UAB
Universitat Autònoma
de Barcelona

Universitat Autònoma de Barcelona
Faculty of Science
Department of Physics

Bachelor's Thesis in Physics

Characterisation of the Opposite-Side Flavour Taggers in LHCb Run 3.

Roger Feliu Vert
Supervised by Prof. Dr. Stephanie Hansmann-Menzemer and Dr. Sara Celani
with administrative supervision by Dr. José Flix Molina

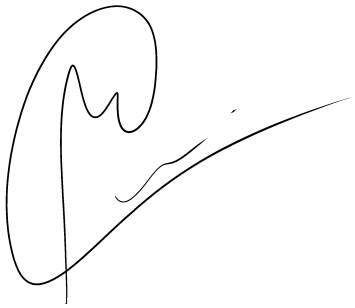
– July 2025 –

DECLARACIÓ D'AUTORIA DEL TREBALL DE GRAU

Jo, Roger Feliu Vert, amb Document Nacional d'Identitat 47741736S, i estudiant del Grau en Física de la Universitat Autònoma de Barcelona, en relació amb la memòria del treball de final de Grau presentada per a la seva defensa i avaluació durant la convocatòria de Juliol del curs 2024-2025, declara que

- El document presentat és original i ha estat realitzat per la seva persona.
- El treball s'ha dut a terme principalment amb l'objectiu d'avaluar l'assignatura de treball de grau en física en la UAB, i no s'ha presentat prèviament per ser qualificat en l'avaluació de cap altra assignatura ni aquesta ni en cap altra universitat.
- En el cas de continguts de treballs publicat per terceres persones, l'autoria està clarament atribuïda, citant les fonts degudament.
- En els casos en els quals el meu treball s'ha fet en col·laboració amb altres investigadors i/o estudiants, es declara amb exactitud quines contribucions es deriven del treball de tercers i quines es deriven de la meva contribució.
- A l'excepció dels punts esmentat anteriorment, el treball presentat és de la meva autoria.

Signat:



_____

Abstract

This thesis presents a detailed study of flavour tagging performance in Run 3 of the LHCb experiment, focusing on calibration and characterisation of opposite-side (OS) tagging algorithms trained on Run 2 data. Utilizing control and signal channels such as $B^+ \to J/\psi\, K^+$ and $B_s^0 \to J/\psi\, \phi$, the tagger outputs are recalibrated using the `FTCalib` framework with sWeight methods and kinematic reweighting. Similarly, the same-side Kaon tagging algorithm was independently calibrated on $B_s^0 \to D_s^{\mp}\pi^{\pm}$ decays.

In addition to this, a machine-learning-based approach to meta-tagging is developed, where the output of individual OS tagging algorithms are merged both with a gradient boosted decision tree (BDT) and a Transformer neural network. After calibration, the BDT outperforms the official combination algorithm and the Transformer, reaching a tagging power of 7.08%. The official `FTCalib` combination algorithm and the Transformer achieve 6.97% on the same data set.

The results confirm the transferability of Run 2 taggers to Run 3 conditions under proper calibration, and demonstrate the utility of meta-tagging as a flexible and robust flavour tagging substitute for LHCb. These developments pave the way for more inclusive, data-driven approaches in future tagging strategies.

Resum

Aquest treball presenta un estudi detallat del rendiment de l'etiquetatge de sabor (*flavour tagging*) durant el Run 3 de l'experiment LHCb, centrant-se en la calibració i caracterització dels algorismes etiquetadors del costat oposat (*OS*) entrenats amb dades del Run 2. Mitjançant canals de control i senyal com ara $B^+ \to J/\psi\, K^+$ i $B_s^0 \to J/\psi\, \phi$, les sortides dels etiquetadors han estat recalibrades amb el paquet `FTCalib`, utilitzant tècniques de sWeights i reponderació cinemàtica. L'algorisme d'etiquetació per kaons del mateix costat (*SS Kaon*) també s'ha calibrat de manera independent en desintegracions $B_s^0 \to D_s^{\mp}\pi^{\pm}$.

A més a més, s'ha desenvolupat un enfocament d'aprenentatge automàtic (*machine learning*) per a la meta-etiquetació, combinant les sortides dels algorismes d'etiquetació OS mitjançant un *gradient boosted decision tree* (BDT) i una xarxa neuronal *Transformer*. Després de la calibració, el BDT ha superat l'algorisme de combinació oficial i el *Transformer*, assolint una potència d'etiquetatge del 7.08%. L'algoritme de combinació oficial de `FTCalib` i el *Transformer* han assolit un valor de 6.97% sobre el mateix conjunt de dades.

Els resultats confirmen l'aplicabilitat dels etiquetadors del Run 2 a les condicions del Run 3 quan són adequadament calibrats, i demostren el potencial de la meta-etiquetació com a alternativa flexible i robusta per a l'etiquetatge de sabor a LHCb. Aquests desenvolupaments obren el camí a enfocaments basats en dades i més inclusius en les estratègies futures d'etiquetatge.

Contents

1 Introduction

The Standard Model of Particle Physics (SM) is our current best theory for understanding elementary particles and their interactions. It is the most complete one we have up to today, and has shown to have a great predictive power on most experimental data. A well-known example of this is the discovery of the Higgs boson in the year 2012, after its existence and mass were predicted by the SM.

Despite the success of the SM in describing particle interactions, it cannot account for several astronomical observations, such as the dominance of matter over antimatter in the universe. One promising approach to searching for New Physics phenomena is through so-called indirect searches, which provide access to energies far beyond direct reach. Precision tests of quantum-loop–induced processes may reveal contributions from new heavy particles. At the LHCb experiment, located at the Large Hadron Collider (LHC) at CERN, one of the most promising indirect search for NP phenomena is the study of time-dependent CP asymmetry in $B_s^0 \to J/\psi\,\phi$ decays[1], induced by the interference of the amplitude of the decay and the mixing with subsequent decay.

The data-taking periods at the LHC are divided into distinct runs, with shutdowns in between for maintenance and upgrades. Run 2 refers to the period between 2015 and 2018 during which the LHC was operated at a centre-of-mass energy of 13 TeV and the LHCb detector operated in its original configuration. Run 3, which began in 2022 following a major upgrade, features important improvements to the resolution of the detector and data acquisition, such as a complete software-based trigger and increased luminosity. These improvements render LHCb increasingly capable of physics discovery, but necessitate the re-assessment of already trained algorithms, such as the flavour tagging algorithms. Flavour tagging algorithms are designed to determine the production flavour of B mesons (i.e, if they contain a b or a $\bar{b}$), exploiting information of the production on the hadronisation process.

Machine learning techniques are already used in the LHCb experiment, as well as in many other research fields. In particular, Boosted Decision Trees (BDT) are fast models with easy implementation with a wide use in LHCb. Nevertheless, recent developments in machine learning are being used to accomplish significant improvements in the design and performance of flavour tagging algorithms. In particular, Transformer-based architectures (originally developed for natural language processing) offer a flexible and highly expressive framework for modelling sequential and relational data. This thesis explores the application of both Transformer and BDTs models to the test of B_s^0 flavour tagging, with the goal of constructing a new meta-tagging algorithm to combine the decisions of existing individual taggers.

A key aspect of this study involves the calibration of the model's output predicted mistag probabilities on recorded or simulated LHCb data. The calibration procedure ensures that the predicted mistag are reliable and that the tagging performance can be accurately quantified using the established control channels. In this work, we focus on the $B_s^0 \to J/\psi\,\phi$ decay as the signal channel and make use of $B^+ \to J/\psi\,K^+$ as the primary calibration mode, following LHCb's standard methodology.

This thesis builds upon on various internal tools, standard procedures, and previous works developed within the LHCb collaboration which are carefully adapted, applied, and extended to Run 3 data as part of this work. All external contributions are explicitly acknowledged throughout the text. In addition, the development, training, and calibration of machine-learning-based meta-taggers, described in Section 6, constitute original work entirely carried out by the author, representing a novel approach to inclusive flavour tagging within the experiment.

By combining modern deep learning techniques with established calibration strategies, this thesis aims to contribute to the ongoing development of high-performance flavour tagging at LHCb, ultimately enabling more precise measurements of fundamental parameters in the B sector and enhancing the sensitivity to potential New Physics phenomena.

[1] Charge-conjugated decays are implied unless specified.

2 Theoretical Foundations

The current theoretical framework used to describe the fundamental interactions of the Standard Model (SM) is Quantum Field Theory (QFT). These interactions are described as exchanges of particles with integer spin (bosons), rather than as interactions through classical potentials. There are four fundamental interactions in nature: gravitational, electromagnetic, strong, and weak. The SM, however, only includes the latter three, and the weak interaction is of particular relevance for the decay modes studied in this thesis.

2.1 Weak interaction

The weak interaction is mediated by three different, massive bosons: the $W^{\pm}$ and Z^0 bosons, for charged or neutral current interactions.

The SM's fundamental building blocks of matter are fermions, particles with half-integer spin, which are split into leptons and quarks. Quarks form composite particles: mesons (quark-antiquark pairs) and baryons (three-quark states). Each particle has an antiparticle counterpart with opposite quantum numbers, such as electric charge or colour, but otherwise has the same properties as the particle itself. Also, quarks may change flavour via charged-current weak interactions.

Flavour-changing transitions between quarks are controlled by the unitary Cabibbo–Kobayashi–Maskawa (CKM) matrix [1, 2]

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \tag{1}$$

which describes the strength of the coupling for different flavour-changing interaction and introduces charge-parity (CP) violation into the SM as well. CP violation, observed only in the weak interaction, breaks symmetry under joint charge conjugation and space inversion, turning particles into antiparticles.

In hadronic weak decays, interference between particle-antiparticle amplitudes can lead to CP violation. In the B_s^0 system, it is expressed through the weak phase ϕ_s, whose precise measurement could reveal a difference with respect to SM calculations and a hint of New Physics.

2.2 $B_s^0 - \bar{B}_s^0$ mixing and taggers overview

Neutral B mesons oscillate between being a particle and an anti-particle through a quantum loop, a process called mixing. In this loop, New Physics may contribute in the form of new particles and modify the CP-violating phase ϕ_s [3, 4, 5].

The $B_s^0 \to J/\psi\,\phi$ decay is very clean (without undetectable neutrinos) and relatively common, which allows for a nice reconstruction of this channel. In this decay, the B_s^0 travels a distance of about 1 cm [6] (while oscillating back and forward into a $\bar{B}_s^0$) before decaying into a J/ψ meson and a ϕ meson which further decay into $\mu^+\mu^-$ and K^+K^- respectively. This decay chain is the same for the $\bar{B}_s^0$, so it is not a trivial task to know the flavour of the B_s^0 meson at production nor at decay. Flavour tagging algorithms are used to decide the production flavour.

LHCb employs a combination of opposite-side (OS) and same-side (SS) tagging methods, which rely on the identification of decay products from the other b hadron in the event or from the hadronization process of the signal B meson. These include electron, muon, kaon, and vertex-charge taggers (charm tagger [7] also exists, but it is not included here), whose decisions are combined using the `FTCalib` algorithm [8] and calibrated using self-tagging control channels. Figure 1 shows a visual scheme of the distribution of the SS and OS taggers along the hadronisation process of a $B_{(s)}^0$.

