{\text{electron}} = \frac{N_{\text{data}}\left[D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+\right]}{N_{\text{data}}\left[D^+_{(s)} \to (\phi \to e^-e^+)\,\pi^+\right]} \cdot \frac{\epsilon_{\text{MC}}\left[D^+_{(s)} \to (\phi \to e^-e^+)\,\pi^+\right]}{\epsilon_{\text{MC}}\left[D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+\right]}$$

Attributing the difference primarily to the modelling of electrons, would mean that values greater than one correspond to a higher electron efficiency in simulation than in data.

The corrections applied are listed in Table 5.12 and the values for D^+ and D_s^+ agree at $1.5\ \sigma$. Further data simulation differences that arise from the offline selection are corrected for using the reweighting method described in Section 5.5.1 and are probed using the cross-check described in Section 6.3. Corrections for 2016 are used for the 2015 cross-check as the systematic uncertainty on the fit to the 2015 $\phi \to e^-e^+$ data is large due to the limited statistics.

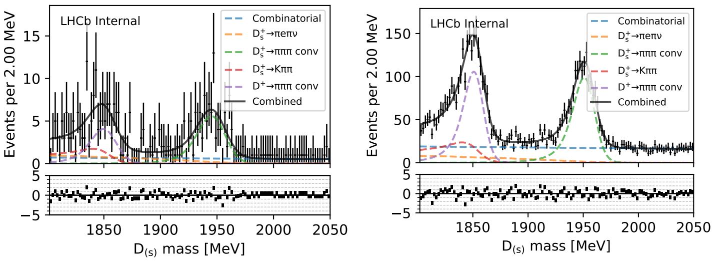
Year	Parent	Per electron correction (%)
2016	D^+	103.4±4.7
2016	D_s^+	110.5±4.2

Table 5.12: Per electron corrections made to the signal efficiency. The uncertainty includes the systematic uncertainty on the normalisation fit yield as described in Section 6.1.5.

5.6 Summary

A search has been prepared using LHCb data for 28 rare and forbidden decays of the form $D^+_{(s)} \to h^\pm l^+ l'^\mp$, where h is a charged kaon or pion and l is a electron or muon. Candidates are selected with multivariate methods with the efficiency of these selections is corrected using full detector Monte Carlo simulation. Corrections are applied to these simulation samples to further improve the agreement with real data. Yields are extracted using extended maximum likelihood fits with the background distributions being generated using fast simulation techniques and validated against real data.

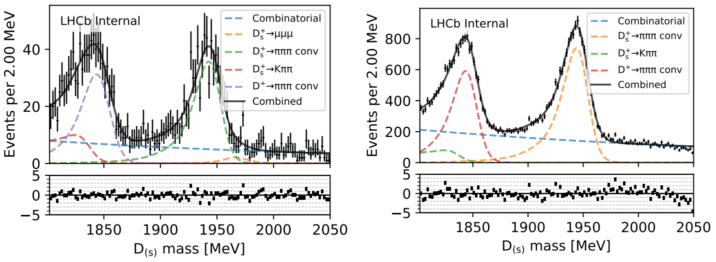
In the next chapter, the systematic uncertainties associated with these methods are estimated. Cross-checks are then performed using both reference samples and a $0.3\,\text{fb}^{-1}$ dataset. Finally, new upper limits are obtained for 25 of the 28 channels with 22 of these improving upon the previous world's best limits. Results for the remaining three channels are not available at the time of thesis submission.



(a) $\pi^+\pi^+\pi^-$ in 2015

(b) $\pi^+\pi^+\pi^-$ in 2016

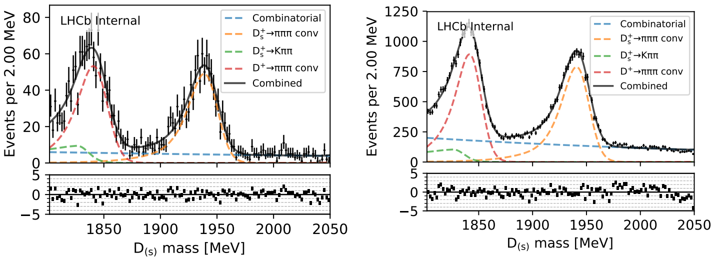
Figure 5.19: Reconstructed mass distributions for backgrounds to $D^+ \to \pi^+\mu^+\mu^-$ enhanced using PID cuts and fitted using `RooKeysPDF`. Components that have a convolution applied are denoted by *conv* in the plot legend.



(a) $\pi^+\pi^+\pi^-$ in 2015

(b) $\pi^+\pi^+\pi^-$ in 2016

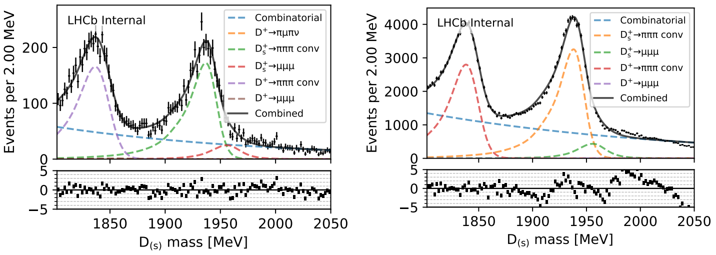
Figure 5.20: Reconstructed mass distributions for backgrounds to $D^+ \to \pi^+\mu^+e^-$ enhanced using PID cuts and fitted using `RooKeysPDF`. Components that have a convolution applied are denoted by *conv* in the plot legend.



(a) $\pi^+\pi^+\pi^-$ in 2015

(b) $\pi^+\pi^+\pi^-$ in 2016

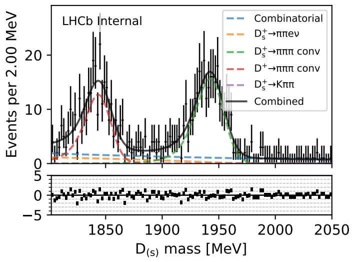
Figure 5.21: Reconstructed mass distributions for backgrounds to $D^+ \to \pi^+e^+\mu^-$ enhanced using PID cuts and fitted using `RooKeysPDF`. Components that have a convolution applied are denoted by *conv* in the plot legend.



(a) $\pi^+\pi^+\pi^-$ in 2015

(b) $\pi^+\pi^+\pi^-$ in 2016

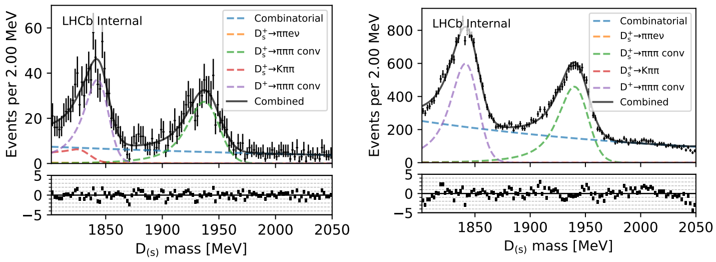
Figure 5.22: Reconstructed mass distributions for backgrounds to $D^+ \to \pi^+e^+e^-$ enhanced using PID cuts and fitted using `RooKeysPDF`. Components that have a convolution applied are denoted by *conv* in the plot legend.



(a) $\pi^+\pi^+\pi^-$ in 2015

(b) $\pi^+\pi^+\pi^-$ in 2016

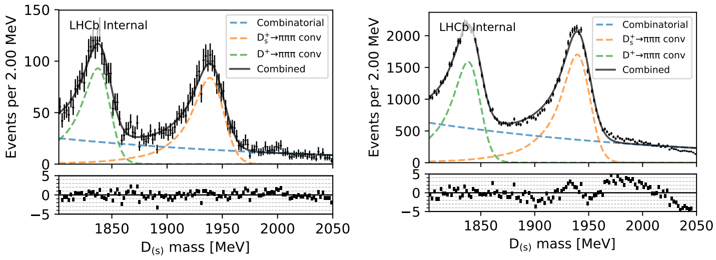
Figure 5.23: Reconstructed mass distributions for backgrounds to $D^+ \to \pi^-\mu^+\mu^+$ enhanced using PID cuts and fitted using `RooKeysPDF`. Components that have a convolution applied are denoted by *conv* in the plot legend.



(a) $\pi^+\pi^+\pi^-$ in 2015

(b) $\pi^+\pi^+\pi^-$ in 2016

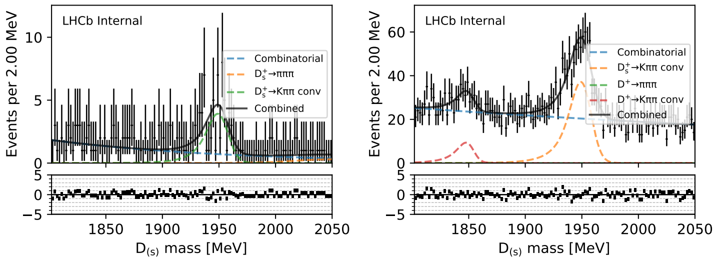
Figure 5.24: Reconstructed mass distributions for backgrounds to $D^+ \to \pi^-\mu^+e^+$ enhanced using PID cuts and fitted using `RooKeysPDF`. Components that have a convolution applied are denoted by *conv* in the plot legend.



(a) $\pi^+\pi^+\pi^-$ in 2015

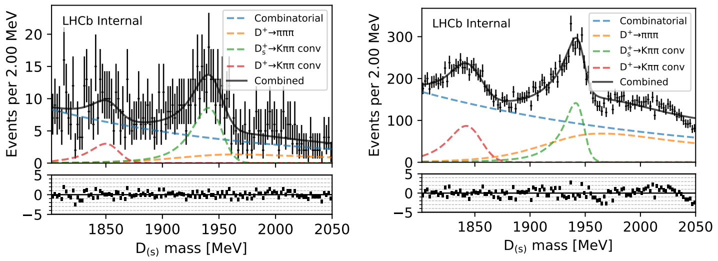
(b) $\pi^+\pi^+\pi^-$ in 2016

Figure 5.25: Reconstructed mass distributions for backgrounds to $D^+ \to \pi^-e^+e^+$ enhanced using PID cuts and fitted using `RooKeysPDF`. Components that have a convolution applied are denoted by *conv* in the plot legend.



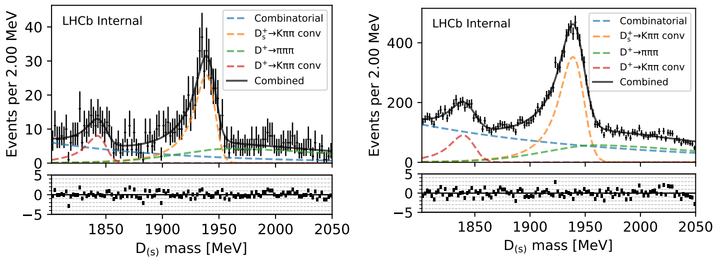
(a) $K^+\pi^+\pi^-$ in 2015 (b) $K^+\pi^+\pi^-$ in 2016

Figure 5.26: The $D^+ \to K^+\mu^+\mu^-$ decay channel data with background components enhanced using PID cuts (see text).



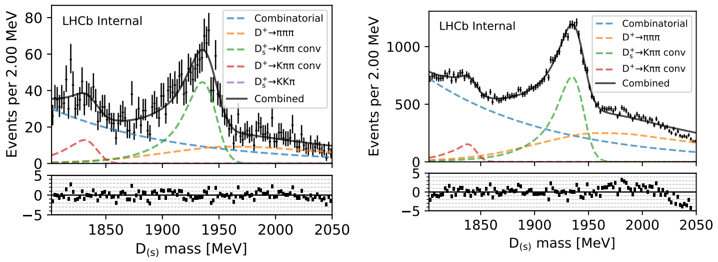
(a) $K^+\pi^+\pi^-$ in 2015 (b) $K^+\pi^+\pi^-$ in 2016

Figure 5.27: The $D^+ \to K^+\mu^+e^-$ decay channel data with background components enhanced using PID cuts (see text).



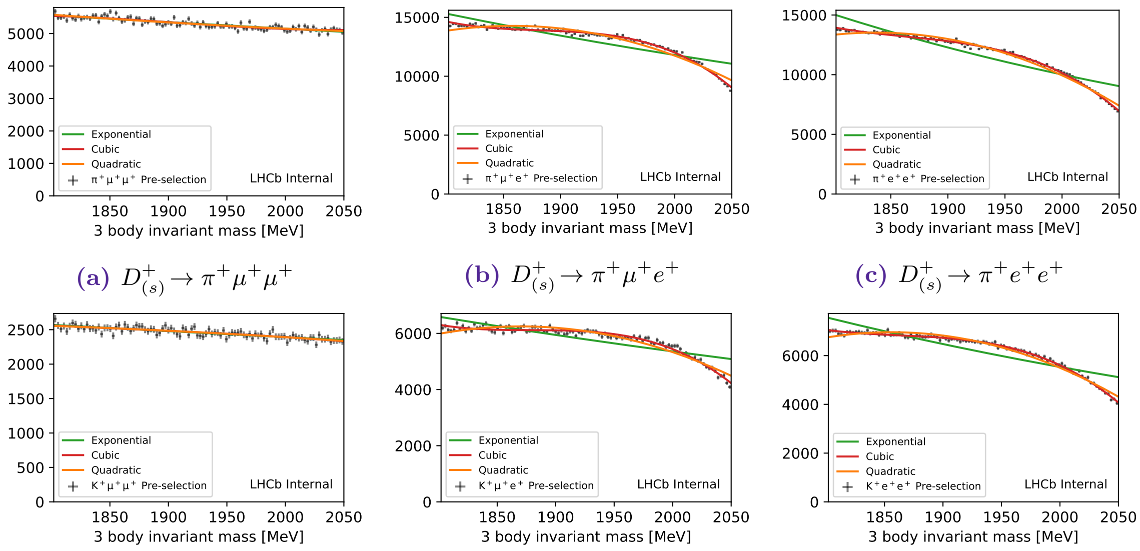
(a) $K^+\pi^+\pi^-$ in 2015 (b) $K^+\pi^+\pi^-$ in 2016

Figure 5.28: The $D^+ \to K^+e^+\mu^-$ decay channel data with background components enhanced using PID cuts (see text).



(a) $K^+\pi^+\pi^-$ in 2015 (b) $K^+\pi^+\pi^-$ in 2016

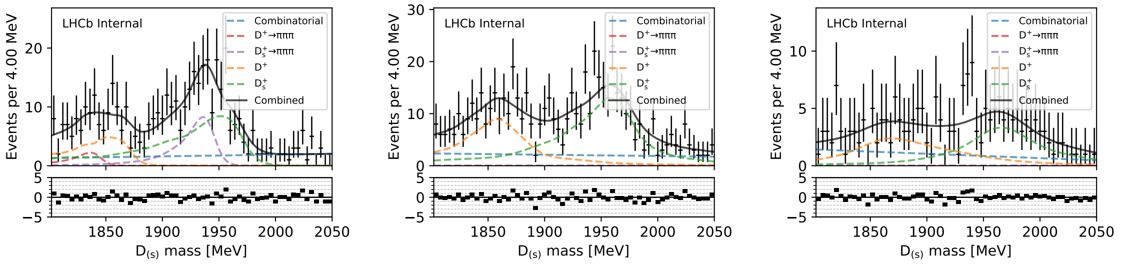
Figure 5.29: The $D^+ \to K^+e^+e^-$ decay channel data with background components enhanced using PID cuts (see text).



(a) $D^+_{(s)} \to \pi^+\mu^+\mu^+$ **(b)** $D^+_{(s)} \to \pi^+\mu^+e^+$ **(c)** $D^+_{(s)} \to \pi^+e^+e^+$

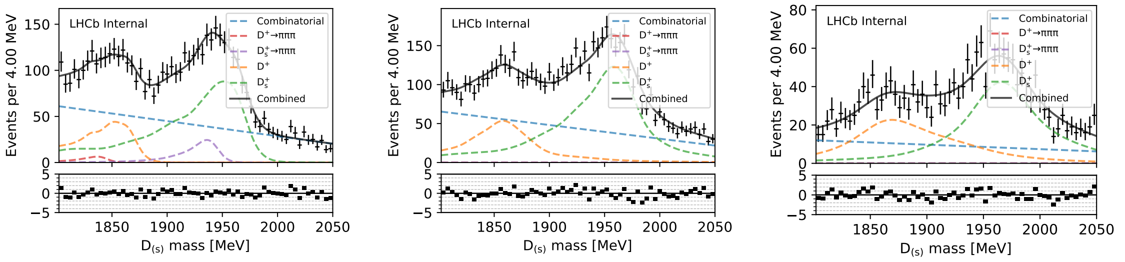
(d) $D^+_{(s)} \to K^+\mu^+\mu^+$ **(e)** $D^+_{(s)} \to K^+\mu^+e^+$ **(f)** $D^+_{(s)} \to K^+e^+e^+$

Figure 5.30: Mass distributions in pure combinatorial background samples, where all three final state hadrons have the same charge, with a loose selection applied. Each sample has a fitted exponential, quadratic and cubic distribution overlaid and a clear difference can be seen when electrons are present in the final state.



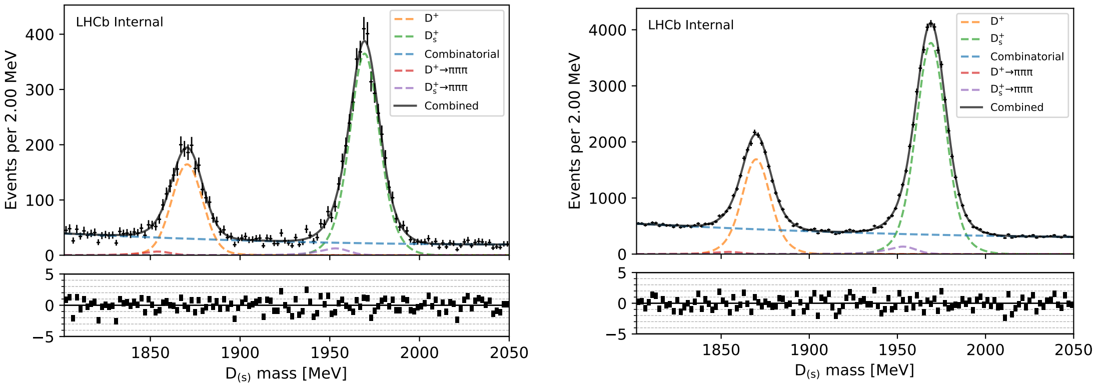
(a) 0 photons added **(b)** 1 photon added **(c)** 2+ photons added

Figure 5.31: Components of the fit to the invariant mass distribution of the normalisation channel $D^+_{(s)} \to (\phi \to e^-e^+)\,\pi^+$ for 2015, split by the number of bremsstrahlung photons that have been added to the candidates.



(a) 0 photons added **(b)** 1 photon added **(c)** 2+ photons added

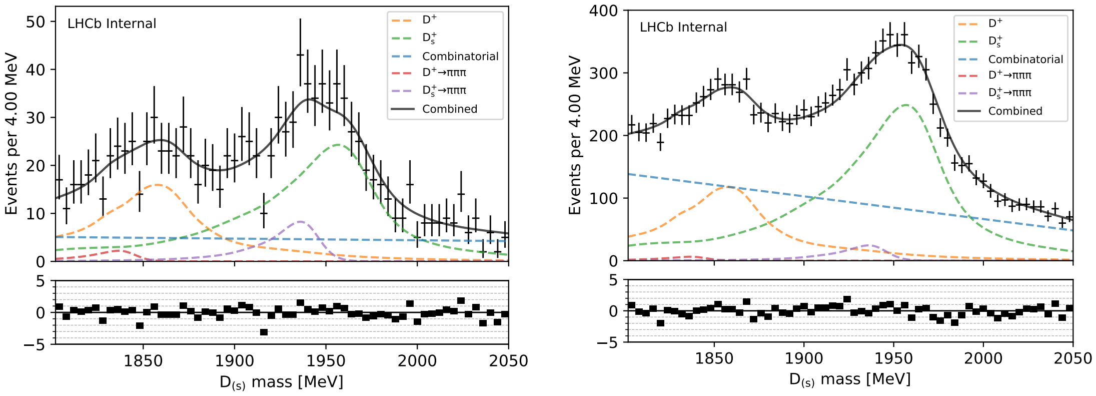
Figure 5.32: Components of the fit to the invariant mass distribution of the normalisation channel $D^+_{(s)} \to (\phi \to e^-e^+)\,\pi^+$ for 2016, split by the number of bremsstrahlung photons that have been added to the candidates.



(a) 2015

(b) 2016

Figure 5.33: Fits to the invariant mass distribution of the normalisation channel $D^+_{(s)} \to (\phi \to \mu^- \mu^+)\ \pi^+$ using data from the years indicated.



(a) 2015

(b) 2016

Figure 5.34: Combined fits to the invariant mass distribution of the normalisation channel $D^+_{(s)} \to (\phi \to e^- e^+)\ \pi^+$ using data from the years indicated.

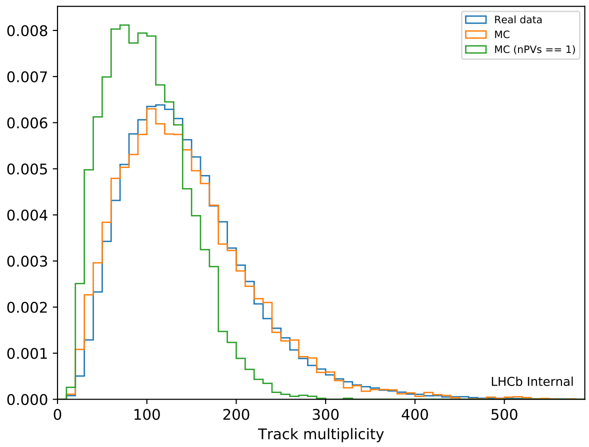


Figure 5.35: Comparison between track multiplicity distributions in data and simulation for 2016 $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+$ showing the effect of requiring one reconstructed primary vertex is poorly modelled in the distribution of track multiplicity. No background subtraction is applied to the data sample.

Blank page

Chapter 6

Cross-checks and results for $D^+_{(s)} \to h^{\pm}l^+l'^{\mp}$

The previous chapter described the method used in the search for 28 rare and forbidden decays of the form $D^+_{(s)} \to h^{\pm}l^+l'^{\mp}$, where h is a charged kaon or pion and l is a electron or muon. This chapter continues this work by studying possible systematic effects and assigning uncertainties where appropriate. The technique for obtaining an upper limit using the CL$_s$ method[193, 194] is then described in Section 6.2, followed by a further cross-check with $0.3\,\text{fb}^{-1}$ of LHCb data that was collected in 2015. Finally the full $1.5\,\text{fb}^{-1}$ dataset is unblinded with upper limits obtained for 25 of the signal channels, 23 of which represent the world's most precise measurements. The remaining three signal channels remain in review at the time of thesis submission.

6.1 Systematic uncertainties

When performing a measurement of a quantity there can be unknown effects that modify the final result. To obtain a best estimate for the true value it is necessary to quantitatively estimate these effects and assign *systematic uncertainties* where appropriate.

For this analysis there are three main sources of systematic uncertainty; those from the signal and background fit shapes used, those from inaccuracies in the signal efficiency determination and those from the normalisation channel branching fraction. These are included in the likelihood as nuisance parameters when setting the limit in Section 6.2.

6.1.1 Decay model used for simulated data

The decay model used when generating the signal simulation samples assumes the decay has equal probabilities in all areas of the Dalitz plane. This may not be representative of the real decay. However, the results of this analysis are quoted under this assumption, as the true decay model is not known and hence no systematic is ascribed.

6.1.2 Finite simulated sample size

All efficiencies derived from simulation are subject to a systematic from the finite size of the sample used. The size of this uncertainty varies between and 1 % and 7 % and can be found in Section 5.5.

6.1.3 Track reconstruction

It is known that the tracking efficiency reproduced in simulated data is not fully representative of the efficiency in the detector. As a result the LHCb tracking group provide tables to correct simulated data[195]. These are generated using a tag-and-probe technique where $J/\psi \to \mu^+\mu^-$ decays are selected by only making requirements on one muon. Three methods are then used to imply the presence of the other muon:

- **VELO method**: Only require a track in the T stations.
- **T-station method**: Only require a track in the VELO and muon stations.
- **Long method**: Only require a track in TT and muon stations.

These partially reconstructed tracks can then be compared with unused track segments to estimate the efficiency of each sub-detector.

The tables used for this analysis can be found at Reference [196] and give correction factors that should be applied on a track by track basis depending on the momentum and pseudorapidity of the track. To account for the correlated uncertainty the Python package `mcerp`[197] was used to propagate the uncertainty on the corrections, and their correlations, using pseudoexperiments. Additionally, two systematic uncertainties are also provided by the tracking group:

- A per track uncertainty is taken to be 0.8 % for 2016 and accounts for other differences in data and simulation that could be correlated with the tracking efficiency. For this analysis it is assumed to be uncorrelated between tracks of different particle species.
- An additional uncertainty of 1.1 % to 1.5 % for the uncertainty in the LHCb material causing hadronic interaction rates to differ in simulation. This is caused by the simplified nature of the simulated detector and conservatively taken to be 1.5 % and is only included for channels containing a kaon; i.e. where the final state hadron differs from that in the normalisation channel.

The relative tracking correction factors and their uncertainties are listed in Table 6.1.

6.1.4 Reweighting classifier

As explained in Section 5.5.1 a multivariate classifier is used for reweighting the simulated data to match the variable distributions seen in real data. For dimuon and dielectron channels this is validated with the cross-check described in Section 6.3, however, for channels with mixed leptons a choice must be made for which classifier to use for reweighting. Due to the much larger sample of $\phi \to \mu^-\mu^+$ events the dimuon classifiers were chosen.

Channel	Absolute efficiency	Relative efficiency
$D^+ \to \pi^+\mu^+\mu^-$	(96.17±2.23)%	(99.87±0.05)%
$D^+ \to K^+\mu^+\mu^-$	(96.50±2.24)%	(100.24±2.27)%
$D^+ \to \pi^-\mu^+\mu^+$	(96.20±2.24)%	(99.90±0.01)%
$D^+ \to K^-\mu^+\mu^+$	(96.52±2.24)%	(100.26±2.27)%
$D^+ \to \pi^-e^+e^+$	(95.82±2.20)%	(99.54±2.25)%
$D^+ \to K^-e^+e^+$	(96.21±2.19)%	(99.96±3.18)%
$D^+ \to K^+e^+e^-$	(96.16±2.18)%	(99.92±3.18)%
$D^+ \to \pi^+e^+e^-$	(95.86±2.20)%	(99.58±2.26)%
$D^+ \to \pi^+\mu^+e^-$	(95.78±1.92)%	(99.48±1.14)%
$D^+ \to \pi^-\mu^+e^+$	(95.85±1.92)%	(99.55±1.14)%
$D^+ \to \pi^+e^+\mu^-$	(95.84±1.92)%	(99.54±1.14)%
$D^+ \to K^+\mu^+e^-$	(96.10±1.91)%	(99.84±2.52)%
$D^+ \to K^-\mu^+e^+$	(96.11±1.91)%	(99.85±2.52)%
$D^+ \to K^+e^+\mu^-$	(96.12±1.91)%	(99.85±2.52)%
$D_s^+ \to \pi^+\mu^+\mu^-$	(96.36±2.24)%	(99.87±0.06)%
$D_s^+ \to K^+\mu^+\mu^-$	(96.66±2.26)%	(100.21±2.27)%
$D_s^+ \to \pi^-\mu^+\mu^+$	(96.35±2.25)%	(99.86±0.03)%
$D_s^+ \to K^-\mu^+\mu^+$	(96.68±2.26)%	(100.23±2.27)%
$D_s^+ \to K^+e^+e^-$	(96.29±2.19)%	(99.85±3.18)%
$D_s^+ \to \pi^+e^+e^-$	(95.96±2.21)%	(99.48±2.26)%
$D_s^+ \to K^-e^+e^+$	(96.32±2.20)%	(99.88±3.18)%
$D_s^+ \to \pi^-e^+e^+$	(96.03±2.21)%	(99.55±2.26)%
$D_s^+ \to K^+e^+\mu^-$	(96.25±1.92)%	(99.79±2.52)%
$D_s^+ \to K^+\mu^+e^-$	(96.27±1.92)%	(99.82±2.52)%
$D_s^+ \to K^-\mu^+e^+$	(96.27±1.91)%	(99.82±2.52)%
$D_s^+ \to \pi^+e^+\mu^-$	(95.98±1.92)%	(99.50±1.15)%
$D_s^+ \to \pi^+\mu^+e^-$	(96.03±1.93)%	(99.54±1.14)%
$D_s^+ \to \pi^-\mu^+e^+$	(96.00±1.92)%	(99.51±1.15)%

Table 6.1: Tracking correction factors for 2016 simulation as both absolute correction factors and relative to $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\,\pi^+$ simulation.

To obtain a systematic uncertainty on this choice of classifier, the efficiency is calculated with the dielectron classifier. In both cases the electron correction is included. Figure 6.1 shows the change in efficiency in units of the uncertainty and observed difference is systematically smaller in magnitude for D_s^+ than D^+, likely due to the large uncertainty on the `sWeights` from the fit to $D^+ \to (\phi \to e^-e^+)\,\pi^+$. To account for any possible bias in this distribution the RMS of the points in Figure 6.1 (7.6 %) is taken as a systematic uncertainty when computing limits.

6.1.5 Normalisation channel yield

The signal yield of the normalisation channel is extracted using a maximum likelihood fit as described in Section 5.4.4. The most likely cause of a sizeable systematic uncertainty is the incorrect treatment of physical backgrounds near the peak. To obtain a maximum possible uncertainty for these the normalisation channel fits are repeated with the $D^+_{(s)} \to$

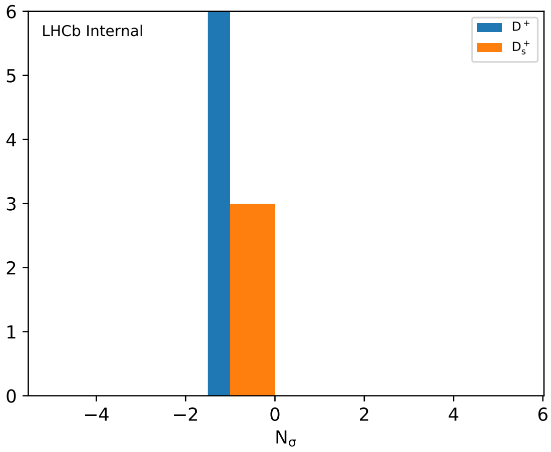


Figure 6.1: Change in efficiency, in units of σ, from using a different classifier when computing the efficiency for the 12 mixed lepton channels when using the $\phi \to e^-e^+$ reweighting classifiers instead of $\phi \to \mu^-\mu^+$.

Channel	Year	Nominal model				
		N_{D^+}	$N_{D_s^+}$	$N_{D^+\to\pi^+\pi^+\pi^-}$	$N_{D_s^+\to\pi^+\pi^+\pi^-}$	N_{comb}
$D^+ \to \pi^+ (\phi \to e^+e^-)$	2015	290±52	518±50	17±15	79±23	289±77
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	2015	1817±109	3994±99	76±68	143±78	3241±140
$D^+ \to \pi^+ (\phi \to e^+e^-)$	2016	2156±175	5323±178	49±55	235±72	5789±273
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	2016	18105±341	42042±378	460±250	1542±287	49008±486

Channel	Year	Simplified model		
		N_{D^+}	$N_{D_s^+}$	N_{comb}
$D^+ \to \pi^+ (\phi \to e^+e^-)$	2015	286±52	602±46	305±77
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	2015	1849±77	4102±87	3320±123
$D^+ \to \pi^+ (\phi \to e^+e^-)$	2016	2172±160	5595±152	5787±249
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	2016	18211±249	43222±323	49723±456

Table 6.2: Fit component yields for the normalisation channels using the nominal and simplified fit models.

$\pi^+\pi^+\pi^-$ background components removed, henceforth known as the "simplified model". The component yields are given in Table 6.2.

The difference in signal yield between the nominal and simplified model is then taken as a systematic uncertainty and is given in Table 6.3. For the 2015 cross-check this systematic introduces a prohibitively large uncertainty on the electron correction factors due to fit instabilities with the limited statistics. The systematic uncertainty obtained for the 2016 data sample is also applied to the cross-check performed with the 2015 data.

6.1.6 Normalisation branching fraction

The uncertainty on the normalisation branching fraction of the normalisation channel is taken from the PDG[150]. The uncertainty is somewhat smaller that than in the previous LHCb analysis[155], as the branching fractions are calculated relative to the measurements for $\phi \to K^-K^+$, as described in Section 5.1. The relative uncertainties of the branching fractions are given in Table 6.4.

Channel	Year	Signal yield difference $\Delta_{N_{D^+}}$	$\Delta_{N_{D_s^+}}$	Relative yield uncertainty $\epsilon_{N_{D^+}}$	$\epsilon_{N_{D_s^+}}$
$D^+ \to \pi^+ (\phi \to e^+e^-)$	2015	2.45%	17.32%	17.98%	9.70%
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	2015	2.12%	2.77%	6.01%	2.48%
$D^+ \to \pi^+ (\phi \to e^+e^-)$	2016	1.43%	5.22%	8.12%	3.34%
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	2016	0.62%	2.82%	1.88%	0.90%

Table 6.3: Systematic uncertainty on the signal yield in each of the normalisation channels. The "relative yield uncertainty" columns show the relative statistical uncertainty on the yield for reference. The uncertainties are propagated using `mcerp`[197] and the mean of the resulting distributions are reported.

