Measurement of the charge asymmetry in beauty-dijet production at the LHCb experiment

Matthew Bradley

High Energy Physics

Blackett Laboratory

Imperial College London

A thesis submitted to Imperial College London

for the degree of Doctor of Philosophy

February 14, 2023

Abstract

A measurement of the charge asymmetry in beauty-dijet production at LHCb is presented in this thesis. The measurement uses the 2016 dataset of proton-proton collisions gathered by the detector, corresponding to an integrated luminosity of $1.7\,\text{fb}^{-1}$. The charge asymmetry is measured in three bins of the invariant dijet mass, with bin edges at: $[50, 75, 105, 150]$ GeV. This represents the first measurement of the charge asymmetry in beauty-dijet production in proton-proton collisions at a centre-of-mass energy of 13 TeV. To make the measurement a charge tagging method is first developed and tested in order to distinguish between the presence of b- and $\bar{b}$-quarks in jets. The calibration of simulated events is then carried out to correct for mismodelling when compared to the data. Next, fits are performed to extract the yield of beauty-dijets in the data sample. These yields are then corrected for detector and mistagging effects. The values of the asymmetry are then calculated. Results are compared to Standard Model predictions and are found to agree to within one standard deviation in all three bins.

Declaration of originality

The analysis presented in this thesis is the result of the research I performed between January 2019 to December 2022, with the support of the Imperial College High Energy Physics group and members of the LHCb Collaboration. The analysis work presented in this thesis was performed by myself with some exceptions. The size of the χ^2_{FD} correction applied to simulated heavy flavour jets in Section 6.3 was calculated by Dr William Barter. The one-dimensional fits carried out in Section 7.1.1 to help understand light-jet contributions to the sample were also carried out by Dr William Barter. Both of these pieces of work were carried out in close association with myself. These contributions are also explicitly noted in the body of the text for the avoidance of doubt.

This thesis has not been submitted for any other qualification.

Matthew Bradley, February 2022

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).
Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.
When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.
Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

This thesis is the culmination of four years' worth of work and I am so grateful for the numerous individuals who have supported and encouraged me along the way. While it is impossible to name them all here, their contributions are not forgotten. There are a few thanks in particular that I would like to make to those who have been particularly instrumental in this chapter of my life.

First and foremost I would like to thank my supervisor Will Barter who has been a robust source of knowledge and support throughout my time researching and working on this thesis over the last four years. It still strikes me as an incredible coincidence that you went from interviewing me for my undergraduate degree to mentoring me through my PhD several years later! Your tremendous expertise and wisdom have served me immeasurably well and I am sure they will do so again for many other students in the future.

I would like to extend my thanks to members of the LHCb Imperial Group, both past and present, for creating a supportive and enjoyable environment in which to carry out my studies. I am also grateful to the wider Imperial HEP Group for giving me a place to complete my research; the knowledge and experience within the group have been invaluable in shaping my work. Thanks also to the wider LHCb community at CERN, with particular thanks to Lorenzo Sestini for your guidance and support.

Thank you to all my friends and family who have supported me during this journey. Your encouragement and tolerance of my prolonged status as a student is appreciated. Thank you for keeping me laughing.

Finally, to Iz. I could not have done this without your love, patience and unwavering support. Thank you for being by my side through everything. I am so excited to see where life takes us next.

To mum,

You made me who I am today and I am eternally grateful for everything you have ever done for me.

My successes are your successes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Standard Model (SM) is a theoretical framework that describes the fundamental particles and how they interact. The theory was developed in the 1960s and 70s by (amongst others) Glashow, Salam and Weinberg, who won the 1979 Nobel Prize in Physics for their contributions to the theory of the unified electromagnetic and weak interactions. It is a huge feat of modern science that this theory has been able to accurately predict several phenomena, such as the existence and properties of the W and Z bosons [1–4], the six quarks [5–7], and the Higgs boson [8–10], the latter of which was discovered nearly 50 years after it was first theorised [11, 12]. That one can uncover so much about the structure of the universe using little more than a pen and paper, and that predictions made can be tested experimentally with incredible precision, is a triumph that cannot be overstated.

Although experimental results have been able to confirm several predictions made by the SM, there remains much that cannot be explained by the theory. Why is there a huge imbalance in the amount of matter and antimatter present in the universe, when the SM predicts these should be created and destroyed in nearly equal amounts? What is the underlying mechanism for the force of gravity? Is there a particle responsible for the observations of dark matter and, if so, how does it interact with other particles in the universe? These are just a few examples of questions that continue to puzzle those in the field.

As part of the work carried out by theoretical physicists, extensions to the SM to include New Physics (NP) are continuously being developed in the hope of achieving a potential resolution to these questions. It is therefore imperative that as experimental physicists we continue to test and probe the SM rigorously from all angles to constrain

NP models and potentially expose any potential weaknesses of the SM.

Such a test of the SM is the topic of this thesis. The aim of the analysis presented is to provide a measurement of the charge asymmetry in beauty-quark production at the LHCb detector. This measurement probes two interesting regions of the SM. First, the measurement probes quantum chromodynamics (QCD), the theory underlying the so-called strong force, since the main mechanism for the production of beauty quarks at the LHC is described by this theory. Second, the measurement probes the electroweak sector, since Z bosons decay to produce pairs of beauty quarks. These beauty quarks will quickly hadronise to form jets and so it is the angle of these jets that is measured. The measurement is provided in bins of the dijet invariant mass, since the rates of the different processes that contribute to the asymmetry vary with this property. The measurement uses the 2016 dataset gathered with the LHCb detector, corresponding to an integrated luminosity of $1.7\,\text{fb}^{-1}$. This paves the way for an analysis using the full Run II dataset in the future.

The thesis has two main parts: Part I covers the theoretical and experimental background for the thesis and Part II details the measurement itself. These are further broken down into chapters as follows. Chapter 2 will explore the theoretical background by examining the mathematical structure of the SM and its relevance to the studies set out in this thesis. Chapter 3 lays out the experimental background, describing the LHCb detector and related features. Chapter 4 introduces the measurement and provides some of the theoretical predictions for the value of the asymmetry measured. The samples on which the measurement is made and the selection of events within those samples are laid out in Chapter 5. The calibration of simulated jets used in the analysis is then described in Chapter 6. Fits to the data are performed to obtain signal yields in Chapter 7, and the results are corrected for detector effects in Chapter 8. Systematic uncertainties for the measurement are set out in Chapter 9. Finally, results and future related measurements are discussed in Chapter 10, with concluding remarks in Chapter 11.

A note on units and conventions

Throughout this thesis natural units, where $\hbar = c = 1$, are used unless specified otherwise. The Einstein summation convention is also used. In addition, 4-vectors will be given Greek letter indices (e.g. x^μ) and 3-vectors will be given Latin letter indices (e.g. x^i).

It is also useful at this point to define the coordinate system commonly used at hadron

collider experiments, as well as some variables that are often encountered. Colliding beams are defined to travel in the $\pm z$ direction. Angles in the x–y plane are referred to as azimuthal angles, ϕ. Angles to the z-axis are defined as θ, but such angles at hadron colliders are typically referred to in terms of the pseudorapidity,

$$\eta \equiv -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]. \tag{1.1}$$

Pseudorapidity is a useful variable as particle production is typically flat as a function of η. Regions of phase space with high pseudorapidity (corresponding to small angles θ) are referred to as the "forward" or "backward" regions for positive and negative z directions respectively. The absolute value of the momentum of a particle in the x–y plane is referred to as the transverse momentum, p_{T}.

When logarithms are taken in the thesis, unless explicitly stated, these are natural logarithms and will be written as "$\log(x)$".

Finally, a note on the rounding of numbers in this thesis. Where results from studies carried out in this thesis are quoted, the rounding rules from the Particle Data Group (PDG) have been applied [13]. This states that if the three highest-order digits of the uncertainty lie between 100 and 354, then rounding will be carried out to two significant figures. If the digits lie between 355 and 949, then rounding will be carried out to one significant figure. For digits between 950 and 999, the value is rounded up to 1000 and two significant figures are kept. For results with multiple sources of uncertainty reported, the number of significant figures reported will be set using the largest uncertainty.

Part I

Theoretical and experimental overview

Chapter 2

Theoretical overview

The following chapter explores the theories and underlying concepts that are relevant for motivating and understanding the work set out in this thesis and placing it in the wider context. The first part introduces the SM, detailing the fundamental particles and their interactions. This is followed by a closer look at the mathematical formalism that describes these interactions. The chapter is concluded by discussing the general theory relevant for LHC collisions.

2.1 The Standard Model

The SM is introduced and summarised in the following section, based on Refs. [14–16]. The SM governs the known fundamental particles that make up matter in the universe and the interactions between these particles. Interactions in the SM are comprised of three fundamental forces: the electromagnetic, the weak and the strong forces. These forces are mediated by the exchange of spin-1 particles called gauge bosons. The exchange of photons leads to the electromagnetic force; W and Z bosons are responsible for the weak force; and gluons carry the strong force. There is also a fourth fundamental force of nature, gravity, but this is not accounted for in the SM.

There are then 12 spin-half particles that constitute matter, known as fermions. These fermions can be further divided into two categories: leptons and quarks. Leptons have integer charge and interact through the weak force. If their charge is non-zero, they also interact through the electromagnetic force. There are three generations of lepton, each consisting of one charged and one neutral particle (a neutrino). These are the electron (e^-) and electron neutrino (ν_e), the muon (μ^-) and muon neutrino (ν_μ), and the tau

Table 2.1: The twelve fermions of the Standard Model. These have been split into the fractionally-charged quarks and integer-charged leptons, and each of these is further divided into three generations. The charges of each particle are shown along with their masses. Measurements of neutrino flavour oscillations (such as in Ref. [17]) have demonstrated the existence of neutrino mass, hence the neutrinos listed here are the mass eigenstates labelled numerically. The masses have not been measured, hence they are not given here. Limits have, however, been placed on the sum of neutrino masses from cosmological observations, giving $\sum_i m_i < 0.12\,\text{eV}$ [13]. The KATRIN collaboration also recently provided the most stringent limit on the effective mass of the electron neutrino at $m_\nu < 0.8\,\text{eV}$ [18]. Table reproduced from Ref. [13].

	Quarks			Leptons		
Generation	Particle	Mass	Charge	Particle	Mass	Charge
1	u	2.2 MeV	$+\frac{2}{3}$	e	511 keV	-1
	d	4.7 MeV	$-\frac{1}{3}$	ν_1	–	0
2	c	1.3 GeV	$+\frac{2}{3}$	μ	106 MeV	-1
	s	93 MeV	$-\frac{1}{3}$	ν_2	–	0
3	t	173 GeV	$+\frac{2}{3}$	τ	1.8 GeV	-1
	b	4.2 GeV	$-\frac{1}{3}$	ν_3	–	0

(τ^-) and tau neutrino (ν_τ). Quarks feel all three of the forces in the SM and have a fractional charge. These can also be grouped into three generations: the up (u) and down (d) quarks, the charm (c) and strange (s) quarks, and the top (t) and beauty[1] (b) quarks. The fermions and their properties are summarised in Table 2.1.

There is one final particle in the SM, the Higgs boson. This is a spin-0 particle that does not carry a force like the other bosons. This boson is an excitation of the Higgs field. The field is required in order to preserve the gauge invariance of the weak force in the presence of massive bosons. The Higgs field is also responsible for giving particles their mass via its interactions with them. The bosons and their properties are summarised in Table 2.2.

[1]The beauty quark is often referred to as the bottom quark. This naming emphasises its partnership with the top quark but sounds less elegant, hence the former is chosen here.

Table 2.2: Table detailing the charge and mass of bosons in the SM. Table reproduced from Ref. [13].

Particle	Mass / GeV	Spin	Charge	Force
γ	0	1	0	Electromagnetism
$W^{\pm}$	80.4	1	± 1	Weak
Z	91.2	1	0	
g	0	1	0	Strong
H^0	125	0	0	-

It should be noted here that a recent measurement of the W mass by the CDF collaboration [19] is in significant tension with previous measurements. The PDG average [13] and the measured CDF values are

$$m_W^{\mathrm{PDG}} = 80.377 \pm 0.012\,\mathrm{GeV}, \tag{2.1}$$

$$m_W^{\mathrm{CDF}} = 80.434 \pm 0.009\,\mathrm{GeV}. \tag{2.2}$$

As yet this discrepancy is unexplained, with further work required to understand the results.

2.2 Interactions in the Standard Model

The interactions between particles via the electromagnetic, weak and strong forces can be understood through the mathematical formalism of the SM, where the forces are each described by a quantum field theory (QFT). The simplest force is that of electromagnetism, described by quantum electrodynamics (QED), and so this example will be explored first.

2.2.1 Quantum electrodynamics

QED describes interactions between charged particles. As a starting point, the motion of a free, charged particle can be described by the Dirac Lagrangian

$$\mathcal{L} = \overline{\psi}(i\gamma^{\mu}\partial_{\mu} - m)\psi, \tag{2.3}$$

where γ^μ are the Dirac matrices and ψ is a spin-half field ($\overline{\psi} = \psi^\dagger\gamma^0$) describing a particle with mass m. Considering the field $\overline{\psi}$ and using the Euler–Lagrange equations, the equation of motion for the free particle can be derived as

$$i\gamma^\mu\partial_\mu\psi - m\psi = 0. \tag{2.4}$$

The Lagrangian is invariant under the transformation

$$\psi \to e^{i\alpha}\psi, \tag{2.5}$$

where $\alpha \in \mathbb{R}$ is a constant number. This is a global gauge transformation, as we are changing the phase of the field in the same way at every point in space. However, when considering a local gauge transformation where the phase depends on the spacetime coordinate, such as

$$\psi \to e^{i\alpha(x)}\psi, \tag{2.6}$$

the Lagrangian transforms as

$$\mathcal{L} \to \mathcal{L} - (\partial_\mu\alpha)\overline{\psi}\gamma^\mu\psi, \tag{2.7}$$

which is clearly not invariant. To achieve a gauge symmetry, a new field, A^μ, is introduced, which transforms under the local gauge transformation as

$$A^\mu \to A^\mu - \frac{1}{e}\partial^\mu\alpha(x). \tag{2.8}$$

As a result, the derivative term in the Lagrangian can be rewritten to include this extra field as

$$\partial_\mu \to \partial_\mu + ieA_\mu = D_\mu. \tag{2.9}$$

The full form of this Lagrangian is then invariant under local gauge transformations. This is written as

$$\mathcal{L} = \overline{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu}, \tag{2.10}$$

where $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$ and the $F^{\mu\nu}F_{\mu\nu}$ term in the Lagrangian accounts for the kinetic energy of the field A_μ. Local gauge invariance is therefore not possible for a free particle. Instead, an additional field is required that couples to the particle via the electric charge (in the case of an electron this has magnitude e). This is the photon field and is responsible for interactions between charged particles, shown in the expanded Lagrangian

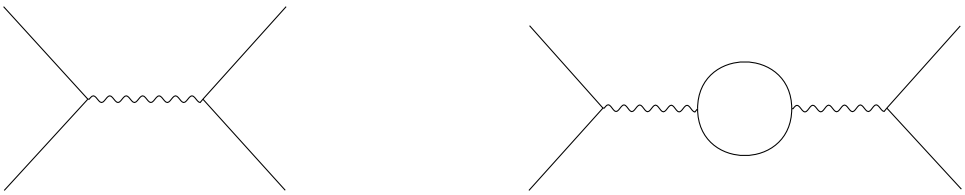

Figure 2.1: Feynman diagrams for a simple process in QED. The presence of the loop in the right-hand diagram leads to integrals that diverge in the calculations of the amplitudes.

as the term $-e\overline{\psi}\gamma^{\mu}A_{\mu}\psi$. The photon cannot have a similar form for its kinetic energy as the electron (i.e. a term quadratic in the field A_{μ}) as this would not be invariant under the local gauge transformation, hence the photon must be massless.

This Lagrangian can be used to make predictions about processes occurring within the theory. The calculations are governed by the Feynman rules, which can be used to calculate the probability of a certain process occurring. The calculations themselves are a set of integrals in the momentum space of the particles involved in the interaction, depicted pictorially by Feynman diagrams. These calculations are made using perturbation theory, with the order of the calculation quantified in terms of powers of the coupling strength, $\alpha = e^2/4\pi$. However, where loops occur in the Feynman diagrams, the integrals in momentum space can become divergent, leading to apparent infinite probabilities — this is clearly unphysical. Examples of Feynman diagrams for a simple QED process with and without a loop are shown in Figure 2.1. To calculate finite expectation values a process called "renormalisation" is used. This involves summing all of the divergent diagrams and absorbing these divergences into the parameters of the theory, such as the mass of the electron and QED coupling strength. This leads to the parameters of the theory seeming to change with energy, this is referred to as the parameters "running". The energy scale at which theoretical calculations are made with respect to renormalisation is called the renormalisation scale. This is typically chosen to correspond to some energy scale in the process of interest. The renormalisation scale and its impact on theoretical predictions will be discussed more in Section 2.5.

A comparison of the theoretical predictions of QED with the measured values shows how powerful the theory can be. The prediction [20] and measured value [21] for the anomalous magnetic moment of the electron are

$$a_{\text{theory}} = 1\,159\,652\,181.61 \pm 0.23 \times 10^{-12}, \tag{2.11}$$

$$a_{\text{meas}} = 1\,159\,652\,180.73 \pm 0.28 \times 10^{-12}, \tag{2.12}$$

representing an agreement of better than one part in a billion, an astonishing level of accuracy.

2.2.2 Spontaneous symmetry breaking

Within QED we are forced to have a massless, force-carrying boson in order to maintain gauge invariance, as discussed in Section 2.2.1. Since the weak force is propagated by massive gauge bosons, the same mechanism cannot be used to preserve the gauge invariance of the weak force. An alternative mechanism is therefore desired that allows for the preservation of gauge invariance but provides massive bosons. Spontaneous symmetry breaking provides such a route and a simple example of this mechanism is provided here. It will be shown how this gives rise to massive gauge bosons in the weak force in Section 2.2.3.

The example begins with a Lagrangian describing a complex scalar field

$$\mathcal{L} = (\partial^\mu \phi)^\dagger (\partial_\mu \phi) - V(|\phi|), \tag{2.13}$$

where $V(|\phi|)$ is the potential energy of the field and is given by

$$V(|\phi|) = \mu^2 |\phi|^2 + \lambda |\phi|^4. \tag{2.14}$$

This Lagrangian is invariant under the global gauge transformation

$$\phi \to e^{i\alpha}\phi, \tag{2.15}$$

similar to the QED Lagrangian. If $\mu^2 > 0$ then the potential has a minimum at $|\phi| = 0$. However, if the case of $\mu^2 < 0$ is considered then the minima occurs instead at

$$|\phi| = \sqrt{\frac{-\mu^2}{2\lambda}} \equiv \frac{v}{\sqrt{2}}. \tag{2.16}$$

The potential in question still has a stationary value at $|\phi| = 0$, but this is now a local maximum, with a ring of minima at $|\phi| = v/\sqrt{2}$. The field has acquired a non-zero vacuum expectation value (VEV). Choosing v to be along the real axis, the substitution

$$\phi = \frac{1}{\sqrt{2}}(v + \phi_1 + i\phi_2), \tag{2.17}$$

can be made in the Lagrangian, where $\phi_{1,2}$ are real scalar fields and represent perturbations from the VEV in the real and imaginary directions, respectively. If this is done, there emerges a quadratic term in ϕ_1 indicating a massive excitation with mass, $m = \sqrt{-2\mu^2}$. There are no quadratic terms for the field ϕ_2, hence it must be massless.

2.2.3 Electroweak unification

The unification of the electromagnetic and weak forces gives rise to the electroweak theory, for which Glashow, Salam and Weinberg were awarded the 1979 Nobel Prize [1–3]. This is a gauge invariant theory, with symmetry group $SU(2) \times U(1)$. Experiments on beta decay in 1957 by Wu showed that parity is violated in the weak interaction [22], meaning that the Lagrangian is not invariant under the spatial inversion $\mathbf{x} \to -\mathbf{x}$. The fermion fields are therefore divided into states of definite chirality

$$\psi_L = P_L\psi = \frac{1-\gamma^5}{2}\psi, \tag{2.18}$$

$$\psi_R = P_R\psi = \frac{1+\gamma^5}{2}\psi, \tag{2.19}$$

where $P_{L,R}$ are operators that project out the left- and right-handed chiral states of the fermion field, and $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$. A parity transformation would switch the states between left- and right-handed: $\psi_L \leftrightarrow \psi_R$. As parity is violated, the two fields must behave differently in the theory and hence may transform in different ways under a gauge transformation. To examine the effects of this, the Dirac Lagrangian can be written in terms of these left- and right-handed fields which gives

$$\mathcal{L} = i\overline{\psi_L}(\gamma^\mu\partial_\mu)\psi_L + i\overline{\psi_R}(\gamma^\mu\partial_\mu)\psi_R - m(\overline{\psi_L}\psi_R + \overline{\psi_R}\psi_L). \tag{2.20}$$

For massless fermions, the Lagrangian has completely separate terms for ψ_L and ψ_R, but for massive fermions, the left- and right-handed components are coupled. This would break the gauge invariance of the Lagrangian if the two fields transform differently. As gauge invariance is still required, the fermion masses are instead generated using a different mechanism. The SM makes use of spontaneous symmetry breaking as is discussed in Subsection 2.2.4.

We will start by considering the interactions of the fermions with the electroweak bosons. The charge associated with the $SU(2)$ subgroup of the theory is called the weak

isospin, T, and particles form multiplets with different values of weak isospin. The left-handed fermion fields form doublets. There is one doublet for each generation of the leptons and quarks for a total of six as follows:

$$L = \left[\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L \right], \tag{2.21}$$

$$Q = \left[\begin{pmatrix} u \\ d \end{pmatrix}_L, \begin{pmatrix} c \\ s \end{pmatrix}_L, \begin{pmatrix} t \\ b \end{pmatrix}_L \right]. \tag{2.22}$$

Then L_i and Q_i are used to refer to the left-handed doublets of the i^{th} generation. The vectors for right-handed fields are

$$E_R = \begin{pmatrix} e_R \\ \mu_R \\ \tau_R \end{pmatrix}, \quad \nu_R = \begin{pmatrix} \nu_{e,R} \\ \nu_{\mu,R} \\ \nu_{\tau,R} \end{pmatrix}, \quad U_R = \begin{pmatrix} u_R \\ c_R \\ t_R \end{pmatrix}, \quad D_R = \begin{pmatrix} d_R \\ s_R \\ b_R \end{pmatrix}. \tag{2.23}$$

The $U(1)$ subgroup of electroweak theory is similar to the QED theory discussed in Section 2.2.1, however, the charge here is called the weak hypercharge, Y, rather than electric charge as previously. There are two covariant derivatives that maintain the gauge invariance of the theory, one for left-handed and one for right-handed fields. These are defined as

$$D_\mu^L = \partial_\mu + ig\frac{\sigma_i}{2}W_\mu^i + ig'\frac{Y}{2}B_\mu, \tag{2.24}$$

$$D_\mu^R = \partial_\mu + ig'\frac{Y}{2}B_\mu, \tag{2.25}$$

where W_μ^i are the gauge fields associated with the $SU(2)$ subgroup and B_μ is the gauge field associated with the $U(1)$ subgroup. The matrices σ_i are the Pauli spin matrices, where the index i sets a direction in weak isospin space. The constants g and g' are coupling strengths between fermions and the gauge fields. The right-handed covariant derivative has no W_μ^i term as the weak isospin of the right-handed singlets is zero. The components of the electroweak Lagrangian relating to the fermions can then be written

$$\mathcal{L}_{\text{fermions}} = \sum_{\text{fermions}} \overline{\psi_L}\gamma^\mu D_\mu^L \psi_L + \overline{\psi_R}\gamma^\mu D_\mu^R \psi_R. \tag{2.26}$$

The gauge fields themselves have kinetic terms in the Lagrangian

$$\mathcal{L}_{\text{bosons}} = -\frac{1}{4}B^{\mu\nu}B_{\mu\nu} - \frac{1}{4}W^{i\mu\nu}W^{i}_{\mu\nu}, \tag{2.27}$$

where $B_{\mu\nu}$ and $W^{i}_{\mu\nu}$ are the field strength tensors associated with the respective fields and are defined

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu, \tag{2.28}$$

$$W^{i}_{\mu\nu} = \partial_\mu W^{i}_{\nu} - \partial_\nu W^{i}_{\mu} - g\varepsilon^{ijk}W^{j}_{\mu}W^{k}_{\nu}, \tag{2.29}$$

where ε^{ijk} are structure constants.

The terms in Eq. (2.27) give rise to massless bosons (as seen with QED), but the Z and $W^{\pm}$ bosons are known to be massive. The symmetry of the group is therefore spontaneously broken to give rise to these massive bosons and the massless photon. This is similar to the process described in Section 2.2.2. To do this the scalar Higgs is introduced as a doublet of scalar fields

$$\Phi = \begin{pmatrix} \phi^{+} \\ \phi^{0} \end{pmatrix} \tag{2.30}$$

We can then write a Lagrangian term for the Higgs

$$\mathcal{L} = (D_\mu \Phi)^{\dagger}(D^\mu \Phi) - V(\Phi), \tag{2.31}$$

where the Higgs potential is given by

$$V(\Phi) = \mu^2 \Phi^{\dagger}\Phi + \lambda(\Phi^{\dagger}\Phi)^2. \tag{2.32}$$

As in Section 2.2.2 the symmetry of the Higgs potential is spontaneously broken by choosing $\mu^2 < 0$. The minimum of $V(\Phi)$ then occurs at a non-zero VEV given by

$$|\Phi| = \sqrt{\frac{-\mu^2}{2\lambda}} = \frac{v}{\sqrt{2}}. \tag{2.33}$$

Without loss of generality, the VEV of the Higgs field can be chosen to lie along the ϕ_0 direction, and hence the field at this minimum can be written as

$$\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.34}$$

Substituting this expression into the covariant derivative term of the Higgs Lagrangian gives

$$\frac{1}{8}\begin{pmatrix}0, v\end{pmatrix}\begin{pmatrix} gW_\mu^3 + g'B_\mu & g(W_\mu^1 - iW_\mu^2) \\ g(W_\mu^1 - iW_\mu^2) & -gW_\mu^3 + g'B_\mu \end{pmatrix}^2 \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.35}$$

The physical gauge fields are then identified as

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2), \tag{2.36}$$

$$Z_\mu^0 = \frac{1}{\sqrt{g^2 + g'^2}}(gW_\mu^3 - g'B_\mu), \tag{2.37}$$

$$A_\mu = \frac{1}{\sqrt{g^2 + g'^2}}(g'W_\mu^3 + gB_\mu). \tag{2.38}$$

It was shown previously that the right-handed fermion states have no interaction with the W_μ^i terms due to their weak isospin being zero. The $W^\pm$ bosons can therefore only interact with left-handed fermions. It can be shown similarly that the W bosons only interact with right-handed antifermions. The Z boson is able to couple to both left- and right-handed (anti)fermion states, due to the contribution from B_μ. The A_μ term represents the photon. The terms for the physical bosons can now be inserted into Eq. (2.35) which then reduces to

$$\frac{v^2 g^2}{4} W_\mu^+ W^{-\mu} \; + \; \frac{v^2(g^2 + g'^2)}{8} Z_\mu Z^\mu. \tag{2.39}$$

The mass terms for the physical gauge bosons can then be extracted[2]

$$m_W = \frac{vg}{2}, \quad m_Z = \frac{v\sqrt{g^2 + g'^2}}{2}. \tag{2.40}$$

There is no mass term for the field A_μ, as expected: the photon is massless. The Z boson and photon fields can also be written in terms of the weak mixing angle θ_W,

$$\begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix} \tag{2.41}$$

[2]There are different normalisations for the two fields here. $W^\pm$ are complex fields which have mass terms $m^2\phi\phi^*$. Z_μ is a real field which has a mass term $\frac{m^2}{2}\phi^2$.

where

$$\cos\theta_W = \frac{g}{\sqrt{g^2+g'^2}}, \quad \sin\theta_W = \frac{g'}{\sqrt{g^2+g'^2}}. \tag{2.42}$$

The masses of the weak bosons can then be related by $m_W = m_Z \cos\theta_W$. The weak mixing angle has been measured experimentally as $\sin\theta_W = 0.23153 \pm 0.00016$ [23]. This agrees with the latest world averages of the boson masses provided by the PDG [13], and hence the values summarised in Table 2.2. These averages do not include the latest CDF measurement of the W mass, which shows significant tensions with previous measurements. More work is needed to understand this discrepancy. Furthermore, as set out in Ref. [13], a greater understanding of the correlations between the results is required before an updated average is provided in the next iteration of the PDG.

2.2.4 Fermion masses

As discussed in Section 2.2.3 the fermions also obtain their masses through coupling with the Higgs field. These coupling terms are Yukawa couplings and, for the leptons, are of the form

$$\mathcal{L}_{\text{Yuk},L} = -\sum_i \lambda_i \overline{L_i} \Phi e_R^i + \text{h.c.}, \tag{2.43}$$

where λ_i is the coupling strength and h.c. refers to the Hermitian conjugate. If the VEV of the Higgs field is inserted into this equation then mass terms for the leptons can be extracted. Taking the electron as an example we see a mass term

$$\mathcal{L}_{e_{\text{mass}}} = -\frac{\lambda_e\, v}{\sqrt{2}} \left(\overline{e_L} e_R + \overline{e_R} e_L \right), \tag{2.44}$$

where the mass of the electron can now be identified as $\lambda_e v/\sqrt{2}$, with similar terms for the muon and tau. It can be seen from this that the mass of a particle relates to the strength of its coupling with the Higgs field. The neutrino terms remain massless here as only the bottom half of the $SU(2)$ doublets are projected out. The quarks have similar mass terms, although the inclusion of Pauli spin matrices is required to project out the upper half of the quark $SU(2)$ doublets and hence generate mass for the up-type quarks. The charge conjugate of the Higgs field is therefore defined

$$\Phi^c = i\sigma_2 \Phi^* = \frac{1}{\sqrt{2}} \begin{pmatrix} v \\ 0 \end{pmatrix} \tag{2.45}$$

The most general mass terms that can then be written for the quarks are

$$\mathcal{L}_{\text{Yuk},Q} = -\frac{v}{\sqrt{2}}\left(\lambda_d^{ij}\,\overline{Q^i}\,\Phi\,D_R^j + \lambda_u^{ij}\,\overline{Q^i}\,\Phi^c\,U_R^j + \text{h.c.}\right), \tag{2.46}$$

where λ_{ij} are 3×3 complex Yukawa coupling matrices. These matrices can be diagonalised by a change of basis, which also entails a rotation of the $SU(2)$ doublets. The diagonalisation will then leave mass terms similar to those seen in the lepton sector. It is, however, impossible to diagonalise both λ_d^{ij} and λ_u^{ij} simultaneously without altering the form of the Lagrangian in Eq. (2.26). When these Yukawa matrices are simultaneously diagonalised, the weak interaction term corresponding to the W boson then includes a matrix referred to as the Cabibbo–Kobayashi–Maskawa (CKM) matrix [7,24]. This gives the possibility of mixing between the generations of quarks, as it allows for one quark to decay to another belonging to a different generation via the W boson. This effect shows that the mass eigenstates of the quarks are not equivalent to the eigenstates through which the weak interaction acts, the so-called "flavour" eigenstates. The CKM matrix can be parameterised by three mixing angles and a CP-violating phase [7].

The same effect wasn't shown in the lepton sector due to the lack of a neutrino mass term in the Lagrangians discussed here. This leaves only one Yukawa coupling matrix to diagonalise, which allows for the Langrangian in Eq. (2.26) to remain unchanged with respect to the leptons, giving no intergenerational interactions. This is the origin of a strict form of lepton flavour conservation, a rule of thumb where the difference between the number of leptons and antileptons of a given flavour generation is constant within an interaction. Of course, in reality, neutrinos do have mass as demonstrated by oscillation experiments [25, 26]. This means that mixing between the different lepton generations is indeed possible and, hence, lepton numbers are not conserved within a generation. This also gives rise to a matrix similar to the CKM matrix in the lepton sector: the Pontecorvo–Maki–Nakagawa–Sakata (PMNS) matrix [27, 28]. One open question is whether neutrinos are Dirac particles like the other fermions. This would imply the existence of right-handed neutrinos which are yet to be observed[3] (these were, however, introduced in Eq. (2.23) as they fit naturally within the theory described). Alternatively, the neutrino might be its own antiparticle, a hypothetical type of particle called a Majorana particle [29], although there has been no experimental evidence for this.

[3]Observation would be astoundingly difficult, as right-handed neutrinos would only interact through gravity or the Higgs field without NP theories. Both of these forces would couple very weakly to the neutrino due to its near-zero mass.

2.2.5 Interactions with the Higgs boson

It has been shown that the masses of the bosons and fermions are generated through their interaction with the Higgs field but so far interactions of the Higgs boson itself have not been considered. The Higgs boson is a massive excitation of the Higgs field, which can be considered as a perturbation to the VEV of the field. The expression for the Higgs field in Eq. (2.34) can be modified to include this perturbation

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \tag{2.47}$$

where $h(x)$ is the excitation corresponding to the Higgs boson. This form for the Higgs field is referred to as the unitary gauge. If this excitation is included in the Lagrangian shown in Eq. (2.43) terms are recovered such as

$$\mathcal{L}_{h-e} = -\frac{\lambda_e h}{\sqrt{2}} \left(\overline{e_L} e_R + \overline{e_R} e_L \right). \tag{2.48}$$

This corresponds to an interaction between an electron and a Higgs boson. The coupling strength of this interaction is proportional to the mass of the electron which can be seen by comparison with Eq. (2.44). Similarly, every particle that gains a mass through interaction with the Higgs field also gains a coupling to the Higgs boson itself. As the gluon and photon are massless they do not couple directly to the Higgs.

2.2.6 Quantum chromodynamics

The final force in the SM is the strong force. This is described by the theory of quantum chromodynamics (QCD). QCD is a gauge field theory with the symmetry group $SU(3)$. The charge of this theory is called colour charge and there are three charges labelled red, green and blue. Quarks carry one of these colours and antiquarks carry an anticolour. The gauge boson that is responsible for carrying the force is the gluon. Gluons carry a colour and an anticolour (or linear combinations of this type). The QCD Lagrangian is given by

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4} G^i_{\mu\nu} G^{\mu\nu}_j + \overline{\psi_a} \left(i\gamma^\mu D^{ab}_\mu - m\delta^{ab} \right) \psi_b, \tag{2.49}$$

where the gluon field strength tensor $G^a_{\mu\nu}$ is given by

$$G^i_{\mu\nu} = \partial_\mu A^i_\nu - \partial_\nu A^i_\mu - g_s f^{ijk} A^j_\mu A^k_\nu, \tag{2.50}$$

the covariant derivative is given by

$$D^{ab}_\mu = \delta^{ab}\partial_\mu + i g_s A^i_\mu T^{ab}_i, \tag{2.51}$$

and δ^{ab} is the Kronecker delta. The gluon fields in these equations are A^i_μ, and g_s is the gluon coupling strength. The structure constants of the $SU(3)$ group are written f^{ijk}. The T^{ab}_i are the Gell-Mann matrices which are the generators of the $SU(3)$ group. Eight matrices are needed to span the $SU(3)$ group, hence i runs from $1-8$. This shows the existence of eight separate gluons, identified by varying colour combinations. The indices $a, b \in [1, 2, 3]$ indicate the colour charge.

Similarly to QED, QCD contains divergences when considering higher-order Feynman diagrams. Renormalisation is again used to remove these and give meaningful predictions. This again leads to a running of the strong coupling constant, α_S. However, whilst the coupling constant in QED gets smaller at higher energies, the strong coupling constant does the opposite. At low energies α_S is too large to consider a perturbative expansion approach to calculations and so non-perturbative QCD calculations are used instead.

The strength of the strong coupling constant at low energies leads to the concept of confinement. Particles that carry colour can only be observed as colourless combinations. A colour and its corresponding anticolour combine to form a colourless object as does the combination of all three colours: red, green and blue. The combination of a quark and an antiquark is therefore allowed (these are the mesons), as is the combination of three quarks (the baryons). Conversely, a single, free quark cannot exist alone. If it were possible to extract a single quark from a baryon, the amount of energy put into the system would lead to the creation of new quarks in such a way that the resultant particles would still be colourless. This is the basis of hadronisation, an important concept for jets and their formation which will be discussed in Section 2.4.

At higher energies, α_S becomes sufficiently small for perturbative calculations to be used, called perturbative QCD (pQCD). The smaller value of the coupling constant means that quarks are more weakly interacting, such that at high energies quarks behave as though they are free particles. This property is called asymptotic freedom [30, 31]. For the discovery of this Gross, Politzer and Wilczek won the Nobel Prize in 2004.

2.3 LHC collisions

At the Large Hadron Collider (LHC), beams of protons are collided together at high energies and the particles produced in these collisions are detected and measured. A basic view of the proton is that its quark composition is two up quarks and one down quark (uud). However, these quarks are constantly interacting through the strong force, exchanging gluons. These gluons in turn lead to the creation of quark-antiquark pairs. This means that we cannot simply think of the proton using the basic uud view.

The constant interaction of particles via QCD within the proton poses a problem for making theoretical predictions at hadron colliders. The low-energy QCD interactions within the proton cannot be handled using pQCD. This problem is addressed by using the factorisation theorem. Simply put, this states that emission and interactions below a certain energy threshold are considered to be a part of the proton. The partons created in these interactions are referred to as sea quarks, with valence quarks being the original two u quarks and one d quark. When two protons collide the interactions between the partons that make them up also include sea quarks or gluons. In the factorisation theorem, emission above this energy threshold is instead considered to be separate from the proton. This must then be included when considering the Feynman diagrams for the process of interest. The energy threshold that separates these two regions is referred to as the factorisation scale. The factorisation scale is set by hand, typically to the value of an energy scale in the process of interest.

The contributions from different sea and valence partons to the proton are described by parton distribution functions (PDFs). These describe the probability density of finding a given parton with a given Bjorken-x value [32] at an energy scale Q^2. In the limit of high energy, x is the fraction of the proton's momentum carried by the parton. The proton must contain two up valence quarks and one down valence quark implying

$$\int_0^1 u_v(x)\,\mathrm{d}x = 2, \tag{2.52}$$

$$\int_0^1 d_v(x)\,\mathrm{d}x = 1, \tag{2.53}$$

where u_v and d_v are the PDFs for the up and down valence quarks respectively. As sea quarks and antiquarks of the same flavour will be created and destroyed in pairs, the PDF for sea quarks is expected to be equal to that of sea antiquarks. As the mass of

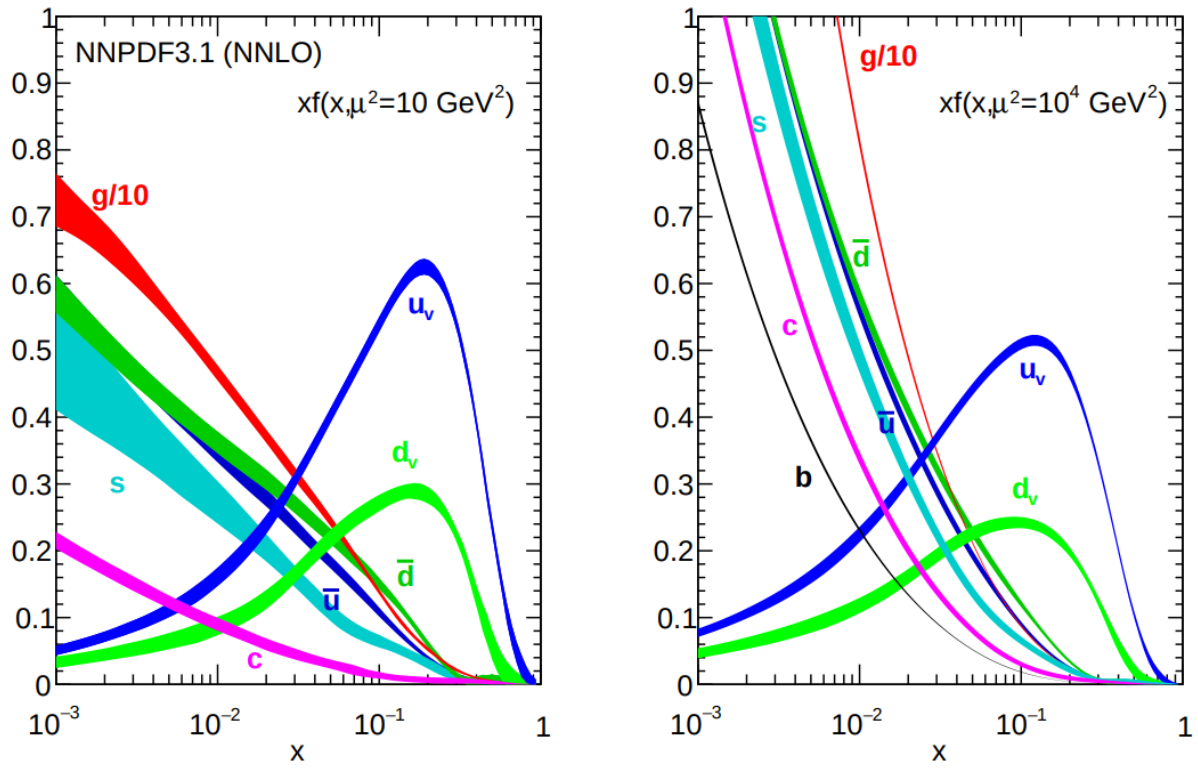

Figure 2.2: Parton distribution functions from the NNPDF collaboration at different values of Bjorken-x and factorisation scale, μ^2. On the y-axis is $xf(x, \mu^2)$. From Ref. [33].

the up and down quarks are similar, and the production of sea quarks is dominated by QCD, it can also be assumed that the contributions from up and down sea quarks are equivalent. Combining these equivalencies gives

$$u_s = \overline{u}_s \approx d_s = \overline{d}_s. \tag{2.54}$$

There are then also sea quark contributions from the other quark flavours and gluons. The complete set of PDFs is also constrained such that

$$\sum_i \int_0^1 x f_i(x)\, \mathrm{d}x = 1, \tag{2.55}$$

which states the sum of momentum fractions carried by each particle must sum to one. Here $f_i(x)$ is the PDF for a particle of flavour i.

PDFs are determined using global fits to experimental data. For example, the values calculated from the NNPDF collaboration [33] are shown in Figure 2.2. It can be seen

that the gluon contribution is largely dominant, and as such gluon-gluon interactions make up most of the collisions at the LHC.

These PDFs can then be used alongside the factorisation theorem and asymptotic freedom to make predictions for proton-proton collisions. To calculate the cross-section to produce some final state at the LHC, firstly the calculation for each possible initial state is made. These are then weighted by the probability of having the partons of the correct energy in the proton-proton collision. The cross-section for a process is then expressed mathematically as

$$\sigma(P_1, P_2) = \sum_{i,j} \int_0^1 \mathrm{d}x_1 \int_0^1 \mathrm{d}x_2 \, f_i(x_1, \mu^2) \, f_j(x_2, \mu^2) \, \sigma_{ij}(p_1, p_2, \boldsymbol{\lambda}(\mu^2), Q^2, \mu^2), \qquad (2.56)$$

where $\sigma(P_1, P_2)$ is the cross-section for a given process starting with two protons of momenta P_1 and P_2. The two interacting partons are of type i and j and have Bjorken-x values of $x_{1,2}$. All possible combinations of initial state partons are summed over. The cross-section for the parton level process is given by σ_{ij}. This cross-section depends on: the parton momenta, p_i; the coupling constants of any interactions in the process, represented here by $\boldsymbol{\lambda}$; the energy scale of the interaction, Q^2; and the renormalisation and factorisation scales. These scales are set here to the same value μ^2. The parton distribution functions for the partons are $f_{i,j}$ and are also dependent on the scale μ^2.

The PDFs can be seen to depend on the factorisation scale that is chosen. This is due to the fact that as the factorisation scale changes, the radiation that is considered to be a part of the proton also changes. This must be factored into the PDFs as is shown in Figure 2.2. It can be seen that as the factorisation scale increases, the relative contributions from the valence quarks decreases, and the contributions from sea quarks increases. This can be explained by the increase in radiation that is considered to be a part of the proton with the higher energy threshold.

2.4 Jets

Proton-proton collisions often lead to the production of highly energetic partons. These partons will then emit radiation, in the form of other partons, in a process called showering. As a parton showers, it loses energy and at some point the partons belonging to the shower drop below the energy scale which separates the perturbative and non-perturbative regions of QCD. At this point the size of the strong coupling constant becomes large and

the partons can no longer be considered as individual particles. Instead, they join together to form hadrons in a process called hadronisation. For sufficiently energetic partons in the final state, these showers are highly collimated and are referred to as jets.

To compare jets that are detected with theoretical predictions, the concept of a jet needs to be more rigorously defined. Procedures for deciding which final state hadrons form a jet are called jet clustering algorithms. The same algorithms can be used in theoretical predictions as well as in detector experiments to ensure that the results are directly comparable.

There are several potential ways to achieve this clustering, but it is vital that the algorithms make physical sense. One key point is that the algorithms must exhibit infrared and collinear (IRC) safety. This is the concept that the theoretical predictions of hard jets should not change if extra emission of soft particles or collinear splitting of partons is added. Examples of such an approach can proceed as follows:

1. Calculate the distance between all pairs of particles, d_{ij}, and the distance between each particle and the beam, d_{iB}.

2. Find the smallest of all the calculated distances,

$$\min(\{d_{ij}\}, \{d_{iB}\}),$$

   if this corresponds to the distance for a particle pair then the two particles are combined into one overall particle. Summing the four momenta of the particles is the typical way to do this. The merged particle replaces the pair. If instead, the minimum distance corresponds to a particle and the beam, then this particle is declared a jet and removed from the list of particles for further clustering.

3. Go to Step 1 and repeat until no particles remain.

The distance measures are defined differently for different algorithms, but for the commonly used k_T family of algorithms the distances are defined as

$$d_{iB} = \left(p_{Ti}^2\right)^n \; ; \quad d_{ij} = \min\left[\left(p_{Ti}^2\right)^n, \left(p_{Tj}^2\right)^n\right] \frac{(\Delta R_{ij})^2}{R^2}, \tag{2.57}$$

for some integer n. Here p_{Ti} is the transverse momentum of particle i and ΔR_{ij} is the distance between the two particles in η–ϕ space, where η is the pseudorapidity and ϕ the

azimuthal angle. This distance is defined by

$$\Delta R_{ij} = \sqrt{(\Delta\eta_{ij})^2 + (\Delta\phi_{ij})^2}. \tag{2.58}$$

Finally, R defines the jet radius in $\eta - \phi$ space. Changing the parameter n gives different algorithms: $n = 1$ gives the k_T algorithm [34]; $n = 0$ is the Cambridge–Aachen algorithm [35]; and $n = -1$ gives the anti-k_T algorithm [36]. These algorithms are all IRC safe. If, for example, a particle undergoes collinear splitting, the two daughter particles will have $\Delta R_{ij} = 0$, and so will be recombined in the algorithms. As well as this, if a hard parton radiates a soft parton with a non-zero angular separation the emission will either be clustered last back into the same jet, or it will create a new jet containing only itself. Neither of these will affect the properties of the jet, due to the emission being soft. Another IRC safe algorithm is the SISCone algorithm [37]. The basis of this is to define jets with fixed radius R around particles in the event, which are then iteratively updated following checks for the stability of each jet. This algorithm is also IRC safe.

The shapes of jets formed using the different algorithms are compared in Figure 2.3. One nice feature of the anti-k_T algorithm is that the jets are mostly highly conical in shape, with a radius close to that defined by R. This is due to the fact that highly energetic particles are clustered first, with soft partons being included up to this radius. For soft partons at values $\Delta R_{ij} > R$ the jet will have a distance measure which is closer to the beam than to the soft partons, and hence no further partons will be clustered into the jet. This is the most commonly used algorithm at the LHC.

2.5 Making theoretical predictions

For it to be possible to test a theory and its mathematical underpinning, it must provide testable predictions. For the SM, this includes predicting the results of measurements made at particle colliders, such as the LHC. In the modern day, SM predictions are typically handled computationally. One such way to make predictions is the use of Monte Carlo event generators. These programs aim to simulate the scattering events at colliders, with the name Monte Carlo taken from the statistical technique used in the generation. An overview of the required steps to make predictions of SM processes using event generators is given here, with Pythia 8 [38] followed as a specific example. This is due to the fact that Pythia 8 is widely used by the LHCb collaboration.