To quantify the effective statistical weight added by tagging that directly impacts the sensitivity to CP-violating parameters such as ϕ_s, the effective tagging power is used. It is provided by

$$\varepsilon_{\text{eff}} = \varepsilon_{\text{tag}} D^2 = \varepsilon_{\text{tag}} (1 - 2\omega)^2, \tag{2}$$

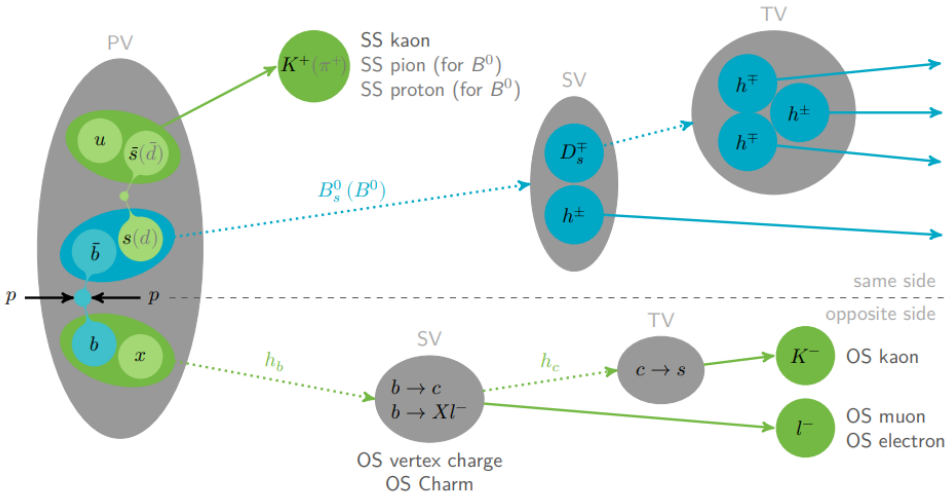


Figure 1: Illustration of the hadronisation process of a $B^0_{(s)}$ meson in a pp collision at LHCb, from [9]. The signal B-meson decay chain is highlighted in blue.

where ε_{tag} is the tagging efficiency, ω is the mistag fraction and D is the dilution. The tagging efficiency and the mistag fraction are expressed as

$$\varepsilon_{\text{tag}} = \frac{R+W}{R+W+U} \quad \text{and} \quad \omega = \frac{W}{R+W}, \tag{3}$$

where R, W, U are the number of correctly tagged, incorrectly tagged and untagged B candidates, respectively.

The mistag fraction can be extracted from flavour-specific decay modes (called "control channels" in this thesis), i.e. those decay modes in which the final state particles individually specify the quark/antiquark composition of the signal B. In this study the decay modes $B^+ \to J/\psi\, K^+$, and $B_s^0 \to D_s^+\pi^-$ are used to calibrate the OS and SS taggers, respectively. For charged mesons, the mistag fraction is determined by directly comparing the tagging decision with the B flavour of the signal since there is no mixing. For neutral mesons, it is determined by fitting the B^0 flavour oscillation as a function of the decay time.

The chance for the choice of a given tag to be accurate is calculated from the event and kinematic characteristics of the tagging particle by means of a neural network that has been trained on Monte Carlo (MC) events in order to identify the correct flavour of the signal B. When more than one tagging algorithm responds to a B candidate, the probabilities provided by each of the algorithms are combined into a single probability and the decisions are combined into a single decision. The combined probability can be utilized on an event-by-event basis in order to assign a higher weight to events with small probability of mistag and thus to increase the overall significance of an asymmetry measurement [10].

2.3 Motivation for recalibration and meta-tagging in Run 3

The flavour tagging algorithms that are currently in use at LHCb were initially developed and optimised on the data taken during Run 2. The start of Run 3, however, has brought a number of major changes within the experiment, namely a substantial upgrade of the detector and the introduction of a fully software-based trigger system. These changes affect the kinematic and topological distributions of the reconstructed particles, and could influence the efficiency and reliability of the current tagging systems when projected onto novel datasets.

It is thus necessary to re-tune the predicted mistag probability of Run 2 taggers (taggers trained with Run 2 data) with control channels in Run 3. Re-tuning is critical in order to achieve precise flavour tagging in physics analyses, especially in tests of time-dependent CP asymmetries, where mistagging biases directly affect the extraction of physical parameters.

In addition, the default technique used to combine tagger responses in LHCb, the `FTCalib` algorithm, relies on a likelihood-based combination algorithm to optimally merge the different taggers under the assumption of conditional independence that, while very resilient, cannot model inter-tagger correlations

or dynamically adapt to event-level context. This scenario paves the way for the application of machine learning-based strategies (meta-taggers), which can learn an optimal combination of the tagger decisions based on a data-driven approach.

This study investigates both directions: it carries out the calibration of the opposite-side taggers built in Run 2 on Run 3 control and signal channels, and studies the development of new meta-taggers with the aim of substituting the default combiner. The purpose is to evaluate the feasibility and potential gain of their integration in the LHCb flavour tagging system for upcoming analyses.

3 The LHCb detector in Run 3

The LHCb experiment is one of the four major experiments at CERN's Large Hadron Collider (LHC) dedicated to precise studies of hadrons involving b and c quarks. The main objective is to test the Standard Model predictions for CP violation and rare decays in the heavy flavour system, and to search for indirect signatures of New Physics phenomena.

In contrast to multipurpose detectors such as ATLAS or CMS, LHCb is a single-arm forward spectrometer. This exploits the kinematics of $b\bar{b}$ production at the LHC, where b hadrons are primarily emitted in the forward or backward directions due to their small mass and the properties of parton distribution functions. The pseudorapidity[2] acceptance of the detector is approximately $2 < \eta < 5$ [11], optimal for heavy flavour decays.

The LHCb detector consists of a series of subsystems along the beam line. The closest to the interaction point is the Vertex Locator (VELO), composed of silicon pixel sensors, which carries out accurate reconstruction of the primary and secondary vertices. Such capability plays an important role in time-dependent measurements, as it makes it possible to reconstruct the decay length and decay time of long-lived particles such as B mesons.

Downstream of the VELO is the tracking system, made up of silicon-strip detectors preceding and succeeding a dipole magnet that has an integrated field of approximately 4 Tm [11]. This enables the measurement of charged particles with great accuracy in momentum, with resolutions of approximately 0.5% at low momenta and up to 1.0% at 200 GeV/c [12].

For particle identification (PID), LHCb uses two Ring Imaging Cherenkov (RICH) detectors for the discrimination of pions, kaons, and protons over a wide momentum range. They are complemented by electromagnetic and hadronic calorimeters for identification of photons, electrons, and hadrons and by a muon system consisting of layers of absorbers and multi-wire proportional chambers or triple-GEM detectors.

The experiment has a two-level trigger system, entirely software-based. The first stage, HLT1, performs a fast reconstruction of key signatures using information from the full detector, while the second stage, HLT2, performs a more refined and complete event reconstruction in real time. This fully software-based approach enables greater flexibility and precision in selecting events for offline analysis.

Since its major upgrade during the Long Shutdown 2 of the LHC, the LHCb detector runs at an instantaneous luminosity of up to 2×10^{33} cm^{-2}s^{-1} and a readout rate of 40 MHz [12]. The data collected since then (Run 3 data) is collected in pp collisions at a maximum of $\sqrt{s} = 13.6$ TeV. The improved data rate increases the experiment's sensitivity to rare processes and small asymmetries that are the subject of CP violation and flavour tagging research.

The unique capabilities of the LHCb detector specifically make it ideal for algorithm research and experimentation of flavour tagging. The robust tracking, high-resolution vertexing, and powerful PID allow reconstruction of decay chains and tagging particle identification on signal and opposite sides of the $b\bar{b}$ event. These are required for using and experimenting with sophisticated machine learning techniques such as those found in this thesis.

[2]The pseudorapidity is defined as $\eta = -\ln(\tan(\theta/2))$, where θ is the angle from the beam pipe

4 Decay channels and datasets

To correctly tune the tagging algorithms and assess their performance on Run 3 data, it is required to identify some decay channels that allow for a clear determination of the initial flavour of the B meson. In this section, the decay modes exploited in the analysis, the relevant data samples, and the performance yielded by Run 2 trained taggers applied to these events are explained. These items constitute the building blocks of the calibration and meta-tagging activities outlined in the subsequent sections.

4.1 Selected decay channels

As stated, the signal channel of this thesis is the $B_s^0 \to J/\psi\,\phi$ channel, since it is a rather common and clear decay. There are two other specific decay channels used in this analysis to evaluate and calibrate the tagging algorithms on Run 3 data. These other channels are selected based on their ability to unambiguously determine the production flavour of the B meson and their relevance within the LHCb flavour tagging strategy.

The already presented decay $B^+ \to J/\psi K^+$, where the J/ψ subsequently decays to a muon pair, is the used as the control channel of OS taggers. The original B^+ is not neutral, so it does not undergo mixing, and the initial flavour can be known with certainty from the charge of the final-state kaon. It shares the OS characteristics as the signal channel and, additionally, this channel has a relatively high branching fraction and a clean signature with muons in the final state, which facilitates its selection and reconstruction in the LHCb detector, making it ideal for mistag calibration.

This analysis is mostly focused on the OS taggers, but also evaluates and calibrates the most relevant SS tagger, the SS Kaon, for completeness of the study of flavour tagging power. For that, the decay $B_s^0 \to D_s^{\mp}\pi^{\pm}$ is used as the control channel of SS Kaon, where the $D_s^{\mp}$ decays into a fully hadronic final state such as $K^+K^-\pi^-$. This mode is also flavour-specific, in the sense that the final-state pion charge identifies the flavour of the neutral B_s^0 meson at the time of its decay. However, due to the possibility of B_s^0–$\bar{B}_s^0$ mixing, the mistag probability must be extracted from a time-dependent fit to the oscillation pattern. This channel closely resembles the topology of the signal decay $B_s^0 \to J/\psi\phi$, making it a natural proxy for evaluating tagging performance in physics analyses targeting CP violation in the B_s system.

4.2 Data samples and preselection

For this analysis, data taken in 2024 (Run 3) at the LHCb experiment in the measurement periods block 1 and 2 is used. It includes only the upwards magnet polarity and corresponds to an integrated luminosity of around 1.78 fb^{-1}. The data samples used for both the control and signal channels are selected from ntuples provided as preselected datasets containing reconstructed candidates for both the $B^+ \to J/\psi K^+$ and $B_s^0 \to D_s^{\mp}\pi^{\pm}$ decay modes, respectively. These samples include events that already satisfy the trigger requirements and basic quality criteria, as defined by the corresponding central productions.

Although the trigger selection was not directly implemented in this analysis, it is important to acknowledge its role in shaping the final dataset. For the control channel $B^+ \to J/\psi K^+$, the selection typically relies on dimuon-based trigger lines designed to select events with a clean muon pair from the J/ψ decay while minimizing decay-time biases. These trigger lines include requirements on the muon kinematics and vertex quality, such as the muon momenta of $p > 6$ GeV/c and a requirement of the invariant dimuon mass $m(\mu^+\mu^-) > 2.7$ GeV/c^2. The lines do not introduce stringent cuts on the flight distance or impact parameter that would distort the decay-time distribution.

The ntuples contain all relevant reconstructed variables needed for candidate selection, kinematic cuts, particle identification, and tagging evaluation. From these, the analysis proceeds with an additional offline selection, described in detail in Table 1 below, to further refine the candidate purity. This includes constraints on vertex quality, mass windows, decay times, and PID variables.

Additionally, the outputs of the Run 2-trained flavour tagging algorithms are added to the ntuples via the official LHCb tagging framework. In addition to the above offline selection, a BDT classifier is trained on the candidates to $B^+ \to J/\psi K^+$ decay events combined with sideband data, optimized to distinguish between signal and combinatorial background while keeping an unbiased decay time distribution. Classifier creation and training had been conducted by a fellow LHCb collaboration member as part of her Bachelor thesis [14]. A fixed cut applied to the BDT response serves the purpose of keeping high signal efficiency while minimizing background events. The internal architecture, training parameters, and

condition	range
decay time t	$0.3 < t < 15$ ps
decay time error σ_t	$\sigma_t < 0.15$ ps
invariant mass $m(J/\psi K^+K^-)$	$5200 < m < 5550$ MeV/c^2
invariant mass $m(K^+K^-)$	$990 < m < 1050$ MeV/c^2

Table 1: Offline selection from [13].

performance verification of the BDT are discussed in [14], and the identical model is applied directly in this analysis.