Source	This analysis
$\mathcal{B}(D^+ \to (\phi \to \mu^-\mu^+)\,\pi^+)$	7.4 %
$\mathcal{B}(D_s^+ \to (\phi \to \mu^-\mu^+)\,\pi^+)$	7.5 %

Table 6.4: Systematic uncertainties on the normalisation channels branching fractions in this analysis.

6.1.7 Fit shapes

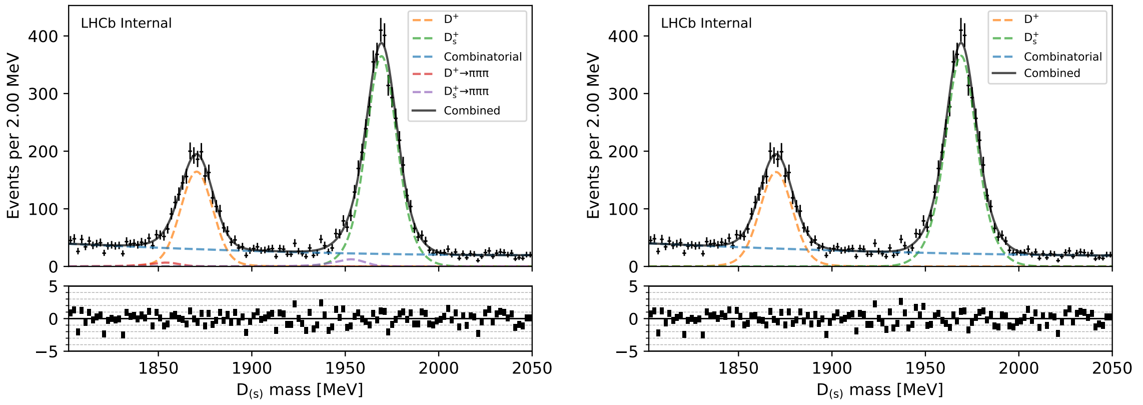
Section 5.4.1 describes the method used for obtaining the PDF that describes the signal shape. This method is entirely reliant on simulated data and may not perfectly represent real data. To account for potential differences the normalisation channels are fitted with an alternative model:

- **$D_{(s)}^+ \to (\phi \to \mu^-\mu^+)\;\pi^+$:** The yield obtained with the kernel density estimations (the nominal signal model) are compared to the yield obtained with a triple Gaussian (the nominal normalisation channel model). The fits are shown in Figure 6.2 and 6.3.
- **$D_{(s)}^+ \to (\phi \to e^-e^+)\;\pi^+$:** A simultaneous fit is used for these normalisation channels using the `RooKeysPDF` shapes extracted from simulation, splitting by the number of bremsstrahlung photons that are reconstructed. The nominal normalisation model allows the relative yield between the categories to vary. For the comparison the fraction of events in each category is fixed from fully selected simulation. The fits are shown in Figure 6.4 and 6.5.

Table 6.5 and 6.6 show the relative change in signal yield. For dimuon final states, the maximum observed difference in yield for $D_{(s)}^+ \to (\phi \to \mu^-\mu^+)\,\pi^+$ is used for the systematic uncertainty. In all other cases the maximum observed difference in yield for all channels is used for the systematic uncertainty.

6.1.8 Particle identification

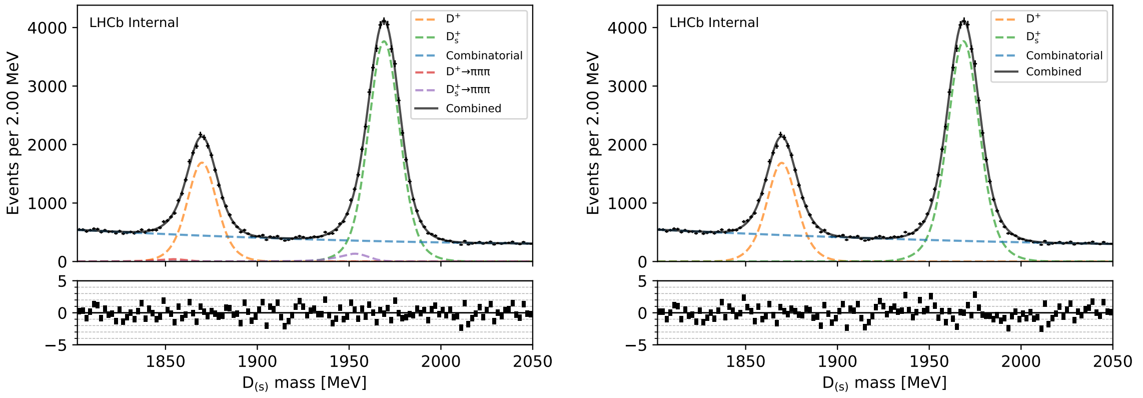
The simulated particle identification variables are corrected using `PIDGen` and the Kernel Density Estimation (KDE) method has associated systematic uncertainties. The first



(a) Triple Gaussian

(b) Kernel density estimation

Figure 6.2: Fitted reconstructed mass distributions for 2015 $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\ \pi^+$ with different models used for the signal shape between (a) and (b) as indicated by the captions.



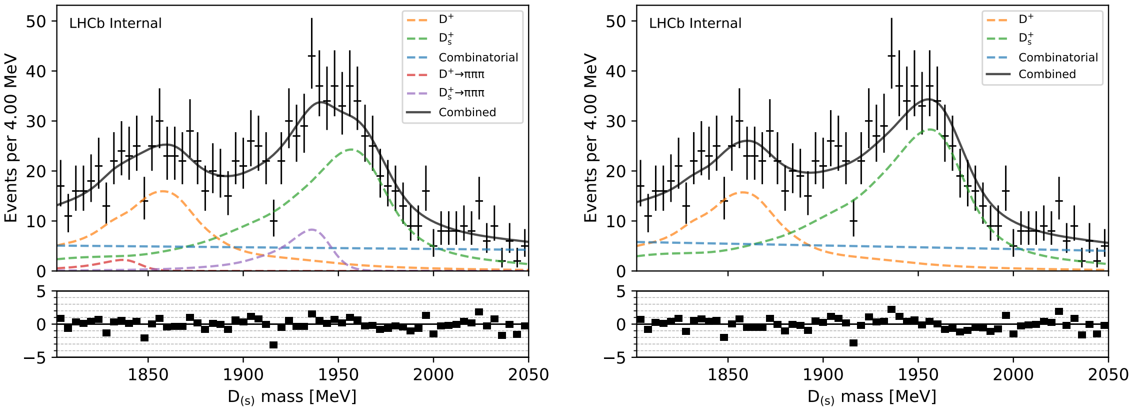
(a) Triple Gaussian

(b) Kernel density estimation

Figure 6.3: Fitted reconstructed mass distributions for 2016 $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\ \pi^+$ with different models used for the signal shape between (a) and (b) as indicated by the captions.

source is from the finite size of the calibration samples and this is estimated by bootstrapping[198] the calibration samples to create five additional KDEs. These variations are known as `stat_0`, `stat_1`, `stat_2`, `stat_3`, `stat_4`. An additional variation is also available that uses a 50 % larger bandwidth when fitting the KDE and this is known as `syst_1`.

To check how these variations might affect this analysis the efficiency of the final selection was recomputed for each of these variations. The uncertainty on these values arises from the finite size of the simulated sample and does not account for the correlation that is present due to the same simulated events being used for computing the efficiency with both the nominal and varied PID variables. Table 6.7 shows these differences in units of the statistical uncertainty on the final selection efficiency and Figure 6.6 shows these values as a histogram. This shows that any systematic uncertainty from the use of particle identification is negligible in comparison to the systematic from the finite MC size that is



(a) Floating fractions

(b) Fixed fraction

Figure 6.4: Fitted reconstructed mass distributions for 2015 $D^+_{(s)} \to (\phi \to e^- e^+)\, \pi^+$ with different models used for the signal shape. (left) allows the fraction between the bremsstrahlung categories to float and (right) fixes the fraction from fully selected simulation.

Channel	Relative yield difference
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	1.3 %
$D_s^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	0.7 %
$D^+ \to \pi^+ (\phi \to e^+e^-)$	6.2 %
$D_s^+ \to \pi^+ (\phi \to e^+e^-)$	3.6 %

Table 6.5: Relative change in signal fit yield for $D^+_{(s)} \to (\phi \to l^- l^+)\, \pi^+$ when fitting the reconstructed mass distributions with the `RooKeysPDF` template instead of a triple Gaussian for 2015 data.

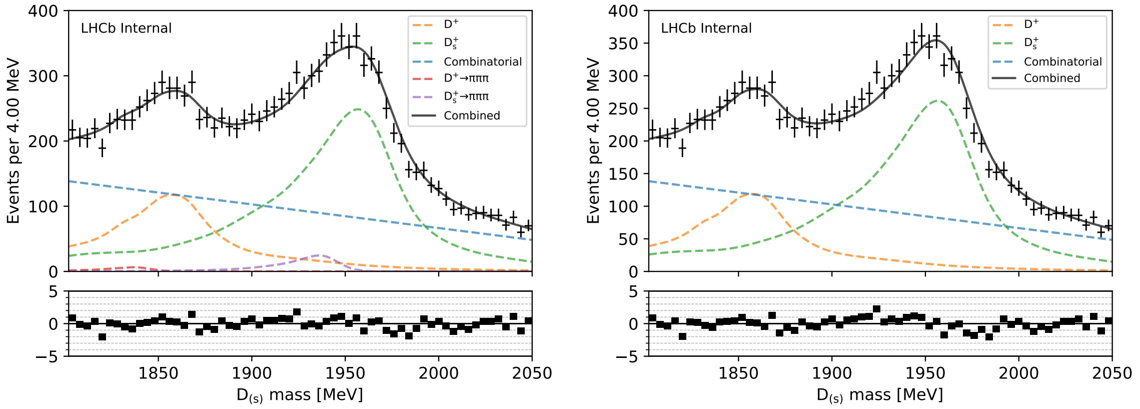
included in the final results. The finite MC systematic is itself a sub-dominant systematic uncertainty.

6.1.9 Summary tables

Table 6.8 gives a summary of the systematics used for the 2015 dataset. Note these values are not accurate for 2015, since simulation was only available for 2016, and the 2015 results are provided only as a cross-check. The systematic uncertainties for the 2016 dataset, on which the measurements are made, are given in Table 6.9. These tables do not include the small systematic uncertainty for the background model, as this is not applied as an efficiency variation, this is discussed in the section below.

6.1.10 Background model

It is possible that the background model chosen is not suitable for describing the backgrounds actually present in the dataset. To ensure that the limit is not significantly affected limits are calculated for the nominal model and for four additional parametrisations, three of which test alternative combinatorial background models:



(a) Floating fractions

(b) Fixed fraction

Figure 6.5: Fitted reconstructed mass distributions for 2016 $D^+_{(s)} \to (\phi \to e^- e^+)\ \pi^+$ with different models used for the signal shape. (left) allows the fraction between the bremsstrahlung categories to float and (right) fixes the fraction from fully selected simulation.

Channel	Relative yield difference
$D^+ \to \pi^+ (\phi \to \mu^+\mu^-)$	0.3 %
$D^+_s \to \pi^+ (\phi \to \mu^+\mu^-)$	0.8 %
$D^+ \to \pi^+ (\phi \to e^+e^-)$	7.0 %
$D^+_s \to \pi^+ (\phi \to e^+e^-)$	5.2 %

Table 6.6: Relative change in fit yield for $D^+_{(s)} \to (\phi \to l^- l^+)\ \pi^+$ when fitting with the `RooKeysPDF` template instead of a triple Gaussian for 2016 data. These values are used to assign a systematic uncertainty for the signal fit shapes.

- **Nominal combinatorial** This is the nominal model described in Section 5.4.3 with a 3$^{\text{rd}}$ order Chebyshev polynomial distribution being used for channels with one or two electrons and an exponential distribution being used for dimuon final states.
- **Alternate combinatorial** The opposite model to the nominal case, i.e. an exponential distribution is used for channels with electrons and a cubic is used for dimuon final states.
- **Quadratic combinatorial** The combinational background is described by a 2$^{\text{nd}}$ order Chebyshev polynomial.
- **Linear combinatorial** The combinational background is described by a 1$^{\text{st}}$ order Chebyshev polynomial.

A fourth parametrisation was used to check an alternative model for the physical backgrounds observed. Nominally the physical backgrounds are modelled using `RooKeysPDF` shapes that are fitted to the output of RapidSim (see Section 5.4.2). These are then split into two different classes of final states:

- **Kaon + same charge leptons** Nominally the `RooKeysPDF` shapes are convoluted with a Gaussian with both the mean and width left floating in the final fit due to the large contribution from $D^+ \to K^-\pi^+\pi^+$. For the "modified convolutions" model, the width of the Gaussian is fixed to its starting position of 1 MeV to cause it to have negligible impact on the shape of the peak.

- **Pion + any** Nominally the `RooKeysPDF` shapes are convoluted with a Gaussian with both the mean floating in the final fit due to the large contribution from $D^+_{(s)} \to \pi^+\pi^+\pi^-$. For the "modified convolutions" model, the width of the Gaussian is also floating.
- **Other final states** For these channels the background contribution is small and therefore the nominal fit used the `RooKeysPDF` objects directly. For the "modified convolutions" model this shape is convoluted with a Gaussian with a fixed mean of zero and floating width.

The variation between alternative background models can be found in Figure 6.10 of Section 6.4 and is small compared to the size of the statistical uncertainty. To assign a systematic uncertainty for this effect, an equal number of toy datasets are produced for each of the five background models. The test statistics from each of these toy datasets are then combined to estimate the distribution of the test statistic. The observed test statistic is computed using the likelihood that was fitted to the real dataset using the nominal model. See Section 6.2 for more details about the limit setting procedure.

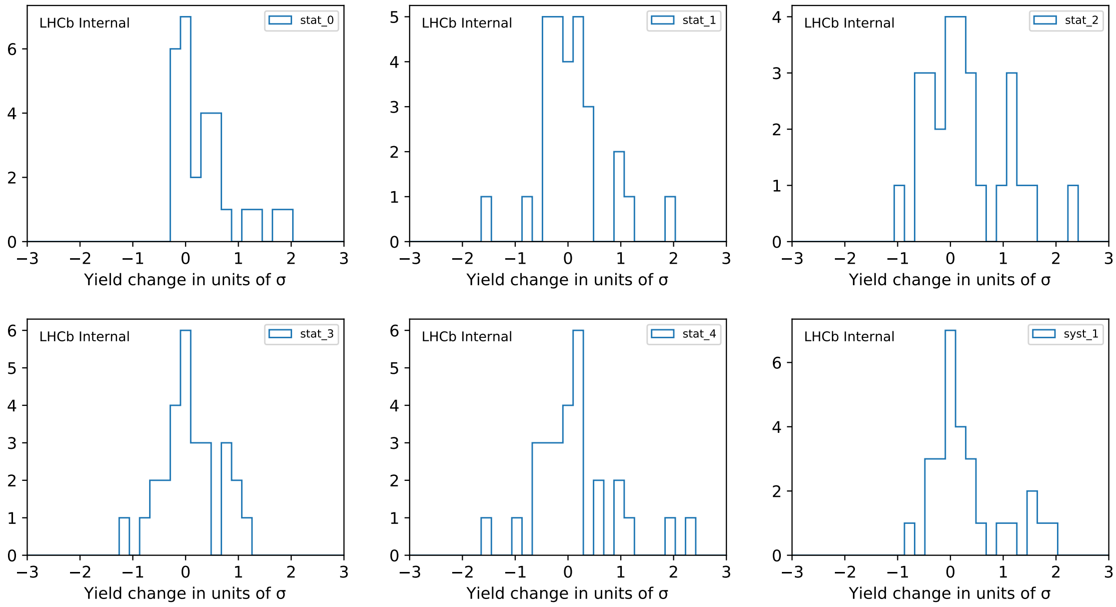


Figure 6.6: Histograms of the change of the final PID selection efficiency for each variation of the kernel destiny estimation used by `PIDGen`. The change is given in units of the statistical uncertainty of the efficiency that is described in Section 6.1.2.

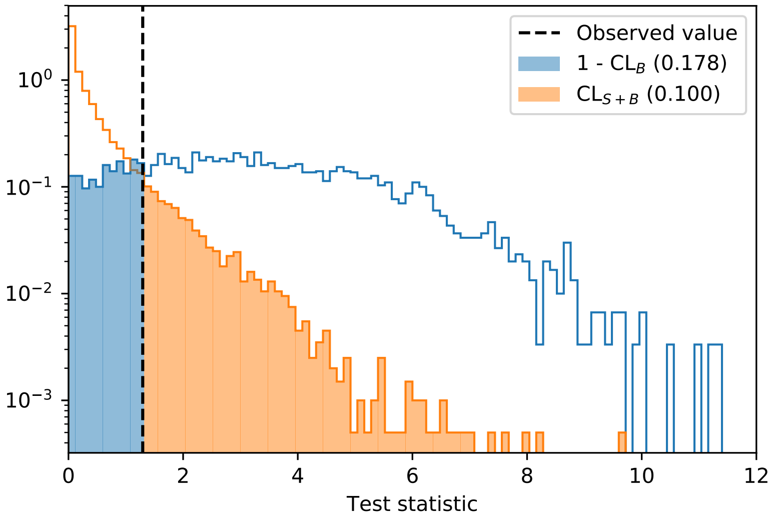


Figure 6.7: Example of the CL$_s$ method. The distribution of the test statistic from toys generated using two branching fraction hypotheses are shown and the dashed line corresponds to the observed value of the test statistic.

6.2 Limit setting

In this analysis the CL$_s$ method[193, 194] is used to compute an upper limit on the absolute branching fraction of each signal decay. As part of this, it is necessary to define a test statistic with which to compute the statistical significance. For this analysis the chosen test statistic is the *profile likelihood ratio*,

$$\lambda(\mathcal{B}, \hat{\theta}) = \frac{\mathcal{L}(\mathcal{B}, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mathcal{B}}, \hat{\theta})}, \tag{6.1}$$

where the likelihood is as defined in Section 5.4, $\mathcal{B}$ is the parameter of interest (the branching fraction), $\hat{\mathcal{B}}$ and $\hat{\theta}$ are the maximum likelihood estimates of the parameters in the dataset and $\hat{\hat{\theta}}$ is the maximum likelihood estimate of the parameters for a given value of $\mathcal{B}$. This then allows two hypotheses to be compared as shown in Figure 6.7 where

$$\mathrm{CL}_s = \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b}. \tag{6.2}$$

For this analysis a scan is performed for many possible branching fraction values and each of these is compared with the distribution for $\mathcal{B} = 0$.

This is achieved using the `RooStats` implementation of CL$_s$ and the improved version of the `StandardHypoTestInverter` demonstration that can be found in Reference [199]. This final limit obtained is set under the assumption that the signal is uniformly distributed across the Dalitz plane of the decay productions. Additionally, for those channels where there are SM contributions, several regions in dilepton q^2 are removed to avoid the resonant contributions, as described in Section 5.2.2. The efficiency is corrected for these removed regions. This analysis is not sensitive to any new physics contributions that might be present in or under the resonances.

If a signal peak had been seen after unblinding the 2016 dataset the measured branching fraction would instead be published with an associated significance calculated using the toys originally intended for the limit. Additionally, the Dalitz plane of the decay products would be split into four equal sized bins to check if the signal seen is uniformly distributed. This would result in 8 additional limits (4 for D^+ and 4 for D_s^+) or branching fraction measurements potentially being published.

The branching fraction measured in this analysis is defined as

$$\mathcal{B} = \frac{N_{D^+_{(s)} \to h^\pm l^+ l'^\mp}}{N_{D^+_{(s)} \to \pi^+ \phi(\mu^+\mu^-)}} \cdot \frac{\epsilon_{D^+_{(s)} \to \pi^+ \phi(\mu^+\mu^-)}}{\epsilon_{D^+_{(s)} \to h^\pm l^+ l'^\mp}} \cdot \mathcal{B}\left(D^+_{(s)} \to \pi^+ \phi\right) \cdot \mathcal{B}\left(\phi \to \mu^+\mu^-\right)$$

and is obtained from a fit to the invariant mass of the $D^+_{(s)}$ meson. The systematic uncertainties described in Section 6.1 are included with a log-normal distribution. To account for the potential overlap between the signal peaks, the yield of the other meson is floating in the fit and treated as an additional nuisance parameter when computing the significance of the signal peak.

6.3 Cross-checks of the ϕ branching ratio

To check if the efficiency of the offline selection criteria is sufficiently well understood a cross-check measurement is made of the branching ratio of $D^+_{(s)} \to \pi^+\mu^+\mu^-$ and $D^+_{(s)} \to \pi^+e^+e^-$ in the ϕ bin. The cross-check also serves to increase confidence that the analysis framework is free from bugs that would affect the final result. This measurement does not aim to be an independently useful result, it only serves to validate the main result. All branching ratios are measured relative to $D^+_{(s)} \to (\phi \to \mu^-\mu^+)\ \pi^+$, which is processed separately within the framework with a loose selection applied as described in Section 5.3.3. In contrast, the signal channel has a BDT requirement and tighter PID applied. As a result, this cross-check probes that the offline (post stripping) selection is well understood.

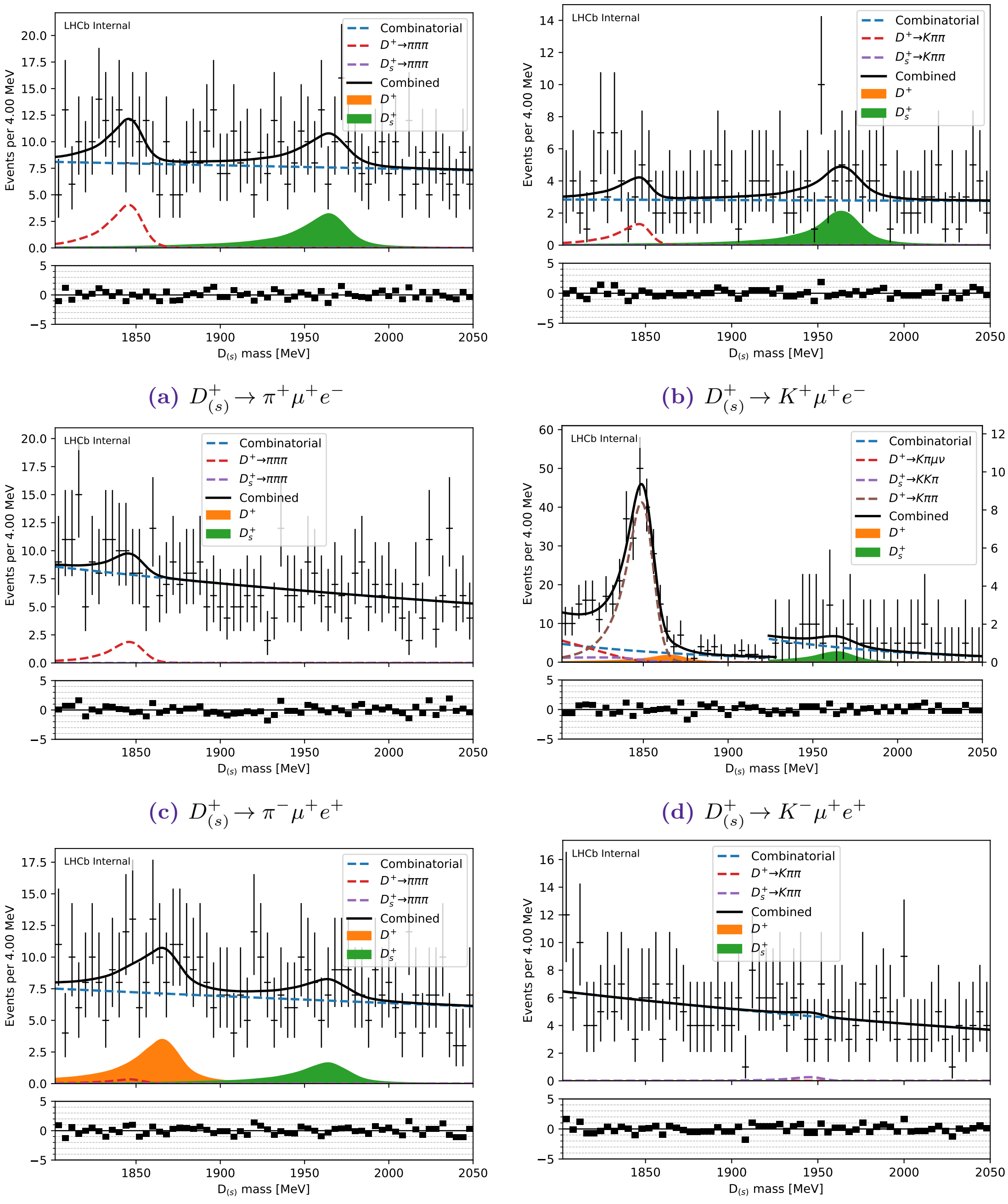
Table 6.10 shows the result of the cross-check measurements. All values agree within their uncertainties at the 1.5 sigma level or better. The fit results give only statistical uncertainties and are considered only as cross check-results.

6.4 Cross-check results using 2015 data

As discussed in Section 5.2.1, it was decided to unblind the 2015 dataset to ensure that the proposed analysis strategy worked for all final states. This was useful as there was no suitable sample that could be used to fully validate the expected background contributions and their shapes. Due to the smaller size of this sample, the convolution parameters of the $D^+_{(s)} \to \pi^+\pi^+\pi^-$ backgrounds were held fixed and an exponential distribution was used to describe the combinatorial background component for all channels. While this sample is significantly less sensitive than the full 2016 dataset, unbinding this sample could have potentially shown signal contributions that were smaller than the previous world's best limits. All results shown in this section are only indicative due to the use of inputs that are not applicable to this dataset, such as the simulated dataset and various efficiency corrections.

The fitted mass distributions are shown in Figure 6.8 and no obvious signal contributions can be seen. Approximate limits are obtained using the CL_s method and no significant deviations from the background only hypothesis are observed. These results are summarised in Figure 6.9 with the observed 90 % confidence upper limit and the previous world's best limits. The median expected limit is shown in orange, the $\pm 1\sigma$ band is shown as the central black box and the $\pm 2\sigma$ band is shown as the extended whiskers. Results using the five background models are shown independently in Figures 6.10 and all models give a reasonable description of the data and a similar result for the upper limit.

To estimate the 2016 results the expected sensitivity in the unblinded 2015 can be scaled by the square root of the luminosity difference ($\sqrt{5}$) to give an approximate sensitivity. Additionally, the blinded dimuon sideband datasets for 2016 can be fitted to given an alternative estimated limit. The approximate 2015 results and their extrapolations to the 2016 sample are shown in Table 6.11. The final results are given in Section 6.5 and are a small improvement over these estimations.



(a) $D^+_{(s)} \to \pi^+\mu^+e^-$

(b) $D^+_{(s)} \to K^+\mu^+e^-$

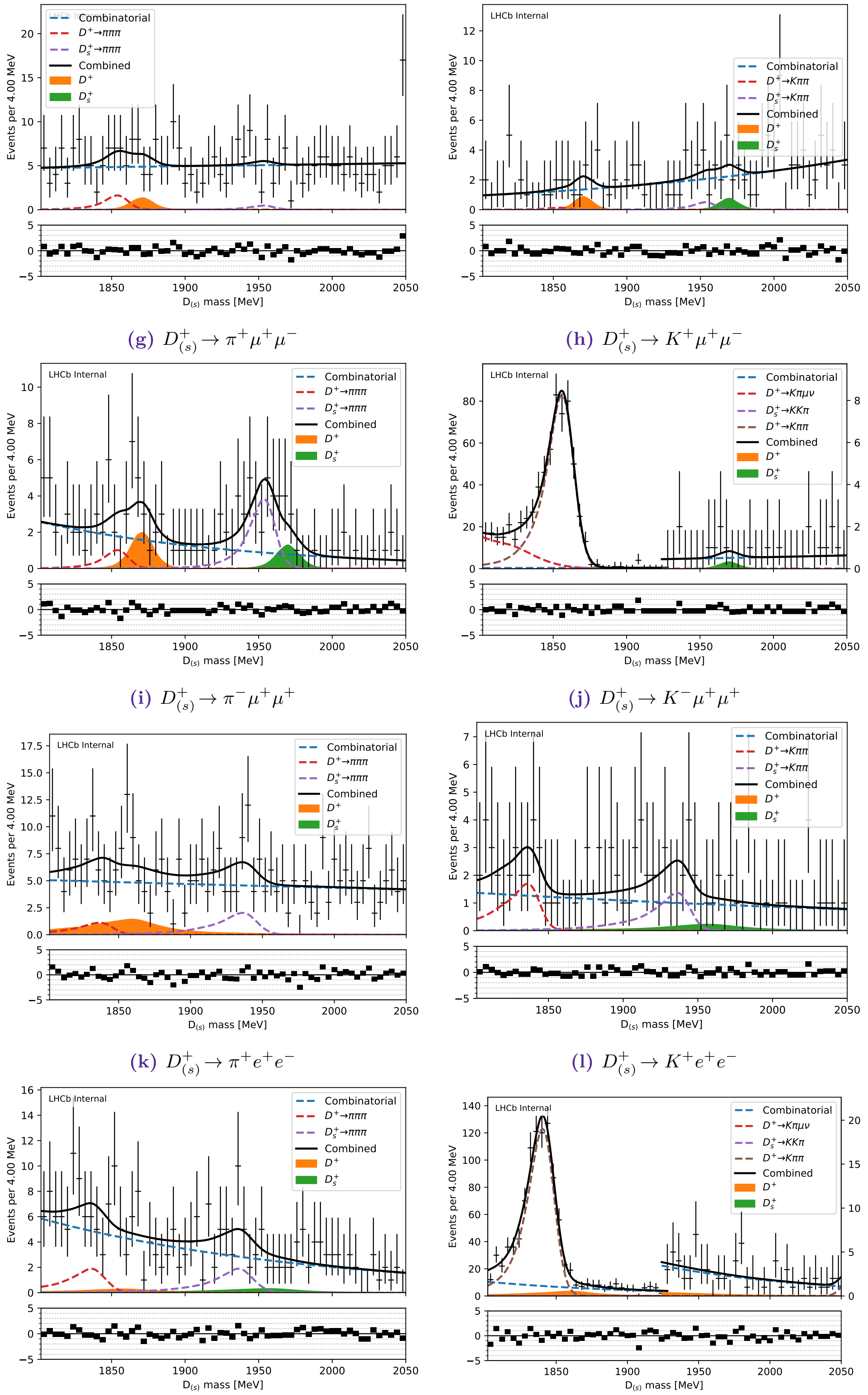
(c) $D^+_{(s)} \to \pi^-\mu^+e^+$

(d) $D^+_{(s)} \to K^-\mu^+e^+$

(e) $D^+_{(s)} \to \pi^+e^+\mu^-$

(f) $D^+_{(s)} \to K^+e^+\mu^-$

Figure 6.8: Fitted $D^+_{(s)}$ invariant mass distributions for the 2015 dataset. Due to the large difference between the background contribution to $D^+ \to K^-l^+l'^+$ and $D_s^+ \to K^-l^+l'^+$ the axis range is split into two ranges with different normalisations.