Event generation begins with an evaluation of the cross-section for a particular final

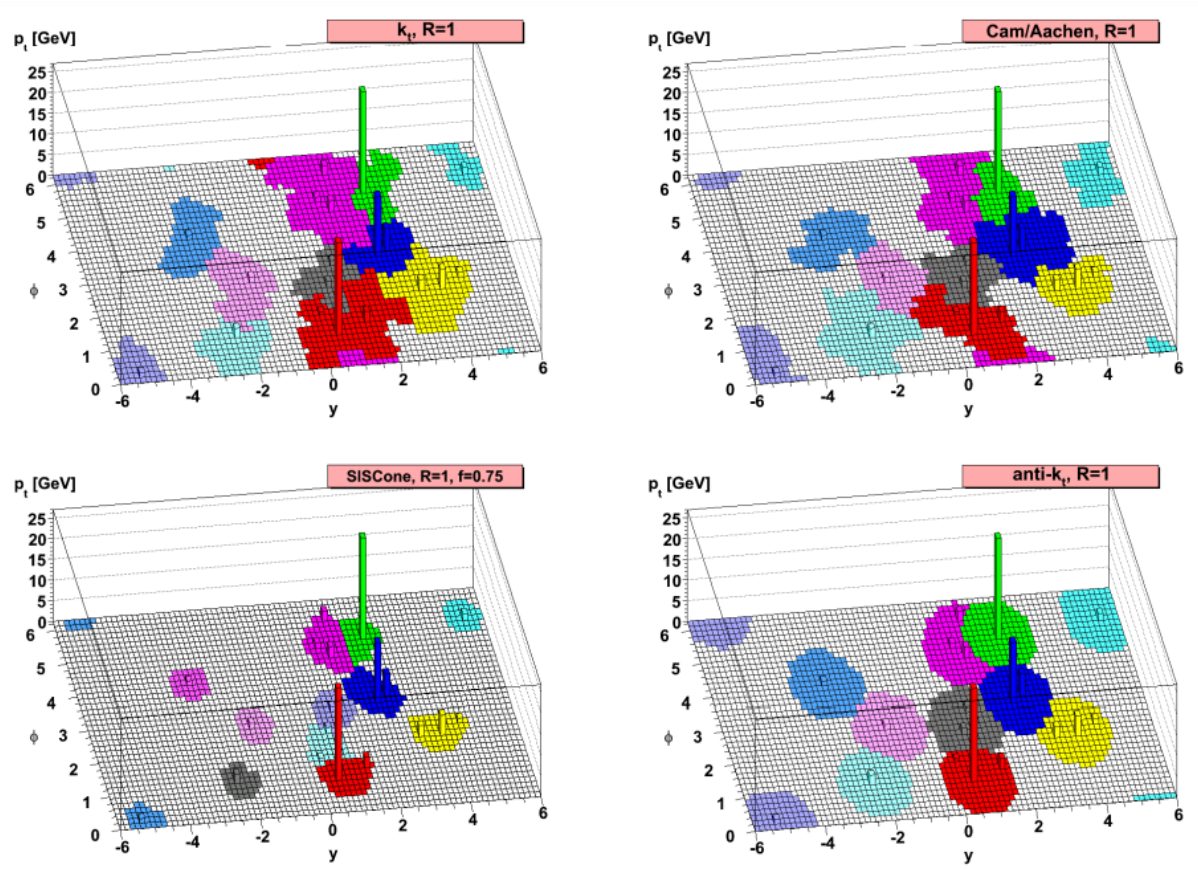

Figure 2.3: Jet clustering is shown for four different clustering algorithms. The simulated event is the same for all four algorithms so the shapes of the jets can be directly compared. From Ref. [36].

state to be reached from a relevant parton-parton collision. This cross-section is considered up to a given order in perturbation theory, for Pythia 8 this is a leading order (LO) calculation. The exact process that this calculation corresponds to is referred to as the "hard process". In Section 2.3, a mathematical description of proton-proton collisions and the hard process in an event at the LHC is given. The reality of an event at the LHC is, of course, more complicated than just the hard process. The effects of additional partons being emitted following the hard process must also be considered. This can be handled using so-called "parton showers" [38]. Parton showers are responsible for simulating the emission of particles describing the effects of higher-order processes than those that are considered in the cross-section of the hard process. These parton showers are constructed recursively with a strict ordering. In Pythia 8, the highest p_{T} partons are emitted first [38]. Emission continues until no final state particle has an energy above a given threshold

value, this is 1 GeV for Pythia 8 [38]. The emissions are random in nature and so a wide range of final states is therefore simulated.

These final state particles are now of low energy and, as such, hadronise into bound states (or hadrons). The low energy of these partons means that pQCD cannot be used to predict this hadronisation process. Hadronisation is handled using the Lund String Model [39, 40] in Pythia 8. Unstable hadrons then undergo decay. This decay is typically modelled using measured branching fractions for particles. Decay channels can be assigned a weight proportional to their branching fractions, with these weights then used to randomly select a channel for the simulation [38].

It is also possible that the partons within the original proton-proton collision have more than one interaction. If these interactions are significant, the events are labelled as containing multiple partonic interactions (MPIs). These interactions are also modelled in event generators such as Pythia 8, and the products of these interactions undergo showering and hadronisation alongside the hard process. What is left of the two colliding protons after the removal of the hard process and any MPIs is called the beam remnant. This itself can undergo interactions, although these are of low energy (else they would be classified within an MPI) and hence are non-perturbative in nature. The distinction between MPI and beam remnant is somewhat arbitrary and dependent upon the model employed; these are merely labels used to describe different parts of the event. Beam remnants and MPIs are typically included together (along with other physics effects) as the "underlying event". This term refers to the multitude of physics processes that go on alongside the hard process. The modelling of the hard process and underlying event (followed by parton showers, hadronisation and decays) then gives a whole event. Jets can then be defined within this event, as described in Section 2.4.

Finally, it is also possible to simulate the response of the detector to the event. This is useful for enabling direct comparisons between data taken using a detector with theoretical predictions, as well as for analysing detector performance or understanding the sensitivity of measurements at a detector to particular processes. This requires careful simulation of how the particles interact with the detector. This is typically implemented using Geant4 [41, 42].

In this section, event generators such as Pythia 8 have been discussed. These are capable of providing LO predictions of the hard process being simulated [38]. Other event generators able to simulate processes at NLO include Powheg [43–45] and MadGraph5_aMC@NLO [46]. It is then possible to interface these with Pythia 8 in order to simulate the parton shower.

It is also possible to make predictions of processes using other methods that do not rely on the direct generation of events. For example, these can use numerical integration to calculate the integrals that describe the cross-sections of hard processes up to a given order in perturbation theory. These alternative methods still often rely on Monte Carlo techniques to evaluate calculations.

As shown in Eq. (2.56), the cross-sections that are calculated in these predictions are dependent upon the renormalisation and factorisation scales. These scales were introduced to deal with divergences in the calculations: the factorisation scale to deal with divergences from low energy radiation within the proton; and the renormalisation scale to deal with divergences associated with loops in the Feynman diagrams. If it were possible to calculate the cross-sections of processes to all orders in perturbation theory, the result would not be dependent upon these scales. It is therefore possible to gather information about the size of effects associated with higher-order terms not included in the perturbative calculation by varying these scales. This scale variation therefore provides an estimate of the uncertainty on the final theoretical prediction associated with the choice of scale and hence the higher-order corrections that are not included in the calculation. Theoretical calculations are usually carried out using the same nominal value for the renormalisation and factorisation scales (although there is no inherent reason that this must be the case). To estimate the uncertainty on a prediction, these scales are typically varied to half and double their values independently. This can be done using the seven-point scheme as prescribed in Ref. [47], where if one scale is doubled, the other cannot be halved simultaneously (leaving seven points at which to evaluate the cross-section). The maximal change in the final prediction then gives an estimate of the uncertainty on the measurement.

Chapter 3

LHCb Detector

The work carried out in this thesis uses data collected with the LHCb detector to study proton-proton collisions at the Large Hadron Collider (LHC). In this chapter, a brief overview of the LHC will be given as well as a more detailed description of the LHCb detector. In addition, the LHCb computing and software environments used to simulate events are set out. This chapter concludes with a discussion of how jets are reconstructed at LHCb.

3.1 The Large Hadron Collider

The LHC [48, 49] is located at the European Organisation for Nuclear Research (CERN) and is designed to accelerate and collide protons at a centre-of-mass energy of up to 14 TeV. In addition, the complex is used to accelerate and collide lead ions. The LHC is a 27km long, circular collider which spans the Swiss–French border near Geneva. Before being injected into the LHC, protons are first accelerated by a series of boosters and synchrotrons up to 450 GeV, as shown in Figure 3.1. The LHC then further accelerates the proton beams using 16 radio frequency chambers. The beam trajectories are controlled by 1232 superconducting dipole magnets, each producing a field of 8.3 T. The beams are collided at four crossing points. Quadrupole magnets are used to squeeze and collimate the beams, focusing them to a width of around 14 μm at the crossing points. At each of these, there is a particle detector designed with specific physics goals in mind. The ATLAS [50] and CMS [51] detectors are general-purpose experiments, with a focus on searching for direct evidence of NP in the form of heavy particles and for the exploration of the Higgs sector. These detectors were jointly responsible for the discovery of the Higgs Boson in

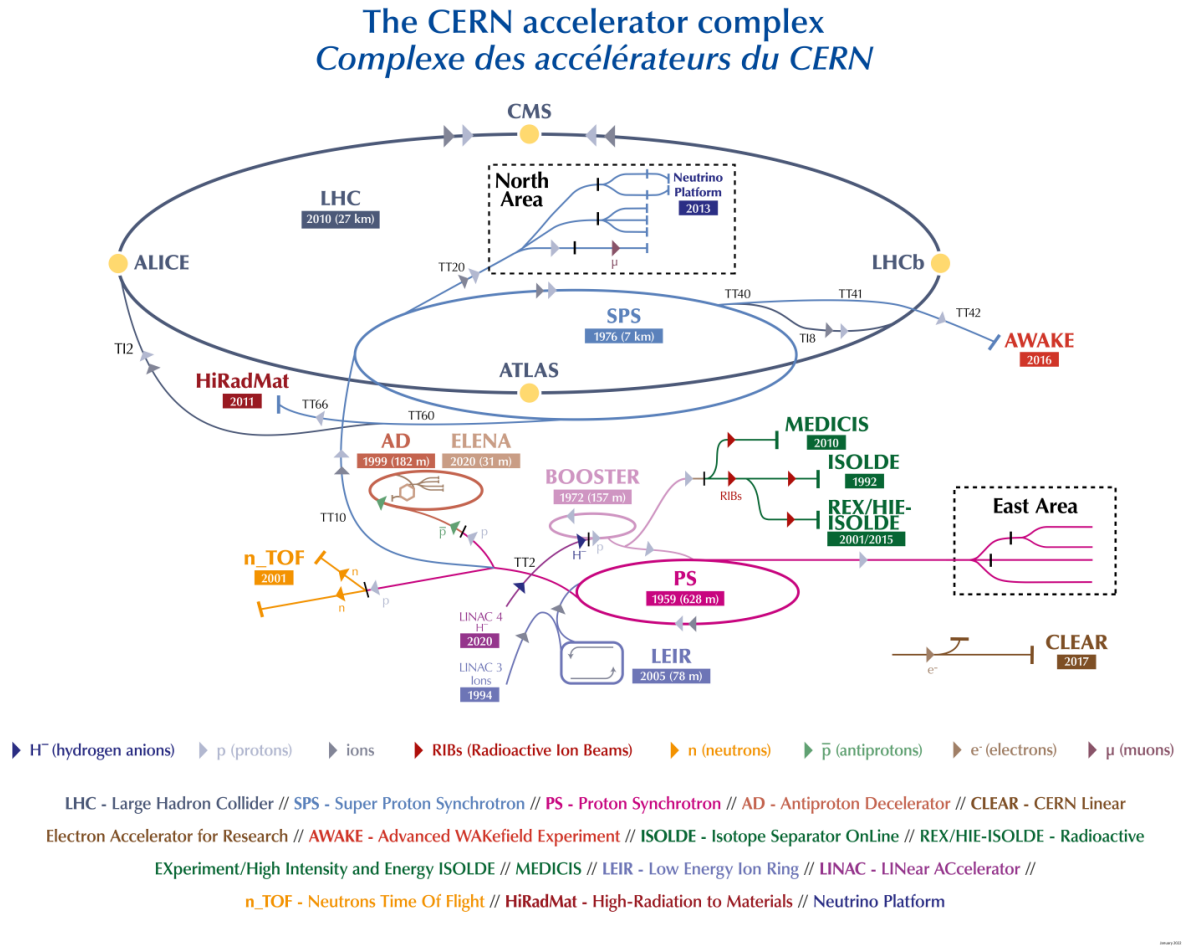

Figure 3.1: The CERN accelerator complex. The smaller machines are used in a chain to help accelerate the particles up to their final energies. Beams are also provided to a set of smaller experiments shown. From [54].

2012 [11,12] and have subsequently made precision measurements of the properties of this particle. The ALICE detector [52] is designed for the study of the Quark Gluon Plasma, utilising lead ion collisions for this purpose. Finally, the LHCb detector [53] is dedicated to the study of heavy flavour physics and is designed to look for indirect evidence for NP in CP violation and rare decays of hadrons containing b and c quarks.

3.2 Layout of the LHCb detector

The LHCb detector is a forward-arm spectrometer located at Point 8 of the LHC ring. The description given here follows Ref. [53].

A schematic of the full detector is shown in Figure 3.2, where the coordinate system is also defined. The z-axis points along the beamline in the forward direction, the y-axis points vertically upwards and the x-axis then points in such a direction to create a right-handed coordinate system. The angle between a particle and the beamline is referred to as θ, while the azimuthal angle in the x–y plane is referred to as ϕ.

Production of b quarks at the LHC[1] predominantly occurs through gluon fusion, gluon-gluon splitting and quark-antiquark annihilation [53]. The angles of the pair-produced b and $\bar{b}$ quarks are highly correlated and production is strongly favoured in the forward and backward regions [55], as is shown in Figure 3.3. The LHCb detector is therefore designed to capture a good proportion of these quarks, whilst being constrained by the limited space within the cavern hosting the detector. The detector covers the forward region of phase space, with an angular acceptance of $10 < \theta < 250\,\text{mrad}$ or, equivalently, a pseudorapidity range of approximately $2 < \eta < 4.5$. Electing not to also cover the backward region of phase space reduces the number of $b\bar{b}$ events within the acceptance by a factor of two, but this allows for a longer forward arm to be built. This in turn allows for a greater number of detector subsystems. Also, the longer lever arm allows for more precise measurements of the particle momenta to be made.

Proton-proton collision data were collected with the LHCb detector at a centre-of-mass energy $\sqrt{s} = 7\,\text{TeV}$ in 2011 and $\sqrt{s} = 8\,\text{TeV}$ in 2012, collectively known as Run I. Run II occurred between 2015 and 2018, with data collected at $\sqrt{s} = 13\,\text{TeV}$. The production cross-section for $pp \to b\bar{b}X$ within the LHCb acceptance has been measured as $\sigma(pp \to b\bar{b}X) = 72.0 \pm 0.3 \pm 6.8\,\mu\text{b}$ at $\sqrt{s} = 7\,\text{TeV}$ and $\sigma(pp \to b\bar{b}X) = 154.3 \pm 1.5 \pm 14.3\,\mu\text{b}$ at $\sqrt{s} = 13\,\text{TeV}$ [56]. Therefore, for an equal number of proton-proton collisions, there are roughly twice as many $b\bar{b}$ pairs produced at $\sqrt{s} = 13\,\text{TeV}$ as opposed to $\sqrt{s} = 7\,\text{TeV}$.

The following sections will detail the detector subsystems.

3.2.1 Vertex locator

The VErtex LOcator (VELO) surrounds the point where the proton beams collide. It is designed to locate the position where particles are produced and decay, via precise measurements of charged particle trajectories (tracks). Of particular importance is the ability to separate the primary proton-proton collision points (primary vertices, or PVs) from secondary vertices (SVs) where beauty and charm hadrons decay. These SVs are

[1]Here the general picture at LHCb is being described. The details of production methods focusing on the study of production asymmetries in this thesis will be covered in later sections.

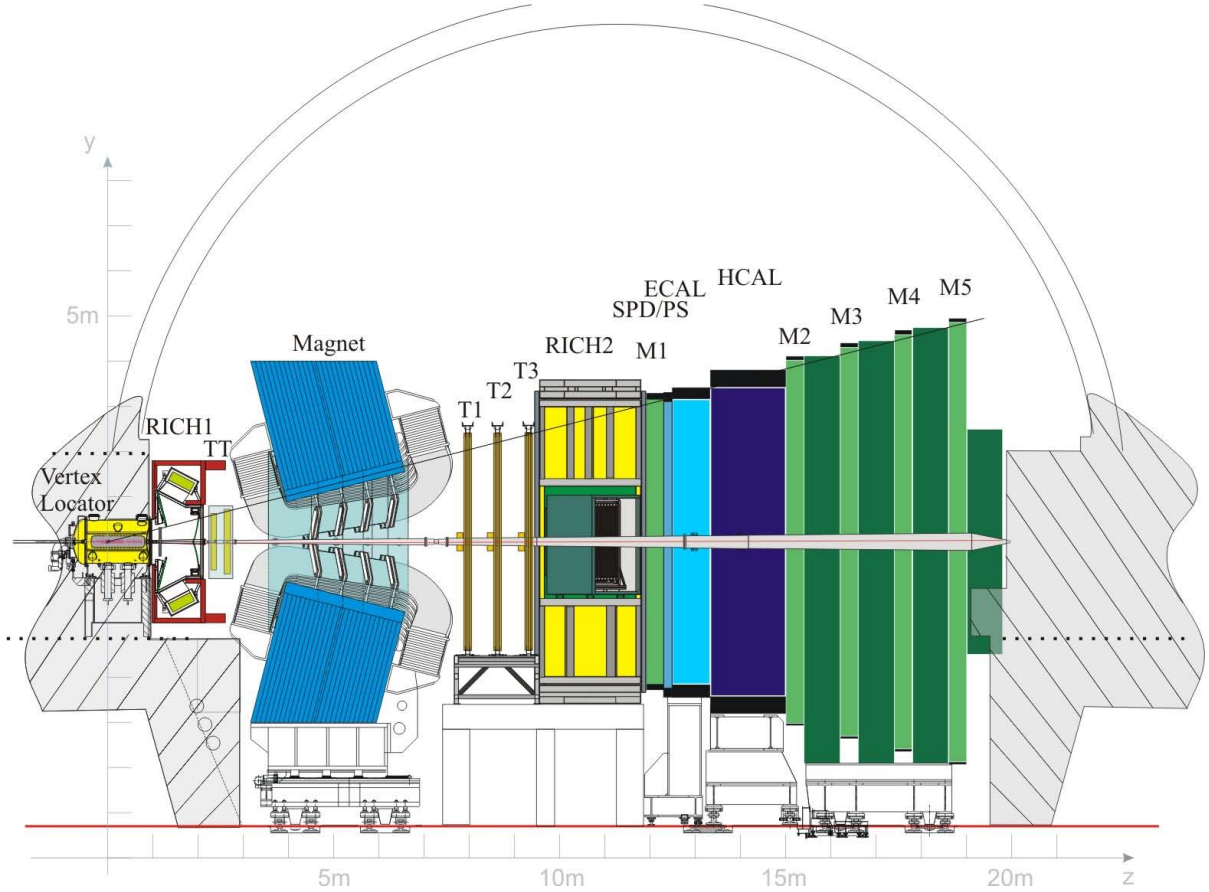

Figure 3.2: A cross-sectional view of the LHCb detector in the $y - z$ plane. The z-axis points along the beam line and the x-axis into the page. The subdetectors are discussed in detail in the main text. From Ref. [53].

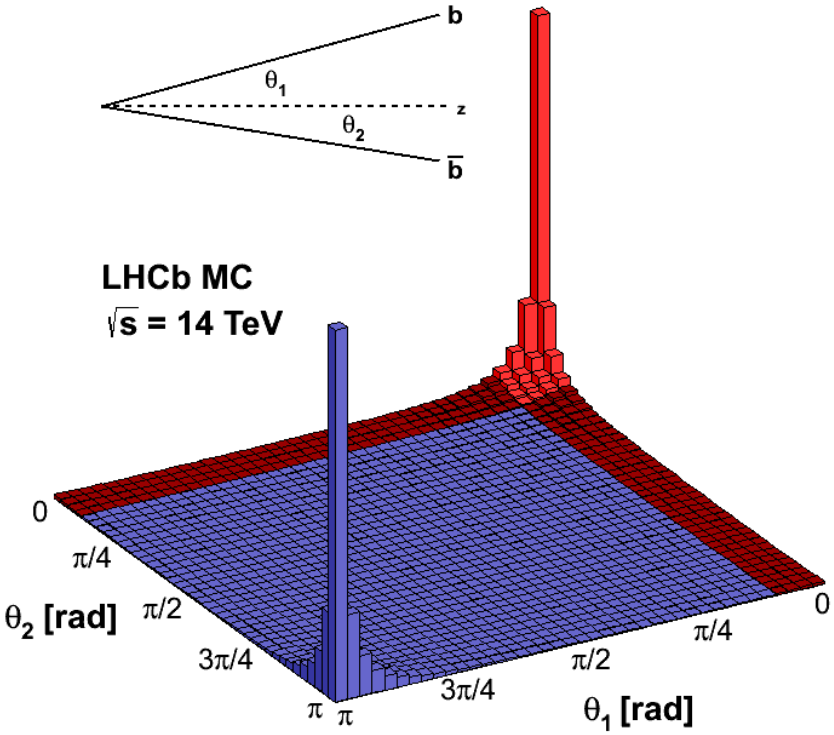

Figure 3.3: Production angles of all $b\bar{b}$ pairs at the LHC. The bins coloured in light red show the events where both of the quarks fall within the LHCb acceptance. Those bins in dark red show where only one quark falls in the acceptance, and bins in blue are where neither quark is in the acceptance. From Ref. [55].

present due to the heavy flavour hadrons travelling significant distances from the PV before decaying via the weak force. For B-hadrons, this distance is typically around 1 cm, whereas for D-hadrons it is typically around 4 mm.

The VELO, illustrated in Figure 3.4, is made up of 21 stations. Each of these stations in turn is made up of two silicon modules (left and right). These modules consist of two silicon strip sensors, one for providing information on the radial coordinate r and one for the azimuthal angle ϕ of a particle travelling through them. A charged particle travelling through the VELO deposits small amounts of energy in the sensors, called 'hits', which are read out by electronic chips associated with the sensors. This information is combined with further information from the other subdetectors to reconstruct the particle's path. There are an additional two stations backward from the interaction point, called pileup

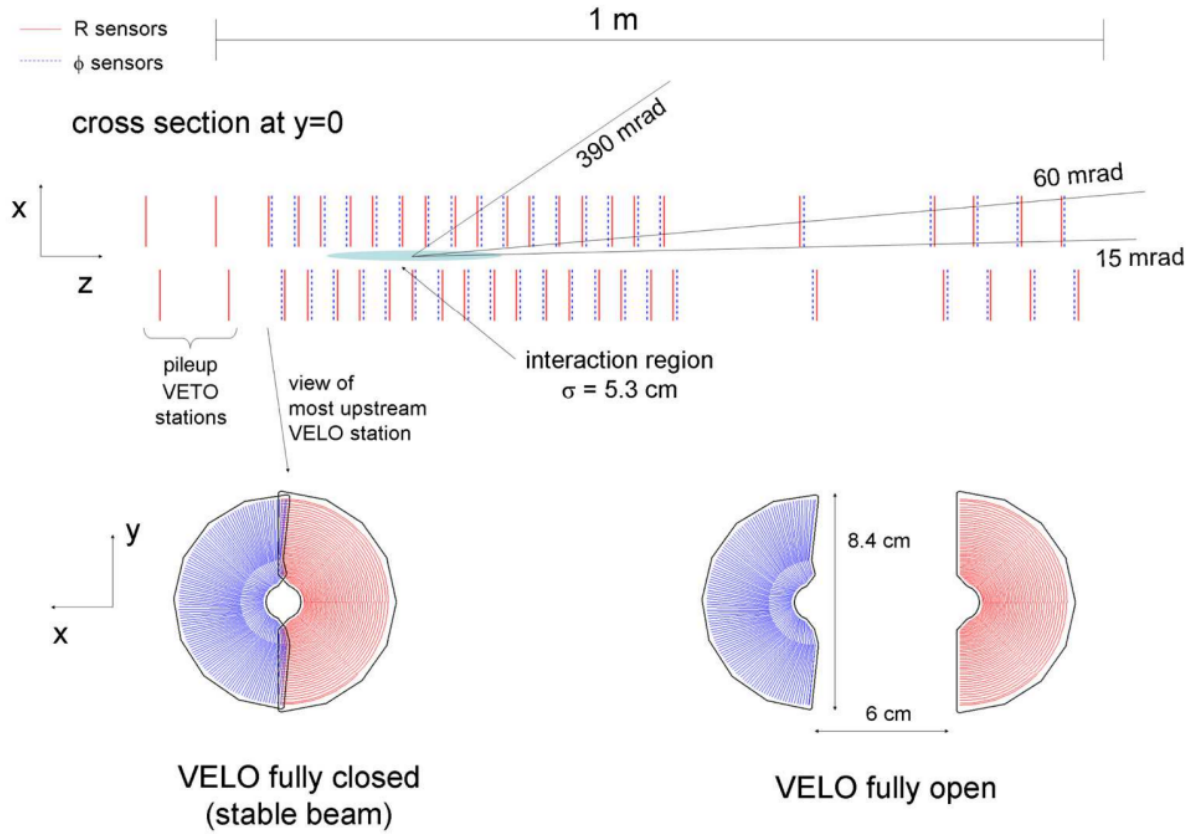

Figure 3.4: Top: cross-section of the VELO detector in the x-z plane, during stable beam conditions. Bottom: VELO in the open and closed positions. The VELO opens outside of stable beam conditions to protect from increased radiation damage. From Ref. [53].

stations. These are used to reject events with too many PVs.

When beams are first injected into the LHC they are wider than when stable running is achieved. To protect from radiation damage during the time between injection and stability, the VELO modules are designed to retract to around 3 cm from the beam axis. During stable running the VELO closes to a distance of 8 mm from the beamline. This closer proximity allows for more precise measurements of the vertices and impact parameters of tracks with respect to these vertices. The hit resolution of the VELO varies between approximately 4–40 μm. This gives a typical resolution on the position of vertices of 10–20 μm in the transverse plane and 50–100 μm in the z direction, as shown in Figure 3.5. Typical resolutions for the impact parameter of tracks to these vertices range from around 12 μm for very high momentum tracks to around 80 μm for low momentum tracks [57].

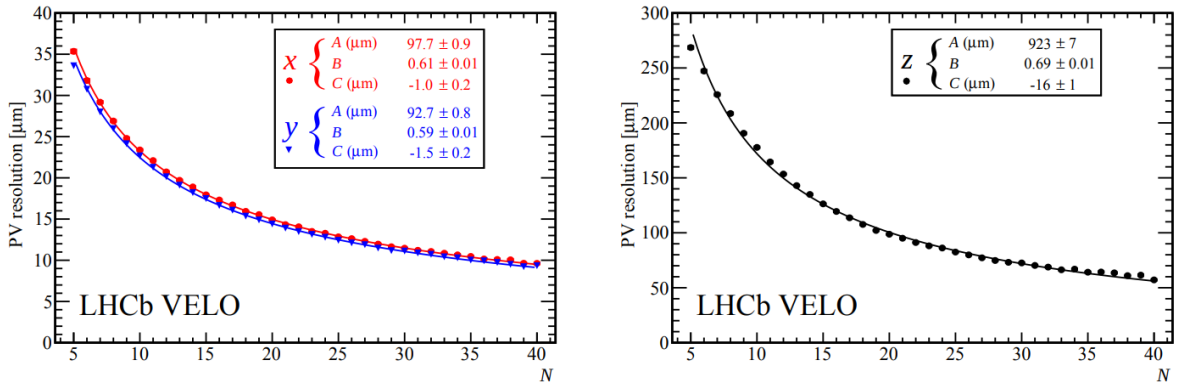

Figure 3.5: Resolution of the PV location in the VELO as a function of the number of tracks associated with the PV, N. Left: the resolution of PV location in x (red) and y (blue) directions. Right: the resolution of PV location in the z direction. From Ref. [57].

3.2.2 Tracking and magnet

The LHCb detector has several components that make up its tracking system, which are used to reconstruct the trajectories of electrically charged particles as they travel through the detector. These components are the VELO, the TT, and the T1, T2 and T3 stations [53]. The T1–T3 stations can be split into the inner tracker (IT) and the outer tracker (OT) with the IT lying closer to the beamline. The IT and TT are both silicon trackers and the outer tracker (OT) is a drift-time detector. Silicon trackers have a higher resolution than the drift-time detectors, meaning they are better able to resolve the increased number of hits that occur close to the beamline.

Between the TT and the T1 station, there is a dipole magnet which produces a magnetic field along the y-axis with a bending power of around 4 Tm. The highest magnetic field strength produced by the magnet is 1.1 T, occurring at the centre of the magnet. The magnetic field causes the charged particles to follow a curved trajectory, which allows for the determination of the momenta of particles (since the radius of curvature depends on the momentum of the particle). The direction of the magnetic field is reversed periodically to reduce the effects of any asymmetries occurring from oppositely charged particles bending in different directions.

The tracking system has excellent performance with an average efficiency of reconstructing tracks better than 95% for tracks with momenta, p, in the range $5\,\text{GeV} < p < 200\,\text{GeV}$ and pseudorapidity $2 < \eta < 5$ [58].

3.2.3 Cherenkov detectors and particle identification

The LHCb detector contains two Ring Imaging Cherenkov (RICH) detectors which are a crucial element of the particle identification (PID) system of the detector. Cherenkov radiation occurs when a particle travels through a medium faster than the speed of light in that medium. This results in the emission of a cone of light which has an angle θ_c from the direction of travel of the particle. This angle is related to the mass, m, and momentum, p, of the particle as well as the refractive index, n, of the medium through the formula

$$\cos(\theta_c) = \frac{\sqrt{m^2 + p^2}}{pn} = \frac{1}{n\beta}, \tag{3.1}$$

where $\beta = v/c$, and v is the velocity of the particle. The RICH detectors are able to determine the angles of emitted Cherenkov radiation and therefore provide information on the mass, given the momenta of particles travelling through the detector. The RICH1 and RICH2 detectors use different media with different refractive indices, optimised for different momentum ranges. The RICH1 detector uses C_4F_{10} gas which has a refractive index $n = 1.0014$, providing good PID in the momentum range $1 < p < 60\,\text{GeV}$. The RICH2 detector uses CF_4 gas with $n = 1.0005$ and provides good PID in the range $15 < p < 100\,\text{GeV}$. The reconstructed Cherenkov angle for different particle species in RICH1 is shown in Figure 3.6. Separation of the particle species is visible in the middle of the momentum range, with kaons and protons forming distinct clusters. As the momenta of particles become higher ($p \to \infty$) the mass of the particle becomes unimportant with $cos(\theta_c) \to 1/n$. This is called saturation and leaves particle species indistinguishable from one another. The information from the RICH detectors is combined with information from the calorimeter systems and the measured particle momentum to give a mass hypothesis for each particle and thus identify the species.

3.2.4 Calorimeters

The calorimeter system at the LHCb detector is comprised of four parts: the scintillating pad detector (SPD), the pre-shower detector (PS), the electromagnetic calorimeter (ECAL) and the hadronic calorimeter (HCAL). The different subdetectors all work using the same basic principle: as particles pass through them, energy is deposited which is measured by the detectors. The elements cause different particles to deposit energy in different ways, which allows for different species of particles to be distinguished. In

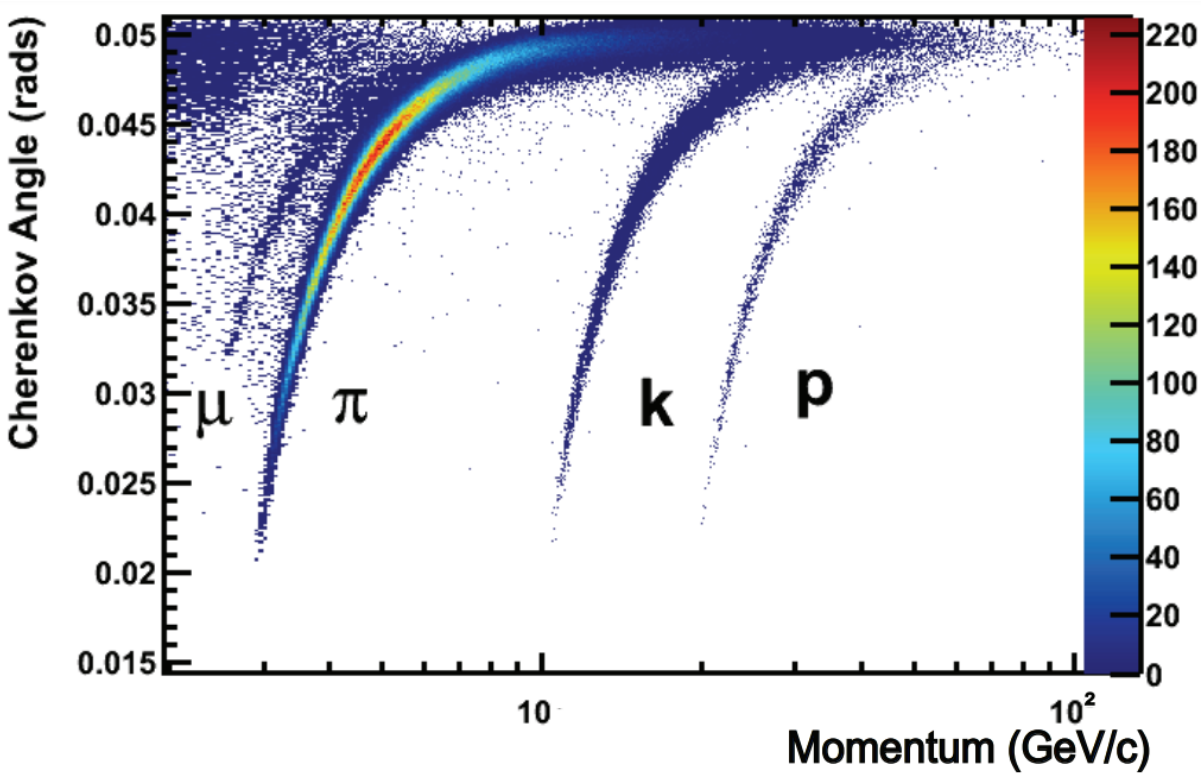

Figure 3.6: Reconstructed Cherenkov angle as a function of track momentum in RICH1 with the medium C_4F_{10}. Different particle species are labelled. From Ref [59].

particular, the LHCb calorimeter system allows for hadrons, electrons and photons to be distinguished from one another since they leave different signatures in the detectors, as shown in Figure 3.7.

Both the SPD and PS consist of scintillating pads with a thickness of 15mm and are the first two elements of the calorimeter system. They are interspaced by a lead converter with a thickness of 15mm equivalent to 2.5 radiation lengths for electrons. This lead causes photons to convert into electron-positron pairs, initiating an electromagnetic shower. This shower is measured by the PS detector. The SPD, sitting before the lead converter, typically only interacts with charged particles, so the SPD and PS together can be used to distinguish between photons and electrons.

The next element of the calorimeter system is the ECAL, which employs 'shashlik' technology of alternating scintillating tiles and lead plates. The ECAL has a length of 25 electron radiation lengths allowing it to fully contain electrogmagnetic showers. The ECAL consequently has an energy resolution of $\frac{\sigma_E}{E} = \frac{10\%}{\sqrt{E}} \oplus 1\%$ with E measured in GeV.

The HCAL is the final element of the calorimeter system and is designed for the

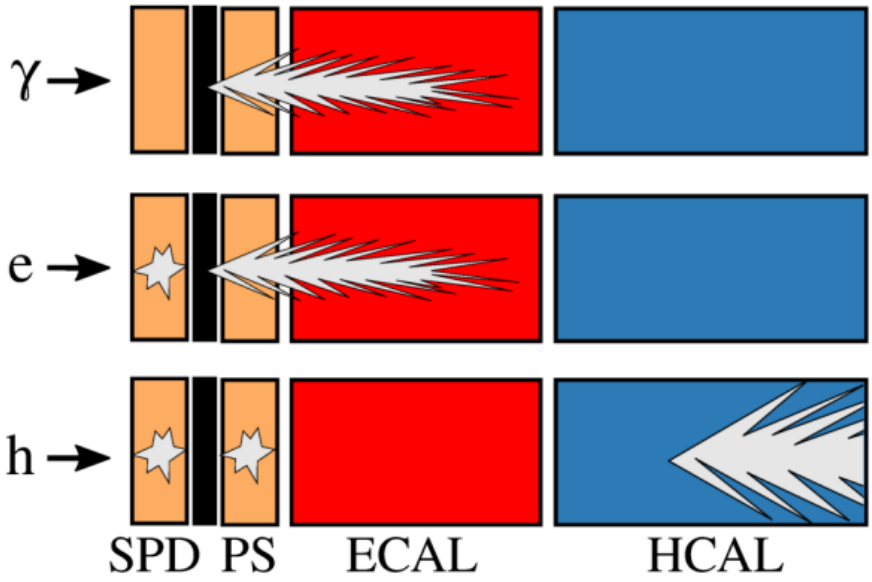

Figure 3.7: Illustration showing the energy deposits of photons (γ), electrons (e) and hadrons (h) through the calorimeter system at the LHCb detector. From Ref [60].

detection of hadrons. Similarly to the ECAL, it is made up of alternating scintillating tiles but with iron plates. It has a length of 5.6 hadron radiation lengths, which is set by the length of the LHCb cavern. Consequently, hadron showers are not always contained. This gives an energy resolution of $\frac{\sigma_E}{E} = \frac{70\%}{\sqrt{E}} \oplus 10\%$ with E measured in GeV. This is less precise than the ECAL but provides sufficient information for the LHCb trigger and for use in particle reconstruction.

3.2.5 Muon stations

Identification of muons forms an important part of this thesis due to their use in charge tagging beauty-jets, discussed later in Section 5.3. The muon system consists of five muon stations (M1–M5) with M1 situated before the calorimeters and M2–M5 behind the calorimeters as shown in Figure 3.2. Muons are able to penetrate a much higher amount of material compared to other particles. Therefore a particle that traverses the whole detector and leaves hits in the later muon stations is overwhelmingly likely to be a muon. It is this property of muons that therefore allows for their identification.

A side view of the muon stations is shown in Figure 3.8a. Separating the stations are 80 cm thick iron absorbers. These absorbers stop particles other than muons from travelling through all of the stations and hence filter these particles out. Muons with a

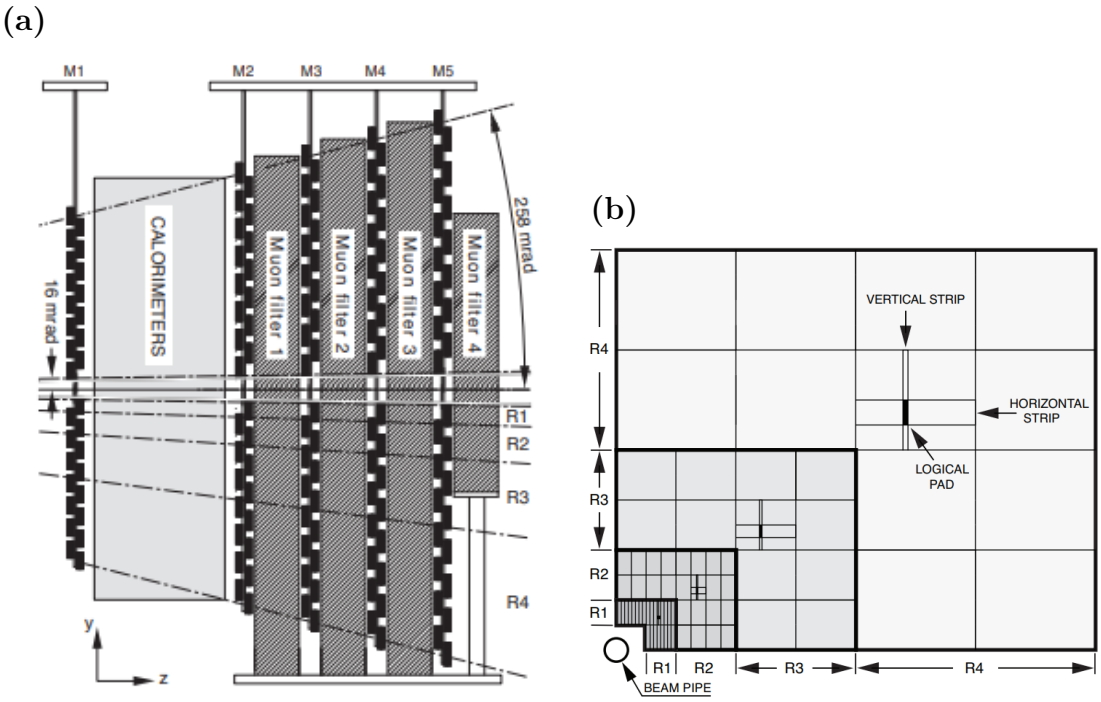

Figure 3.8: (a) Side view of the muon stations. Muon chambers are separated by 80 cm thick iron absorbers. (b) Cross-section of a muon chamber, with only one quadrant shown for simplicity. Rectangles show the 69 MWPCs. In M1 the 3 most central MWPCs shown here are replaced by GEMs. From Ref [53].

momentum greater than 6 GeV are able to successfully traverse all of the stations [61].

The muon stations are each made up of 276 multi-wire proportional chambers (MWPCs) with the exception of M1, where the inner 12 MWPCs are replaced with triple-gas electron multiplier (triple-GEM) detectors. This is due to the high radiation experienced in this region, which would exceed safety limits for the MWPCs. The make-up of the stations is shown in Figure 3.8b. The stations are broken down into four regions R1–4, with each region an increasing distance from the beam pipe. The MWPCs can be further segmented into logical pads, which define the partial resolution of the muon system. The size of these logical pads decreases closer to the beam pipe to help keep their occupancy consistent across the detector.

The MWPCs are filled with a gas mixture consisting of Ar, CO_2 and CF_4. Ionisation of the gas by muons passing through creates electrons which are collected and used to detect the muons.

The muon stations are important both for triggering events (discussed in more detail

Table 3.1: Stations which require hits to give `isMuon=TRUE` for a track. Table reproduced from [62].

Track momentum / GeV	Required stations hit
$p < 3$	Always false
$3 < p < 6$	M2 & M3
$6 < p < 10$	M2 & M3 & (M4 || M5)
$p > 10$	M2 & M3& M4 & M5

in Section 3.2.6) and for the identification of reconstructed tracks as muons. Whether a track is identified as a muon or not is described in the data by the variable `isMuon`. This is a binary variable and takes the value `TRUE` if tracks are associated with hits in the muon stations. The specific criteria depend on the track momentum [62] and are set out in Table 3.1.

Performance of the `isMuon` variable is excellent. The efficiency is around 97% for muons with transverse momentum of around 1 GeV, increasing to around 99% at higher transverse momenta. Kaon and pion misidentification is typically below 1% [62].

The charge of muon candidates can be determined by the direction with which their trajectory curves in the magnetic field. Incorrect reconstruction of the charge is incredibly rare due to the high magnetic field causing significant curvature.

3.2.6 Trigger system

The LHC provides proton bunch crossings at a rate of 40 MHz, or once every 25 ns. Not all of these result in visible proton-proton interactions, which occur at a rate of ~ 30 MHz at the LHCb detector [63]. Recording every proton-proton collision is not feasible[2]. Instead, a trigger is implemented to select which events are of sufficient interest for further study. These events are then written out and stored. In Run II, the LHCb trigger system reduced the detector output to around 12.5 kHz.

Events that meet the criteria of the trigger are said to have "fired" it and are subse-

[2]A typical event takes up 100kB of data. Therefore if every visible proton-proton interaction were read out and stored it would result in roughly 3TB per second being accumulated. If this were the case LHCb alone would produce in one day a similar amount of data to what is actually recorded by the four LHC detectors in a year.

quently classified according to what in the event is responsible for the selection. When looking at a candidate object within an event, the event is classified as TOS (Trigger On Signal) with respect to this candidate if it is responsible for the trigger firing. If instead the trigger is fired by a particle not belonging to the candidate object then the event is classified as TIS (Trigger Independent of Signal) with respect to the candidate.

The trigger is split into two stages: first, there is the low level trigger (L0) which is implemented through hardware; second is the high level trigger (HLT), which is software based. These will be discussed in detail in the next sections.

Low level trigger

The L0 trigger reduces the data rate to 1 MHz, a scale at which the whole LHCb detector can be read out. In order to achieve this, information from the calorimeters and muon systems is used, since both can be read out at high speed. The information from these systems is used separately to select events with high p_{T} muons or large transverse energy deposits in the calorimeters.

The muon trigger is implemented in four L0 muon processors, with each processor associated with a quadrant of the muon detector. Each of these processors identifies the two highest p_{T} tracks in its quadrant. As there is no exchange of information between the processors, muons that traverse multiple quadrants cannot be reconstructed in the L0 trigger. The processors look for hits through the muon stations that form a straight line and point back towards the interaction point in the y–z plane[3]. An estimate of the p_{T} is achieved by assuming that each track is a muon originating from the nominal interaction point and that it traversed the known integrated magnetic field. This gives a fast estimate of the p_{T} but with a poor resolution ($\mathcal{O}(25\%)$) compared to tracks that are fully reconstructed later in the LHCb analysis chain. The procedure yields eight muon candidates (two from each quadrant) from which selection criteria can be applied as shown in Table 3.2. For `L0Muon` and `L0DiMuon`, momentum requirements are applied alongside a limit in the number of hits in the SPD. By selecting only events with a low SPD occupancy, the complexity of the events selected is reduced, thereby allowing faster reconstruction in the subsequent software-based triggers. As the `L0DiMuon` trigger has a lower rate than the single muon line, the SPD criterion is relaxed. The final line, `L0MuonEW`, requires a higher momentum muon with no limit on the SPD hits. This is useful for analyses where the

[3]These lines are not expected to point back to the interaction point in the x–z plane due to the bending of particle trajectories by the magnet in this plane.

Table 3.2: L0 muon trigger requirements for 2016 data taking conditions. $p_{\mathrm{T}}^{(1,2)}$ are the highest and second highest p_{T} tracks found by the L0 muon processors [64].

L0 trigger line	Momentum threshold	SPD hits threshold
L0Muon	$p_{\mathrm{T}}^{(1)} > 1.8\,\mathrm{GeV}$	< 450
L0DiMuon	$p_{\mathrm{T}}^{(1)} \times p_{\mathrm{T}}^{(2)} > 2.25\,\mathrm{GeV}^2$	< 900
L0MuonEW	$p_{\mathrm{T}}^{(1)} > 6.0\,\mathrm{GeV}$	N/A

entire event is of interest, rather than just the decay of a particular heavy flavour hadron (for example, those concerning the production of particles). Selection requirements on the number of hits in the SPD can bias such analyses by removing events of interest in particular regions of the phase space. The exact momentum thresholds given in Table 3.2 are varied for the L0Muon and L0DiMuon triggers for different data taking conditions in such a way that their total output rate is consistent at around 400 kHz. The L0MuonEW trigger has a relatively low rate in comparison, and its thresholds were kept constant throughout Run II [64].

The calorimeter information is split into the hadron trigger (L0Hadron) and electromagnetic trigger (L0Electron and L0Photon). The former selects events that have deposited a significant amount of transverse energy, $E_{\mathrm{T}} > 3.7\,\mathrm{GeV}$, in the HCAL, and the latter selects events where electrons or photons deposit $E_{\mathrm{T}} > 2.4\,\mathrm{GeV}$ or $E_{\mathrm{T}} > 2.78\,\mathrm{GeV}$ in the ECAL respectively. The calorimeter trigger thresholds provided here are for 2016 data-taking conditions [64]. The output rates of the hadron and electromagnetic triggers are 450 kHz and 150 kHz respectively [63], where again the momentum thresholds were tuned throughout Run II to keep these output rates consistent. This brings the total output of the L0 trigger to around 1 MHz.

High level trigger

Events that pass the L0 trigger are then processed by the software-based trigger (HLT). Events are transferred to the Event Filter Farm, which consists of approximately 1700 computing nodes with 27000 cores. The HLT further reduced the event rate from 1 MHz to 12.5 kHz throughout Run II. The software trigger consists of two stages: HLT1 and HLT2.

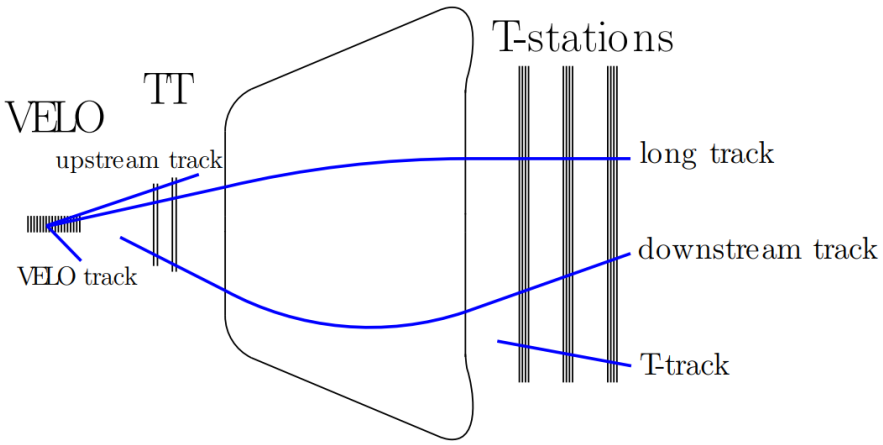

Figure 3.9: Sketch of different track types in the LHCb detector. From Ref. [64].