An additional noteworthy cut was applied to the predicted mistags of the OS Vertex Charge. After the analysis, some outliers hard to explain were found in the calibration curve of this tagger, and when plotting a histogram[3] of the values in their whole range, 3 spikes (values with abnormal repetition in the data) were discovered. Specifically, the values 0.3250, 0.4141 and 0.3500, and since this did not have any physical explanation, those candidates were just removed from the dataset, fixing the outliers observed on the calibration curve.

Both control channels are subject to basic preselection criteria, offline selection and a BDT cut to remove poorly reconstructed candidates and reduce combinatorial background. Since this analysis is part of a larger analysis within the LHCb collaboration, and its results are used by other members of the research group for their own investigation (same as other members' results are used here), specific ntuples for each channel are used to ensure coherence in the overall analysis. The resulting samples contain over 300,000 candidates for the OS control channel and over 200,000 candidates for the signal and SS control channel.

4.3 Run 2 opposite-side flavour tagging algorithms

The flavour tagging information used in this analysis is given by the OS taggers (and SS Kaon), which have been trained and calibrated on Run 2 data. The purpose of these taggers is to identify the initial flavour of the signal B meson by tagging the flavour of the accompanying b-hadron produced at the event (while SS taggers identify particles produced in the hadronisation of the signal B meson itself). The OS tagging algorithm merges information from the individual taggers, each making use of a different tagging particle, including muons, electrons, kaons, and inclusive secondary vertices (vertex charge).

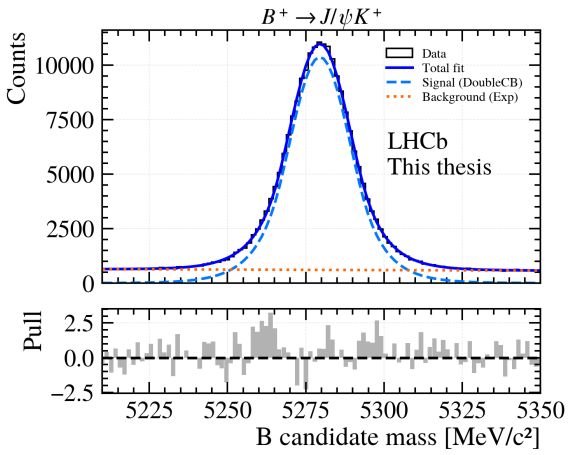
Each sub-tagger returns a tagging decision $d_i \in \{-1, 0, +1\}$, $\bar{b}$, untagged, or b flavour, respectively, and a predicted mistag probability $\eta_i \in [0, 0.5]$. From these outputs, there are several performance metrics that can be defined, as previously mentioned in Section 2.2. The tagging efficiency ε_{tag} is the fraction of events that receive a non-zero decision; the mistag rate ω is the fraction of tagged events whose decision is incorrect; and the tagging power, defined in equation 2.2, quantifies the statistical impact of the tagger in oscillation or time-dependent CP violation measurements. These metrics will be evaluated after calibration in Section 5.4. In this analysis, the tagger outputs are directly taken from the official LHCb Flavour Tagging software package trained on the control channels in Run 2 and are used on Run 3 data without retraining or recalibration.

The combination of multiple tag decisions, d_i, and calibrated mistags, $\omega_i(\eta_i)$, is carried out using the likelihood-based method implemented in the `FTCalib` algorithm [8]. For clarification, η_i is the predicted mistag (by the i-th tagger), while $\omega_i(\eta_i)$ is the calibrated mistag, which is explained more in depth in Section 5.1. For each candidate, the probability of containing a b or $\bar{b}$ quark is computed as

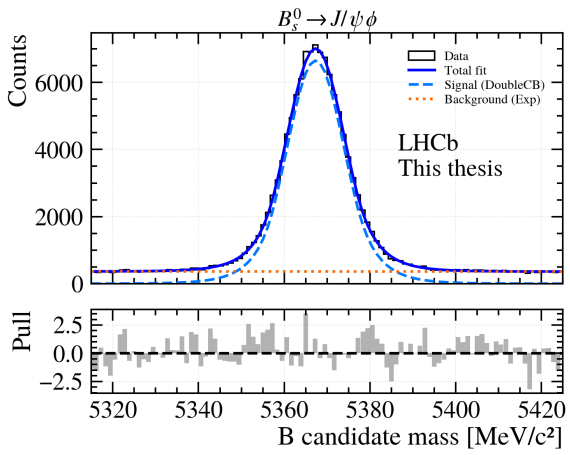
$$P_b(p_b, p_{\bar{b}}) = \frac{p_b}{p_b + p_{\bar{b}}}, \qquad P_{\bar{b}} = 1 - P_b, \tag{4}$$

where the quantities p_b and $p_{\bar{b}}$ aggregate the individual tagger decisions and predicted mistags as

[3] Those can be found in Appendix C, along a more extensive explanation on the matter.



(a) Mass fit to the $B^+ \to J/\psi\, K^+$ candidates.



(b) Mass fit to the $B_s^0 \to J/\psi\, \phi$ candidates.

Figure 2: Invariant mass fits used for signal extraction. The fits are used to derive sWeights for background subtraction in the tagging calibration.

$$p_b(\vec{\omega}, \vec{d}) = \prod_i \left(\frac{1 + d_i}{2} - d_i \left(1 - \omega_i(\eta_i)\right) \right), \tag{5}$$

$$p_{\bar{b}}(\vec{\omega}, \vec{d}) = \prod_i \left(\frac{1 - d_i}{2} + d_i \left(1 - \omega_i(\eta_i)\right) \right). \tag{6}$$

The final combined tag decision d_{comb} and predicted combined mistag probability η_{comb} are then calculated as

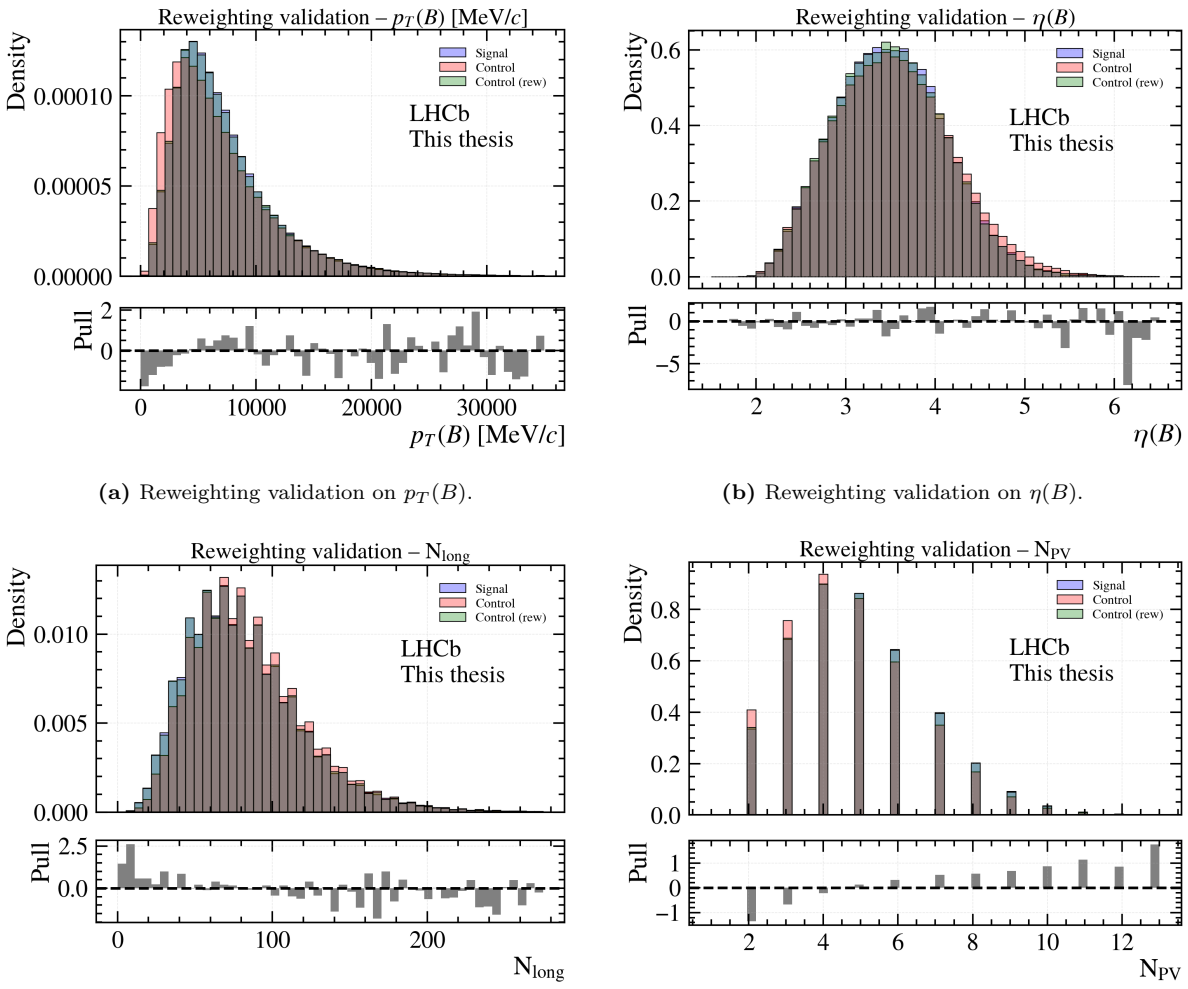
$$d_{\text{comb}} = \text{sign}(P_b - P_{\bar{b}}), \qquad \eta_{\text{comb}} = 1 - \max(P_b, P_{\bar{b}}). \tag{7}$$

This method assumes conditional independence between taggers and provides a smooth way to incorporate the information from multiple sub-taggers into a unified tagging decision. As said, one objective of this analysis is to investigate data-driven alternatives to this combination method, as presented in Section 6.

4.4 Mass fits and sWeight extraction

In order to perform a correct and unbiased calibration of the mistag probability on recorded data, it is necessary to isolate the component of the background in the selected data samples. This is done by using an unbinned maximum likelihood fit method on the invariant mass distribution of the reconstructed B meson candidates.

In both OS control channels and signal channel, the distribution of invariant mass is modelled by a Double Crystal Ball [15] with a shared mean to represent the signal peak, and an exponential function to represent the combinatorial background. The mass fits are carried out by using the `zfit` [16] package, and displayed in Figure 2. The mass fit for the SS control channel is more complex and was developed by a fellow member of the LHCb collaboration, whose work is currently ongoing and therefore cannot be cited at this time, and was performed using the `B2DXFitters`[4] package. The signal probability density function is modelled as the sum of an Ipatia function and a Johnson U distribution, while a double exponential describes the background. In addition, the lower sideband region is modelled using templates derived from simulated events.



(a) Reweighting validation on $p_T(B)$.

(b) Reweighting validation on $\eta(B)$.

(c) Reweighting validation on `NLongTracks`.

(d) Reweighting validation on `NPVs`.

Figure 3: Validation of the reweighting procedure for $B^+ \to J/\psi\, K^+$ candidates across the four input variables.

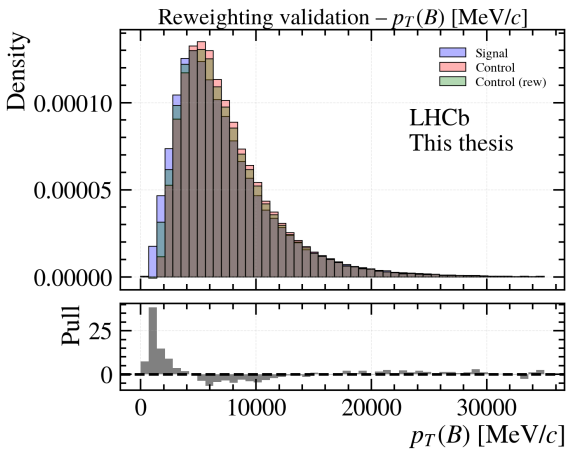
4.5 Reweighting of the control channels

Once the fitting process is completed, the *sPlot* method [17] is applied to statistically subtract the background component and thereby facilitate the extraction of the genuine distribution of the signal events, using the `HepC`[5] package. It attributes an event-by-event weight called sWeight, derived from the covariance matrix of the fitted yields and the total likelihood. The sWeights values are constructed such that, when applied to a distribution uncorrelated with the feature used for their extraction (the invariant mass), the resulting weighted histogram corresponds to the signal distribution. It is important to note that the sWeights do not represent the probability of each event being signal, since they can take negative values (where the background PDF is dominating over the signal PDF) or above one, and only have meaning when applied collectively over many events.