(g) $D^+_{(s)} \to \pi^+\mu^+\mu^-$

(h) $D^+_{(s)} \to K^+\mu^+\mu^-$

(i) $D^+_{(s)} \to \pi^-\mu^+\mu^+$

(j) $D^+_{(s)} \to K^-\mu^+\mu^+$

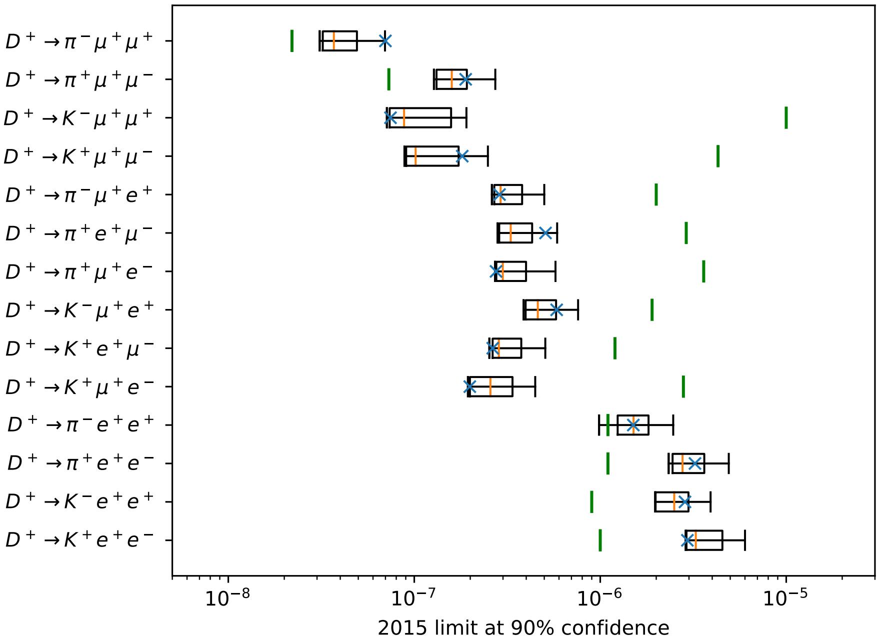
(k) $D^+_{(s)} \to \pi^+e^+e^-$

(l) $D^+_{(s)} \to K^+e^+e^-$

(m) $D^+_{(s)} \to \pi^-e^+e^+$

(n) $D^+_{(s)} \to K^-e^+e^+$

Figure 6.8: $D^+_{(s)}$ invariant mass distributions for the 2015 dataset (continued)



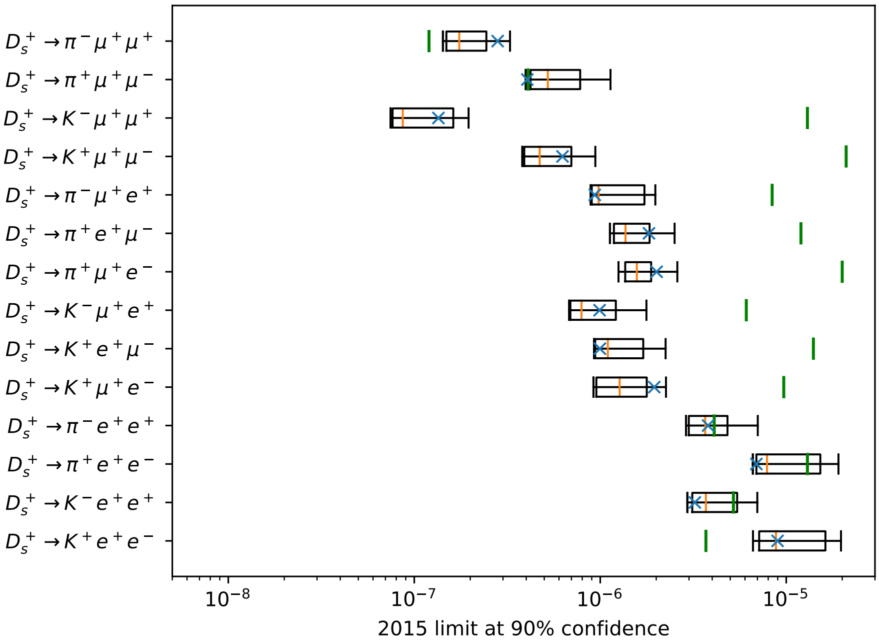


Figure 6.9: Approximate 90 % upper limits on the $D^+_{(s)}$ signal channels using the 2015 data. The median (orange), $\pm 1\sigma$ and $\pm 2\sigma$ expected limits are shown as box plots and the observed limit is given by a blue cross. The semi-transparent and dotted lines show the limit when calculated without systematic uncertainties. The green line shows the prior world's best limit for each channel. As discussed in the text as a consequence of approximations made for this data sample these results are cross-checks only and not final results of the analysis.

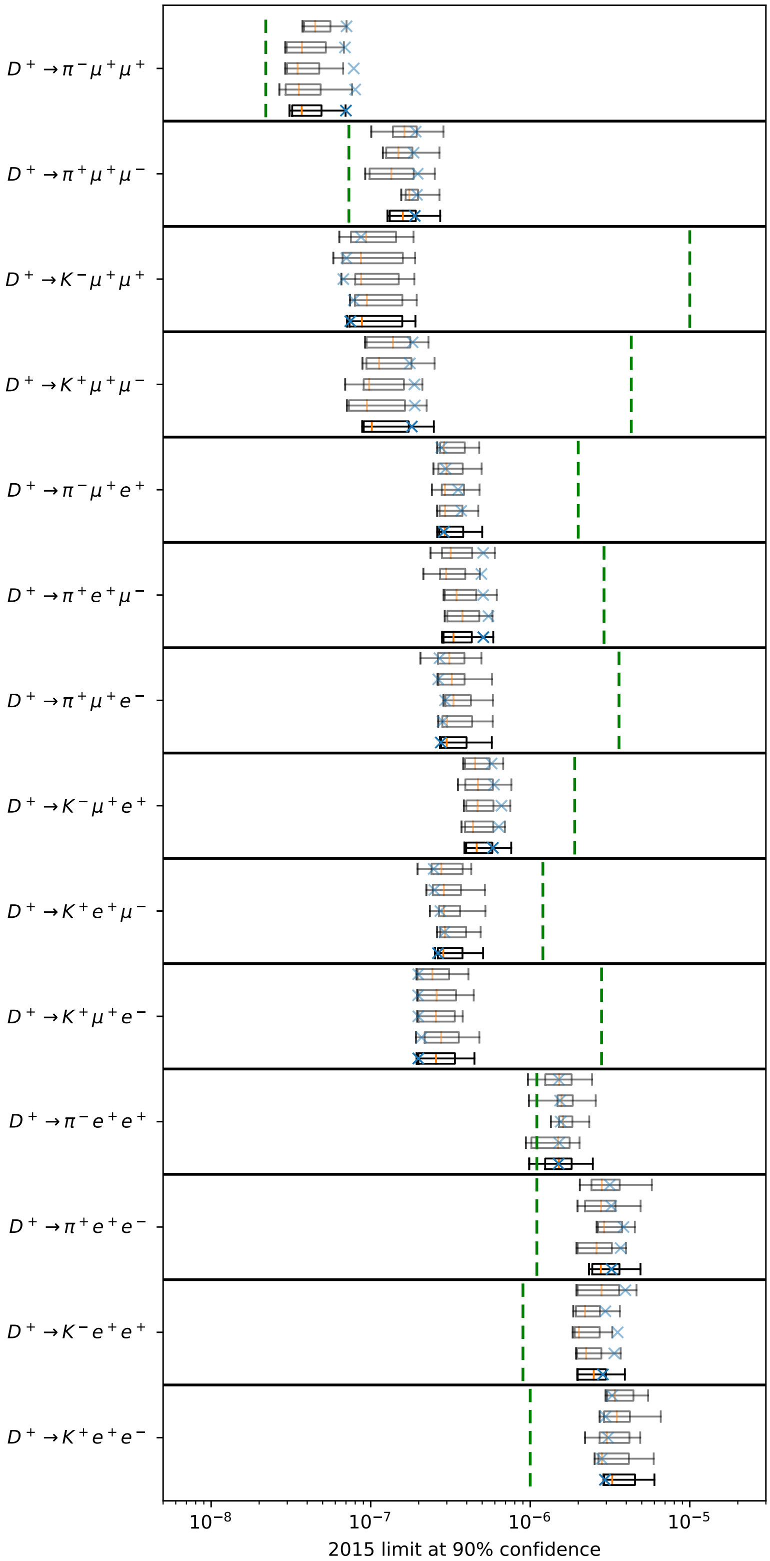


Figure 6.10: 90 % upper limits on the signal channels using the 2015 data. For each decay channels results are given for the exponential, cubic, quadratic and linear background models as well as the modified convolutions (bottom to top). The median (orange), $\pm 1\sigma$ and $\pm 2\sigma$ expected limits are shown as box plots and the observed limit is given by a blue cross. The alternative background models only use 100 toys for each point. The green line shows the prior world's best limit for each channel. As discussed in the text as a consequence of approximations made for this data sample these results are cross-checks only and not final results of the analysis.

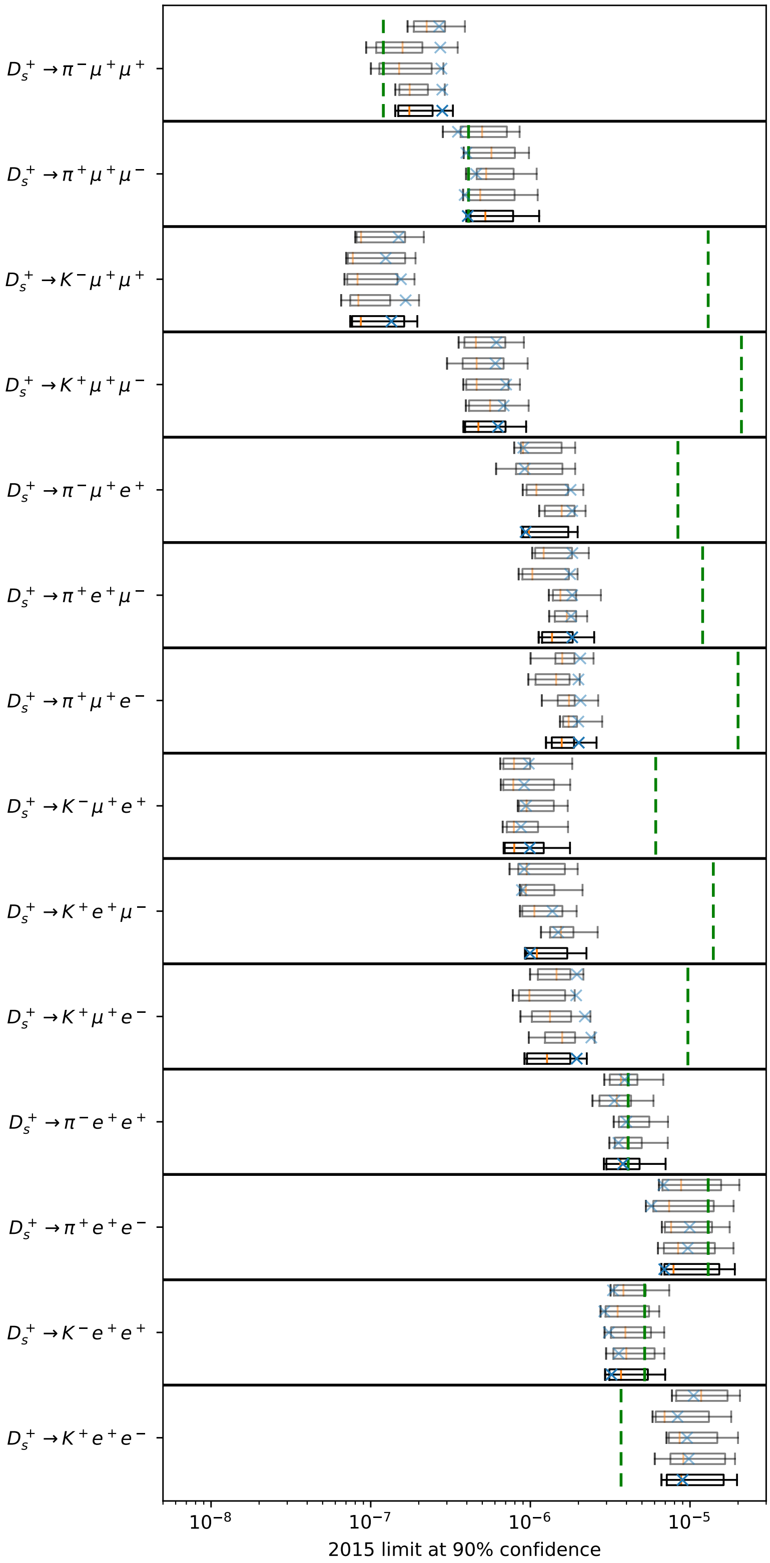


Figure 6.10: 90 % upper limits on the signal channels using the 2015 data (continued)

Channel	stat_0	stat_1	stat_2	stat_3	stat_4	syst_1
$D^+ \to K^+ e^+ e^-$	0.06σ	-0.31σ	0.32σ	-0.74σ	-0.62σ	-0.43σ
$D^+ \to K^- e^+ e^+$	0.07σ	-0.33σ	-0.50σ	-0.13σ	0.20σ	0.00σ
$D^+ \to K^+ e^+ \mu^-$	1.12σ	0.21σ	1.27σ	0.93σ	-0.38σ	1.80σ
$D^+ \to K^+ \mu^+ e^-$	0.34σ	-0.04σ	1.01σ	0.86σ	1.92σ	1.20σ
$D^+ \to K^- \mu^+ e^+$	0.50σ	0.93σ	-0.29σ	0.48σ	-0.54σ	0.55σ
$D^+ \to K^+ \mu^+ \mu^-$	0.60σ	0.45σ	-0.27σ	0.15σ	-0.29σ	0.24σ
$D^+ \to K^- \mu^+ \mu^+$	-0.28σ	-0.16σ	0.10σ	0.25σ	-0.18σ	0.03σ
$D^+ \to \pi^+ e^+ e^-$	0.08σ	-0.17σ	0.21σ	-0.50σ	-0.66σ	-0.25σ
$D^+ \to \pi^- e^+ e^+$	0.38σ	0.10σ	-0.38σ	0.06σ	0.19σ	-0.38σ
$D^+ \to \pi^+ e^+ \mu^-$	0.80σ	-0.78σ	-0.63σ	1.00σ	0.89σ	0.11σ
$D^+ \to \pi^+ \mu^+ e^-$	1.77σ	1.14σ	1.14σ	0.45σ	2.37σ	1.89σ
$D^+ \to \pi^- \mu^+ e^+$	-0.02σ	-1.61σ	1.16σ	-0.44σ	0.21σ	0.48σ
$D^+ \to \pi^+ \mu^+ \mu^-$	0.14σ	0.14σ	0.33σ	0.22σ	0.24σ	0.33σ
$D^+ \to \pi^- \mu^+ \mu^+$	-0.26σ	-0.15σ	-0.06σ	-0.02σ	-0.11σ	0.16σ
$D_s^+ \to K^+ e^+ e^-$	0.31σ	-0.30σ	0.18σ	-0.12σ	0.06σ	0.00σ
$D_s^+ \to K^- e^+ e^+$	0.63σ	-0.04σ	0.67σ	-0.36σ	-0.08σ	0.00σ
$D_s^+ \to K^+ e^+ \mu^-$	0.30σ	-0.10σ	1.13σ	0.72σ	0.96σ	1.53σ
$D_s^+ \to K^+ \mu^+ e^-$	-0.26σ	0.15σ	-0.95σ	-0.64σ	-1.01σ	-0.09σ
$D_s^+ \to K^- \mu^+ e^+$	1.36σ	1.87σ	1.53σ	0.48σ	0.51σ	1.02σ
$D_s^+ \to K^+ \mu^+ \mu^-$	0.03σ	0.07σ	-0.07σ	-0.20σ	-0.13σ	-0.23σ
$D_s^+ \to K^- \mu^+ \mu^+$	-0.04σ	0.16σ	0.25σ	-0.18σ	-0.31σ	-0.39σ
$D_s^+ \to \pi^+ e^+ e^-$	-0.05σ	-0.46σ	-0.18σ	0.00σ	-0.09σ	0.19σ
$D_s^+ \to \pi^- e^+ e^+$	-0.15σ	0.45σ	-0.45σ	-0.00σ	0.19σ	0.34σ
$D_s^+ \to \pi^+ e^+ \mu^-$	1.98σ	1.02σ	2.29σ	1.18σ	1.21σ	1.45σ
$D_s^+ \to \pi^+ \mu^+ e^-$	-0.16σ	-0.32σ	-0.50σ	-1.25σ	-1.59σ	-0.73σ
$D_s^+ \to \pi^- \mu^+ e^+$	0.59σ	0.34σ	0.36σ	0.70σ	0.59σ	0.03σ
$D_s^+ \to \pi^+ \mu^+ \mu^-$	0.10σ	0.21σ	0.08σ	0.03σ	0.10σ	-0.05σ
$D_s^+ \to \pi^- \mu^+ \mu^+$	-0.16σ	-0.27σ	-0.02σ	0.07σ	0.07σ	-0.23σ

Table 6.7: Relative change in the final selection efficiency for 2016 data when using the variations of the particle identification variables, given in units of the statistical uncertainty of the finite MC statistics. Values greater than one are shown in red and 29 out of 168 values exceed an absolute value of 1 sigma, as expected this is less than would be obtained for uncorrelated values.

Channel	$\epsilon_{\text{electron}}$	Norm BR	Norm MC statistics	Norm yield	Reweighting	Signal MC statistics	Signal shape	Tracking
$D^+ \to K^+e^+e^-$	9.1 %	7.7 %	1.1 %	6.4 %	-	11.3%	6.2 %	3.2 %
$D^+ \to K^-e^+e^+$	9.1 %	7.7 %	1.1 %	6.4 %	-	4.6 %	6.2 %	3.2 %
$D^+ \to K^+e^+\mu^-$	4.6 %	7.6 %	1.1 %	6.4 %	7.6 %	2.1 %	6.2 %	2.5 %
$D^+ \to K^+\mu^+e^-$	4.6 %	7.7 %	1.1 %	6.4 %	7.6 %	2.1 %	6.2 %	2.5 %
$D^+ \to K^-\mu^+e^+$	4.6 %	7.6 %	1.1 %	6.4 %	7.6 %	2.3 %	6.2 %	2.5 %
$D^+ \to K^+\mu^+\mu^-$	-	7.6 %	1.1 %	6.4 %	-	3.4 %	1.3 %	2.3 %
$D^+ \to K^-\mu^+\mu^+$	-	7.6 %	1.1 %	6.4 %	-	1.7 %	1.3 %	2.3 %
$D^+ \to \pi^+e^+e^-$	9.1 %	7.5 %	1.1 %	6.4 %	-	6.3 %	6.2 %	2.3 %
$D^+ \to \pi^-e^+e^+$	9.1 %	7.6 %	1.1 %	6.4 %	-	4.2 %	6.2 %	2.3 %
$D^+ \to \pi^+e^+\mu^-$	4.6 %	7.6 %	1.1 %	6.4 %	7.6 %	1.9 %	6.2 %	1.1 %
$D^+ \to \pi^+\mu^+e^-$	4.6 %	7.6 %	1.1 %	6.4 %	7.6 %	1.7 %	6.2 %	1.1 %
$D^+ \to \pi^-\mu^+e^+$	4.6 %	7.6 %	1.1 %	6.4 %	7.6 %	1.7 %	6.2 %	1.1 %
$D^+ \to \pi^+\mu^+\mu^-$	-	7.6 %	1.1 %	6.4 %	-	2.4 %	1.3 %	0.1 %
$D^+ \to \pi^-\mu^+\mu^+$	-	7.6 %	1.1 %	6.4 %	-	1.6 %	1.3 %	0.0 %

Channel	$\epsilon_{\text{electron}}$	Norm BR	Norm MC statistics	Norm yield	Reweighting	Signal MC statistics	Signal shape	Tracking
$D_s^+ \to K^+e^+e^-$	7.5 %	7.6 %	1.1 %	3.7 %	-	8.9 %	6.2 %	3.2 %
$D_s^+ \to K^-e^+e^+$	7.5 %	7.6 %	1.1 %	3.7 %	-	4.5 %	6.2 %	3.2 %
$D_s^+ \to K^+e^+\mu^-$	3.8 %	7.6 %	1.1 %	3.7 %	7.6 %	2.8 %	6.2 %	2.5 %
$D_s^+ \to K^+\mu^+e^-$	3.8 %	7.6 %	1.1 %	3.7 %	7.6 %	3.2 %	6.2 %	2.5 %
$D_s^+ \to K^-\mu^+e^+$	3.8 %	7.6 %	1.1 %	3.7 %	7.6 %	3.4 %	6.2 %	2.5 %
$D_s^+ \to K^+\mu^+\mu^-$	-	7.5 %	1.1 %	3.7 %	-	3.9 %	1.3 %	2.3 %
$D_s^+ \to K^-\mu^+\mu^+$	-	7.6 %	1.1 %	3.7 %	-	2.2 %	1.3 %	2.3 %
$D_s^+ \to \pi^+e^+e^-$	7.5 %	7.5 %	1.1 %	3.7 %	-	5.9 %	6.2 %	2.3 %
$D_s^+ \to \pi^-e^+e^+$	7.5 %	7.5 %	1.1 %	3.7 %	-	4.1 %	6.2 %	2.3 %
$D_s^+ \to \pi^+e^+\mu^-$	3.8 %	7.6 %	1.1 %	3.7 %	7.6 %	2.5 %	6.2 %	1.2 %
$D_s^+ \to \pi^+\mu^+e^-$	3.8 %	7.6 %	1.1 %	3.7 %	7.6 %	2.5 %	6.2 %	1.1 %
$D_s^+ \to \pi^-\mu^+e^+$	3.8 %	7.6 %	1.1 %	3.7 %	7.6 %	2.5 %	6.2 %	1.2 %
$D_s^+ \to \pi^+\mu^+\mu^-$	-	7.5 %	1.1 %	3.7 %	-	3.3 %	1.3 %	0.1 %
$D_s^+ \to \pi^-\mu^+\mu^+$	-	7.6 %	1.1 %	3.7 %	-	2.3 %	1.3 %	0.0 %

Table 6.8: Summary of systematic uncertainties used for the cross-check measurement with the 2015 data sample. All values are given as a fractional uncertainty on the signal yield.

Channel	$\epsilon_{\text{electron}}$	Norm BR	Norm MC statistics	Norm yield	Reweighting	Signal MC statistics	Signal shape	Tracking
$D^+ \to K^+e^+e^-$	9.1 %	7.6 %	1.0 %	2.0 %	-	6.4 %	7.0 %	3.2 %
$D^+ \to K^-e^+e^+$	9.1 %	7.6 %	1.0 %	2.0 %	-	2.7 %	7.0 %	3.2 %
$D^+ \to K^+e^+\mu^-$	4.6 %	7.6 %	1.0 %	2.0 %	7.6 %	2.1 %	7.0 %	2.5 %
$D^+ \to K^+\mu^+e^-$	4.6 %	7.6 %	1.0 %	2.0 %	7.6 %	1.8 %	7.0 %	2.5 %
$D^+ \to K^-\mu^+e^+$	4.6 %	7.6 %	1.0 %	2.0 %	7.6 %	1.7 %	7.0 %	2.5 %
$D^+ \to K^+\mu^+\mu^-$	-	7.6 %	1.0 %	2.0 %	-	3.0 %	0.8 %	2.3 %
$D^+ \to K^-\mu^+\mu^+$	-	7.6 %	1.0 %	2.0 %	-	1.4 %	0.8 %	2.3 %
$D^+ \to \pi^+e^+e^-$	9.1 %	7.6 %	1.0 %	2.0 %	-	3.9 %	7.0 %	2.3 %
$D^+ \to \pi^-e^+e^+$	9.1 %	7.7 %	1.0 %	2.0 %	-	2.6 %	7.0 %	2.3 %
$D^+ \to \pi^+e^+\mu^-$	4.6 %	7.6 %	1.0 %	2.0 %	7.6 %	1.9 %	7.0 %	1.1 %
$D^+ \to \pi^+\mu^+e^-$	4.6 %	7.6 %	1.0 %	2.0 %	7.6 %	1.5 %	7.0 %	1.1 %
$D^+ \to \pi^-\mu^+e^+$	4.6 %	7.6 %	1.0 %	2.0 %	7.6 %	1.6 %	7.0 %	1.1 %
$D^+ \to \pi^+\mu^+\mu^-$	-	7.6 %	1.0 %	2.0 %	-	1.9 %	0.8 %	0.1 %
$D^+ \to \pi^-\mu^+\mu^+$	-	7.6 %	1.0 %	2.0 %	-	1.5 %	0.8 %	0.0 %

Channel	$\epsilon_{\text{electron}}$	Norm BR	Norm MC statistics	Norm yield	Reweighting	Signal MC statistics	Signal shape	Tracking
$D_s^+ \to K^+e^+e^-$	7.5 %	7.6 %	1.0 %	3.0 %	-	6.0 %	7.0 %	3.2 %
$D_s^+ \to K^-e^+e^+$	7.5 %	7.6 %	1.0 %	3.0 %	-	3.2 %	7.0 %	3.2 %
$D_s^+ \to K^+e^+\mu^-$	3.8 %	7.6 %	1.0 %	3.0 %	7.6 %	3.0 %	7.0 %	2.5 %
$D_s^+ \to K^+\mu^+e^-$	3.8 %	7.6 %	1.0 %	3.0 %	7.6 %	2.5 %	7.0 %	2.5 %
$D_s^+ \to K^-\mu^+e^+$	3.8 %	7.6 %	1.0 %	3.0 %	7.6 %	2.4 %	7.0 %	2.5 %
$D_s^+ \to K^+\mu^+\mu^-$	-	7.6 %	1.0 %	3.0 %	-	3.3 %	0.8 %	2.3 %
$D_s^+ \to K^-\mu^+\mu^+$	-	7.6 %	1.0 %	3.0 %	-	1.8 %	0.8 %	2.3 %
$D_s^+ \to \pi^+e^+e^-$	7.5 %	7.6 %	1.0 %	3.0 %	-	4.3 %	7.0 %	2.3 %
$D_s^+ \to \pi^-e^+e^+$	7.5 %	7.6 %	1.0 %	3.0 %	-	3.1 %	7.0 %	2.3 %
$D_s^+ \to \pi^+e^+\mu^-$	3.8 %	7.6 %	1.0 %	3.0 %	7.6 %	2.7 %	7.0 %	1.2 %
$D_s^+ \to \pi^+\mu^+e^-$	3.8 %	7.6 %	1.0 %	3.0 %	7.6 %	2.3 %	7.0 %	1.1 %
$D_s^+ \to \pi^-\mu^+e^+$	3.8 %	7.6 %	1.0 %	3.0 %	7.6 %	2.2 %	7.0 %	1.2 %
$D_s^+ \to \pi^+\mu^+\mu^-$	-	7.6 %	1.0 %	3.0 %	-	2.5 %	0.8 %	0.1 %
$D_s^+ \to \pi^-\mu^+\mu^+$	-	7.6 %	1.0 %	3.0 %	-	2.2 %	0.8 %	0.0 %

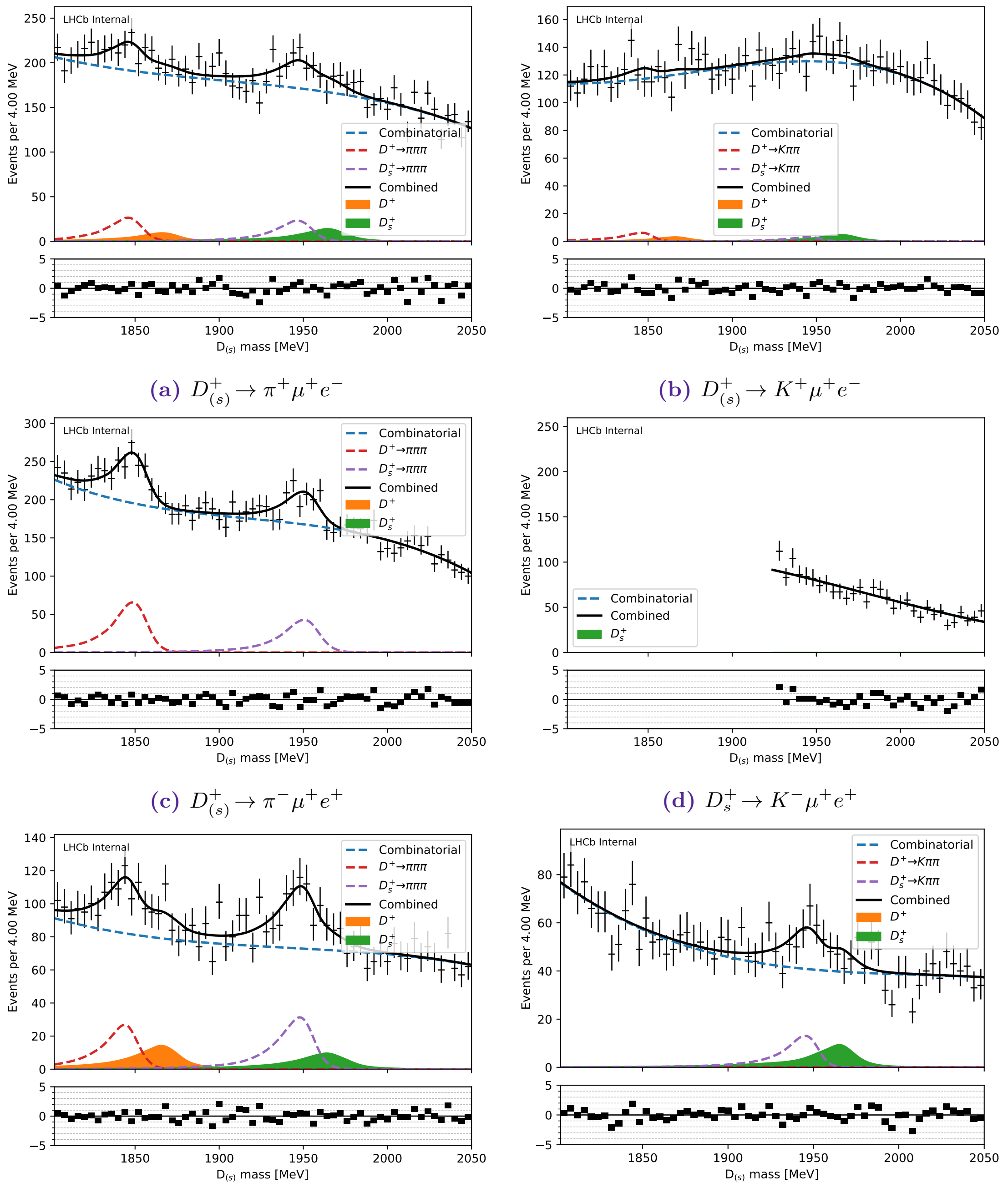
Table 6.9: Summary of systematic uncertainties for the 2016 dataset. All values are given as a fractional uncertainty on the signal yield.

Year	Channel	PDG BR	Fit BR	Difference [%]	σ
2015	$D^+ \to \pi^+\mu^+\mu^-$	$(1.63 \pm 0.12) \times 10^{-6}$	$(1.66 \pm 0.15) \times 10^{-6}$	1.85±9.12	0.2
2015	$D^+ \to \pi^+e^+e^-$	$(1.67 \pm 0.06) \times 10^{-6}$	$(1.80 \pm 0.54) \times 10^{-6}$	7.62±32.07	0.2
2015	$D_s^+ \to \pi^+\mu^+\mu^-$	$(1.33 \pm 0.10) \times 10^{-5}$	$(1.50 \pm 0.11) \times 10^{-5}$	12.57±8.31	1.5
2015	$D_s^+ \to \pi^+e^+e^-$	$(1.37 \pm 0.05) \times 10^{-5}$	$(1.70 \pm 0.32) \times 10^{-5}$	24.23±22.98	1.1
2016	$D^+ \to \pi^+\mu^+\mu^-$	$(1.63 \pm 0.12) \times 10^{-6}$	$(1.60 \pm 0.07) \times 10^{-6}$	-1.75±4.03	-0.4
2016	$D^+ \to \pi^+e^+e^-$	$(1.67 \pm 0.06) \times 10^{-6}$	$(1.37 \pm 0.28) \times 10^{-6}$	-17.95±17.03	-1.1
2016	$D_s^+ \to \pi^+\mu^+\mu^-$	$(1.33 \pm 0.10) \times 10^{-5}$	$(1.35 \pm 0.07) \times 10^{-5}$	1.69±4.99	0.3
2016	$D_s^+ \to \pi^+e^+e^-$	$(1.37 \pm 0.05) \times 10^{-5}$	$(1.36 \pm 0.25) \times 10^{-5}$	-1.05±17.95	-0.1

Table 6.10: Measured values of the ϕ resonance branching ratio with the full selection, measured relative to $D_{(s)}^+ \to (\phi\to \mu^-\mu^+)\ \pi^+$ with a loose selection applied. As explained in the text, these values do not include all uncertainties and serve only as a cross-check of the analysis.