High level trigger 1

HLT1 partially reconstructs events in order to reduce the event rate to around 110 kHz [64]. Charged particles that have traversed the full detector are reconstructed. These are referred to as "long tracks" and are shown in Figure 3.9, along with other track types. Tracks reconstructed in the VELO are also used to determine the location of PVs. Full particle identification isn't possible owing to tight time constraints. Muon identification is the exception to this, however, due to the clear signature in the detector [64].

To reconstruct long tracks, hits in the VELO are first combined to form straight lines that roughly point back to the beamline [65]. This is followed by a straight-line extrapolation of these tracks to the TT location. If three hits are found in a small search window around this extrapolation then "upstream tracks" are formed [66]. Finally, these tracks are linked with further hits in the T1–T3 trackers, allowing long tracks to be reconstructed. These are subject to a transverse momentum requirement of $p_{\mathrm{T}} > 500\,\mathrm{MeV}$. The tracks are then fit using a Kalman filter [67, 68] and fake tracks[4] are rejected.

Events are selected to pass the trigger based on criteria called trigger lines. There are several such lines and those used vary from analysis to analysis. The specifics of the trigger lines used in this thesis will be discussed in Part II.

[4]Tracks may be faked by a random combination of hits that are aligned or a mismatch of segments from the different tracking stations.

High level trigger 2

Reconstruction in the HLT2 trigger is the first time a full event reconstruction is performed. This uses information from all of the subdetectors and is made possible by the low input rate to HLT2. HLT2 itself has an output rate of 12.5 kHz, with these events written out for storage and future data analysis [64]. The reconstruction used by HLT2 is the same as is used in the full offline analysis of data. There are three major steps to the full event reconstruction in HLT2: the track reconstruction of charged particles, the reconstruction of neutral particles, and PID.

While track reconstruction at HLT1 features a momentum threshold, the goal at HLT2 is to fully reconstruct all tracks in LHCb with no such requirement. The same general approach as HLT1 is followed at HLT2, however, lower momentum tracks are no longer filtered. Long-lived particles may decay outside the VELO and these are reconstructed by searching for tracks in the T-stations which are extrapolated backwards and combined with hits in the TT. Fake tracks are then rejected. This is achieved by applying a Kalman filter [67, 68] and using the output of this along with other track information to train a neural net to perform this rejection [69].

Information from the calorimeters, muon stations and RICH detectors is used in combination for PID. The RICH reconstruction takes all of the available hits in RICH1 and RICH2, from all of the reconstructed tracks, and analyses them simultaneously. The hypothesised species of each particle (and therefore mass) is varied, allowing the likelihood that a track is associated with a particular particle species to be found. The species considered are pions, kaons, protons, deuterons, electrons and muons. In addition, the calorimeter and muon systems also contribute to PID information. Energy deposits from the different calorimeter subsystems are combined to determine the total energy of each particle. Neutral particles can then be identified by their isolation with respect to the reconstructed charged tracks [64].

Similarly to HLT1, there exist multiple trigger lines implemented in HLT2. The details of the HLT2 lines used in this thesis will be discussed in Part II.

3.2.7 Data storage

The Data Acquisition (DAQ) system [70, 71] is responsible for interfacing with the detector, the Event Filter Farm and the HLT to transfer events to permanent storage. The DAQ system has a 40 TB buffer storage system as mitigation if the link to permanent

storage fails.

All events that pass the triggers are saved at CERN on tape. By the end of 2021 LHCb was responsible for 75.4 PB of stored data [72]. To allow for easier access to the data it is also stored on disk at CERN and other sites around the world. As disk space is limited the data first undergoes some selections in a process known as 'stripping'. Stripping lines consist of loose requirements being placed on the data which select events of interest for studies and ensure some level of quality in the candidates selected. Some groups of lines have a large overlap in the events selected, so are grouped together into streams. All events from lines within a given stream are then recorded together, reducing the number of times the same event is written to disk in multiple locations. In total LHCb had 39.2 PB of data written to disk by the end of 2021 [72].

In Run II the Turbo stream was also introduced [73]. This stream bypasses the stripping but only saves information about the particles responsible for firing specific HLT2 lines. This stream is therefore only used where information about the full event is not needed for the analysis.

3.3 Luminosity at LHCb

The instantaneous luminosity, $\mathcal{L}$, is a measure of how frequent proton-proton collisions are. The rate with which a given process occurs can then be calculated using the cross-section, σ, and the luminosity

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \mathcal{L}\,\sigma. \tag{3.2}$$

The nominal operating luminosity of the LHC is $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, however, the peak operational luminosity during Run II was recorded at double this [74]. At the LHCb detector a lower instantaneous luminosity is desired, as this enables the best separation of SVs from the proton-proton interaction points. In addition, the HLT is already dominated by beauty and charm meson decays; higher pileup[5] would require tighter requirements in the trigger and not lead to significant increases in signal yield recorded for many decay modes. Finally, the low pileup also reduces radiation damage to the LHCb detector.

To achieve this lower luminosity the proton beams are defocused in a process referred to as luminosity levelling. The amount of levelling is adjusted during a fill to keep the instantaneous luminosity in the detector roughly constant. This adjustment is required because the proton beams deplete throughout the course of a fill. The nominal instanta-

[5]The pileup is the number of proton-proton interactions in an event.

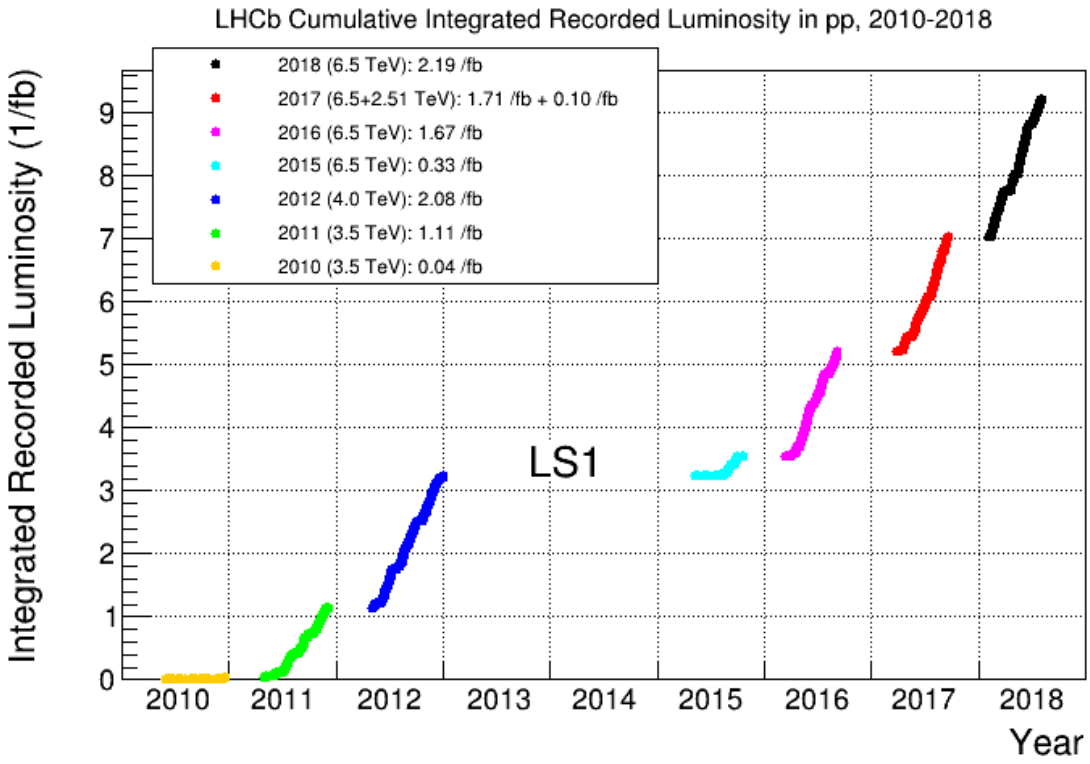

Figure 3.10: The cumulative integrated luminosity delivered to the LHCb detector. Source: LHCb.

neous luminosity delivered at the LHCb detector is $5 \times 10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The actual instantaneous luminosity delivered can be measured using methods such as a "van der Meer scan" which scans the overlap of the two beams at different beam positions [75]. This can then be integrated over time which gives a measure of the size of the dataset gathered by the detector. The integrated luminosity delivered to LHCb each year is shown in Figure 3.10. The luminosity levelling results in an average pileup of 1.3 at LHCb, significantly lower than that at CMS and ATLAS.

3.4 Generation of simulated events

The goal of a particle physics experiment is to make measurements that correspond as closely as possible to the 'true' values. In this description, 'true' means the values that would be obtained by a measurement under ideal conditions, where backgrounds, detector effects and other issues are completely removed. This is not possible in the real world, and the measurements that are made are imperfect observations of these true values. It then

becomes important to correct for these imperfections and infer the true underlying values from the observed one as much as possible. For this purpose, Monte Carlo simulations are performed, similar to those discussed in Section 2.5. This involves simulating everything from the initial proton-proton interaction to the subsequent particle decays and their interaction with the detector. The result of this provides simulated events with results of their properties as if read out by the real detector, along with the true underlying values that the particles are generated with. Comparison of these 'measured' and 'true' values can then be used to both quantify and control effects that influence the measured values.

Event simulation at LHCb is managed by the Gauss framework [76, 77]. The initial hard process of the proton-proton interaction is typically simulated using Pythia 8 [38], with Photos [78] used for the inclusion of QED final-state radiation (FSR). Pythia 8 is a general-purpose event generator simulating the hard processes at LO. Decays of beauty and charm hadrons are handled by EvtGen [79]. EvtGen is an event generator specifically designed by the BaBar collaboration for handling the decay of heavy flavour particles. It is important to use a generator more specifically designed for this purpose as this allows for the processes studied at LHCb to be more accurately modelled and hence provide more precise results. The detector geometry, material content and interaction with particles are simulated by Geant4 [41, 42]. Next, the simulation is digitised whereby the simulated energy deposits are converted into a detector response, using a package called Boole [80]. Simulated events are then handled in the same way as real data gathered by the detector.

3.5 Jet reconstruction

The charge asymmetry measurement presented in this thesis considers pairs of beauty-jets, referred to here as dijets. Therefore, in this section, the reconstruction of jets at LHCb is discussed. The formation of jets from a theoretical perspective is covered in Section 2.4.

Jet reconstruction begins with the standard particle flow algorithm used by LHCb [81–83]. The goal of the algorithm is to take information from all of the subdetectors and use it to provide a list of particles that can subsequently be clustered into jets.

First, the PVs in the event must be reconstructed. The process for this is detailed fully in Ref. [84] but is discussed very briefly here. The PVs are reconstructed by first finding points in the detector where there are a high number of tracks passing close together.

The Tukey biweight method [85] is then used to assign a weight to each track according to the value of its χ^2_{IP}. This variable measures the significance of the impact parameter with the estimated PV location. High values are penalised with a low, or zero, weight (as these are less likely to truly originate from the PV). Once the weights have been assigned, a weighted least squares fit is performed to determine the location of the PV.

Once the PVs have been reconstructed, the next step is to select all of the long and downstream tracks for a given PV[6], where these track types are shown in Figure 3.9. Standard LHCb track quality requirements are applied when selecting these inputs: the fit used to reconstruct tracks is required to return $\chi^2/\text{ndf} < 3$, where ndf is the number of degrees of freedom. Input from the detector systems is used to identify the particle species of each track, which is then used to set the mass of the associated particle. Due to the shorter length of upstream tracks, their momentum resolution is much poorer than that of long tracks, so their momentum is set to zero. Although this may sound extreme, it is acceptable as these tracks are low momentum (having been bent out of the detector acceptance by the magnetic field). Additionally, the jet clustering algorithms used observe IR safety, meaning very low momentum particles will not alter the final jet properties. These tracks are still included as they provide useful information on which proton-proton interaction produced the jet. If a group of tracks is compatible with a neutral resonance (often K^0_{S} or Λ) they are replaced as inputs with the appropriate resonance. This is because the resonant particle is more likely to point back to the PV compared to the individual tracks, meaning the particles are less likely to be assigned to the wrong interaction. This recombination has little effect on the jet energy [82].

The algorithm also uses neutral particles as input and so these must also be identified. The first step here is to select energy deposits in the calorimeter system that are isolated from charged tracks. The conditions to consider a deposit sufficiently isolated are listed in Table 3.3. After applying these criteria, neutral particles that are not as well isolated from charged tracks must also be identified. This is crucial because jets are highly collimated streams of particles, such that individual particles may not be well isolated but may still carry a significant amount of energy, meaning they must be "recovered" and included in the reconstruction. To achieve this, the expected calorimeter energy deposits for charged tracks are calculated using the track momentum and particle species hypothesis. These expected values are subtracted from the measured deposits associated with the tracks. If there is significant excess energy in the deposits then these are associated with neutral particles. To be considered as an input for jet reconstruction, this excess must be greater

[6]The full jet reconstruction is repeated for each PV in turn.

Table 3.3: Calorimeter cluster isolation requirements used in the standard LHCb particle flow algorithm. Here the χ^2 is a measure of the significance of the isolation of the cluster from any track, which is defined using the track and cluster coordinates as described in Ref. [86]. Table reproduced from Ref. [82].

Calorimeter	E_{T} Threshold / GeV	Isolation Criteria
ECAL	Any	$\chi^2 > 25$
HCAL	$2 < E_{\mathrm{T}} < 5$	$\chi^2 > 25$
HCAL	$E_{\mathrm{T}} > 5$	$\chi^2 > 16$
HCAL	$E_{\mathrm{T}} < 2$	Rejected

than 1 GeV [83].

At this point, the particle flow algorithm will have produced a list of charged and neutral particles in the detector for clustering into jets. The clustering is then performed using the FastJet [87, 88] interface. The anti-k_T algorithm [36] is used with a radius parameter, $R = 0.5$. Jets are typically reconstructed in the pseudorapidity range $2.2 < \eta < 4.2$ since this ensures that edge effects of the detector are reduced.

In simulated events, it is also possible to define truth jets. These are the jets that would be reconstructed under ideal conditions with a perfect detector. Inputs for clustering into truth jets are defined at hadron level (also referred to as particle level). Inputs are used if they are stable (with a mean lifetime greater than 10^{-8} s [83]), with the exception that neutrinos are not considered. Alternatives are possible, for example considering the fundamental partons produced. However, such approaches are no longer favoured, as they involve greater model dependence in correcting what the detector measures at hadron level.

Due to effects such as track-finding inefficiencies and imperfect neutral recovery, not all of the energy associated with a jet is reconstructed in the detector. It may also be the case that, owing to pileup[7], additional particles that originate from a different proton-proton interaction than the one of interest are clustered into jets [81]. It is therefore necessary to correct the measured energy of jets by scaling this quantity to obtain the best estimate of the true jet energy. This scale factor is determined from simulation as a

[7]Due to the significantly lower instantaneous luminosity at LHCb compared to the general-purpose detectors at the LHC, pileup is lower than experienced elsewhere. It is however still important to consider the effects associated with it.

function of the jet pseudorapidity, the fraction of the jet transverse momentum measured with the tracking systems, and the number of proton-proton interactions in the event. The scaling factor is typically between 0.9 and 1.1. This indicates that the base inputs are well measured, such that the subsequent required correction is small. Furthermore, previous studies have shown that the jet energy scale is well modelled in $b\bar{b}$-dijet events after this correction is applied [89]. The jet energy resolution is around 15% for jets with transverse momentum between 10 and 100 GeV as measured in simulation [81, 83], though it is typical for individual analyses at LHCb to smear this resolution to achieve good agreement between data and simulation [90].

There is an uncertainty associated with the derivation of the scale factor, which has contributions from different sources. There are systematic uncertainties associated with the method used and the flavour content of the jets. As well as this, an uncertainty is obtained by deriving a subsequent correction scale factor for jets that have already been corrected. Deviations from unity in this second factor are used as the uncertainty. Finally, there is a statistical uncertainty due to the limited simulated sample size. The relative size of these sources of uncertainty varies across the phase space of the jets and is discussed in more detail in Ref. [83], however the overall size of the uncertainty is typically less than 5%. There is an additional uncertainty associated with how well the simulation models the detector response to jets. This uncertainty is evaluated explicitly for the analysis discussed in this thesis in Chapter 6.

Jets are also identified by their flavour at LHCb. If a jet contains a b hadron it is labelled a b-jet, if it contains a c hadron and no b hadron then it is a c-jet and otherwise, it is a light-jet (initiated from a u, d, s or g). Together, b- and c-jets are referred to as heavy flavour (HF) jets. HF jets contain a secondary vertex (SV) due to the HF hadrons flying some distance from the PV before decaying via the weak force. For simulated b-jets used in this thesis, this distance is typically ~ 3 cm. A schematic of a jet containing a SV is shown in Figure 3.11. The tracks associated with the decay of the HF hadrons point to the location of this SV and not the PV. This feature is therefore utilised and HF jets are tagged by the presence of a SV within the jet.

A detailed description of the SV-tagger algorithm can be found in Ref. [91] but it is described briefly here, making note of important selection criteria. Tracks with significant p_{T} and displacement from the PV are used as input for the algorithm. Particle identification is not carried out and, instead, all particles are assigned the pion mass. All possible two-track SVs are built where the distance of closest approach between the tracks is less than 0.2 mm and where the vertex is of good quality. The two-body mass of the tracks

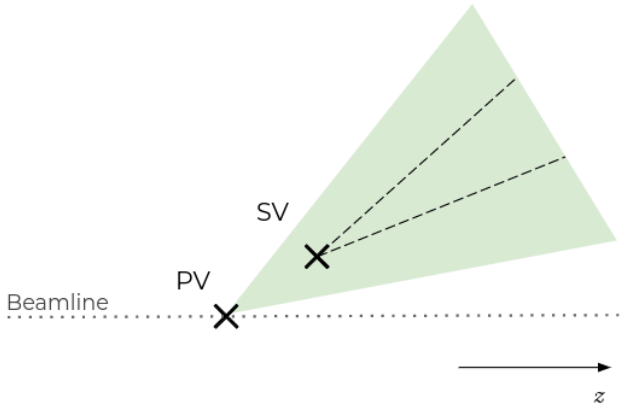

Figure 3.11: Diagram showing a jet tagged with a SV. The SV is due to the decay of a heavy flavour hadron after flying through the detector. The tracks associated with the HF decay therefore point to the SV location and not the PV. The PV and SV are shown as black crosses. The tracks making up the SV are depicted as dashed black lines and the jet is the green-shaded region, other tracks within the jet are not shown. This diagram is not to scale.

forming the SV must have a mass satisfying, $M > 0.4\,\text{GeV}$. This requirement removes SVs from most strange-particle decays, which would cause contamination of the HF jet samples. SVs that share a track are then combined to create n-body SVs. These SVs are in turn required to have a transverse momentum $p_{\text{T}} > 2\,\text{GeV}$, as well as significant spatial separation from the PV. The location of the SV is also required to be in a region that is compatible with the decay of HF hadrons. The longitudinal direction of the SV must be in the forward region relative to the PV: $z(\text{SV}) > z(\text{PV})$. Interactions between the detector material and particles are also able to fake a SV. These are suppressed by requiring that the flight distance of the SV from the PV divided by the momentum of the SV is less than $1.5\,\text{mm}/\,\text{GeV}$. This quantity serves as a proxy for the hadron lifetime. This also further suppresses strange-particle decays. If the SV contains only two tracks and the mass is compatible with the K^0_{S}, then the SV is rejected. Finally, if a SV is then found within the jet ($\Delta R < 0.5$) the jet is labelled as SV-tagged.

In Run II the efficiency of the SV-tagger is about 50% for b-jets and about 15% for c-jets. The probability of misidentifying a light-jet is about 0.5%. These results are estimated in simulated samples produced specifically for this thesis, which are discussed in Chapter 5. The efficiency for SV tagging jets increases with the momentum of the jet as this increases the likelihood that the SV passes the selection requirements. This effect is particularly noticeable for charm-jets.

Part II

Measurement of the charge asymmetry in beauty-dijet events

Chapter 4

Introduction to the charge asymmetry

Measurements in High Energy Physics aim to test the SM. If results are not in agreement with predictions from the SM this can indicate the need for NP models to explain them. Conversely, if results are in agreement with predictions then they can help to constrain NP models that are used to explain anomalous results elsewhere. Precise measurements of SM phenomena therefore serve as indirect searches, allowing investigation of the effects of NP without the need to directly produce NP particles.

Precision measurements of heavy-quark production play an important role in understanding the landscape of flavour physics. Such measurements have allowed precision tests of the SM and have led to constraints on parameters in the SM. The measurement presented in this thesis is associated with the production of beauty-dijets. In the SM, pairs of beauty quarks can be produced at LO via gluon fusion, or quark-antiquark annihilation via the Z boson, as shown in Figure 4.1. Each beauty quark in the pair then hadronises, leading to the formation of a $b\bar{b}$-dijet system. It is this system that is considered in this thesis.

This thesis seeks so measure the charge asymmetry in the $b\bar{b}$-dijet system, defined as

$$A^{b\bar{b}} \equiv \frac{N(\Delta\eta > 0) - N(\Delta\eta < 0)}{N(\Delta\eta > 0) + N(\Delta\eta < 0)}, \tag{4.1}$$

where $\Delta\eta = \eta_b - \eta_{\bar{b}}$ is the pseudorapidity difference between the jet containing the b quark and the jet containing the $\bar{b}$ quark. Events where the b quark has a greater pseudorapidity are referred to here as "forward" events, and those where the $\bar{b}$ quark has

Figure 4.1: Pair production of $b\bar{b}$ via gluon fusion (left) and via the Z boson (right). The gluon fusion diagram gives rise to no charge asymmetry in the produced $b\bar{b}$ pair since QCD terms that contribute to the asymmetry begin at $\mathcal{O}(\alpha_s^3)$ [92]. The production via the Z boson is a LO source of asymmetry.

a greater pseudorapidity are referred to as "backward" events.

One of the benefits of such a measurement stems from the inherent level of uncertainty present in QCD predictions, which is discussed in Section 2.5. This uncertainty means it is not possible to make cross-section predictions for many processes at hadron colliders to a high level of precision. However, large parts of these theoretical uncertainties in the calculations of $b\bar{b}$-dijet production cancel when evaluating the asymmetry. It is therefore possible to obtain a more precise prediction of this quantity, allowing for a more rigorous test of the SM.

There have been similar measurements at hadron colliders in the past. At the Tevatron, the D0 collaboration measured the asymmetry in B-hadron production [94], and the CDF collaboration measured the asymmetry in beauty-dijets [93]. For example, the CDF result is shown in Figure 4.2. The LHCb collaboration also made a measurement in proton-proton collisions at a centre-of-mass energy of 7 TeV [95]. These measurements were all made with precision at the level of $\mathcal{O}(0.5 - 1\%)$ and are consistent with the SM predictions. The measurement presented here constitutes the first measurement of the asymmetry in $b\bar{b}$-dijet production in proton-proton collisions at a centre-of-mass energy of 13 TeV at the LHC.

As well as these measurements at hadron colliders, a similar asymmetry was measured at LEP in e^+e^- collisions [23]. The production here was entirely through the Z boson and virtual photon, with no equivalent QCD contribution and therefore the measurement is not directly comparable with those above. However, the result is in tension with the SM at the level of $2 - 3\sigma$. This effect is shown in Figure 4.3, where the effective couplings of the b-quark have been extracted and compared to the SM prediction [23]. Given these tensions, it remains important to provide new measurements of the asymmetry to gain further information about this potential discrepancy.

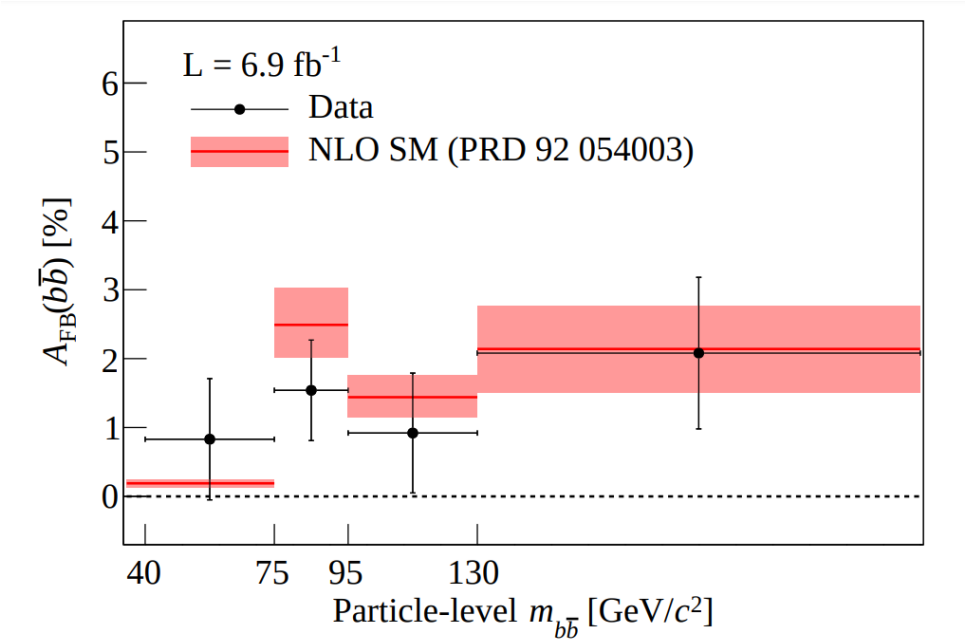

Figure 4.2: The charge asymmetry in $b\bar{b}$-dijet events at CDF. From [93].

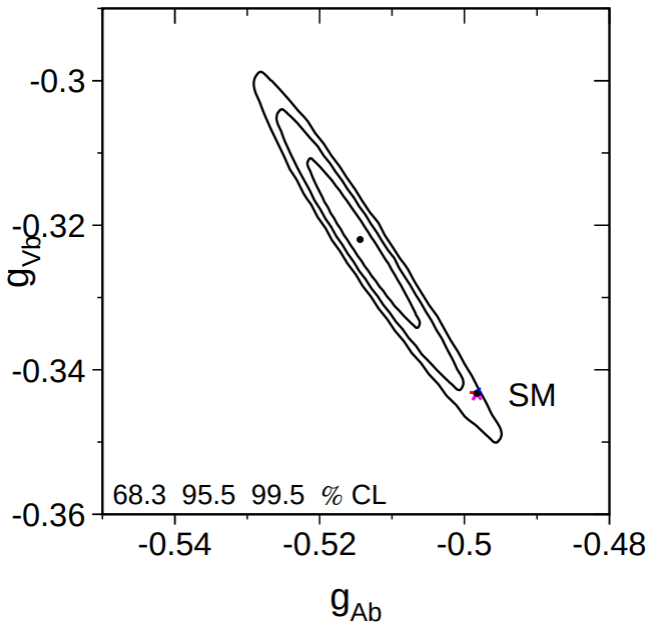

Figure 4.3: The effective vector and axial-vector couplings of the b-quark, measured by LEP [23]. The SM prediction is shown along with confidence intervals on the measurement. From [23].

The measurement presented here is sensitive to the couplings of the b quarks to both the Z boson and the gluon and so it is an important test of both electroweak and QCD theory. A measurement of the asymmetry is also an indirect measurement of some NP models which can induce contributions to the asymmetry [96–99]. These NP models often invoke the existence of new particles such as the axigluon, for example. In the future, the measurement could be combined with a similar measurement of the charge asymmetry in charm-dijet production. As laid out in Ref. [92], such a combination could lead to the most stringent constraints on the Zbb and Zcc couplings to date if there are also improvements in the theoretical understanding. Potential future work is discussed in more detail in Section 10.2.

The LHCb detector is well placed to provide a measurement of this charge asymmetry at the LHC. For a symmetric initial state there can be no observable charge asymmetry defined using the direction of one beam relative to the other. Consequently, in the central region at the LHC, where there is no "breaking" of this symmetry in the $\pm z$ direction, no asymmetry would exist. However, events that are within the acceptance of the LHCb detector have been boosted into the forward region. This requirement implies that the underlying partons in the initial state must have had an imbalance in their momenta, with the forward-going parton having a significantly higher Bjorken-x than that of the backward-going parton. By examining the PDFs in Figure 2.2 it can be seen that the forward-going parton is therefore more likely to be a valence quark, and the backward-going parton a sea quark or gluon. This breaks the symmetry of the initial state and allows a charge asymmetry to be defined and measured relative to the direction of the forward-going beam. The LHCb detector is unique amongst the detectors around the LHC ring in that its design provides dedicated high precision coverage of the forward region, making it an excellent place to make this measurement at the LHC.

Theoretical predictions have been made for the asymmetry at LHCb at NLO [92]. The LO terms are defined as the lowest-order terms that introduce a non-zero asymmetry, occurring at $\mathcal{O}(\alpha^2)$ or $\mathcal{O}(\alpha_s^3)$. There is also $b\bar{b}$ production from an $\mathcal{O}(\alpha_s^2)$ term, shown in Figure 4.1, but this gives rise to symmetric production and hence effectively reduces the value of the asymmetry measured. The forward region of the LHCb detector also benefits from having less of this production method relative to the asymmetric contributions when compared to regions covered by other detectors at the LHC, meaning that the asymmetric production methods are enhanced. The $\mathcal{O}(\alpha^2)$ term corresponds to $b\bar{b}$ production via the Z boson, and the $\mathcal{O}(\alpha_s^3)$ term corresponds to QCD processes such as flavour excitation. Examples of processes that contribute at this level are shown in Figure 4.4. Predictions of

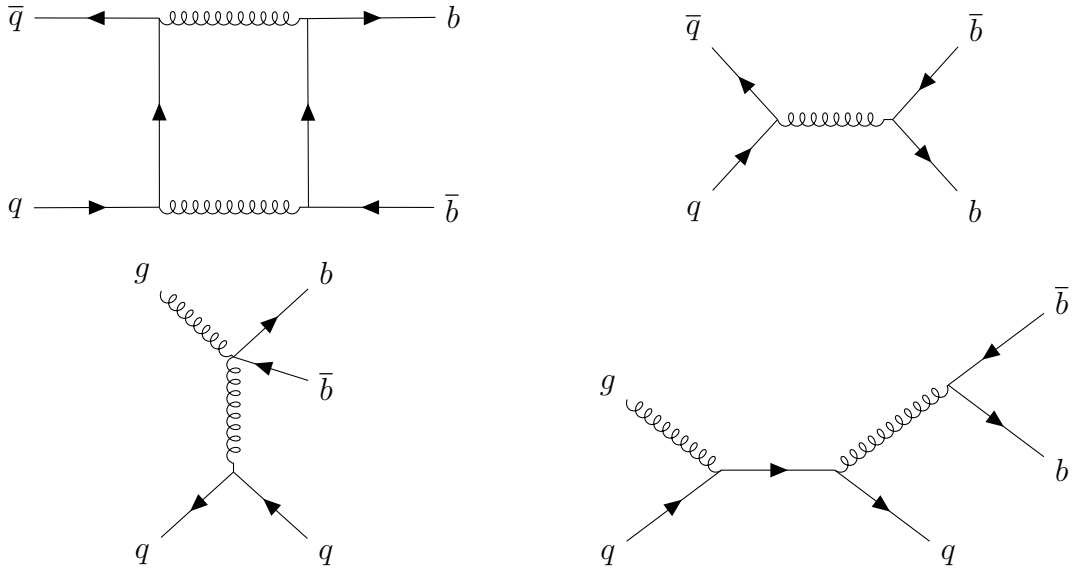

Figure 4.4: QCD processes begin to contribute to the asymmetry at $\mathcal{O}(\alpha_s^3)$. Shown here are examples of such processes. The top two diagrams generate an asymmetry via their interference whereas the bottom two directly generate an asymmetry [92].

the asymmetry are provided in Ref. [92] in bins of the invariant mass of the dijet system which is given by

$$m_{jj}^2 = (E_0 + E_1)^2 - \sum_{i \in x,y,z} (p_0^i + p_1^i)^2, \tag{4.2}$$

where $E_{0,1}$ is the energy of each jet[1] and $p_{0,1}^i$ is the component of the three momentum in the ith direction. This allows for the contributions via the Z boson and from QCD to be enhanced in different regions. For $m_{jj} \sim m_Z$ the dominant contribution to asymmetric production will occur via the Z boson [92]. Outside of this region, QCD effects will instead be dominant. The predictions obtained from Ref. [92] are shown in Figure 4.5 as a function of m_{jj}. These predictions have been made for two choices of the renormalisation and factorisation scales, which in both cases are varied using the seven-point scheme as prescribed in Ref. [47] to estimate the theoretical uncertainties[2]. The variation of scales for the estimation of theoretical uncertainties is discussed in Section 2.5.

[1]The jet energy is defined by the summing of the four momenta of the particles contained by the jet. The energy of these particles is determined in turn using PID information and momentum measurements.

[2]These predictions do not take into account uncertainties within the PDFs.

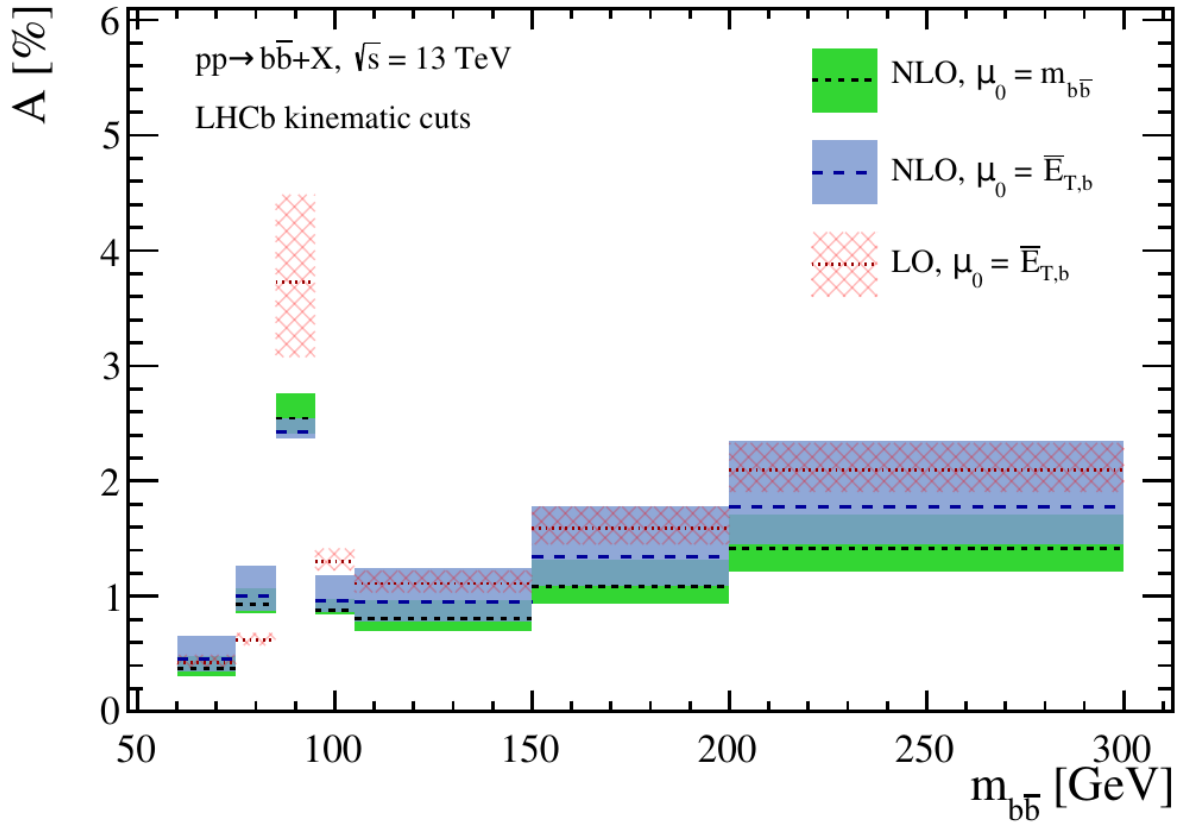

Figure 4.5: Theoretical predictions for the charge asymmetry in beauty dijet production at LHCb. Here the invariant mass of the dijet system is labelled $m_{b\bar{b}}$. Plot from Ref. [92].

4.1 Analysis strategy

This analysis is based on the data gathered at LHCb during 2016 at a larger centre-of-mass energy ($\sqrt{s}$=13 TeV) than the previous measurement at LHCb[3]. To make a measurement of the charge asymmetry in $b\bar{b}$-dijet production the jets associated with the b and $\bar{b}$ quarks

[3]The choice to only use the 2016 dataset as opposed to the full Run II dataset is well motivated. The reasons are discussed throughout the thesis when appropriate but are briefly laid out here. It is necessary to carry out significant calibration work on the simulated samples used in this analysis. This leads to the total systematic uncertainty that is of a similar size to the statistical uncertainty on the final measurement. Therefore increasing the data sample size will not impact the overall uncertainty significantly. As such, as well as providing an interesting test of the SM, this analysis aims to give a suitable understanding of the dijet systems used in the measurement and their behaviour within the detector, alongside the relevant work that must be carried out on the simulated samples to deliver this measurement. In this context, suggestions towards a more precise analysis (using the full Run II dataset) are made later in the thesis, using the lessons learned in this study.

must be detected, and the charge of at least one must be determined. The analysis is carried out in bins of the dijet invariant mass, with bin edges at $m_{jj} = [50, 75, 105, 150]$ GeV. These widths are set by the resolution possible with the LHCb detector. They are similar[4] to the bins used in the previous analysis [95]. The authors of Ref. [92] are thanked for providing theoretical predictions shown in Figure 4.5 in a form that allows the bins in the original paper to be combined. The steps followed in this thesis to measure the asymmetry for $b\bar{b}$-dijet production are set out here for reference:

- Data and simulated samples are prepared and events are selected. The jets in the event are also tagged to determine the charge of the originating b quark (distinguishing between b and $\overline{b}$ quarks), by directly measuring muons produced in HF decays in the jet. Jets containing a b quark will preferentially contain a negatively charged muon, while jets containing a $\overline{b}$ quark will preferentially contain a positively charged muon. The events can then be categorised as forward or backward depending on whether the b quark or $\overline{b}$ quark has a greater pseudorapidity.
- Simulated samples are calibrated using data to correct for mismodelling.
- Fits to the data are performed to determine the signal yield, using the separation of the decay location of the HF candidate in the jet from the PV to discriminate signal from background. Simulated events are used to populate template shapes for this fit. Forward and backward events are fit separately.
- These results are then unfolded to correct for detector effects. The detector response is modelled using simulated events.
- The unfolded distributions for forward and backward events are combined to make a measurement of the asymmetry.
- Mistagging the charge of a jet leads to a dilution in the value of the asymmetry. This is corrected for to give a final measurement of the asymmetry in bins of the dijet invariant mass.

Alongside the theoretical uncertainties being smaller for predictions of the asymmetry when compared to cross-section predictions, the data analysis approach taken here

[4]The lower bin edge has been increased from 40 GeV to 50 GeV in this analysis. In addition, there is the inclusion in this analysis of underflow and overflow bins for events that are below 50 GeV and above 150 GeV respectively. The asymmetry is not formally calculated in these bins. This is done to improve the handling of events that migrate into and out of the analysis due to detector effects during the unfolding process.

is similarly robust, featuring significant cancellations of potential systematic effects. Effects such as limited knowledge of the sample size (as determined from the integrated luminosity) completely cancel in asymmetry measurements, whereas this remains a large and significant uncertainty for measurements of production rates and cross-sections. Additionally, this analysis uses a final state that considers both a particle and its oppositely charged antiparticle (b and $\overline{b}$ quarks) in every event. This is in contrast to other asymmetry measurements that consider the difference between the decays of particles and antiparticles (such as measurements of CP-violation). This ensures that the effects of any detection asymmetries between matter and antimatter particles are minimised in this measurement. The matter/antimatter content of jets in this analysis is distinguished by locating muons from HF decays. The detection and reconstruction efficiency of muons within the LHCb detector is known to be charge symmetric to better than $\mathcal{O}(0.1\%)$ [100]. Furthermore, it is confirmed explicitly in Section 5.3 that the efficiency of tagging the jets with the presence of a muon is consistent in this analysis for jets initiated by b and $\overline{b}$ quarks. It is also possible to tag both forward and backward events with either charge of muon, such that any effect of an asymmetry in detecting muons with different charges (were it present) is further reduced. Finally, the magnetic polarity is periodically reversed as discussed in Chapter 3. This further suppresses any remnant asymmetries that could arise if the behaviour of particles and antiparticles differed across the detector [100]. The previous analysis of the charge asymmetry in beauty-dijet production at LHCb, studying 7 TeV proton-proton collisions, also considered such instrumentation effects and found their impact to be negligible [95].

The remainder of this thesis describes the work carried out in the bullet points above in order to make a measurement of the asymmetry. Chapters 5–8 discuss these steps in turn. Chapter 9 then considers systematic uncertainties and cross-checks and Chapter 10 discusses the final measurement and potential future improvements. Finally, concluding remarks are provided in Chapter 11.

Chapter 5

Samples and selection

This chapter details the samples that are used to make a measurement of the charge asymmetry in beauty-dijet production. This will be followed by details on how $b\bar{b}$-dijet candidate events are selected in data and simulation.

5.1 Samples

The analysis is performed using the 2016 dataset collected by LHCb of proton-proton collisions at a centre-of-mass energy of 13 TeV, corresponding to a total integrated luminosity of $1.7\,\text{fb}^{-1}$.

For the analysis, simulated samples of $b\bar{b}$-, $c\bar{c}$- and $q\bar{q}$-dijets (where q stands for a light parton u, d, s or g) are required. The simulated jets are required to:

- Model the distributions of observables related to SV tagging for the different flavours. These are used in the fit for measuring sample purity.
- Unfold detector effects from the measured distributions.
- Estimate and help optimise the performance of the charge tagging procedure.

The simulated events used in this analysis are produced using Pythia 8 at LO in pQCD, applying the standard process employed at LHCb, which is laid out in Section 3.4. This simulation has been extensively validated for use with the LHCb detector and furthermore this allows the use of results from other LHCb analyses within this work where applicable. It is noted here that whenever simulated jets are used, their properties are calibrated using data samples. These calibrations will be discussed throughout the thesis when

relevant. Results obtained in this analysis will be compared to state-of-the-art theoretical predictions [92], which have a much higher formal accuracy than the LO simulation of the hard event in PYTHIA 8 [38].

For $b\bar{b}$- and $c\bar{c}$-dijets, simulated samples are both produced with the condition that a muon is present in the final LHCb acceptance. This is due to the selection requirements placed on the data and simulated samples since jets are tagged for their charge by using muons from HF decays. The condition that a muon is present in the simulated event increases the fraction of events that pass this selection requirement and hence provides a more efficient way to generate simulated samples for the analysis. The condition for a muon to be present is not used for $q\bar{q}$-dijets, due to the fact that light-flavoured jets will very rarely contain a muon. The light-jets that are identified as containing a muon, usually do so due to the misidentification of some other particle as a muon. This happens rarely, which, together with the low rate of SV tagging for light-jets, means that light backgrounds to the analysis are negligible. This will be further demonstrated in Chapter 7. The simulated light-dijets are still required to confirm the low efficiency of the charge-tagging selection for this background, as is discussed in Section 5.3.

Further to this, additional $b\bar{b}$ samples are generated with stricter kinematic requirements at production to ensure a greater number of events remains within the LHCb acceptance. This helps to ensure that there is a sufficient yield available for the analysis. The requirements are shown in Table 5.1. The tighter production requirements for the b and $\bar{b}$ quarks are $24 < \theta < 270\,\text{mrad}$ ($2.0 < \eta < 4.4$). As the fiducial selections on the jets require that $2.2 < \eta^{\text{jet}} < 4.2$, the loss of any events that might be studied in the analysis by applying these tighter generation requirements is negligible.

All samples are produced in three bins of the Mandelstam variable $\hat{s} = (p_1 + p_2)^2$, where $p_{1,2}$ are the four momenta of the partons involved in the interaction. The bin edges used to produce this simulation are $\sqrt{\hat{s}} = [30, 70, 110, 13 \times 10^3]\,\text{GeV}$, where the upper limit is set by the centre-of-mass energy of the colliding beams. This ensures there is a significant number of events across the full mass range of the measurement, as this variable correlates significantly with the dijet invariant mass. When samples from the three mass bins are combined together they must be given weights that take into account that the cross-sections vary across these bins. All events in a given mass bin are assigned the same weight. These weights are determined by combining the simulated events and setting the weights such that the distribution of $\hat{s}$ is smooth.

The procedure for this is to fit the distributions of $\hat{s}$ for two samples across the mass boundary. The fit is performed to histograms of the simulated samples, with a bin width

Table 5.1: Details of simulated samples that are used in the analysis. Here, $\hat{s}$ is the invariant mass of the two underlying partons that generate the jets, θ(parton) and θ(muon) are the angles from the beam pipe that the partons that initiate the jets, and the muon in the event have.

Flavour	Generator requirements	Number of events	Weight
$b\bar{b}$	$\hat{s} \in [30, 70]\,\text{GeV};\ \ 24 < \theta(\text{parton}) < 270\,\text{mrad};\ \ \theta(\mu) < 350\,\text{mrad}$	39.2M	1.000
$b\bar{b}$	$\hat{s} \in [70, 110]\,\text{GeV};\ \ 24 < \theta(\text{parton}) < 270\,\text{mrad};\ \ \theta(\mu) < 350\,\text{mrad}$	32.4M	0.173
$b\bar{b}$	$\hat{s} > 110\,\text{GeV};\ \ 24 < \theta(\text{parton}) < 270\,\text{mrad};\ \ \theta(\mu) < 350\,\text{mrad}$	32.1M	0.044
$b\bar{b}$	$\hat{s} \in [30, 70]\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad};\ \ \theta(\mu) < 400\,\text{mrad}$	2.23M	1.000
$b\bar{b}$	$\hat{s} \in [70, 110]\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad};\ \ \theta(\mu) < 400\,\text{mrad}$	2.07M	0.153
$b\bar{b}$	$\hat{s} > 110\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad};\ \ \theta(\mu) < 400\,\text{mrad}$	2.02M	0.036
$c\bar{c}$	$\hat{s} \in [30, 70]\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad};\ \ \theta(\mu) < 400\,\text{mrad}$	2.39M	1.000
$c\bar{c}$	$\hat{s} \in [70, 110]\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad};\ \ \theta(\mu) < 400\,\text{mrad}$	2.04M	0.333
$c\bar{c}$	$\hat{s} > 110\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad};\ \ \theta(\mu) < 400\,\text{mrad}$	2.45M	0.049
$q\bar{q}$	$\hat{s} \in [30, 70]\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad}$	2.17M	1.000
$q\bar{q}$	$\hat{s} \in [70, 110\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad}$	2.15M	0.236
$q\bar{q}$	$\hat{s} > 110\,\text{GeV};\ \ \theta(\text{parton}) < 400\,\text{mrad}$	2.02M	0.306

of 1 GeV. For a given boundary the fit is performed on the final five bins of the lower mass sample. The simulated sample in the upper mass bin is then reweighted to minimise the mean squared error with the fit prediction for the first five bins. For both boundaries at 70 GeV and 110 GeV three fits are performed: a linear, a quadratic and an exponential fit. The best fit is then chosen by finding the fit with the lowest mean squared error in the five bins previous to the fit range. A linear fit is chosen over the boundary at $\sqrt{\hat{s}} = 70\,\text{GeV}$ and an exponential fit chosen at $\sqrt{\hat{s}} = 110\,\text{GeV}$. The final fits and rescaled simulation for the main $b\bar{b}$ samples are shown in Figure 5.1. The number of events, weights and production requirements for all simulated samples are shown in Table 5.1. For the purposes of calibrating simulated events, specific datasets are also utilised throughout this thesis. These will be introduced and explained as necessary.