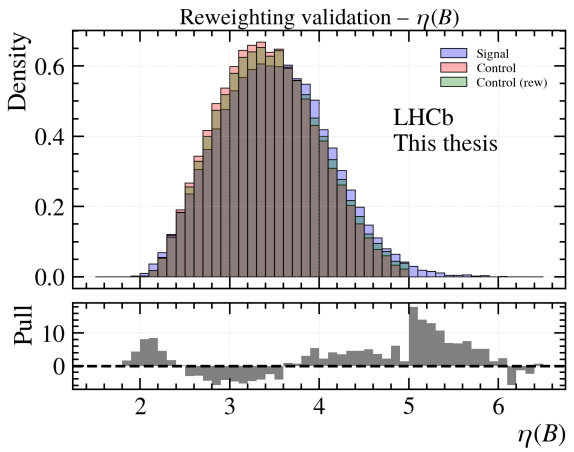
This method relies on the hypothesis that the discriminative variable is uncorrelated with variables of interest, i.e., the tagging decision and the predicted mistag probability. In flavour tagging, this hypothesis is true in general because the tagging variables rely on the opposite-side b-hadron and not on the signal-side mass reconstruction.

The usage of control channels is key for the calibration process, since they are flavour-specific decay-modes, and the ones chosen share similar characteristics with the signal channel. Even though these similarities, there are some differences between them, so a direct calibration on the control channel may

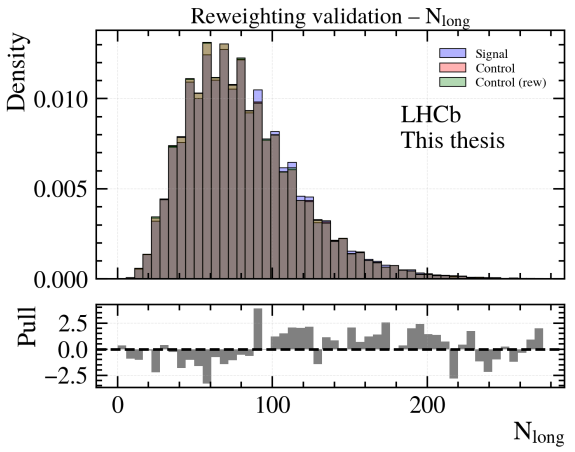
[4]It is an LHCb-internal tool developed for mass and time-dependent fits involving $B \to DX$ decays.
[5]It is also an internal tool provided by the LHCb group.



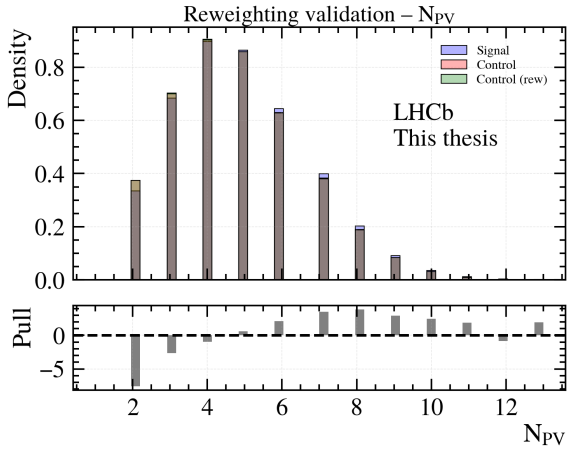
(a) Reweighting validation on $p_T(B)$.



(b) Reweighting validation on $\eta(B)$.



(c) Reweighting validation on `NLongTracks`.



(d) Reweighting validation on `NPVs`.

Figure 4: Validation of the reweighting procedure for $B_s^0 \to D_s^{\mp}\pi^{\pm}$ candidates across the four input variables.

lead to biased results. For example, the kinematic properties of the control and signal samples may differ, and flavour tagging algorithms can depend on the kinematics of the signal B meson. To mitigate this, a reweighting procedure is applied to the control channels to match the kinematic distributions of signal and control channels as much as possible before calibration.

The reweighting is performed using the kinematic variables that are expected to influence the tagging response but remain uncorrelated with the tagging decision or predicted mistag probability. In this analysis, the transverse momentum, $p_T(B)$; pseudorapidity ,$\eta(B)$, of the reconstructed B candidate; the number of primary vertices, `NPVs`, and the number of tracks that are reconstructed in all tracking stations of the LHCb spectrometer,`NLongTracks`, are used.

A multidimensional reweighting algorithm is used to derive a weight function that transforms the distribution of the control sample to match that of the signal. In this analysis, it is implemented through the use of `GBReweighter`[18], a Gradient Boosted Regression (GBR) based reweighting tool often used in High Energy Physics from the `hep_ml` library, trained to model the density ratio between the control and signal samples, wherein the sWeights from previous section are employed both in the origin and target distribution, to ensure reweighting of pure the signal samples.

To validate the results of the reweighting, the distributions of the reweighted control sample and the original signal sample are compared, along the original control sample, to visualize the change. In Figure 3, the blue columns represent the original signal sample, the red ones represent the original control sample and the green ones represent the reweighted sample for the $B^+ \to J/\psi\, K^+$ candidates. The superposition of all three displays the gray colour. A good agreement is observed in all input variables, indicating that

the control sample provides a suitable proxy for the signal in terms of kinematic properties.

The reweighting validation on the $B_s^0 \to D_s^{\mp}\pi^{\pm}$, shown in Figure 4, shows certain disagreement in the input variables, specially in the $p_T(B)$ with huge pulls for low values and $\eta(B)$ with moderate pulls along the whole range. This is related to the difficulty of a multidimensional reweighing when the input variables differ, highlighting the higher complexity of this channel.

5 Calibration of the tagging algorithms

5.1 Motivation and calibration formalism

The output of a flavour tagging algorithm is a tag decision $d \in \{-1, 0, +1\}$ and a predicted mistag probability $\eta \in [0, 0.5]$. Whereas the tag decision indicates the inferred flavour of the B meson initially, the mistag probability quantifies the uncertainty of the algorithm in this assignment.

Nevertheless, the predicted η values are obtained from classifiers that have been trained with Run 2 data, and might not be optimal estimates of true mistag rates when extrapolating to Run 3 samples due to detector condition differences, event topologies, and trigger settings.

To correct for this, a calibration step is performed to ensure that the predicted mistag probability, η, corresponds to the actual measured mistag rate, ω. The calibration is performed on both control channels, one for OS taggers and the other one for SS Kaon. The calibrated mistag probability, ω, is determined in bins (precisely 10 bins in this analysis) of predicted η, using the known flavour from the kaon to determine the production flavour.

The relationship between η and ω is defined by a polynomial function. A polynomial of third order with a logit link function is employed in this study for both sides:

$$\text{logit}(\omega(\eta)) = p_0 + p_1(\eta - \bar{\eta}) + p_2(\eta - \bar{\eta})^2 + p_3(\eta - \bar{\eta})^3, \tag{8}$$

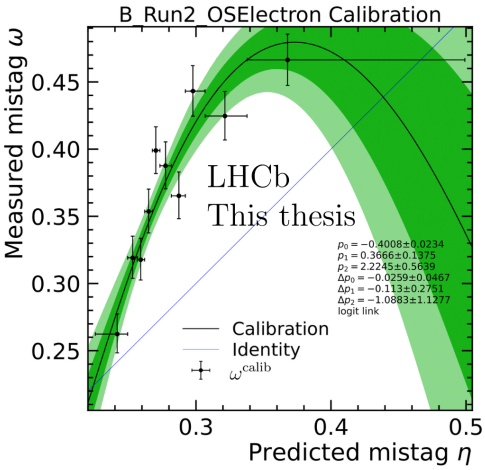
where $\bar{\eta}$ is the predicted sample mean mistag and p_i are fit-determined calibration parameters. The logit link guarantees that the calibrated mistag is physically within the range $(0, 1)$. This calibration is performed with the `FTCalib` package, which carries out this fit process via a binned maximum likelihood method, wherein the reweighting factors from the above section are incorporated.

During calibration, separate fits are performed for B and $\bar{B}$ candidates, as well. The differences between the corresponding polynomial coefficients are denoted as $\Delta p_i = p_i^{(B)} - p_i^{(\bar{B})}$, and are reported in calibration outputs together with the global parameters p_i. The Δp_i parameters capture potential tagging asymmetries, and their values are important for estimating systematic uncertainties and ensuring unbiased flavour tagging in time-dependent analyses.

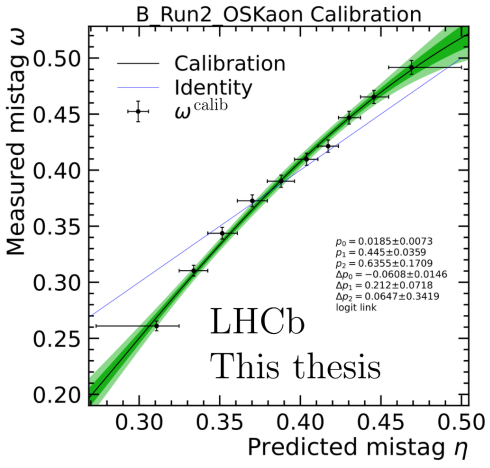
5.2 Calibration of the opposite-side tagging algorithms

The calibration of the Run 2 OS taggers applied to Run 3 data is made on the $B^+ \to J/\psi\, K^+$ channel, with the corresponding reweighting to $B_s^0 \to J/\psi\, \phi$ to emulate its kinematics. A calibration is made for each single OS tagger, then combined using the `FTCalib` algorithm as described in Section 4.3, and a final calibration is made to the taggers' combination. The resulting individual taggers' calibration curves are shown in Figure 5a, 5b, 5c and 5d, while the calibration curve of the combination can be seen in Figure 5e, where the improved precision of the predicted mistag is evident. The data points represent the measured mistag rates in bins of predicted η, while the fitted function (green band) captures the overall behaviour. Deviations from the identity line (blue) indicate the degree of miscalibration present in the raw tagger outputs of Run 2 when transferred to Run 3 conditions.

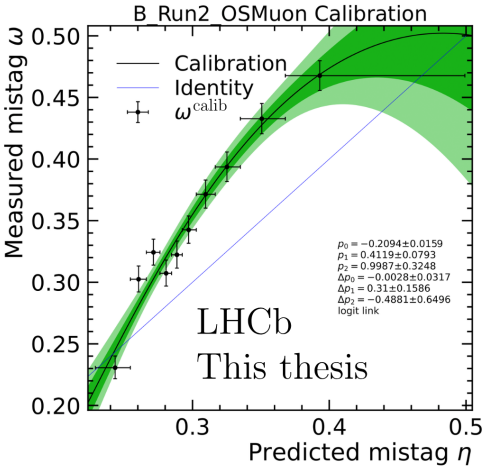
As seen in Figure 5, the individual OS taggers exhibit varying levels of performance. The combination algorithm takes this into account by assigning higher weights to taggers with higher discriminative power, such as the OS Kaon, in the final combined decision. The combination shows a reduction on the size of the green bands, meaning a more precise prediction of the mistag probability in the entire range.