Channel	Expected 2015	Observed 2015	World's best	Scaled 2015	Improvement	Expected 2016	Improvement
$D^+ \to K^+e^+e^-$	3.5×10^{-6}	2.9×10^{-6}	1.0×10^{-6}	1.5×10^{-6}	0.8 - 0.5	-	-
$D^+ \to K^-e^+e^+$	2.4×10^{-6}	2.8×10^{-6}	9.0×10^{-7}	1.1×10^{-6}	0.9 - 0.7	-	-
$D^+ \to K^+e^+\mu^-$	2.9×10^{-7}	2.6×10^{-7}	1.2×10^{-6}	1.3×10^{-7}	10.4 - 7.0	-	-
$D^+ \to K^+\mu^+e^-$	2.6×10^{-7}	2.0×10^{-7}	2.8×10^{-6}	1.2×10^{-7}	30.7 - 20.2	-	-
$D^+ \to K^-\mu^+e^+$	4.3×10^{-7}	5.6×10^{-7}	1.9×10^{-6}	1.9×10^{-7}	11.0 - 7.7	-	-
$D^+ \to K^+\mu^+\mu^-$	1.1×10^{-7}	1.8×10^{-7}	4.3×10^{-6}	5.0×10^{-8}	114.0 - 53.5	4.1×10^{-8}	125.9 - 77.9
$D^+ \to K^-\mu^+\mu^+$	8.7×10^{-8}	7.7×10^{-8}	1.0×10^{-5}	3.9×10^{-8}	290.3 - 144.7	1.8×10^{-8}	583.7 - 502.1
$D^+ \to \pi^+e^+e^-$	2.8×10^{-6}	3.3×10^{-6}	1.1×10^{-6}	1.2×10^{-6}	1.0 - 0.7	-	-
$D^+ \to \pi^-e^+e^+$	1.5×10^{-6}	1.5×10^{-6}	1.1×10^{-6}	6.9×10^{-7}	1.9 - 1.3	-	-
$D^+ \to \pi^+e^+\mu^-$	3.5×10^{-7}	5.2×10^{-7}	2.9×10^{-6}	1.5×10^{-7}	22.4 - 14.8	-	-
$D^+ \to \pi^+\mu^+e^-$	3.0×10^{-7}	2.8×10^{-7}	3.6×10^{-6}	1.4×10^{-7}	28.2 - 19.3	-	-
$D^+ \to \pi^-\mu^+e^+$	2.9×10^{-7}	2.9×10^{-7}	2.0×10^{-6}	1.3×10^{-7}	16.2 - 11.2	-	-
$D^+ \to \pi^+\mu^+\mu^-$	1.6×10^{-7}	1.9×10^{-7}	7.3×10^{-8}	7.1×10^{-8}	1.5 - 0.8	4.2×10^{-8}	1.9 - 1.3
$D^+ \to \pi^-\mu^+\mu^+$	3.6×10^{-8}	6.9×10^{-8}	2.2×10^{-8}	1.6×10^{-8}	1.7 - 1.0	1.4×10^{-8}	2.0 - 1.2
$D_s^+ \to K^+e^+e^-$	9.5×10^{-6}	9.3×10^{-6}	1.0×10^{-6}	4.2×10^{-6}	0.3 - 0.1	-	-
$D_s^+ \to K^-e^+e^+$	3.7×10^{-6}	3.2×10^{-6}	9.0×10^{-7}	1.7×10^{-6}	0.7 - 0.4	-	-
$D_s^+ \to K^+e^+\mu^-$	1.1×10^{-6}	1.1×10^{-6}	1.2×10^{-6}	5.0×10^{-7}	3.0 - 1.5	-	-
$D_s^+ \to K^+\mu^+e^-$	1.3×10^{-6}	2.0×10^{-6}	2.8×10^{-6}	5.8×10^{-7}	6.7 - 3.6	-	-
$D_s^+ \to K^-\mu^+e^+$	8.1×10^{-7}	10.0×10^{-7}	1.9×10^{-6}	3.6×10^{-7}	6.1 - 3.1	-	-
$D_s^+ \to K^+\mu^+\mu^-$	4.8×10^{-7}	6.2×10^{-7}	4.3×10^{-6}	2.1×10^{-7}	24.6 - 13.9	1.7×10^{-7}	27.7 - 20.2
$D_s^+ \to K^-\mu^+\mu^+$	8.5×10^{-8}	1.3×10^{-7}	1.0×10^{-5}	3.8×10^{-8}	295.4 - 140.9	3.0×10^{-8}	369.0 - 247.3
$D_s^+ \to \pi^+e^+e^-$	8.2×10^{-6}	6.8×10^{-6}	1.1×10^{-6}	3.6×10^{-6}	0.4 - 0.2	-	-
$D_s^+ \to \pi^-e^+e^+$	3.6×10^{-6}	3.8×10^{-6}	1.1×10^{-6}	1.6×10^{-6}	0.8 - 0.5	-	-
$D_s^+ \to \pi^+e^+\mu^-$	1.5×10^{-6}	1.8×10^{-6}	2.9×10^{-6}	6.9×10^{-7}	5.3 - 3.4	-	-
$D_s^+ \to \pi^+\mu^+e^-$	1.6×10^{-6}	2.0×10^{-6}	3.6×10^{-6}	7.1×10^{-7}	6.2 - 4.3	-	-
$D_s^+ \to \pi^-\mu^+e^+$	1.2×10^{-6}	9.6×10^{-7}	2.0×10^{-6}	5.4×10^{-7}	4.8 - 2.5	-	-
$D_s^+ \to \pi^+\mu^+\mu^-$	5.1×10^{-7}	4.4×10^{-7}	7.3×10^{-8}	2.3×10^{-7}	0.4 - 0.2	1.8×10^{-7}	0.5 - 0.3
$D_s^+ \to \pi^-\mu^+\mu^+$	1.8×10^{-7}	2.8×10^{-7}	2.2×10^{-8}	7.8×10^{-8}	0.3 - 0.2	4.9×10^{-8}	0.5 - 0.3

Table 6.11: Estimated sensitivity of the 2016 dataset for decays of a $D_{(s)}^+$ meson. The first estimate ("Scaled 2015") is based on scaling the expected limit from Section 6.4 using the 2015 dataset by the difference in luminosity ($\sqrt{5}$). The second estimate ("Expected 2016") is based on the expected limit based on the fits to the blinded sidebands for 2016 data. The estimated limit and improvement factor are shown in green when the ±1σ expected limit is an improvement on the current world's best limit. All values are given at 90 % confidence. As discussed in the text as a consequence of approximations made for this data sample these results are cross-checks only and not final results of the analysis.



(a) $D^+_{(s)} \to \pi^+\mu^+e^-$

(b) $D^+_{(s)} \to K^+\mu^+e^-$

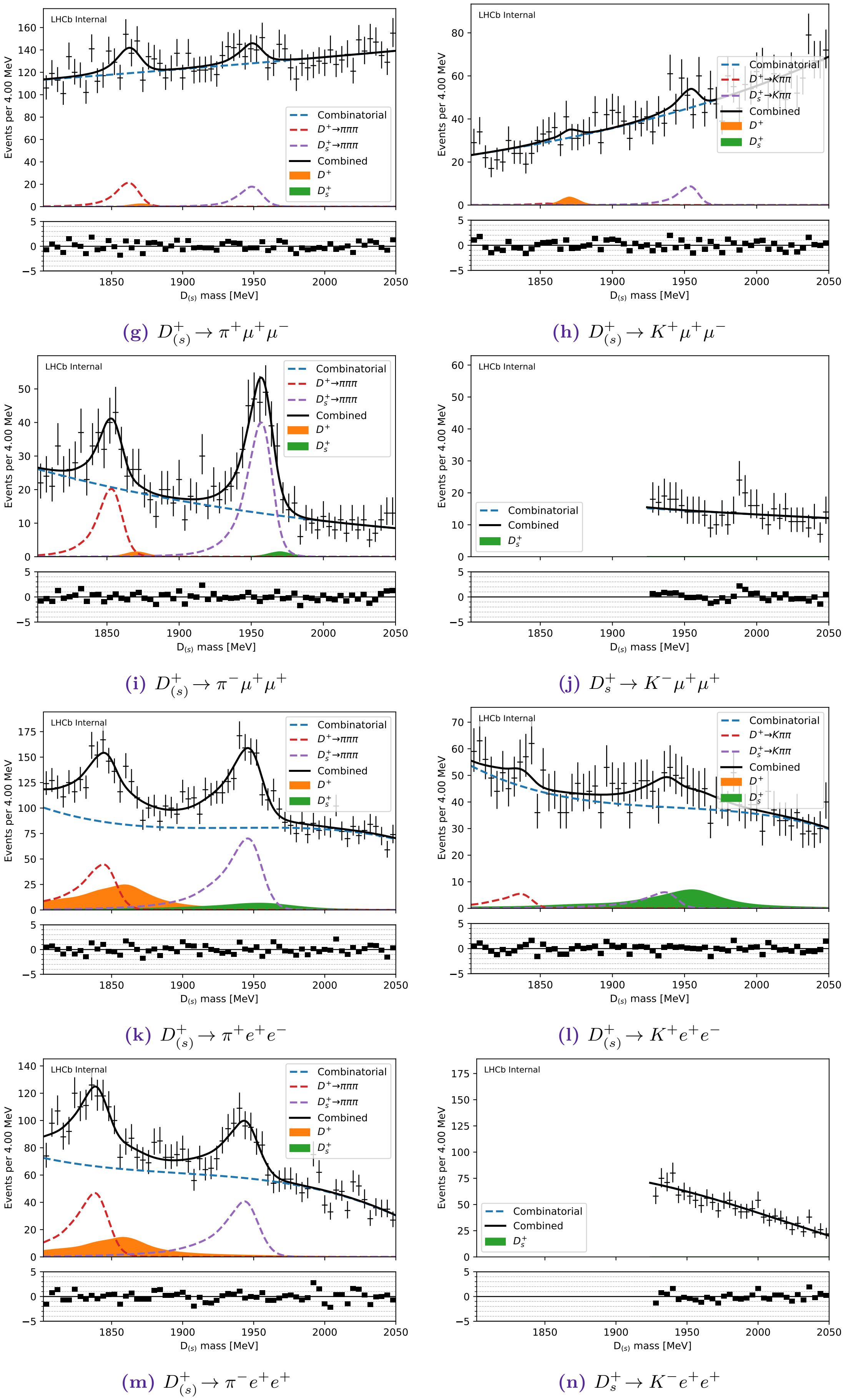
(c) $D^+_{(s)} \to \pi^-\mu^+e^+$

(d) $D^+_s \to K^-\mu^+e^+$

(e) $D^+_{(s)} \to \pi^+e^+\mu^-$

(f) $D^+_{(s)} \to K^+e^+\mu^-$

Figure 6.11: Fitted $D^+_{(s)}$ invariant mass distributions in 2016 data that are used for the final result. The regions corresponding to $D^+ \to K^-l^+l'^+$ are not available at the time of submission and are omitted.



(g) $D^+_{(s)} \to \pi^+\mu^+\mu^-$

(h) $D^+_{(s)} \to K^+\mu^+\mu^-$

(i) $D^+_{(s)} \to \pi^-\mu^+\mu^+$

(j) $D^+_s \to K^-\mu^+\mu^+$

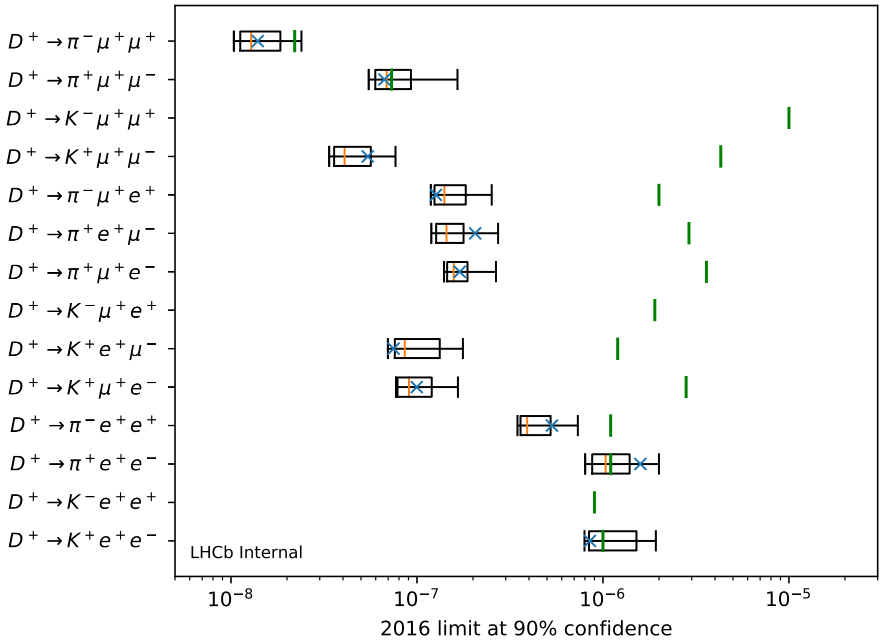
(k) $D^+_{(s)} \to \pi^+e^+e^-$

(l) $D^+_{(s)} \to K^+e^+e^-$

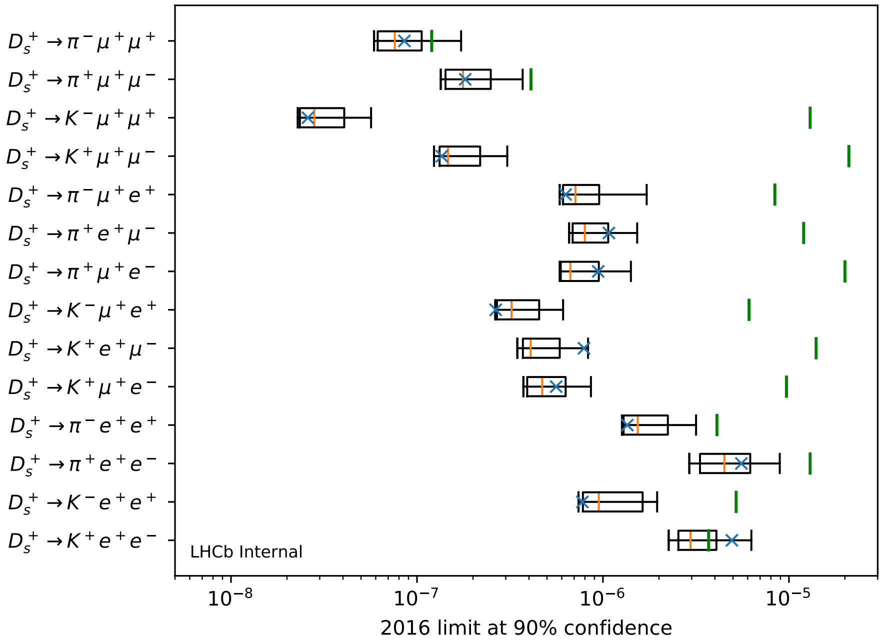
(m) $D^+_{(s)} \to \pi^-e^+e^+$

(n) $D^+_s \to K^-e^+e^+$

Figure 6.11: Fitted $D^+_{(s)}$ invariant mass distributions for the 2016 dataset (continued)



(a) D^+



(b) D_s^+

Figure 6.12: 90 % upper limits on the $D^+_{(s)}$ signal channels using the 2016 data. The median (orange), $\pm1\sigma$ and $\pm2\sigma$ expected limits are shown as box plots and the observed limit is given by a blue cross. The semi-transparent and dotted lines show the limit when calculated without systematic uncertainties. The green line shows the prior world's best limit for each channel.

Decay	D^+		D_s^+		Improvement	
	$90\,\%$ $[\times 10^{-9}]$	$95\,\%$ $[\times 10^{-9}]$	$90\,\%$ $[\times 10^{-9}]$	$95\,\%$ $[\times 10^{-9}]$	D^+	D_s^+
$D^+_{(s)} \to \pi^+\mu^+\mu^-$	67	74	180	210	1.1	2.3
$D^+_{(s)} \to \pi^-\mu^+\mu^+$	14	16	86	96	1.6	1.4
$D^+_{(s)} \to K^+\mu^+\mu^-$	54	61	140	160	79.0	150.0
$D^+_{(s)} \to K^-\mu^+\mu^+$	-	-	26	30	-	500.0
$D^+_{(s)} \to \pi^+e^+\mu^-$	210	230	1100	1200	14.0	11.0
$D^+_{(s)} \to \pi^+\mu^+e^-$	220	220	940	1100	16.0	21.0
$D^+_{(s)} \to \pi^-\mu^+e^+$	130	150	630	710	16.0	13.0
$D^+_{(s)} \to K^+e^+\mu^-$	75	83	790	880	16.0	18.0
$D^+_{(s)} \to K^+\mu^+e^-$	100	110	560	640	28.0	17.0
$D^+_{(s)} \to K^-\mu^+e^+$	-	-	260	320	-	23.0
$D^+_{(s)} \to \pi^+e^+e^-$	1600	1800	5500	6400	0.7	2.3
$D^+_{(s)} \to \pi^-e^+e^+$	530	600	1400	1600	2.1	3.0
$D^+_{(s)} \to K^+e^+e^-$	850	1000	4900	5500	1.2	0.8
$D^+_{(s)} \to K^-e^+e^+$	-	-	770	840	-	6.7

Table 6.12: Upper limits obtained from $1.5\,\text{fb}^{-1}$ of LHCb data using the CL_s method with improvement by the given factor relative to the previous best results.

6.5 Results and conclusions

A search has been made of 25 previously unobserved three body semileptonic decays of a $D^+_{(s)}$ meson using $1.5\,\text{fb}^{-1}$ of data that was collected by the LHCb experiment during 2016. The results for the three channels of the form $D^+ \to K^- l^+ l'^+$ are unavailable at the time of thesis submission. Fits to the three body invariant mass distribution are shown in Figure 6.11. No significant deviations from the background only hypothesis are seen and world's best limits have been obtained for 23 of these decays as shown in Figure 6.12 and Table 6.12. The two channels where the results are not better than the previous determinations contain a kaon and two electrons.

Blank page

Chapter 7

Tools for analysis

While the work documented in this thesis has been performed efforts have been made to improve the ecosystem for HEP data analysis. This has included many small contributions to open source projects, such as: adding support for the `XRootD` protocol in `Snakemake`[200], maintaining the popular `root_pandas` package as a member of the `scikit-hep` organisation[201] and packaging `ROOT`,[1] `XRootD` and `GEANT4` for the conda[202] package manager.

This chapter documents some of the larger contributions that have been made to the analysis ecosystem. Section 7.1 discusses work performed on analysis preservation that can benefit both the collaboration and the wider public. Section 7.2 shows an evaluation of the `nix` package manager[203] that was performed by the author with the view to improving the flexibility and reliability of software distribution. Section 7.3 describes the extension of the LHCb working group production system as a replacement for distributed computing activities that are currently performed by individual analysts. Section 7.4 discusses the LHCb Starterkit initiative that provides software training for new members of the LHCb collaboration. Finally, Section 7.5 summarises work published in a paper[204] on a hypothesis testing technique known as the *energy test*.

7.1 Analysis preservation

It is becoming increasingly difficult to reproduce results in high energy physics due to the high cost of modern experiments and the large number of people involved.[205] To counteract this, funding bodies and journals are requiring that data, and the tools required to process it, are archived as part of the publication procedure to ensure the maximal benefit can be obtained from these publicly funded datasets.[206] As part of this, the four main LHC experiments have policies to release their data to the public domain after a fixed a period of time.[207–210] Furthermore, the software that is developed by HEP collaborations is highly specialised and contains a vast quantity of expert knowledge. This software must also be released to allow public data to be useful to researchers from outside the original collaboration. To allow for this, the source code for the experiment's software stacks is being released under a variety of open source licenses.[211–214] As a

[1] This has been downloaded over 10 000 times in the first three months after it became available.

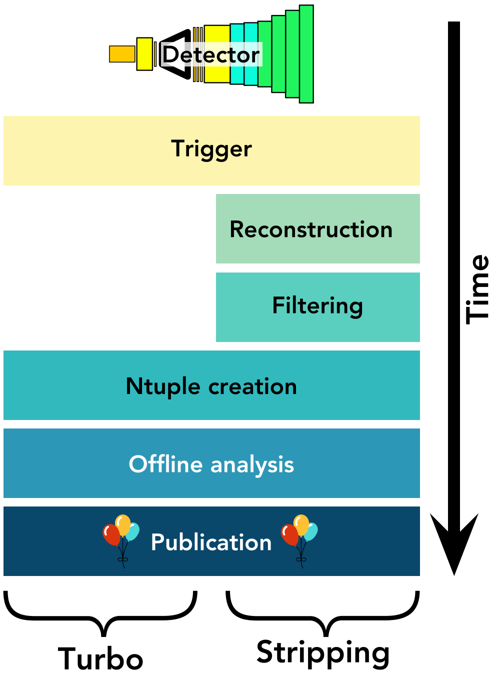


Figure 7.1: High level overview of the steps involved in preparing LHCb data for publication. The stages written in black text are normally performed using the worldwide LHC computing grid and the stages written in white are typically performed using other computing resources. The left (right) side of the diagram shows the steps involved when using turbo (stripped) data, as explained in Section 2.11.

result of these efforts, a peer reviewed publication has already been made by a group of external researchers[2] using CMS open data[215], providing a hint as to the potential value of these releases. The datasets have also been used for other purposes, including tutorials, examples and benchmarking of software.[216, 217] Being able to use real LHC data for these purposes can be both easier and more engaging than generating realistic toy samples. Additionally, having open data polices could allow independent researchers to try and resolve tensions between results. One example of a discrepancy is that of the Ξ_{cc}^{+} baryon where the results of SELEX[218, 219] are in disagreement with those of other experiments.[220–223]

Even if data is not released publicly, mandating the archival of all stages of each analysis provides benefits to collaborations. A large proportion of analysis is performed by young researchers, the majority of whom will leave academic research after completing a PhD. This can also result in data that was previously in their personal storage areas being lost when their accounts become inactive.[224] Being able to exactly reproduce the steps that were taken during an analysis has several other benefits:

- **Assessing the impact of bugs:** Due to the complexity of modern HEP software it is almost certain that bugs will be present in the code. Most of these will be harmless for published analyses. However, in rare cases it is useful to explicitly check that

[2] Admittedly with assistance from CMS colleagues.

a result is not affected. Without access to the original code it is difficult and time consuming to guarantee that a published result is unaffected.

- **Testing software developments:** Experimental software stacks are rapidly developing to incorporate new improvements. Testing these changes can be difficult due to the complex correlations between different components. When major changes are made, for example updated simulation versions, it would be beneficial to validate the changes by rerunning existing analyses.
- **Reinterpret results:** Some analyses produce results that are dependent on specific new physics models. The choice of models to try is generally guided by the currently available results. However this can rapidly change after or even during an analysis. Having the code available can make it easier to try additional models and a robust pipeline (see Section 7.1.3) can be used in conjunction with platforms such as `REANA`[225] to allow results to be updated relatively easily.[3]
- **Repeat analyses:** Analyses are often updated with larger datasets. Having an archive that can reproduce the original result can simplify this process, especially if the original analysis was written with reuse in mind. This allows subsequent analyses to focus on developing improved methods instead of repeating the same work.
- **Documentation:** Analysis code can serve as documentation of how to implement techniques and use other software tools. This is especially important when the tools used are poorly documented as is often the case.

To allow LHCb to gain the advantages of analysis preservation, the author was involved in writing a roadmap[226] that provides recommendations. This is focused on the *offline analysis* stage of Figure 7.1. This is typically performed outside of the central LHCb software stack over the course of one or more years. The following sections split this into smaller archival problems that, when combined, can allow the entire offline analysis step to be reproduced. The *trigger*, *reconstruction* and *filtering* steps are already effectively preserved within `Dirac` thanks to their centralised nature. See Section 7.3 for a proposal for preserving the *ntuple creation* step.

7.1.1 Storage

The problem of storing data can be split into two separate problems. For input datasets, `EOS`[133] has been developed at CERN for providing access to many petabytes of storage. LHCb is split into *working groups* that are each responsible for a different area of the physics programme. These groups each have storage quotas on `EOS` that can be used for permanently preserving large collections of data.

For storing analysis code it is logical to separate this into a separate system. The files are small, rarely exceeding a few kilobytes, making it practical to store every change that is made. This can be achieved using *version control systems* (VCS) such as `CVS`, `SVN`, `Git`,

[3] In practice a knowledgeable analyst will always have to review the results to ensure that the analysis is still valid. This can however reduce or remove the need to develop code to process the data.

Mercurial and darcs. The current recommendation at CERN is to use Git as part of the GitLab instance. This also has additional benefits such as providing tools for issue tracking, continuous integration and peer review.

7.1.2 Software environments

While being able to access the code that was used to perform an analysis is useful, it is insufficient to unambiguously know what was done without knowledge of the software versions that were used. Many popular packages have rapid development cycles and frequently break backwards compatibility, hence making it difficult to execute code without knowledge of the original versions. Furthermore, having this information easily available makes it possible to check if any analyses are affected when potentially serious bugs are found. The ability to archive a software environment is also useful when developing an analysis as it allows the code to be more easily moved between different computing resources and analysts, without time being wasted trying to replicate the original setup.

Container technologies are helpful here as they can package an arbitrary software environment into a single large executable "blob" that is easily portable between different systems. This model has rapidly risen in popularity in commercial software deployment. Docker[227] is particularly popular thanks to it being freely available and open source. The company behind Docker, Docker Inc., also provides a commercially supported edition and is currently valued at over a billion US dollars.[228] One key limitation however, is that it is currently impossible for users to use Docker without having complete access to the host machine. This makes it unsuitable for the shared computing environments that are commonly used by analysts. Fortunately, the containerisation technology Docker is built upon is based upon open standards making it possible to have alternative implementations. Singularity[229] is one such alternative that has proven popular in high performance computing community and the CMS experiment uses Singularity containers for all production jobs on the grid.

While containers are excellent for preserving environments, software must still be installed inside them. This can be done by compiling packages from source. However, this is time consuming and can require specialised knowledge depending on the robustness of the build system of the given software. It also does not provide an easy way for a user to know what software is installed inside the container. The solution to this is to use a package manager that is able to manage dependencies and ideally provide binaries that can be installed. The Conda[202] package manger is very popular in a wide range of both academic and commercial data analysis fields and many data analysis libraries have it listed as the recommended installation method thanks to its portability between machines. It also provides a simple interface that allows users to run `conda install SOMETHING` with the reasonable expectation that a working installation of SOMETHING will be available soon after. Large and well maintained repositories of up-to-date packages are available for both general purpose[230, 231] and field specific[232, 233] software. This makes conda ideal[4] for

[4] A robust build of ROOT was unavailable until recently and Conda is expected to be officially supported platform by the ROOT developers in the near future.

analysts to reliably manage local installations without sacrificing portability. It has also been used for the production software stack of some smaller experiments[234]. However concerns have previously been raised as to its suitability for the core software needs of the large LHC experiments. Section 7.2 describes an alternative package manager, suggested and evaluated by the author, that is better suited to this use case and is currently being considered.

7.1.3 Pipelines

Once the data and code has been archived it is necessary to know how to execute the code in order to correctly produce the result. This can be achieved using workflow management systems, that control the execution of a sequence of tasks. This is then known as a *workflow* or a *pipeline*. Most systems use a custom syntax to define each stage that is then used to schedule tasks as appropriate. Due to the varying requirements of different applications a wide range of systems have been developed, of which only a few shown here, see Reference [235] for a more complete list. In order to select a system to recommend for the partial reproducibility of LHCb analyses the following requirements were considered:

- **Simple** For a pipeline to be widely adopted it must be easy to implement for users, ideally making use of knowledge that is already required from analysis; such as python or shell scripting.
- **Scriptable** To be used in a system that provides full reproducibility it must be possible for the pipeline to be executed as a script that can then be integrated with continuous integration or larger pipeline, like the CERN `REANA` project[225].
- **Dependency tracking and caching** Pipelines should be aware of changing inputs to the analysis, such as updated ntuples on EOS, and act accordingly. If a stage of an analysis has not changed, then it should not be recomputed on each execution of the pipeline to save computing resources and prevents analysts waiting unnecessarily for small changes.
- **Easy to debug** The use of a pipeline should not prevent the analysis from being executed manually to allow analyses that are in development to be debugged. In addition, errors in the pipeline should be clearly presented.
- **Community** To ease the adoption of a particular system it should be actively developed and have an existing community to support users. Having an existing user base also helps ensure solutions for common problems are easy to find by searching, reducing the barrier to entry.

Many systems have been considered and, for the most promising, pipelines were implemented using a simplified version of the analysis that was used to measure polarisation in $\mathrm{B}_s^0 \to \mathrm{J}/\psi\overline{\mathrm{K}}^{*0}$ decays using Run 1 LHCb data.[236] A summary of how the short-listed

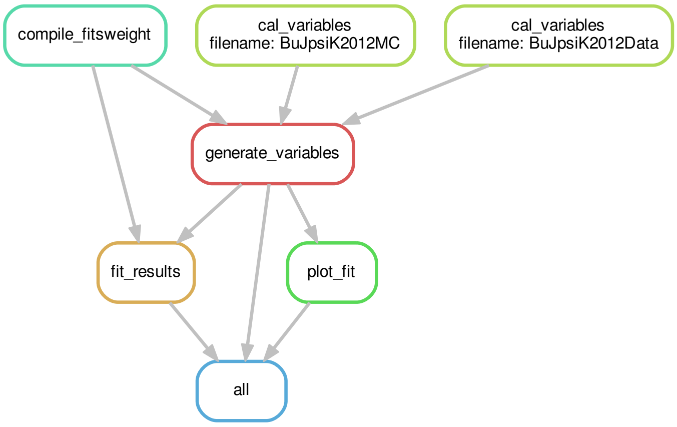


Figure 7.2: Pipeline showing the steps involved in running a simplified version of the $B_s^0 \to J/\psi \overline{K}^{*0}$ analysis.

systems performed relative to the aforementioned requirements is shown in Table 7.1. As a result it was recommended that analysts first consider using `Snakemake`[237].[5]

	Simple	Scriptable	Caching	Debugging	Community
`Bash`	✓	✓	✗	✗	✓
`Make`	✗	✓	✓	✗	✓
`Snakemake`	✓	✓	✓	✓	✓
`Yadage`	✗	✓	✓	✗	✗
`Luigi`	✗	✓	✗	✓	✓
`Fabricate`	✓	✓	✓	✓	✗
`CWLTool`	✗	✓	✓	✓	✗

Table 7.1: Overview of how various workflow management systems meet the requirements that were chosen for LHCb analysis preservation.

`Snakemake`

`Snakemake`[237] is a workflow management system that aims to simplify the process of creating workflows. Workflows are written in `Python` with a small number of additions to the syntax that are inspired by `make`. The system has found widespread use in bioinformatics research and has been cited by over 60 published papers. In the context of LHCb analysis preservation, `Snakemake` also has the following useful features:

- Remote access to files via `XRootD`
- Local and cluster based execution
- Rules can depend on arbitrary resources, such as GPUs or RAM
- Support for importing shared subworkflows, for example to compute PID efficiencies

[5] While `Snakemake` is the recommended system, those who wish to use alternative systems are free to do so and none of the other recommendations presented in the analysis preservation roadmap[226] are dependent on the choice of pipeline system.

Snakemake is now being used for many analyses in LHCb and was used in the search for $D^+_{(s)} \to h^\pm l^+ l'^\mp$ described in Section 5 and 6. In this analysis Snakemake was able to perform the entire analysis, with the input being taken directly from set of working group productions (Section 7.3), using over 12 000 individual steps.

7.1.4 Conclusion

The author and his colleagues on LHCb have produced an analysis preservation roadmap[226] and the main analysis documented in this thesis has made use of these recommendations. As of November 2018 the minimal recommendations are being enforced within LHCb, with all analyses being required to have a Git repository containing all code that is required to reproduce an analysis. These repositories must include written instructions are documenting the steps that can be used to execute it. More complete levels of analysis preservation are being promoted during the software training events discussed in Section 7.4. These aim to encourage more complete preservation for analyses that are still in their early stages.

7.2 Software packaging

Computationally intensive areas of modern research, such as high energy physics, provide unique challenges for software packaging. Software is used at a massive scale for processing large datasets using heterogeneous resources, such as the Woldwide LHC Computing Grid[90]. The simulation of samples and reprocessing of data can continue to use software for decades after the software was originally written. For example, the Large Electron Positron collider (LEP) continued to publish results for over 20 years after the start of data taking and the Large Hadron Collider (LHC) will have an even longer lifetime.

In order to facilitate this use, software must be stable for long periods; much longer than even Long Term Support operating systems are available. Additionally, the software should reproduce any and all bugs that were present in the original version to ensure the accuracy of the final results. Builds should be reproducible to allow for patches to carefully introduced.

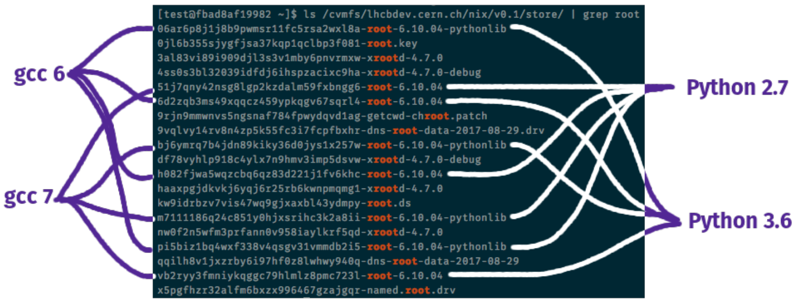
Contradictorily, analysts of data often want to experiment with using modern or even prerelease software to make analysing data easier or to improve final results. However, once a method has been finalised, the environment is expected to stay stable for the remainder of the analysis that can often take multiple years. Even after a result is published, it can still be necessary to rerun the code to combine older results with newer ones and to ensure the best possible combined result is obtained.

Finally, most analysts are physicists with little training in software engineering practices and should not be expected to build and preserve complex software stacks.

7.2.1 nix

Nix[203] is a “purely functional package manager” that was started as a research project in 2003[238]. It has since grown to become both a full Linux based operating system (NixOS) as well as an independent package manager that supports both Linux and macOS. It can build and run software for the `i686`, `x86_64` and `arm64` architectures, either directly or with the use of cross-compilation. `Nix` is used for a wide range of use cases including managed hosting[239], high performance computing (HPC)[240, 241], financial services companies and embedded systems[242].