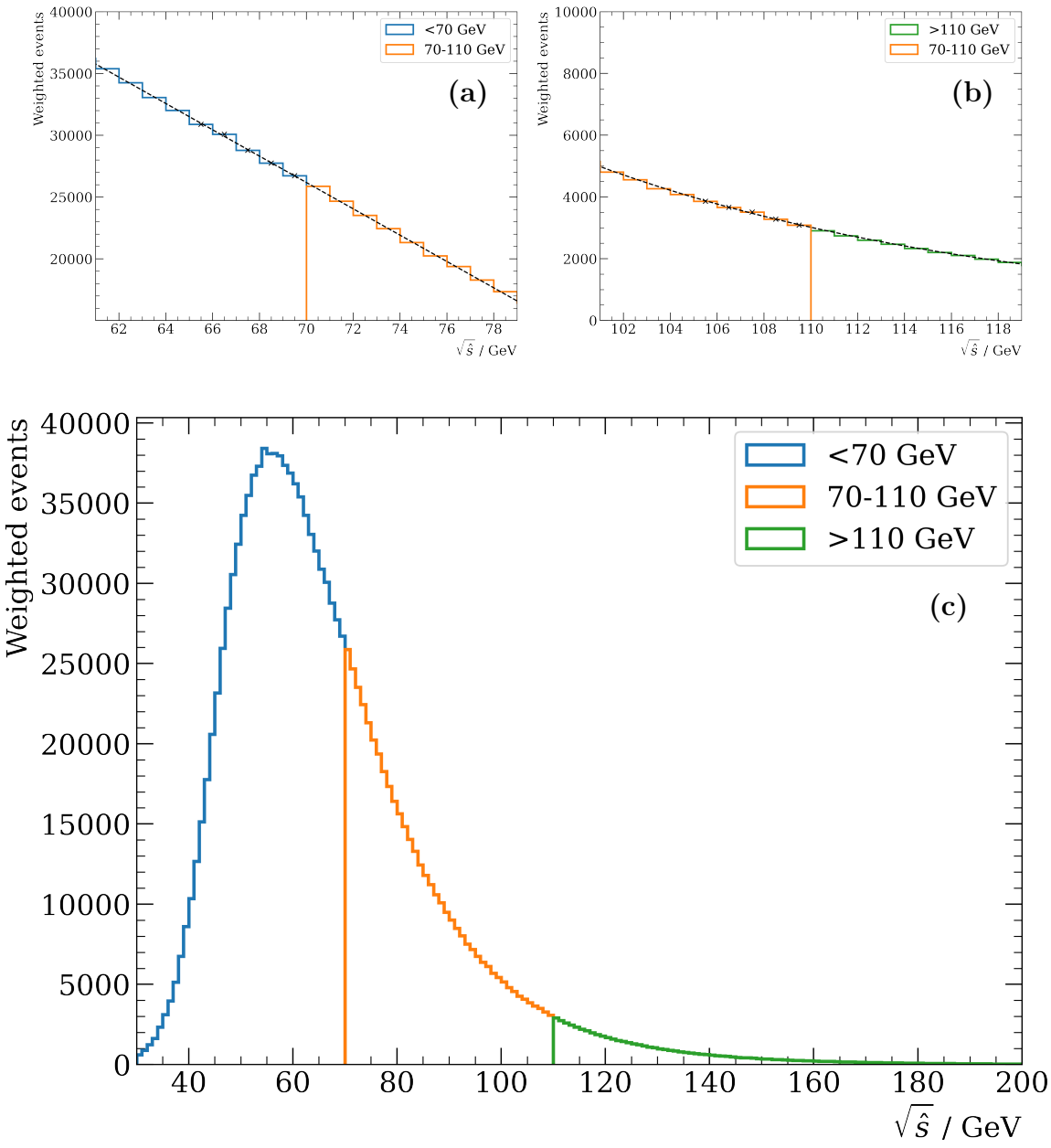

Figure 5.1: These plots show the weighting procedure for the $b\bar{b}$ simulation. This simulation is originally produced in three bins of the Mandelstam variable $\sqrt{\hat{s}}$ with bin edges of $[30, 70, 110, 13 \times 10^3]$ GeV, where the upper limit is set by the centre-of-mass energy of the colliding beams. As the number of events and the cross-section for production is different in each bin weights are applied to the upper two bins to produce a smooth distribution. This is done by fitting the binned data with a straight line over the boundary at 70 GeV and an exponential over the 110 GeV boundary. These fits are shown in plots (a) and (b) respectively. The full weighted distribution is shown in (c).

5.2 Selection requirements

The selection requirements that are common between data and simulated samples are first laid out. These can be broken down into two categories:

1. Fiducial selections. These are the selection requirements within which the measurement is defined. They are used in theoretical predictions of the measurement [92] and as such the exact choice of these selection criteria is already predetermined. Fiducial selections ensure that the jets are both experimentally and theoretically well defined such that appropriate comparisons with theory can be made.

2. Detector selections. To select beauty-jets, identify their charge and measure the charge asymmetry in dijet production it is also required to apply experimentally necessary selections. As these are not defined in the theoretical predictions of the measurement, care must also be taken that such selections do not bias the measurement of the asymmetry in a way which cannot be corrected accurately.

Selections from these two categories will be detailed before giving a description of the data and simulated samples.

Fiducial selections

The jets are subject to the following fiducial selections:

- $p_{\mathrm{T}}^{\mathrm{jet}} > 20\,\mathrm{GeV}$,
- $2.2 < \eta^{\mathrm{jet}} < 4.2$,
- $\Delta\phi^{\mathrm{jet-jet}} > 2.6\,\mathrm{rad}$.

The first two of these criteria ensure that the jets are well defined and fall within the detector. The last requirement was originally introduced in the first measurement of the asymmetry at LHCb, motivated by enhancing the contribution of non-gg production mechanisms [95]. It is also used in the theoretical predictions to which the measured results here are compared [92], hence it must be included here. Theoretical studies have since suggested that the threshold value for this requirement does little to impact the statistical sensitivity of the measurement to the value of the asymmetry [92]. The selection requirement is still useful however for two reasons. First, as set out in Ref. [92], some jet

separation is required as it avoids the configuration that can happen for example at $\mathcal{O}(\alpha_s^3)$, where both heavy quarks are contained within the same jet. Second, the requirement helps to remove backgrounds from mixed-flavour jets. Mixed-flavour events are typically produced through double gluon splitting or through multi-parton interactions, so the directions of the jets are less correlated and hence are less likely to be found with a large angular separation. In comparison true $b\bar{b}$-dijet events are more likely to have a large angular separation. This can be seen by considering the LO production mechanisms of $b\bar{b}$-dijet events as shown in Figure 4.1. The two b quarks are produced from the splitting of a single gluon. In the rest frame of this gluon the quarks must be produced perfectly back-to-back in order to conserve momentum. Switching back to the lab-frame requires a Lorentz boost. To first order the direction of this boost is along the z-axis with little transverse component, since the gluon has little p_{T}. A boost along the z-axis leaves the p_{T} of the created quarks unchanged and hence they are also back-to-back in the x–y plane of the lab frame. Effects such as initial and final state radiation, and inclusion of the underlying event and out-of-cone radiation, will mean that jets are not perfectly back-to-back. However, it remains that they are strongly peaked with $\Delta\phi \sim \pi$, especially when compared to mixed-flavour dijets which have different underlying production mechanisms.

Detector selections

An event is required to have two SV-tagged jets. The reconstruction of these jets and the identification of SVs are discussed in Section 3.5. Next, the events must pass the following trigger requirements:

- `L0MuonEW`. This trigger line requires a muon to be found with a momentum of greater than 6 GeV [64]. This trigger line also contains no global event selection requirements which have the potential to bias the measurement. For example, other L0 trigger lines have a selection requirement on the maximum number of hits in the SPD. This selects events with lower activity. As the theoretical predictions for this measurement are inclusive, such a selection is not desired. The `L0MuonEW` trigger line is around 76% efficient on the samples used in this analysis. The events not passing the trigger line are those where the muon identified by qTAG is not determined to have a high enough momentum to pass the trigger requirement.

- `Hlt1TrackMVA`. This trigger line uses multivariate analysis on the properties of tracks (such as the momentum and impact parameter) to select events that are

likely to contain a HF particle [101]. It is around 99% efficient on the $b\bar{b}$-dijet signal, given the tight selection requirements elsewhere. Therefore, this gives an unbiased route through the trigger.

- `Hlt2JetsDiJetSVSV`. This trigger line selects events which contain two SV-tagged jets. Each jet is required to have a transverse momentum $p_{\mathrm{T}} > 17\,\mathrm{GeV}$, although this is superseded by the fiducial requirement of $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ which is discussed above. The stripping applied to data selects all events that pass this trigger line.

At least one jet in the event is required to be TOS (defined in Section 3.2.6) for both of the L0 and HLT1 trigger lines above. The HLT2 line is applied to the entire event.

Jets in the analysis are also tagged for their charge by locating muons from HF decays. One jet is required to be tagged in this way. This process is called qTAG and is discussed in Section 5.3. The use of muons for charge tagging further motivates the usage of the `L0MuonEW` trigger line.

The simulated events also undergo a process of truth flavour matching, whereby the two jets generated in an event are checked to ensure they are of the correct truth flavour, where the flavour of a jet was defined in Section 3.5. Both jets in the event are required to match the simulated flavour.

In total after all of the selections are applied there are around 1.45 million events in the data sample[1]. There are around 4.25 million events across the $b\bar{b}$ simulated samples produced with tighter generation requirements, which are those used for the main analysis.

[1]The fraction of events with multiple dijet candidates after all of the selections are applied is at the level of $\mathcal{O}(10^{-4})$, so no special treatment of such events is made.

5.3 Charge tagging (qTAG)

Crucial to the main analysis presented in this thesis is the ability to distinguish between b-jets and $\bar{b}$-jets. This can be achieved by identifying the charge of the quark that originates the jet, referred to here as charge tagging. It is desired that the charge tagging method employed will not have a significant contribution to the uncertainty on the final measured value of the charge asymmetry. Current LHCb methods that tag the charge of B mesons are not optimised for the very high momentum tracks[2] that are required to make this measurement across the phase space considered in this analysis [102–104]. For this reason, a specialised charge tagging procedure is developed for use here. This will be referred to as qTAG.

The qTAG method must be appropriate for high-momentum tracks and must not introduce fake asymmetries into the measurement. For this reason, the same approach as the previous analysis at LHCb [95] is used. The basis of this approach is to use only the muons from semi-leptonic b decays to tag the charge of the b quark. An example of such a decay is shown in Figure 5.2a. The presence of a negatively charged muon is used to tag the jet as containing a b quark and a positively charged muon is used to tag the jet as containing a $\bar{b}$ quark. As the muon comes directly from a decay of the b quark it is likely to carry a significant amount of the jet's momentum. Therefore, for the qTAG method, the displaced track[3] with the highest p_{T} in the jet is taken, and if this is identified as a muon following the criteria set out in Section 3.2.5 then the quark that originated the jet is tagged by the muon's charge. Only one jet in each event needs to have its quark tagged, as the charge of the quark originating the other jet can be inferred. By comparing the pseudorapidities of the jets, the event can then be tagged as forward or backward, where forward (backward) events are those where the b-jet ($\bar{b}$-jet) has a greater pseudorapidity. Positively and negatively charged muons are therefore each able to tag both forward and backward events. If the quarks of both jets in the event are tagged by this procedure with the same charge, one must be incorrect if the event is a true $b\bar{b}$ event. As it is not possible to tell which tag is correct and which is incorrect, the event is tagged by one of the jets at random, with the other tag discarded. This leaves the correct tag rate of forward/backward events the same as the correct tag rate for an individual

[2]For example, the RICH detectors are not able to accurately determine particle species at high momentum due to the saturation of the Cherenkov radiation. Information from the RICH detectors is used in many charge tagging methods in analyses at LHCb.

[3]Displaced here means that the track's impact parameter with the PV associated with the dijet system is significant. The specific requirement placed on the muon here is that $\chi^2_{\mathrm{IP}} > 16$.

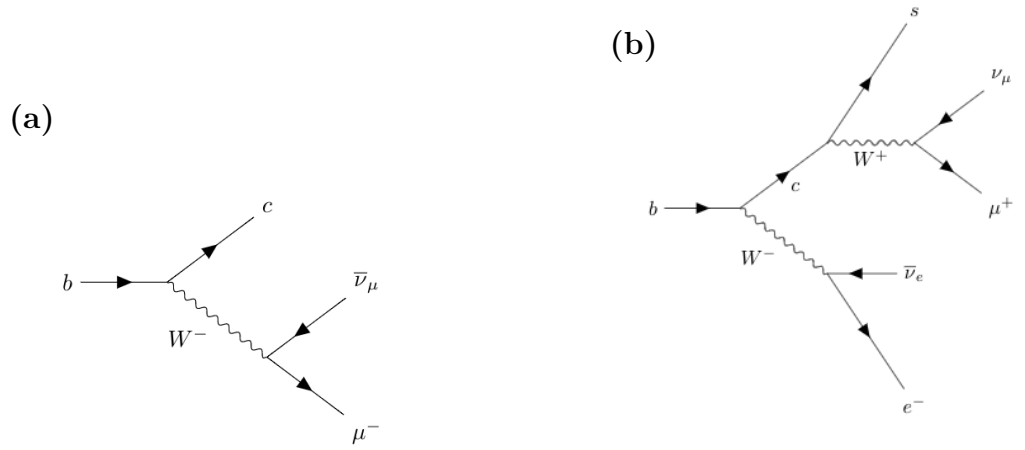

Figure 5.2: Feynman diagrams for: (a) Semileptonic b-decay resulting in the correct tag of a b quark; and (b) a $b \to c \to \mu X$ type decay resulting in the incorrect tag of a b quark.

quark's charge. Such an event is rare as muonic b decays have a branching fraction of around 11% [13].

Mistagging of events is clearly important to understand and evaluate. The two main sources of mistagging the $b\overline{b}$ events are as follows:

- $b \to c \to \mu X$ decays. An example of such a decay is shown in Figure 5.2b. The charge of the muon is now opposite to that of a muon originating directly from a decay of the b quark, thus leading to an incorrect tag if this muon is identified. Muons from this source will, on average, have a lower momentum than a muon coming directly from a b quark and so the qTAG procedure of selecting the highest momentum track should already suppress this source of mistag. It may also be possible to place requirements on the p_{T} of the muon identified by qTAG in an effort to further suppress this source of mistagging.

- B-meson oscillations. A hadron containing a b quark in a jet may oscillate before decaying to a muon, leading to the incorrect assignment of a forward or backward event. It is quarks produced in the hard event that are important for this analysis. It is difficult to efficiently suppress this source of mistagging, whilst satisfying the requirements that the method is effective and simple to model.

These sources of mistagging have implications for the performance of the charge tagging procedure and, hence, the analysis as a whole. It is therefore required to measure the performance of qTAG and ensure that this is well understood. The performance of a

tagging method is typically measured by two important variables. First, its efficiency

$$\varepsilon = \frac{N_{\text{tagged}}}{N_{\text{total}}}, \tag{5.1}$$

where N_{tagged} is the number of tagged events and N_{total} is the total number of events in the $b\bar{b}$ sample before this tagging is applied. And second, the rate at which events are incorrectly tagged

$$\omega = \frac{N_{\text{incorrect}}}{N_{\text{tagged}}}, \tag{5.2}$$

where in this analysis $N_{\text{incorrect}}$ is the number of events where the charges of the quarks originating the jets are misidentified. Mistagging of the quark charge impacts the measured value of the asymmetry and has the effect of diluting the result when compared to the true value. This dilution is derived in Appendix A with the result is given here

$$A^{\text{meas.}} = (1 - 2\omega)A^{\text{true}}, \tag{5.3}$$

showing the true value of the asymmetry is reduced by the factor $D = 1 - 2\omega$, referred to as the dilution factor. Therefore, the measured value of the asymmetry needs to be corrected by this dilution factor to recover the true value. Furthermore, it can be shown that to minimise the statistical uncertainty on this estimate of the true asymmetry, the figure of merit εD^2 must be maximised. This is also shown in Appendix A.

The performance of qTAG will now be roughly estimated using knowledge of the sources of mistag. This estimation will then be compared with mistagging rates measured in simulation to ensure that the method is well understood and performing as expected.

5.3.1 Estimate of performance

Using the branching fractions for b and c decays along with parameters for B^0–$\overline{B}^0$ and B_s^0–$\overline{B}_s^0$ mixing it is possible to estimate the performance of the charge tagging method. Although such calculations are approximate, they will provide a useful benchmark when testing the performance of the charge tagging procedure and will help to ensure that the performance is not behaving dramatically differently than expected.

First, in b hadron decays a muon will be produced around 11% of the time [13]. Studies on simulation during the previous measurement at LHCb found that this muon will be the highest p_{T} particle in roughly 60% of semileptonic b decays [95]. Putting these two numbers together gives the total efficiency of tagging a muon originating from a b decay

in a single jet at around

$$\varepsilon_b \approx 6.5\%. \tag{5.4}$$

For jets not tagged by a muon originating directly from a semileptonic b decay, the mistagging due to muons originating from a $b \to c \to \mu^+ X$ type decay must be considered. To analyse this, the branching fraction $\mathcal{B}(b \to c \to \mu^+ X) \approx 8\%$ [13] is used. Studies on simulation from the previous measurement suggest that the probability of the muons from these c decays being the highest momentum particle in the jet is around 10% [95]. Altogether this means that the efficiency of mislabelling the charge of the quark due to a $b \to c \to \mu$ decay is

$$\begin{aligned} \varepsilon_c &= (1 - \varepsilon_b) \times 0.10 \times 0.08, \\ &\approx 0.75\% \end{aligned} \tag{5.5}$$

Further to this, B^0–$\overline{B}^0$ and B_s^0–$\overline{B}_s^0$ oscillations must be considered. These oscillations lead to some B^0 (B_s^0) mesons converting into $\overline{B}^0$ ($\overline{B}_s^0$) mesons before decay, and vice versa. The B_s^0–$\overline{B}_s^0$ oscillation rate is fast enough that around 50% of mesons will decay as the other flavour from which they are produced [13]. This probability is referred to as the lifetime oscillation probability and is denoted here as χ_s for B_s^0–$\overline{B}_s^0$ oscillations. For the B^0–$\overline{B}^0$ oscillations, the lifetime oscillation probability, $\chi_d = 0.186$ [13]. These values can be combined with the relative B-hadron production fractions as measured at LHCb [105] and shown in Table 5.2. This gives an overall probability of oscillation before decay of

$$\chi_{\text{total}} = f_d \times \chi_d + f_s \times \chi_s \approx 0.16. \tag{5.6}$$

The b-jets with the **R**ight **S**ign of muon identified by qTAG are then the events where there is no b oscillation before a decay directly to a muon, summed with those where there is an oscillation but the decay is to a c and then a muon. The fraction of events is given by

$$\text{RS} = (1 - \chi_{\text{total}}) \times \varepsilon_b + \chi_{\text{total}} \times \varepsilon_c \approx 0.056. \tag{5.7}$$

The term corresponding to oscillation followed by a $b \to c \to \mu$ decay is found to be negligible for this rough estimate. The number of **W**rong **S**ign muons is similarly evaluated

$$\text{WS} = (1 - \chi_{\text{total}}) \times \varepsilon_c + \chi_{\text{total}} \times \varepsilon_b \approx 0.017. \tag{5.8}$$

In total, if all appropriate muons were identified, approximately 7.3% of b-jets would

Table 5.2: The proportion of different B hadrons produced at LHCb [105].

	Hadron	Fraction
f_u	B^-	0.362
f_d	$\overline{B}^0$	0.362
f_s	$\overline{B}^0_s$	0.187
f_b	Λ^0_b	0.088

be tagged by qTAG. As either jet in the event can be tagged by qTAG, the probability of tagging an event is estimated at about 14%. The incorrect tagging rate is given by $\omega = \mathrm{WS}/(\mathrm{WS} + \mathrm{RS}) \approx 23\%$. This gives a dilution factor in the context of the analysis of $1 - 2\omega \approx 54\%$. Next, the performance of the charge tagging procedure will be measured in simulation and compared to the benchmarks set by the rough values estimated here.

5.3.2 Performance in simulation

The performance of the charge tagging procedure is now measured in simulated events to confirm that the method performs as expected.

First, the efficiency is calculated. It is noted here that the value of the efficiency is not used directly in the analysis, apart from in measuring the performance of qTAG here and to provide approximate estimates of values elsewhere. To calculate the efficiency it is not possible to simply calculate the fraction of events that are tagged in simulated samples. This is due to the fact that the simulated samples are produced with the requirement that a muon is found in the final state within the acceptance of the detector. This requirement artificially increases the fraction of jets in the sample that are tagged by qTAG. Instead, qTAG is applied so that at least one jet in every event is tagged. From there it is possible to extract the efficiency for tagging an individual b-jet by evaluating the ratio of events where both jets are tagged to those where only one jet is tagged. First, the efficiency for tagging two jets in the event is given by

$$\epsilon_{\mathrm{double}} = \varepsilon_{\mathrm{b}}^2, \tag{5.9}$$

where ε_b is the efficiency of tagging an individual b-jet. The efficiency for events that have

only one jet tagged in the event is given by

$$\epsilon_{\text{single}} = 2\,\varepsilon_b\,(1-\varepsilon_b). \tag{5.10}$$

This means that after applying qTAG, the fraction of events in the sample that have both jets tagged is given by

$$f_{\text{double}} = \frac{\epsilon_{\text{double}}}{\epsilon_{\text{double}} + \epsilon_{\text{single}}}, \tag{5.11}$$

$$= \frac{\varepsilon_b}{2-\varepsilon_b}. \tag{5.12}$$

Therefore, it is possible to extract the efficiency for tagging an individual jet as

$$\varepsilon_b = \frac{2f_{\text{double}}}{1+f_{\text{double}}}. \tag{5.13}$$

The fraction of events with both jets tagged, f_{double}, is measured in simulated $b\bar{b}$-dijet events. The efficiency of tagging an individual jet is then evaluated as

$$\varepsilon_b = (5.753 \pm 0.019)\%, \tag{5.14}$$

where the uncertainty here is the statistical uncertainty associated with having a finite sample size. This value of the efficiency agrees with expectations from Section 5.3.1 at the level of 20%, reasonable agreement given the "back-of-the-envelope" nature of the estimate. The efficiency of tagging b-jets and $\bar{b}$-jets is also found to be compatible in simulation. When calculating the asymmetry the originating quark of only one jet in the event needs to be tagged, as the charge of the other can be inferred from this. Hence the total efficiency of tagging an event would be given by

$$\epsilon_{b\bar{b}\text{-event}} = 1-(1-\varepsilon_b)^2 \tag{5.15}$$

$$= (11.21 \pm 0.04)\%$$

The rate of mistagging can then be evaluated by comparing the charge assigned by qTAG with the truth level information on the charge of the b quark that originated the jet. In simulated dijet events the mistag rate is measured at

$$\omega = (22.52 \pm 0.06)\%, \tag{5.16}$$

where again the uncertainty here is associated with the finite size of the simulated sample. This rate of mistagging is also in line with expectations of performance, which predicted a value of around 23%. The good agreement between predictions and measured performance demonstrates that qTAG is well understood and is working as expected. The measured dilution factor is evaluated as

$$\begin{aligned} D &\equiv 1 - 2\omega \\ &= 0.5495 \pm 0.0012. \end{aligned} \tag{5.17}$$

The efficiencies are also measured for different jet flavours with qTAG having a $(2.38 \pm 0.06)\%$ efficiency on single c-jets and an efficiency of $(0.50 \pm 0.05)\%$ on single light jets. These values are much lower than those seen for b-jets. This is useful as qTAG then further suppresses contributions from charm- and light-jet backgrounds in the analysis.

One of the main sources of mistagging identified in Section 5.3.1 is that of muons coming from $b \to c \to \mu X$ type decays. It is expected that these muons have a lower momentum when compared to those coming directly from a b decay, and hence that applying some momentum criteria on the muon selected by qTAG might improve the performance. To test this, a requirement is included that the muon identified has a transverse momentum greater than some threshold value. The efficiency, dilution factor and figure of metric εD^2 are compared for a range of threshold values, using the simulated samples for this analysis. The results are shown in Figure 5.3. As the threshold value is increased the dilution factor also increases, showing that the rate of mistagging is indeed lower. However, the efficiency also falls sharply, and the value of εD^2 also decreases, indicating that there is no benefit to this requirement when considering the statistical uncertainty on the final value of the asymmetry. For this reason, no p_{T} threshold is required in the qTAG method beyond the implicit requirement of only considering the highest-p_{T}, displaced track in the jet. This figure also shows the value of the dilution factor reaching some upper limit for increasing threshold requirements on the muon p_{T}. This is due to the B oscillation effects already discussed. Even if all of the $b \to c \to \mu X$ type decays are removed from qTAG, there would still be an incorrect tag rate given by the oscillation probability. This is given in Eq (5.6) as $\chi_{\mathrm{total}} \approx 0.16$. This shows that if only B oscillations were responsible for mistagging, the expected dilution factor would be $D = 1 - 2 \times \chi_{\mathrm{total}} \approx 0.68$. Again this is in line with what is seen in Figure 5.3.

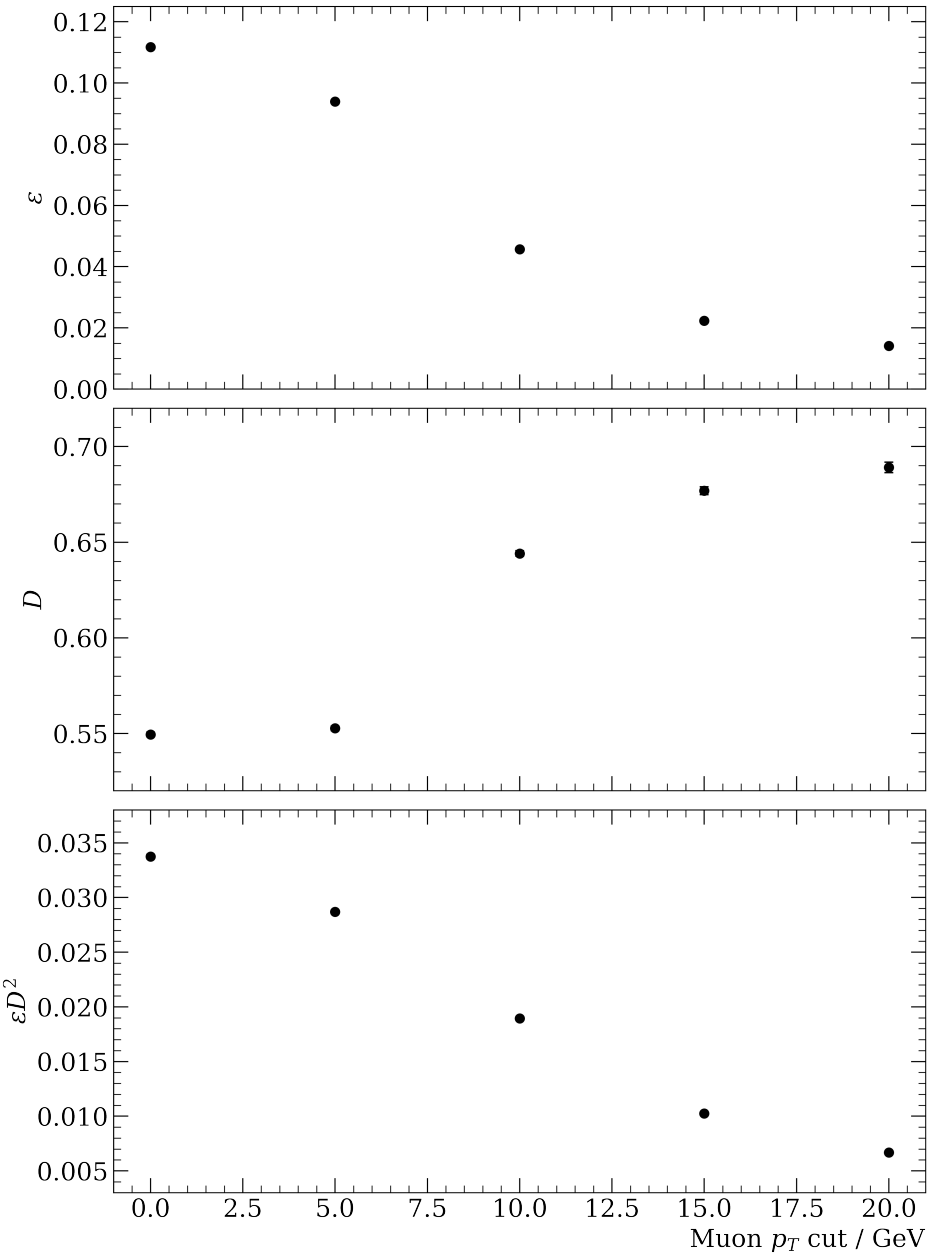

Figure 5.3: Plots showing qTAG performance estimated in simulation when applying p_{T} selections to the muon found by qTAG. The x-axis values denote the minimum transverse momentum that must be carried by the muon identified by qTAG. The aim of this is to remove backgrounds from $b \to c \to \mu X$ type decays. Top: Efficiency, ε, of tagging a $b\overline{b}$-dijet event with qTAG. Either one or both jets in the event can be tagged. Middle: The dilution factor for qTAG. Bottom: The total measure of performance εD^2. It can be seen that although the rate of mistagging is decreased by increasing the p_{T} threshold required by the muon, the loss in overall efficiency would give a higher overall statistical uncertainty in a measurement of the asymmetry, as seen by a lower value of εD^2.

5.4 Jet properties and kinematics

The kinematics of the simulated $b\bar{b}$-dijets[4] selected using the requirements in Section 5.2 are shown in Figure 5.4. Here Jet0 and Jet1 are ordered by their transverse momentum, with Jet0 having the larger of the two. The jet with the higher transverse momentum is often referred to as the "leading jet" and that with the lower momentum as the "subleading jet". This ordering sculpts the distributions of the transverse momenta, as can be seen clearly for Jet0. The ordering of the jets also sculpts the pseudorapidity distributions, as the jet with higher transverse momentum is more likely to have a lower pseudorapidity.

The distribution of the dijet invariant mass, m_{jj}, falls towards higher values as higher energies are needed to produce systems at high mass; this is less likely to be achieved than a low mass system. The shape at the lower end of this distribution is also sculpted by the requirements on the transverse momenta of the jets. This sculpting can be understood with an example of $b\bar{b}$ pair production at parton level. Using the conservation of energy and momentum it can be shown that in the rest frame of the $b\bar{b}$ pair

$$p = \frac{1}{2}\sqrt{m_{bb}^2 - 4m_b^2}, \tag{5.18}$$

where p is the magnitude of the three momentum of one b quark, m_{bb} is the invariant mass of the entire system and $m_b \approx 4\,\text{GeV}$ is the invariant mass of a b quark. Applying the selection requirement that $p_\text{T} > 20\,\text{GeV}$ then gives

$$m_{bb}^2 > 1600 + 4m_b^2, \tag{5.19}$$

$$\implies m_{bb} \gtrsim 40\,\text{GeV}. \tag{5.20}$$

The momentum selection requirements therefore create a minimum value for the dijet invariant mass. As the transverse momentum is invariant under Lorentz boosts along the z-axis, this momentum requirement in the rest frame is equivalent to one placed in the lab frame. Masses slightly above this threshold are made less likely as the probability one jet failed to pass the transverse momentum requirement is increased close to this boundary. This argument neglects momentum in the longitudinal direction, the presence of which also makes the most likely values of m_{bb} larger than 40 GeV.

[4]Whilst the data carries broadly the same features as the simulation, full fits and studies of the detector modelling would be necessary for a direct comparison between the kinematic distributions in data and simulated events. Instead, such studies can be considered in the context of the differential cross-section measurement [106]. In work carried out here, the simulation is calibrated where it is used.

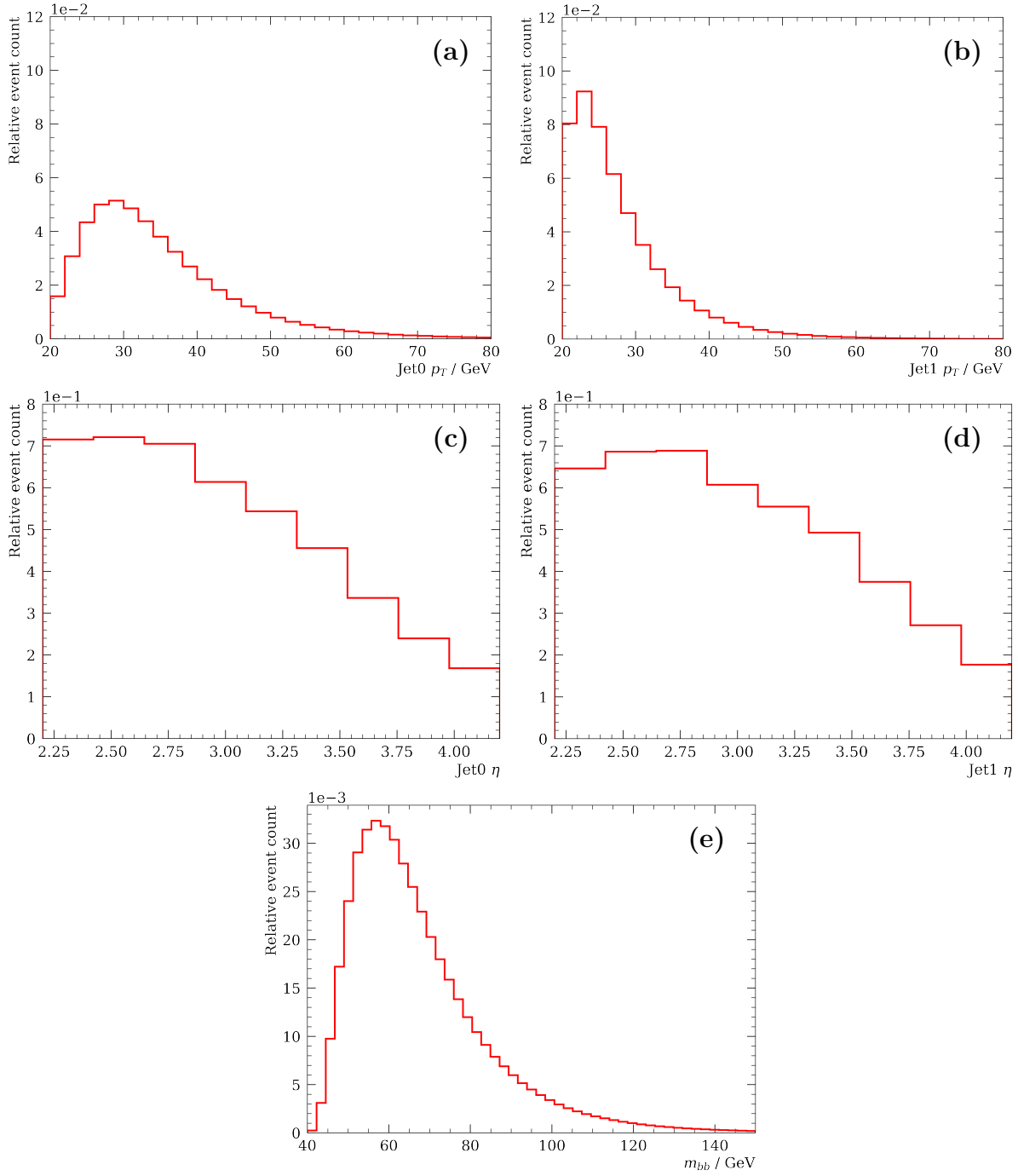

Figure 5.4: Plots showing the kinematics of simulated $b\bar{b}$-dijet events. The jets are ordered by their transverse momentum with Jet0 having the higher p_{T} and Jet1 having the lower.

Chapter 6

Simulated jet calibration

Imperfect modelling of the detector response to jets in simulated events can lead to discrepancies in the jet properties between data and simulation. To quantify and correct for these effects, jets in the simulated events must be calibrated by comparing them to jets in data. This section lays out the work that has gone into carrying out these calibrations for the jet energy and for variables that are used to identify heavy flavour jets. Whilst originally this calibration work was being carried out centrally within the LHCb collaboration by multiple people, much of this effort was lost as a consequence of the COVID-19 pandemic. It is therefore necessary to derive and apply corrections to the simulation specifically for this analysis. Other analyses, such as the Z+c production analysis [90], also found it necessary to carry out their own calibrations. The calibrations performed here are at the level of precision required by the analysis, with a conservative approach taken to the systematic uncertainties. A less conservative approach should be considered in future iterations of this measurement, particularly if the full Run II dataset is analysed.

Calibrations are performed for two properties of the simulated jets. First, the jet energy, which is comprised of a calibration for the resolution and the scale of the energy reconstructed in simulated jets. As discussed in Section 3.5, reconstructed jets in both data and simulation have their energy scaled using a calculation derived in simulated jets that is standard for LHCb [82]. This ensures that the reconstructed energy and momentum of a jet is the best estimate of the true jet properties, accounting for detector effects. To correct for potential mismodelling of the detector response, the resolution and scale of the jet energy are explicitly calibrated in this analysis. This calibration mostly impacts the analysis when unfolding the dijet invariant mass in Chapter 8. The effects are, however,

shown in that chapter to be small. Second, calibrations are performed on variables used in the identification of HF jets. It is imperative for this analysis to distinguish between beauty and charm-jets. The properties of SVs found within jets are typically used for this purpose at LHCb. Due to detector degradation that is not accurately modelled in the simulation, the properties of these SVs show discrepancies between events in data and simulation. This mismodelling is not exclusive to this analysis [107, 108], but the calibration is performed explicitly for the samples used here.

This section will address each of these calibrations in turn: the jet energy resolution, the jet energy scale and finally, the SV properties.

6.1 Jet energy resolution

The first calibration for the jet energy carried out here is the resolution of the measured jet energy. The goal is to correct for differences between the resolution in data and its modelling in simulated events.

To investigate this mismodelling, the balance of the magnitude of the p_{T} between the two jets in the event is considered, this is referred to here as the p_{T} magnitude balance. The dijet events and their selection are discussed in Chapter 5. Fits are performed in the main analysis on this same sample and it is found that the vast majority ($> 95\%$) of these events are $b\bar{b}$-dijet events as shown in Section 7.2. For this reason, a direct comparison between simulated events of this flavour composition and the data events is justified[1]. The p_{T} magnitude balance is defined as

$$B_{\mathrm{dijet}} = \frac{|\boldsymbol{p}_{\mathbf{T}}^{j_m}| - |\boldsymbol{p}_{\mathbf{T}}^{j_n}|}{|\boldsymbol{p}_{\mathbf{T}}^{j_m}| + |\boldsymbol{p}_{\mathbf{T}}^{j_n}|}, \tag{6.1}$$

where $j_{m,n}$ are the two jets in the event with an arbitrary ordering. The balance takes a form similar to previous studies at LHCb [83]. At leading order, the transverse momenta of the partons that interacted to create the pair of jets should be zero. Then by the conservation of momentum, the jet pair should have a transverse momentum that also sums to zero, this is to say that the jet pair is produced back-to-back in the azimuthal plane. The distribution of the p_{T} magnitude balance should therefore peak at zero while the width of this distribution will depend upon the resolution of the transverse momentum[2].

[1] In addition, conservative uncertainties are assigned to the corrections derived in these sections.

[2] This also only considers leading order production methods of the dijet system. The inclusion of initial and final state radiation, and other higher-order diagrams, can lead to the production of additional jets in

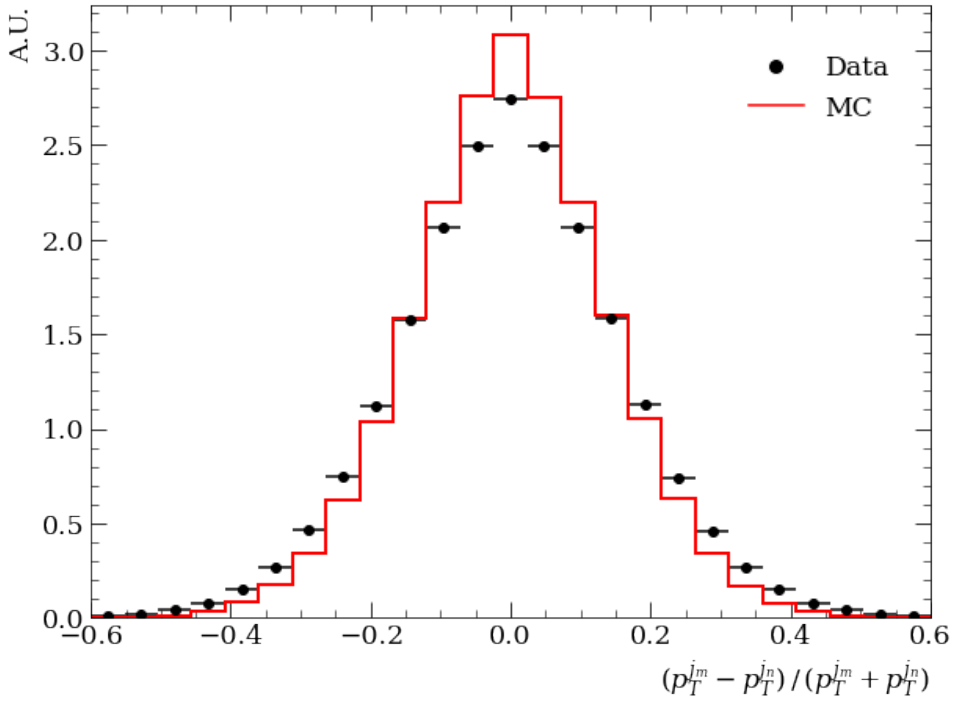

Figure 6.1: Comparison between data (black) and simulation (red) for the p_{T} magnitude balance in dijet events. Histograms have been scaled such that the area contained by them is equal to unity.

Comparing the distributions of B_{dijet} in data and simulated events, therefore, indicates how well the resolution is modelled in simulation.

6.1.1 Correcting the dijet simulation

The p_{T} magnitude balance for dijet events in simulation and data is shown in Figure 6.1. The distribution in simulation is visibly narrower than that of data, indicating the detector performance is being overestimated in simulation.

To correct for this, a Gaussian smearing is applied to the p_{T} of both jets in the event with

$$p_{\mathrm{T}}^{\mathrm{smeared}} = N(p_{\mathrm{T}},\ \alpha\, p_{\mathrm{T}}), \tag{6.2}$$

where $N(\mu, \sigma)$ draws from a normal distribution with mean μ and standard deviation σ, and α is a parameter that dictates the level of smearing, referred to as the resolution smearing factor. The form of this correction is the same as used in previous LHCb

the event. The momentum of the two selected jets in the system would then not be expected to balance.

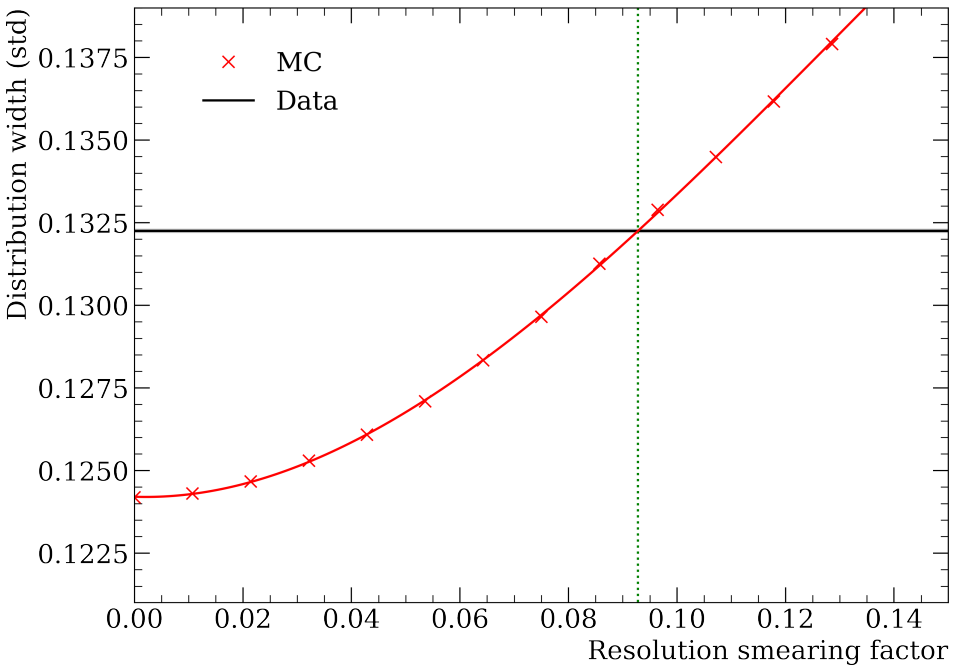

Figure 6.2: Comparison of the width of the p_{T} magnitude balance distribution for dijet events in data (black) and simulation (red) for different smearing factor applied to the jet p_{T} in simulation. The points generated from simulation are fit using a third-order polynomial. The green dashed line shows where the width of the distribution in simulation is equal to that of data. Statistical uncertainties in both distributions are negligible. A large systematic uncertainty is applied to this smearing factor which is discussed in the main body of the text.

analyses, such as in Ref. [83]. To determine the value of α, different smearing factors are applied to the jets in simulation and the RMS of the B_{dijet} distribution is compared to that in data. During this process, a requirement is also placed on the events in both data and simulation that $|B_{\mathrm{dijet}}| < 0.3$. This removes events in the tails of the distribution where there may be multiple jets in the event and hence the two selected jets would not be expected to balance. This requirement follows similar applied in previous analyses at LHCb [90]. The value of α is then chosen such that the RMS values of the two distributions agree. The RMS of the two distributions are shown in Figure 6.2 where the results from simulation have been fit using a third-order polynomial with the mean squared error of the fit minimised. Due to the large sample sizes used here, the statistical uncertainties are negligible for this result. The resolution smearing factor is evaluated as $\alpha = 0.0927$.

The p_{T} magnitude balance in dijet events after applying this smearing is shown in Figure 6.3. A systematic uncertainty is also shown in this plot which is calculated as the

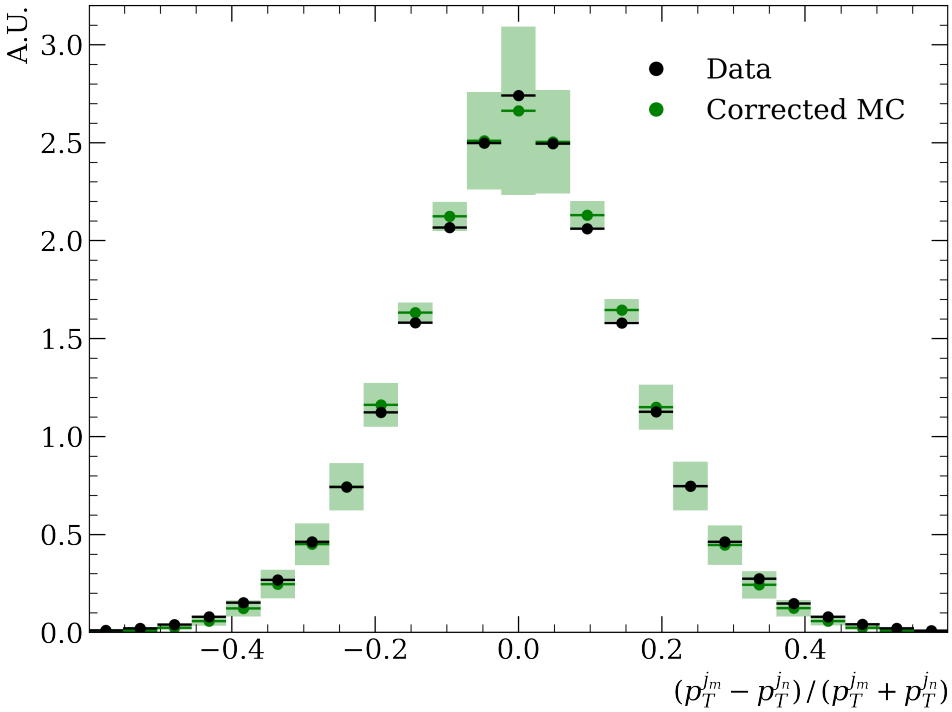

Figure 6.3: Comparison between data (black) and corrected simulation (green) for the p_{T} magnitude balance in dijet events. Corrected simulation has had the p_{T} of both jets smeared as discussed in the main body of the text. Green bands indicate the effects of systematic uncertainty assigned to this smearing factor. Histograms have been scaled such that the area contained by them is equal to unity. The lack of exact agreement between data and simulation may motivate the study of alternative correction parameterisations in future analyses.

difference between distributions before and after applying the smearing. This systematic uncertainty will also be propagated through the entire analysis of the charge asymmetry measurement, which will be discussed in Chapter 9. The systematic uncertainty is applied symmetrically to the nominal value of the smearing used. This is a very conservative approach, however the uncertainty on the final value of the asymmetry remains subdominant[3]. In future analyses, it may be possible to explore different forms for the correction and derive a less conservative systematic uncertainty from this. Overall, a much better agreement between data and simulation is seen once the corrections have been carried out. The value of this smearing factor is similar in size to that seen in the Z+c production analysis [90], with the value obtained here slightly smaller.

[3]This is due to an appropriate choice of width for the bins of dijet invariant mass in which the measurement is made.

The value of the smearing factor used is verified in a beauty-jet enriched region of the data, to further suppress the low level of backgrounds that are present in the data sample. To do this a requirement is placed on the corrected mass of each of the two jets in the event. The corrected mass is an important quantity in distinguishing between jet flavours and is defined as

$$M_{\text{cor}} = \sqrt{M^2 + p^2 \sin^2\theta} + p \sin\theta, \tag{6.3}$$

where M and p are the invariant mass and momentum of particles that form the SV in the jet and θ is the angle between the momentum and direction of flight of the SV [91]. This variable describes the minimum mass the particle that generated the SV can have that is consistent with the direction of flight. The distribution of the corrected mass is shown for simulated beauty and charm-jets in Figure 6.4. To reduce charm backgrounds a requirement is placed on each jet in the event that the corrected mass, $M_{\text{cor}} > 2.2\,\text{GeV}$. This requirement is measured in simulated events to be around 75% efficient for beauty-jets and 13% efficient for charm-jets. The requirement is placed on both the simulated and data samples. The value of the smearing factor found using these beauty-enriched samples is found to be $\alpha = 0.0898$. This agrees extremely well with the nominal value.