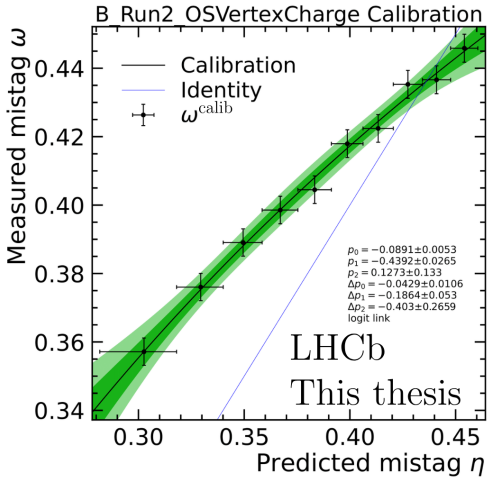
(a) Calibration curve for the Run 2 OS Electron tagger.



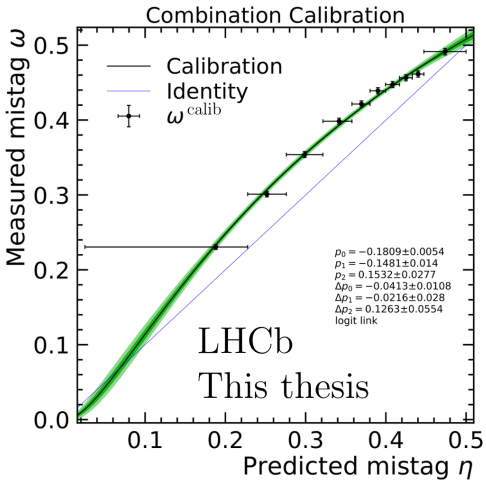
(b) Calibration curve for the Run 2 OS Kaon tagger.



(c) Calibration curve for the Run 2 OS Muon tagger.



(d) Calibration curve for the Run 2 OS Vertex Charge tagger.



(e) Calibration curve for the Run 2 combination of OS taggers.

Figure 5: Calibration curves for the Run 2 OS taggers applied to Run 3 data. The measured mistag ω is shown as a function of the predicted mistag η. The green band shows the uncertainty in the fitted calibration function, while the blue line indicates the identity line $\omega = \eta$.

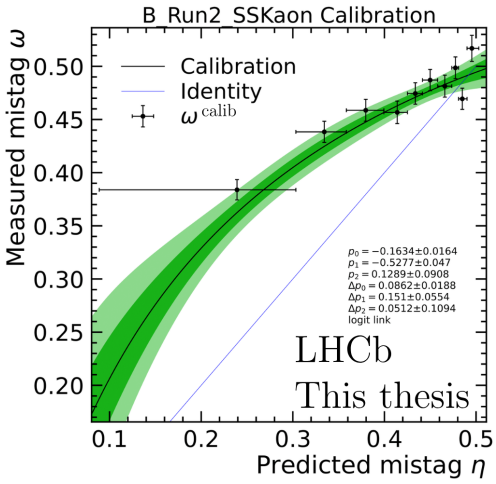


Figure 6: Calibration curve for the Run 2 SS Kaon tagger, using the $B_s^0 \to D_s^{\mp}\pi^{\pm}$ control channel. The measured mistag ω is extracted from a time-dependent fit to the oscillation pattern.

5.3 Calibration of the same-side Kaon tagging algorithm

The calibration of the SS Kaon tagger is performed using the $B_s^0 \to D_s^{\mp}\pi^{\pm}$ decay channel. Among the different SS taggers available (SS pion, SS proton), only the SS Kaon tagger is used in this analysis. This choice is motivated by the fact that this analysis is centred on OS taggers, and SS tagger are only used here to grasp an idea of the full power of the flavour tagging mechanism. Out of all SS taggers, the SS Kaon is the most relevant because kaons are the dominant tagging particle associated with B_s^0 production, and the other taggers either contribute marginally or are not fully commissioned for Run 3 at the time of this study.

Since the B_s^0 is a neutral meson, it undergoes flavour oscillations before decaying, which does not happen in the OS control channel. As a result, a time-dependent decay analysis is required. The `FTCalib` package takes that into account using the decay time of the B meson, computed by Decay Tree Fit, and fitting the flavour oscillation amplitude as a function of decay time. Since there is only on tagger in this SS tagging analysis, there is no combination involved.

The calibration is again performed using a third-order polynomial and a logit link function, following the same approach used for the OS taggers. The resulting calibration curve is shown in Figure 6. A considerable deviation from the identity line, particularly for low η, can be seen, reflecting a significant miscalibration of the uncorrected SS Kaon tagger. After calibration, the corrected mistag probability is consistent with the measured values across the full η range.

5.4 Tagging performance results

After the calibration procedure, the final tagging metrics explained in Section 4.3 are extracted for each individual tagger and their combination. These include the tagging efficiency ε_{tag}, the average mistag rate, $\bar{\omega}$, and the effective tagging power, ε_{eff}, which quantifies the statistical power of the tagger in time-dependent analyses and combines the other two metrics, giving it special relevance.

Table 2 summarises the performance of the calibrated Run 2 OS taggers when applied to Run 3 data, both individually and when combined. The combined tagger benefits from the complementary information provided by the sub-taggers, resulting in improved overall performance, as expected from the calibration curves. Also, the performance of the SS Kaon tagger calibrated on the $B_s^0 \to D_s^{\mp}\pi^{\pm}$ channel is reported in Table 2 as a separated row.

Overall, the combination tagger with OS has the best performance with an effective tagging power of approximately 7%. This is mainly caused by the sub-taggers' complementary nature. Among all single OS taggers, the SS Kaon tagger plays the most important role, whereas the Vertex Charge, efficient though it is, suffers from a relatively high mistag rate. Their calibration curves shown in Figure 5b and Figure 5d are also the cleanest and closest to the identity line. The independently trained SS Kaon

Tagger	ε_{tag} [%]	$\bar{\omega}$ [%]	ε_{eff} [%]
OS Electron	5.26 ± 0.06	$35.5 \pm 0.1\,(\text{stat}) \pm 0.5\,\text{syst}$	$0.443 \pm 0.006\,(\text{stat}) \pm 0.032\,\text{syst}$
OS Muon	12.33 ± 0.09	$33.1 \pm 0.04\,(\text{stat}) \pm 0.34\,\text{syst}$	$1.41 \pm 0.01\,(\text{stat}) \pm 0.06\,\text{syst}$
OS Vertex Charge	100.00 ± 0.00	$40.5 \pm 0.01\,(\text{stat}) \pm 0.13\,\text{syst}$	$3.65 \pm 0.00\,(\text{stat}) \pm 0.10\,\text{syst}$
OS Kaon	54.79 ± 0.13	$36.1 \pm 0.02\,(\text{stat}) \pm 0.17\,\text{syst}$	$4.23 \pm 0.02\,(\text{stat}) \pm 0.10\,\text{syst}$
OS Combination	97.69 ± 0.04	$36.6 \pm 0.02\,(\text{stat}) \pm 0.12\,\text{syst}$	$7.05 \pm 0.02\,(\text{stat}) \pm 0.12\,\text{syst}$
SS Kaon	73.75 ± 0.17	$44.2 \pm 0.03\,(\text{stat}) \pm 0.36\,\text{syst}$	$0.978 \pm 0.010\,(\text{stat}) \pm 0.123\,\text{syst}$

Table 2: Tagging performance of the Run 2 OS and SS Kaon taggers after calibration on Run 3 data. Shown are the tagging efficiency (ε_{tag}), average mistag rate ($\bar{\omega}$), and tagging power (ε_{eff}) for each individual tagger and their combination.

tagger on a different control channel performs worse, as one would expect from the intrinsic challenge of SS tagging and the higher mistag probability.

6 Development of meta-taggers

6.1 Motivation

As said in previous sections, the current combining algorithm in `FTCalib` takes a likelihood combination of the predicted mistag probabilities and decisions, merging the outputs of the sub-taggers. Although successful and popular, this procedure has some shortcomings: the combination is not data-learned, based on prechosen functional forms, and does not explicitly use potential correlations among taggers or nonlinear features that may enhance mistag probability prediction overall.

A different strategy is called *inclusive flavour tagging* [19, 20], where a machine learning model is trained from low- or high-level event data itself to predict the production flavour of the B meson, usually taking OS and SS information at once, allowing it to learn sophisticated relations across all the features that may contain information about the original flavour. It is very strong and promising, and a lot of efforts are currently brought together into this research field. But this method is in practice difficult to develop: the model needed is usually big, complicated, and hard to train. An approach that can partially benefit from the advantages of inclusive flavour tagging and avoid such complexity, and at the same time benefit from the great performance of current Run 2 taggers, is *meta-tagging*.

In meta-tagging, some sort of classifier is trained to infer the flavour of the B meson from inputs that come from the output of the current taggers i.e., their tagging assignments and mistag probability predictions. It provides a data-driven, flexible, and potentially more powerful combination compared to `FTCalib`, with a simple and interpretable structure. In this analysis, two different machine learning models are used for meta-tagging, and comparing their results: a BDT, and a Transformer model.

The BDT is specifically a gradient boosting classifier (XGBoost), since they are widely used in the LHCb context due to its capability to learn to classify complex but tabular data. The eight outputs of the taggers (four decision tags and four mistags probabilities, one for each sub-tagger), make this idea feasible. Additionally, the BDTs are really fast at training and evaluation.

The Transformer-based neural network is a more sophisticated model that was originally developed for natural language processing [21]. This model is potentially more powerful than the BDT thanks to its deep structure and attention mechanism. The simplified idea behind its usage is to treat the outputs of the sub-taggers as a sentence, where each decision tag and predicted mistag probability form a composed word, and the Transformer tries to learn the whole meaning of the flavour, which represent the true flavour. This model is much slower at training while fast at evaluation, and its flexibility in the input length allows the usage of extra features (like kinematics of the event) to deal with cases where the tagger outputs by themselves are not enough to decide on the true flavour.

6.2 Meta-taggers: training and structure

The inputs to the meta-taggers are, for each tagger, the tagging decision and the predicted mistag probability as tuples to allow more expressiveness of the models in learning the relation between the predicted mistag probabilities of the different sub-taggers. Apart from that, in the Transformer model case, the extra features used are: the B meson transverse momentum ($p_T(B)$), the pseudorapidity ($\eta(B)$) and the number of long tracks (`NLongTracks`).

The full set of hyperparameters and loss functions of both models can be found in the Appendix A to ensure reproducibility. For the BDT, a search of over 900 combination of hyperparameters is made to maximise its performance, while this is infeasible with the Transformer. Here, the hyperparameters are adjusted manually depending on the results.

In simple words, to train a supervised machine learning model as the ones described, it is necessary to have the ground truth or true labels for each event, so the model can learn from it. Therefore, the models could not be trained in the $B_s^0 \to J/\psi\,\phi$ decay due to the lack of information about the true flavour. That would only be possible in a simulation, which was not available for this analysis. Therefore, the models are trained in the $B^+ \to J/\psi K^+$ channel, and calibrated to the signal channel with the same procedure as for the sub-taggers. But to ensure fair comparison, it is necessary to evaluate the performance of the meta-taggers on the same data as the classic combination.

Thus, the data samples of the $B^+ \to J/\psi K^+$ decay channel are split in two, 50% for training and 50% for evaluation, leaving a total of about 160,000 events for each part. The training of the BDT typically

takes less than 5 minutes on multiple CPUs, while the Transformer can take between 3 and 7 hours on a single GPU, depending on the early stopping due to no improvement. For the evaluation, the BDT takes less than a minute on a single CPU, whereas the Transformer takes less than 2 minutes to predict on a single CPU. After applying the models on half of the data and measuring their performance, the sWeights are extracted, the same kinematic variables described in Section 4.5 are reweighted, and the individual calibration to the signal channel of both meta-taggers is performed, as described in Section 5. The same process is done all over again for the sub-taggers and the classic combination algorithm, but using only the half of the data that the models are evaluated on.

For the Transformer, it is relevant to mention early signs of overfitting, due to no improvement on evaluation while the training loss drops significantly in the first 5 epoh and diminishes slowly from there. This is stopped by the early stopping, that typically breaks the training after the first 5 to 20 epochs. This is not surprising since the amount of data for training is relatively low to fully train a complex model like a Transformer. This suggests that model capacity can be bigger with more extensive datasets.