A strong focus of `nix` is on the purity, reproducible and portability of the builds. Packages are built as deep stacks, with every dependency being defined within `nix` down to the `libc` and `ELF` interpreter. Installed packages are kept in a unique subdirectory of the *store*, typically `/nix/store/`. This subdirectory is named using a cryptographically secure hash of all inputs to the build, including the build sources, configuration and dependencies to allow for an unlimited number of versions and configurations to be available simultaneously, without any risk of conflicts between installations. For example if `ROOT` and `XRootD` are each built with different `Python` and `gcc` versions they each end up in a different directory as shown in Figure 7.3.



```
[test@fbad8af19982 ~]$ ls /cvmfs/lhcbdev.cern.ch/nix/v0.1/store/ | grep root
06ar6p8j1j8b9pwmsr11fc5rsa2wxl8a-root-6.10.04-pythonlib
0jl6b355sjygfjsa37kqp1qclbp3f081-root.key
3al83vi89i909djl3s3v1mby6pnvrmxw-xrootd-4.7.0
4ss0s3bl32039idfdj6ihspzacixc9ha-xrootd-4.7.0-debug
51j7qny42nsg8lgp2kzdalm59fxbngg6-root-6.10.04
6d2zqb3ms49xqqcz459ypkqgv67sqrl4-root-6.10.04
9rjn9mmwnvs5ngsnaf784fpwydqvd1ag-getcwd-chroot.patch
9vqlvy14rv8n4zp5k55fc3i7fcpfbxhr-dns-root-data-2017-08-29.drv
bj6ymrq7b4jdn89kiky36d0jys1x257w-root-6.10.04-pythonlib
df78vyhlp918c4ylx7n9hmv3imp5dsvw-xrootd-4.7.0-debug
h082fjwa5wqzcbq6qz83d221j1fv6khc-root-6.10.04
haaxpgjdkvkj6yqj6r25rb6kwnpmqmg1-xrootd-4.7.0
kw9idrbzv7vis47wq9gjxaxbl43ydmpy-root.ds
m7111186q24c851y0hjxsrihc3k2a8ii-root-6.10.04-pythonlib
nw0f2n5wfm3prfann0v958iaylkrf5qd-xrootd-4.7.0
pi5biz1bq4wxf338v4qsgv31vmmdb2i5-root-6.10.04-pythonlib
qqilh8v1jxzrby6i97hf0z8lwhwy940q-dns-root-data-2017-08-29
vb2ryy3fmniykqggc79hlmlz8pmc723l-root-6.10.04
x5pgfhzr32alfm6bxzx996467gzajgqr-named.root.drv
```

Figure 7.3: Example contents of a `nix` store directory containing `ROOT` and `XRootD` built using `gcc 6` and `gcc 7` against `Python 2.7` and `Python 3.6` in a build matrix. This results in four unique builds of both `ROOT` and `XRootD`, with each having a different dependency hash in their installation directory.

Source files, such as tarballs and patches, are defined using a hash of their content. These are downloaded and imported into the store directory to ensure that all required inputs are available indefinitely, or until they are explicitly deleted.

To ensure builds remain pure and do not have dependencies that have not been explicitly specified, `nix` typically uses a sandbox[243] to isolate builds. This uses modern Linux kernel features, such as user namespaces, to restrict the build to only access the directories within the `nix` store directory that have been specified as a dependency. Additionally, downloaded inputs such as source tarballs and patches are downloaded to the store directory and network access is restricted to prevent builds from downloading files that may change or be removed in future. Builds aim to be bit-for-bit reproducible, though this is an ongoing effort with the wider community to remove non-deterministic elements from the build process[244].

The primary source of `nix` expressions is the `nixpkgs` git repository[245], which contains definitions for $\mathcal{O}(14\,000)$ packages. The `Git` commit hash of a particular revision can be used to pin a snapshot of this repository. Simple modifications can be made when installing by overriding attributes on the deviation that is to be installed. Additionally, *overlays* can be used to modify any part of `nixpkgs`; from a minor configuration change when building, to replacing a low level dependency such as the `gcc` version and flags used for building all known packages hence triggering a rebuild of the entire system.

7.2.2 Hydra

One of the disadvantages of building deep stacks is that it is time consuming and inconvenient for many use cases. To mitigate this issue `nix` can query static web servers using the package's hash to download a signed tarball of the build products. The servers hosting this content are known as binary caches.

Binary caches can be managed using `Hydra`[246], a continuous build system that can be used to build software after every change, after releases or periodically. It has deep integration with `nix` and is primarily built for the testing and deployment of the official `nix` binary cache, though it can also be used to provide build and continuous integration for any project. Private instances are used by several organisations that build the entirety of `nixpkgs` either to apply low level customisations, such as changing the default compiler, or out of security concerns when using externally provided binaries. `Hydra` is also used to provide continuous integration for `nix` projects such as `nix`, `hydra`, `patchelf` as well as some `GNU` projects.

`Hydra` can either be run using a single machine or use `SSH` to distribute builds over a cluster of build machines and has mitigations built to fix common issues, such as misbehaving workers, network issues or random failures. A web interface is provided for configuring `Hydra`, managing builds and viewing build logs. Binaries can be served directly, or uploaded using a plugin system (most commonly to a `S3` compatible endpoint).

7.2.3 Defining packages

`nix` packages are defined using a custom functional language though knowledge of this language is not needed for almost all use cases. The (`nixpkgs`) repository contains many helper functions to simplify defining packages, while also performing actions to help ensure the builds are pure. Package definitions already exist with support for most build systems as well as binary releases. Adding new packages can generally be done by creating a file contain a URL and hash for the source, listing the packages's dependencies and then adding one line to `pkgs/all-packages.nix` to make `nix` aware of the new package. The default build script hides almost all of the complexity of correctly building packages for `nix`. It is highly configurable and splits the build into phases:

- `unpackPhase`: Unpack the archives from the `src` variable.
- `patchPhase`: Apply any patches that are required, taken from the `patches` variable.
- `configurePhase`: Prepare the source tree for building. By default this assumes an Autoconf script and runs `./configure.sh` however including dependencies like `cmake` overrides this as appropriate.

- **`buildPhase`** Compile the package, by default this simply calls `make` provided a suitable `Makefile` is present.
- **`checkPhase`**: Run tests against the build output to avoid broken builds. Defaults to being disabled.
- **`installPhase`**: Install software to the default store directory, typically by running `make install`.
- **`installCheckPhase`**: Similar to `checkPhase` except test against the installed binaries. Also disabled by default.
- **`fixupPhase`**: Perform `nix` specific post-processing. This involves stripping or splitting debug information, patching interpreter paths, minimising runtime dependencies by simplifying the `RPATH` in `ELF` files and splitting the output into multiple parts. Much of this is achieved using patchelf, which is also a `nix` project.

An example derivation that is used to build the base `LHCb` software package is shown in Figure 7.4.

7.2.4 Defining environments

Environments can be defined using `nix` as meta-packages that are “built” by creating a directory of symlinks. `nixpkgs` contains several helpful functions to help with this, the most important of these is `buildEnv`. `Nix` includes an executable (`nix-shell`) that can be used to setup environments, including non standard environment variables such as `ROOTSYS` and `CMAKE_MODULE_PATH`. See the HSF packaging group’s testdrives for an example of using `buildEnv` to define a deep stack[247].

7.2.5 Tests building LHCb software

The LHCb software stack is made up of around 20 separate packages that are typically distributed as binary releases on `CVMFS`[248]. For testing `nix` it was decided to build up to the reconstruction package (`Brunel`) that depends on four other LHCb packages, several “data packages” containing non-executable dependencies like the magnetic field map, as well as many external packages. See Section 2.11 for more details.

To ensure `nix` is suitable for use with the current distribution model, the store directory was changed to a mocked directory representing `CVMFS`, `/cvmfs/lhcbdev.cern.ch/nix/`. This can be done by setting environment variables that override the install directory. As this contributes to the hash that is used to define a package this results in all packages having to be rebuilt.

Initial developments relied upon building all software from source at install time, however it was soon found that setting up a custom `Hydra` instance to serve a binary cache is simple and dramatically improves the experience of using `nix`. This instance is hosted on CERN’s OpenStack cloud and is backed with a Postgres DataBase on Demand (DBoD) instance. `Hydra` is installed inside a minimal docker container running Alpine Linux and

```
{ stdenv, fetchurl, boost, cmake, python, ninja, root, gaudi
, clhep, xercesc, cppunit, libxml2, openssl, relax, gsl, eigen, aida, graphviz
, qt5, mysql57, sqlite, hepmc, cool, coral, libgit2, pkgconfig, vdt, cpp-gsl
, oracle-instant-client, xrootd
# Data packages
, det-sqldddb, fieldmap, gen-decfiles, paramfiles, prconfig, raweventformat
, tck-hlttck, tck-l0tck }:

stdenv.mkDerivation rec {
  name = "LHCb-${version}";
  version = "v44r0";

  src = fetchurl {
    url = "https://gitlab.cern.ch/lhcb/LHCb/repository/${version}/archive.tar.gz";
    sha256 = "0h5wph3p3ha7h34byyamd1dlvb27hs5xpjbfff363y8r43dsk4pa";
  };

  buildInputs = [
    cmake ninja boost gaudi clhep xercesc cppunit libxml2 openssl relax eigen
    gsl aida graphviz qt5.qtbase mysql57 sqlite hepmc cool coral libgit2
    pkgconfig vdt cpp-gsl oracle-instant-client xrootd root
    (python.withPackages (ps: with ps; [ xenv pyqt5 lxml ]))
    det-sqldddb fieldmap gen-decfiles paramfiles prconfig
    raweventformat tck-hlttck tck-l0tck
  ];

  propagatedBuildInputs = [ python ];

  cmakeFlags = [
    "-GNinja"
    "-DMYSQL_INCLUDE_DIR=${mysql57}/include/"
    "-DGRAPHVIZ_INCLUDE_DIR=${graphviz}/include/"
    "-DCOOL_PYTHON_PATH=${cool}/python"
    "-DCORAL_PYTHON_PATH=${coral}/python"
  ];

  checkPhase = ''
    ninja test
  '';
  doCheck = true;

  postInstall = ''
    for fn in $out/lib/lib*.so; do \
      ${gaudi}/bin/listcomponents.exe $fn >> "'${fn%.so}.components"
    done
  '';

  enableParallelBuilding = true;

  meta = {
    homepage = http://lhcbdoc.web.cern.ch/lhcbdoc/lhcb/;
    description = "General purpose classes used throughout the LHCb software.";
    platforms = stdenv.lib.platforms.unix;
  };
}
```

Figure 7.4: `nix` expression of defining `LHCb`, the base library of the LHCb experiment's software stack.

uses SSH to connect to a docker container running on a powerful build machine. Additional build machines were easy to add at times of high load.

The unstable branch of the upstream `nixpkgs` repository was forked to allow easy experimentation with building entirely custom stacks on top of `nix`, such as rebuilding all packages under different `gcc` versions. Maintaining this fork was simple, with `Hydra` automatically monitoring for changes and making new builds as appropriate. An even simpler method was later found known as "pinning" `nixpkgs` which allows an upstream git revision to be specified along with a list of `patch` files.

While most dependencies of the LHCb software stack are already included in `nixpkgs`; `CatBoost`, `COOL`, `CORAL`, `CLHEP`, `frontier`, `pacparser`, `RELAX`, `REFLEX` and `VDT` were missing. Most were trivial to define with only two requiring notable effort:

- `CatBoost` has a closed source build system that depends on glibc. Once this was identified, `patchelf` could be used to modify the provided binaries to find the `ELF` interpretor from a non-standard location.

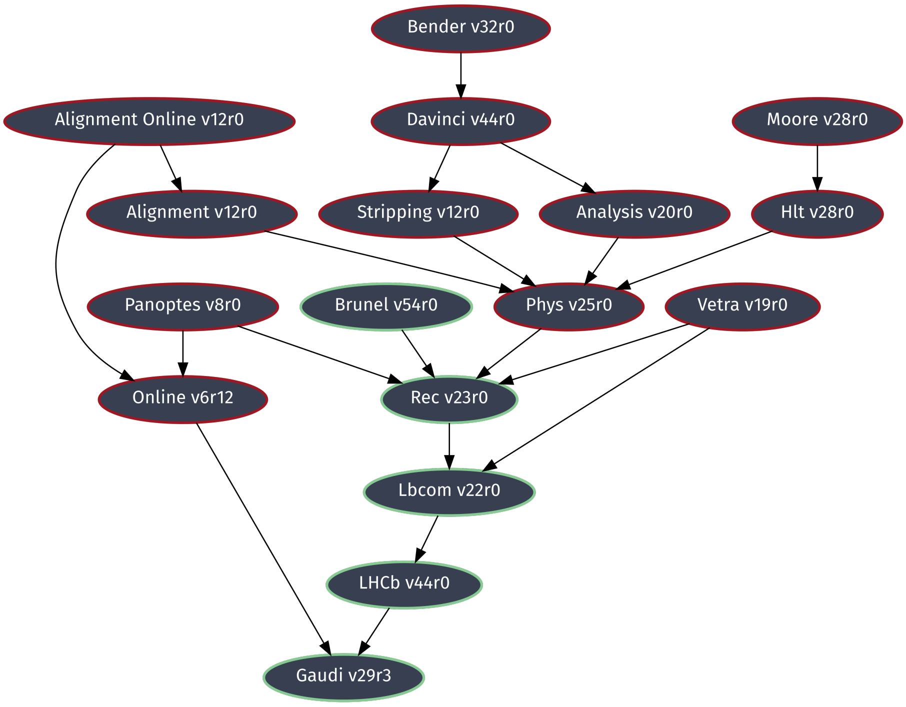
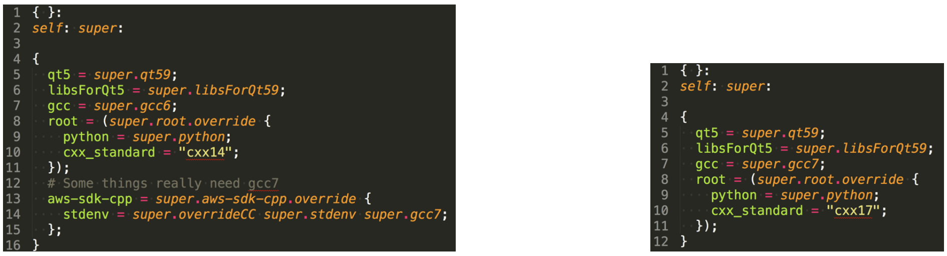


Figure 7.5: Dependency graph containing most of LHCb's software packages. The packages built as part of this work are shown in green.



```
{ }:
self: super:

{
  qt5 = super.qt59;
  libsForQt5 = super.libsForQt59;
  gcc = super.gcc6;
  root = (super.root.override {
    python = super.python;
    cxx_standard = "cxx14";
  });
  # Some things really need gcc7
  aws-sdk-cpp = super.aws-sdk-cpp.override {
    stdenv = super.overrideCC super.stdenv super.gcc7;
  };
}
```

(a) `gcc6.nix`

```
{ }:
self: super:

{
  qt5 = super.qt59;
  libsForQt5 = super.libsForQt59;
  gcc = super.gcc7;
  root = (super.root.override {
    python = super.python;
    cxx_standard = "cxx17";
  });
}
```

(b) `gcc7.nix`

Figure 7.6: Example overlay definition files used to change the default qt and compiler version as well as the `c++` standard. For `aws-sdk-cpp` it was necessary to override the compiler back to `gcc 7` as `gcc 6` is not supported.

- **Oracle Instant Client** is included in `nixpkgs` however licensing issues prevent `nix` from automatically downloading and distributing the source binaries. This required manually downloading/importing the source and enabling builds of non-free software in `Hydra`.

LHCb's software is typically built for a selection of platforms that are defined according to the HSF platform naming convention[249]. This defines a string of the form `architecture-OS-compiler-buildtype` such as `x86_64+avx2-centos7-gcc7-opt` and `x86_64-slc6-gcc49-dbg`. A similar system was achieved within `nix` with the use of overlays, with exception of the `OS` component that is redundant when using `nix` as binaries can be used on any Linux distribution. This allows for modifications to `nixpkgs` to be defined in an external file, such as globally replacing the default `gcc` version. Examples are shown in Figure 7.6.

7.2.6 Containers

Container technologies, such as docker[227] and singularity[229], are seen as a likely solution to many software preservation problems as they providing a simple way to provide a self contained and system independent binary (see Section 7.1.2). Despite this, the problem of how to build a container remains. Typically the build process is effectively a shell script and often downloads dependencies using a package manager with no guarantee that same script will continue to work indefinitely. Nixpkgs provides a solution to this in the form of functions that can build images according to `v1.2.0` of the Docker Image Specification[250] from a `nix` expression. This ensures the reproducibility of the build as the configuration and source dependencies will have been fully cached in the `nix` store of the build machine.

Additionally, containers are often relatively large binary blobs that add overhead when starting jobs and significantly increase the amount of storage required. This situation can be improved by using layers in the container to share a common basis between containers. This basis is not ideal however as each layer in the image is dependent on the previous layer leading to duplication between layers.

Further improvements are possible with `nix` thanks to the fact that each directory within the `nix` store is immutable after installation and has an exactly known set of dependencies that are also store directories. This, combined with the fact that dependencies are defined by images rather than the layers themselves means that each store directory can be placed into a separate layer. Docker images can then be created that depend on arbitrary combinations of these layers to give maximal caching between images using the `pkgs.dockerTools.buildLayeredImage` function from `nixpkgs`. A more advanced algorithm can be used to work around the limit on the number of layers that can be used by an image[251].

7.2.7 Conclusion

High energy physics, and HPC in general, requires highly configurable package management that is able to produce efficient and reproducible binaries. `nix` is an ideal candidate for this task and there is interest from the wider HPC community in `nix`, with several organisations working to improve relevant parts of `nix` such as support for InfiniBand networking, the Intel Math Kernel Library and the Intel Compiler Collection. As software continues to become more and more complex shared effort is becoming essential to ensure builds remain up to date and reliable: especially as many popular ecosystems, such as Python, contain many small packages that can be difficult to distribute in a reliable way. `nix` has been used to successfully build part of the LHCb experiment's software stack and this effort will continue, with changes being pushed upstream in collaboration with the wider community[252].

7.3 Working group productions

The massive computing resources required to process LHC data make it necessary to use distributed computing resources. As previously discussed in Section 2.10.2, this is mostly achieved using the Worldwide LHC Computing Grid with LHCb's job brokerage and data management being performed by Dirac. While the majority of the available resources are utilised by centralised *production* jobs, around 8 % of jobs are submitted and managed directly by physicists and are known as *user* jobs.[253]

Physicists typically interact with this system by submitting and managing jobs using Ganga to: produce ROOT ntuples of candidates, generate specialised Monte Carlo productions, perform large scale toy studies and other computationally intensive tasks. Jobs are executed on a single logical x86_64 processor core and limits for other resources, such as time, disk or RAM, can be specified during submission. Development efforts to add support for executing jobs using multiple CPU cores[254] and GPUs[255] are in the advanced stages, however at the time of writing they are not supported by LHCb's production instance of Dirac.

Unlike the offerings of most commercial cloud providers, that provide reliability and uptime guarantees of 99.99 %[256–260], the grid is optimised for cost efficiency with and has an availability target of 97 % to 99 % depending on the sites purpose.[261] The system is designed to tolerate this, with jobs being retried as necessary. These concessions are acceptable as the offline processing of specific files is not time sensitive at the order of hours or days.

Failure rates for user jobs are considerably higher than production jobs, with over 30 % of wasted CPU time originating from user jobs.[253] This is mostly caused by misconfiguration and/or insufficient testing of jobs. The failures are also more intrusive as retries are not managed by Dirac[91] and a user's Ganga[95] session has to be running in order to monitor and resubmit failed jobs. Additionally, the API exposed by Dirac does not currently scale well when many thousands of jobs are monitored as each job must be individually queried for status updates in a forked subprocess, typically taking hundreds of milliseconds. These problems can be mitigated by moving the management of how to run a job into Dirac, known within the collaboration as *working group productions*. Instead of submitting a series of jobs, each with a subset of the input dataset hard coded, a *transformation* is created that specifies a command and a location within the data management system. Dirac is then able to dynamically split the dataset into subsets and configure jobs to each process subsets of the data. Failures can then be handled automatically by creating smaller subsets or creating jobs that can be executed at a different site.[6] Unlike user productions, output data is stored within the LHCb's central data management system and is therefore accessible to all members of the collaboration. Furthermore, the processing steps applied to the data are automatically and permanently stored which assists with analysis preservation (Section 7.1).

[6] Typically the input data and execution of a job are in the same geographic location to minimise the load on the network connections between sites. However, this is not a requirement and it is possible to remotely stream data.

The code executed in a working group production must be deployed across LHCb's distributed filesystem, `CVMFS`[248]. With the exception of the master node, known as the *Stratum 0*, `CVMFS` is read only and therefore it is only possible to add files by making a centralised request. To simplify management of this system, code[7] that is used for configuration are stored in *data packages* that are stored within CERN's instance of GitLab. Around 50 data packages currently exist.

7.3.1 Automated testing and submission

This system of working group productions has been in place for many years and has been predominately used by the physics performance working groups of the collaboration, for tasks such as calibrating the track reconstruction as well as particle identification algorithms. Creating productions could only be done by members of the collaboration with special privileges within `Dirac` and the submission involved manually completing several web forms. While this is acceptable for a small number of productions, it does not scale for the wide range of jobs that are currently submitted as user productions.

A new system has been implemented in a new data package using GitLab and its Continuous Integration (CI) functionality. Groups of productions are created by making a pull request[8] to the new data package. This contains the configuration code for the desired software application, along side a `JSON` file that contains the configuration for one or more productions. For each production, the following information is specified:

- **Application:** The name of the application within LHCb's software stack to execute.
- **Application Version:** The version of the requested application.
- **Options files:** A list of one or more file that are used to configure the application.
- **Bookkeeping path:** A string corresponding to a dataset within LHCb's data management system.
- **Output data type:** A string that will be used, in combination with the input data, to identify the output data in LHCb's data management system.

Each time the pull request is updated, tests are run automatically for each production that will be created. Each test involves checking the configuration for common mistakes and processing 1000 events. The status, execution logs and output data are reported to the user. At this stage, the production can be reviewed or modified by collaborators that are interested in working with the same dataset. Once ready, the pull request is merged by the maintainer of the data package, triggering the second stage of the continuous integration:

1. **Testing:** More complete testing is performed, with a full data file being processed.[9] The logs and output data of this test from this stage are available to the user.

[7] Data packages are also used to store other assets, such as the map of the detectors magnetic field.

[8] Knowledge of how to create a pull request in GitLab is already required to contribute to the experiments software stack.

[9] The number of events in a file varies significantly between datasets however, $\mathcal{O}\,(100\,000)$ is typical.

2. **Release:** The version number is incremented and a new release of the data package, and the corresponding `Git` tag, is created.
3. **Deployment:** A `JIRA` task is created to request that the new release be installed on `CVMFS`. All deployments to the main `CVMFS` instance are manually triggered by an on-call expert.
4. **Submission:** Once the release has been deployed, the productions are created using the `Dirac` API. The results of the testing step are used to estimate the time required and output size per input file, so this can be used to guide the job submission.

Every six hours, a monitoring job is used to extract the status of the submitted production from `Dirac`. This information is added to a publicly accessible spreadsheet on Google Docs to allow users to quickly see the status of all currently available productions.

7.3.2 Conclusion and ongoing developments

This system has now been used to create over 600 productions and has many benefits over the traditional `Ganga` based approach:

- **Less work for users:** Users do not need to manage job splitting, the resubmission of failed jobs or keep a long running session of `Ganga` open.
- **Better testing:** Jobs are automatically tested prior to large scale submission. Additionally, the configuration is checked for common mistakes.
- **Analysis preservation:** The output data is stored in a centrally known location that is accessible to all members of the collaboration. Additionally code and configuration used to process data are preserved indefinitely ensuring data provenance is maintained. This can also serve as reference material for other analysts.
- **Trivial resubmission:** It is often necessary to reprocess data to include bug fixes, modify the configuration to include new datasets. This can be done by simply opening a pull request to modify the original configuration as required.
- **More efficient use of resources:** Improved testing reduces the likelihood that jobs will fail. The validation and pull request review can reduces the need to recreate productions due to misconfiguration.

Further developments are ongoing to further simplify the workflow:

- **More validation:** Some misconfiguration, such as creating a variable with the result of an incorrectly named trigger line, can be difficult to quickly identify by examining the output of tests with a small data sample. These mistakes can be checked for automatically and with warnings shown to users. Additionally, when issues are found with existing productions, new validations can be added to prevent mistakes being repeated.
- **Additional flexibility:** The current submission system is limited to productions that only contain a single processing step and one merging step that takes data files

as input. This is not an inherent limitation of the system and interest has been expressed to have multi-step productions and those that do not use input data, such as toy studies. Additionally, support should be added for running LHCb applications that have been compiled using additional patches.

- **Automated deployment:** Deploying the data package to `CVMFS` currently involves the on-call software deployment expert manually triggering the installation. Work to fully automate this step is ongoing.
- **Automated starting of productions:** After productions are submitted they must be further configuration must be added by the an expert within the distributed data processing team. The information required for this is available to the submission step in GitLab CI, therefore this can also be automated.
- **Analysis trains:** Submitting jobs using WG productions adds the potential for jobs to be grouped together into what are often referred to as *analysis trains*. These are used by many other HEP collaborations[262, 263] and can reduce the computing resources required, in particular by sharing the streaming and unpacking of data.
- **Data access:** The LHCb data management system[163] requires[10] users to have a local installation of `Dirac` to query and obtain a path that can be used to access the data. This should be replaced with a simpler and more robust system for analysts.

These developments are seen as an important first step to a new data processing paradigm for LHCb. This will be essential as the LHCb Upgrade programmes will generate larger and larger datasets.

7.4 Software training

In 2015 a group of 11 young researchers, including both Masters and PhD students, decided there was a need for better software training within LHCb. A five day workshop was organised at CERN in July 2015[264, 265] where attendees were taught basic software skills as well as how to use the LHCb experiment's core software. The first two days were taught in conjunction with Software Carpentry[266], whom provided guidance that continues to heavily influence Starterkit activities. Thanks to the success of the first workshop, this week long workshop has continued and now takes place in October or November of each year to coincide with when most students join the LHCb collaboration. A follow up workshop, known as the Impactkit, has been added that builds upon the material taught in the Starterkit. Furthermore, the 2017 Starterkit was organised in conjunction with the ALICE experiment for the first time with 49 LHCb students and 25 ALICE students. This collaboration was very successful and was repeated for the 2018 Starterkit, with the smaller SHiP collaboration also being involved. It is expected that this collaboration between experiments will continue. The author of this thesis has been heavily involved since the second Starterkit and was responsible for spreading this

[10] Workarounds exist to avoid this limitation however, they sacrifice the redundancy of being able to access data from an alternative site if one replica is unavailable.

initiative to the ALICE and SHiP collaborations. Additionally, the author co-organised the 2017 Starterkit the 2018 Impactkit.

An important principle when teaching was that copying “magic” lines of code must be avoided. Instead, code that is able to perform a minimal task is written with each line being explained. The example can then be built upon, with the aim of the exercises being to explain the core concepts that underpin the software that is being used. Lessons are taught interactively, with the students following along on their laptops. When teaching in this style, it is essential that all students are able to follow and that they do not get left behind while fixing issues. This is achieved by keeping the group sizes small, ideally with no more than 25 students in a room. Additionally, two *helpers* are available to assist students with problems they encounter and to answer questions when they arise. For these helpers, it is important that they remain engaged during the lesson and actively monitor the progress of students. This particularly helps quieter students and also allows them to provide real-time feedback to the teacher if issues arise that affect many students.

7.4.1 Lesson Material

Despite the lessons being taught interactively, standalone material is made available on a website. This includes written explanations of everything that is covered in the lessons and is designed for new members of the collaboration that are unable to attend one of the workshops. Most supervisors now encourage their new students to follow the lessons on the webpage. This is also beneficial if they later attend a Starterkit as being already familiar with the material helps allow more subtle details to be understood. Additionally, the Starterkit lesson material has now become the standard documentation for many common tasks thanks to it being extensively reviewed and updated annually in preparation for each Starterkit.

7.4.2 Agenda

Prior to the Starterkit commencing, a set of prerequisites are set to students. The most important of these is that they must be able to use `SSH` connect to CERN’s interactive logon service, `lxplus`. This helps ensure all students have a similar environment for the lessons. Additionally, `lxplus` is an important part of the LHCb software environment and is used by the majority of the collaboration.

The agenda of the Starterkit has remained largely unchanged since the first workshop and the 2017 schedule is shown in Figure 7.7. During the first two days `Python`, `bash` and `Git` are taught. One of the difficulties with teaching this material is the variation in the experience of the Starterkit’s participants. Earlier Starterkits taught these basic skills from a very low level, with even trivial tasks being explained such as changing directories with `cd` and what a `for` loop and `if` statement are in `Python`. These lessons have since evolved to skip this lower level content and cover more complex tasks such as using `tmux` or training a multivariate classifier. However, a small minority of participants require the introductory content and it is important that these people are included. To date the best

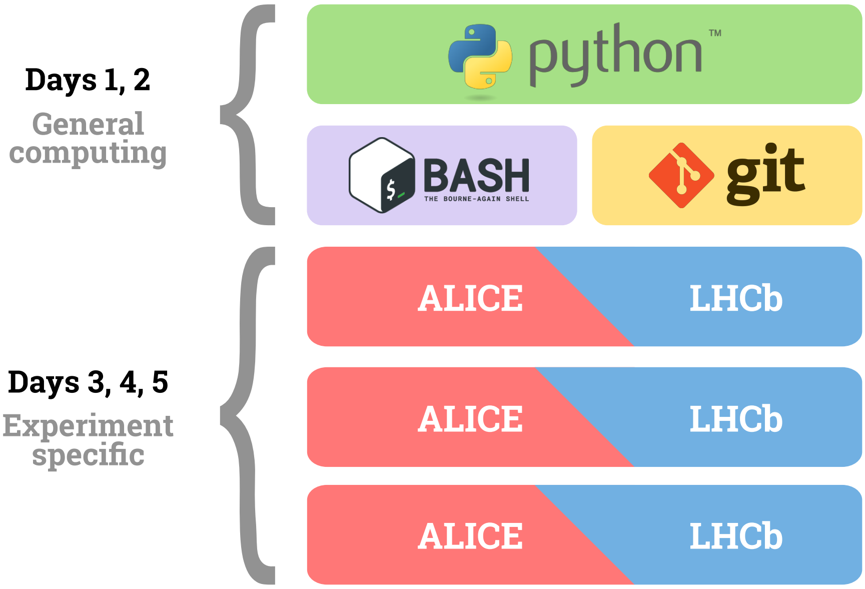


Figure 7.7: Schedule of the 2017 Starterkit. This was the first event that was ran in conjunction with another high energy physics collaboration.

solution has been for helpers to identify these students and quietly provide a one-on-one lesson. Since 2017, this portion of the Starterkit has been taught in three parallel sessions; with each session containing a mixture of students from each participating experiment.

The final three days of the Starterkit start with a short introduction by a chairperson from the LHCb Early Career Gender and Diversity office.[267] The group is then split into two parallel sessions that each cover the same content, starting with an approximately one hour overview of how data is processed prior to being made available to analysts. The remainder of the workshop is then devoted to teaching how to configure, develop and run the LHCb software, both locally and on the grid. During the evening of the penultimate day, a social event is organised with food and refreshments provided. This gives the various experiments and parallel groups the opportunity to network.

The three day follow up Impactkit workshop in May is split into two halves. The first 1.5 days extends the LHCb specific Starterkit content by teaching about simulation[11], the trigger and several advanced analysis tools. In 2018 an analysis preservation lesson was added and it is expected this will continue.

A “hackathon” then takes place during the second half of the ImpactKit event with a list of projects being presented to the attendees. Attendees then choose a project to work on in small groups with helpers being available. As well as teaching the attendees, these projects had led to many contributions to the core software stack. At the end of the hackathon, the students present lightning talks about their projects. After these talks, a social event in the form of a barbecue is held at the site of the LHCb experiment.

[11] Most LHCb simulation is produced centrally with analysts making requests, via liaisons, to the simulation and production experts. However, it is occasionally necessary to make small private productions for specialised studies.

7.4.3 Organisation and sustainability

Often software training in high energy physics collaborations becomes the responsibility of a small number of volunteer experts. This can be beneficial and allow for feedback to be collected during the lesson to guide future developments. However, providing lessons is a time consuming activity that detracts from the time available for other tasks. To counteract this the Starterkit encourages participants to act as a helper or teacher in the following years. This also has the benefit of requiring the teachers and helpers to learn the material in more detail and gain a better understanding.

Each workshop is organised by two people[12] and efforts are made to select new people each time. This reduces the burden each person involved and makes the Starterkit initiative more resilient to people leaving the LHCb collaboration. It also provides an opportunity for young people to become better known and gain experience organising events. This effort to ensure the sustainability of the Starterkit has been successful so far, with one of the 2017 organisers having attended the first Starterkit and both 2018 organisers having attended a Starterkit during the first few months of their PhDs.

7.4.4 Conclusion

The Starterkit initiative has been hugely successful within LHCb and has continued despite many of the original organisers leaving academia. The material that is used for teaching has become the standard documentation for many common tasks and had remained up to date since it was first written. Additionally, it has been noticed that new members of the collaboration are more comfortable configuring the software.