6.2 Jet energy scale

The next calibration for the jet energy is the scale. As discussed in Section 3.5 the jet energy scale can be impacted by effects such as the track finding inefficiencies or incomplete neutral recovery. Mismodelling of these effects in simulation leads to a difference in the jet energy and momentum distributions compared to data.

The dijet p_{T} magnitude balance is not sensitive to the jet energy scale as any global scaling of the jet p_{T} will cancel in this measure. For this reason, a different event type is needed. Z+jet events are therefore used where the Z subsequently decays to a pair of muons. These events are used because the Z transverse momentum is precisely measured when decaying to two muons, with a resolution of $\mathcal{O}(1\%)$, and this is well modelled in simulation [109]. Therefore any mismodelling effects seen will be dominated by the modelling of the jet reconstruction. Similarly to the dijet case, the Z and jet are expected to be produced back-to-back in the azimuthal plane and hence they should have equal transverse momenta. In events where multiple jets are produced in association with the Z, the balance between the leading jet and the Z need not be equal to satisfy momentum

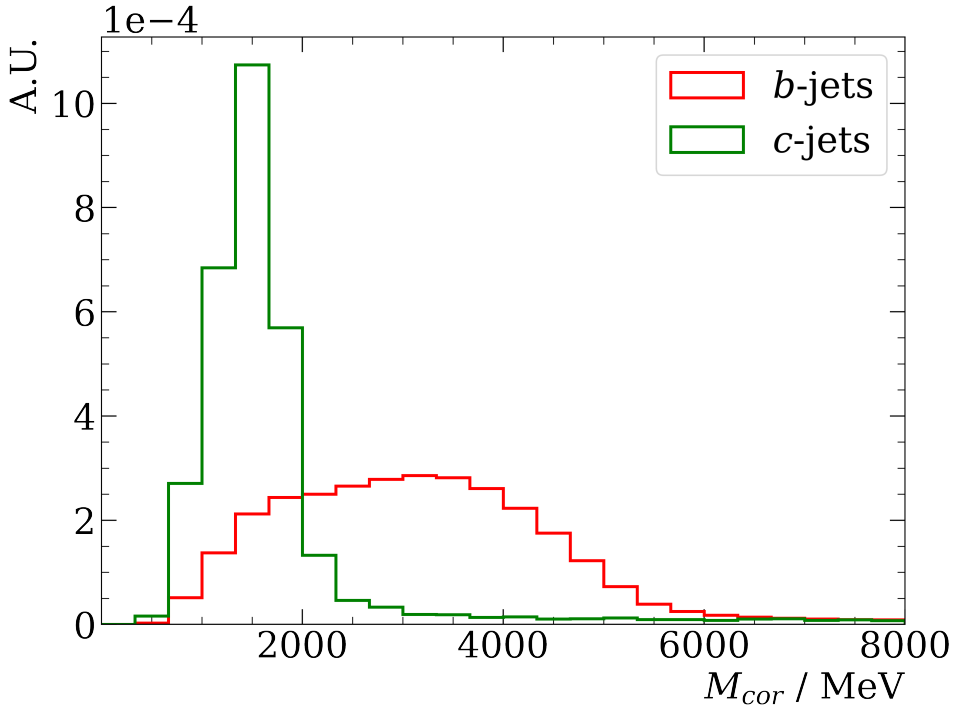

Figure 6.4: Corrected mass distribution for b and c jets as measured in simulation.

conservation and hence in these events, the balance would differ from unity. Selection criteria are therefore applied to help remove such events, as discussed below. This p_{T} magnitude balance in Z+jet events is defined as

$$B_Z = \frac{|\boldsymbol{p}_{\mathbf{T}}^{\mathrm{jet}}|}{|\boldsymbol{p}_{\mathbf{T}}^{Z}|}. \tag{6.4}$$

Where this measure of the p_{T} magnitude balance takes the same form as used in previous LHCb analyses [83]. This measure is sensitive to mismodelling of the detector response with regard to the jet energy scale. Comparison of B_Z in simulation and data allows this to be investigated. The energy scale derived here will be based on light-jets, however, the particle flow algorithm shows very similar performance for light and HF jets [110]. Furthermore, previous studies at LHCb have confirmed that the energy scale is consistent between HF and light-jets to within $\mathcal{O}(2\%)$ [89].

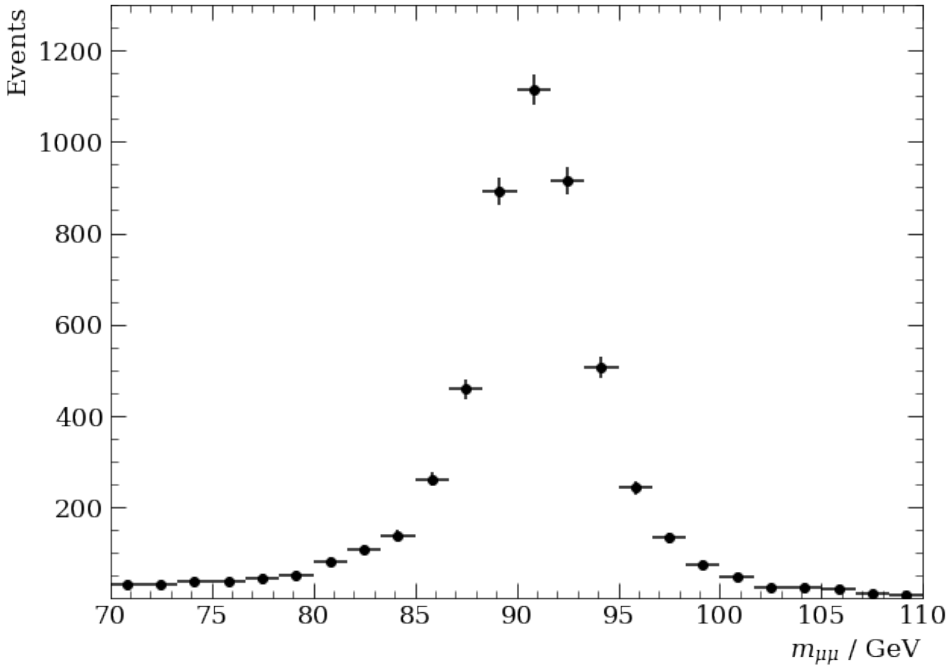

Figure 6.5: Distribution of the invariant mass of the two muons in Z+jet data events where the Z subsequently decays to two muons. A clear peak is seen around the Z mass.

6.2.1 Event samples

This calibration is performed using the 2016 dataset collected with the LHCb detector for proton-proton collisions at a centre-of-mass energy of 13 TeV. The Z+jet process is also simulated in Pythia 8 [38], where the LO production of the Z produced in association with a single jet is specified. The selections used for the Z+jet events in this section of work are similar to a previous LHCb analysis which can be found in Ref. [81]. Firstly, both muons from the $Z \to \mu^+\mu^-$ decay are required to have transverse momenta $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and pseudorapidities in the range $2.0 < \eta < 4.5$. The invariant mass of the dimuon pair must also be in the range $60 < m_{\mu\mu} < 120\,\mathrm{GeV}$. The event is further required to have a reconstructed jet with transverse momentum $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and pseudorapidity in the range $2.2 < \eta < 4.2$. This jet must then be separated from each of the reconstructed muons by a radius $\Delta R > 0.5$. There is also a 'back-to-back' selection placed on the reconstructed Z and the jet, by requiring that their azimuthal angular separation, $\Delta\phi > 2.9\,\mathrm{rad}$. As well as this, any additional jet in the event is required to have a transverse momentum less than 25% of that of the highest p_{T} jet in the event. These final two criteria help to remove Z+multi-jet events, where B_Z would

not be expected to be unity. A total of 3,843 events are selected in data and 6,497 in simulation. The purity of the selected Z+jet sample is greater than 99% [81,90,111]. The reconstructed dimuon mass can be seen in Figure 6.5.

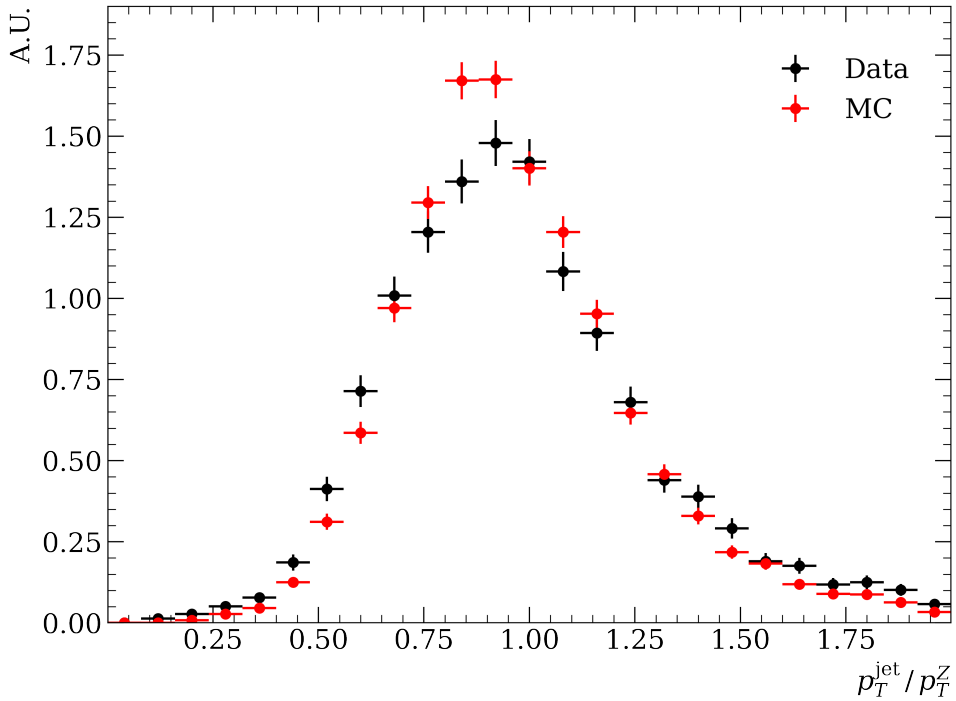

Figure 6.6: Comparison between data (black) and simulation (red) for the p_{T} magnitude balance in Z+jet events before scaling or smearing.

6.2.2 Correcting the Z+jet simulation

The p_{T} magnitude balance for simulation and data in Z+jet events is shown in Figure 6.6. To determine the level of scaling required, the means of the p_{T} magnitude balance distributions are compared. It is important here to remove events where the Z boson is produced in association with multiple jets. This is due to the fact that during the production of the simulated samples, the LO production of a Z boson with a single jet is specified. Multi-jet events are therefore not expected to be well modelled. The selection discussed above removes these events, ensuring that the calibration process tests the modelling of the detector response, as opposed to the modelling of the hard process in the simulation. In addition to the requirements placed on the presence of a second jet above, a selection of $0.5 < B_Z < 1.3$ is applied to events following similar criteria set out in previous LHCb analyses [90]. The jet energy resolution must also be taken into

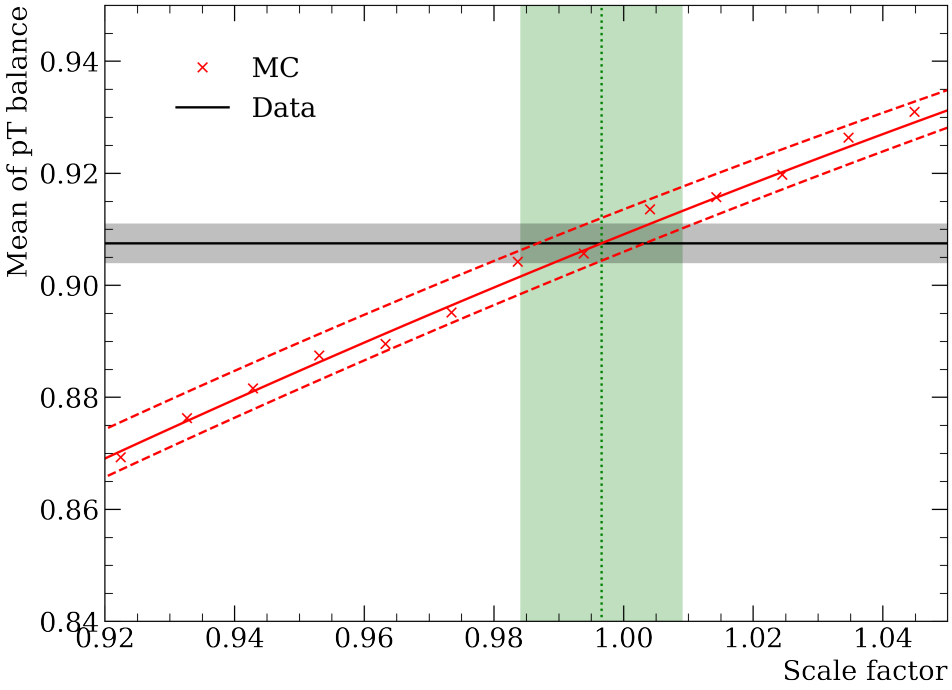

Figure 6.7: Comparison of the mean in the p_{T} magnitude balance distribution of Z + jet events in data (black) and simulation (red) for different scale factors applied to the jet p_{T} in simulation. The data points generated from simulation are fit using a third-order polynomial. The green dashed line shows where the mean value in simulation is equal to that of data. A statistical uncertainty band of one standard deviation is shown for data in grey. There is an uncertainty on the scaling applied to all jets to correct for detector effects, as discussed in Section 3.5.. This uncertainty is shown as the dashed red lines for the mean values in simulation. The statistical uncertainty band from the data and the uncertainty on the jet energy scale can be converted to statistical and systematic uncertainties, respectively, on the jet energy scale derived here as discussed in the main body of the text. The scale factor is evaluated at $s = 0.997 \pm 0.012$.

account, and so the p_{T} of jets in the simulated sample are scaled and smeared as

$$p_{\mathrm{T}}^{\mathrm{smeared}} = N(s\,p_{\mathrm{T}},\ \alpha\,(s\,p_{\mathrm{T}})), \tag{6.5}$$

where $\alpha = 0.0927$ is the smearing factor determined in $b\bar{b}$-dijets and the value of s is chosen such that the distributions of B_Z have the same mean in simulation and data. The form of this smearing ensures that jets are smeared according to their rescaled momentum[4]. The means of the distributions are shown for a range of scale factors in

[4]The p_{T} magnitude balance in dijet events is unchanged by global scaling of the jet p_{T}, so the value of the smearing factor was insensitive to this scaling.

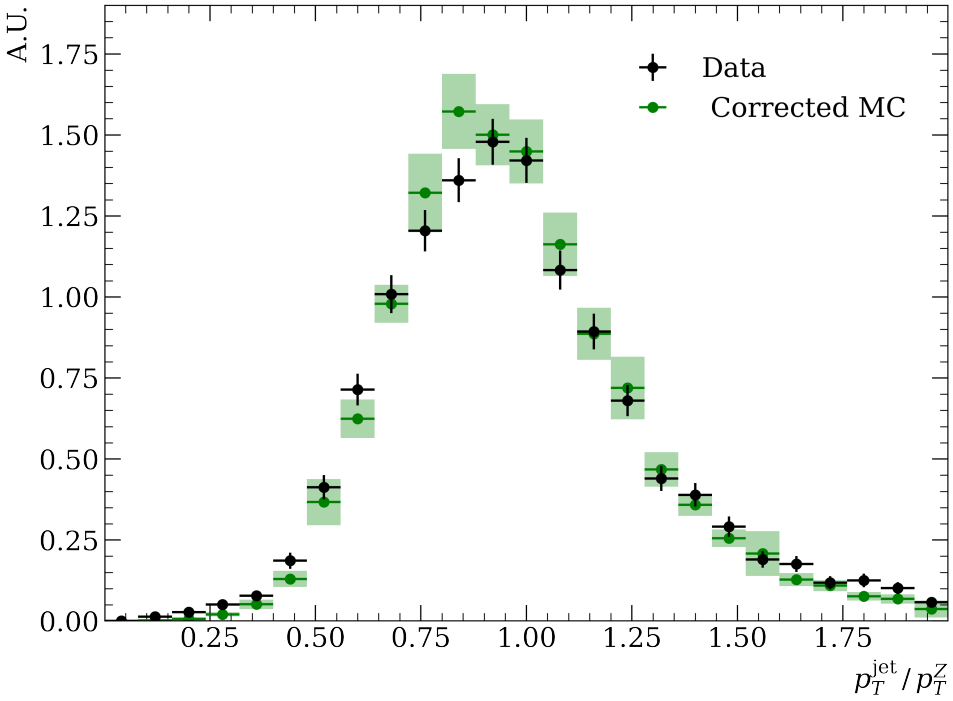

Figure 6.8: Comparison between data (black) and simulation (green) for the p_{T} balance in Z+jet events after scaling and smearing the jet energy. A scale factor of unity is used with a smearing factor of 0.0928, as discussed in the main body of the text. Uncertainty bands shown for simulated events are a combination of the statistical and systematic uncertainties owing to the uncertainty in the scale factor applied.

Figure 6.7. The results from simulation are fit using a third-order polynomial, with the mean squared error of the fit minimised. From these results, it can be found that a scale factor, $s = 0.997 \pm 0.007$, needs to be applied to the p_{T} of the simulated jets to match the p_{T} magnitude balance mean seen in data. The uncertainty here is a statistical uncertainty associated with the determination of the mean in data. This is evaluated by finding the maximum and minimum scale factors that can be applied whilst still agreeing with the mean value in data to within one standard deviation. There is also an uncertainty on the jet energy scale common to all analyses at LHCb, directly evaluated when determining the jet energy correction in simulation, as discussed in Section 3.5. This also gives an uncertainty on the mean of the p_{T} distribution which is shown for simulation by the dashed red lines in Figure 6.7. This uncertainty is then converted to a systematic uncertainty on the scale factor by considering what scale factor would be required to give agreement between data and simulation if the simulated jets were scaled up and down by one standard deviation. The systematic and statistical uncertainties are

Table 6.1: This table shows the scale factors needed to rescale the transverse momentum of the jet in Z+jet events for different regions of the phase space. Uncertainties shown are statistical from the determination of the mean transverse momentum balance in data. Results are consistent with the nominal value of 0.997 across the phase space studied. nPV is the number of primary vertices associated within the event.

Phase space requirement	Scale factor
$\eta < 3.3$	0.993 ± 0.009
$\eta > 3.3$	1.012 ± 0.016
nPV $= 1$	1.005 ± 0.012
nPV $= 2$	0.993 ± 0.013
nPV ≥ 3	0.981 ± 0.032
$20 < p_{\mathrm{T}}(\mathrm{jet}) < 30\,\mathrm{GeV}$	1.000 ± 0.018
$30 < p_{\mathrm{T}}(\mathrm{jet}) < 40\,\mathrm{GeV}$	0.997 ± 0.014
$p_{\mathrm{T}}(\mathrm{jet}) > 40\,\mathrm{GeV}$	0.988 ± 0.011

combined in quadrature. To give a symmetric error, the largest out of the positive and negative standard deviations is taken in both directions. The scale factor is evaluated at $s = 0.997 \pm 0.012$. The distribution of B_Z is shown in data and simulated jets after smearing and scaling in Figure 6.8. The systematic uncertainty associated with the scale is shown as a green band. There is a significant improvement in the agreement between the data and simulated events after scaling and smearing.

The scale factor is also measured in different regions of the phase space used in the analysis. There is no significant variation of the scale factor with the jet p_{T}, jet pseudorapidity or the number of PVs in the event. The results for these studies are shown in Table 6.1.

As this scale factor is consistent with unity, no additional scaling will be applied to jets for the main analysis in this thesis. The uncertainty in this value will instead be considered as a systematic uncertainty in Chapter 9. This is in addition to systematic uncertainties associated with a calculation of the jet energy scaling applied to all reconstructed jets at LHCb which will also be included in Chapter 9.

The smearing is applied to all simulated jets in the analysis. As the resolution on the pseudorapidity and azimuthal angle of these jets is good and well modelled in simulation

[83], this is applied in such a way as to keep these unchanged. The p_{T} is scaled and smeared as in Eq. (6.5) and the longitudinal momentum of the jet is then scaled in such a way as to keep the direction of the jet the same. The jet energy is also scaled in this way as it is expected that the resolution on this should be similar to the momentum.

6.3 Heavy flavour identification

This subsection reports work conducted by Dr William Barter carried out in close association with the author.

The identification of heavy flavour jets is an important part of jet analyses carried out with the LHCb detector. This is done by the identification of SVs as discussed in Section 3.5. The properties of a SV in a jet are then used to identify whether the jet originated from the hadronisation of beauty, charm or light quarks. This identification has been previously carried out using boosted decision trees (BDTs) that were trained on simulated samples of b-, c- and light-jets with the full list of inputs detailed in Ref. [91]. There are two BDTs that were utilised: one that distinguishes between heavy (b- or c-jets) and light-jets, and one that distinguishes between b- and c-jets. The scores outputted by these BDTs were then used to distinguish the flavour of the jets and used to perform fits to ascertain the flavour composition of data samples. An example of this is shown in Figure 6.9 where the flavour composition of jets in B+jet events is obtained [91]. This figure also shows imperfections in the modelling of the BDT score distributions. This is due to the difficulties in modelling the various properties of the SV which leads to slight discrepancies in the outputs of the BDT. Over the lifetime of the detector, the mismodelling of the detection and properties of SVs has unfortunately worsened. One of the major reasons for this is that the VELO has received significant radiation damage, leading to a decrease in its performance [112]. This decrease in performance is not modelled well in the simulation which has led to larger discrepancies between data and simulation [107,108]. For this analysis, it is important to accurately obtain the purity of the selected dijet sample and for this reason, it is necessary to calibrate the simulation such that it more accurately models the data.

Attempting to correct the full set of inputs used by the BDTs would be a significant undertaking. Instead, a single variable is chosen based on two requirements. First, it must have good distinguishing power between charm-, beauty- and light-jets. Second, a simple method for calibrating this variable should be obtainable. For these reasons,

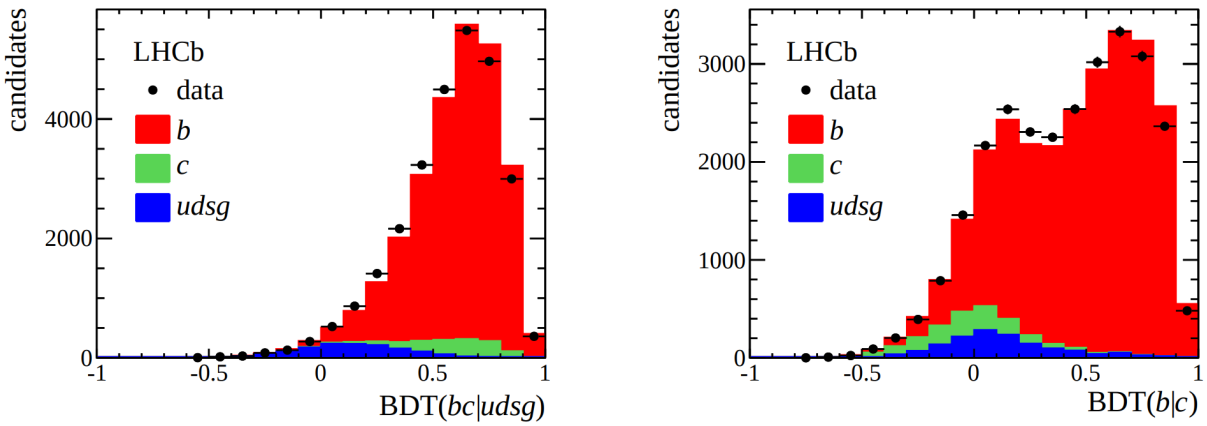

Figure 6.9: Fits to identify jet flavour in B+jet events using standard LHCb flavour tagging BDTs. From Ref. [91].

the $\log(\chi^2_{\text{FD}})$ of the SV is chosen. FD here stands for the flight distance, which is the distance separating the SV from the PV associated with the dijet system. The χ^2_{FD} gives the significance of the flight distance and its compatibility with zero. This is an important variable in identifying heavy flavour jets as heavy flavour hadrons are long-lived and travel some distance in the detector before decaying. Distributions of this variable for beauty-, charm- and light-jets as used in the fits for the main analysis can be found in Figure 7.2a. This variable, and its power to distinguish between jet flavours, relies heavily on the excellent VELO performance and the ability to find and separate secondary decays from the primary interactions. Inputs for this variable include the location of the vertices and associated uncertainties, the latter of which are underestimated in simulation due to mismodelling of the VELO degradation [107, 108]. Work carried out and detailed in Ref. [108] indicates that it is possible to calibrate this variable and obtain good agreement between data and simulation using an appropriate control channel. For these reasons, this is the variable that is chosen in this analysis for distinguishing between jet flavours. The work into calibrating this variable is detailed in the remainder of this section.

Calibrations of SV properties have been carried out before using dijet data at LHCb, notably in Ref. [90]. The method involved a "tag-and-probe" approach, where a stringent flavour (b or c) selection was applied to the tag jet leaving an enhanced flavour sample in the probe. For the analysis cited this approach was useful as a correction could be derived in dijet samples and then applied to Z+c-jet events. Unfortunately, this work cannot easily be applied here as the corrections are derived in the samples on which the measurement of the charge asymmetry is carried out. This would lead to strong correlations between distributions in the simulated samples and the data that would be

difficult to control. Instead, events containing $B^+ \to J/\psi K^+$ and $B^0 \to J/\psi K^*$ decays (and their antimatter conjugate processes) are used for the calibration process and the $\log(\chi^2_{\rm FD})$ of the reconstructed B meson is compared in simulation and data[5]. These events form a good proxy for the dijet events as they contain a SV with a clear, high-purity signal providing easy selection. Importantly, they also have similar values of $\log(\chi^2_{\rm FD})$ to jets in $b\bar{b}$-dijet systems.

6.3.1 Event samples

To derive a correction for the $\chi^2_{\rm FD}$ variable $B^+ \to J/\psi K^+$ decays are used where the J/ψ subsequently decays to two muons. The events used were studied in Ref. [113] and the data files were kindly provided by the authors. The full LHCb 2016 dataset is used to derive this calibration. The B^+ mass reconstructed from the $K^+\mu\mu$ decay products can be seen in Figure 6.10. For deriving the correction, the selection on the B^+ mass is tightened by requiring that its mass is in the region $5230 < m_B < 5320\,\text{MeV}$.

The correction derived is tested using $B^0 \to J/\psi K^*$ decays where the J/ψ subsequently decays to two muons and the excited kaon undergoes a $K^* \to K^+\pi^-$ decay. Again, the data files used here were kindly provided by authors of Ref. [114].

6.3.2 Deriving a $\chi^2_{\rm FD}$ correction

In previous studies of the VELO degradation at LHCb and associated underestimation of uncertainties in the simulation, it has been seen that a scaling of the $\chi^2_{\rm FD}$ variable delivers good agreement between data and simulation [107, 108] and so this method is considered here. The scaling takes the form

$$\chi^2_{\rm FD} \to \lambda\, \chi^2_{\rm FD}, \tag{6.6}$$

for some constant factor λ. The logarithm of the $\chi^2_{\rm FD}$ is often used which transforms as

$$\log(\chi^2_{\rm FD}) \to \log(\chi^2_{\rm FD}) + \log(\lambda). \tag{6.7}$$

The value of λ is chosen such that the mean value of $\log(\chi^2_{\rm FD})$ agrees between data and simulation. This is measured in bins of $p_{\rm T}$ of the B^+ candidate. This is fit with a function of the type $a_0 + a_1\, \exp(a_2\, p_{\rm T})$ so that the correction can be applied on an event-by-event

[5]Whilst these events may also be in the dijet sample, they will make up a very small proportion of the overall dataset. This due to the low branching fractions for these processes.

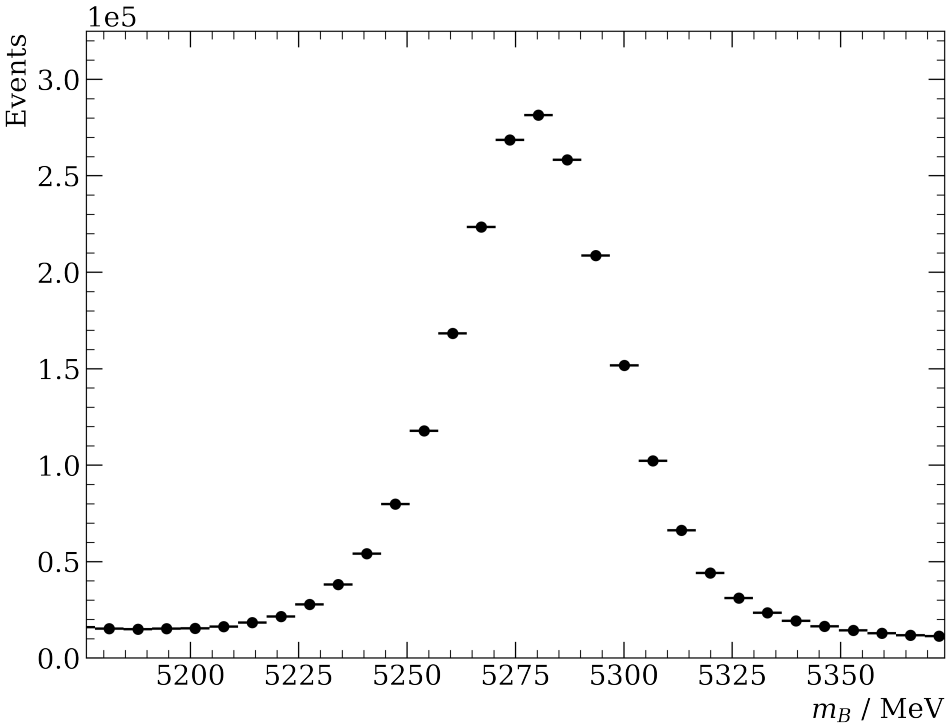

Figure 6.10: The distribution of the B^+ mass in $B^+ \to J/\psi K^+$ decays. Work for this plot was carried out by Dr William Barter, in discussions with the author.

basis. The choice of shape for this fit is an empirical choice after observing the shape of the correction. Parameterising the correction using the transverse momentum also means that this correction is directly transferable from the B-meson decay to the SV. The SVs of the jets used in the main analysis typically have a transverse momentum in the range $5 < p_{\mathrm{T}} < 20\,\mathrm{GeV}$. The parameters for the fit result are resampled using their covariance matrix to find an uncertainty on the fit which is shown in Figure 6.11. The fit yields the correction factor

$$\log(\lambda) = -0.146 + 0.215 \times \exp(-0.409\, p_{\mathrm{T}}/\,\mathrm{GeV}). \tag{6.8}$$

For all but the lowest p_{T} mesons, this correction is negative, showing that the value of the $\log(\chi^2_{\mathrm{FD}})$ is overestimated in simulation. This implies more significant separation of the SVs from the PV than is found in data. This is as expected, as the uncertainties in simulation are underestimated leading to an overestimate of the significance of separation. The correction factor is then applied to the $\log(\chi^2_{\mathrm{FD}})$ distribution in the $B^+ \to J/\psi K^+$ events and the distributions in simulation and data are compared in Figure 6.12. There

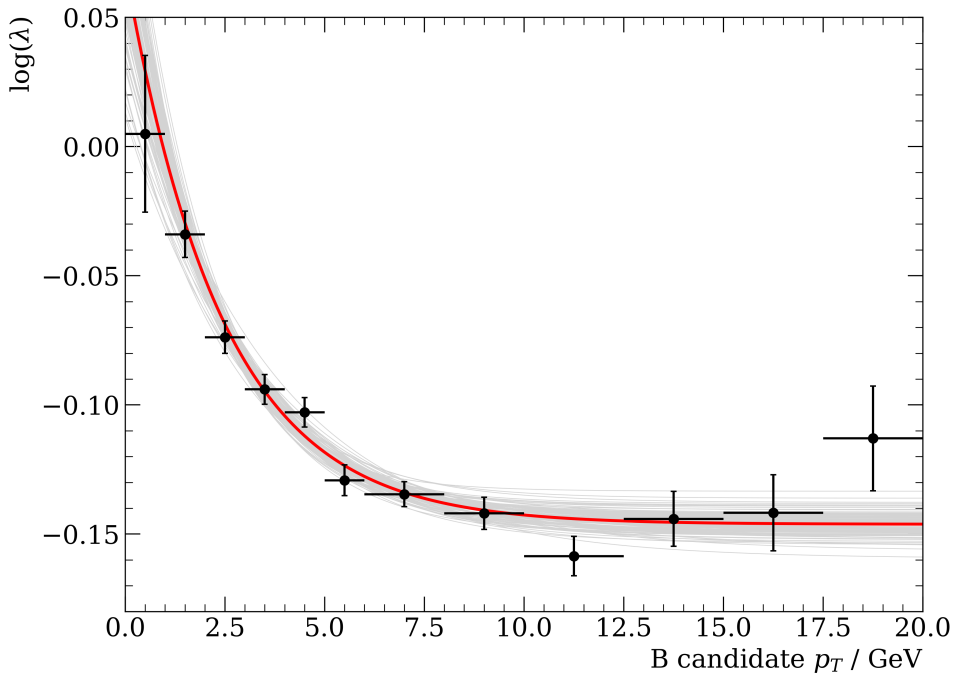

Figure 6.11: The correction factor required to give agreement between simulation and data is shown as black data points in bins of the p_{T} of the B candidate. This is fit with a function of the type $y = a_0 + a_1 \exp(a_2\, p_{\mathrm{T}})$, this fit is shown as the red line. Using the covariance matrix for this fit, the parameters a_i are varied within their uncertainty bounds and the resultant fits are plotted as grey lines. This indicates the level of uncertainty on the fit. Work for this plot was carried out by Dr William Barter, in discussions with the author.

is a clear improvement in the agreement between data and simulation after applying the correction.

6.3.3 Cross-checks using $B^0 \to J/\psi K^*$ decays

To provide a cross-check for the correction, $B^0 \to J/\psi K^*$ decays are used. This constitutes a four-body decay as opposed to the three-body decay used to derive the correction. This helps confirm that the correction derived is robust to changes in the momentum profile of the particles produced in the B decay. The correction derived in $B^+ \to J/\psi K^+$ is first applied directly to simulated $B^0 \to J/\psi K^*$ decays. The results of this are shown in Figure 6.13. There is again a good improvement seen between the data and simulation after the correction is applied, showing that the correction is robust. This is important as it shows that the correction derived in $B^+ \to J/\psi K^+$ decays is more generally applicable.

An alternative correction can also be derived from the $B^0 \to J/\psi K^*$ decays using the same method as for the $B^+ \to J/\psi K^+$ decays. If this is done the fitted correction yields a correction factor

$$\log(\lambda) = -0.082 + 0.399 \times \exp(-1.15\, p_{\mathrm{T}}/\,\mathrm{GeV}). \tag{6.9}$$

This shows a difference from that derived previously in Eq. (6.8). Although the correction is in the same direction, it is smaller here indicating the simulation better models these decays. One reason for this may be the 4-body nature of this decay as opposed to the 3-body decay used previously. This difference perhaps indicates that further work needs to be carried out for correcting the simulation to a level of high precision. However, this is beyond the scope of this analysis and instead, both corrections can be implied independently with a systematic uncertainty being assigned based on the difference in the final results. For typical jets used in the analysis, this amounts to varying the correction applied by roughly $\pm 40\%$. This will be discussed in greater detail in Chapter 9.

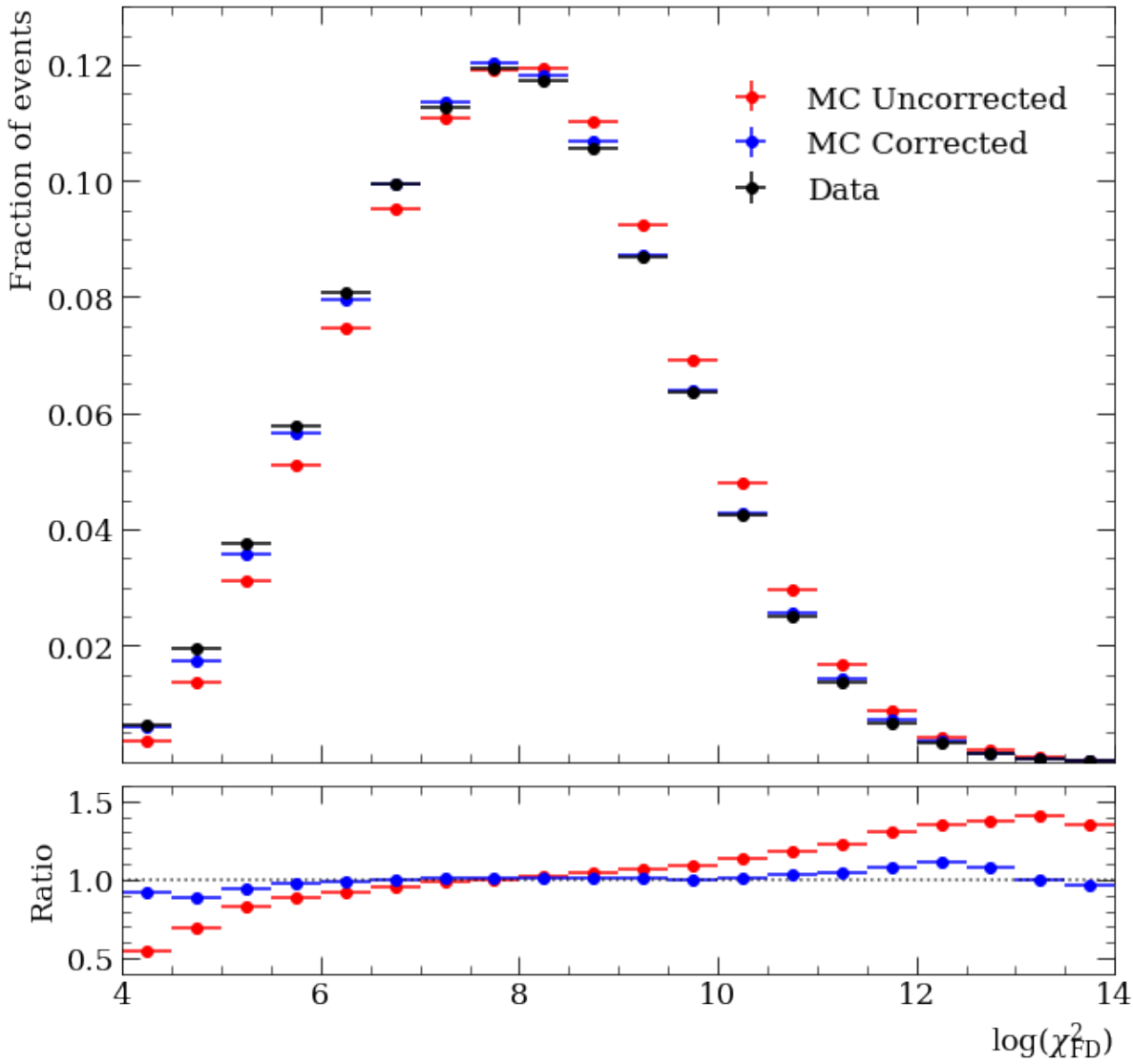

Figure 6.12: Applying a correction for the χ^2_{FD} in simulated $B^+ \to J/\psi K^+$ decays. A comparison is shown between the distributions of $\log(\chi^2_{\mathrm{FD}})$ in data (black), uncorrected simulation (red) and simulation with the correction applied (blue). The ratio of the simulation to data is shown for the corrected and uncorrected simulation.

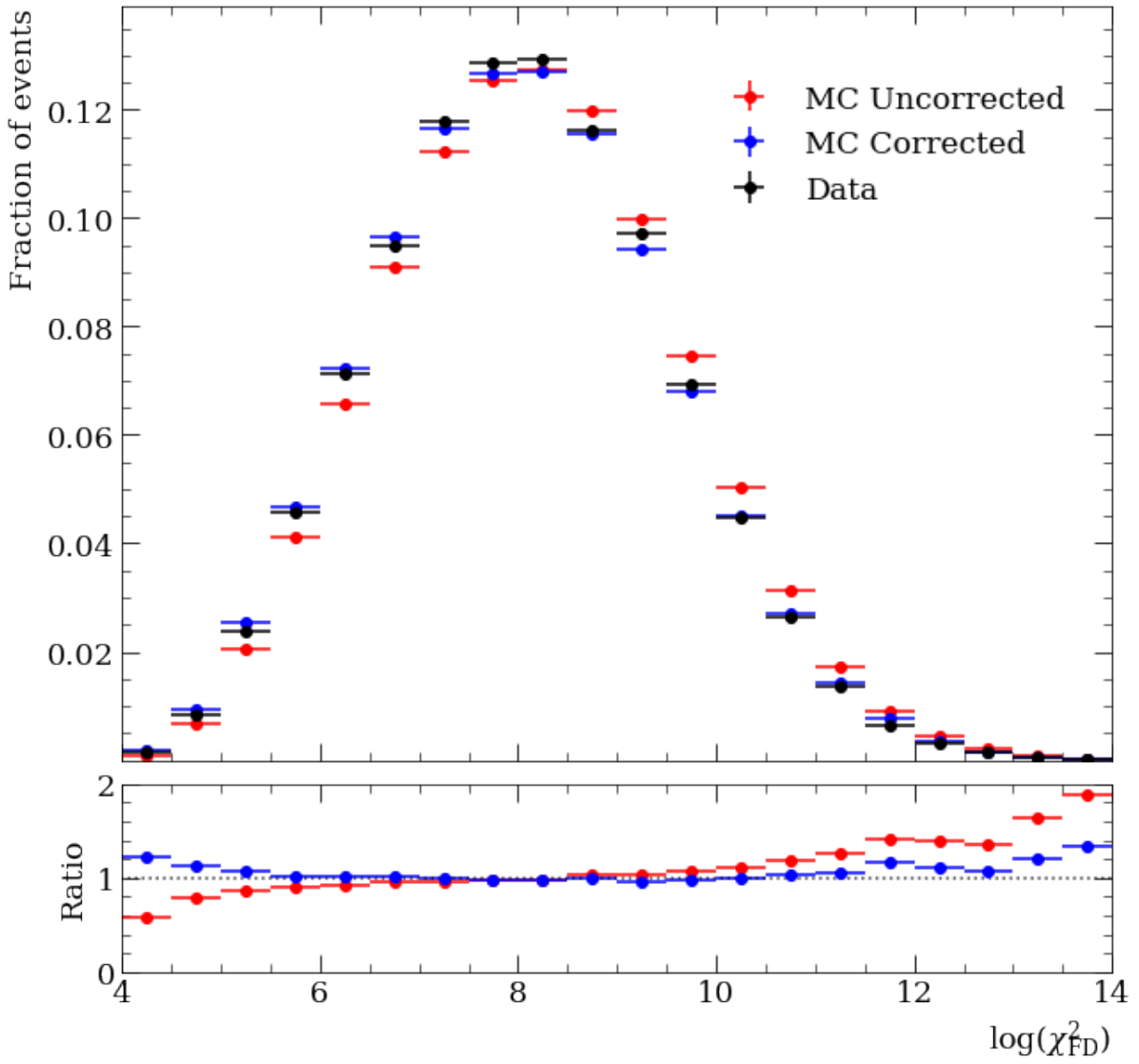

Figure 6.13: Applying a correction for the $\log(\chi^2_{\mathrm{FD}})$ in simulated $B^0 \to J/\psi K^*$ decays as a cross-check. The correction here is derived in $B^0 \to J/\psi K^+$ decays. A comparison is shown between the distributions of $\log(\chi^2_{\mathrm{FD}})$ in data (black), uncorrected simulation (red) and simulation with the correction applied (blue). The ratio of the simulation to data is shown for the corrected and uncorrected simulation.

Chapter 7

Determining the sample purity

Now that the simulation has been corrected for mismodelling effects, it is possible to begin to determine the yield of forward and backward $b\bar{b}$-dijet events in the data. This is done by fitting the observed distribution of the variable $\log(\chi^2_{\mathrm{FD}})$ for forward and backward events in the data using the simulated shapes for the signal and background distributions.

The data are fit to extract the number of $b\bar{b}$-dijet events in each bin of m_{jj}, using extended log likelihood fits, as implemented in the RooFit package [115]. The bins of dijet invariant mass have edges $m_{jj} = [50, 75, 105, 150]\,\mathrm{GeV}$. Fits are also performed to underflow and overflow bins for events below the lower mass limit and above the upper mass limit respectively[1]. Events in the underflow and overflow bins pass all of the other selection requirements used in the analysis. The minimum dijet invariant mass in the underflow bin is around 40 GeV due to the transverse momentum selection requirements applied to the jets. Forward and backward events are considered separately since this allows the asymmetry to be explicitly found. The fits are carried out using the variable $\log(\chi^2_{\mathrm{FD}})$. The χ^2_{FD} of a jet measures the compatibility of the flight distance of the SV with zero, the calibration of which is discussed in Section 6.3. The main sources of contamination are expected to be from charm-jets, either in $c\bar{c}$-dijet events or bc-dijet events[2]. Backgrounds from events containing light-jets, labelled q-jets, are also considered. These backgrounds are possible as q-jets can potentially be mistagged with a SV and hence incorrectly identified as HF jets. As light quarks don't fly through the detector before decaying in the same way that heavy quarks do, these SVs identified in the light-jets must

[1] A value of the asymmetry will not be explicitly determined in these bins. However, they are included to allow the handling of events that migrate into and out of the analysis during the unfolding procedure.

[2] Although not explicitly included in the label, either, both or neither of the quarks in a bc-dijet could be an antiquark. The label used is a useful shorthand to include all such events.

be fake. For example, this can happen because of the finite detector resolution; occasionally multiple particles associated with the PV are misreconstructed and a false secondary vertex is identified. Using results from the previous analysis of the charge asymmetry in beauty-dijets at LHCb, the contribution from q-jets is expected to be negligible.

The work in this chapter is laid out as follows. First, to confirm the negligible contribution from q-jets, a one-dimensional fit is carried out to data using a single jet in the event. This is followed by a theoretical estimation of the sample composition using other results at the LHC, along with knowledge of the relevant efficiencies of selection requirements used in this analysis. Other possible backgrounds are then discussed. Finally, a two-dimensional fit to the data will be performed using both jets in the event, in order to extract the full event yields.

7.1 Estimation of dijet compositions

7.1.1 Light-jet fits

This subsection reports work conducted by Dr William Barter carried out in close association with the author.

To confirm that the light-jet component of the analysed data sample is negligible, a template fit is performed using a b-jet, c-jet and light-jet component. Fits performed here are one-dimensional with the $\log(\chi^2_{\text{FD}})$ of a randomly selected jet from the dijet system used for the fit. Templates are defined as one-dimensional histograms of this variable. The fit is carried out in bins of the dijet invariant mass. To fill the b and c templates, jets are taken from the simulated samples discussed in Section 5.1. To fill the light-jet template, events are required that accurately modelled the behaviour of light-jets making it into the final selection. To do this a technique previously used in LHCb analyses is employed by finding jets with so-called "backwards travelling SVs" [91]. A backwards travelling SV is one that lies in the negative z-direction relative to the PV associated with the event, such that if the coordinates of the "SV−PV" position vector were inverted they would lie within the jet. This configuration is shown in Figure 7.1. All other SV-tagging requirements are applied as normal. The tracks that form these fake SVs will be mismeasured tracks associated with the PV. As a light-jet should be equally likely to fake a SV in front of the PV as behind it [91], jets with backwards travelling SVs are used from data to fill the light-jet template shapes. These SVs must be faked as a backward travelling SV can-

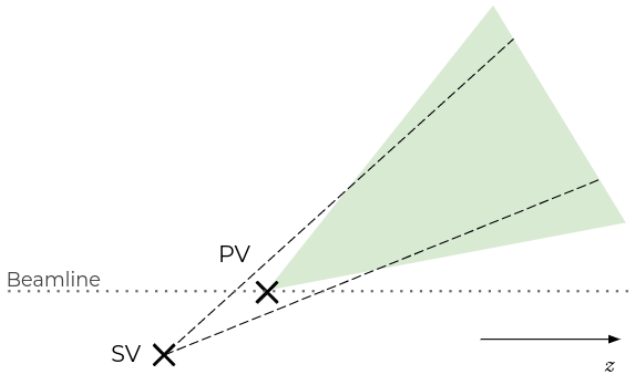

Figure 7.1: Diagram showing the positions of the PV, SV and jet for a backward travelling SV. The PV and SV are shown as black crosses. The tracks making up the SV are depicted as dashed black lines and the jet is the green-shaded region, other tracks within the jet are not shown. Tracks forming the SV truly originate from the PV but are reconstructed as a SV due to mismeasurement. This diagram is not to scale.

not initiate forward-going particles in the jet. Events with backwards travelling SVs are normally removed by selection requirements in the SV-tagging procedure. The backward travelling SV events were confirmed to well model light, SV-tagged jets in Ref. [91]. The events are taken from data meaning that there will be no issues with mismodelling the jets in simulation. The shapes of the templates used are shown in Figure 7.2.