It is also important to mention that, since the models are trained on recorded data, there can be a mix of background and signal samples despite the offline selection and the BDT cut mentioned in Section 4.2. Thus, the training of both the BDT and Transformer meta-taggers are performed using the control channel, with sWeights applied to statistically subtract the background contribution. The XGBoost allows a special input for weighting the data, but the negative sWeights have to be fixed to a small number[6] in order to do so. For the Transformer, during training, the clipped sWeights are used as per-event loss weights to prioritize signal-like events by scaling each example's contribution to the loss function, ensuring the model focuses more on learning from signal-rich regions of the data.

6.3 Results of the meta-taggers

The output score of each classifier is a number in $(-\infty, \infty)$, and after applying the sigmoid function, it is mapped to $(0, 1)$, representing the conditional probability $P(\text{flavour} = +1 \,|\, \text{inputs})$, meaning that flavour is 1 when the meson is a B^+ and 0 when it is a B^-. Then, a small undecided band with half-width of 0.015 (maintains high tagging efficiency while decreases the mistag rate) is set for getting the decision tag. If the score of an event is in $[0, 0.485]$, the event is tagged as -1 and the predicted mistag probability is directly the score. If the score is in $[0.515, 1]$, the decision tag is set to 1 and the predicted mistag probability is set to $1 - \text{score}$. If the score is in the undecided band $(0.485, 0.515)$, the decision tag is set to 0 and the predicted mistag probability to 0.5. With this as output of the meta-tagger, it is calibrated using the `FTCalib` with polynomial (order 3) and logit link for consistency.

The calibration curves for both the XGBoost and Transformer meta-taggers, together with the distribution of the scores of both models, can be found in Appendix B, Figures 10 and Figure 8 respectively. The mass fits and the reweighting validation plots of the data used for the validation (half of the one used for the previous analysis) of the models can also be found in Appendix B, Figure 7, to avoid redundancy. The tagging metrics of the models can be seen in Table 3, together with the metrics of the `FTCalib` combination for the data used in this second analysis.

Tagger	ε_{tag} [%]	$\bar{\omega}$ [%]	ε_{eff} [%]
`FTCalib` Combination	96.69 ± 0.07	$36.57 \pm 0.03\,(\text{stat}) \pm 0.17\,\text{syst}$	$6.97 \pm 0.03\,(\text{stat}) \pm 0.18\,\text{syst}$
BDT	92.59 ± 0.07	$36.17 \pm 0.02\,(\text{stat}) \pm 0.12\,\text{syst}$	$7.08 \pm 0.02\,(\text{stat}) \pm 0.13\,\text{syst}$
Transformer	93.28 ± 0.07	$36.33 \pm 0.02\,(\text{stat}) \pm 0.12\,\text{syst}$	$6.97 \pm 0.02\,(\text{stat}) \pm 0.13\,\text{syst}$

Table 3: Tagging performance comparison between the meta-taggers and the `FTCalib` combination.

Both models exhibit similar performance to the `FTCalib` combination algorithm, being the BDT slightly better. The tagging efficiency of the meta-taggers is lower than the one from the combination, probably due to the undecided band, but at the same time, the models show lower mistag rate, likely due to the same reason. It is worth to mention that the tagging power of the `FTCalib` combination is lower in this comparison because only half of the data is used.

[6] This is reasonable, as negative sWeights arise when the PDF exceeds that of the signal. By replacing these negative values with $1 \cdot 10^{-6}$, events predominantly associated with the background are effectively excluded from the training process, thereby preventing them from biasing the model. This has actually not a great impact since only 15.81% of the sWeights are negative.

It is also important to mention that the performances of the two models in the control channel (before calibration) differ from their performance in the signal channel (after calibration). The BDT has an AUC[7] value of 0.6322 while the Transformer has AUC = 0.6346 (both values are sWeight-corrected), so the Transformer actually performed slightly better before calibration. This flipped behaviour is probably caused because the Transformer learns better the data it is trained on (overfits) and lacks generalisation, while the BDT does not.

Further investigation is carried out with the Transformer. It is trained using more extra features as inputs. Specifically, the three that are already being used ($P_T(B)$, $\eta(B)$ and`NLongTracks`) and the number of reconstructed primary vertices (N_{PV}), the decay time of the B meson (t_B), the χ^2 per degree of freedom of the B decay vertex fit (χ^2_{vtx}/ndf), the impact parameter χ^2 of the B meson ($\chi^2_{\mathrm{IP},B}$) and the invariant mass of the B meson (m_B). The results of the new retraining show a better performance on the control channel, suggesting that using extra features helps the model to deal with events where the sub-taggers are unclear. But on the signal channel, the performance is worse, probably due to the fact that these extra features are inherent to the control channel, and the relations the model learns cannot be completely extrapolated to the signal channel through the reweighting and calibration.

[7]The AUC is the Area Under the ROC Curve computed on the validation set, a metric to measure the model's ability to distinguish between classes. AUC = 1.0 implies the perfect classifier, while AUC = 0.5 is just random guessing. Said ROC curves for both models can be found in Appendix B, Figure 9.

7 Discussion and Future Directions

The results show that the `FTCalib` combination, Transformer meta-tagger, and BDT meta-tagger achieve very similar performance, with tagging powers of 6.97%, 6.97%, and 7.08% respectively. Interestingly, the BDT-based meta-tagger slightly outperforms both the Transformer and the official combination in terms of tagging power, despite its simpler architecture. This suggests that a data-driven combination can effectively capture the relevant correlations between taggers and, in some cases, outperform even more complex models when properly tuned and calibrated, although, as discussed, a larger dataset could potentially give greater advantage to the Transformer model.

One of the most significant strengths of the meta-tagging approach is that it is scalable and flexible. In contrast to `FTCalib`, where one works with predefined functional forms and have limited power in modelling correlations, the meta-taggers can learn high-order nonlinear dependencies among tagger outputs and model subtle correlations. This makes them very appealing in anticipation of future expansion, where the number and variety of tags are to be expanded. Following this line of thought, training the Transformer on an extensive $B_s^0 \to J/\psi\,\phi$ simulated dataset with a different set of extra features could be also be a direction to explore.

The models' practicality is also relevant for the overall picture. The `FTCalib` algorithm is already implemented in the package and does not require any type of training. The BDT is easy to implement through existing libraries, it does not take long for training time and is CPU-based. Finally, the Transformer has a complex structure and preprocessing, requires time for training and preferably a GPU, although this increased workload only has to be done once. When it comes to applying or evaluating, all three models are very similar.

One limitation of the current work is the use of control channels for training and validation. Although reweighting removes some kinematic mismatches, there can still be residual biases. Another limitation factor is time. Some promising investigation directions remain unexplored, and can potentially uncover better performances. In future research, an obvious extension is to pursue semi-supervised or weakly supervised learning methods that make more direct use of signal data.

Additionally, the promise of combined OS+SS meta-tagging, in which both groups of taggers are merged within a single model, has been limitedly investigated and may bring about additional performance gains.

Lastly, the concept of inclusive flavour tagging is still intellectually attractive, yet computationally and systematically demanding. Further studies along these lines, perhaps with hybrid models or attention-guided mechanisms, may provide a way towards a unified and very expressive tagging system for LHCb.

8 Conclusions

In this thesis, a thorough investigation and characterisation of flavour tagging performance at LHCb Run 3, with special focus on calibration and combination of Run 2 taggers to Run 3, is presented. The study is mainly driven by focusing on OS taggers, although SS taggers were included for completeness, and calibration via control channels and sWeight-based background subtraction was performed. A kinematic reweighting scheme was also implemented to improve compatibility with the signal channel.

The official `FTCalib` combination shows a tagging power of 7.05% for the tuned performance of the Run 2 OS taggers on Run 3 data, which is stable and robust. SS Kaon tagger was independently tuned with $B_s^0 \to D_s^{\mp}\pi^{\pm}$ decay, achieving a tagging power of 0.98%. These results confirm that Run 2 taggers are still able to provide good inputs for flavour tagging in Run 3 conditions, provided that appropriate calibration is used.

In addition, two machine learning-based meta-taggers, a gradient boosted decision tree and a Transformer neural network, were trained to combine the OS taggers in a data-driven manner. The BDT, after calibration, achieves a tagging power of 7.08%, outperformed both `FTCalib` and the Transformer, which are tied with a tagging power 6.97% over the evaluation dataset. These results show a moderate gain in performance due to the limitations exposed through the whole analysis. Even so, the increased performance indicates that meta-tagging is an adaptable and effective alternative to the standard combination algorithm, which can match or surpass performance when finely tuned.

The meta-tagger strategy shows a promising direction for future improvements in inclusive or hybrid flavour tagging. It provides a flexible structure that can be utilized to incorporate new taggers, additional features, or even joint OS+SS information without requiring radical alteration of the fundamental combination procedure. The potential for additional improvements lies as more Run 3 data are incorporated and model training can be applied to bigger and more representative samples.

A Meta-tagger hyperparameters

XGBoost parameters

The following hyperparameters were used for training the XGBoost meta-tagger:

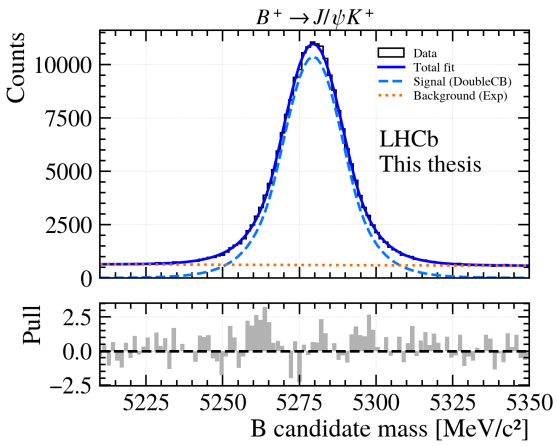
- `n_estimators` = 1200
- `learning_rate` = 0.005
- `max_depth` = 3
- `min_child_weight` = 5
- `subsample` = 0.6
- `colsample_bytree` = 0.8
- `reg_lambda` = 0.0
- `reg_alpha` = 1.0
- `eval_metric` = `"logloss"`
- `n_jobs` = -1 (use as many CPUs as possible)
- `random_state` = 42

Transformer parameters

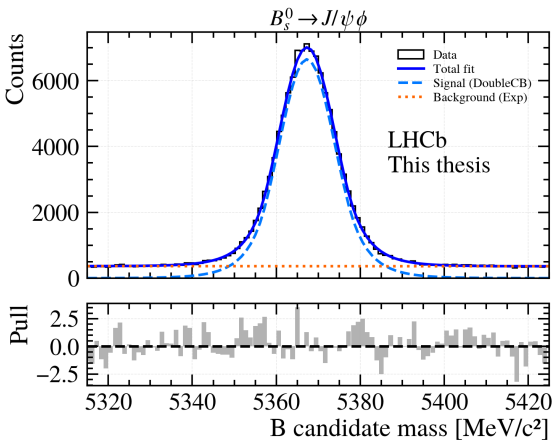
The Transformer meta-tagger was trained with the following configuration:

- 12 layers
- 16 attention heads
- Embedding dimension: 256
- Dropout: 0.15
- Batch size: 2048
- Optimizer: AdamW
- Learning rate scheduler: One-cycle policy
- Loss function: Binary cross entropy with logits
- Early stopping after convergence (typically within 20 epochs)

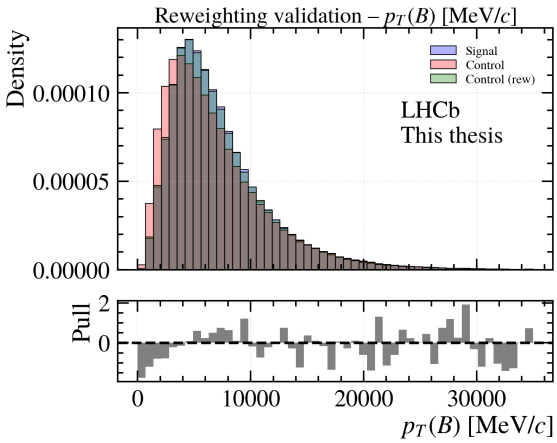
B Supplementary figures for the meta-tagging algorithms



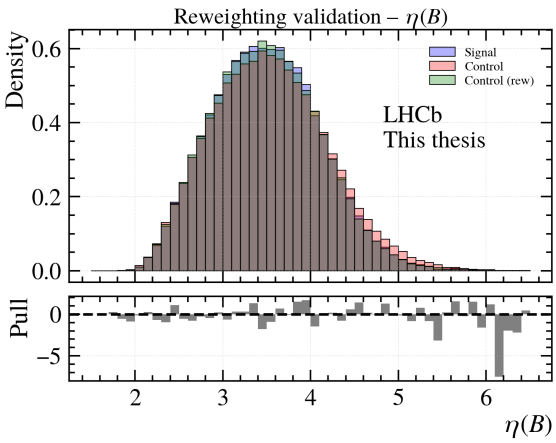
(a) Mass fit to the $B^+ \to J/\psi\, K^+$.