Unfortunately, all Starterkit events to date have been hosted at CERN with no remote participation possible. This disadvantages members of the collaboration that are less able to travel to CERN, especially those from outside of Europe. Expanding the initiative to the wider HEP community would be beneficial and potentially make it feasible to have workshops hosted elsewhere.

7.5 The Energy test

Hypothesis testing is essential to many areas of research and has already been used for this thesis in Section 6.2, for computing limits using the $\mathrm{CL_s}$[268] method. Another common application of hypothesis testing is to determine how likely it is that two ensembles of points, that are randomly sampled from an N-dimensional space, originate from the same underlying probability density function. An example of this within the LHCb physics program is in the search for CP violation. In the most naive approach, CP violation can be tested by counting the number of observed decays, N, through a given channel and its charge conjugate, N'. If production and experimental asymmetries can be neglected, the

[12] While only two people are formally responsible for organising, lots of support is available from across the collaboration. From experience, volunteers can be easily found for any task even if no notice is given.

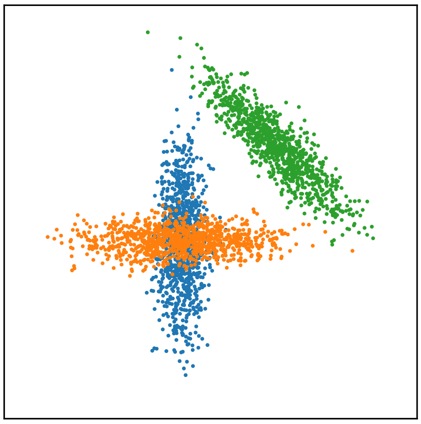


Figure 7.8: Three example distributions where $\frac{1}{n(n-1)}\sum\psi_{ij}$ is the same, despite not originating from the same underlying probability density function.

asymmetry is then given by

$$\mathrm{A_{CP}} = \frac{\mathrm{N} - \mathrm{N}'}{\mathrm{N} + \mathrm{N}'}. \tag{7.1}$$

This is known as the *global CP asymmetry* however, in many decays the asymmetry varies across the phase space of the decay products.[269] This can result in *local CP asymmetries* that are much larger than the global asymmetry. To measure this experimentally the phase space is often binned into regions that are either uniformly distributed or optimised to avoid the cancellation of positive and negative CP violation. While this has the benefit of probing for local CP violation without introducing model dependence, some information is still lost due to the binning procedure.

Recently the energy test[270, 271] has risen in popularity[269, 272–274] in the study of CP violation. A distance function, ψ, is used to compute the "energy" between the two samples. This is analogous to the electrostatic potential energy of a system containing positive and negatively charged particles, where the position for each of the particles is chosen randomly from an underlying distribution that could be different for positive and negatively charged particles. In the case that the underlying distributions are identical, the total electrostatic energy will converge to a minimum as the number of particles increases.

In the energy test, the test statistic

$$T = \frac{1}{n\,(n-1)}\sum_{i>j}^{n}\psi_{ij} + \frac{1}{\bar{n}\,(\bar{n}-1)}\sum_{i>j}^{\bar{n}}\psi_{ij} - \frac{1}{n\bar{n}}\sum_{i,j}^{n,\bar{n}}\psi_{ij} \tag{7.2}$$

is used, where n is the number of points in sample one and $\bar{n}$ is the number of points in sample two. The first term in the sum corresponds to the average distance between the points in sample one. Correspondingly, the second term is the average distance between points in sample two. The third term is then the average distance between points from different samples. For observing differences, first two terms ensure both samples have the same relative distribution and the third term is sensitive to cases when the second distribution is a translation and/or rotation of the first. See Figure 7.8 for a pictorial representation of this.

Once a value of the test statistic has been computed for the given dataset, the underlying distribution must be known in order to compute a p-value that rejects the hypothesis that both originate from the same underlying distribution.[13] This can be done by using a bootstrapping method; where the two samples are combined and labels are randomly assigned to each point. It is important to ensure that the number of points in each sample remains the same to account for the reduced sensitivity that is present when one sample is smaller. The test statistic is then computed and this is commonly referred to as a *permutation*. Additional permutations are computed until the underlying distribution is sufficiently well understood. In the case of $p = 0.999\,999\,42$, corresponding to a 5σ observation, many millions of test statistics must be computed to accurately determine the observed p-value.

7.5.1 Computational complexity

One of the key limitations of the energy test is the computation complexity of calculating the test statistic. This is normally described using big O notation to show the limiting behaviour of the function describing the runtime of the function. For the energy test, the complexity is given by

$$\mathcal{O}\left(n^2 + \bar{n}^2 + n\bar{n}\right) = \mathcal{O}\left(n^2 + \bar{n}^2\right). \tag{7.3}$$

If $n \approx \bar{n}$, as is normally the case in CP violation studies, this reduces to

$$\mathcal{O}\left(n^2\right). \tag{7.4}$$

In order to compute a p-value this must be repeated m times, where m is often dependent on n as measurements with larger samples are sensitive to smaller p-values. Additionally the dataset must be shuffled for each permutation introducing an additional n term[275] in each permutation. The total computational complexity of making a measurement with the energy test is therefore

$$\mathcal{O}\left(n^2 + m\left(n + n^2\right)\right)\right) = \mathcal{O}\left(mn^2\right). \tag{7.5}$$

Even with a relatively modest sample with $n = 100\,000$, around 1×10^{17} distances must be computed and summed to compute a 5σ p-value. For a 4 body decay, a trivial distance function and the fastest currently available workstation GPU[14] operating at its theoretical peak performance, this calculation would take over two days. In practice, this is an underestimate by an order of magnitude and there is also interest in applying this method to samples containing tens of millions of events making it impractical, for charm physics, with all but the smallest of samples.

[13] In the case of CP violation studies, this corresponds to and asymmetry between the two samples. It does not give a measure of the magnitude or phase space dependence of the observed asymmetry. See Reference [272] for a discussion of variations of the energy test that visualise the phase space dependence.

[14] The Nvidia V100 has a theoretical peak performance of 7 Pflop s^{-1}.

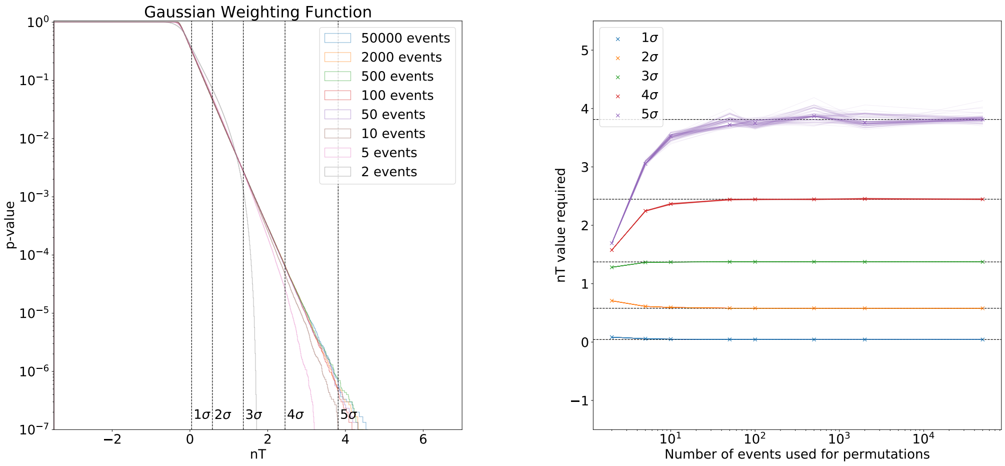


Figure 7.9: Validation of the scaling property of the energy test, when comparing samples that each contain 500 000 points using a Gaussian weighting function. In each 8 different scaling factors are used and in each case 30 000 000 permutations are used to compute the test statistic distribution. The left plot shows the p-value as a function of the test statistic, also known as the survival function. The right plot shows the variation of the test statistic that would be required for a $n\sigma$ rejection of the hypothesis that both samples originate from the same underlying probability distribution. The dashed black lines show the test statistic required in the test case with the largest number of events used for permutations. The coloured lines are displayed by bootstrapping 100 times to show the uncertainty on the required test statistic. The distribution quickly converges with fewer than 100 events being required to accurately model the distribution.

In References [271] and [272] it was observed that the distribution can be approximated by a fitting a small number of permutations with a Generalised Extreme Value (GEV)[276] function. This was only a casual observation and it was explicitly stated that this approximation must be validated each time it is applied. If a weighting function is chosen that is bounded in the range $[0, 1]$, such as a Gaussian function, the test statistic must be contained by the interval $[-1, 1]$. The GEV function always predicts a finite probability outside of this range making it pathological at some level. Furthermore, Reference [204] was able to demonstrate this failure using a representative toy model.

7.5.2 Mathematical optimisation

The three terms in Equation 7.2 each correspond to computing the mean of the weighting function applied to n^2 distances. If the distances are each statistically independent when computing the permutations, the variance on each of these means is inversely proportional to n and therefore the number of points used when computing the permutations can be reduced by a factor k and the test statistic can be scaled by a factor of $\frac{1}{k}$. This is known as the *scaling property* of the energy test.

A key assumption of this approximation is that the distances used are all statistically independent. This is not the case in the energy test as each point is reused n times. A proof that this correlation can be neglected was not available, instead a physically

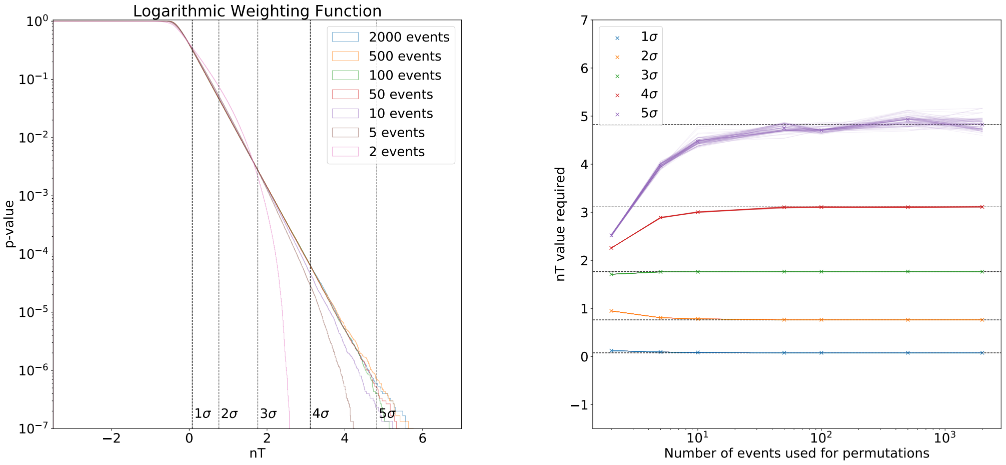


Figure 7.10: Validation of the scaling property of the energy test with a logarithmic weighting function. See Figure 7.9 for a full description.

motivated toy model was used to validate this assumption. `Laura++`[277] was used to simulate a 1 000 000 three body decays, including intermediate resonances. The points are then the invariant masses between each pair of particles, these are also known as the Dalitz variables. This reduces down to a 2-D dataset as one of the three invariant masses is redundant, due to momentum conservation constraining the phase space of the decay. The choice of weighting function is arbitrary and can be tuned to optimise the sensitivity of the test. Two common choices are a Gaussian function, $e^{-\frac{d^2}{\delta^2}}$, or a logorithmic function, $-\log \|d + \epsilon\|$; where d is the Euclidean distance and δ/ϵ are free parameters.

Figures 7.9 and 7.10 show a comparison of the test statistic distribution obtained with 30 000 000 permutations with the toy dataset, for a selection of scaling factors. From this it is clear that calculating permutations with fewer than 100 events and then scaling can accurately reproduce this distribution.

7.5.3 Computational optimisation

Using the previously discussed scaling method the computational complexity can be reduced from

$$\mathcal{O}\left(mn^2\right) \tag{7.6}$$

to

$$\mathcal{O}\left(n^2 + mnn'\right) = \mathcal{O}\left(n^2 + mn\right), \tag{7.7}$$

where n' is a constant, denoting the number of events used for permutations. Additionally, it is no longer necessary to shuffle the entire array for each permutation as only the first n' points have been used. Instead the dataset can be shuffled once for each $\lfloor \frac{n}{n'} \rfloor$ permutations, making the time complexity reduce to

$$\mathcal{O}\left(n^2 + mnn'\left(\lfloor \frac{n}{n'} \rfloor\right)^{-1}\right) = \mathcal{O}\left(n^2 + m\right). \tag{7.8}$$

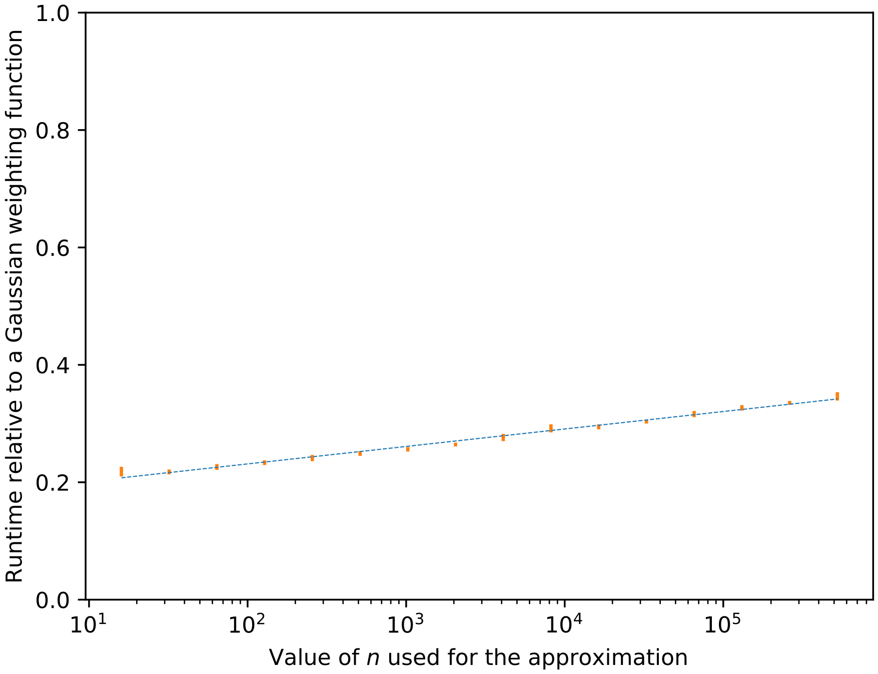


Figure 7.11: Time taken to compute the energy test's test statistic using $\left(1 - \frac{d^2}{\delta^2 n}\right)^n$, relative to a Gaussian weighting function, as a function of n. Each of the two samples used contain 25 000 entries. The calculation is repeated 25 times, with the mean and standard deviation being reported in the plot. Note n is shown on a logarithmic scale as it can be efficiently implemented using $\log_2(n)$ repeated multiplications.

Therefore the cost of computing the distribution of the test statistic is independent of the sample size.

Comparing two samples using the energy test can be further optimised with the choice of weighting function. For example, using a Gaussian weighting function requires an exponential with a negative exponent to be computed. This can be rewritten in an easier to compute form as

$$e^{-\frac{d^2}{\delta^2}} = \lim_{n\to\infty}\left(1 - \frac{d^2}{\delta^2 n}\right)^n, \tag{7.9}$$

which only requires one addition and $2 + \log_2(n)$ multiplication operations to compute. The improvement performance when calculating the energy test statistic is shown in Figure 7.11. This approximation is also most accurate for when $\|d\|$ is small and this corresponds to the dominant terms in the evaluation of the test statistic. Additionally, as the choice of weighting function is arbitrary, an approximated weighting function does not introduce a systematic uncertainty on final p-value that is computed.

7.5.4 Conclusion

The energy test is a powerful, though computationally expensive, tool for comparing multidimensional datasets. The scaling property makes it possible for very small p-values to be computed for samples containing many tens of millions of points, especially when using modern GPU hardware and a carefully chosen weighting function. This property has subsequently been proven in Reference [278] and further developed in Reference [279].

Chapter 8

Summary and Conclusions

In this thesis three main bodies of work are presented. The first section summarises alignment studies that were performed for the LHCb VELO Upgrade programme. Next, a search for rare charm decays of the form $D^+_{(s)} \to h^\pm l^+ l'^\mp$ is described and upper limits are given for the branching fraction of each decay. Finally, the various developments that have been made to improve the effectiveness of HEP data analysis are descried. This chapter summarises the main results and considers what the future may hold in each area.

8.1 Searching for $D^+_{(s)} \to h^\pm l^+ l'^\mp$

Upper limits have been obtained for 25 rare $D^+_{(s)}$ decays and, as can be seen in Figure 8.1, the majority of these represent an order of magnitude improvement with respect to previous results. This analysis is the first of this kind at LHCb where a significant number of channels with a similar topology are analysed simultaneously. The analysis was performed in a highly automated fashion and is also amongst the first in the collaboration to adopts some analysis preservation strategies.

In the near future, the 2017 and 2018 LHCb datasets can contribute an additional $2.9\,\text{fb}^{-1}$ of data that will allow for even smaller branching fractions to be probed. Furthermore, the LHCb Upgrade I and Upgrade II programmes are expected to expand this dataset to a total of $50\,\text{fb}^{-1}$ and $300\,\text{fb}^{-1}$ respectively. The future branching fractions to which LHCb will be sensitive are estimated in Figure 8.2 under the assumption the analysis performance scales with the square root of the collected luminosity. The LHCb Upgrade programme is also expected to improve the efficiency of selecting charm hadrons by removing the Level 0 hardware trigger and this is accounted for in the prediction by doubling the effective luminosity for Run 3 onwards. As shown in Figure 5.5, LHCb will continue to be able to probe the branching fraction of $D^+ \to \pi^+\mu^+\mu^-$ for the foreseeable future without the resonant contributions dominating. In addition to branching fraction searches, some theorists have suggested CP violation studies of the ϕ resonance and the high q^2 region of $D^+_{(s)} \to \pi^+\mu^+\mu^-$ could be sensitive to BSM contributions[24].

In $D^+ \to \pi^+\mu^\pm e^\mp$ the results of this thesis already exclude some leptoquark scenarios proposed by Reference [24] and the full upgrade datasets will continue to expand these

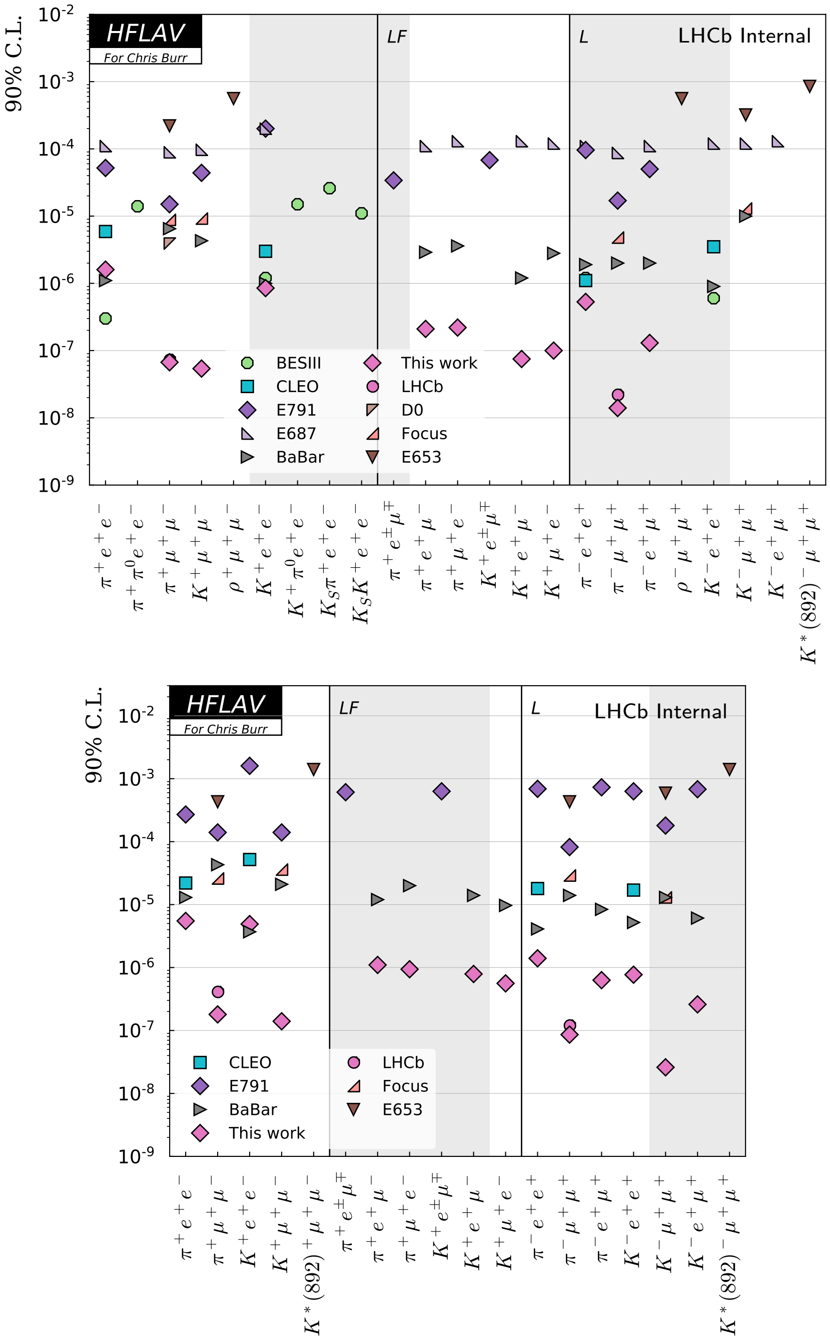


Figure 8.1: Unofficial HFLAV plots showing this results of this thesis alongside other measurements of rare $D^+_{(s)}$ decays. The middle section shows lepton flavour violating decays (*LF*) and the rightmost column shows decays that are both lepton flavour and number violating (*LF*). The 25 new results from this analysis are shown, 23 of which improve on the previous world best limits.

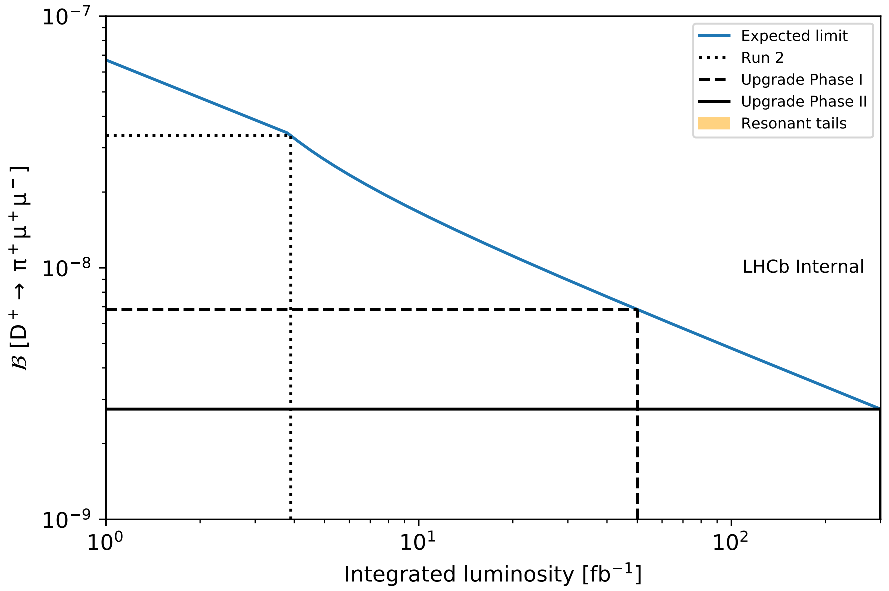


Figure 8.2: Estimated sensitivity for excluding $D^+ \to \pi^+\mu^+\mu^-$ at 90 % confidence as the LHCb dataset grows, see text for details. The uncertainty on the prediction from Reference [24] for when the resonant contributions will dominate in the low q^2 region is represented by the orange band.

constraints. For final states arising from $c \to ue^+e^-$ transitions, where this analysis is less sensitive, there are no BSM contributions foreseen.[24] These measurements are however useful as they contain resonant contributions that can be used for validating the treatment of electrons. Furthermore, parts of this work are being used by other analysts in LHCb to provide validation for lepton flavour universality measurements in B decays. The LHCb experiment is ideally suited to making measurements such as these in this thesis thanks to its excellent reconstruction performance for charged final states. Combined with the huge samples that are produced by the LHC, this makes LHCb a true *charm factory* that will continue to define flavour physics for the next decade and beyond.

8.2 Alignment studies for the LHCb VELO Upgrade

A comprehensive study has been performed detailing the potential impact of misalignment in the LHCb VELO Upgrade. This study helped guide the choice of module substrate at the VELO Upgrade Mechanical Module EDR in 2017 and has since resulted in modification to the manufacturing and quality assurance procedures to ensure measurements are made of each module. These measurements will be useful for imposing constraints on the module positions to ensure the best possible detector performance is obtained. Such information can be used directly as a Lagrange constraint on the module position or parametrically with the temperature readout being used as an input to the reconstruction procedure. The quality of the detector alignment is particular important in the LHCb Upgrade as it represents a radical shift in how HEP data is analysed, with the majority of data being reconstructed in close to real time before permanently discarding the raw detector readout.

The ability to quickly align and calibrate the detector in this situation is essential as many corrections cannot be reapplied if issues are found at a later date. The LHCb VELO upgrade is currently in construction, with installation starting in late 2019. The detector will then be commissioned and initial data taking will occur in 2021. It is expected that the detector will operate for around a decade.

8.3 Analysis tools for HEP

Many improvements to the wider HEP analysis ecosystem have been made while the work in this thesis was performed. This has helped advise new policies within LHCb that mandate the archival of all datasets and software that are used for published analyses. These requirements will help ensure results can be reproduced if required and it is hoped such efforts can improve the efficiency of the collaboration by increasing the reusability of analysis code. Developments in packaging are helping preserve the software environments required to reproduce analyses, while also providing analysts with the means to quickly and reliably use newly released tools. Contributions have been made to this area by evaluating the Nix package manager for use in LHCb and by the addition of various packages to conda, the most notably ROOT. The LHCb Starterkit initiative is helpful as a method of promoting these best practises. The author was heavily involved in this effort and has helped expanded this initiative to the ALICE and SHiP collaborations. Working group productions are a high level abstraction within LHCb that allows distributed computing resources to be used without consideration of how the processing is performed. The work of this thesis has allowed for them to grow in popularity and continues to provide improvements in both scope and ease of use. It is expected working group productions will form a central part of the LHCb Upgrade I analysis model. These various developments will lead the way for the more efficient use of both human and computing resources that will become increasingly important as the LHCb dataset is expected to grow exponentially for the foreseeable future.

Bibliography

[1] R. Aaij *et al.*, “Measurements of prompt charm production cross-sections in pp collisions at $\sqrt{s} = 13$ TeV”, *JHEP*, vol. 03, p. 159, 2016. DOI: 10.1007/JHEP03(2016)159. arXiv: 1510.01707 [hep-ex].

[2] R. Aaij *et al.*, “Measurements of prompt charm production cross-sections in pp collisions at $\sqrt{s} = 5$ TeV”, *JHEP*, vol. 06, p. 147, 2017. DOI: 10.1007/JHEP06(2017)147. arXiv: 1610.02230 [hep-ex].

[3] A. Pearce, G. Lafferty, and S. Easo, “Measurements of charm production and CP violation with the LHCb detector”, Presented 09 Dec 2016, Oct. 2016. [Online]. Available: https://cds.cern.ch/record/2254814.

[4] D. Muller, M. Gersabeck, and C. Parkes, “Measurements of production cross-sections and mixing of charm mesons at LHCb”, Presented 23 Oct 2017, Nov. 2017. [Online]. Available: https://cds.cern.ch/record/2297069.

[5] D. Griffiths, *Introduction to elementary particles*. Weinheim Germany: Wiley-VCH, 2008, ISBN: 9783527406012.

[6] M. Bona *et al.*, “Model-independent constraints on $\Delta F = 2$ operators and the scale of new physics”, *JHEP*, vol. 03, p. 049, 2008. DOI: 10.1088/1126-6708/2008/03/049. arXiv: 0707.0636 [hep-ph].

[7] J. F. Kamenik, “Flavour Physics and CP Violation”, in *Proceedings, 2014 European School of High-Energy Physics (ESHEP 2014): Garderen, The Netherlands, June 18 - July 01 2014*, 2016, pp. 79–94. DOI: 10.5170/CERN-2016-003.79. arXiv: 1708.00771 [hep-ph].

[8] M. Tanabashi, K. Hagiwara, K. Hikasa, *et al.*, “Review of Particle Physics”, *Phys. Rev. D*, vol. 98, p. 030 001, 3 Aug. 2018. DOI: 10.1103/PhysRevD.98.030001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[9] F. Englert and R. Brout, “Broken Symmetry and the Mass of Gauge Vector Mesons”, *Phys. Rev. Lett.*, vol. 13, pp. 321–323, 9 Aug. 1964. DOI: 10.1103/PhysRevLett.13.321. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[10] P. W. Higgs, “Broken Symmetries and the Masses of Gauge Bosons”, *Phys. Rev. Lett.*, vol. 13, pp. 508–509, 16 Oct. 1964. DOI: 10.1103/PhysRevLett.13.508. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[11] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles", *Phys. Rev. Lett.*, vol. 13, pp. 585–587, 20 Nov. 1964. DOI: `10.1103/PhysRevLett.13.585`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevLett.13.585`.

[12] G. Aad *et al.*, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett.*, vol. B716, pp. 1–29, 2012. DOI: `10.1016/j.physletb.2012.08.020`. arXiv: `1207.7214 [hep-ex]`.

[13] S. Chatrchyan *et al.*, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett.*, vol. B716, pp. 30–61, 2012. DOI: `10.1016/j.physletb.2012.08.021`. arXiv: `1207.7235 [hep-ex]`.

[14] M. Kobayashi and T. Maskawa, "CP-Violation in the Renormalizable Theory of Weak Interaction", *Progress of Theoretical Physics*, vol. 49, pp. 652–657, Feb. 1973. DOI: `10.1143/PTP.49.652`.

[15] J. H. Christenson, J. W. Cronin, V. L. Fitch, *et al.*, "Evidence for the 2π Decay of the K_2^0 Meson", *Phys. Rev. Lett.*, vol. 13, pp. 138–140, 4 Jul. 1964. DOI: `10.1103/PhysRevLett.13.138`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevLett.13.138`.

[16] B. Aubert, D. Boutigny, I. De Bonis, *et al.*, "Measurement of *CP*-Violating Asymmetries in B^0 Decays to *CP* Eigenstates", *Phys. Rev. Lett.*, vol. 86, pp. 2515–2522, 12 Mar. 2001. DOI: `10.1103/PhysRevLett.86.2515`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevLett.86.2515`.

[17] K. Abe, K. Abe, R. Abe, *et al.*, "Observation of Large *CP* Violation in the Neutral *B* Meson System", *Phys. Rev. Lett.*, vol. 87, p. 091 802, 9 Aug. 2001. DOI: `10.1103/PhysRevLett.87.091802`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevLett.87.091802`.

[18] "Observation of *CP* violation in charm decays", no. CERN-EP-2019-042. LHCB-PAPER-2019-006, Mar. 2019. [Online]. Available: `http://cds.cern.ch/record/2668357`.

[19] L. Wolfenstein, "Parametrization of the Kobayashi-Maskawa Matrix", *Phys. Rev. Lett.*, vol. 51, pp. 1945–1947, 21 Nov. 1983. DOI: `10.1103/PhysRevLett.51.1945`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevLett.51.1945`.

[20] J. Charles, A. Hocker, H. Lacker, *et al.*, "CP violation and the CKM matrix: Assessing the impact of the asymmetric *B* factories", *Eur. Phys. J.*, vol. C41, no. 1, pp. 1–131, 2005. DOI: `10.1140/epjc/s2005-02169-1`. arXiv: `hep-ph/0406184 [hep-ph]`.

[21] S. L. Glashow, J. Iliopoulos, and L. Maiani, "Weak Interactions with Lepton-Hadron Symmetry", *Phys. Rev. D*, vol. 2, pp. 1285–1292, 7 Oct. 1970. DOI: `10.1103/PhysRevD.2.1285`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevD.2.1285`.

[22] A. Lenz, “Heavy flavour physics and effective field theories”, 2017. [Online]. Available: `http://www.ippp.dur.ac.uk/~lenz/Lecture_Flavour_2017.pdf`.

[23] S. Fajfer and N. Košnik, “Prospects of discovering new physics in rare charm decays”, *Eur. Phys. J.*, vol. C75, no. 12, p. 567, 2015. DOI: `10.1140/epjc/s10052-015-3801-2`. arXiv: `1510.00965 [hep-ph]`.

[24] S. de Boer and G. Hiller, “Flavor and new physics opportunities with rare charm decays into leptons”, *Phys. Rev.*, vol. D93, no. 7, p. 074 001, 2016. DOI: `10.1103/PhysRevD.93.074001`. arXiv: `1510.00311 [hep-ph]`.