The fit results show that the light-jet component is consistent with zero in each of the three central mass bins used in the analysis. The fit provides the largest light-jet component in the 105–150 GeV bin, which is at the level of $0.15^{+0.80}_{-0.15}\%$ of the total events. The result for this bin is shown in Figure 7.2. All other bins in the fit are also consistent with no light-jet component. The light-jet component is therefore expected to be negligible and hence is not considered for the full analysis fits. A systematic uncertainty associated with the possible presence of a light-jet component will be evaluated in Chapter 9.

7.1.2 Theoretical estimation

The results set out in Section 7.1.1 can be cross-checked against other results at the LHC, using knowledge of the various SV-tag and qTAG efficiencies. The production ratios of dijets with different flavour compositions were measured at ATLAS in pp collisions with a centre-of-mass energy of $\sqrt{s} = 7\,\text{TeV}$ [116]. This was carried out in bins of the leading jet p_{T}, with the lowest bin being 40–60 GeV. Whilst this, and the angular coverage, do not match the kinematics in the main analysis here, these numbers allow for a back-of-the-

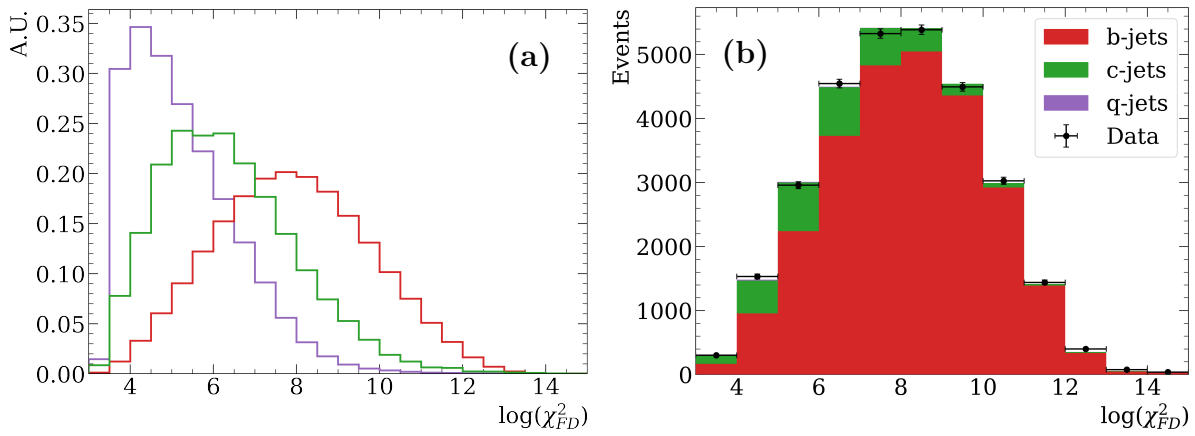

Figure 7.2: (a) Single jet template shapes for b, c and light (labelled q) jets. (b) Fit result from the 105-150 GeV bin using a single random jet from the event. A light-jet template is included in the fit. This mass bin yielded the highest light-jet component, which is still negligible and compatible with zero. The low level of light jets means the component is not visible in the plot. Work for these plots was carried out by Dr William Barter, in discussions with the author.

envelope check of the expected light quark contribution (as well as the contribution from other dijet flavour compositions). Numbers from ATLAS are used as there is no analysis that provides the production ratios of dijets at LHCb that includes mixed-flavour dijets. The production ratios from ATLAS are shown in Table 7.1, with the results normalised to the $b\bar{b}$ production fraction.

All events in the analysis are subject to the requirement that both jets are tagged with a SV, the process for which is discussed in Section 3.5. The efficiency for a single jet to be SV-tagged is measured in simulated jets produced for this analysis as: $\sim 50\%$ for a b-jet, $\sim 15\%$ for a c-jet, and $\sim 0.5\%$ for a q-jet. As both jets are required to be SV tagged these probabilities must be combined for dijets of different flavour compositions. The efficiency of tagging the dijets is shown in Table 7.1.

Next, the impact of requiring one jet in the event to pass the qTAG requirement is considered. The efficiency of qTAG is determined using simulation for individual jets of different flavours as described in Section 5.3. The efficiency of tagging a dijet event with one jet flavour i and the second flavour j is given by

$$\begin{aligned} \varepsilon_{ij} &= 1 - (1 - \varepsilon_i)(1 - \varepsilon_j), \\ &= \varepsilon_i + \varepsilon_j - \varepsilon_i \varepsilon_j, \end{aligned} \tag{7.1}$$

since only one jet needs to be tagged in the event. The results for dijets of different flavour

Table 7.1: Production ratios of dijets with different flavour compositions as measured at ATLAS with the leading jet having a transverse momentum in the range $40 < p_{\mathrm{T}}(\mathrm{jet}) < 60\,\mathrm{GeV}$ [116]. These have been normalised to the $b\bar{b}$ production fraction and the uncertainties are the combination of statistical and systematic effects. The final column is normalised to sum to unity. Trigger efficiencies are not considered here.

Dijet flavour	Production ratio	Dijet SV-tag efficiency	qTAG efficiency	Back-to-back probability	Total Ratio
$b\bar{b}$	1.00 ± 0.19	0.25	0.11	1	0.909
$c\bar{c}$	1.7 ± 0.7	0.023	0.047	1	0.058
$q\bar{q}$	127.8 ± 3.2	2.5×10^{-5}	0.010	1	< 0.001
bc	0.8 ± 0.4	0.075	0.080	0.17	0.025
bq	6.3 ± 0.7	2.5×10^{-3}	0.062	0.17	0.005
cq	16.3 ± 2.7	7.5×10^{-4}	0.029	0.17	0.002

compositions are again listed in Table 7.1.

Finally, for the dijets which have a mixed flavour (bc, bq and cq) the probability that these jets are found back-to-back must be considered as this is included in the fiducial requirements applied to the events. Whilst the jets in same-flavour dijet events are generally both produced in the same process this is not necessarily true for mixed-flavour events. The expectation that these jets are back-to-back therefore no longer holds. It is assumed here for simplicity that the two jets are completely uncorrelated in the direction that they are produced. Whilst this is the most extreme possible case such an assumption is appropriate for the rough calculation carried out here. The probability that two uncorrelated jets satisfy the criteria $\Delta\phi^{\text{jet-jet}} > 2.6\,\text{rad}$ is given by $p = (\pi - 2.6)/\pi \approx 17\%$. For the purpose of this estimate, it is assumed that all same-flavour dijets will be back-to-back in their azimuthal angle.

Using a combination of the values above it is possible to calculate a rough estimate of the level of each dijet flavour in the final data sample. The total ratios are shown in Table 7.1. The predicted light-jet contribution is below the level of 1%. This is in good agreement with the single jet fits performed in Section 7.1.1. The result is also in line with what was seen in the previous analysis of the charge asymmetry in beauty-jet production at LHCb, where light-jet contributions were found to be negligible [95]. It is important to note that the calculation provided here is approximate and relies upon ATLAS measurements in a different region of the phase space. The final composition ratios provided in Table 7.1 should not be taken to be an exact prediction. However, it remains clear that dijets containing a light-jet are heavily suppressed by the requirements of SV tagging and qTAG. This leaves the prediction that only $b\bar{b}$, $c\bar{c}$ and bc components are present at any noticeable level in the final sample, of which $b\bar{b}$ is dominant. This further confirms the results of Section 7.1.

7.1.3 Contribution from $t\bar{t}$ decays

Decays of the top quarks in $t\bar{t}$ events can also yield $b\bar{b}$-dijets and so their contribution is discussed here. The contribution from $t\bar{t}$ events was considered in the first measurement of the charge asymmetry at LHCb and it was found that this was negligible, with $\mathcal{O}(10)$ $t\bar{t}$ events expected within the LHCb acceptance in the entire sample [95]. The cross-section at the LHCb for $t\bar{t}$ is around 10 times larger at the centre-of-mass energy $\sqrt{s} = 13\,\text{TeV}$ compared to $\sqrt{s} = 7\,\text{TeV}$ [117]. Therefore it is expected that the contribution to the sample here is of $\mathcal{O}(100)$ events, making up $< 0.01\%$ of the total sample. This is still

clearly a negligible contribution and $t\bar{t}$ events are not considered in this analysis[3].

7.1.4 Contribution from τ decays

Another potential source of contamination in the sample is from τ decays. As these particles have a similar lifetime to D mesons [13] it is possible that the decay might be identified as a SV and lead to the reconstruction of a SV tagged jet. This may lead to a background for the analysis if the τ is produced in association with another jet, or if a τ pair is produced and both are identified as SV-tagged jets. Even in these cases is unlikely for a τ to pass the qTAG requirements. To be tagged as a jet containing a SV the τ could decay to three or more hadrons, but this does not also produce a muon and so qTAG would not tag such a jet. Instead, the τ could decay directly to a muon, but this muon is typically isolated and the misidentification of a lepton as a jet at LHCb has a very low probability. Furthermore, the production mechanisms for τ leptons are electroweak. This means the cross-section of τ production is suppressed relative to dijet production mechanisms which include strong force processes. These potential sources of τ background have been considered and it is found that the level of contamination is less than 100 events, making up $< 0.01\%$ of the total sample. This level is also clearly negligible and hence no further consideration is given to these types of event.

7.2 Two-dimensional fit procedure

A fit to the data is performed to extract the number of $b\bar{b}$-dijet events in each bin of the dijet invariant mass, for the forward and backward events separately. The only significant sources of background contamination are expected to be from charm-jets, either in $c\bar{c}$-dijet events or mixed-flavour bc-dijet events.

Although flavour tagging BDTs have been used at LHCb [91], it is found in this analysis that the data-simulation agreement is not sufficient in the output of these BDTs to perform an accurate fit. This is largely due to detector degradation, which is discussed in Section 6.3. Instead, the $\log(\chi^2_{\text{FD}})$ of each jet is used to perform template fits. This

[3]It is noted that if the contribution were larger it would be important to discern where these effects should be considered in the overall measurement. As the decay from $t\bar{t}$ events still produces a $b\bar{b}$-dijet event, it may be that this production method should be included in theoretical calculations, at which point a measurement of the asymmetry would not need to remove these events from the sample. If instead they remained excluded from theoretical calculations then these events would have to be accounted for appropriately in the measurement, either by removing them or assigning some systematic uncertainty.

variable has been calibrated using $B^+ \to J/\psi K^+$ decays with the procedure laid out in Section 6.3. As shown in Figure 7.2 this gives good distinguishing power between the three flavours of jets. Attempts were also made to calibrate the corrected mass of the SVs, defined in Eq. (6.3), to carry out a fit to the data, following the strategy in the first measurement of the charge asymmetry at LHCb [95]. However, this approach is not useful here due to the level of mismodelling seen in simulation for this variable and the significant difficulty associated with accurately calibrating it.

The fit is performed to a two-dimensional histogram of the data with variables $[\log(\chi^2_{\rm FD}(j_0)), \log(\chi^2_{\rm FD}(j_1))]$, where $j_{0,1}$ are the two SV tagged jets in the event with a random ordering. Random ordering is useful as for the mixed flavour bc events the jets are equally likely to be of each flavour. This means that the template shapes are not dependent upon the modelling of any variable used to order the jets. There are 20×20 bins[4] with a range for each variable set to [3,14] which covers over 99% of events in the data sample. The fits are performed for forward and backward events separately in each bin of the dijet invariant mass. For the $b\bar{b}$-dijet signal, the template PDF is sampled from a two-dimensional histogram filled with simulated events, treating the simulation identically to data when producing the templates. For the $c\bar{c}$-dijet and the mixed flavour bc-dijet backgrounds, the two-dimensional templates used in the fit are taken by convolving two one-dimensional distributions drawn from the relevant beauty and charm simulated samples. This follows the approach considered in Ref. [106]. The templates are shown in Figure 7.3. It can be seen from these distributions that the templates occupy different regions of the phase space. It should be noted that in one dimension the bc component cannot be distinguished from an equal combination of $b\bar{b}$- and $c\bar{c}$-dijets. The power to distinguish these mixed-flavour dijets only comes with the full two-dimensional template. The two-dimensional distribution in data is also shown in Figure 7.4.

As a cross-check, the $c\bar{c}$ template is also explicitly constructed as a two-dimensional PDF in the same manner as the $b\bar{b}$ template[5]. This returns consistent values of the asymmetry and is further documented in Section 9.2. It is challenging to produce a template in the same manner for the bc component as there is no efficient option in Pythia 8 to produce mixed-flavour dijets across the full phase space of the analysis [38]. A cross-check is also performed using one-dimensional templates of beauty and charm

[4]The number of bins was varied and the results remain consistent, with no significant change in sensitivity seen.

[5]This is not used as the main analysis strategy due to the lower number of events in the $c\bar{c}$ simulation for the 2016 run conditions. Additional simulation, produced using 2017 and 2018 run conditions, is used for this cross-check.

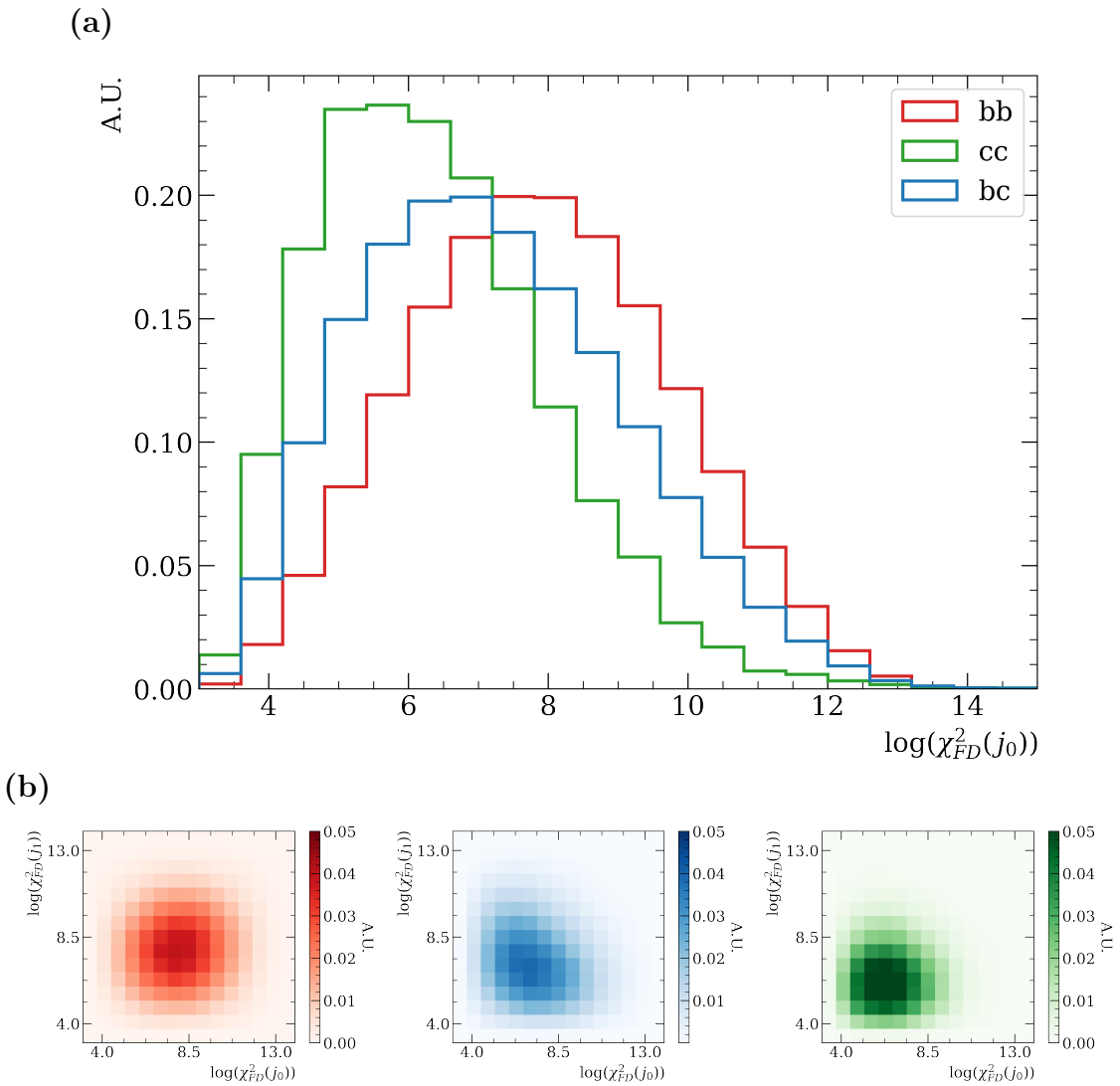

Figure 7.3: Template shapes used in the fit. Templates are defined as two-dimensional histograms $[\log(\chi^2_{\mathrm{FD}}(j_0)), \log(\chi^2_{\mathrm{FD}}(j_1))]$ with 20×20 bins and a fit range of $[3, 14]$ for each direction. Here j_0 and j_1 are the two jets in the event with a random ordering. (a) A projection into the j_0 direction. Due to the random ordering, this will be identical to the j_1 projection. (b) The two-dimensional template shapes.

which require no convolution to produce. This again returns consistent values of the asymmetry and is documented further in Section 9.2.

Although forward and backward events are fit independently, the template shapes are shared across the two directions and are filled with both forward and backward events from the simulation. This is done as the template shapes have negligible differences between the forward and backward directions and sharing the events ensures that

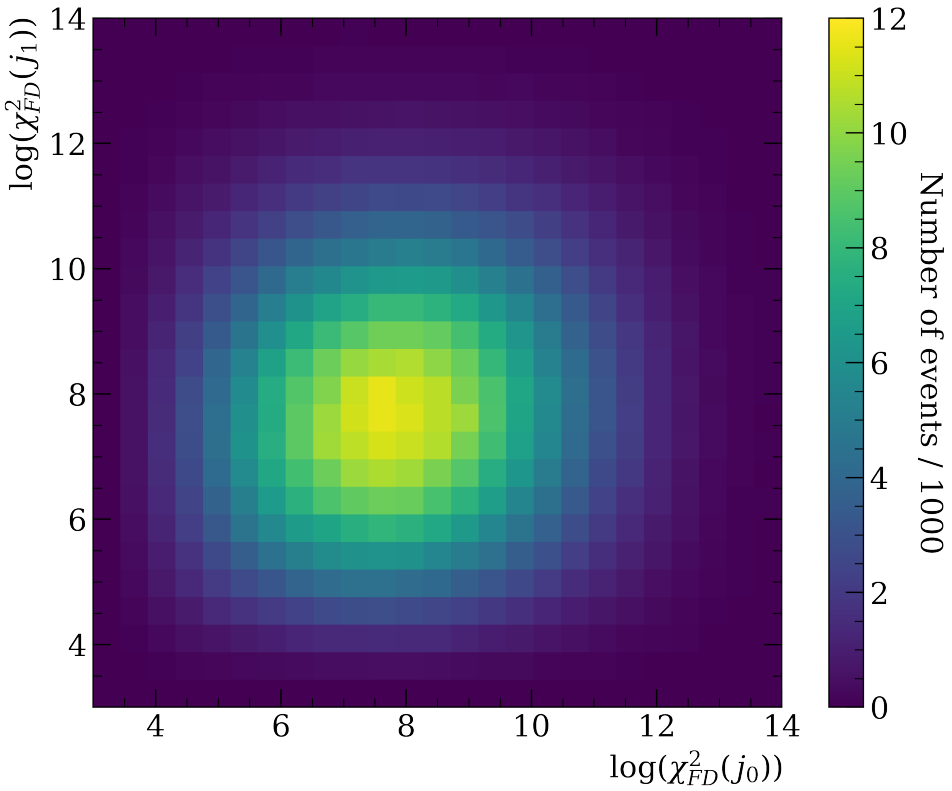

Figure 7.4: The two-dimensional distribution of the total data sample for the variables used in the fit: $[\log(\chi^2_{\mathrm{FD}}(j_0)), \log(\chi^2_{\mathrm{FD}}(j_1))]$, where $j_{0,1}$ are the two jets in the event with a random ordering.

the statistical uncertainty associated with using a finite number of simulated events is minimised. The variation of the template shapes between forward and backward events is shown in Figure 7.5, as estimated in simulated $b\bar{b}$-dijet events. Any variation of the template shapes with the pseudorapidity of the jets is assumed to be well modelled by the simulation.

The templates are used to determine the flavour composition of events in the data using an extended log likelihood fit. The distribution of events in data is modelled as

$$\mathcal{P}_{\mathrm{data}} = N_{\mathrm{bb}}\overline{\mathcal{P}}_{\mathrm{bb}} + N_{\mathrm{bc}}\overline{\mathcal{P}}_{\mathrm{bc}} + N_{\mathrm{cc}}\overline{\mathcal{P}}_{\mathrm{cc}}, \tag{7.2}$$

where $\overline{\mathcal{P}_{\mathrm{ij}}}$ is the normalised distribution of events in the template with jet flavour composition ij and N_{ij} is the number of these events. The number of each event type is

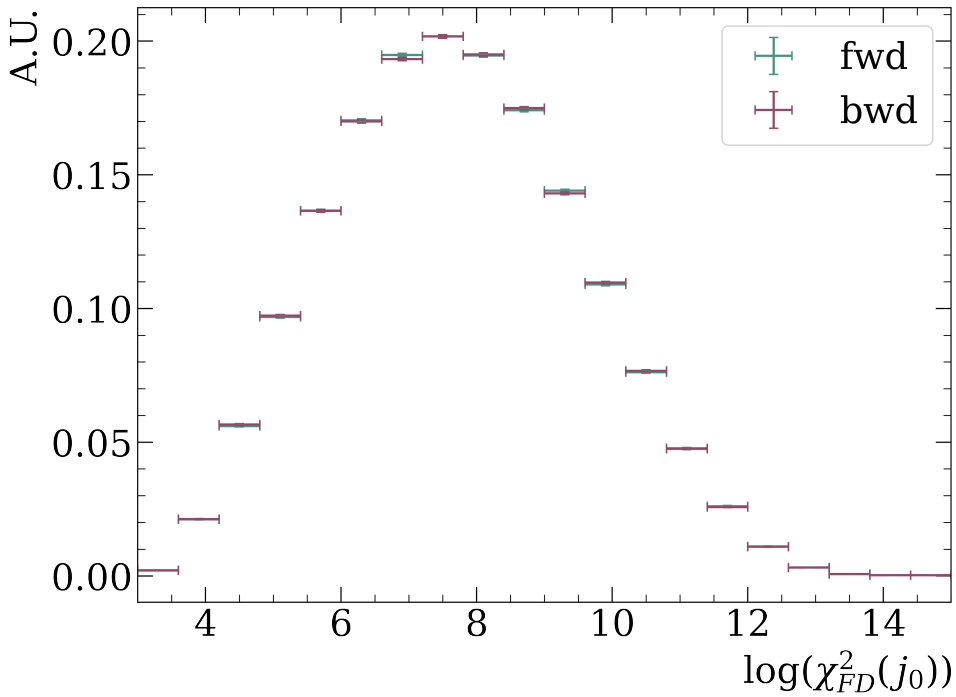

Figure 7.5: Plot showing a projection of the template shape into the j_0 axis and how this varies for forward and backward events. These template shapes were measured in $b\bar{b}$ dijet simulation. There is no significant difference between the shapes, motivating the sharing of templates across both directions.

initialised based on the results from Section 7.1 and these are set to

$$N_{\mathrm{bb}} = 0.90\, N_{\mathrm{data}}, \tag{7.3}$$

$$N_{\mathrm{bc}} = 0.05\, N_{\mathrm{data}}, \tag{7.4}$$

$$N_{\mathrm{cc}} = 0.05\, N_{\mathrm{data}}. \tag{7.5}$$

The fit accounts for the statistical uncertainty in the shapes of the distributions observed in data but the uncertainty on the templates generated from simulation must also be taken into account. As it cannot be assumed that the template shapes are perfect, these uncertainties and their impact on the fit results are discussed here, rather than leaving this until a discussion of the full systematic uncertainties in the analysis. There are three sources of uncertainty in the template shapes that are considered:

1. A systematic uncertainty in the form of the correction applied to the $\log(\chi^2_{\mathrm{FD}})$ distribution derived in Section 6.3.

2. The application of the $\log(\chi^2_{\text{FD}})$ correction to charm-jets.

3. The statistical uncertainty on the template shapes in simulation.

To account for uncertainty on the form of the correction derived from $B^+ \to J/\psi K^+$ decays, two other corrections are applied in turn and the fit results are compared to the nominal fit. The first alternative correction is that derived using $B^0 \to J/\psi K^*$ decays detailed in Section 6.3, referred to as the "K^* correction". The second alternative is obtained by shifting the nominal correction an equal amount in the opposite direction to that seen when using $B^0 \to J/\psi K^*$ decays, referred to as the "$2K - K^*$ correction". This is broadly equivalent to varying the correction applied by $\pm 40\%$. A systematic uncertainty is then assigned by evaluating the maximal difference from the nominal fit compared to the two alternatives.

It is expected that the $\log(\chi^2_{\text{FD}})$ correction should also apply to charm-jets as the issues that motivate the correction in beauty-jets are universal [107, 108]. A systematic uncertainty is considered to account for this assumption. This is evaluated by not applying any correction to charm-jets whilst still using the nominal correction for beauty-jets[6]. Since charm-jets do not form a large proportion of jets in data, it is expected that this effect will be small. A systematic uncertainty is then assigned by evaluating the difference from the nominal fit compared with a fit using the uncalibrated charm-jet templates.

Finally, the template shapes in simulation are generated from a finite number of events and, as such, will have a statistical uncertainty associated with them. The Beeston–Barlow method [118] for propagating the uncertainties associated with a finite number of simulated events is sometimes used in HEP analyses. This method can suffer from issues when using weighted events, as is done here, which is recognised in the original paper. For these reasons and owing to the large number of simulated events compared to data, the uncertainty is instead approximated using fit results from so-called toy studies. The fits are repeated 500 times with the template shapes fluctuated within their statistical uncertainties. Each of these repeats is referred to here as a toy. Here a systematic uncertainty is assigned by evaluating the standard deviation on the results from these 500 toy fits.

These three sources of uncertainty all provide an uncertainty on the template shapes. As such they are added in quadrature to give a total systematic uncertainty on the fit. This systematic uncertainty is then added in quadrature with the statistical uncertainty

[6]Whilst this is a conservative approach the uncertainty on the asymmetry from this source remains small when compared to other sources of uncertainty.

from the fit to yield an overall uncertainty in each mass bin of the fit.

The fit results for forward events are shown in Figure 7.6 and for backward events in Figure 7.7, with the total uncertainty in each bin overlaid. The total yields from the nominal fit are shown also in Table 7.2. The projections of the fit result have a good agreement with data within the uncertainties associated with the finite size of the data sample and variations in the template shapes from simulation. As well as this, the numerical results are broadly consistent with the production cross-sections being similar for $b\bar{b}$- and $c\bar{c}$-dijets given the much lower SV-tag and qTAG efficiencies in $c\bar{c}$-dijets. Fit results from the underflow and overflow bins are shown in Figures 7.8 and 7.9 for forward and backward events respectively. The results in these underflow and overflow bins have a minimal impact on the analysis as a whole. They are utilised only when considering events that migrate into and out of the analysis during the unfolding process. Excluding these bins from the analysis is found to have a negligible impact on the overall measurements.

It is necessary to correct the dijet yields for detector and mistagging effects (discussed in the following chapter). However, it is possible at this stage to see the level of uncertainty in an uncorrected measure of the asymmetry, here referred to as the "raw" asymmetry. This is the asymmetry in the yield of forward and backward events taken directly from the fit results. To achieve this, raw values of the asymmetry are calculated under each of the above systematic variations considered and the shift in the value of the raw asymmetry is taken as the systematic uncertainty. While the systematic variations can cause $\mathcal{O}(50-100\%)$ fluctuations in the background yield, the changes in the value of the asymmetry are much smaller. This is due to the low level of background present in the sample and because the variations in yield are highly correlated for the forward and backward categories. For the K^* and $2K - K^*$ corrections, the largest deviation from the nominal asymmetry is taken out of the two as both of these are concerned with the size of the correction applied to the templates. The deviations from the nominal raw asymmetry are shown in Table 7.3. For the uncertainty generated by the finite number of events in simulated templates, the templates are fluctuated within their uncertainties for 500 toys and the fit is repeated to extract raw asymmetries. The standard deviation is taken as the systematic uncertainty from this source. The uncertainty from each systematic source is considered separately and added in quadrature to yield a total uncertainty. This gives

the values of the raw asymmetry as

$$A^{\text{raw}}_{50-75} = (0.21 \pm 0.13 \pm 0.13)\%, \tag{7.6}$$

$$A^{\text{raw}}_{75-105} = (0.7 \pm 0.4 \pm 0.2)\%, \tag{7.7}$$

$$A^{\text{raw}}_{105-150} = (0.1 \pm 0.9 \pm 0.6)\%, \tag{7.8}$$

where the subscript indicates the mass window, the first uncertainty is statistical and the second is systematic. From this, it can be seen that the systematic and statistical uncertainties associated with the fit in the lowest mass bin are roughly equivalent, whereas, in the other two bins, statistical uncertainties are dominant. This is due to the fact the lowest mass bin has a much higher overall yield of events than the other two bins as shown in Table 7.2. The statistical uncertainties reported by the fit are also validated by running 500 toy fits with the data distributions fluctuated within their statistical uncertainties. The results from this gave statistical uncertainties consistent with those returned from the fit and provided above.

The distributions of forward and backward events in the mass windows now need to be unfolded to correct for detector effects, which will be detailed in Chapter 8. These unfolded yields will then be used to calculate the final value of the asymmetry, with the systematic and statistical uncertainties propagated through this calculation. Further cross-checks for these fit results will also be considered in Chapter 9.

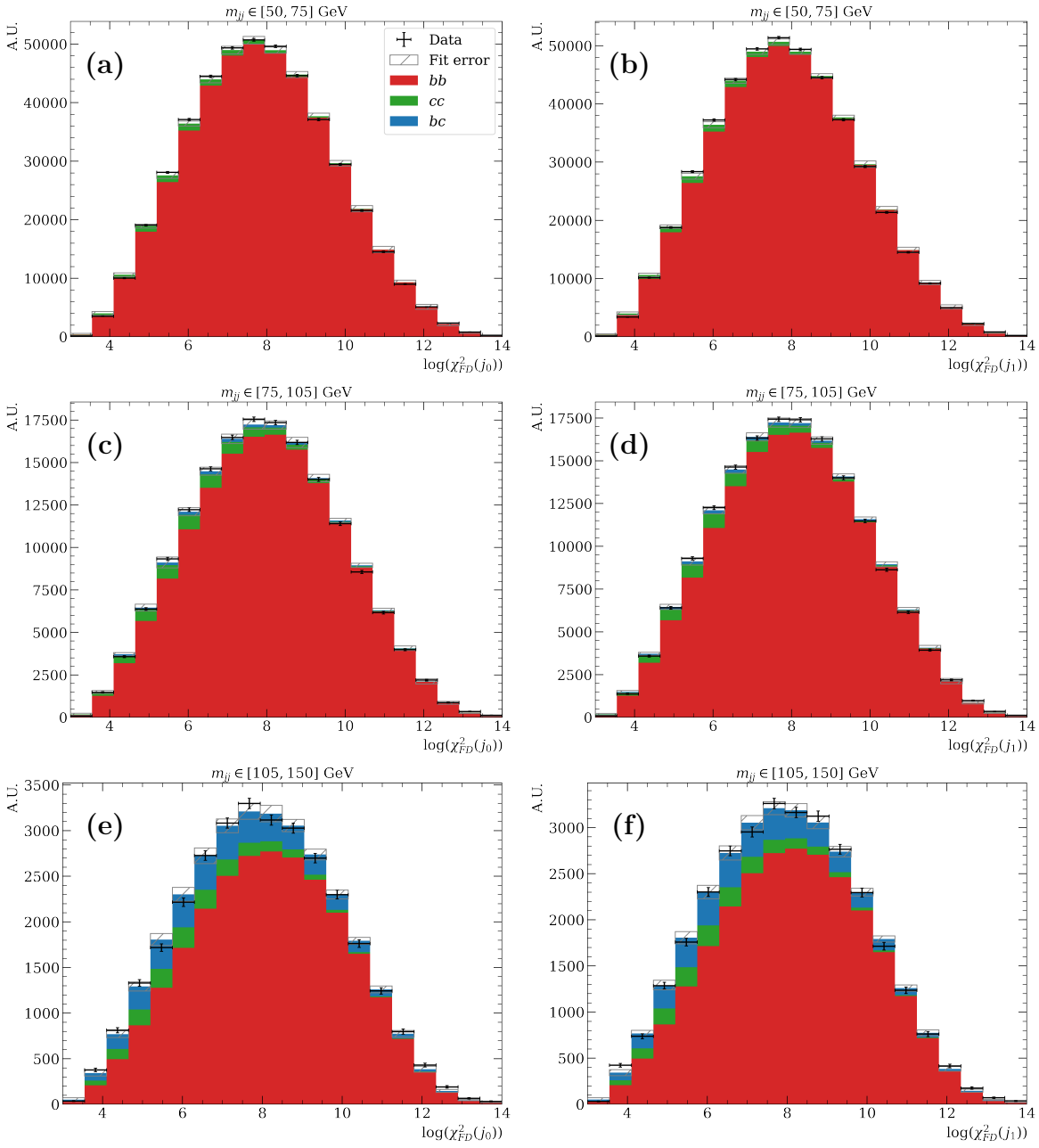

Figure 7.6: Plots showing the fit results for the forward events projected into the two axes. The uncertainty on the total fit is shown as a grey, hatched area and is a combination of the statistical and systematic uncertainties discussed in the main body of the text. Dijet invariant mass ranges are: top row: $50 - 75$ GeV, middle row: $75 - 105$ GeV, bottom row: $105 - 150$ GeV.

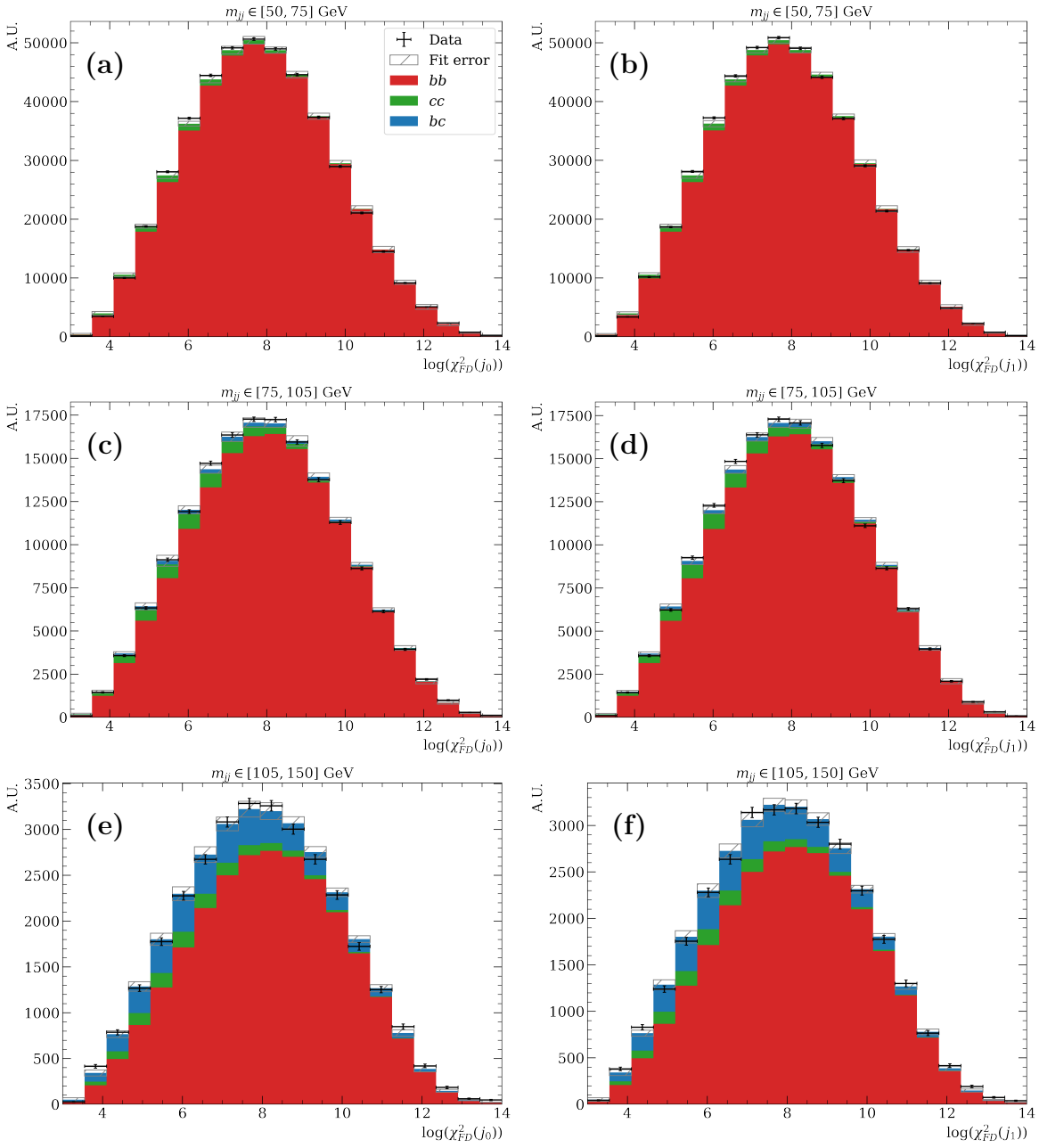

Figure 7.7: Plots showing the fit results for the backward events projected into the two axes. The uncertainty on the total fit is shown as a grey, hatched area and is a combination of the statistical and systematic uncertainties discussed in the main body of the text. Dijet invariant mass ranges are: top row: $50 - 75$ GeV, middle row: $75 - 105$ GeV, bottom row: $105 - 150$ GeV.

Table 7.2: Dijet flavour yields from the nominal fits to data. The uncertainties shown here are statistical. Fit components that float to zero are shown as dashes. Forward and backward events are labelled F and B respectively.

Mass range / GeV	F/B	Number of data events	$b\bar{b}$ component	$c\bar{c}$ component	bc component
Underflow	F	71903	70060±350	1840±230	–
	B	71833	69900±400	1890±270	–
50–75	F	456809	448400±800	8400±500	–
	B	454844	446500±800	8400±500	–
75–105	F	162881	155000±900	6200±600	1700±1300
	B	161383	152900±900	6300±600	2300±1300
105–150	F	31255	26090±340	1640±250	3500±500
	B	31342	26050±340	1240±240	4000±500
Overflow	F	3934	3060±90	650±70	220±90
	B	3907	3030±90	550±70	320±90

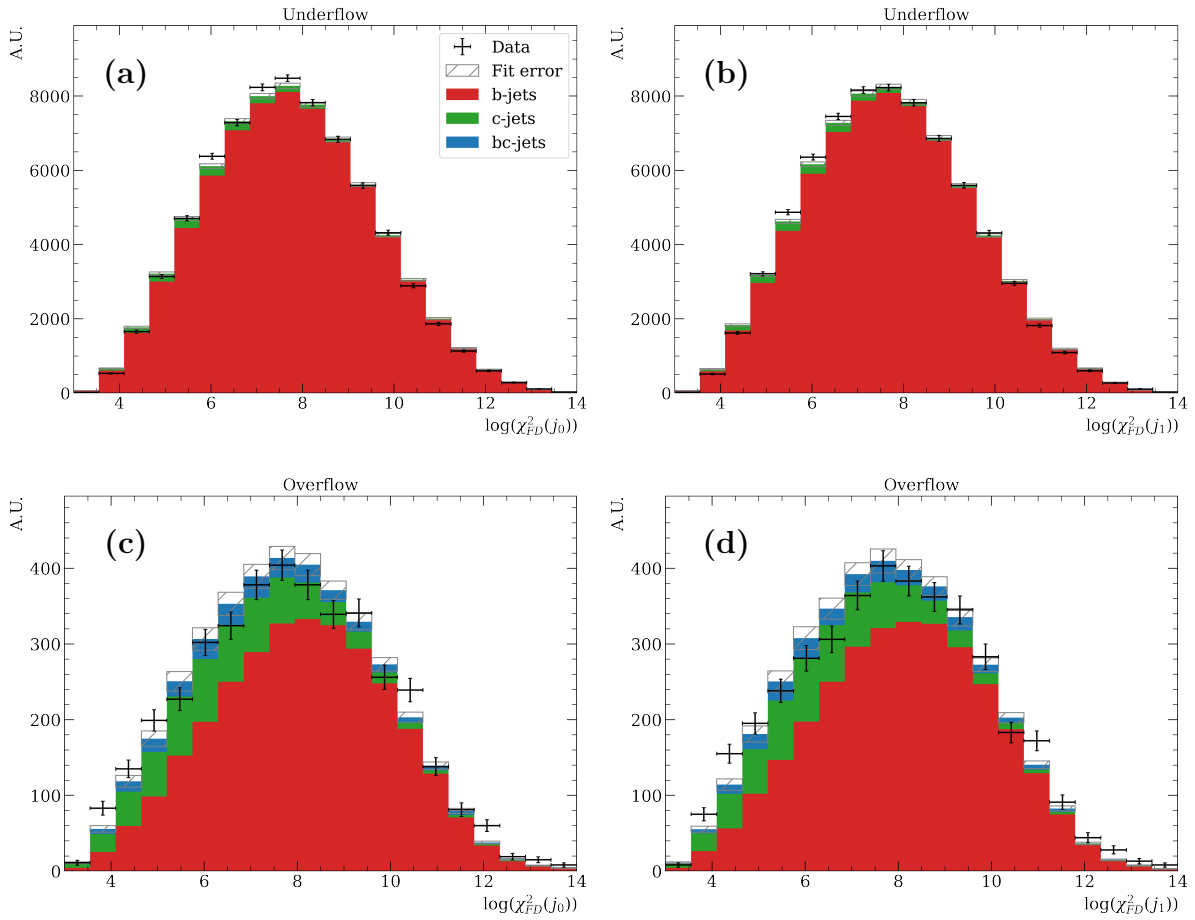

Figure 7.8: Plots showing the fit results for the forward events projected into the two axes, in the underflow (top row) and overflow (bottom row) bins of dijet invariant mass. The uncertainty on the total fit is shown as a grey, hatched area and is a combination of the statistical and systematic uncertainties discussed in the main body of the text. Events in the underflow and overflow bins pass all of the analysis selections but have a dijet invariant mass reconstructed outside of that considered for the measurement of the asymmetry. Events in the underflow bin are those reconstructed with a dijet invariant mass below the analysis threshold of 50 GeV; these have a minimum reconstructed dijet invariant mass of around 40 GeV due to selection requirements on the jet transverse momentum. Events in the overflow bin are reconstructed with a dijet invariant mass above 150 GeV.

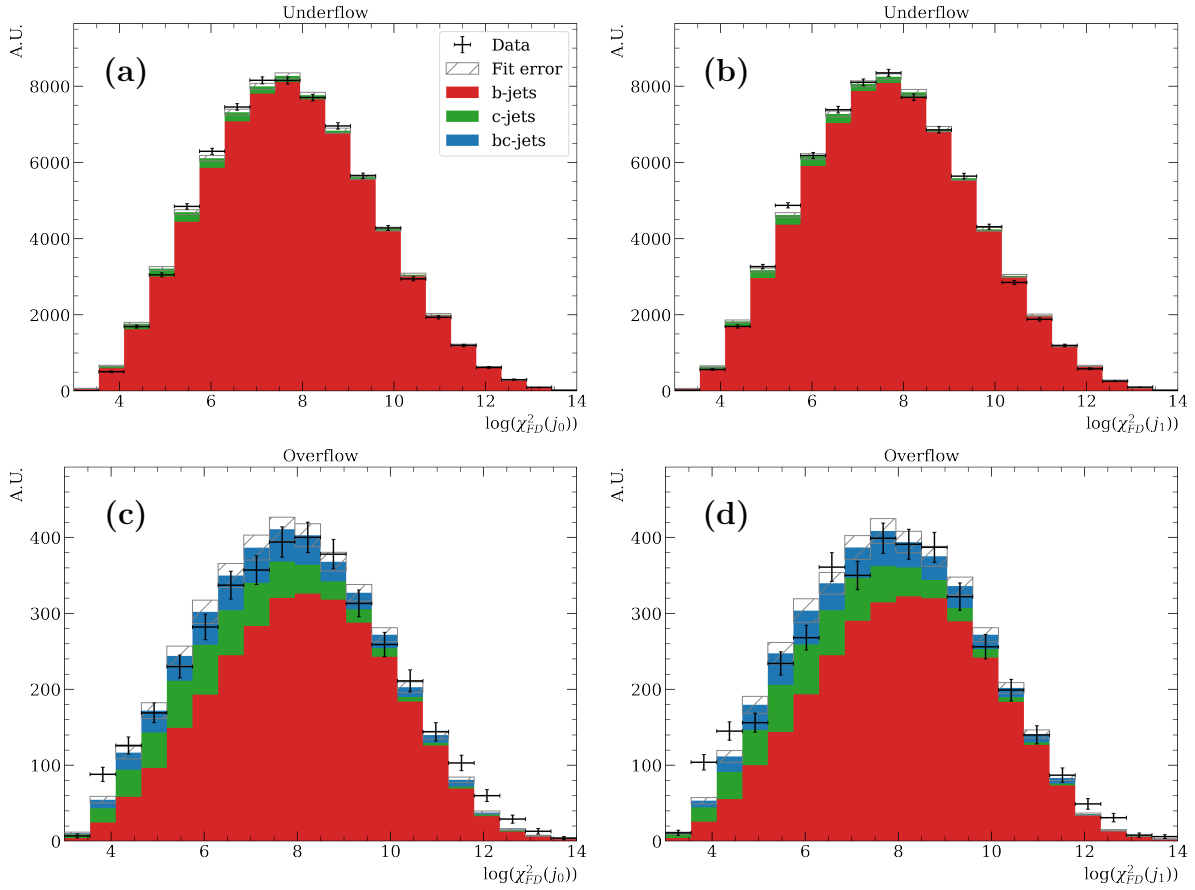

Figure 7.9: Plots showing the fit results for the backward events projected into the two axes, in the underflow (top row) and overflow (bottom row) bins of dijet invariant mass. The uncertainty on the total fit is shown as a grey, hatched area and is a combination of the statistical and systematic uncertainties discussed in the main body of the text. Events in the underflow and overflow bins pass all of the analysis selections but have a dijet invariant mass reconstructed outside of that considered for the measurement of the asymmetry. Events in the underflow bin are those reconstructed with a dijet invariant mass below the analysis threshold of 50 GeV; these have a minimum reconstructed dijet invariant mass of around 40 GeV due to selection requirements on the jet transverse momentum. Events in the overflow bin are reconstructed with a dijet invariant mass above 150 GeV.

Table 7.3: Results from systematic uncertainty studies of corrections applied to templates. Values quoted are the absolute values of the differences in raw asymmetry from the nominal fit result. The maximum value is taken between K^* and $2K - K^*$ for the systematic uncertainty as both of these consider the size of the correction applied. The level of precision quoted for the individual sources is set by the dominant systematic uncertainty for a given dijet invariant mass bin.

Dijet mass / GeV	$\lvert\Delta A\rvert$ / %					
	K^*	$2K - K^*$	Maximum	Uncorrected charm	Simulated sample size	Total
50–75	0.12	0.00	0.12	0.01	0.06	0.13
75–105	0.06	0.12	0.12	0.04	0.18	0.22
105–150	0.1	0.5	0.5	0.0	0.3	0.6

Chapter 8

Unfolding the measured results

Unfolding is a general term used to describe corrections that are applied to measured results and distributions in order to recover the "true" values. These true values are the values that would be obtained by a measurement under idealised conditions, not distorted by effects such as the finite detector resolution. This is studied using simulated events which model the detector effects on the true underlying information. It is assumed in this work that the resolution on the dijet invariant mass does not change with the pseudorapidity of the jets. This is because the jets are chosen to be significantly far from the detector limits that edge effects should not be important. This assumption could be studied in more detail in future iterations of this measurement.

The results of the fits performed in Chapter 7 provide a distribution of $b\bar{b}$-dijet events in bins of the dijet invariant mass in both the forward and backward directions. Here, there are two aspects of the measurement that must be corrected for. First, bin migrations in the distribution of the dijet invariant mass which is caused by the finite detector resolution of this variable. Second, mistagging the charge of the event using qTAG which causes migrations between the forward and backward categories.

It is found that the unfolding of the dijet invariant mass only has a small impact on the final value of the asymmetry. This is driven by the width of the bins used in the dijet invariant mass and the fact that the asymmetry does not vary significantly between these bins. Multiple variations of the procedure were tested and gave differences in the asymmetry that are small when compared with the statistical uncertainties. It is noted therefore that whilst there may exist more rigorous testing and optimisation methods for the unfolding, these are often much more complex and their use is unwarranted in this case. Even though the impact on the measurement is minimal, work into the unfolding

procedure is still laid out here to demonstrate this fact, and for completeness. When this measurement is repeated using the full Run II data (or beyond that in Run III) it may be necessary to analyse again the impact of the unfolding and the optimisation with regards to the statistical and systematic uncertainties.