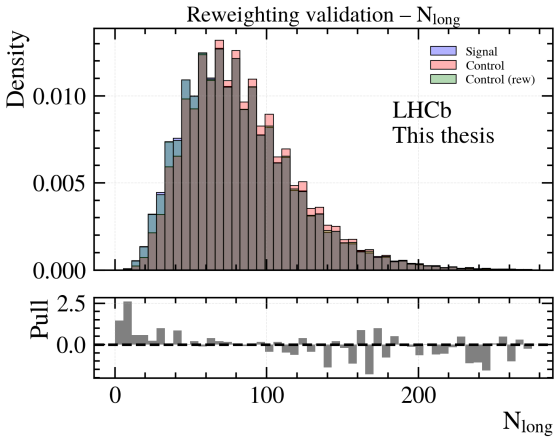
(b) Mass fit to the $B_s^0 \to J/\psi\, \phi$.



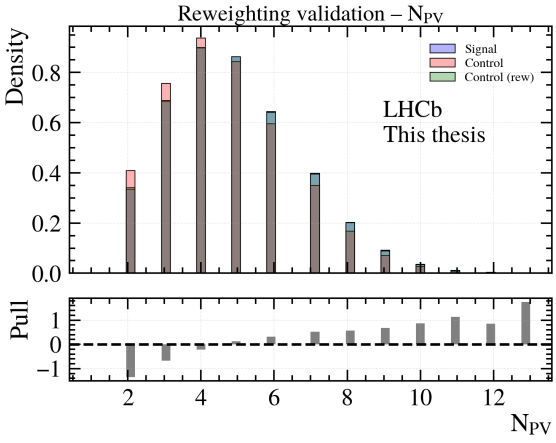
(c) Reweighting on $p_T(B)$.



(d) Reweighting on $\eta(B)$.



(e) Reweighting on `NLongTracks`.



(f) Reweighting on `NPVs`.

Figure 7: Top row: invariant mass fits used to extract signal and background components for sWeight computation. Middle and bottom rows: validation of the reweighting procedure across the four input variables used to match the control and signal channels. For all these plots, only half of the data was used, the one used to evaluate the meta-taggers performance.

Analysis before evaluation

In the Figure 7, the mass fits for the extraction of the sWeights on both the signal and control channel and the validation plots of the kinematic reweighting are shown. To maintain consistency for the comparison between the meta-taggers and the `FTCalib` algorithm performance, the whole analysis was repeated on half of the original dataset, in those sample where the evaluation of the meta-taggers was done. There is no appreciable change between the plots in Figure 7 and the ones in Figure 2 and Figure 3.

Performance visualization of the meta-taggers in the control channel

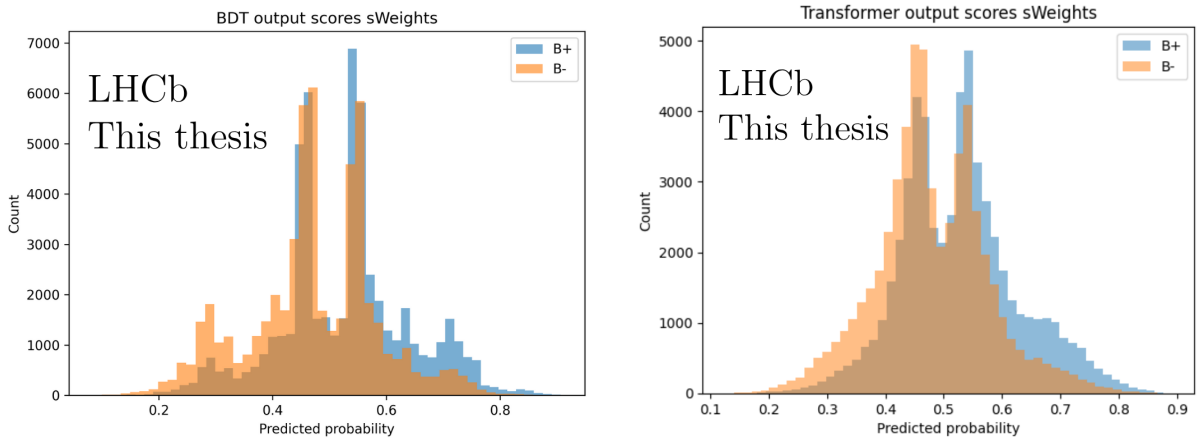
The meta-taggers are trained with the outputs of the sub-taggers applied to the control channel $B^+ \to J/\psi\, K^+$. Thus, the output or scores of the meta-taggers are optimized for this channel. In Figure **??**, the distribution of the output scores of both the BDT and Transformer are plotted, weighting each event with the corresponding sWeight, i.e, suppressing the background. Since in this channel the true flavour is known, this allows a differentiation in the plots between the true flavours of the scores.

The histograms of Figure 8 show, as one would expect, that B^+ and B^- events congregate at high and low predicted values, respectively. But it is also shown a big overlap between distributions, specially close to the 0.5 region (expected as well). This explains the high mistag rates (and low tagging power) of the flavour tagging mechanisms, underscoring the complexity of the task.

The BDT score distribution is steeper in peaks, meaning the model is inclined to make strongly confident predictions where there is clear separation or splitting. This has a cost, however, since the distribution is less smooth with sudden changes that could be constraining generalisation or calibration robustness. The Transformer output is continuous and more symmetric by comparison, showing that the model learns a softer boundary among the classes. This makes sense because this behaviour can lead to more robust and smoother calibration even at slightly worse tagging power. These plots illustrate how each model qualitatively conveys uncertainty and class separation.

To reflect quantitatively how every model conveys class separation, Figure 9 shows the ROC curve of both the BDT and the Transformer models' outputs, corrected with the sWeights.

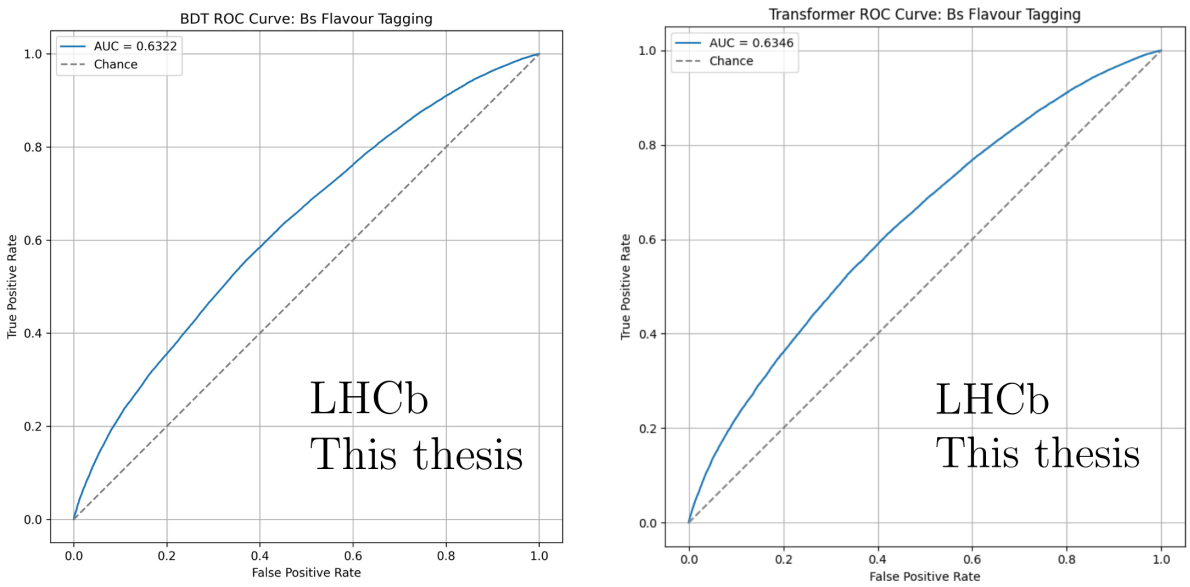
Both models perform significantly above chance (AUC = 0.5), around 0.63. The Transformer does slightly better than the BDT on AUC (0.6346 vs. 0.6322) by a minimal amount and quite possibly not translating to ultimately tagging capability. These results are consistent with the empirical experience that both models are capable of learning the tagging pattern adequately, and fine-grained variation in performance depends less on raw score distinction and more on calibration and mistag control.



(a) BDT meta-tagger output scores.

(b) Transformer meta-tagger output scores.

Figure 8: Distributions of predicted output scores (sWeight-corrected) from the BDT and Transformer meta-taggers for the B^+ and B^- classes.



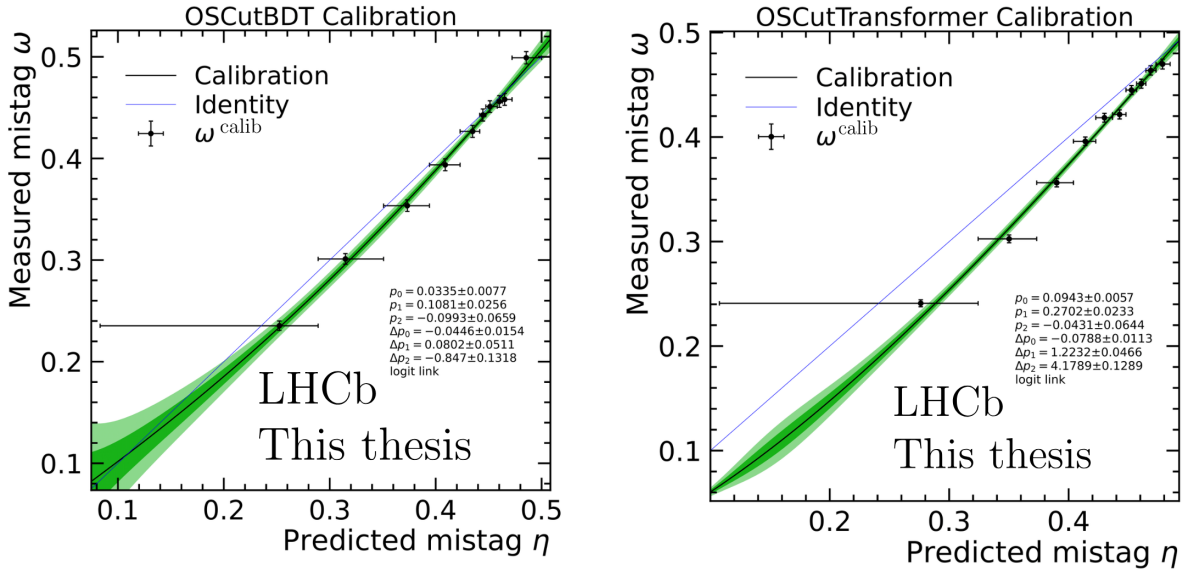
(a) ROC curve of the BDT output scores.

(b) ROC curve of the Transformer output scores.