[25] A. J. Buras, “Flavour Visions”, *PoS*, vol. BEAUTY2011, p. 008, 2011. DOI: `10.22323/1.129.0008`. arXiv: `1106.0998 [hep-ph]`.

[26] D. Ambrose *et al.*, “Improved branching ratio measurement for the decay $K^0_{(L)} \to \mu^+\mu^-$”, *Phys. Rev. Lett.*, vol. 84, pp. 1389–1392, 2000. DOI: `10.1103/PhysRevLett.84.1389`.

[27] T. Akagi *et al.*, “Experimental study of the rare decays $K^0_{(L)} \to \mu e$, $K^0_{(L)} \to ee$, and $K^0_{(L)} \to \mu\mu$ and $K^0_{(L)} \to eeee$”, *Phys. Rev.*, vol. D51, pp. 2061–2089, 1995. DOI: `10.1103/PhysRevD.51.2061`.

[28] A. Heinson *et al.*, “Measurement of the branching ratio for the rare decay $K^0_{(L)} \to \mu{+}\mu{-}$”, *Phys. Rev.*, vol. D51, pp. 985–1013, 1995. DOI: `10.1103/PhysRevD.51.985`.

[29] L. M. Sehgal, “Electromagnetic Contribution to the Decays $K_S \to l\bar{l}$ and $K_L \to l\bar{l}$”, *Phys. Rev.*, vol. 183, pp. 1511–1513, 5 Jul. 1969. DOI: `10.1103/PhysRev.183.1511`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRev.183.1511`.

[30] L. M. Sehgal, “Electromagnetic Contribution to the Decays $K_S \to ll$ and $K_L \to ll$”, *Phys. Rev. D*, vol. 4, pp. 1582–1582, 5 Sep. 1971. DOI: `10.1103/PhysRevD.4.1582.4`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevD.4.1582.4`.

[31] A. V. Artamonov *et al.*, “New measurement of the $K^+ \to \pi^+\nu\bar{\nu}$ branching ratio”, *Phys. Rev. Lett.*, vol. 101, p. 191 802, 2008. DOI: `10.1103/PhysRevLett.101.191802`. arXiv: `0808.2459 [hep-ex]`.

[32] R. Aaij *et al.*, “Measurement of the $B^0_s \to \mu^+\mu^-$ branching fraction and effective lifetime and search for $B^0 \to \mu^+\mu^-$ decays”, *Phys. Rev. Lett.*, vol. 118, no. 19, p. 191 801, 2017. DOI: `10.1103/PhysRevLett.118.191801`. arXiv: `1703.05747 [hep-ex]`.

[33] V. Khachatryan *et al.*, “Observation of the rare $B^0_s \to \mu^+\mu^-$ decay from the combined analysis of CMS and LHCb data”, *Nature*, vol. 522, pp. 68–72, 2015. DOI: `10.1038/nature14474`. arXiv: `1411.4413 [hep-ex]`.

[34] S. Chatrchyan *et al.*, “Measurement of the $B^0_s \to \mu^+\mu^-$ Branching Fraction and Search for $B^0 \to \mu^+\mu^-$ with the CMS Experiment”, *Phys. Rev. Lett.*, vol. 111, p. 101 804, 2013. DOI: `10.1103/PhysRevLett.111.101804`. arXiv: `1307.5025 [hep-ex]`.

[35] R. Aaij *et al.*, “Observation of CP violation in charm decays”, *arXiv e-prints*, arXiv:1903.08726, arXiv:1903.08726, Mar. 2019. arXiv: `1903.08726 [hep-ex]`.

[36] R. Aaij *et al.*, “Test of lepton universality with $B^0 \to K^{*0}\ell^+\ell^-$ decays”, *JHEP*, vol. 08, p. 055, 2017. DOI: `10.1007/JHEP08(2017)055`. arXiv: `1705.05802 [hep-ex]`.

[37] R. Aaij *et al.*, “Angular analysis of the $B^0 \to K^{*0}\mu^+\mu^-$ decay using 3 fb^{-1} of integrated luminosity”, *JHEP*, vol. 02, p. 104, 2016. DOI: `10.1007/JHEP02(2016)104`. arXiv: `1512.04442 [hep-ex]`.

[38] W. Altmannshofer, P. Stangl, and D. M. Straub, “Interpreting Hints for Lepton Flavor Universality Violation”, *Phys. Rev.*, vol. D96, no. 5, p. 055 008, 2017. DOI: `10.1103/PhysRevD.96.055008`. arXiv: `1704.05435 [hep-ph]`.

[39] B. Capdevila, A. Crivellin, S. Descotes-Genon, *et al.*, “Patterns of New Physics in $b \to s\ell^+\ell^-$ transitions in the light of recent data”, *JHEP*, vol. 01, p. 093, 2018. DOI: `10.1007/JHEP01(2018)093`. arXiv: `1704.05340 [hep-ph]`.

[40] T. Hurth, F. Mahmoudi, D. Martinez Santos, *et al.*, “Lepton nonuniversality in exclusive $b{\to}s\ell\ell$ decays”, *Phys. Rev.*, vol. D96, no. 9, p. 095 034, 2017. DOI: `10.1103/PhysRevD.96.095034`. arXiv: `1705.06274 [hep-ph]`.

[41] G. D’Amico, M. Nardecchia, P. Panci, *et al.*, “Flavour anomalies after the R_{K^*} measurement”, *JHEP*, vol. 09, p. 010, 2017. DOI: `10.1007/JHEP09(2017)010`. arXiv: `1704.05438 [hep-ph]`.

[42] L.-S. Geng, B. Grinstein, S. Jäger, *et al.*, “Towards the discovery of new physics with lepton-universality ratios of $b \to s\ell\ell$ decays”, *Phys. Rev.*, vol. D96, no. 9, p. 093 006, 2017. DOI: `10.1103/PhysRevD.96.093006`. arXiv: `1704.05446 [hep-ph]`.

[43] M. Ciuchini, A. M. Coutinho, M. Fedele, *et al.*, “On Flavourful Easter eggs for New Physics hunger and Lepton Flavour Universality violation”, *Eur. Phys. J.*, vol. C77, no. 10, p. 688, 2017. DOI: `10.1140/epjc/s10052-017-5270-2`. arXiv: `1704.05447 [hep-ph]`.

[44] V. G. Chobanova, T. Hurth, F. Mahmoudi, *et al.*, “Large hadronic power corrections or new physics in the rare decay $B{\to}K^*\mu^+\mu^-$?”, *JHEP*, vol. 07, p. 025, 2017. DOI: `10.1007/JHEP07(2017)025`. arXiv: `1702.02234 [hep-ph]`.

[45] G. Burdman, E. Golowich, J. L. Hewett, *et al.*, “Rare charm decays in the standard model and beyond”, *Phys. Rev.*, vol. D66, p. 014 009, 2002. DOI: `10.1103/PhysRevD.66.014009`. arXiv: `hep-ph/0112235 [hep-ph]`.

[46] R.-M. Wang, J.-H. Sheng, J. Zhu, *et al.*, “Decays $D^+_{(s)} \to \pi(K)^+\ell^+\ell^-$ and $D^0 \to \ell^+\ell^-$ in the MSSM with and without R-parity”, *Int. J. Mod. Phys.*, vol. A30, no. 12, p. 1 550 063, 2015. DOI: `10.1142/S0217751X15500633`. arXiv: `1409.0181 [hep-ph]`.

[47] S. Fajfer, S. Prelovsek, and P. Singer, “Rare charm meson decays D → P lepton+ lepton- and c → u lepton+ lepton- in SM and MSSM”, *Phys. Rev.*, vol. D64, p. 114 009, 2001. DOI: 10.1103/PhysRevD.64.114009. arXiv: hep-ph/0106333 [hep-ph].

[48] S. Fajfer, N. Kosnik, and S. Prelovsek, “Updated constraints on new physics in rare charm decays”, *Phys. Rev.*, vol. D76, p. 074 010, 2007. DOI: 10.1103/PhysRevD.76.074010. arXiv: 0706.1133 [hep-ph].

[49] C. Delaunay, J. F. Kamenik, G. Perez, *et al.*, “Charming CP Violation and Dipole Operators from RS Flavor Anarchy”, *JHEP*, vol. 01, p. 027, 2013. DOI: 10.1007/JHEP01(2013)027. arXiv: 1207.0474 [hep-ph].

[50] A. Paul, A. De La Puente, and I. I. Bigi, “Manifestations of warped extra dimension in rare charm decays and asymmetries”, *Phys. Rev.*, vol. D90, no. 1, p. 014 035, 2014. DOI: 10.1103/PhysRevD.90.014035. arXiv: 1212.4849 [hep-ph].

[51] A. Paul, I. I. Bigi, and S. Recksiegel, “On $D \to X_u l^+ l^-$ within the Standard Model and Frameworks like the Littlest Higgs Model with T Parity”, *Phys. Rev.*, vol. D83, p. 114 006, 2011. DOI: 10.1103/PhysRevD.83.114006. arXiv: 1101.6053 [hep-ph].

[52] S. Fajfer and S. Prelovsek, “Effects of littlest Higgs model in rare D meson decays”, *Phys. Rev.*, vol. D73, p. 054 026, 2006. DOI: 10.1103/PhysRevD.73.054026. arXiv: hep-ph/0511048 [hep-ph].

[53] J. P. Lees *et al.*, “Searches for Rare or Forbidden Semileptonic Charm Decays”, *Phys. Rev.*, vol. D84, p. 072 006, 2011. DOI: 10.1103/PhysRevD.84.072006. arXiv: 1107.4465 [hep-ex].

[54] P. Rubin *et al.*, “Search for rare and forbidden decays of charm and charmed-strange mesons to final states $h^+ - e^- + e^+$”, *Phys. Rev.*, vol. D82, p. 092 007, 2010. DOI: 10.1103/PhysRevD.82.092007. arXiv: 1009.1606 [hep-ex].

[55] P. L. Frabetti *et al.*, “Search for rare and forbidden decays of the charmed meson D+”, *Phys. Lett.*, vol. B398, pp. 239–244, 1997. DOI: 10.1016/S0370-2693(97)00229-3.

[56] E. M. Aitala *et al.*, “Search for rare and forbidden dilepton decays of the D+, D+(s), and D0 charmed mesons”, *Phys. Lett.*, vol. B462, pp. 401–409, 1999. DOI: 10.1016/S0370-2693(99)00902-8. arXiv: hep-ex/9906045 [hep-ex].

[57] K. Kodama *et al.*, “Upper limits of charm hadron decays to two muons plus hadrons”, *Phys. Lett.*, vol. B345, pp. 85–92, 1995. DOI: 10.1016/0370-2693(94)01610-O.

[58] J. M. Link *et al.*, “Search for rare and forbidden three body dimuon decays of the charmed mesons D+ and D(s)+”, *Phys. Lett.*, vol. B572, pp. 21–31, 2003. DOI: 10.1016/j.physletb.2003.07.079. arXiv: hep-ex/0306049 [hep-ex].

[59] V. M. Abazov *et al.*, “Search for flavor-changing-neutral-current D meson decays”, *Phys. Rev. Lett.*, vol. 100, p. 101 801, 2008. DOI: 10.1103/PhysRevLett.100.101801. arXiv: 0708.2094 [hep-ex].

[60] R. Aaij *et al.*, “Search for D+(s) to pi+ mu+ mu- and D+(s) to pi- mu+ mu+ decays”, *Phys. Lett.*, vol. B724, pp. 203–212, 2013. DOI: 10.1016/j.physletb.2013.06.010. arXiv: 1304.6365 [hep-ex].

[61] C. Burr, “Measuring open charm hadron production at 7 TeV using the LHCb detector”, Master’s thesis, Southampton U., 2015. [Online]. Available: http://inspirehep.net/record/1381292/files/CERN-THESIS-2015-074.pdf.

[62] B. Wolf, *Handbook of ion sources.* Boca Raton, Fla: CRC Press, 1995, ISBN: 9780849325021.

[63] J. Haffner, “The CERN accelerator complex”, Oct. 2013. [Online]. Available: http://cds.cern.ch/record/1621894/.

[64] LHCb Collaboration. (2011). “Nu, Mu and Pile-Up”, [Online]. Available: https://web.archive.org/web/20190319134629/https://twiki.cern.ch/twiki/bin/view/LHCb/NuMuPileUp (visited on 03/19/2019).

[65] R. Aaij *et al.*, “LHCb Detector Performance”, *Int. J. Mod. Phys.*, vol. A30, no. 07, p. 1 530 022, 2015. DOI: 10.1142/S0217751X15300227. arXiv: 1412.6352 [hep-ex].

[66] The LHCb Collaboration, “bb production angle plots”, 2008. [Online]. Available: http://lhcb.web.cern.ch/lhcb/speakersbureau/html/bb%5C_ProductionAngles.html.

[67] R. Lindner, “LHCb Detector Layout”, Feb. 2008. [Online]. Available: https://cds.cern.ch/record/1087860.

[68] A. A. Alves Jr. *et al.*, “The LHCb Detector at the LHC”, *JINST*, vol. 3, S08005, 2008. DOI: 10.1088/1748-0221/3/08/S08005.

[69] R. Aaij *et al.*, “Performance of the LHCb Vertex Locator. Performance of the LHCb Vertex Locator”, *JINST*, vol. 9, no. CERN-LHCB-DP-2014-001. CERN-LHCB-DP-2014-001. LHCB-DP-2014-001, P09007. 61 p, May 2014, Comments: 61 pages, 33 figures. [Online]. Available: https://cds.cern.ch/record/1707015.

[70] R. Aaij *et al.*, “Performance of the LHCb Vertex Locator”, *JINST*, vol. 9, p. 09 007, 2014. DOI: 10.1088/1748-0221/9/09/P09007. arXiv: 1405.7808 [physics.ins-det].

[71] R. Aaij *et al.*, “Precision measurement of the B^0_s-$\bar{B}^0_s$ oscillation frequency with the decay $B^0_s \to D^-_s \pi^+$”, *New J. Phys.*, vol. 15, p. 053 021, 2013. DOI: 10.1088/1367-2630/15/5/053021. arXiv: 1304.4741 [hep-ex].

[72] R. Arink *et al.*, “Performance of the LHCb Outer Tracker”, *JINST*, vol. 9, no. 01, P01002, 2014. DOI: 10.1088/1748-0221/9/01/P01002. arXiv: 1311.3893 [physics.ins-det].

[73] G. Knoll, *Radiation Detection and Measurement.* Wiley, 2000, ISBN: 9780471073383. [Online]. Available: `https://books.google.ch/books?id=HKBVAAAAMAAJ`.

[74] M. Adinolfi *et al.*, "Performance of the LHCb RICH detector at the LHC", *Eur. Phys. J. C*, vol. 73, no. arXiv:1211.6759. CERN-LHCb-DP-2012-003. LHCb-DP-2012-003, 2431. 25 p, Nov. 2012. [Online]. Available: `https://cds.cern.ch/record/1495721`.

[75] P. Perret, "First Years of Running for the LHCb Calorimeter System", *PoS*, vol. TIPP2014, p. 030, 2014. arXiv: `1407.4289 [physics.ins-det]`.

[76] R. Antunes-Nobrega, A. França-Barbosa, I. Bediaga, *et al.*, *LHCb reoptimized detector design and performance: Technical Design Report*, ser. Technical Design Report LHCb. Geneva: CERN, 2003. [Online]. Available: `https://cds.cern.ch/record/630827`.

[77] J. Alves A.A. *et al.*, "Performance of the LHCb muon system", *JINST*, vol. 8, no. LHCB-DP-2012-002. CERN-LHCb-DP-2012-002. LHCb-DP-2012-002, P02022. 32 p, Nov. 2012. [Online]. Available: `https://cds.cern.ch/record/1492807`.

[78] M. Hushchyn, D. Derkach, and N. Kazeev, "Machine Learning based Global Particle Identification Algorithms at the LHCb Experiment", in *CHEP2018*, 2018. [Online]. Available: `https://indico.cern.ch/event/587955/contributions/2937578/`.

[79] T. Sjöstrand, S. Mrenna, and P. Skands, "A brief introduction to PYTHIA 8.1", *Computer Physics Communications*, vol. 178, no. 11, pp. 852–867, 2008, ISSN: 0010-4655. DOI: `http://dx.doi.org/10.1016/j.cpc.2008.01.036`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0010465508000441`.

[80] I. Beiyaev, T. Brambach, N. H. Brooke, *et al.*, "Handling of the generation of primary events in Gauss, the LHCb simulation framework", in *IEEE Nuclear Science Symposuim Medical Imaging Conference*, Oct. 2010, pp. 1155–1161. DOI: `10.1109/NSSMIC.2010.5873949`.

[81] D. J. Lange, "The EvtGen particle decay simulation package", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 462, no. 1–2, pp. 152–155, 2001, BEAUTY2000, Proceedings of the 7th Int. Conf. on B-Physics at Hadron Machines, ISSN: 0168-9002. DOI: `http://dx.doi.org/10.1016/S0168-9002(01)00089-4`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0168900201000894`.

[82] P. Golonka and Z. Was, "PHOTOS Monte Carlo: A Precision tool for QED corrections in Z and W decays", *Eur. Phys. J.*, vol. C45, pp. 97–107, 2006. DOI: `10.1140/epjc/s2005-02396-4`. arXiv: `hep-ph/0506026 [hep-ph]`.

[83] S. Agostinelli, J. Allison, K. Amako, *et al.*, "Geant4—a simulation toolkit", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, pp. 250–303, 2003, ISSN: 0168-9002. DOI: `http://dx.doi.org/10.1016/S0168-9002(03)01368-8`.

[Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900203013688.

[84] M. Clemencic, G. Corti, S. Easo, *et al.*, “The LHCb Simulation Application, Gauss: Design, Evolution and Experience”, *Journal of Physics: Conference Series*, vol. 331, no. 3, p. 032 023, 2011. [Online]. Available: http://stacks.iop.org/1742-6596/331/i=3/a=032023.

[85] T. Head, “The LHCb trigger system”, *JINST*, vol. 9, p. C09015, 2014. DOI: 10.1088/1748-0221/9/09/C09015.

[86] R. Aaij *et al.*, “The LHCb Trigger and its Performance in 2011”, *JINST*, vol. 8, P04022, 2013. DOI: 10.1088/1748-0221/8/04/P04022. arXiv: 1211.3055 [hep-ex].

[87] G. Dujany and B. Storaci, “Real-time alignment and calibration of the LHCb Detector in Run II”, *J. Phys.: Conf. Ser.*, vol. 664, no. LHCb-PROC-2015-011. CERN-LHCb-PROC-2015-011, 082010. 8 p, Apr. 2015. [Online]. Available: http://cds.cern.ch/record/2017839.

[88] R. Aaij *et al.*, “Tesla : an application for real-time data analysis in High Energy Physics”, 2016. arXiv: 1604.05596 [physics.ins-det].

[89] R. Aaij *et al.*, “Measurement of forward J/ψ production cross-sections in pp collisions at $\sqrt{s} = 13$ TeV”, *JHEP*, vol. 10, p. 172, 2015. DOI: 10.1007/JHEP10(2015)172. arXiv: 1509.00771 [hep-ex].

[90] I. Bird, P. Buncic, F. Carminati, *et al.*, “Update of the Computing Models of the WLCG and the LHC Experiments”, CERN-LHCC-2014-014 LCG-TDR-002, Apr. 2014, [Online]. Available: https://cds.cern.ch/record/1695401.

[91] F. Stagni, A. Tsaregorodtsev, ubeda, *et al.*, “DIRACGrid/DIRAC: v6r20p15”, Oct. 2018. DOI: 10.5281/zenodo.1451647. [Online]. Available: https://doi.org/10.5281/zenodo.1451647.

[92] Y. Kato, K. Hayasaka, T. Hara, *et al.*, “Job monitoring on DIRAC for Belle II distributed computing”, *Journal of Physics: Conference Series*, vol. 664, no. 6, p. 062 023, 2015. [Online]. Available: http://stacks.iop.org/1742-6596/664/i=6/a=062023.

[93] S. Belov, B. Suo, Z. Deng, *et al.*, “Design and Operation of the BES-III Distributed Computing System”, *Procedia Computer Science*, vol. 66, pp. 619–624, 2015, 4th International Young Scientist Conference on Computational Science, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.11.070. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915034195.

[94] A. Sailer, M. Petric, and C. collaboration, “Using OSG Computing Resources with (iLC)Dirac”, *Journal of Physics: Conference Series*, vol. 898, no. 9, p. 092 013, 2017. [Online]. Available: http://stacks.iop.org/1742-6596/898/i=9/a=092013.

[95] K. Harrison, W. T. L. P. Lavrijsen, P. Mato, *et al.*, “GANGA: a user-Grid interface for Atlas and LHCb”, *arXiv e-prints*, cs/0306085, cs/0306085, Jun. 2003. arXiv: `cs/0306085 [cs.SE]`.

[96] F. Pérez and B. E. Granger, “IPython: a System for Interactive Scientific Computing”, *Computing in Science and Engineering*, vol. 9, no. 3, pp. 21–29, May 2007, ISSN: 1521-9615. DOI: `10.1109/MCSE.2007.53`. [Online]. Available: `https://ipython.org`.

[97] IBM. (2019). “IBM Spectrum LSF Suites”, [Online]. Available: `https://web.archive.org/web/20190104125200/https://www.ibm.com/us-en/marketplace/hpc-workload-management` (visited on 01/04/2019).

[98] T. Tannenbaum, D. Wright, K. Miller, *et al.*, “Condor – A Distributed Job Scheduler”, in *Beowulf Cluster Computing with Linux*, T. Sterling, Ed., MIT Press, Oct. 2001.

[99] M. Ellert, M. Grønager, A. Konstantinov, *et al.*, “Advanced Resource Connector Middleware for Lightweight Computational Grids”, *Future Gener. Comput. Syst.*, vol. 23, no. 2, pp. 219–240, Feb. 2007, ISSN: 0167-739X. DOI: `10.1016/j.cam.2006.05.008`. [Online]. Available: `http://dx.doi.org/10.1016/j.cam.2006.05.008`.

[100] C. Aiftimiei, P. Andreetto, S. Bertocco, *et al.*, “Design and implementation of the gLite CREAM job management service”, *Future Generation Computer Systems*, vol. 26, no. 4, pp. 654–667, 2010, ISSN: 0167-739X. DOI: `https://doi.org/10.1016/j.future.2009.12.006`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0167739X0900185X`.

[101] LHCb Collaboration, “Computing Model of the Upgrade LHCb experiment”, CERN, Geneva, Tech. Rep. CERN-LHCC-2018-014. LHCB-TDR-018, May 2018. [Online]. Available: `https://cds.cern.ch/record/2319756`.

[102] G. Barrand *et al.*, “GAUDI - A software architecture and framework for building HEP data processing applications”, *Comput. Phys. Commun.*, vol. 140, pp. 45–55, 2001. DOI: `10.1016/S0010-4655(01)00254-5`.

[103] R. Brun and F. Rademakers, “ROOT: An object oriented data analysis framework”, *Nucl. Instrum. Meth.*, vol. A389, pp. 81–86, 1997. DOI: `10.1016/S0168-9002(97)00048-X`.

[104] LHCb Collaboration, “Letter of Intent for the LHCb Upgrade”, CERN, Geneva, Tech. Rep. CERN-LHCC-2011-001. LHCC-I-018, Mar. 2011. [Online]. Available: `https://cds.cern.ch/record/1333091`.

[105] LHCb Collaboration, “LHCb Tracker Upgrade Technical Design Report”, Tech. Rep. CERN-LHCC-2014-001. LHCB-TDR-015, Feb. 2014. [Online]. Available: `https://cds.cern.ch/record/1647400`.

[106] LHCb Collaboration, “LHCb PID Upgrade Technical Design Report”, Tech. Rep. CERN-LHCC-2013-022. LHCB-TDR-014, Nov. 2013. [Online]. Available: `https://cds.cern.ch/record/1624074`.

[107] LHCb Collaboration, "LHCb VELO Upgrade Technical Design Report", Tech. Rep. CERN-LHCC-2013-021. LHCB-TDR-013, Nov. 2013. [Online]. Available: `https://cds.cern.ch/record/1624070`.

[108] T. Poikela *et al.*, "VeloPix: the pixel ASIC for the LHCb upgrade", *JINST*, vol. 10, no. 01, p. C01057, 2015. DOI: `10.1088/1748-0221/10/01/C01057`.

[109] R. R. Schaller, "Moore's Law: Past, Present, and Future", *IEEE Spectr.*, vol. 34, no. 6, pp. 52–59, Jun. 1997, ISSN: 0018-9235. DOI: `10.1109/6.591665`. [Online]. Available: `http://dx.doi.org/10.1109/6.591665`.

[110] LHCb Collaboration, "Upgrade Software and Computing", CERN, Geneva, Tech. Rep. CERN-LHCC-2018-007. LHCB-TDR-017, Mar. 2018. [Online]. Available: `https://cds.cern.ch/record/2310827`.

[111] K. Rupp. (2018). "42 Years of Microprocessor Trend Data", [Online]. Available: `https://web.archive.org/web/20181130232430/https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/` (visited on 01/06/2019).

[112] R. E. Kálmán, "A new approach to linear filtering and prediction problems" transaction of the asme journal of basic", 1960.

[113] E. Bos and E. Rodrigues, "The LHCb Track Extrapolator Tools", CERN, Geneva, Tech. Rep. LHCb-2007-140. CERN-LHCb-2007-140, Nov. 2007. [Online]. Available: `https://cds.cern.ch/record/1070314`.

[114] J. Amoraal, J. Blouw, S. Blusk, *et al.*, "Application of vertex and mass constraints in track-based alignment", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 712, pp. 48–55, 2013, ISSN: 0168-9002. DOI: `http://dx.doi.org/10.1016/j.nima.2012.11.192`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0168900213001861`.

[115] W. D. Hulsbergen, "The global covariance matrix of tracks fitted with a Kalman filter and an application in detector alignment", *Nuclear Instruments and Methods in Physics Research A*, vol. 600, pp. 471–477, Mar. 2009. DOI: `10.1016/j.nima.2008.11.094`. arXiv: `0810.2241 [physics.ins-det]`.

[116] F. James, "MINUIT Function Minimization and Error Analysis: Reference Manual Version 94.1", 1994.

[117] F. E. Curtis and X. Que, "A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees", *Mathematical Programming Computation*, vol. 7, no. 4, pp. 399–428, Dec. 2015, ISSN: 1867-2957. DOI: `10.1007/s12532-015-0086-2`. [Online]. Available: `https://doi.org/10.1007/s12532-015-0086-2`.

[118] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, *et al.*, "Equation of State Calculations by Fast Computing Machines", *J. Comput. Phys.*, vol. 21, pp. 1087–1092, Jun. 1953. DOI: `10.1063/1.1699114`.

[119] S. Borghi, C. Hombach, and C. Parkes, "AIDA Alignment package user guide", Apr. 2014. [Online]. Available: https://cds.cern.ch/record/1701341.

[120] V. Blobel and C. Kleinwort, "A New method for the high precision alignment of track detectors", in *Advanced Statistical Techniques in Particle Physics. Proceedings, Conference, Durham, UK, March 18-22, 2002*, 2002, URL–STR(9). arXiv: hep-ex/0208021 [hep-ex]. [Online]. Available: http://www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings//blobel1.pdf.

[121] V. Blobel, "Millepede II - Draft Manual", University Hamburg, Tech. Rep., 2007. [Online]. Available: https://web.archive.org/web/20190213101934/http://www.desy.de/~kleinwrt/MP2/doc/html/draftman_page.html.

[122] C. Kleinwort, "H1 alignement experience", 9 p, 2007. [Online]. Available: http://cds.cern.ch/record/1047100.

[123] R. Mankel, "Alignment of the ZEUS micro-vertex detector", 8 p, 2007. [Online]. Available: http://cds.cern.ch/record/1047101.

[124] "Alignment of the ATLAS Inner Detector Tracking System with 2010 LHC proton-proton collisions at $\sqrt{s} = 7$ TeV", CERN, Geneva, Tech. Rep. ATLAS-CONF-2011-012, Mar. 2011. [Online]. Available: https://cds.cern.ch/record/1334582.

[125] S. Chatrchyan *et al.*, "Alignment of the CMS tracker with LHC and cosmic ray data", *JINST*, vol. 9, P06009, 2014. DOI: 10.1088/1748-0221/9/06/P06009. arXiv: 1403.2286 [physics.ins-det].

[126] S. Borghi *et al.*, "First spatial alignment of the LHCb VELO and analysis of beam absorber collision data", *Nucl. Instrum. Meth.*, vol. A618, pp. 108–120, 2010. DOI: 10.1016/j.nima.2010.02.109.

[127] S. Viret, C. Parkes, and M. Gersabeck, "Alignment procedure of the LHCb Vertex Detector", *Nucl. Instrum. Meth.*, vol. A596, pp. 157–163, 2008. DOI: 10.1016/j.nima.2008.07.153. arXiv: 0807.5067 [physics.ins-det].

[128] C. Hombach, "Search for the Rare Baryonic B^+ to proton anti-Lambda Decay with the LHCb Detector and Alignment of Pixel Detectors", Presented 20 May 2016, Mar. 2016. [Online]. Available: https://cds.cern.ch/record/2284934.

[129] A. Gerbershagen, "Overview over CERN SPS secondary beams", in *6th Beam Telescopes and Test Beams Workshop 2018*, 2018. [Online]. Available: https://indico.desy.de/indico/event/18050/session/7/contribution/7.

[130] K. Akiba *et al.*, "LHCb VELO Timepix3 Telescope", 2019. arXiv: 1902.09755 [physics.ins-det].

[131] T. Poikela, J. Plosila, T. Westerlund, *et al.*, "Timepix3: a 65K channel hybrid pixel readout chip with simultaneous ToA/ToT and sparse readout", *Journal of Instrumentation*, vol. 9, no. 05, pp. C05013–C05013, May 2014. DOI: 10.1088/1748-0221/9/05/c05013. [Online]. Available: https://doi.org/10.1088%2F1748-0221%2F9%2F05%2Fc05013.

[132] J. Visser, M. van Beuzekom, H. Boterenbrood, *et al.*, "SPIDR: a read-out system for Medipix3 & Timepix3", *Journal of Instrumentation*, vol. 10, no. 12, pp. C12028–C12028, Dec. 2015. DOI: `10.1088/1748-0221/10/12/c12028`. [Online]. Available: `https://doi.org/10.1088%2F1748-0221%2F10%2F12%2Fc12028`.

[133] A. Peters, E. Sindrilaru, and G. Adde, "EOS as the present and future solution for data storage at CERN", *J. Phys. Conf. Ser.*, vol. 664, no. 4, p. 042 042, 2015. DOI: `10.1088/1742-6596/664/4/042042`.

[134] The LHCb Collaboration, *Kepler*, version v3r0, Nov. 24, 2016. [Online]. Available: `https://gitlab.cern.ch/lhcb/Kepler`.

[135] Pallets team, *Flask*, version 0.11, May 29, 2016. [Online]. Available: `http://flask.pocoo.org/`.

[136] S. Ritt, *ELOG*, version 3.1.3, Apr. 21, 2017. [Online]. Available: `https://midas.psi.ch/elog`.

[137] E. Buchanan, "Spatial Resolution Studies for the LHCb VELO Upgrade", Presented 31 Oct 2018, Oct. 2018. [Online]. Available: `https://cds.cern.ch/record/2653356`.

[138] *LHCb VELO Upgrade Technical Design Report*, LHCb-TDR-013, Geneva, 2013.

[139] S. D. Capua, "Velo Upgrade Module", 83rd LHCb week, 2017, [Online]. Available: `https://indico.cern.ch/event/616118/`.

[140] University of Manchester, "VELO Upgrade Mechanical Module EDR", 2017. [Online]. Available: `https://indico.cern.ch/event/657986/`.

[141] C. Burr, C. Parkes, S. Borghi, *et al.*, "How sensitive are physics measurements to VELO upgrade misalignment?", CERN, Geneva, Tech. Rep. LHCb-INT-2017-022. CERN-LHCb-INT-2017-022, Sep. 2017. [Online]. Available: `https://cds.cern.ch/record/2283202`.

[142] S. Ponce, "Detector Description Framework in LHCb", CERN, Geneva, Tech. Rep. LHCb-PROC-2003-004. CERN-LHCb-PROC-2003-004, Mar. 2003. [Online]. Available: `https://cds.cern.ch/record/1496875`.

[143] M. Gersabeck, J. Nardulli, and E. Rodrigues, "Impact of misalignments on the analysis of B decays", CERN, Geneva, Tech. Rep. LHCb-2008-012. CERN-LHCb-2008-012, Aug. 2008. [Online]. Available: `https://cds.cern.ch/record/1119088`.

[144] The LHCb Collaboration, *Brunel*, version v51r1, Sep. 20, 2016. [Online]. Available: `https://gitlab.cern.ch/lhcb/Brunel`.

[145] The LHCb Collaboration, *Panoramix*, version v23r2p2, Oct. 7, 2016. [Online]. Available: `https://gitlab.cern.ch/lhcb/Panoramix`.

[146] J. Amoraal *et al.*, "Application of vertex and mass constraints in track-based alignment", *Nucl. Instrum. Meth.*, vol. A712, pp. 48–55, 2013. DOI: `10.1016/j.nima.2012.11.192`. arXiv: `1207.4756 [physics.ins-det]`.

[147] R. Aaij *et al.*, "Precision measurement of the B_s^0–$\overline{B}_s^0$ oscillation frequency in the decay $B_s^0 \to D_s^- \pi^+$", *New J. Phys.*, vol. 15, p. 053 021, 2013. DOI: 10.1088/1367-2630/15/5/053021. arXiv: 1304.4741 [hep-ex].

[148] S. Hansmann-Menzemer, G. Krocker, and S. Wandernoth, "Measurement of Δm_s in the decay $B_S^0 \to D_s^- \pi^+$ using 1fb^{-1}", Apr. 2014. [Online]. Available: https://cds.cern.ch/record/1445564.

[149] S. Borghi and W. Hulsbergen, "Systematic uncertainties due to Velo length scale", 2017, [Online]. Available: https://indico.cern.ch/event/605176/contributions/2554522/.

[150] C. Patrignani *et al.*, "Review of particle physics", *Chin. Phys.*, vol. C40, p. 100 001, 2016. DOI: 10.1088/1674-1137/40/10/100001.

[151] M. Williams, "Velo Upgrade Module Simulation", 2016, [Online]. Available: https://indico.cern.ch/event/590985/contributions/2391922/.

[152] G. Zunica, "Alignment of the VELO detector on the Z axis using secondary vertexes reconstruction", Sep. 2018. [Online]. Available: https://cds.cern.ch/record/2640281.

[153] R. Aaij *et al.*, "Test of lepton universality using $B^+ \to K^+ \ell^+ \ell^-$ decays", *Phys. Rev. Lett.*, vol. 113, p. 151 601, 2014. DOI: 10.1103/PhysRevLett.113.151601. arXiv: 1406.6482 [hep-ex].

[154] D. Wang, "Searches for Flavor Changing Neutral Currents at BESIII", ICHEP2018, 2018, [Online]. Available: https://indico.cern.ch/event/686555/contributions/2982771/.

[155] R. Aaij *et al.*, "Search for $D^+_{(s)} \to \pi^+ \mu^+ \mu^-$ and $D^+_{(s)} \to \pi^- \mu^+ \mu^+$ decays", *Phys. Lett.*, vol. B724, p. 203, 2013. DOI: 10.1016/j.physletb.2013.06.010. arXiv: 1304.6365 [hep-ex].

[156] A. Rogozhnikov, L. Tatiana, A. Ustyuzhanin, *et al.*, "arogozhnikov/hep_ml: Bump release for DOI", May 2018. DOI: 10.5281/zenodo.1247391. [Online]. Available: https://doi.org/10.5281/zenodo.1247391.

[157] J. R. Klein and A. Roodman, "Blind Analysis In Nuclear And Particle Physics", *Annual Review of Nuclear and Particle Science*, vol. 55, no. 1, pp. 141–163, 2005. DOI: 10.1146/annurev.nucl.55.090704.151521. eprint: https://doi.org/10.1146/annurev.nucl.55.090704.151521. [Online]. Available: https://doi.org/10.1146/annurev.nucl.55.090704.151521.

[158] "Data from Table 3 from: Measurement of charged particle multiplicities and densities in *pp* collisions at $\sqrt{s} = 7$ TeV in the forward region", DOI: 10.17182/hepdata.63498.v1/t3.

[159] R. Aaij *et al.*, "Performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC", 2018. arXiv: 1812.10790 [hep-ex].

[160] M. De Cian, S. Farry, P. Seyfert, *et al.*, "Fast neural-net based fake track rejection in the LHCb reconstruction", CERN, Geneva, Tech. Rep. LHCb-PUB-2017-011. CERN-LHCb-PUB-2017-011, Mar. 2017. [Online]. Available: `http://cds.cern.ch/record/2255039`.

[161] A. Gulin, I. Kuralenok, and D. Pavlov, "Winning The Transfer Learning Track of Yahoo!'s Learning To Rank Challenge with YetiRank", in *Proceedings of the Learning to Rank Challenge*, O. Chapelle, Y. Chang, and T.-Y. Liu, Eds., ser. Proceedings of Machine Learning Research, vol. 14, Haifa, Israel: PMLR, Jun. 2011, pp. 63–76. [Online]. Available: `http://proceedings.mlr.press/v14/gulin11a.html`.

[162] O. Lupton. (2018). "`AALLSAMEBPV LoKi` bug", [Online]. Available: `https://indico.cern.ch/event/715191/contributions/2960112/` (visited on 04/10/2018).

[163] J. P. Baud, P. Charpentier, K. Ciba, *et al.*, "The LHCb Data Management System", *Journal of Physics: Conference Series*, vol. 396, no. 3, p. 032 023, 2012. [Online]. Available: `http://stacks.iop.org/1742-6596/396/i=3/a=032023`.

[164] J. Beringer *et al.*, "Review of particle physics", *Phys. Rev.*, vol. D86, p. 010 001, 2012, and 2013 partial update for the 2014 edition. DOI: `10.1103/PhysRevD.86.010001`.

[165] S. Beranek, S. Bifani, A. Davis, *et al.*, "Measurement of the electron detection and reconstruction efficiency", https://twiki.cern.ch/twiki/bin/view/LHCb/Velo2longElectronTrackingEfficiency, 2018.

[166] M. Kearns and L. G. Valiant, "Learning Boolean Formulae or Finite Automata is as Hard as Factoring", Harvard University Aiken Computation Laboratory, Tech. Rep. TR 14-88, 1988.

[167] R. E. Schapire, "The strength of weak learnability", *Machine Learning*, vol. 5, no. 2, pp. 197–227, Jun. 1990, ISSN: 1573-0565. DOI: `10.1007/BF00116037`. [Online]. Available: `https://doi.org/10.1007/BF00116037`.

[168] Y. Freund, "Boosting a Weak Learning Algorithm by Majority", *Information and Computation*, vol. 121, no. 2, pp. 256–285, 1995, ISSN: 0890-5401. DOI: `https://doi.org/10.1006/inco.1995.1136`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0890540185711364`.

[169] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *J. Comput. Syst. Sci.*, vol. 55, p. 119, 1997. DOI: `10.1006/jcss.1997.1504`.

[170] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, San Francisco, California, USA: ACM, 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: `10.1145/2939672.2939785`. [Online]. Available: `http://doi.acm.org/10.1145/2939672.2939785`.

[171] A. Veronika Dorogush, V. Ershov, and A. Gulin, “CatBoost: gradient boosting with categorical features support”, *arXiv e-prints*, arXiv:1810.11363, arXiv:1810.11363, Oct. 2018. arXiv: `1810.11363 [cs.LG]`.

[172] J. Friedman, T. Hastie, and R. Tibshirani, “Additive Logistic Regression: a Statistical View of Boosting”, *Annals of Statistics*, vol. 28, p. 2000, 1998.

[173] L. Mason, J. Baxter, P. Bartlett, *et al.*, “Boosting Algorithms as Gradient Descent”, in *In Advances in Neural Information Processing Systems 12*, MIT Press, 2000, pp. 512–518.

[174] R. Aaij *et al.*, “Observation of *CP* violation in $B^{\pm} \to DK^{\pm}$ decays”, *Phys. Lett.*, vol. B712, p. 203, 2012. DOI: `10.1016/j.physletb.2012.04.060`. arXiv: `1203.3662 [hep-ex]`.

[175] xgboost developers. (2016). “XGBoost Parameters”, [Online]. Available: `https://web.archive.org/web/20190303173229/https://xgboost.readthedocs.io/en/latest/parameter.html` (visited on 03/03/2019).

[176] M. Nielsen. (2017). “Improving the way neural networks learn”, [Online]. Available: `http://neuralnetworksanddeeplearning.com/chap3.html` (visited on 07/17/2018).

[177] M. Perrin-Terrin and G. Mancinelli, “Optimisation of the binning of the discriminating variables used in the computation of $\mathcal{B}(B^0_s \to \mu^+\mu^-)$ upper limits with the modified frequentist approach”, CERN, Geneva, Tech. Rep. LHCb-INT-2012-003. CERN-LHCb-INT-2012-003, Feb. 2012. [Online]. Available: `https://cds.cern.ch/record/1419784`.

[178] S. Bachmann, A. Davis, A. Di Canto, *et al.*, “Measurement of charm mixing and CP violation parameters from wrong-sign $D^{*+} \to D^0(\to K^+\pi^-)\pi^+$ decays”, Sep. 2013. [Online]. Available: `https://cds.cern.ch/record/1504387`.

[179] G. Punzi, “Sensitivity of searches for new signals and its optimization”, *eConf*, vol. C030908, MODT002, 2003, [,79(2003)]. arXiv: `physics/0308063 [physics]`.

[180] R. J. Barlow, “Extended maximum likelihood”, *Nucl. Instrum. Meth.*, vol. A297, pp. 496–506, 1990. DOI: `10.1016/0168-9002(90)91334-8`.

[181] IEEE Computer Society Standards Committee. Working group of the Microprocessor Standards Subcommittee and American National Standards Institute, *IEEE standard for binary floating-point arithmetic*, ser. ANSI/IEEE Std 754-1985. 1109 Spring Street, Suite 300, Silver Spring, MD 20910, USA: IEEE Computer Society Press, 1985, p. 18.

[182] K. S. Cranmer, “Kernel estimation in high-energy physics”, *Comput. Phys. Commun.*, vol. 136, pp. 198–207, 2001. DOI: `10.1016/S0010-4655(00)00243-5`. arXiv: `hep-ex/0011057 [hep-ex]`.

[183] F. E. James, “Monte Carlo phase space”, CERN, Geneva, 1 May 1968, CERN, Geneva: CERN, 1968, 41 p. [Online]. Available: `https://cds.cern.ch/record/275743`.

[184] R. Aaij *et al.*, “Measurements of prompt charm production cross-sections in *pp* collisions at $\sqrt{s} = 13\,\text{TeV}$”, *JHEP*, vol. 03, p. 159, 2016. DOI: `10.1007/JHEP03(2016)159`. arXiv: `1510.01707 [hep-ex]`.

[185] G. A. Cowan, D. C. Craik, and M. D. Needham, “RapidSim: an application for the fast simulation of heavy-quark hadron decays”, *Comput. Phys. Commun.*, vol. 214, pp. 239–246, 2017. DOI: `10.1016/j.cpc.2017.01.029`. arXiv: `1612.07489 [hep-ex]`.

[186] M. Cacciari, M. Greco, and P. Nason, “The P(T) spectrum in heavy flavor hadroproduction”, *JHEP*, vol. 05, p. 007, 1998. DOI: `10.1088/1126-6708/1998/05/007`. arXiv: `hep-ph/9803400 [hep-ph]`.

[187] M. Cacciari, S. Frixione, and P. Nason, “The p(T) spectrum in heavy flavor photoproduction”, *JHEP*, vol. 03, p. 006, 2001. DOI: `10.1088/1126-6708/2001/03/006`. arXiv: `hep-ph/0102134 [hep-ph]`.

[188] A. Rogozhnikov, “Reweighting with Boosted Decision Trees”, *J. Phys. Conf. Ser.*, vol. 762, no. 1, p. 012 036, 2016. DOI: `10.1088/1742-6596/762/1/012036`. arXiv: `1608.05806 [physics.data-an]`.

[189] M. Pivk and F. R. Le Diberder, “SPlot: A Statistical tool to unfold data distributions”, *Nucl. Instrum. Meth.*, vol. A555, pp. 356–369, 2005. DOI: `10.1016/j.nima.2005.08.106`. arXiv: `physics/0402083 [physics.data-an]`.

[190] A. Poluektov, “Correction of simulated particle identification response in LHCb using transformation of variables”, CERN, Geneva, Tech. Rep. LHCb-INT-2017-007. CERN-LHCb-INT-2017-007, Apr. 2017. [Online]. Available: `https://cds.cern.ch/record/2260281`.

[191] A. Poluektov, “Kernel density estimation of a multidimensional efficiency profile”, *JINST*, vol. 10, no. 02, P02011, 2015. DOI: `10.1088/1748-0221/10/02/P02011`. arXiv: `1411.5528 [physics.data-an]`.

[192] D. Müller, M. Clemencic, G. Corti, *et al.*, “ReDecay: A novel approach to speed up the simulation at LHCb”, *Eur. Phys. J.*, vol. C78, no. 12, p. 1009, 2018. DOI: `10.1140/epjc/s10052-018-6469-6`. arXiv: `1810.10362 [hep-ex]`.

[193] A. L. Read, “Presentation of search results: The CL(s) technique”, *J. Phys.*, vol. G28, pp. 2693–2704, 2002, [,11(2002)]. DOI: `10.1088/0954-3899/28/10/313`.

[194] G. Cowan, K. Cranmer, E. Gross, *et al.*, “Asymptotic formulae for likelihood-based tests of new physics”, *Eur. Phys. J.*, vol. C71, p. 1554, 2011, [Erratum: Eur. Phys. J.C73,2501(2013)]. DOI: `10.1140/epjc/s10052-011-1554-0,10.1140/epjc/s10052-013-2501-z`. arXiv: `1007.1727 [physics.data-an]`.

[195] R. Aaij *et al.*, “Measurement of the track reconstruction efficiency at LHCb”, *JINST*, vol. 10, no. 02, P02007, 2015. DOI: 10.1088/1748-0221/10/02/P02007. arXiv: 1408.1251 [hep-ex].

[196] LHCb Collaboration. (2017). “2016 tracking efficiency tables”, [Online]. Available: https://twiki.cern.ch/twiki/bin/view/LHCbInternal/TrackingEffStatus2016_Sim09b?rev=5 (visited on 02/07/2018).

[197] A. Lee, “https://github.com/tisimst/mcerp”, 2014. [Online]. Available: https://github.com/tisimst/mcerp.

[198] M. Chernick, *Bootstrap methods : a guide for practitioners and researchers*. Hoboken, N.J: Wiley-Interscience, 2008, ISBN: 9780470192573.

[199] A. Morris, “https://github.com/abmorris/StandardHypoTestInverter”, 2018. [Online]. Available: https://github.com/abmorris/StandardHypoTestInverter.

[200] C. Burr. (2017). “Add support for xrootd backend and using remote data without a local copy”, [Online]. Available: https://web.archive.org/web/20190121080516/https://bitbucket.org/snakemake/snakemake/pull-requests/195 (visited on 01/21/2019).

[201] C. Burr, I. Babuschkin, D. Remenska, *et al.*, “scikit-hep/root_pandas v0.6.1”, Oct. 2018. DOI: 10.5281/zenodo.1469122. [Online]. Available: https://doi.org/10.5281/zenodo.1469122.

[202] Anaconda Inc., “Conda 4.5.12”, Dec. 2018. [Online]. Available: https://conda.io/.

[203] E. Dolstra, “Nix: The Purely Functional Package Manager”, Accessed on: 01/12/2018, [Online]. Available: https://nixos.org/nix/.

[204] W. Barter, C. Burr, and C. Parkes, “Calculating p-values and their significances with the Energy Test for large datasets”, *JINST*, vol. 13, no. 04, P04011, 2018. DOI: 10.1088/1748-0221/13/04/P04011. arXiv: 1801.05222 [physics.data-an].

[205] Z. Akopov *et al.*, “Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics”, 2012. arXiv: 1205.4667 [hep-ex].

[206] Research Nature, “Data availability statements and data citations policy”, Sep. 2016. [Online]. Available: https://web.archive.org/web/20190121202204/https://www.nature.com/documents/nr-data-availability-statements-data-citations.pdf.

[207] ALICE Collaboration, “ALICE data preservation strategy”, Tech. Rep., Oct. 2016. DOI: 10.7483/OPENDATA.ALICE.54NE.X2EA. [Online]. Available: http://doi.org/10.7483/OPENDATA.ALICE.54NE.X2EA.

[208] ATLAS Collaboration, “ATLAS Data Access Policy”, Tech. Rep., May 2014. DOI: 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ. [Online]. Available: http://doi.org/10.7483/OPENDATA.ATLAS.T9YR.Y7MZ.

[209] CMS Collaboration, "2018 CMS data preservation, re-use and open access policy", Tech. Rep., Apr. 2018. DOI: 10.7483/OPENDATA.CMS.7347.JDWH. [Online]. Available: http://doi.org/10.7483/OPENDATA.CMS.7347.JDWH.

[210] LHCb Collaboration, "LHCb External Data Access Policy", Tech. Rep., Apr. 2013. DOI: 10.7483/OPENDATA.LHCb.HKJW.TWSZ. [Online]. Available: http://doi.org/10.7483/OPENDATA.LHCb.HKJW.TWSZ.

[211] ATLAS Collaboration. (2018). "Software license for Athena", [Online]. Available: https://web.archive.org/web/20190121212906/https://gitlab.cern.ch/atlas/athena/blob/master/LICENSE (visited on 01/21/2019).

[212] ALICE Collaboration. (2016). "Software license for AliPhysics", [Online]. Available: https://web.archive.org/web/20190121213108/https://github.com/alisw/AliPhysics/blob/master/LICENSE (visited on 01/21/2019).

[213] CMS Collaboration. (2017). "Software license for CMSSW", [Online]. Available: https://web.archive.org/web/20190127133516/https://github.com/cms-sw/cmssw/blob/master/LICENSE (visited on 01/27/2019).

[214] LHCb Collaboration. (2018). "Software license for DaVinci", [Online]. Available: https://web.archive.org/web/20190121213158/https://gitlab.cern.ch/lhcb/DaVinci/blob/master/COPYING (visited on 01/21/2019).

[215] A. Larkoski, S. Marzani, J. Thaler, *et al.*, "Exposing the QCD Splitting Function with CMS Open Data", *Phys. Rev. Lett.*, vol. 119, p. 132 003, 13 Sep. 2017. DOI: 10.1103/PhysRevLett.119.132003. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.119.132003.

[216] T. Head. (2015). "scikit-learn for TMVA Users", [Online]. Available: https://web.archive.org/web/20190127140036/https://betatim.github.io/posts/sklearn-for-TMVA-users/ (visited on 01/27/2019).

[217] Unknown. (2018). "Computational Data Analysis Workflow Systems", [Online]. Available: https://web.archive.org/web/20190127135914/https://root.cern.ch/doc/master/df102__NanoAODDimuonAnalysis_8C.html (visited on 01/27/2019).

[218] M. Mattson *et al.*, "First Observation of the Doubly Charmed Baryon Ξ_{cc}^{+}", *Phys. Rev. Lett.*, vol. 89, p. 112 001, 2002. DOI: 10.1103/PhysRevLett.89.112001. arXiv: hep-ex/0208014 [hep-ex].

[219] A. Ocherashvili *et al.*, "Confirmation of the double charm baryon Xi+(cc)(3520) via its decay to p D+ K-", *Phys. Lett.*, vol. B628, pp. 18–24, 2005. DOI: 10.1016/j.physletb.2005.09.043. arXiv: hep-ex/0406033 [hep-ex].

[220] B. Aubert *et al.*, "Search for doubly charmed baryons Xi(cc)+ and Xi(cc)++ in BABAR", *Phys. Rev.*, vol. D74, p. 011 103, 2006. DOI: 10.1103/PhysRevD.74.011103. arXiv: hep-ex/0605075 [hep-ex].

[221] R. Chistov *et al.*, "Observation of new states decaying into Lambda(c)+ K- pi+ and Lambda(c)+ K0(S) pi-", *Phys. Rev. Lett.*, vol. 97, p. 162 001, 2006. DOI: 10.1103/PhysRevLett.97.162001. arXiv: hep-ex/0606051 [hep-ex].

[222] Y. Kato *et al.*, "Search for doubly charmed baryons and study of charmed strange baryons at Belle", *Phys. Rev.*, vol. D89, no. 5, p. 052 003, 2014. DOI: 10.1103/PhysRevD.89.052003. arXiv: 1312.1026 [hep-ex].

[223] R. Aaij *et al.*, "Search for the doubly charmed baryon Ξ_{cc}^{+}", *JHEP*, vol. 12, p. 090, 2013. DOI: 10.1007/JHEP12(2013)090. arXiv: 1310.2538 [hep-ex].

[224] CERN Computer Security Office. (2014). "Computer Accounts Rules", [Online]. Available: https://web.archive.org/web/20190122151728/https://security.web.cern.ch/security/rules/en/accounts.shtml (visited on 01/22/2019).

[225] T. Bell, L. Canali, E. Grancher, *et al.*, "Web-based Analysis Services Report", CERN, Geneva, Tech. Rep. CERN-IT-Note-2018-004, Nov. 2017, Editor: Massimo Lamanna, CERN IT. [Online]. Available: https://cds.cern.ch/record/2315331.

[226] S. Neubert, A. Trisovic, A. Pearce, *et al.*, "LHCb Analysis Preservation Roadmap", CERN, Geneva, Tech. Rep. LHCb-INT-2017-021. CERN-LHCb-INT-2017-021, Aug. 2017. [Online]. Available: https://cds.cern.ch/record/2280615.

[227] Docker Inc., "Enterprise container platform", Accessed on: 01/12/2018, [Online]. Available: https://www.docker.com/.

[228] Crunchbase Inc. (2018). "Series E - Docker", [Online]. Available: https://web.archive.org/web/20190127145121/https://www.crunchbase.com/funding_round/docker-series-e--a283f818 (visited on 01/27/2019).

[229] P. Ewels, S. Fillinger, A. Peltzer, *et al.*, "Containerized Genomic Workflows with Singularity", Nov. 2018. [Online]. Available: https://www.sylabs.io/wp-content/uploads/2018/11/Singularity-NextFlow-Whitepaper-ebook-Nov2018.pdf.

[230] Anaconda Inc., "Anaconda Software Distribution", Nov. 2016. [Online]. Available: https://anaconda.com.

[231] conda-forge, "A community led collection of recipes, build infrastructure and distributions for the conda package manager.", Dec. 2018. [Online]. Available: https://conda-forge.org.

[232] B. Grüning, R. Dale, A. Sjödin, *et al.*, "Bioconda: A sustainable and comprehensive software distribution for the life sciences", *bioRxiv*, 2017. DOI: 10.1101/207092. eprint: https://www.biorxiv.org/content/early/2017/10/27/207092.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2017/10/27/207092.

[233] Space Telescope Science Institute, “Conda channel providing tools and utilities to process and analyze data from the Hubble Space Telescope, James Webb Space Telescope, and others.”, Dec. 2018. [Online]. Available: https://astroconda.readthedocs.io/.

[234] D. Remenska, C. Tunnell, J. Aalbers, *et al.*, “Giving pandas ROOT to chew on: experiences with the XENON1T Dark Matter experiment”, *Journal of Physics: Conference Series*, vol. 898, p. 042 003, Oct. 2017. DOI: 10.1088/1742-6596/898/4/042003. [Online]. Available: https://doi.org/10.1088%2F1742-6596%2F898%2F4%2F042003.

[235] Unknown. (2018). “Computational Data Analysis Workflow Systems”, [Online]. Available: https://web.archive.org/web/20190122154206/https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems (visited on 01/22/2019).

[236] R. Aaij *et al.*, “Measurement of CP violation parameters and polarisation fractions in $\mathrm{B}_s^0 \to \mathrm{J}/\psi\overline{\mathrm{K}}^{*0}$ decays”, *JHEP*, vol. 11, p. 082, 2015. DOI: 10.1007/JHEP11(2015)082. arXiv: 1509.00400 [hep-ex].

[237] J. Köster and S. Rahmann, “Snakemake—a scalable bioinformatics workflow engine”, *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, 2012. DOI: 10.1093/bioinformatics/bts480. eprint: /oup/backfile/content_public/journal/bioinformatics/28/19/10.1093/bioinformatics/bts480/2/bts480.pdf. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts480.

[238] E. Dolstra, “The Purely Functional Software Deployment Model”, PhD thesis, Universiteit Utrecht, 2006. [Online]. Available: https://nixos.org/~eelco/pubs/phd-thesis.pdf.

[239] C. Kauhaus, “Migrating a Hosting Infrastructure from Gentoo”, in *Nixcon*, 2018. [Online]. Available: https://www.youtube.com/watch?v=5GtOAaqqNGU.

[240] P.-A. Bouttier, “Nix as HPC package management system”, in *Nixcon*, 2018. [Online]. Available: https://www.youtube.com/watch?v=s5iY3CsdSfQ.

[241] B. Oldeman, “Combining CVMFS, Nix, Lmod, and EasyBuild at Compute Canada”, in *FOSDEM*, 2018. [Online]. Available: https://archive.fosdem.org/2018/schedule/event/computecanada/.

[242] D. Barlow, “NixWRT: purely functional firmware images for IoT”, in *Nixcon*, 2018. [Online]. Available: https://www.youtube.com/watch?v=0K1qn60X2HI.

[243] J. Thalheim, “About Nix sandboxes and breakpoints”, in *Nixcon*, 2018. [Online]. Available: https://www.youtube.com/watch?v=ULqoCjANK-I.

[244] H. Levsen and C. Lamb, “Reproducible Buster and beyond”, in *DebConf18*, 2018. [Online]. Available: https://debconf18.debconf.org/talks/80-reproducible-buster-and-beyond/.

[245] E. Dolstra and the Nixpkgs/NixOS contributors, “The Nix Packages collection”, Accessed on: 01/12/2018, [Online]. Available: `https://github.com/NixOS/nixpkgs`.

[246] E. Dolstra and E. Visser, “Hydra: A Declarative Approach to Continuous Integration”, Dec. 2018. [Online]. Available: `https://nixos.org/~eelco/pubs/hydra-scp-submitted.pdf`.

[247] H. S. Foundation, “HSF Packaging Working Group”, Accessed on: 01/12/2018, [Online]. Available: `https://github.com/HSF/packaging`.

[248] J. Blomer, “Decentralized Data Storage and Processing in the Context of the LHC Experiments at CERN”, PhD thesis, Technical University of Munich, 2012. [Online]. Available: `https://cdsweb.cern.ch/record/1462821/`.

[249] B. Hegner, “HSF Platform Naming Conventions - A Proposal”, Jun. 2018. [Online]. Available: `https://hepsoftwarefoundation.org/notes/HSF-TN-2018-01.pdf`.

[250] Docker Inc., “Docker Image Specification `v1.2.0`”, Accessed on: 01/12/2018, [Online]. Available: `https://github.com/moby/moby/blob/master/image/spec/v1.2.md#docker-image-specification-v120`.

[251] G. Christensen, “Optimising Docker Layers for Better Caching with Nix”, Accessed on: 01/12/2018, [Online]. Available: `https://grahamc.com/blog/nix-and-layered-docker-images`.

[252] C. Burr, “Nix for software deployment in high energy physics”, in *Nixcon*, 2018. [Online]. Available: `https://www.youtube.com/watch?v=Ee8k97Rx3DA`.

[253] C. Bozzi, “LHCb Computing Resource usage in 2017”, CERN, Geneva, Tech. Rep. LHCb-PUB-2018-002. CERN-LHCb-PUB-2018-002, Mar. 2018. [Online]. Available: `https://cds.cern.ch/record/2307500`.

[254] A. McNab, “Resources (plans): multiprocessor resources”, in *10th LHCb Computing Workshop*, 2017. [Online]. Available: `https://indico.cern.ch/event/561982/contributions/2786937/`.

[255] A. McNab, “GPU support in LHCb DIRAC”, in *pre-GDB on GPU utilisation*, 2018. [Online]. Available: `https://indico.cern.ch/event/689511/contributions/2882019/`.

[256] Microsoft. (2018). “Service Level Agreement”, [Online]. Available: `https://web.archive.org/web/20190120175411/https://azure.microsoft.com/en-gb/support/legal/sla/virtual-machines/v1_8/` (visited on 01/20/2019).

[257] D. Ocean. (2018). “Service Level Agreement”, [Online]. Available: `https://web.archive.org/web/20190120175518/https://www.digitalocean.com/docs/platform/droplet-policies/` (visited on 01/20/2019).

[258] Google. (2018). “Service Level Agreement”, [Online]. Available: `https://web.archive.org/web/20190120175600/https://cloud.google.com/compute/sla` (visited on 01/20/2019).

[259] Amazon. (2018). “Service Level Agreement”, [Online]. Available: https://web.archive.org/web/20190120175636/https://aws.amazon.com/compute/sla/ (visited on 01/20/2019).

[260] Oracle. (2018). “Service Level Agreement”, [Online]. Available: https://web.archive.org/web/20190120175819/https://cloud.oracle.com/en_US/iaas/sla (visited on 01/20/2019).

[261] WLCG. (2018). “Memorandum of Understanding”, [Online]. Available: https://web.archive.org/web/20190120172111/http://wlcg.web.cern.ch/collaboration/mou (visited on 01/20/2019).

[262] M. Zimmermann, “The ALICE analysis train system”, *J. Phys. Conf. Ser.*, vol. 608, no. 1, p. 012 019, 2015. DOI: 10.1088/1742-6596/608/1/012019. arXiv: 1502.06381 [hep-ex].

[263] C. Pinkenburg, “Analyzing ever growing datasets in PHENIX”, *J. Phys. Conf. Ser.*, vol. 331, p. 072 027, 2011. DOI: 10.1088/1742-6596/331/7/072027.

[264] LHCb Starterkit. (2015). “First Starterkit”, [Online]. Available: https://web.archive.org/web/20190125115057/https://lhcb.github.io/starterkit/archive/starterkit/2015/06/02/starterkit1.html (visited on 01/25/2019).

[265] A. Puig Navarro, “The LHCb Starterkit initiative”, *PoS*, vol. EPS-HEP2017, p. 565, 2017. DOI: 10.22323/1.314.0565.

[266] G. Wilson. (Jan. 28, 2016). “Software Carpentry: Lessons Learned”. version 2, [Online]. Available: http://f1000research.com/articles/3-62/v2.

[267] J. Rademacker and B. Sciascia, “The early career, gender, and diversity actions within the LHCb Collaboration”, *PoS*, vol. ICHEP2016, p. 318, 2017. DOI: 10.22323/1.282.0318.

[268] A. L. Read, “Presentation of search results: The CL(s) technique”, *J. Phys.*, vol. G28, pp. 2693–2704, 2002, [,11(2002)]. DOI: 10.1088/0954-3899/28/10/313.

[269] C. Parkes, S. Chen, J. Brodzicka, *et al.*, “On model-independent searches for direct CP violation in multi-body decays”, *J. Phys.*, vol. G44, no. 8, p. 085 001, 2017. DOI: 10.1088/1361-6471/aa75a5. arXiv: 1612.04705 [hep-ex].

[270] B. Aslan and G. Zech, “New test for the multivariate two-sample problem based on the concept of minimum energy”, *Journal of Statistical Computation and Simulation*, vol. 75, no. 2, pp. 109–119, 2005. DOI: 10.1080/00949650410001661440. eprint: https://doi.org/10.1080/00949650410001661440. [Online]. Available: https://doi.org/10.1080/00949650410001661440.

[271] B. Aslan and G. Zech, “Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding”, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 537, no. 3, pp. 626–636, 2005, ISSN: 0168-9002. DOI:

https://doi.org/10.1016/j.nima.2004.08.071. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900204019576.

[272] M. Williams, "Observing CP Violation in Many-Body Decays", *Phys. Rev.*, vol. D84, p. 054 015, 2011. DOI: 10.1103/PhysRevD.84.054015. arXiv: 1105.5338 [hep-ex].

[273] R. Aaij *et al.*, "Search for CP violation in $D^0 \to \pi^-\pi^+\pi^0$ decays with the energy test", *Phys. Lett.*, vol. B740, pp. 158–167, 2015. DOI: 10.1016/j.physletb.2014.11.043. arXiv: 1410.4170 [hep-ex].

[274] R. Aaij *et al.*, "Search for CP violation in the phase space of $D^0 \to \pi^+\pi^-\pi^+\pi^-$ decays", *Phys. Lett.*, vol. B769, pp. 345–356, 2017. DOI: 10.1016/j.physletb.2017.03.062. arXiv: 1612.03207 [hep-ex].

[275] R. Durstenfeld, "Algorithm 235: Random Permutation", *Commun. ACM*, vol. 7, no. 7, pp. 420–, Jul. 1964, ISSN: 0001-0782. DOI: 10.1145/364520.364540. [Online]. Available: http://doi.acm.org/10.1145/364520.364540.

[276] D. McFadden, "Modelling the Choice of Residential Location", Cowles Foundation for Research in Economics, Yale University, Cowles Foundation Discussion Papers 477, 1977. [Online]. Available: https://EconPapers.repec.org/RePEc:cwl:cwldpp:477.

[277] J. Back *et al.*, "LAURA^{++}: A Dalitz plot fitter", *Comput. Phys. Commun.*, vol. 231, pp. 198–242, 2018. DOI: 10.1016/j.cpc.2018.04.017. arXiv: 1711.09854 [hep-ex].

[278] G. Zech, "Scaling property of the statistical Two-Sample Energy Test", 2018. arXiv: 1804.10599 [physics.data-an].

[279] T. P. S. Gillam and C. G. Lester, "Biased bootstrap sampling for efficient two-sample testing", *JINST*, vol. 13, no. 12, P12014, 2018. DOI: 10.1088/1748-0221/13/12/P12014. arXiv: 1810.00335 [physics.data-an].