8.1 Dijet invariant mass

The asymmetry is measured in bins of the dijet invariant mass and for this, the yields of events in these bins must be calculated. The fits documented in Chapter 7 provide this for the invariant masses measured by the detector, but detector resolution effects mean events can be reconstructed in different bins when compared with their true underlying values. It is therefore required to unfold the distribution of events in the dijet invariant mass for both the forward and backward categories.

The method for unfolding used here is Iterative Bayesian Unfolding, also referred to as D'Agostini unfolding [119,120], implemented in Python in the package PyUnfold [121]. The basis for this unfolding method is the use of Bayes' theorem

$$P(t_i|r_j) = \frac{P(r_j|t_i)P(t_i)}{\sum_k [P(r_j|t_k)P(t_k)]}, \tag{8.1}$$

where the terms are defined as follows:

- $P(r_j|t_i)$ is the conditional probability of an event being reconstructed in a dijet invariant mass bin, r_j, given it has a true value lying in the bin t_i and vice versa for $P(t_i|r_j)$. These probabilities form a matrix, which is referred to as the response matrix. This can be calculated using simulated events.

- $P(t_i)$ is the prior probability of having a true dijet invariant mass in a bin t_i. This is again taken initially from simulated events.

The result $P(t_i|r_j)$ can then be applied to the measured distribution (with the measured values forming a column vector) to recover an estimate of the true underlying distribution. The principle of Iterative Bayesian Unfolding is the repeated application of Eq. (8.1), each time updating the prior used to be the unfolded result from the previous iteration.

The bins of dijet invariant mass in which the asymmetry is measured have edges $[50, 75, 105, 150]$ GeV. For the unfolding process it is also beneficial to consider events that migrate into and out of the range defined by these bins, which is done by including

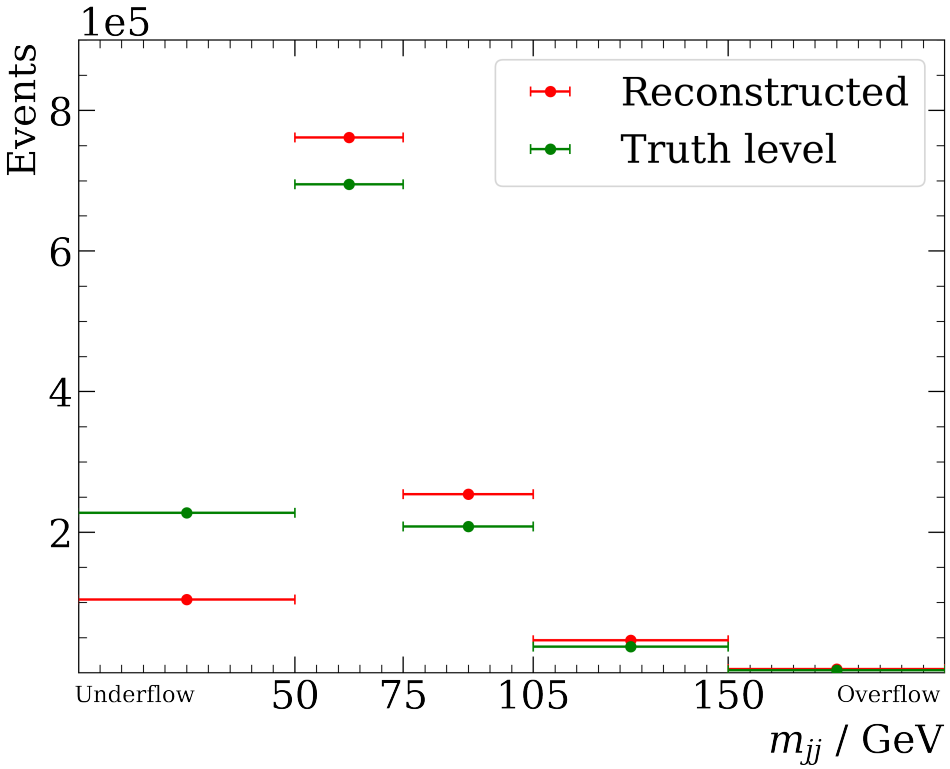

Figure 8.1: True and reconstructed distributions of the dijet invariant mass in simulated events.

an underflow and an overflow bin. This goes beyond considerations made in Ref. [95] which did not consider migrations in and out of the range in which the asymmetry was measured. The initial yield of events in these bins is also determined during the fit procedure with the results in Table 7.2. The underflow and overflow bins are filled with events that pass the fiducial cuts described in Section 5.1 but otherwise have a dijet invariant mass that is outside of the range of the main analysis bins[1].

The distribution of true and reconstructed mass for simulated events is shown in Figure 8.1. The truth distribution shown here is used as the initial prior in the unfolding procedure. The response matrix is taken from the full set of simulated $b\bar{b}$-dijet events discussed in Section 5.1. The forward and backward events are therefore unfolded using the same response matrix, as was also done in the previous measurement of the charge asymmetry at LHCb [95]. This ensures that the impact of the uncertainty associated with the finite size of the simulation sample is minimised in the unfolding procedure.

[1] An upper limit on the dijet invariant mass is set at 500 GeV for the overflow bin. The level of events migrating between the main analysis bins to anything above this threshold is negligible.

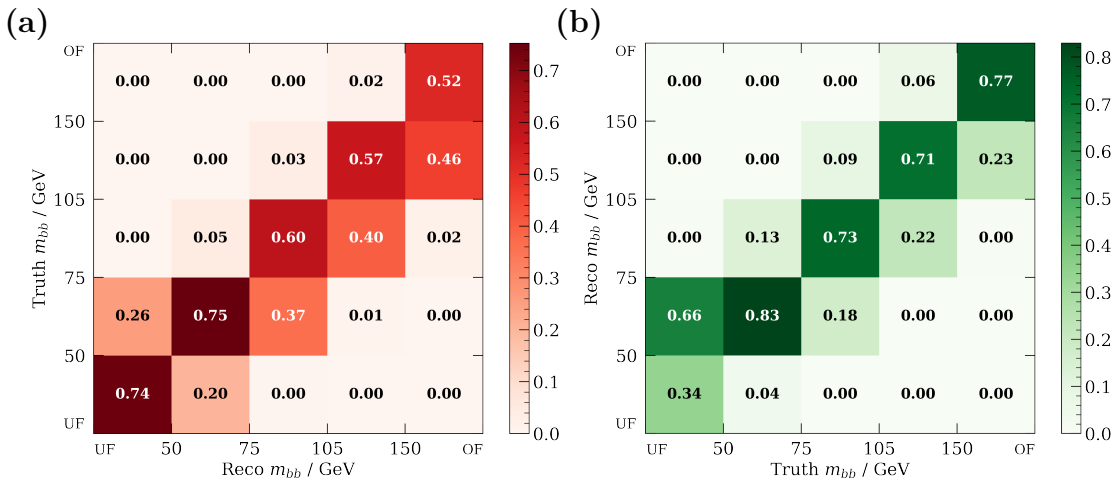

Figure 8.2: (a) The corrections applied in the first round of the iterative unfolding procedure. These are taken directly from simulated $b\bar{b}$-dijet events. The values in the matrix give the probability that an event reconstructed within a dijet invariant mass bin came from the respective true bin. (b) The response matrix taken directly from simulated $b\bar{b}$-dijet events. The values in the matrix give the probability that an event that has a true dijet invariant mass in a given bin is reconstructed in the respective reconstruced bin. In both plots OF and UF stand for underflow and overflow respectively, where all events that pass the fiducial requirements but have a dijet invariant mass below 50 GeV are included in the underflow bin and all those above 150 GeV in the overflow bin. This allows for events to migrate into and out of the main analysis bins.

The corrections derived from this response matrix for the first iteration of the unfolding are shown in Figure 8.2a. These corrections are taken directly from simulation and are subsequently updated using the unfolded results for the following iterations. The values in the matrix shown give the probability that an event reconstructed in a given bin came from the respective truth bin. For bins associated with a reconstructed dijet invariant mass above 75 GeV, the off-diagonal term corresponding to the truth bin immediately below is relatively large. This is due to the falling distribution in the dijet invariant mass, meaning that the number of events migrating across the boundary make up a greater proportion of the bin they enter than the one they came from. An alternative normalisation for this matrix is shown in Figure 8.2b, where the values in this matrix give the probability that an event with a dijet invariant mass in a given truth bin are reconstructed in the respective mass bin. This is the response matrix. From this it can be seen that the probability of being reconstructed in a bin given the event started in the equivalent truth bin is found to be greater than 70% in the three main analysis bins.

The number of iterations for the procedure must also be chosen. Typically two or three iterations are sufficient to accurately recover the true distribution [120]. Large numbers of iterations are undesirable as the resulting distribution shows large fluctuations around the true values with associated large uncertainties [119]. It is expected that the corrections required here are small and that the prior distribution used will reasonably approximate the true underlying distribution of the data. For these reasons, two iterations are chosen as the nominal approach. Larger numbers of iterations are also trialled with the changes in values of the unfolded asymmetry taken as a systematic. The asymmetry values before unfolding and after unfolding with $n = 2, 3$ and 4 iterations are shown in Figure 8.3.

Having unfolded the yields separately for forward and backward events, the asymmetries are then determined to be

$$A_{50-75}^{\text{unfolded}} = (0.25 \pm 0.13 \pm 0.03)\%, \tag{8.2}$$

$$A_{75-105}^{\text{unfolded}} = (0.7 \pm 0.4 \pm 0.2)\%, \tag{8.3}$$

$$A_{105-150}^{\text{unfolded}} = (0.1 \pm 1.0 \pm 0.3)\%, \tag{8.4}$$

where the systematic uncertainties are taken as the difference between the values measured when using two and four iterations. Note that the systematic uncertainties given here only represent those associated with this unfolding procedure. The systematic uncertainties found from the fit procedure have not been propagated through at this stage, and instead, this will be carried out in Chapter 9. It can be seen by comparing these results with the raw asymmetries presented at the end of Chapter 7 that the unfolding procedure has only a small effect on the central values of the asymmetry. The shift, along with any shift seen by varying the number of iterations, is smaller than the statistical uncertainty on the values, as is also shown in Figure 8.3.

It is also possible to unfold distributions in simulated events and compare the result to the true underlying distribution. This is used here to further check that the number of iterations selected gives reasonable results. The sample used to generate the initial prior and response matrix cannot exactly match the distribution to be unfolded, else one iteration alone will give perfect agreement. Instead, half of the simulated events are taken and reweighted such that the distribution of the transverse momentum of the leading jet matches that seen in data. The other half of simulated events are not reweighted and are used to generate the response matrix and prior distribution. The results for this unfolding check are shown in Figure 8.4 for $n = 2, 3$ and 4 iterations. The difference between the yields from two and four iterations is again taken as a systematic, which is shown on the

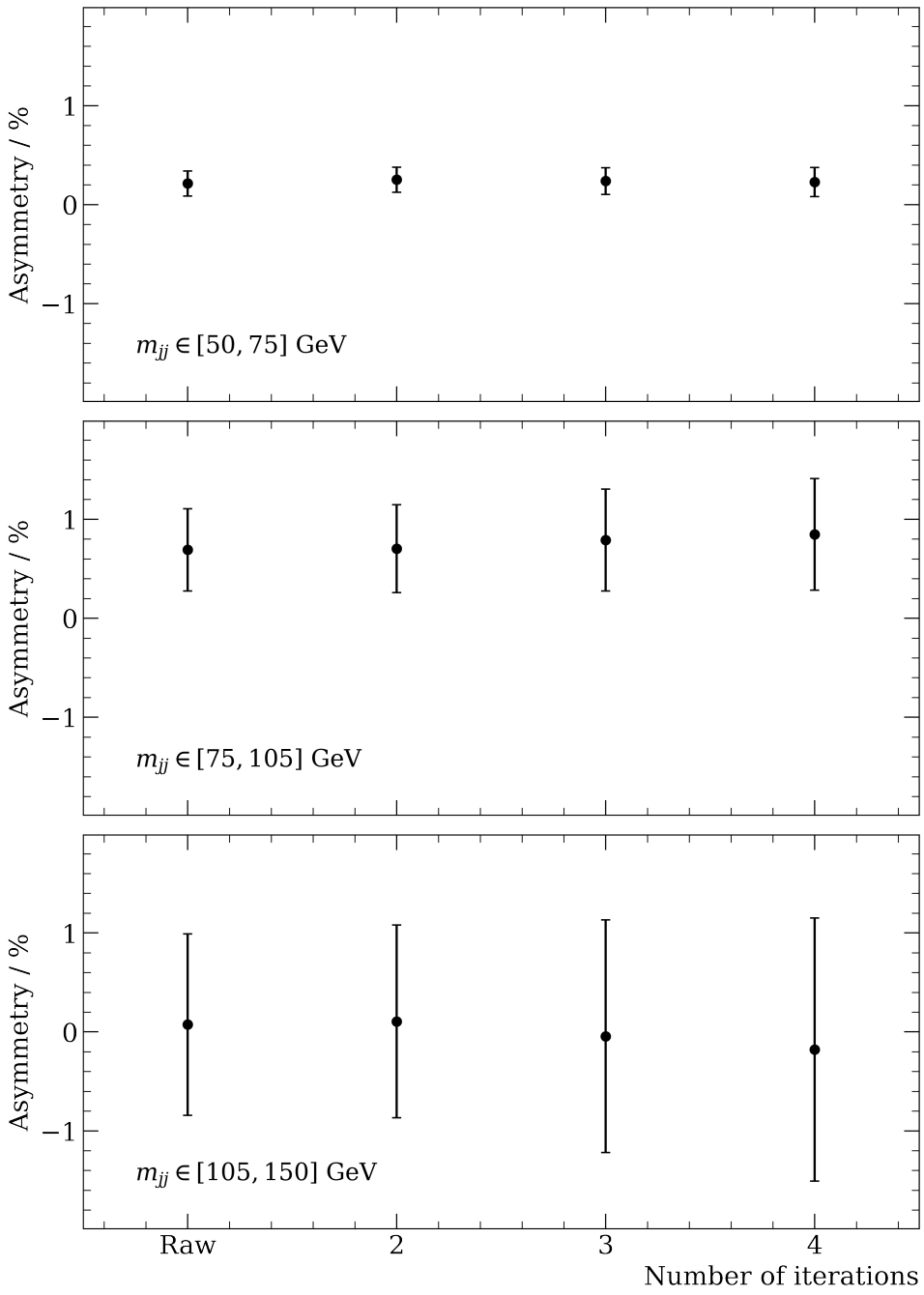

Figure 8.3: The value of the unfolded asymmetry as a function of the number of iterations used. Values of the asymmetry are shown for the three analysis bins of the dijet invariant mass. The result before unfolding is also shown and labelled 'Raw'. The uncertainty shown is the statistical uncertainty associated with the finite size of the data sample. It is clear that the central value of the asymmetry is relatively insensitive to the number of iterations used when compared to the size of the statistical uncertainty shown.

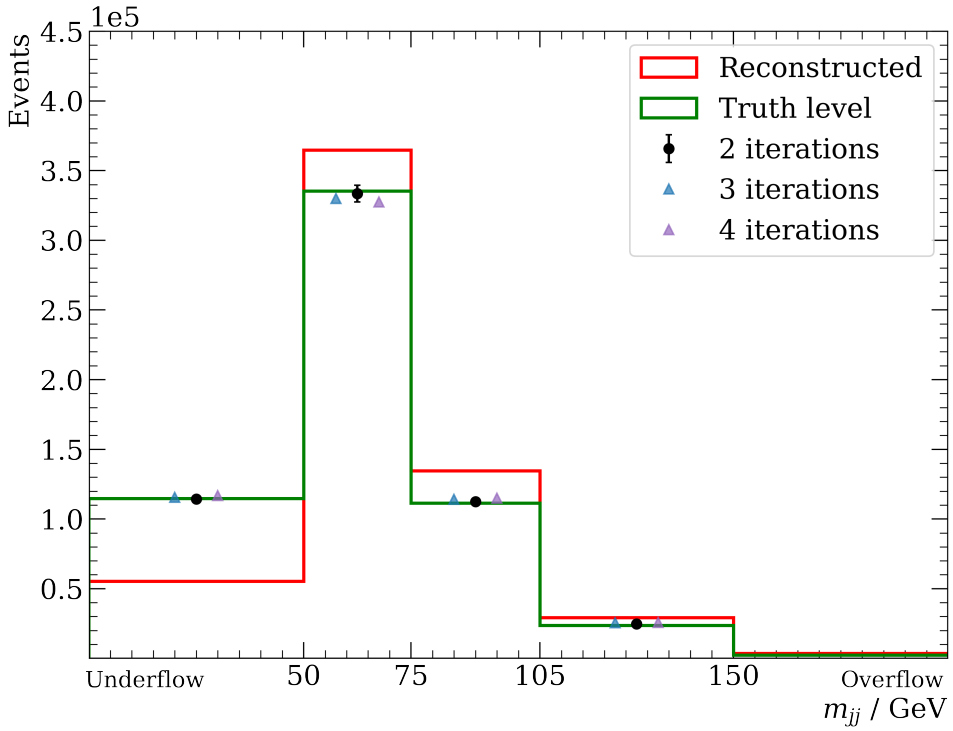

Figure 8.4: To check the unfolding procedure described in the main body of the text, simulated events are unfolded using $n = 2, 3$ and 4 iterations. Two iterations is the nominal value used in the analysis and a systematic uncertainty is calculated as the difference between this result and that using four iterations. This is the uncertainty shown on the results for $n = 2$. Statistical uncertainties are too small to be visible. There is good agreement between the unfolded distribution and the true underlying distribution.

plot. With this, there is agreement at the level of 1σ between the true distribution and the unfolded distribution after two iterations[2].

8.2 Mistagging from qTAG

Tagging the charge of the jets is crucial for determining whether an event is categorised as forward or backward. The procedure used is laid out in Section 5.3 and is referred to as qTAG. Mistagging the charge leads to the incorrect forward-backward assignment which has the effect of diluting the measured value of the asymmetry. It is crucial to correct for

[2]The unfolded results after two iterations are also seen to have the best agreement with the true distribution when compared with a larger number of iterations. This further motivates $n = 2$ as the number of iterations to use for the nominal result.

this effect. The correction factor is defined by

$$D = 1 - 2\omega, \tag{8.5}$$

where ω is the rate of incorrectly tagging events, and the correction factor D is referred to as the dilution factor. The measured value of the asymmetry is corrected by dividing by the dilution factor as shown in Eq. (5.3). As discussed in Section 5.3, the main causes of event mistagging are: $b \to c \to \mu X$ type decays, where the muon identified now originates from a semi-leptonic c decay; and B-meson oscillations. The dilution factor is estimated in the three bins of dijet invariant mass, using simulated events. These dilution factors are shown in Figure 8.5. The values in the three bins of invariant mass are

$$D_{50-75} = 0.5600 \pm 0.0008, \tag{8.6}$$

$$D_{75-105} = 0.5494 \pm 0.0009, \tag{8.7}$$

$$D_{105-150} = 0.5535 \pm 0.0012, \tag{8.8}$$

where the uncertainties given here are statistical. These dilution factors can be compared with the theoretical performance calculated in Section 5.3.1, where the dilution factor is predicted to be around $D^{\text{theoretical}} \approx 54\%$. This gives confidence that the charge tagging procedure is behaving as expected. Applying this dilution correction gives the final unfolded values of the asymmetry as

$$A_{50-75} = (0.45 \pm 0.23)\%, \tag{8.9}$$

$$A_{75-105} = (1.3 \pm 0.8)\%, \tag{8.10}$$

$$A_{105-150} = (0.2 \pm 1.8)\%, \tag{8.11}$$

where the uncertainties given here are statistical. A data-driven validation of the qTAG performance, and the assignment of a systematic uncertainty to cover mismodelling of this in simulation, is discussed in Chapter 9.

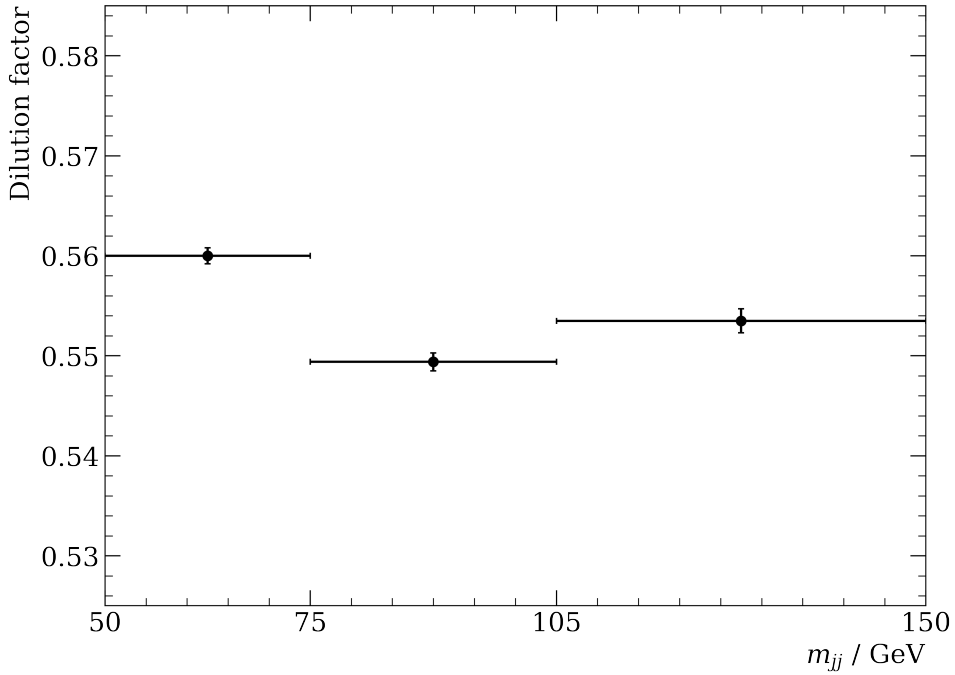

Figure 8.5: The dilution factor in bins of the dijet invariant mass as measured in simulated events.

Chapter 9

Systematic uncertainties and cross-checks

The statistical uncertainty on the asymmetry due to the limited number of events in data has already been accounted for and is given with the central values at the end of Chapter 8. It is also necessary to evaluate the systematic uncertainties on this measurement. Systematic uncertainties quantify possible variations in the measurement originating from other sources. These are difficult to evaluate precisely. If there was a clear systematic bias present then it would be possible to correct for this explicitly. Instead, the size of these uncertainties is estimated and provides an uncertainty band on the final measurement. Some of the sources of systematic uncertainty have already been discussed in the previous chapters, however, there are further sources that have not yet been covered. Each source of systematic uncertainty and methods for determining their size will be discussed in turn, and these will be fully propagated through the analysis to give an uncertainty on the final unfolded value of the asymmetry. Systematic uncertainties here are all presented as absolute shifts in the values of the unfolded, dilution-corrected asymmetry, measured in percent.

In addition, results from cross-checks to the measurement are discussed. These deal with making alternative but related measurements to confirm that the results are as expected and give further confidence in the analysis.

9.1 Systematic uncertainties

9.1.1 Fit template shapes

The fits carried out in Chapter 7 use the $\log(\chi^2_{\rm FD})$ variable. As discussed in the same section, this variable requires correction due to mismodelling effects. The size of this correction is derived using $B^+ \to J/\psi K^+$ decays for the nominal result. To account for the uncertainty on this correction an alternative value is also derived using $B^0 \to J/\psi K^*$ decays and the fits are repeated using this value. A third correction is also obtained by finding the difference between the two derived corrections and adding this difference to the nominal correction. In Chapter 7 the systematic uncertainty on the raw asymmetry is determined. Here, the fit results for the alternative corrections are propagated through the remaining steps in the analysis, being unfolded and corrected for mistagging effects. The difference in the values of the asymmetry is then calculated between the nominal correction and the two alternative approaches. The maximal differences are taken as the systematic uncertainty in each bin. These are evaluated as

$$\sigma^{\chi^2_{\rm FD}}_{50-75} = 0.21\%, \tag{9.1}$$

$$\sigma^{\chi^2_{\rm FD}}_{75-105} = 0.30\%, \tag{9.2}$$

$$\sigma^{\chi^2_{\rm FD}}_{105-150} = 1.1\%. \tag{9.3}$$

These systematic uncertainties are relatively important as the shape of the beauty templates has a significant impact on the yields returned from the template fits.

The corrections to the template shapes are derived in beauty decays but are also applied to charm-jets in the analysis, as it is expected that the mismodelling effects are similar [107,108]. A systematic uncertainty is also assigned to account for this assumption. The analysis is repeated, maintaining the nominal correction for beauty-jets, but applying no correction to the simulated charm-jet events. The difference between the unfolded, dilution-corrected asymmetry for this and the nominal result is taken as the systematic uncertainty. This is evaluated as

$$\sigma^{c\text{-cor}}_{50-75} = 0.0029\%, \tag{9.4}$$

$$\sigma^{c\text{-cor}}_{75-105} = 0.08\%, \tag{9.5}$$

$$\sigma^{c\text{-cor}}_{105-150} = 0.14\%. \tag{9.6}$$

Whilst removing the correction entirely is a very conservative approach, the calculated systematic uncertainty remains small in comparison to other sources. This is due to the fact that the charm component of the data sample is small.

9.1.2 Unfolding procedure

The unfolding procedure used in this analysis to correct for migrations in the dijet invariant mass is Iterative Bayesian Unfolding [119,120]. The number of iterations must be chosen in this method and a systematic uncertainty is assigned by varying the number used and calculating the difference in the corrected asymmetry. The systematic uncertainties are evaluated in Chapter 8 and reported for results before correcting for the qTAG dilution. After applying the dilution correction, the final systematic uncertainties from this source are

$$\sigma^{\text{unf.}}_{50-75} = 0.05\%, \tag{9.7}$$

$$\sigma^{\text{unf.}}_{75-105} = 0.28\%, \tag{9.8}$$

$$\sigma^{\text{unf.}}_{105-150} = 0.5\%. \tag{9.9}$$

9.1.3 Dilution factor

The dilution factor used to correct for misidentifying the charge of the jets with qTAG is measured using simulated $b\bar{b}$-dijet events (as opposed to being directly measured in data). This approach has two advantages. First, the simulation is free from backgrounds which might bias a measurement of the correction. Second, the simulation contains truth information on the charge of the quarks originating the jets which allows for the dilution factor to be found directly and simply. However, it is important to confirm that the value of the correction derived in simulation accurately matches the dilution in data. Despite the lack of truth information for the jets in data, it is also possible to derive a correction using events where both jets are tagged. For events in which both jets are correctly tagged, true $b\bar{b}$ events would receive an opposite sign tag for each jet. Therefore, if both jets are tagged with the same sign of muon then one jet must be correctly tagged and one must be incorrectly tagged. If jets are tagged with an opposite sign then either both are correctly tagged, or both are incorrectly tagged. The rate at which jets are incorrectly tagged, ω, can thus be calculated from the number of events with jets tagged with the same sign

(N_{SS}) and opposite sign (N_{OS}):

$$N_{\mathrm{OS}} = \left((1-\omega)^2 + \omega^2\right) N_{\mathrm{tagged}}, \tag{9.10}$$

$$N_{\mathrm{SS}} = 2\,\omega\,(1-\omega)\,N_{\mathrm{tagged}}. \tag{9.11}$$

These equations are quadratic in ω and, hence, give two solutions. Both of these solutions lie in the range $0-1$ and they are related by $\omega_1 = 1-\omega_2$. In effect, the correct tag rate and the incorrect tag rate are solutions to the equations. For this reason, the solution that satisfies $\omega < 0.5$ is taken when identifying the incorrect tag rate. Requiring that both jets are tagged in the event also reduces the backgrounds as the efficiency of qTAG for charm-jets is lower, as discussed in Section 5.3. After requiring both jets in the event to be tagged by qTAG, there remain 46,715 events in data, around 3% of the single tagged sample. It is important to note that this value does not correspond to the single jet efficiency for qTAG. The probability of tagging both jets in the event given that one is already tagged can be calculated using Bayes' theorem

$$P(N_T = 2 \,|\, N_T \geq 1) = \frac{P(N_T \geq 1 \,|\, N_T = 2) P(N_T = 2)}{P(N_T \geq 1)}, \tag{9.12}$$

$$= \frac{\varepsilon}{2-\varepsilon}, \tag{9.13}$$

where N_T is the number of tagged jets and ε is the efficiency of qTAG tagging a single jet. For small values of ε, this leaves the probability of tagging both jets in the event given one is already tagged as roughly half of the single tag efficiency. To further reduce charm backgrounds, a requirement is placed on the corrected mass of the jets, defined in Eq. (6.3). The distributions of this variable in b- and c-jets are shown in Figure 6.4. To reduce charm backgrounds a requirement is placed on each jet in the event that the corrected mass, $M_{\mathrm{cor}} > 2.2\,\mathrm{GeV}$. In simulation this requirement is measured to be around 55% efficient for $b\bar{b}$-dijets, 2% for $c\bar{c}$-dijets events and 10% efficient for bc-dijets[1]. Having this requirement along with both jets satisfying qTAG will reduce any charm backgrounds to sufficiently low levels that they can be neglected while studying systematic effects.

The dilution factor is then measured in doubly tagged data and simulated $b\bar{b}$-dijet events using the same method, referred to as 'reco'. The values for the dilution factors in data and simulated events that contain two tagged jets are shown in Figure 9.1. The

[1]This is not used as a selection requirement in the main analysis due to the difficulty in accurately modelling it in simulation. The use of such requirements in future updates to this measurement is discussed in Chapter 10.

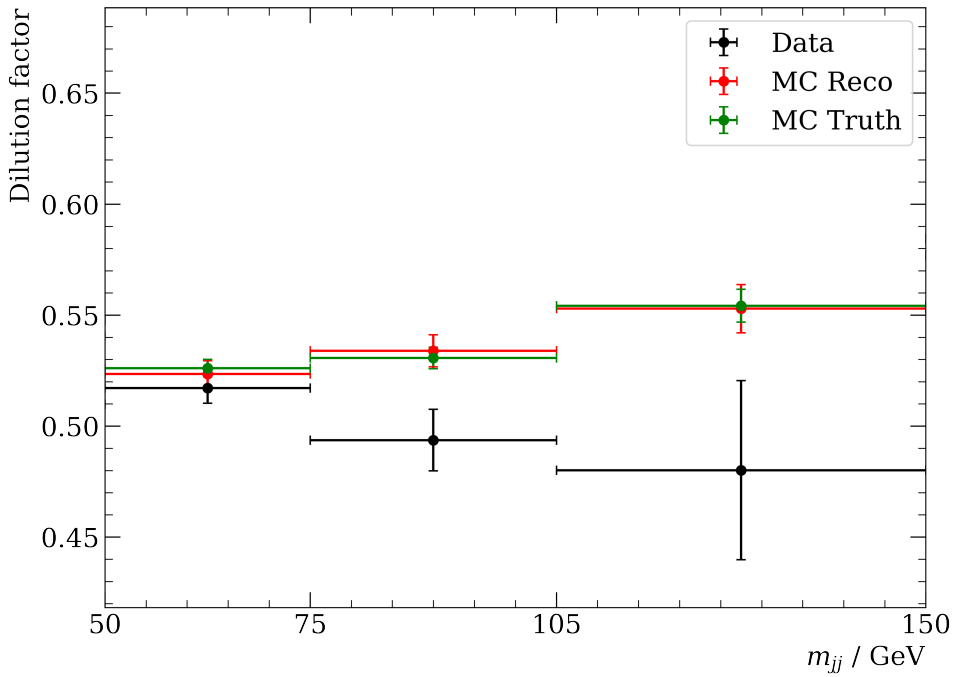

Figure 9.1: Dilution factor in data and simulated events with two qTAG-tagged jets. The dilution factor is measured in simulated events using two methods. The 'Reco' method uses only the reconstructed variables and the dilution factor is evaluated by examining the number of events where both jets carry the same sign of tag. This approach is fully discussed in the main body of the text and is also used for the data events. The 'Truth' method compares the charge of the tag assigned by qTAG with truth variables relating to the charge of b quark that initiated the jet.

value of the dilution factor determined using truth values in the same selected, simulated events with two tagged jets is also measured. There is good agreement between the values obtained using the truth and reco methods which gives confidence in the final values used. The measured dilution factors are listed in Table 9.1. The ratio between the values of the dilution factor measured in data and simulation can then be calculated. These ratios can be used to scale the dilution factor measured in the full set of simulated events. The dilution correction is then repeated with the rescaled values and the difference in the final value of the asymmetry is taken as a systematic uncertainty. The ratios by which the dilution factor in each bin of the dijet invariant mass is scaled are then

$$R_{50-75} = 0.988, \tag{9.14}$$

$$R_{75-105} = 0.925, \tag{9.15}$$

$$R_{105-150} = 0.868. \tag{9.16}$$

Table 9.1: Dilution factors measured in double qTAG events in simulated events (MC) and data. The uncertainties given in this table are statistical. In simulation, two methods are used to evaluate the dilution factor. In the 'MC reco' method the number of dijets with the same flavour tags is used to calculate the mistag rate. In 'MC truth' the true jet flavour is used to evaluate the mistag rate. These methods are in good agreement with each other.

Dijet mass / GeV	Data	MC reco	MC truth
50-75	0.517 ± 0.007	0.523 ± 0.006	0.526 ± 0.004
75-105	0.493 ± 0.014	0.534 ± 0.007	0.531 ± 0.005
105-150	0.48 ± 0.04	0.553 ± 0.011	0.554 ± 0.007

Using these values gives a final systematic uncertainty on the values of the asymmetry as

$$\sigma^{\text{dil.}}_{50-75} = 0.006\%, \tag{9.17}$$

$$\sigma^{\text{dil.}}_{75-105} = 0.10\%, \tag{9.18}$$

$$\sigma^{\text{dil.}}_{105-150} = 0.029\%. \tag{9.19}$$

These are relatively small when compared to other systematic uncertainties in the analysis. The impact in the highest mass bin is small as the asymmetry itself is small in this bin and the dilution correction is multiplicative to the value of the asymmetry. This means that there can be uncertainties on the dilution factor of $\mathcal{O}(10\%)$ without producing a significant systematic uncertainty on the asymmetry. In addition, the dilution factors derived in data could be used directly in the measurement without a significant shift in the final results.

9.1.4 Limited simulated events

The finite number of simulated events gives rise to statistical uncertainties, impacting both the templates used in the fits as well as the response matrix in the unfolding procedure. To account for this, pseudoexperiments are carried out with the templates and the response matrix fluctuated within their uncertainty bands. There are 500 pseudoexperiments carried out and the standard deviation on the final unfolded, corrected value of the asymmetry in these pseudoexperiments is taken as the systematic uncertainty from

this source. These uncertainties are evaluated at

$$\sigma^{\mathrm{MC}}_{50-75} = 0.11\%, \tag{9.20}$$

$$\sigma^{\mathrm{MC}}_{75-105} = 0.34\%, \tag{9.21}$$

$$\sigma^{\mathrm{MC}}_{105-150} = 0.7\%. \tag{9.22}$$

The size of these systematic uncertainties is mostly dominated by the contribution of the uncertainty from the template fits.

9.1.5 Jet energy scale and resolution

As discussed in Section 3.5, there is an uncertainty associated with the jet energy scale (JES) and resolution (JER) in simulated events.

Jets first have their reconstructed energy rescaled to account for detector effects, but this rescaling has an associated uncertainty [83] which must be accounted for in this measurement. The scale and resolution of the jet energy are also investigated explicitly in this analysis to account for mismodelling by using the p_{T} balance in Z+jet and $b\bar{b}$-dijet events. These are both discussed in Section 3.5 and again an associated uncertainty with the required scaling and smearing is derived.

There are then three sources of uncertainty associated with the jet energy:

- Uncertainty on the scale factor derived from simulation to correct for detector effects. This is referred to here as the "JES (simulation)" uncertainty. Events are produced with the uncertainty on the rescaling provided following work in Ref. [83].
- Uncertainty on the required scaling to correct for mismodelling of the JES in simulation. This is investigated explicitly in this analysis by comparing Z+jet events in data and simulation. This is referred to here as the "JES (data-MC)" uncertainty.
- Uncertainty on the required energy smearing to correct for mismodelling of the JER in simulation. This is investigated explicitly in this analysis by comparing $b\bar{b}$-dijet events in data and simulation. This is referred to as the JER uncertainty.

Each of these is considered in turn. For uncertainties on the JES the jet energies in simulation are rescaled to 1σ above and 1σ below their nominal values using the uncertainty on the scale factor. The full analysis is then repeated. To propagate the uncertainty on the

Table 9.2: Systematic uncertainties on the final unfolded, dilution corrected values of the asymmetry associated with the jet energy scale (JES) and resolution (JER). The JES has two sources of uncertainty. The first originates from the scaling that is applied to correct for detector effects derived from simulation: JES (simulation). The second comes from the rescaling that is applied in the Z+jet studies carried out in this thesis to correct for mismodelling: JES (data-MC). The level of precision quoted for the individual sources is set by the dominant systematic for a given dijet invariant mass bin.

	Uncertainty source		
Dijet mass / GeV	JER / %	JES (simulation) / %	JES (data-MC) / %
50-75	0.08	0.02	0.02
75-105	0.22	0.11	0.10
105-150	0.7	0.3	0.4

JER, the smearing is turned off and again the full analysis is repeated[2]. The differences in the final values of the asymmetry between these results and the nominal are then taken as the systematic uncertainty for each of the three sources. The systematic uncertainty associated with each of these is listed in Table 9.2. From this, it can be seen that of the three, the uncertainty associated with the JER is dominant[3]. Taken individually, none of these sources of uncertainty are either notably large or small with regard to other sources of systematic uncertainty in this analysis. They are added in quadrature here to give an overall uncertainty associated with the jet energy and scale in each of the three dijet invariant mass bins, which are calculated as

$$\sigma^{\text{Jet-energy}}_{50-75} = 0.08\%, \tag{9.23}$$

$$\sigma^{\text{Jet-energy}}_{75-105} = 0.26\%, \tag{9.24}$$

$$\sigma^{\text{Jet-energy}}_{105-150} = 0.9\%. \tag{9.25}$$

[2]This is again a conservative approach. In future iterations of this measurement, it may be desirable to derive corrections using alternative parameterisations and use these to calculate a less conservative systematic uncertainty.

[3]It is unsurprising that the uncertainty associated with the JER is dominant relative to the uncertainties associated with the JES, due to the conservative approach taken in estimating this effect. Nevertheless, this uncertainty remains small compared to the statistical uncertainty.

9.1.6 Pseudorapidity resolution

There is good resolution on measurements of the direction of jets in the LHCb detector. Uncertainties on the direction of the jet in $\eta - \phi$ space are typically around $\Delta R \sim 0.05$, with $R = \sqrt{\eta^2 + \phi^2}$ [83]. However, it is still possible that mismeasurement of the angles leads to the incorrect assignment of a forward or backward event. To account for this the fraction of events in simulated $b\bar{b}$-dijet events in which the true pseudorapidity of the two jets are in the opposite order to the reconstructed values is measured. Similarly to the mistagging of events caused by qTAG, this leads to a dilution of the true asymmetry by

$$D^{\eta} = 1 - 2r^{\eta}, \tag{9.26}$$

where r^{η} is the rate at which events are mistagged as forward/backward caused by the incorrect measurement of the jet direction. These mistagging rates are measured in bins of the dijet invariant mass in simulated events, giving

$$r^{\eta}_{50-75} = (1.962 \pm 0.014)\%, \tag{9.27}$$
$$r^{\eta}_{75-105} = (1.239 \pm 0.013)\%, \tag{9.28}$$
$$r^{\eta}_{105-150} = (1.013 \pm 0.015)\%. \tag{9.29}$$

The difference in the final result when correcting for this effect is taken as a systematic uncertainty. This is found to be negligible:

$$\sigma^{\eta}_{50-75} = 0.018\%, \tag{9.30}$$
$$\sigma^{\eta}_{75-105} = 0.033\%, \tag{9.31}$$
$$\sigma^{\eta}_{105-150} = 0.004\%. \tag{9.32}$$

9.1.7 Non-uniform efficiency effects

The next systematic uncertainty that is considered is due to the detector having a non-uniform efficiency for selecting and reconstructing dijet events. This is expected to be small. Events can have a different efficiency through the analysis based on their kinematics. This can create a bias if the selection sculpts the phase-space of events to favour regions where the asymmetry is larger or smaller. To account for this, the efficiency of events through the trigger, the SV tagger and qTAG is measured using simulated events

in bins of the difference of pseudorapidity between the two jets. The overall efficiencies can be seen in Figure 9.2. The shape of this efficiency is driven by a small reduction in the SV tagging efficiency near the detector edge, coupled with the trigger efficiency falling with pseudorapidity. These values are then used to weight the data events by the inverse of the efficiency, as a proxy for a measurement if the efficiency was flat. The full analysis is then repeated, and the difference in the final values of the asymmetry is again taken as a systematic. As expected, the change in the final results is negligible, with the systematic uncertainty evaluated as[4]

$$\sigma^{\text{eff.}}_{50-75} = 0.04\%, \tag{9.33}$$

$$\sigma^{\text{eff.}}_{75-105} = 0.004\%, \tag{9.34}$$

$$\sigma^{\text{eff.}}_{105-150} = 0.07\%. \tag{9.35}$$

9.1.8 Light-jet contamination

Light-jet contributions to the data sample are expected to be negligible, as discussed in Section 7.1. To account for potential contamination from light-jets, for example in bq-dijets, a systematic uncertainty is assigned. For this, it is assumed that the light-jet contribution is at the level of 1% of the total event yield. The theoretical calculations on the dijet composition yield gave an expected light-jet contribution at <1% and this is also the highest upper bound from the single jet fits carried out in Section 7.1. Assuming that these events have no asymmetry and hence are equally split between forward and backward events, and (as a worst case) that all of these events are assigned to the $b\bar{b}$-dijet yield in the fit, the impact on the final asymmetry can be estimated. The presence of this additional background would then reduce the measured value of the asymmetry by a multiplicative factor of 0.99, giving the systematic uncertainty in each bin as

$$\sigma^{\text{light-jet}}_{50-75} = 0.004\%, \tag{9.36}$$

$$\sigma^{\text{light-jet}}_{75-105} = 0.013\%, \tag{9.37}$$

$$\sigma^{\text{light-jet}}_{105-150} = 0.0020\%. \tag{9.38}$$

[4]It should also be noted that instead of a log likelihood fit, a χ^2 fit is carried out here as this correctly handles the weighted data events and uncertainty propagation. This systematic then additionally includes an uncertainty associated with using a different fit procedure. As these uncertainties are found to be small (when compared to others within the analysis) this additional check is helpful as it gives confidence in the procedure used.

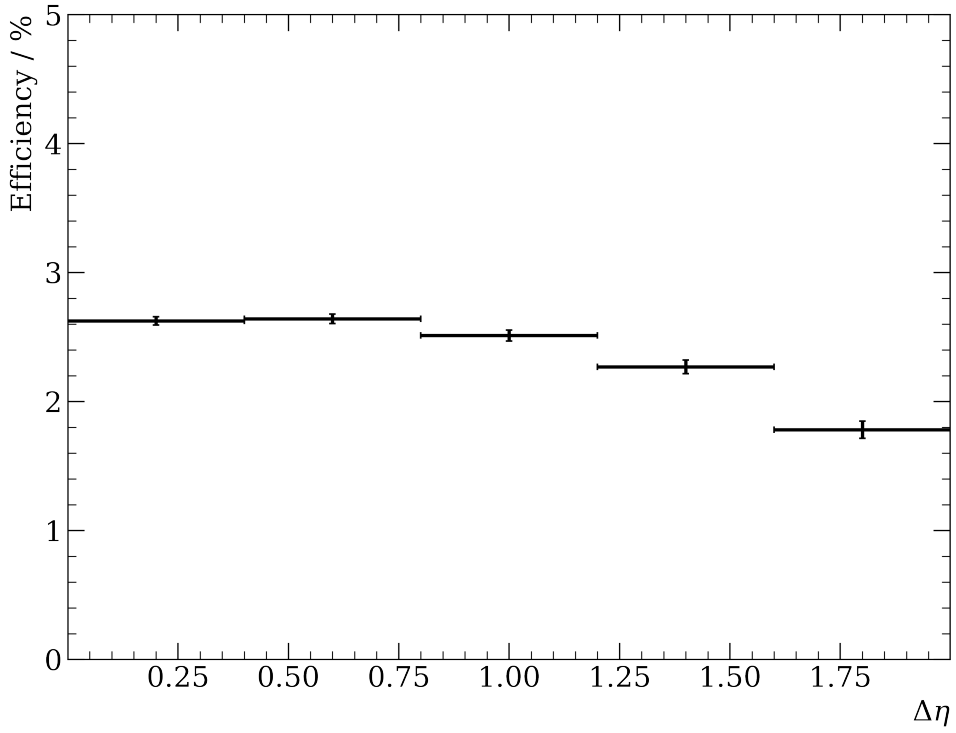

Figure 9.2: The efficiency of the trigger, SV tagger and qTAG combined for events in bins of the difference in jet pseudorapidities, $\Delta\eta$. These efficiencies are measured in simulated events.

These uncertainties are clearly negligible in comparison to the other uncertainties evaluated. Indeed, even if the light-jet contribution is assumed to be much larger, or the bq-dijets had an asymmetry, this would still yield uncertainties that are negligible in comparison to the dominant systematic uncertainties.

9.1.9 Overall systematic uncertainty

The systematic uncertainties from all of the sources are added in quadrature to yield a total systematic uncertainty on the final values of the asymmetry. The individual values and total systematic are listed in Table 9.3. This gives the final measured values of the asymmetry as

$$A_{50-75} = (0.45 \pm 0.23 \pm 0.26)\%, \tag{9.39}$$

$$A_{75-105} = (1.3 \pm 0.8 \pm 0.6)\%, \tag{9.40}$$

$$A_{105-150} = (0.2 \pm 1.8 \pm 1.6)\%, \tag{9.41}$$

Table 9.3: Summary of systematic uncertainties. All values are given as absolute shifts in the asymmetry measured in %. The level of precision quoted for the individual sources is set by the dominant systematic for a given dijet invariant mass bin. For uncertainties which are quoted as zero a more precise value can be found in the main body of the text. The statistical uncertainties in each bin are also shown to aid comparison.

	Dijet invariant mass / GeV		
Uncertainty source	50-75	75-105	105-150
χ^2_{FD} correction	0.21	0.30	1.1
Charm correction	0.00	0.08	0.1
MC Stats	0.11	0.34	0.7
Unfolding procedure	0.05	0.28	0.5
Jet energy resolution	0.08	0.22	0.7
Jet energy scale (data-MC)	0.02	0.10	0.4
Jet energy scale (simulation)	0.02	0.11	0.3
Pseudorapidity resolution	0.02	0.03	0.0
Non-uniform efficiency	0.04	0.00	0.1
Dilution factor	0.01	0.10	0.0
Light-jet contribution	0.00	0.01	0.0
Total systematic uncertainty	0.26	0.60	1.6
Statistical uncertainty	0.23	0.81	1.8

where the first uncertainty is statistical and the second systematic. From this, it can be seen that the systematic and statistical uncertainties are reasonably balanced. It is therefore important, before proceeding to future measurements using the full LHCb Run II dataset, to consider how the systematic uncertainty could be reduced. The largest uncertainty in all three analysis bins originates from the template shapes used in the analysis (this includes 'MC Stats', with the uncertainty in the fit providing the dominant contribution to this source). Therefore, understanding in better detail the detector degradation and associated mismodelling may help in providing corrections that reduce this source of systematic uncertainty. Other approaches are also discussed when considering potential improvements to the analysis in Chapter 10.

9.2 Cross-checks

A number of cross-checks have been carried out to ensure that the results are robust and to give confidence in the methods used. These will be detailed in turn.

9.2.1 Fit procedure

The fits used in the nominal analysis rely on the use of the convolution method to produce background samples, as detailed in Section 7.2. A cross-check is carried out here that does not rely on this method by instead fitting one-dimensional templates to the data. The two jets in the event are ordered by their corrected mass, M_{cor}, where j_1 has a larger M_{cor}. The distributions of this variable in b- and c-jets are shown in Figure 6.4. For a randomly selected b- and c-jet, the probability that the b-jet has a higher corrected mass is greater than 80% as measured in simulation. Therefore the ordering means that for mixed flavour bc-dijets, the b-jet is more likely to be found labelled as j_1 and the c-jet as j_0. Fits are then performed to the $\log(\chi^2_{\text{FD}})$ of j_0, using one-dimensional templates with 20 bins. By fitting the j_0 distributions, bc-dijets are more likely to be represented by the c-jet in the event and hence be included in the c-jet yield. The b-jet yield from the fit will mostly come from $b\bar{b}$-dijets with very small contamination from bc-dijets where the b-jet has a lower corrected mass than the c-jet. An example of the fit results is shown in Figure 9.3 for forward events in the dijet invariant mass range $m_{jj} \in [75, 105]\,\text{GeV}$.

The raw asymmetry values are evaluated for this cross-check directly from the b-jet

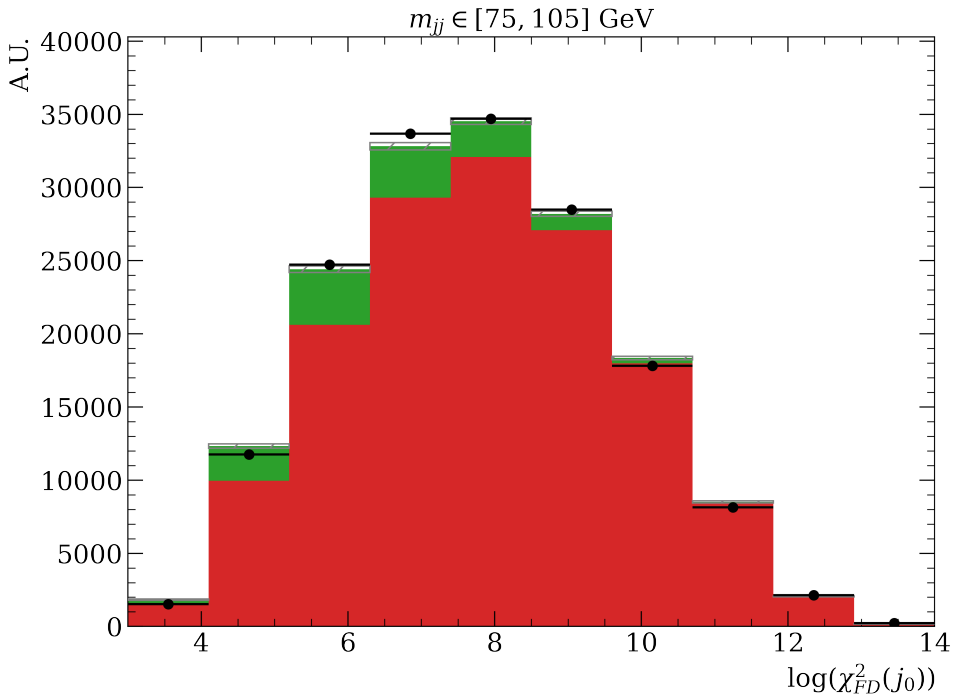

Figure 9.3: Fit result from the one-dimensional fit for forward events in the dijet invariant mass range $m_{jj} \in [75, 105]$ GeV. This one-dimensional fit is performed as a cross-check to the main analysis. The uncertainty in the fit contributions shown here is the statistical uncertainty of the fit itself and does not include any systematic contributions from variations to the template shapes.

yields as

$$A^{1D}_{50-75} = (0.23 \pm 0.18)\% \quad [0.21 \pm 0.13\%], \tag{9.42}$$

$$A^{1D}_{75-105} = (0.70 \pm 0.30)\% \quad [0.7 \pm 0.4\%], \tag{9.43}$$

$$A^{1D}_{105-150} = (0.1 \pm 0.7)\% \quad [0.1 \pm 0.9\%], \tag{9.44}$$

where the uncertainty shown is just the statistical uncertainty associated with the fit, and no unfolding or dilution correction has been applied (this leaves the uncertainties and central values roughly half the size of the final result). The raw asymmetry results from the nominal analysis are shown in square brackets, with just the statistical uncertainty shown here. The central values are in very good agreement, demonstrating the reliability of the two-dimensional fit using the convolution method. Although the statistical uncertainties in this one-dimensional fit are slightly lower than the nominal method in the higher two mass bins, this method is not used for the nominal analysis owing to the lack of a robust

treatment of the bc-dijet component. It would be important to accurately determine the level of bc-dijet contamination in the b-jet yield from the one-dimensional fit. This would rely on good modelling of the corrected mass which, as with other SV-tagging-related observables, is not the case owing to detector degradation not being accurately modelled in simulation. Were this method treated as the nominal approach, a significant additional systematic uncertainty would likely be needed to account for the modelling of the corrected mass variable.

An alternative cross-check is also carried out to check the use of the convolution method for the $c\bar{c}$-dijet sample in the two-dimensional fits used in the nominal analysis. A two-dimensional template is built in the same manner as the $b\bar{b}$-dijet template, constructed directly from the simulated samples without using a convolution. To increase the number of events, and hence decrease statistical fluctuations in the template, extra simulated events are used which are produced under 2017 and 2018 run conditions. These events are otherwise identical to the $c\bar{c}$ simulated sample set out in Section 5.1. A template is constructed from a combination of the 2016, 2017 and 2018 simulated samples. The fit used in the nominal analysis is repeated with the convolved $c\bar{c}$ template replaced using this alternative template. The raw values of the asymmetry after the fit are then extracted as

$$A_{50-75}^{c\bar{c}-no-conv.} = (0.17 \pm 0.12)\%, \tag{9.45}$$

$$A_{75-105}^{c\bar{c}-no-conv.} = (0.7 \pm 0.5)\%, \tag{9.46}$$

$$A_{105-150}^{c\bar{c}-no-conv.} = (-0.1 \pm 0.9)\%, \tag{9.47}$$

these again show little variation from the nominal results found in Chapter 7. The combination of these cross-checks gives confidence that the use of the convolution samples does not significantly bias the measured results.

9.2.2 Magnet polarity

The LHCb magnet bends particles of opposite charge in opposite directions and so they interact with different parts of the detector. Systematic effects could then come from detection or trigger efficiencies varying across the phase space of the detector. The polarity of the magnet is reversed periodically during data taking, with the integrated luminosities achieved in each configuration agreeing at the 2% level. This is important as many systematic effects will cancel out when averaging the results from the two different polarities.

The asymmetry is measured separately for the two different magnet polarities to ensure there are no large systematic biases that could affect the combined measurement. The data sample is split into events where the magnet had the polarity up (MU) and those where it was down (MD). The fits are repeated and the difference in the raw asymmetries is evaluated at

$$\Delta A_{50-75}^{\text{MU-MD}} = (0.7 \pm 0.4)\%, \tag{9.48}$$

$$\Delta A_{75-105}^{\text{MU-MD}} = (1.1 \pm 0.7)\%, \tag{9.49}$$

$$\Delta A_{105-150}^{\text{MU-MD}} = (0.0 \pm 1.8)\%, \tag{9.50}$$

where the uncertainty reported is statistical. There is no deviation from zero larger than 2σ in any single bin. It is potentially of note that the two lowest mass bins do exhibit tensions at the level of 1.5–2σ. However, a χ^2 test for the compatibility of the three results with zero returns a p-value of 10%. Additionally, the systematic uncertainties have not been re-evaluated for these split measurements and hence are not used here to quantify the agreement between the samples. This cross-check therefore potentially underestimates uncertainties on the asymmetry difference between the two polarities, and as such the potential discrepancies estimated here are a worst-case scenario. Despite this, there is no significant evidence here that there is a systematic difference between the datasets collected with the different magnetic field configurations.

9.2.3 Muon transverse momentum

The nominal analysis places no requirements on the muons selected during the qTAG procedure other than that the muon is the highest p_{T}, displaced track in the jet, and that the `L0MuonEW` trigger is responsible for selecting the event[5]. Although previous studies have suggested that muons with a lower transverse momentum should be well behaved, asymmetries in the detection and trigger reconstruction efficiencies are most prevalent at lower momentum [100]. For this reason, the values of the asymmetry are checked by making the measurement with a selection requirement placed on the transverse momentum of muons identified in the qTAG procedure of $p_{\text{T}}(\mu) > 6\,\text{GeV}$. The value is chosen to match the nominal selection placed with the `L0MuonEW` requirement and removes only

[5]The `L0MuonEW` requires the muon to have a p_{T} greater than 6 GeV but has poor momentum resolution, at the level of 25% which allows some lower momentum muons to pass through. See Chapter 3 for more details.

around 4% of events in simulation. This requirement also restricts the measurement to tracks which are bent less by the magnetic field, allowing a further test of the effects set out in the MU vs MD cross-check above. After this requirement is placed on the data and simulated samples the fits are repeated yielding the raw asymmetry values

$$A^{\mu>6}_{50-75} = (0.28 \pm 0.19)\%, \tag{9.51}$$

$$A^{\mu>6}_{75-105} = (0.7 \pm 0.4)\%, \tag{9.52}$$

$$A^{\mu>6}_{105-150} = (0.1 \pm 0.9)\%. \tag{9.53}$$

These can be compared with the raw asymmetries from the nominal analysis, which are found in Chapter 7. This measurement is made on a subset of the data used in the nominal analysis, so the results here are expected to be correlated to those in the nominal approach. Therefore consistency is checked using the formula set out in Ref. [122], with the squared uncertainty on the difference between the two results set by the difference in quadrature of the uncertainties on the two results individually: $\sigma^2 = \sigma^2_{\text{nom}} - \sigma^2_{\mu>6}$. The difference in the central values is then evaluated as

$$\Delta A^{\mu>6}_{50-75} = (-0.07 \pm 0.14)\%, \tag{9.54}$$

$$\Delta A^{\mu>6}_{75-105} = (0.00 \pm 0.15)\%, \tag{9.55}$$

$$\Delta A^{\mu>6}_{105-150} = (-0.05 \pm 0.13)\%. \tag{9.56}$$

It can be seen that in all bins the agreement is better than 1σ.

9.2.4 Trigger requirements

To give further confidence that the potential discrepancies seen in the magnet polarity cross-check carried out above are not of concern a final cross-check is carried out. It is known that muon triggers can exhibit asymmetries between μ^+ and μ^- for a given polarity, although these largely cancel when averaging across the two polarities [100]. Due to this potential asymmetry in the muon trigger used, the analysis is once more repeated using the `L0Hadron` trigger line instead of the `L0MuonEW` trigger line. The values in the raw asymmetries when using this trigger line show agreement between the MU and MD samples at the level of 1σ or better in all bins of the analysis. The results using the combined MU and MD samples are also compared for use of each of the two L0 trigger lines. To do this, the data sample is split in two, with one half requiring the `L0MuonEW`

trigger line and the other half requiring the `L0Hadron` trigger line. This split is done to remove statistical correlations between the two subsamples. The fits are carried out on these samples and the raw asymmetries are extracted from the measured $b\bar{b}$-dijet yields. The difference in the asymmetries for use of the two trigger lines is evaluated as

$$\Delta A^{\text{trigger}}_{50-75} = (-0.1 \pm 0.5)\%, \tag{9.57}$$

$$\Delta A^{\text{trigger}}_{75-105} = (0.7 \pm 0.8)\%, \tag{9.58}$$

$$\Delta A^{\text{trigger}}_{105-150} = (1.7 \pm 1.6)\%, \tag{9.59}$$

which again show agreement at around 1σ or better[6]. These results suggest that even if there were a systematic bias in the muon triggers responsible for the small discrepancies in the magnet polarity cross-check, then these effects cancel when combining the two polarities.

Despite this good performance, the `L0Hadron` trigger line is not used in the nominal analysis since there is a global event cut in the trigger, which may bias the event selection as discussed in Section 5.2. Furthermore, the `L0MuonEW` trigger has a higher efficiency for this analysis. Around 75% of events pass the `L0MuonEW` trigger line, with only around 55% of events passing the `L0Hadron` trigger.

[6]Whilst the uncertainty on this cross-check is larger than the uncertainty on the nominal measurement, this agreement is still encouraging and shows the measurement is robust.

Chapter 10

Discussion

10.1 Results

The charge asymmetry in beauty-dijet production is measured in three bins of the dijet invariant mass. To carry out this measurement data and simulated samples are prepared. Next, a method for distinguishing b- and $\bar{b}$-jets in the event is developed and tested utilising the muons from semileptonic decays. The simulated samples are then calibrated to data to correct for mismodelling. Fits to the data are carried out to extract the $b\bar{b}$-dijet yields and these yields are unfolded for migrations in the dijet invariant mass. This allows for a measurement of the asymmetry to be made from the unfolded yields which is then corrected for the mistagging of the qTAG method. A number of systematic uncertainties are evaluated. The final measured results are

$$A_{50-75} = (0.45 \pm 0.23 \pm 0.26)\%, \tag{10.1}$$

$$A_{75-105} = (1.3 \pm 0.8 \pm 0.6)\%, \tag{10.2}$$

$$A_{105-150} = (0.2 \pm 1.8 \pm 1.6)\%, \tag{10.3}$$

where the first uncertainty is statistical and the second systematic. Theoretical predictions have been made at NLO in the SM [92]. The authors of this prediction are thanked for providing their results in a form that allows bins to be combined. Here, the predictions

using a nominal renormalisation and factorisation scale of $\mu_0 = E_{\bar{T},b}$ are taken. This gives

$$A^{\mathrm{SM}}_{60-75} = (0.43 \pm 0.21)\%, \tag{10.4}$$

$$A^{\mathrm{SM}}_{75-105} = (1.43 \pm 0.19)\%, \tag{10.5}$$

$$A^{\mathrm{SM}}_{105-150} = (0.95 \pm 0.29)\%, \tag{10.6}$$

where the uncertainty here is evaluated using scale variations as discussed in Chapter 4, but where PDF uncertainties are not evaluated [92]. It should be noted that the first mass bin is narrower for the theoretical predictions. The authors of Ref. [92] confirmed in discussion that this is because the asymmetry in the $50 - 60\,\mathrm{GeV}$ region is small but hard to formally predict at NLO. Both the measured results and the theoretical predictions are shown in Figure 10.1. The results are seen to agree to within 1σ with theoretical predictions in all three bins. An additional set of theoretical predictions are also included in this plot from Ref. [96]. These have a lower level of formal accuracy, but predictions at lower order are easier to make in the low mass region. The inclusion of this set of predictions therefore gives coverage for the lowest values of invariant dijet mass considered in this analysis. Again, the agreement between these theoretical predictions and measured values is within 1σ in each bin.

This constitutes the first measurement of the charge asymmetry in beauty-dijet production in proton-proton collisions at a centre-of-mass energy of $13\,\mathrm{TeV}$. As with the previous measurement carried out using the LHCb detector, it shows no tension with the Standard Model predictions. This measurement constitutes a test of SM at a precision of below the percent level. This level of precision can be challenging for measurements involving QCD contributions but is made possible by the nature of asymmetry measurements, which allow for many uncertainties in the measurement and theoretical predictions to cancel.

10.2 Future related measurements

10.2.1 Full Run II measurement

The measurement provided in this thesis uses only the 2016 dataset gathered with the LHCb detector, corresponding to an integrated luminosity of $1.7\,\mathrm{fb}^{-1}$. The clear next step for this measurement would be to use the full Run II data, which would roughly

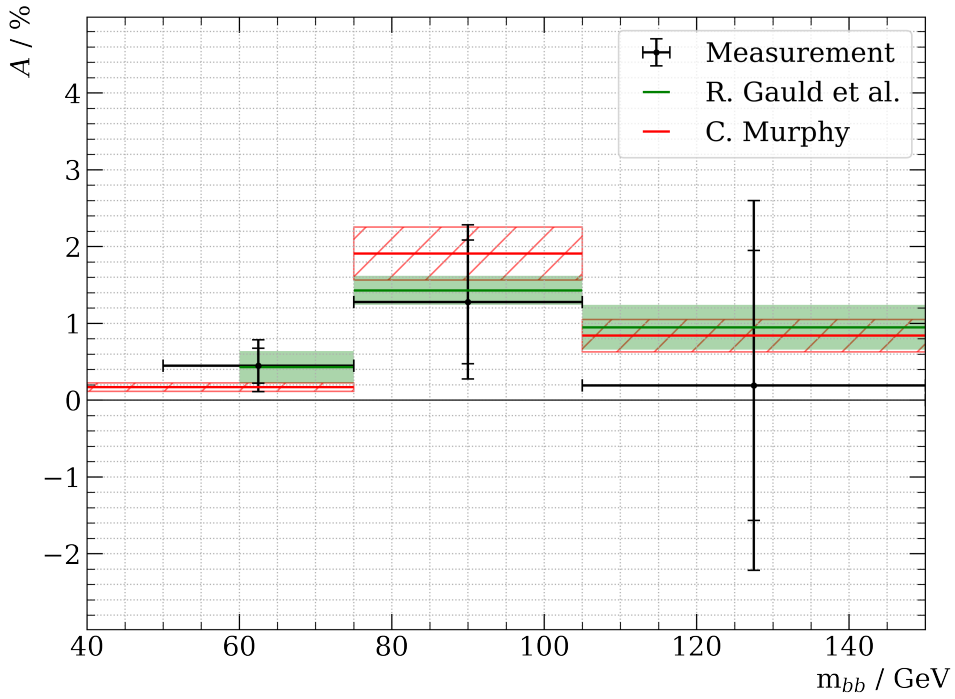

Figure 10.1: The measured values of the charge asymmetry in $b\bar{b}$-dijets compared with theoretical predictions. The outer error bars for data are the statistical and systematic uncertainties added in quadrature, the inner bars show the size of the statistical uncertainty alone. Theoretical predictions shown in green are from Ref. [92], with uncertainties shown as shaded green areas. Theoretical predictions shown in red are from Ref. [96], the uncertainties are shown as red hatched areas. The predictions shown in green are carried out using higher-order calculations than those shown in red.

correspond to tripling the size of the dataset. Whilst this would reduce the statistical uncertainties, additional work would also need to be carried out to reduce the considerable systematic uncertainties seen during this analysis.

Most notable are the uncertainties around the template shapes pertaining to the correction that must be applied to the χ^2_{FD} distribution. It would benefit this proposed update to have a more detailed understanding of the mismodelling seen in the simulation. Ideally, not only would the χ^2_{FD} distribution be corrected, but also other properties of the SV which have also been observed to exhibit significant mismodelling. A full correction of these variables, coupled with a retraining or an updating of the LHCb flavour tagging BDTs could reduce these uncertainties to a level where the overall systematic uncertainty would not be dominant when carrying out a measurement on the full Run II dataset. A larger number of simulated events would also be required, but this is more simple to overcome.

If it is determined that fully correcting the simulation and retraining the BDTs is too demanding of LHCb resources, an alternative approach may also be possible. Backgrounds could be removed from the data sample with a selection requirement on the M_{cor} of the two jets. So far in this thesis, the M_{cor} variable has only been used for the study of systematic uncertainties associated with the qTAG method, to validate the jet p_{T} smearing factor in a beauty-enriched region of the phase space and in a cross-check of the fit results where jets were ordered based on their M_{cor}. It is not used in the nominal analysis presented as there are significant issues with the modelling of the variable in simulation, coupled with the fact that this variable is difficult to accurately calibrate. One reason for this difficulty is that the variable is heavily dependent upon the decay channel examined. For example, in the fully reconstructed b decays that are used to correct the $\log(\chi^2_{\text{FD}})$ in this analysis, the corrected mass will be exactly the mass of the B meson with some width associated with the detector resolution. This is because the corrected mass is accounting for missing momentum in the system and there is none in a fully reconstructed decay. To correct the M_{cor} variable for all jets in the analysis would therefore be a large undertaking, with significant time and resources required. Furthermore, it remains unclear if such a correction would be possible without introducing significant systematic uncertainties into the analysis. However, if such corrections are found to be possible, and allow accurate modelling of the M_{cor} variable, a significant portion of the backgrounds to this analysis could be removed by requiring each jet to have $M_{\text{cor}} > T$, for some threshold value T. The level of background remaining could then be determined by fitting the $M_{\text{cor}} < T$ region and scaling the result to the signal region. This would remove the need for a full fit of the signal region. Assuming the signal region has a high purity, the uncertainty on the level of background would remain relatively small in comparison to the uncertainties seen in the present analysis and hence could help bring systematic uncertainties under greater control. This route may be preferred to a full correction of multiple variables and should be considered in the future alongside the approach given in this thesis.

To provide a quick test of the potential sensitivity of such a method, the asymmetry is measured in a $b\bar{b}$-dijet enriched region in the samples used throughout this analysis. For this example a M_{cor} threshold value of $T = 2.2\,\text{GeV}$ is chosen. This is done without considering background charm-jets and no calibration of the M_{cor} variable has been

carried out. The raw asymmetry of events in this region is measured as

$$A_{50-75}^{M_{\rm cor}} = (0.15 \pm 0.15)\%, \tag{10.7}$$

$$A_{75-105}^{M_{\rm cor}} = (0.30 \pm 0.25)\%, \tag{10.8}$$

$$A_{105-150}^{M_{\rm cor}} = (-0.2 \pm 0.6)\%, \tag{10.9}$$

where the uncertainties provided are statistical. It can be seen by comparing these to the raw asymmetries given in Chapter 7 that there is a slight reduction of the statistical uncertainty in the upper two mass bins using this new method. Using the fits, these uncertainties are at the level of 0.4% and 0.9% for the 75–105 GeV and 105–150 GeV bins respectively. This route therefore shows promise but would require careful correction of the $M_{\rm cor}$ variable, as well as evaluation of the systematic uncertainties associated with using such a method. For now, these numbers serve as one final cross-check on the nominal results presented in this analysis.

Work must also be continued on the theoretical predictions, with the aim of making these even more precise. This is true particularly in the lowest bin of the invariant dijet mass, where the NLO predictions have an uncertainty comparable with the uncertainty provided in this measurement.

10.2.2 Asymmetry in charm-dijet production

A logical extension to the measurement in $b\bar{b}$-dijet production would be to measure the same property in $c\bar{c}$-dijet production. This would allow for a ratio of the two asymmetries to be taken, leading to the cancellation of theoretical uncertainties in the SM prediction [92]. If carried out on the full Run II data, and with additional work on the theoretical predictions, the ratio of the two asymmetries has the potential to provide constraints on the $Zb\bar{b}$ and $Zc\bar{c}$ couplings at a level competitive with LEP measurements [92]. This measurement will unfortunately not be able to follow exactly the same prescription as is used here for the $b\bar{b}$ asymmetry measurement. The charge tagging method used, qTAG, preferentially selects $b\bar{b}$-dijet events, leaving too few $c\bar{c}$-dijet events to make a useful measurement of the charm asymmetry. Instead, some other charge tagging method must be developed. One option would be to use $D^0 \to K^-\pi^+$ decays, which has a 4% branching fraction. This decay can be fully reconstructed and the mass of the $K\pi$ system can be used to select the charm-jets. The charge of the K (or π) can also be used to identify the charge of the c quark. To remove $B \to DX \to K\pi X$ type backgrounds, selection

requirements could be used to remove events where the reconstructed D meson does not point back to the PV, as this would indicate that it had been produced from some secondary decay. The analysis of the charge asymmetry in charm production could then proceed in a similar way to the analysis presented here.

Alternatively, if the simulation of $b\bar{b}$-dijets and $c\bar{c}$-dijets are well corrected, it may be possible to train a machine learning classifier to tag the jets for their flavour with a result that gives more separation between the flavours than is currently available using the LHCb flavour BDTs[1], and with a high efficiency for both beauty and charm-jets. If sufficient separation is achieved it may be possible to use a charge tagging method that need not favourably select beauty or charm-jets, instead relying on the performance of this new classifier to accurately extract the relevant $c\bar{c}$ yield. In this case, it may be possible to relax the criteria used in qTAG and use this relaxed qTAG to tag the jets. For example, currently, the displaced track with the highest transverse momentum is selected and if this is a muon then the event is qTAG. Instead, it may be possible to locate the muon with the highest transverse momentum in the jet and tag the jet if this muon has a significant impact parameter. The level required would not be achievable easily. This would require significant work into potential classifying algorithms and careful calibration of multiple input variables for use in these algorithms to ensure that these variables are accurately modelled in simulated events.

There may also be alternative charge tagging options which have not been explored here. Whichever charge tagging method is chosen, it will clearly be necessary to investigate and optimise this method to ensure the best measurement of the charge asymmetry in charm-jets.

[1]These are currently also not used in this analysis due to the mismodelling of SV properties discussed in this thesis.

Chapter 11

Conclusion

This thesis presents a measurement of the charge asymmetry in beauty-dijet production at LHCb, with all stages of the analysis carried out by the author. The measurement uses the 2016 dataset of proton-proton collisions gathered by the detector, corresponding to an integrated luminosity of $1.7\,\text{fb}^{-1}$. This is the first measurement of the charge asymmetry in beauty-dijet production in proton-proton collisions at a centre-of-mass energy of 13 TeV. The measurement is carried out in three bins of the dijet invariant mass, in order to examine the varying contributions from QCD and electroweak production methods. The results are shown to agree with state-of-the-art NLO theoretical predictions in each of the dijet invariant mass bins that are considered. This represents a test of the SM at a precision at the percent level, made possible by the cancellation of uncertainties in both the theoretical predictions and the measurement of the asymmetry.

Furthermore, this thesis lays the groundwork for a future measurement of the charge asymmetry in beauty-dijet production using the full Run II dataset. It is shown in this analysis that improvements to the modelling of the detector response provided in simulation are necessary in order to reduce systematic uncertainties when a larger dataset is used. Suggestions for alternative analysis strategies have also been laid out in this thesis that may help circumvent mismodelling issues.

A measurement of the charge asymmetry in charm-dijet production may also be possible with the LHCb detector. New methods to tag the charge of charm-jets will likely need to be developed to achieve such a goal, with initial suggestions for this provided in this thesis. If accurate determinations of both the charm- and beauty-dijet production asymmetries are achieved, it may be possible to achieve a sensitivity to the $Zb\bar{b}$ and $Zc\bar{c}$ couplings at a level competitive with that seen in LEP studies. This would also require

further improvement in the theoretical understanding.

New data is now being taken at the LHC during Run III. This will lead to new results that will begin to emerge from a variety of physics programs. What we will learn over the coming years remains to be seen. However, one thing is clear: rigorous testing of the SM and searches for NP will continue across multiple different experiments around the world. It is only with such studies that a greater understanding of the nature of our universe will be achieved.

Bibliography

[1] A. Salam and J. C. Ward, *Electromagnetic and weak interactions*, Phys. Lett. **13** (1964) 168.

[2] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19** (1967) 1264.

[3] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22** (1961) 579.

[4] J. Goldstone, A. Salam, and S. Weinberg, *Broken Symmetries*, Phys. Rev. **127** (1962) 965.

[5] M. Gell-Mann, *A Schematic Model of Baryons and Mesons*, Phys. Lett. **8** (1964) 214.

[6] G. Zweig, in *An SU(3) model for strong interaction symmetry and its breaking. Version 2*, D. B. Lichtenberg and S. P. Rosen, eds., pp. 22–101, 1964.

[7] M. Kobayashi and T. Maskawa, *CP Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49** (1973) 652.

[8] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964) 508.

[9] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13** (1964) 585.

[10] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321.

[11] ATLAS collaboration, 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. B **716** (2012) 1, `arXiv:1207.7214`.

[12] CMS collaboration, S. Chatrchyan *et al.*, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, Phys. Lett. B **716** (2012) 30, arXiv:1207.7235.

[13] Particle Data Group, R. L. Workman *et al.*, *Review of Particle Physics*, PTEP **2022** (2022) 083C01.

[14] M. Thomson, *Modern Particle Physics*, Cambridge University Press, 2013.

[15] A. Buckley, C. White, and M. White, *Practical Collider Physics*, IOP, 2021.

[16] F. Mandl and G. Shaw, *Quantum field theory*, John Wiley & Sons, 2010.

[17] Super-Kamiokande collaboration, Y. Fukuda *et al.*, *Evidence for oscillation of atmospheric neutrinos*, Phys. Rev. Lett. **81** (1998) 1562, arXiv:hep-ex/9807003.

[18] KATRIN collaboration, M. Aker *et al.*, *Direct neutrino-mass measurement with sub-electronvolt sensitivity*, Nature Phys. **18** (2022) 160, arXiv:2105.08533.

[19] CDF collaboration, T. Aaltonen *et al.*, *High-precision measurement of the W boson mass with the CDF II detector*, Science **376** (2022) 170.

[20] T. Aoyama, T. Kinoshita, and M. Nio, *Theory of the Anomalous Magnetic Moment of the Electron*, Atoms **7** (2019) 28.

[21] D. Hanneke, S. Fogwell, and G. Gabrielse, *New Measurement of the Electron Magnetic Moment and the Fine Structure Constant*, Phys. Rev. Lett. **100** (2008) 120801, arXiv:0801.1134.

[22] C.-S. Wu *et al.*, *Experimental test of parity conservation in beta decay*, Physical review **105** (1957) 1413.

[23] ALEPH collaboration, DELPHI collaboration, L3 collaboration, OPAL collaboration, SLD collaboration, LEP Electroweak Working Group, SLD Electroweak Group, SLD Heavy Flavour Group, S. Schael *et al.*, *Precision electroweak measurements on the Z resonance*, Phys. Rept. **427** (2006) 257, arXiv:hep-ex/0509008.

[24] N. Cabibbo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10** (1963) 531.

[25] K2K collaboration, M. H. Ahn *et al.*, *Measurement of Neutrino Oscillation by the K2K Experiment*, Phys. Rev. D **74** (2006) 072003, arXiv:hep-ex/0606032.

[26] Super-Kamiokande collaboration, J. Hosaka *et al.*, *Solar neutrino measurements in super-Kamiokande-I*, Phys. Rev. D **73** (2006) 112001, arXiv:hep-ex/0508053.

[27] Z. Maki, M. Nakagawa, and S. Sakata, *Remarks on the unified model of elementary particles*, Prog. Theor. Phys. **28** (1962) 870.

[28] B. Pontecorvo, *Neutrino Experiments and the Problem of Conservation of Leptonic Charge*, Zh. Eksp. Teor. Fiz. **53** (1967) 1717.

[29] J. Schechter and J. W. F. Valle, *Neutrinoless Double beta Decay in SU(2) x U(1) Theories*, Phys. Rev. D **25** (1982) 2951.

[30] D. J. Gross and F. Wilczek, *Ultraviolet Behavior of Nonabelian Gauge Theories*, Phys. Rev. Lett. **30** (1973) 1343.

[31] H. D. Politzer, *Reliable Perturbative Results for Strong Interactions?*, Phys. Rev. Lett. **30** (1973) 1346.

[32] J. D. Bjorken, *Current algebra at small distances*, Conf. Proc. C **670717** (1967) 55.

[33] NNPDF collaboration, R. D. Ball *et al.*, *Parton distributions from high-precision collider data*, Eur. Phys. J. C **77** (2017) 663, arXiv:1706.00428.

[34] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, Nucl. Phys. B **406** (1993) 187.

[35] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, *Better jet clustering algorithms*, JHEP **08** (1997) 001, arXiv:hep-ph/9707323.

[36] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv:0802.1189.

[37] G. P. Salam and G. Soyez, *A practical seedless infrared-safe cone jet algorithm*, JHEP **2007** (2007) 086.

[38] T. Sjöstrand, S. Mrenna, and P. Skands, *A brief introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, arXiv:0710.3820.

[39] B. Andersson, G. Gustafson, and B. Soderberg, *A General Model for Jet Fragmentation*, Z. Phys. C **20** (1983) 317.

[40] T. Sjostrand, *Jet Fragmentation of Nearby Partons*, Nucl. Phys. B **248** (1984) 469.

[41] Geant4 collaboration, J. Allison *et al.*, *Geant4 developments and applications*, IEEE Trans. Nucl. Sci. **53** (2006) 270.

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

[43] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, JHEP **11** (2004) 040, arXiv:hep-ph/0409146.

[44] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, JHEP **11** (2007) 070, arXiv:0709.2092.

[45] S. Alioli, P. Nason, C. Oleari, and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06** (2010) 043, arXiv:1002.2581.

[46] J. Alwall *et al.*, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv:1405.0301.

[47] K. Hamilton, P. Nason, and G. Zanderighi, *MINLO: Multi-Scale Improved NLO*, JHEP **10** (2012) 155, arXiv:1206.3572.

[48] O. S. Bruning *et al.*, *LHC Design Report Vol.1: The LHC Main Ring*, CERN Yellow Reports: Monographs (2004) .

[49] L. Evans and P. Bryant, *LHC machine*, JINST **3** (2008) S08001.

[50] ATLAS collaboration, G. Aad *et al.*, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003.

[51] CMS collaboration, S. Chatrchyan *et al.*, *The CMS Experiment at the CERN LHC*, JINST **3** (2008) S08004.

[52] ALICE collaboration, K. Aamodt *et al.*, *The ALICE experiment at the CERN LHC*, JINST **3** (2008) S08002.

[53] LHCb collaboration, A. A. Alves Jr. *et al.*, *The LHCb detector at the LHC*, JINST **3** (2008) S08005.

[54] E. Lopienska, *The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022*, accessible at: https://cds.cern.ch/record/2800984, 2022.

[55] C. Elsässer, *Production angle plots of $b\bar{b}$*, accessible at: https://lhcb.web.cern.ch/lhcb/speakersbureau/html/bb_ProductionAngles.html. Accessed: 21/09/22.

[56] LHCb collaboration, R. Aaij *et al.*, *Measurement of the b-quark production cross-section in 7 and 13 TeV pp collisions*, Phys. Rev. Lett. **118** (2017) 052002, Erratum ibid. **119** (2017) 169901, arXiv:1612.05140.

[57] R. Aaij *et al.*, *Performance of the LHCb Vertex Locator*, JINST **9** (2014) P09007, arXiv:1405.7808.

[58] LHCb collaboration, R. Aaij *et al.*, *Measurement of the track reconstruction efficiency at LHCb*, JINST **10** (2015) P02007, arXiv:1408.1251.

[59] M. Adinolfi *et al.*, *Performance of the LHCb RICH detector at the LHC*, Eur. Phys. J. **C73** (2013) 2431, arXiv:1211.6759.

[60] T. Humair, *Testing lepton universality in penguin decays of beauty mesons using the LHCb detector*, CERN-THESIS-2019-044 (2019).

[61] A. A. Alves Jr. *et al.*, *Performance of the LHCb muon system*, JINST **8** (2013) P02022, arXiv:1211.1346.

[62] F. Archilli *et al.*, *Performance of the muon identification at LHCb*, JINST **8** (2013) P10020, arXiv:1306.0249.

[63] T. Head, *The LHCb trigger system*, JINST **9** (2014) C09015.

[64] R. Aaij *et al.*, *Performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC*, JINST **14** (2019) P04013, arXiv:1812.10790.

[65] O. Callot, *FastVelo, a fast and efficient pattern recognition package for the Velo*, LHCb-PUB-2011-001, CERN-LHCb-PUB-2011-001 (2011).

[66] E. E. Bowen, B. Storaci, and M. Tresch, *VeloTT tracking for LHCb Run II*, LHCb-PUB-2015-024, CERN-LHCb-PUB-2015-024, LHCb-INT-2014-040 (2016).

[67] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Journal of Basic Engineering **82** (1960) 35.

[68] R. Fruhwirth, *Application of Kalman filtering to track and vertex fitting*, Nucl. Instrum. Meth. A **262** (1987) 444.

[69] M. De Cian, S. Farry, P. Seyfert, and S. Stahl, *Fast neural-net based fake track rejection in the LHCb reconstruction*, LHCb-PUB-2017-011, CERN-LHCb-PUB-2017-011 (2017).

[70] J. P. Dufey *et al.*, *The LHCb trigger and data acquisition system*, IEEE Trans. Nucl. Sci. **47** (2000) 86.

[71] G. Haefeli *et al.*, *The LHCb DAQ interface board TELL1*, Nucl. Instrum. Meth. A **560** (2006) 494.

[72] C. Bozzi, *LHCb Computing Resource usage in 2021* , CERN, Geneva, 2022.

[73] S. Benson, V. V. Gligorov, M. A. Vesterinen, and M. Williams, *The LHCb Turbo Stream*, J. Phys. Conf. Ser. **664** (2015) 082004.

[74] J. T. Boyd, *LHC Run-2 and Future Prospects*, in *2019 European School of High-Energy Physics*, 2020, `arXiv:2001.04370`.

[75] LHCb collaboration, R. Aaij *et al.*, *Precision luminosity measurements at LHCb*, JINST **9** (2014) P12005, `arXiv:1410.0149`.

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

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

[78] P. Golonka and Z. Was, *PHOTOS Monte Carlo: A precision tool for QED corrections in Z and W decays*, Eur. Phys. J. **C45** (2006) 97, `arXiv:hep-ph/0506026`.

[79] D. J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instrum. Meth. **A462** (2001) 152.

[80] G. Corti *et al.*, *Software for the LHCb experiment*, IEEE Trans. Nucl. Sci. **53** (2006) 1323.

[81] LHCb collaboration, R. Aaij *et al.*, *Study of forward Z +jet production in pp collisions at $\sqrt{s}$ =7 TeV*, JHEP **01** (2014) 033, arXiv:1310.8197.

[82] W. Barter *et al.*, *Jets reconstruction and performances at LHCb*, LHCb-INT-2012-015, LHCb Internal Documentation, CERN, 2012.

[83] W. Barter, *Z boson and associated jet production at the LHCb experiment*, CERN-THESIS-2014-178 (2014).

[84] M. Kucharczyk, P. Morawski, and M. Witek, *Primary Vertex Reconstruction at LHCb*, LHCb-PUB-2014-044, 2014.

[85] P. J. Rousseeuw, F. R. Hampel, E. M. Ronchetti, and W. A. Stahel, *Robust statistics: the approach based on influence functions*, .

[86] LHCb collaboration, R. Aaij *et al.*, *LHCb detector performance*, Int. J. Mod. Phys. **A30** (2015) 1530022, arXiv:1412.6352.

[87] M. Cacciari and G. P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, Phys. Lett. B **641** (2006) 57, arXiv:hep-ph/0512210.

[88] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, Eur. Phys. J. C **72** (2012) 1896, arXiv:1111.6097.

[89] LHCb collaboration, R. Aaij *et al.*, *First observation of forward $Z \to b\bar{b}$ production in pp collisions at $\sqrt{s}$ =8 TeV*, Phys. Lett. **B776** (2018) 430, arXiv:1709.03458.

[90] LHCb collaboration, R. Aaij *et al.*, *Study of Z bosons produced in association with charm in the forward region*, Phys. Rev. Lett. **128** (2022) 082001, arXiv:2109.08084.

[91] LHCb collaboration, R. Aaij *et al.*, *Identification of beauty and charm quark jets at LHCb*, JINST **10** (2015) P06013, arXiv:1504.07670.

[92] R. Gauld, U. Haisch, and B. D. Pecjak, *Asymmetric heavy-quark hadroproduction at LHCb: Predictions and applications*, JHEP **03** (2019) 166, arXiv:1901.07573.

[93] CDF collaboration, T. A. Aaltonen *et al.*, *Measurement of the forward-backward asymmetry in low-mass bottom-quark pairs produced in proton-antiproton collisions*, Phys. Rev. D **93** (2016) 112003, arXiv:1601.06526.

[94] D0 collaboration, V. M. Abazov *et al.*, *Measurement of the Forward-Backward Asymmetry in the Production of $B^{\pm}$ Mesons in $p\bar{p}$ Collisions at $\sqrt{s}$ = 1.96 TeV*, Phys. Rev. Lett. **114** (2015) 051803, arXiv:1411.3021.

[95] LHCb collaboration, R. Aaij *et al.*, *First measurement of the charge asymmetry in beauty-quark pair production*, Phys. Rev. Lett. **113** (2014) 082003, arXiv:1406.4789.

[96] C. W. Murphy, *Bottom-Quark Forward-Backward and Charge Asymmetries at Hadron Colliders*, Phys. Rev. D **92** (2015) 054003, arXiv:1504.02493.

[97] B. Grinstein and C. W. Murphy, *Bottom-Quark Forward-Backward Asymmetry in the Standard Model and Beyond*, Phys. Rev. Lett. **111** (2013) 062003, arXiv:1302.6995, [Erratum: Phys.Rev.Lett. 112, 239901 (2014)].

[98] S. Ipek, *Light Axigluon Contributions to $b\bar{b}$ and $c\bar{c}$ Asymmetry and Constraints on Flavor Changing Axigluon Currents*, Phys. Rev. D **87** (2013) 116010, arXiv:1301.3990.

[99] J. Drobnak, J. F. Kamenik, and J. Zupan, *Flipping $t\bar{t}$ Asymmetries at the Tevatron and the LHC*, Phys. Rev. D **86** (2012) 054022, arXiv:1205.4721.

[100] L. Dufour, *High-precision measurements of charge asymmetries at LHCb*, CERN-THESIS-2019-281 (2019).

[101] V. V. Gligorov, *A single track HLT1 trigger*, LHCb-PUB-2011-003, CERN-LHCb-PUB-2011-003, LHCb-INT-2010-053 (2010).

[102] LHCb collaboration, R. Aaij *et al.*, *A new algorithm for identifying the flavour of B_s^0 mesons at LHCb*, JINST **11** (2016) P05010, arXiv:1602.07252.

[103] LHCb collaboration, R. Aaij *et al.*, *New algorithms for identifying the flavour of B^0 mesons using pions and protons*, Eur. Phys. J. **C77** (2017) 238, arXiv:1610.06019.

[104] LHCb collaboration, R. Aaij *et al.*, *B flavour tagging using charm decays at the LHCb experiment*, JINST **10** (2015) P10005, arXiv:1507.07892.

[105] LHCb collaboration, R. Aaij *et al.*, *Measurement of b-hadron fractions in 13 TeV pp collisions*, Phys. Rev. **D100** (2019) 031102(R), arXiv:1902.06794.

[106] LHCb collaboration, R. Aaij *et al.*, *Measurement of differential $b\bar{b}$ and $c\bar{c}$ dijet cross-sections in the forward region of pp collisions at $\sqrt{s}$ = 13 TeV*, JHEP **02** (2021) 023, arXiv:2010.09437.

[107] M. Alexander and V. Lisovskyi, *Run 1&2 performance: news and update on velo errors issue*, accessible at: https://indico.cern.ch/event/941777/contributions/3998522/attachments/2100355/3530998/20200910-R12PWG-summary.pdf. Talk at LHCb Week (presented: 10/09/20; accessed: 19/05/22).

[108] M. Alexander, *VELO errors: track refitting & recipe for analysis*, accessible at: https://indico.cern.ch/event/955205/contributions/4013664/attachments/2103361/3536759/20200916-VELOErrorsRefitting.pdf. Talk to Performance WG (presented: 16/09/20; accessed: 19/05/22).

[109] LHCb collaboration, R. Aaij *et al.*, *Measurement of the W boson mass*, JHEP **01** (2022) 036, arXiv:2109.01113.

[110] M. Santana Rangel, R. Ticse Torres, V. Coco, and C. Potterat, *Jet energy studies for Run I simulation*, LHCb-INT-2013-036, CERN-LHCb-INT-2013-036 (2013).

[111] LHCb collaboration, R. Aaij *et al.*, *Precision measurement of forward Z boson production in proton-proton collisions at $\sqrt{s}$ = 13 TeV*, JHEP **07** (2022) 026, arXiv:2112.07458.

[112] A. Affolder *et al.*, *Radiation damage in the LHCb vertex locator*, JINST **8** (2013) P08002, arXiv:1302.5259.

[113] LHCb collaboration, R. Aaij *et al.*, *Test of lepton universality in beauty-quark decays*, Nature Physics **18** (2022) 277, arXiv:2103.11769.

[114] LHCb collaboration, R. Aaij *et al.*, *Measurement of CP-averaged observables in the $B^0 \to K^{*0}\mu^+\mu^-$ decay*, Phys. Rev. Lett. **125** (2020) 011802, arXiv:2003.04831.

[115] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*, eConf **C0303241** (2003) MOLT007, arXiv:physics/0306116.

[116] ATLAS collaboration, G. Aad *et al.*, *Measurement of the flavour composition of dijet events in pp collisions at* $\sqrt{s} = 7$ *TeV with the ATLAS detector*, Eur. Phys. J. C **73** (2013) 2301, arXiv:1210.0441.

[117] ALICE collaboration, ATLAS collaboration, CMS collaboration, LHCb collaboration, S. Menke, *Top production and decay*, PoS **LHCP2021** (2021) 116.

[118] R. J. Barlow and C. Beeston, *Fitting using finite Monte Carlo samples*, Comput. Phys. Commun. **77** (1993) 219.

[119] G. D'Agostini, *A Multidimensional unfolding method based on Bayes' theorem*, Nucl. Instrum. Meth. A **362** (1995) 487.

[120] G. D'Agostini, *Improved iterative Bayesian unfolding*, in *Alliance Workshop on Unfolding and Data Correction*, 2010, arXiv:1010.0632.

[121] J. Bourbeau and Z. Hampel-Arias, *Pyunfold: A python package for iterative unfolding*, The Journal of Open Source Software **3** (2018) 741.

[122] R. Barlow, *Systematic errors*, accessible at: https://prezi.com/jonyomm1b-ls/systematic-errors/. Talk at LHCb UK Student Meeting (presented: 09/08/11; accessed: 23/11/22).

Appendices

Appendix A

Effects of mistagging jet charge

Here the effects of mistagging the charge of jets on the measured value of the asymmetry and the associated uncertainty will be derived. Variables will be given the superscript "true" or "meas." for the true underlying values and values that are measured by the detector respectively. The number of forward and backward events measured by the detector is given by

$$N(\Delta\eta > 0) = N_F^{\text{meas.}},$$
$$N(\Delta\eta < 0) = N_B^{\text{meas.}},$$

where $\Delta\eta = \eta_b - \eta_{\overline{b}}$ is the difference in pseudorapidity between the b quark and $\overline{b}$ quark. The total number of tagged events is given by $N_{\text{tagged}} = N_F^{\text{meas.}} + N_B^{\text{meas.}}$. Therefore the probability, p, that an event is assigned into the $N_F^{\text{meas.}}$ set of events is

$$p = \frac{N_F^{\text{meas.}}}{N_{\text{tagged}}}.$$

Similarly, the probability, q, that an event is assigned into the $N_B^{\text{meas.}}$ set of events is

$$q = \frac{N_B^{\text{meas.}}}{N_{\text{tagged}}} = (1 - p).$$

This is a binomial distribution (as an event must be placed in either group), and therefore the variance in $N_F^{\text{meas.}}$ is

$$\sigma^2_{N_F^{\text{meas.}}} = N_{\text{tagged}}\, p\,(1 - p).$$

The measured asymmetry can also be rewritten

$$A^{\text{meas.}} = p - (1-p) = 2p - 1,$$

and hence

$$\begin{aligned}
\sigma_A^2 &= 4\sigma_p^2, \\
&= \frac{4\sigma^2_{N_F^{\text{meas.}}}}{N^2_{\text{tagged}}}, \\
&= \frac{4p\,(1-p)}{N_{\text{tagged}}}, \\
&= \frac{1-(A^{\text{meas.}})^2}{N_{\text{tagged}}}.
\end{aligned}$$

The number of tagged events can be related to the total number of events by

$$N_{\text{tagged}} = \varepsilon N_{\text{tot}},$$

where ε is the efficiency of the tagging procedure. This gives a variance in the measured value of the asymmetry

$$\sigma_A^2 = \frac{1-(A^{\text{meas.}})^2}{\varepsilon N_{\text{tot}}}.$$

The true value of the asymmetry (i.e. if there were no mistagging) in the selected sample is given by

$$A^{\text{true}} = \frac{N_F^{\text{true}} - N_B^{\text{true}}}{N_F^{\text{true}} + N_B^{\text{true}}},$$

where $N_F^{\text{true}} + N_B^{\text{true}} = N_{\text{tagged}}$ still holds. The measured values of $N_F^{\text{meas.}}$ and $N_B^{\text{meas.}}$ are related to these values by

$$\begin{aligned}
N_F^{\text{meas.}} &= (1-\omega)N_F^{\text{true}} + \omega N_B^{\text{true}}, \\
N_B^{\text{meas.}} &= (1-\omega)N_B^{\text{true}} + \omega N_F^{\text{true}},
\end{aligned}$$

where ω is the probability of incorrectly tagging an event, and hence $(1-\omega)$ is the probability of correctly tagging an event. Therefore the measured asymmetry can be

rewritten in terms of these true values

$$\begin{aligned} A^{\text{meas.}} &= \frac{1}{N_{\text{tagged}}}\left[(1-\omega)N_F^{\text{true}} + \omega N_B^{\text{true}}\right] - \left[(1-\omega)N_B^{\text{true}} + \omega N_F^{\text{true}}\right], \\ &= (1-2\omega)\frac{N_F^{\text{true}} - N_B^{\text{true}}}{N_{\text{tagged}}}, \\ &= (1-2\omega)A^{\text{true}}, \\ &= DA^{\text{true}}, \end{aligned}$$

where $D = (1-2\omega)$ is called the dilution factor. This dilution factor allows the measured value of the asymmetry to be corrected to give an estimate on the true value of the asymmetry. The uncertainty on this estimate is then given by

$$\begin{aligned} \sigma^2_{\text{true}} &= \frac{\sigma^2_A}{D^2}, \\ &= \frac{1-(A^{\text{meas.}})^2}{\varepsilon D^2 N_{\text{tot}}}. \end{aligned}$$

In studies presented in this thesis, the asymmetry is expected to be $\mathcal{O}(1\%)$. Therefore, the assumption $(A^{\text{meas.}})^2 \ll 1$ can be made, giving

$$\sigma^2_{\text{true}} = \frac{1}{\varepsilon D^2 N_{\text{tot}}}.$$

The analysis is optimised when this uncertainty is minimised or, equivalently, when the metric εD^2 is maximised. This metric gives a measure of the statistical significance of a result and will depend on the exact selection used on the data. It should give a balance between penalising mistagging events (which will give a small value of D) and placing tight selection requirements on the data (which will give a small value of ε).