Figure 9: Receiver operating characteristic (ROC) curves (sWeight-corrected) for the BDT and Transformer meta-taggers, trained to distinguish between B^+ and B^- candidates using tagging information. The area under the curve (AUC) quantifies the model's ability to separate the two classes.

Calibration curves of the meta-taggers

After all the training and preparation of the data, the calibration of the meta-taggers from the control channel $B^+ \to J/\psi K^+$ to the signal $B_s^0 \to J/\psi \phi$ can be done using the `FTCalib` package, with the same settings as the sub-taggers described in Section 5. The resulting calibration curves for the BDT and the Transformer are shown in Figure 10. It is notable to mention that the names *OSCutBDT* and *OSCutTransformer* mean opposite-tagger (trained only on OS sub-taggers) with undecided band cut (decision tag set to 0 on scores in $(0.485, 0.515)$).



(a) Calibration curve of the BDT meta-tagger.

(b) Calibration curve of the Transformer meta-tagger.

Figure 10: Measured mistag rate ω as a function of predicted mistag η for the two meta-taggers. The green band shows the uncertainty on the fitted calibration model, while the blue line represents the ideal $\omega = \eta$ identity.

The BDT calibration lies very close to the identity line across all of the range of predicted mistag values, and it has very small residuals and a well-behaved polynomial fit. This is consistent with the model's tendency to produce tightly separated score distributions and its ability to make good estimates of uncertainties, and endorses the fact that the BDT has the highest tagging power.

In contrast, the Transformer calibration curve deviates slightly from the identity line, particularly for low η values, where the mistag rate is underpredicted (similarly to the `FTCalib` algorithm, but on the opposite side of the identity line). This is also evident from the steeper slope and larger Δp_2 coefficient. Nevertheless, the overall agreement is not that bad, and the Transformer benefits from smoother score distributions as well as improved generalisation in regions of poorer statistics. In general, but specially in lower η, the Transformer achieves more precision (smaller green bands) than the BDT.

Both models are adequately calibrated following a logit-transformed third-order polynomial fit so that their outputs may be used safely in subsequent physics analyses.

C Selection on opposite-side Vertex Charge mistag probability prediction

As explained briefly in Section 4.2, an extra selection had to be made post-analysis due to inexplicable behaviour. This means that, initially, this selection was not applied, and only after the whole calibration was done, the outlier was spotted. In Figure 11, the calibration curve for the OS Vertex Charge can be seen, together with a red circle drawn since the outlier lies hidden among the legend.

The strange behaviour seen in Figure 11 was attributed to the abnormally repeated values on the distribution of the predicted mistag probability of the OS Vertex Charge. Specifically, the value $\eta = 0.3250$ was repeated $37,984$ times, $\eta = 0.4141$ was repeated $32,618$ times and $\eta = 0.3500$ was repeated $5,832$ times, while any other value had no more than 10 repetitions, as can be seen in Figure 12 (it is important to notice that both distributions are normalised to one, so the spikes do not match with the values just explained).

No explanation was found for this pattern, but the behaviour is more likely due to technical or procedural factors rather than physical reasons. Either way, the safe solution was employed, and those data samples were simply removed from the dataset as can be seen in Figure 12b, fixing the abnormal behaviour.

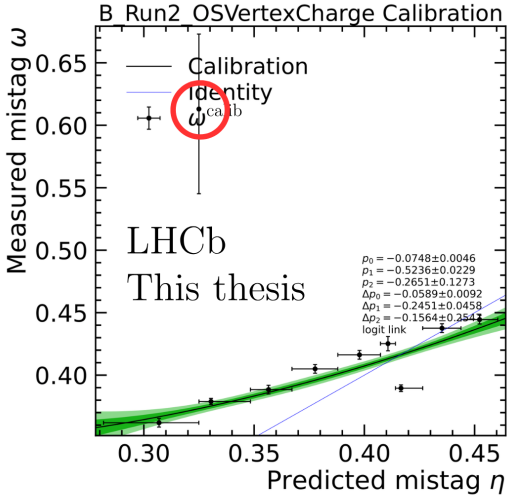
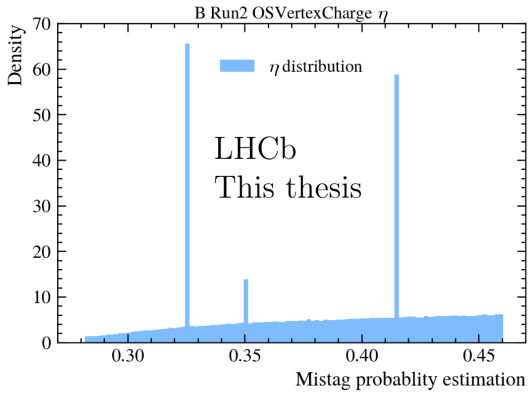
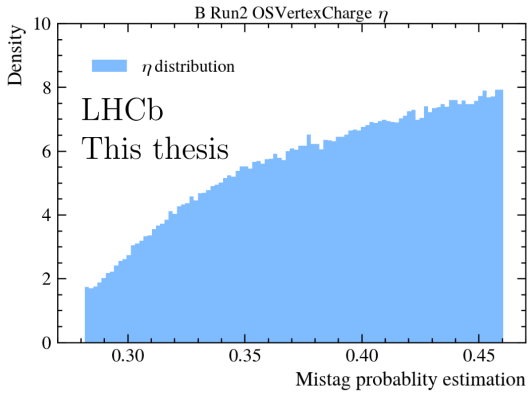


Figure 11: Calibration curve for the Run 2 OS Vertex Charge tagger applied to Run 3 data. A red circle is drawn to highlight the strange behaviour.



(a) Distribution of the η for the OS Vertex Charge before the cut.



(b) Distribution of the η for the OS Vertex Charge after the cut.

Figure 12: Comparison on the η distribution for the OS Vertex Charge tagger before and after the cut. It is important to notice that both histograms are normalised.

References

[1] M. Kobayashi and T. Maskawa. "CP-Violation in the Renormalizable Theory of Weak Interaction". In: *Prog. Theor. Phys.* 49 (1973), p. 652.

[2] N. Cabibbo. "Unitary Symmetry and Leptonic Decays". In: *Phys. Rev. Lett.* 10 (1963), p. 531.

[3] LHCb Collaboration. "Measurement of the B_s^0–$\bar{B}_s^0$ oscillation frequency Δm_s in $B_s^0 \to D_s^-(3)\pi$ decays". In: *Phys. Lett. B* (2011). CERN-PH-EP-2011-194, LHCb-PAPER-2011-010. arXiv: 1112.4311 [hep-ex].

[4] LHCb Collaboration. "Measurement of the CP violating phase ϕ_s in $B_s^0 \to J/\psi f_0(980)$". In: *Phys. Lett. B* (2011). CERN-PH-EP-2011-205, LHCb-PAPER-2011-031. arXiv: 1112.3056 [hep-ex].

[5] LHCb Collaboration. "Measurement of the CP-violating phase ϕ_s in the decay $B_s^0 \to J/\psi\phi$". In: *Phys. Rev. Lett.* arXiv:1112.3183 (2011). CERN-PH-EP-2011-214, LHCb-PAPER-2011-021. arXiv: 1112.3183 [hep-ex].

[6] Penelope Hoffmann. *Precise Determination of the B_s^0–$\bar{B}_s^0$ Oscillation Frequency Δm_s at LHCb.* Seminar Talk. Jan. 25, 2024. URL: https://www.physi.uni-heidelberg.de/.

[7] LHCb Collaboration. "B flavour tagging using charm decays at the LHCb experiment". In: *JINST* 10.10 (2015), P10005. DOI: 10.1088/1748-0221/10/10/P10005. arXiv: 1507.07892 [hep-ex].

[8] LHCb Collaboration. *FTCalib Documentation*. 2024. URL: https://lhcb-ftcalib.readthedocs.io/en/latest/combination.html (visited on 05/10/2025).

[9] Julian Tarek Wishahi. "Measurement of CP Violation in $B^0 \to J/\psi K_S^0$ Decays with the LHCb Experiment". PhD thesis. Dortmund University, 2014. URL: https://repository.cern/record/eas1p-c7p51.

[10] LHCb Collaboration. "Opposite-side flavour tagging of B mesons at the LHCb experiment". In: *Eur. Phys. J. C* 72 (2012), p. 2022. DOI: 10.1140/epjc/s10052-012-2022-1. arXiv: 1202.4979 [hep-ex].

[11] LHCb Collaboration. "The LHCb Detector at the LHC". In: *JINST* 3.08 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005.

[12] The LHC experiments Committee. *LHCb Trigger and Online Upgrade Technical Design Report*. Technical Design Report CERN-LHCC-2014-016, LHCB-TDR-016. CERN, 2014. DOI: 10.17181/CERN.5F5X.FDJM. URL: https://cds.cern.ch/record/1701361.

[13] LHCb Collaboration. *Time-dependent analysis of $B_s^0 \to J/\psi\phi$ decays*. Internal LHCb document. LHCb-ANA-2011-059, Internal Analysis Note. 2011.

[14] Claudia Marie Schudy. "Measurement of the B^+ lifetime in $B^+ \to J/\psi K^+$ decays at LHCb using 2024 data". Physikalisches Institut, Heidelberg. Bachelor's thesis. Heidelberg University, 2025.

[15] Zfit developers. *zfit.pdf.DoubleCB — Double Crystal Ball PDF. zfit Documentation (version 0.6.4)*. Accessed: 2025-05-10. 2023. URL: https://zfit.readthedocs.io/en/0.6.4/user_api/_generated/pdf/zfit.pdf.DoubleCB.html.

[16] Jonas Eschle et al. "zfit: scalable pythonic fitting". In: *Journal of Open Source Software* 5.46 (2020), p. 1954. DOI: 10.21105/joss.01954. URL: https://joss.theoj.org/papers/10.21105/joss.01954.

[17] M. Pivk and F. R. Le Diberder. "sPlot: A statistical tool to unfold data distributions". In: *Nuclear Instruments and Methods in Physics Research Section A* 555.1-2 (2005), pp. 356–369.

[18] Alex Rogozhnikov et al. "New approaches for boosting to uniformity". In: *JINST* 10.03 (2015), T03002. DOI: 10.1088/1748-0221/10/03/T03002. arXiv: 1410.4140 [physics.data-an].

[19] Claire Prouve, Niklas Nolte, and Christoph Hasse. "Fast Inclusive Flavour Tagging at LHCb". In: (2024). DOI: 10.48550/arXiv.2404.14145. arXiv: 2404.14145 [hep-ex]. URL: https://arxiv.org/abs/2404.14145.

[20] Quentin Führing. "Decay-time-dependent studies of strange beauty mesons". Dissertation. Technische Universität Dortmund, 2023. DOI: 10.17877/DE290R-24081. URL: https://doi.org/10.17877/DE290R-24081.

[21] Ashish Vaswani et al. "Attention Is All You Need". In: *arXiv preprint arXiv:1706.03762* (2017). DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

DECLARACIÓ D'EXTENSIÓ DEL TREBALL DE GRAU

Jo, Roger Feliu Vert, amb Document Nacional de Identitat 47741736S, i estudiant del Grau en Física de la Universitat Autònoma de Barcelona, en relació amb la memòria del treball de final de Grau presentada per a la seva defensa i avaluació durant la convocatòria de Juliol del curs 2024-2025, declara que:

- El nombre total de paraules incloses en les seccions des de la introducció a les conclusions és de 7488 paraules.
- El nombre total de figures és de 6, de les quals 1 té 2 subfigures, 2 tenen 4 subfigures i 1 té 5 subfigures.
- El nombre total de línies de formules és de 9.
- El nombre total de línies de taules és de 16.

En total el document, comptabilitza:

7488 paraules + 2 x 200 paraules per figura + 1 x 200 paraules per figura amb dues subfigures + 2 x 400 paraules per figura amb 4 subfigures + 1 x 600 paraules per figura amb 5 subfigures + 25 x 20 paraules per línia de fórmula o taula = 9988

Que compleix amb la normativa al ser inferior a 10000.

Signat: