THE UNIVERSITY *of* EDINBURGH
School of Physics
and Astronomy

Searching for the $X(3823)$ resonance in B meson decays

MPhys Project Report

N. McBlane

October 3, 2020

Abstract

We find optimistic signs of a resonance around the expected mass of the $X(3823)$ at a statistically insignificant 2.4σ in the decay $B^+ \to J/\psi\pi^+\pi^-K^+$, but propose an sWeighting method to improve this. This measurement is made from 6.5fb^{-1} of proton-proton collisions collected by LHCb, roughly 85% of the available data. We estimate a branching fraction of $\mathcal{B}r\big(B^+ \to X(3823)K^+\big) \times \mathcal{B}r\big(X(3823) \to J/\psi\pi^+\pi^-\big) = (2.3 \pm 0.9 \pm 0.3) \times 10^{-7}$ where the first error is statistical and the second systematic. In addition we place a 90% confidence limit at 3.6×10^{-7}.

Supervisors: Dr G. Cowan, Dr M. Needham

Personal statement

The first month or so of the project was spent getting to grips with the libraries that would be used throughout the analysis, predominantly ROOT. I had some issues running these on my own machine but was able to use the PPE computers instead. I had some experience in ROOT from previous projects, but not to the same degree of data exploration and not in its C++ implementation. This period was therefore invaluable in developing an understanding of how the data was stored, what various variables corresponded to, etc. I also used this time to compare the performance of various Multivariate Analysis methods using the ROOT TMVA libraries. As my programming experience was predominantly in Python, I made efforts to progress in C++ by following the Accelerated C++ textbook.

The next two months were spent developing and implementing an initial multivariate selection model using XGBoost. I began with a script used for prior analyses, and developed it to include a command line interface for making simple changes, to produce various plots and log files, and a simple implementation of cross-validation. All of this was useful in assessing performance of the model. Much of this time was spent playing with model parameters, selecting and engineering discriminating variables and determining the best method for estimating efficiencies from which the optimal threshold probability could be determined. In this time I gave my first presentation to the Edinburgh LHCb group, which was a valuable experience.

During this period we also explored the use of RapidSim to attempt to identify potential sources of error in the data due to misidentifications and irreducible backgrounds.

The final week of first semester, then the first few weeks of second semester, were spent developing a resolution model for fitting to the $X(3823)$ resonance. The Christmas break was used for exam and practice presentation preparation. This was the first time I had to develop a relatively complex C++ macro for the analysis so initial progress was slow. Various models were attempted before selecting the double Crystal Ball.

Prior to un-blinding the analysis by studying the $X(3823)$ region of the data in detail, a final further optimisation of the XGBoost classifier hyperparameters was performed. I spent much of my time attempting to use the Spearmint optimisation libraries, but was ultimately unsuccessful. I was pointed towards Skopt, a Python package with which I was able to perform this optimisation with ease. Ultimately, variation of the hyperparameters in the range studied had little effect on classifier performance.

The final months of the project were spent analysing the $X(3823)$ region, and tidying up the rest of the project for the report and presentation. I took some time to get to grips with the likelihood estimation methods for measuring significance and placing confidence limits on the branching fraction. Once I managed this though, I was able to produce results with which I am reasonably happy. I have a solid understanding of where the analysis should be taken next, and I hope to be able to pursue this once exams are complete.

Acknowledgments

I would like to thank Dr Greig Cowan and Dr Matt Needham for their attentive and invaluable support throughout this project. I would also like to thank Marion Lehuraux for her initial help in running XGBoost, and for the Python script from which my analysis was developed. Finally I would like to thank Dr Marco Pappagallo for attending both of my presentations, and for his feedback on each.

Contents

1 Introduction

The Standard Model (Fig.1) is familiar across all levels of physics as the foundation for what makes up everything we observe in matter and the forces of interaction between this matter. The current paradigm has matter taking two forms: bosons and mesons. Bosons consist of a triplet of quarks, for example the proton (uud) and the neutron (udd). Mesons are made up of a quark and an anti-quark, such as the B^+ meson ($u\bar{b}$) and π^- meson (pion, $d\bar{u}$).

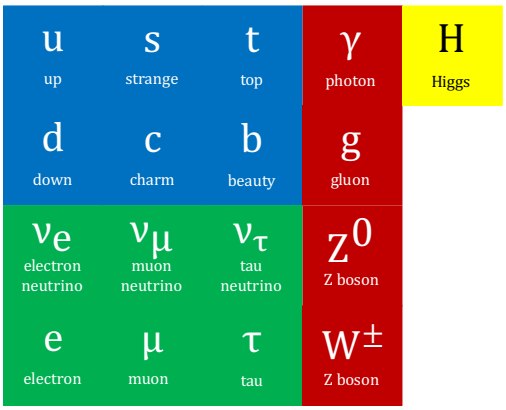


Figure 1. The standard model: quarks (blue) and leptons (green) make up the fermions which form matter. Bosons (red and yellow) mediate the forces of interaction between matter.

Even since the advent of the quark model [1] however, particles containing four quarks (tetraquarks, $qq\bar{q}\bar{q}$), five quarks (pentaquarks, $qqqq\bar{q}$) and so on were proposed. In addition, even more complex structures consisting entirely of gluons [2], so called hybrid mesons [3] and mesonic molecules [4] (analogous to the bosonic molecules which make up the familiar nuclei of atoms) have since been added to the realm of possible particles - with the term "exotic meson" coined to describe those which extend the standard meson picture.

A recent flurry of evidence for exotic meson candidates [5] has reignited interest in the subject. The bulk of this has been associated with masses around the charmonium sector: the area of particle physics concerning the J/ψ meson ($c\bar{c}$) and its various excited energy states. Many of these particles have masses which place them within the sector, but quantum numbers which conflict with what we can construct from the normal meson picture. As such, study of charmonium and so called charmonium-like states (i.e. exotic candidates in the sector) represents an exciting point for the development of particle physics as a whole.

The $X(3823)$ is a resonance which sits well within the charmonium sector. It is expected to be a very normal particle: simply an excited state of the J/ψ meson. Evidence for its existence was first found by Belle in 2013, in the decay $B^+ \to \chi_{c1}\gamma K^+$ [6], and in this paper we present optimistic signs of its presence in the decay $B^+ \to J/\psi\pi^+\pi^- K^+$.

Despite not being an exotic meson candidate, the $X(3823)$ is important to the search for an explanation. A better understanding of the charmonium sector as a whole will allow us to most effectively discern how and where deviations from the standard meson picture occur.

2 Backgrounds

2.1 Charmonium spectroscopy

As mentioned prior, charmonium spectroscopy concerns the study of the J/ψ meson ($c\bar{c}$) and its various excited resonances. In many ways these resonances are analogous to how atoms form excited states and can be predicted with relative simplicity. Assuming the quarks to be sufficiently heavy to approximate as non-relativistic, the problem can be treated as a solution to the Schrödinger equation with the Cornell potential [7] for a two-body charged system:

$$V(r) = -\frac{\kappa}{r} + \frac{r}{a^2} \tag{1}$$

The first term accounts for a gluon exchange process between the quarks: analogous to the electromagnetic attraction of the Coulomb force, and the second for the confinement potential observed between two quarks. In practice, a more complex version of this formula has proved remarkably successful [8]. From this model, we are able to predict the energy and various quantum numbers (Tab.1) of a broad spectrum of resonances.

Charmonium Quantum Numbers		
n	principal quantum number	1, 2, 3, ...
L	orbital angular momentum	S, P, D, ...
S	spin quantum number	0, 1
J	total angular momentum	$\vec{J} = \vec{L} + \vec{S}$
P	parity	$(-1)^{L+1}$
C	charge parity	$(-1)^{L+S}$

Table 1. Quantum numbers used to label charmonium sector resonances.

A key factor to consider when searching for resonances is that their mass does not take on a single value, rather it is distributed about some mean. We thus characterise resonances by both a mean mass (M) and a width (Γ), for example: the J/ψ meson has $M_{J/\psi} = 3096$ MeV/c^2 and $\Gamma_{J/\psi} = 0.093$ MeV/c^2 [9], the $\psi(2S)$ resonance has $M_{\psi(2S)} = 3686$ MeV/c^2 and $\Gamma_{\psi(2S)} = 0.296$ MeV/c^2 [9].

The distribution arises from the uncertainty principle in the form $\Delta E \Delta t \geq \hbar/2$. This is evident when we compare the widths of the resonances mentioned prior: the J/ψ is significantly more stable and therefore has a longer lifetime, its width is thus much smaller. The exact form of this distribution is encapsulated in the Relativistic Breit-Wigner [10]:

$$B(m; M, \Gamma) = \frac{2N}{\pi} \frac{\Gamma^2 M^2}{(m^2 - M^2)^2 + m^4(\Gamma^2/M^2)} \tag{2}$$

Where m is the mass at which we are considering the value of the distribution, N is a normalisation factor accounting for degeneracies and so on, and M and Γ are as defined prior. The shape this takes for the $\psi(2S)$ is shown in Figure 2.

The observed and predicted charmonium sector states are summarised in Figure 3.

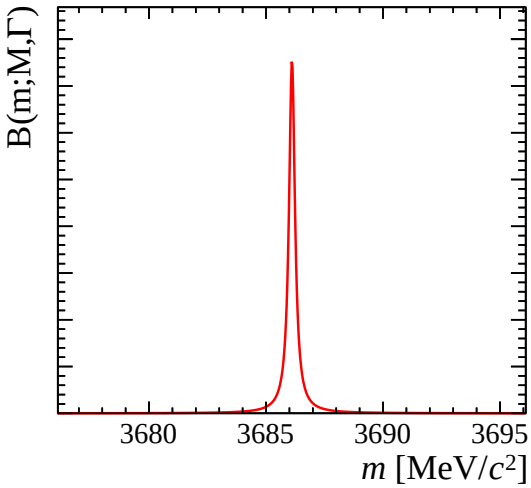


Figure 2. Relativistic Breit-Wigner distribution for the $\psi(2S)$ charmonium resonance.

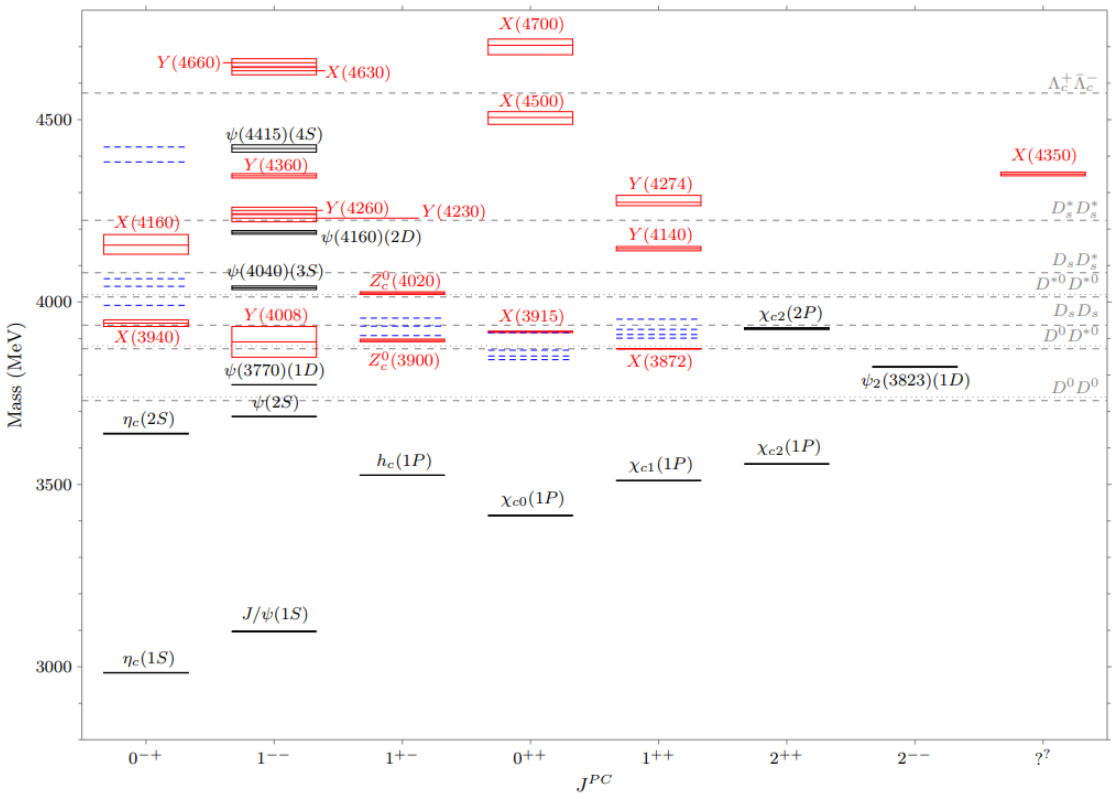


Figure 3. Summary energy level diagram of the charmonium sector. Standard charmonium states are shown in black and labelled with Greek letters (the $X(3823)$ has been labelled $\psi_2(3823)$ to reflect its conventional nature). Predicted standard charmonium states are labelled by blue dashed lines, clustered to reflect variations in model parameters. Observed charmonium-like exotic candidate states (X, Y and Z) are shown in red. The grey dashed lines represent threshold masses of various meson molecules. Taken from [5].

2.1.1 Charmonium resonances

It is worth considering a few examples of these charmonium resonances to cement an understanding of the area, and the role the $X(3823)$ plays within it.

The J/ψ was observed independently by two groups in 1974 [11, 12] (hence its two names). It is considered to be the ground state arrangement of a $(c\bar{c})$ pair due to its relatively long lifespan and quantum numbers. It is the 1^3S_1 resonance and is longer lived than the 1^1S_0 η_c due to a suppression of its decay [13]. It has mass $M_{J/\psi} = 3096$ MeV/c^2 and width $\Gamma_{J/\psi} = 0.093$ MeV/c^2 [9].

In many ways, the $\psi(2S)$ (another very well understood charmonium state) can be considered an excitation of the J/ψ. Its quantum numbers are 2^3S_1 and it has mass $M_{\psi(2S)} = 3686$ MeV/c^2 and width $\Gamma_{\psi(2S)} = 0.296$ MeV/c^2. It is frequently observed decaying to a J/ψ [9].

The resonance of particular concern to us in this analysis is the $X(3823)$: expected to have a mass around 3823 MeV/c^2. Evidence for the $X(3823)$ was obtained by Belle in 2013 [6] and it was observed by BESII in 2015 [14] - in both cases via a different decay route to the one studied in this analysis. The amount of data collected on the resonance thus far has been insufficient to determine its quantum numbers, but the proximity in which its mass lies to the predicted 1^3D_2 state has led to this explanation forming the leading theory for its identity.

2.1.2 Charmonium-like exotics

As mentioned prior, the reason for much of the interest in the charmonium region is due to the broad array of charmonium-like states whose measured qualities require an explanation outwith the normal quark model. Perhaps the most famous example of such a state is the $X(3872)$. Originally observed by Belle in 2003 [15], its mass (around 3872 MeV/c^2) and quantum numbers ($J^{PC} = 1^{++}$ [16]) mean it does not correspond to any existing standard charmonium predictions. In addition, its proximity to the threshold mass required for a $D^0\bar{D}^0$ molecule has led to the leading theory of its existence as an exotic meson in the form of a mesonic molecule.

2.2 LHCb at CERN

All of the data used in this experiment was collected by the LHCb detector [17] on the LHC [18] at CERN [19]. The LHC (Large Hadron Collider) is the largest particle physics experiment in the world: capable of colliding protons at energies up to 13 TeV. It has been running since September 2008 and has produced results sitting at the forefront of contemporary particle physics: most notably observation of the famous Higgs Boson in 2012 [20].

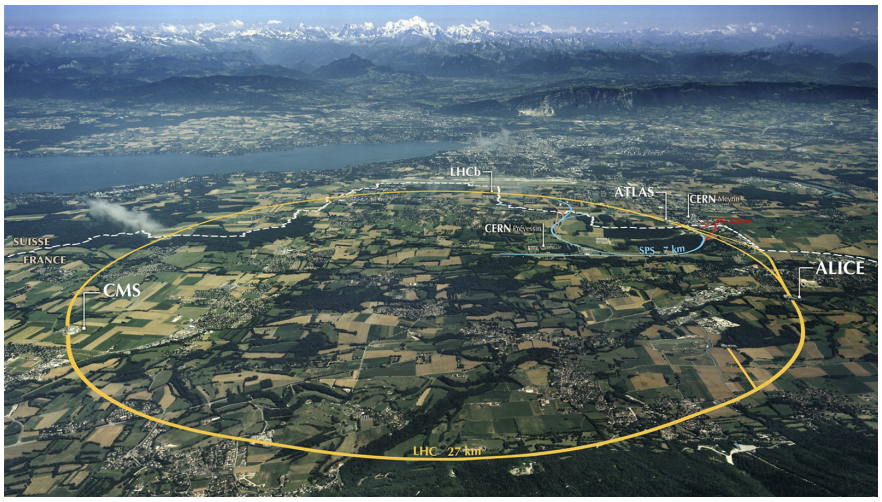


Figure 4. Aerial schematic of the LHC and its four main detectors under Geneva and the surrounding area. Picture by Maximilien Brice, CERN.

LHCb is one of four main detectors on the 27km long LHC ring (Fig.4). For a detailed breakdown of detector goals, structure and performance see here: [21]. It specialises in studying the process of CP violation [22] and rare decays of charm- and beauty-containing particles (hence the interest of this analysis in its output). LHCb has produced the world's largest dataset of charm and beauty decays and is responsible for many key analyses in the sector: [23, 24, 25].

There are both benefits and drawbacks to LHCb as compared to BaBar [26] and Belle [27], the two leading experiments in the study of b-meson decays prior to its completion. Its high energy and pp collisions (as opposed to e^+e^-) give a higher rate of production and wider breadth of possible decays, but also lead to a significant increase in the number of background events that must be contended with. Nevertheless, LHCb has stepped into the fold as the world leading experiment in the area.

2.2.1 Detector design

The physics goals are reflected in the design of the detector (Fig.5). ATLAS and CMS have a symmetric arrangement about their pp collision points, but LHCb has all of its components sitting to one side. This is because collisions resulting in the creation of b and $\bar{\text{b}}$ quarks often produce them as a pair within a narrow cone pointing one way or the other down the beamline. As such, the detector resembles something of a bookshelf arrangement with a series of components covering a relatively narrow solid angle.

It is worth considering, in brief, the role of each subdetector to build a picture of how data is collected by LHCb. The first point of measurement is the VErtex LOcator (VELO), a network of Silicon microchips which can accurately measure the initial collision point of protons, known as the Primary Vertex (PV). This is useful in reconstructing the paths taken by observed particles in the detector.

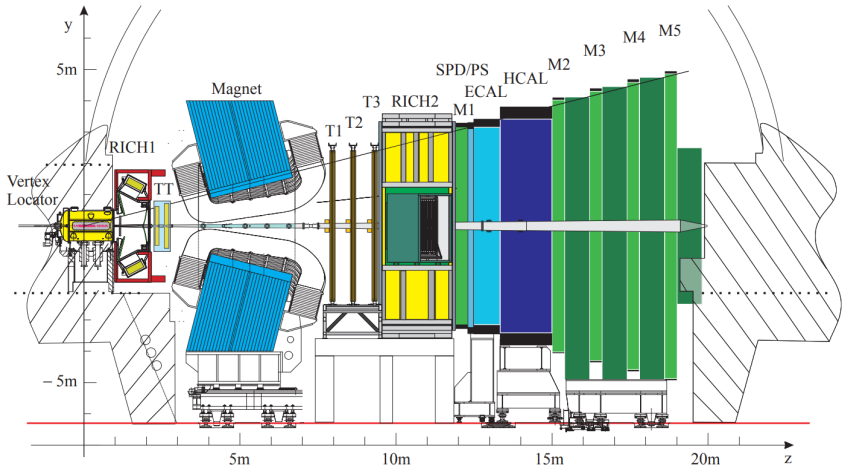


Figure 5. Simplefied schematic diagram of the LHCb detector. Taken from [28].

Subsequently sits the first Ring Imaging CHerenkov detector (RICH1), which is accompanied downstream by RICH2. These exploit the process of Cherenkov Radiation [29] to accurately identify charged hadrons, most often $\pi^{\pm}$, $K^{\pm}$ and p. This is key in recognising decay processes as it provides an important insight into the products of a given collision.

After this sits a series of trackers (TT, then T1, T2 and T3) which straddle a large magnet. The tracking detectors are arrays of silicon microchips or straw chambers, capable of accurately observing the passage of charged particles. The magnet provides an integrated field of 4Tm which deflects charged particles horizontally (the y direction in Fig.5). Combining the information together from the entire arrangement of trackers, the trajectory of particles through the detector can be interpolated and from the deflection due to the magnet their momenta determined.

A series of muon detectors (M1 to M5) at the far end of LHCb are specialised in discerning the presence of muons from the large number of hadrons produced in the collisions. Muons are capable of traversing much further through the detector than the bulk of particles and are important for accurate observation of many b-containing decays, so this arrangement is ideal for the decays we are interested in.

Sandwiched between M1 and M2 is the calorimeter system. Since neutral particles are difficult to detect with the Silicon chip construction of the tracking system and will not be perturbed by the magnet, a combination of the SPD, PS, HCAL and ECAL components is used for their measurement. These are not key to our analysis so are only touched upon briefly, more information can be found here: [21].

2.2.2 Event reconstruction

By combining these components together we can infer a great deal of information about particles produced in pp collisions within the detector: momentum, trajectory, particle identity (PID), energy and so on. Though we can only do this for particles which are relatively

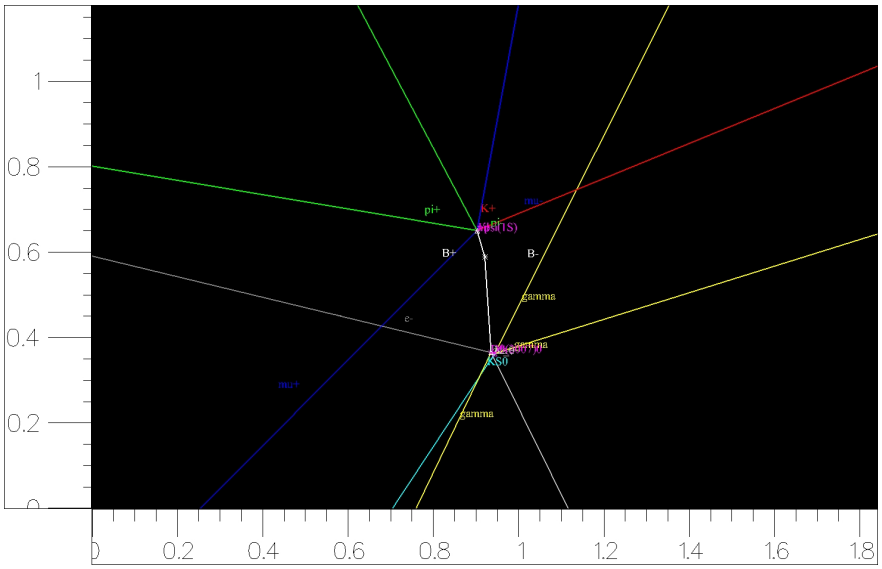


Figure 6. Simulated $B \to X(3872)K^+$, $X(3872) \to J/\psi\pi^+\pi^-$, $J/\psi \to \mu^+\mu^-$ decay, with a corresponding B^- decay. $\mu^\pm$ trajectories are shown in red, $\pi^\pm$ trajectories in green and K^+ in red. Taken from [30].

long-lived, sufficiently so to traverse though multiple layers of the detector, such as p, $K^\pm$, $\pi^\pm$ and γ. Many of the particles which we are interested in studying cannot be seen directly; the J/ψ has a mean lifetime of 7.2×10^{-21}s [9], compared to the $K^\pm$ which is a much more stable 1.2×10^{-8}s [9]. We must therefore indirectly infer their existence by recursively reconstructing the decay path for a given event from the information we do have available.

Figure 6 gives a simplified sense of this process. Isolating a single simulated decay event of $B^+ \to X(3872)K^+$, $X(3872) \to J/\psi\pi^+\pi^-$, $J/\psi \to \mu^+\mu^-$, we can see the particles which dominate the picture are $\mu^\pm$ (red), $\pi^\pm$ (green) and K^+ (red). The interesting processes which we aim to study have occurred via particles so short lived that they are not visible at this scale. We therefore look first at the trajectories of the $\mu^\pm$ and trace them back to the point at which they meet, and by combining this information with details on their four momenta and so on, we can infer that they were likely produced by the decay of a J/ψ meson. Applying conservation of momentum, we can then determine what the trajectory of this parent J/ψ was likely to have been. Repeating this process for the J/ψ and $\pi^\pm$, then the resulting $X(3872)$ and remaining K^+, we can arrive back at the original B^+. We can even trace this back to its original production point from the pp collision (the PV), which we also have information on, to act as a point of reference for the quality of our predicted reconstruction.

For a given collision, we can have hundreds of such reconstructed events to consider, and over the course of many years of running the LHC we therefore build up a dataset of many millions of events.

2.2.3 Background events

Background events come from two main sources in this reconstruction process: mis-reconstructed (i.e. fake) events, and events which very closely resemble the ones we are interested in.

Similar looking particles maybe misidentified, for example an error in measurement may lead to a K^+ being labelled as a π^+. A mistake here could propagate throughout the reconstruction as many inferences will be made assuming a π^+ mass. A coincidental resemblance in spurious detector hits, or a coincidental arrangement of uncorrelated particles downstream can also lead to the incorrect identification of prior particles. To try to account for this, Machine Learning is applied within the detector to best identify particles from a range of variables, but misidentification will still occur to some degree.

The large gap between tracking detectors due to the magnet means that significant interpolation of particle tracks must be performed here. Due to this blind spot, tracks which enter the region can often be incorrectly paired with tracks coming out, leading to erroneous reconstruction of events which do not correspond to true decays. The variables interpreted for these events will thus have unusual (often non-physical) values.

Correctly reconstructed events can also provide a component of background. For instance, in the prior described decay, the parent particle was a B^+ and the daughters were $\mu^\pm$, $\pi^\pm$ and a K^+. The decay $B^+ \to J/\psi(K_1(1270) \to \pi^+\pi^- K^+)$, $J/\psi \to \mu^+\mu^-$ will have the same parent and daughters, but is a fundamentally different process. It will therefore have many features which resemble the former decay, but also many which are different and therefore have a similar effect to the presence of mis-reconstructed background.

We can try to account for this mis-reconstruction by building more complex variables from the observed trajectories, momenta, etc. For example, every decay is assigned a normalised goodness-of-fit value to the model corresponding to it being a true signal event: signal events will have a value close to unity and poorly resembled background events will have a much larger value. We can also study individual parts of the reconstruction: χ^2_{IP} is defined as "the difference in the vertex-fit χ^2 of a given PV reconstructed with and without the track (or composite particle) under consideration". We expect this to be small for true signal parents (e.g. B^+ in the prior decay) as these should be produced at the PV, and large for true signal daughters (e.g. K^+, $\pi^\pm$) as these should be produced at some distance from the PV due to decay of the parent.

2.3 Multivariate analysis

Before studying datasets collected at particle physics experiments in detail, it is often necessary to try to best separate these background events from true signal: especially for rare events where background presence will be sufficient to completely obscure the visibility of any signal.

Often we can perform much of this by imposing simple requirements on the events we use. For example in Figure 6 we infer the presence of a J/ψ by combining the paths of a μ^+ and

a μ^-. We know how the invariant mass of the J/ψ is distributed, so we can impose the $\mu^\pm$ four momenta recombine to give a value around this region to remove any non-physical background given by coincidental alignment, for instance.

We can repeat this for many particles, and many variables, but it does not allow for complete separation of signal and background. We must then turn to the more complex variables - particularly those quantifying goodness-of-fit to the reconstructed signal trajectory. These are often (by design) differently distributed for signal and background, but not perfectly capable of separating the two as some overlap is likely to occur (Fig.14). The challenge of multivariate analysis is therefore to find a way to combine all of these variables together in a model which can then be applied to every event in the dataset in turn to return a single metric capable of discerning signal from background.

2.3.1 Decision trees

Decision trees are a familiar and simple algorithm commonly used to perform such multivariate analyses. They are best explained with a simple worked example (adapted from [31]). For instance, consider a fussy golfer sick of having games he does not enjoy due to poor weather. He decides to collect 14 days worth of information on each game he plays - including outlook, humidity and wind - and to each day he assigns a binary value of whether he enjoyed playing (Tab. 2).

Decision Tree Training Data				
Day	Outlook	Humidity	Wind	Enjoy?
1	Sunny	High	Weak	No
2	Sunny	High	Strong	No
3	Overcast	High	Weak	Yes
4	Rain	High	Weak	Yes
5	Rain	Normal	Weak	Yes
6	Rain	Normal	Strong	No
7	Overcast	Normal	Strong	Yes
8	Sunny	High	Weak	No
9	Sunny	Normal	Weak	Yes
10	Rain	Normal	Weak	Yes
11	Sunny	Normal	Strong	Yes
12	Overcast	High	Strong	Yes
13	Overcast	Normal	Weak	Yes
14	Rain	High	Strong	No
15	Rain	High	Weak	?

Table 2. 14 labelled days of data for training a decision tree. Day 15 represents an unseen event requiring classification. Adapted from [31].

He then recursively splits the dataset on each variable in turn, hoping to eventually achieve a perfect separation of labels into pure subsets. One way to do this would be to first split the

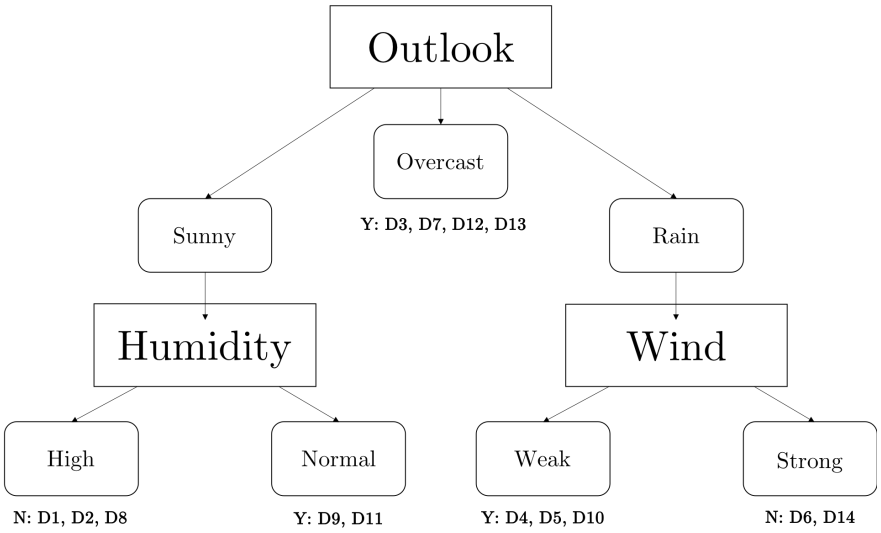


Figure 7. A possible decision tree for perfectly separating the training data in Table 2. Tracing D15 though would result in a "Yes" response. Adapted from [31].

dataset into three subsets based on the value of **Outlook**: with all *Overcast* days resulting in a "Yes" response. The resulting *Sunny* and *Rain* subsets remain impure (i.e. contain both "Yes" and "No" responses) so require further treatment. For *Sunny* days, subsequently splitting on the value of **Humidity** achieves a perfect separation: *High* contains only "No" responses and *Normal* contains only "Yes". Similarly, splitting *Rain* days on **Wind** has all *Weak* days eliciting a "Yes" response and *Strong* days a "No".

The benefit of decision trees is then evident in their action on unseen data. In effect, the golfer has constructed a model which has "learned" his weather preferences. He can then take a day on which he is yet to play, and step it through the model to determine whether or not he historically would enjoy it. Consider a day which is raining and highly humid, but with a weak wind (D15, Tab. 2). Using the tree in Figure 7, he would therefore first follow the right-hand branch, take the subsequent left branch and conclude that it is a good day to play golf. Note that given the outlook is raining, the value of humidity becomes irrelevant.

The variables we have constructed for separating signal from background, though, are numeric rather than categorical. These can still be treated with the decision tree construct by simply splitting on ranges of values. Various algorithms are possible for this, and choose variables and ranges based on different metrics. One such algorithm is C4.5, detailed extensively here: [32].

2.3.2 Overfitting

A key consideration to make when training a Machine Learning model is to avoid overfitting. In effect, these classifiers are developed to draw a boundary through an N-dimensional space (where there are N variables provided to the model) on either side of which each class of data (i.e. signal vs background) sit.

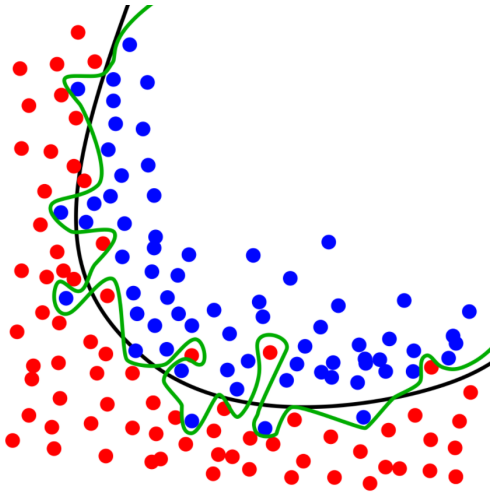


Figure 8. Overfitting of a noisy dataset. The complexity limited model (black) separates data points based on the general trend, where as the more complex model (green) separates on noise. ©Ignacio Icke, Wikimedia Commons.

Numeric values measured experimentally are subject to noise arising from uncertainties, so a small number of events will fluctuate over this boundary - meaning that some signal events may to a small degree resemble background, and vice versa. The distribution of this noise is unique to each dataset, and not representative of the true distribution of events. This means that for a given training set, if we use a model so complex that it is perfectly capable of separating the noise too, it will generalise poorly when applied to a different dataset with its own noise.

In application to the data used in our analysis, this would correspond to fitting so well to the training set that application to real data would result in a poor separation of true signal and background events, due to a different distribution of noise.

To account for this, we try to limit the complexity of a Machine Learning model so that it is capable of accounting for the general trend in the training data, but not capable of perfectly separating all noise (Fig.8). To do this, when developing our model we split the training data in to a training and a validation set, and ensure that for a given set of model parameters performance remains similar between the training set on which the model is developed and the unseen validation set. This process is known as cross validation [33].

3 Methods

In order to search for an $X(3823)$ resonance in the massive quantity of B meson decays recorded by LHCb, there are two key stages of analysis to complete. We must first apply various filters (hereby referred to as selections) to select only the decay route we are interested in, namely $B^+ \to X(3823)K^+$, $X(3823) \to J/\psi\pi^+\pi^-$, $J/\psi \to \mu^+\mu^-$, and to reduce the presence of background events as much as possible. Secondly, we must fit theoretically motivated models to the observed resonance to place limits on its properties.

3.1 Data preparation

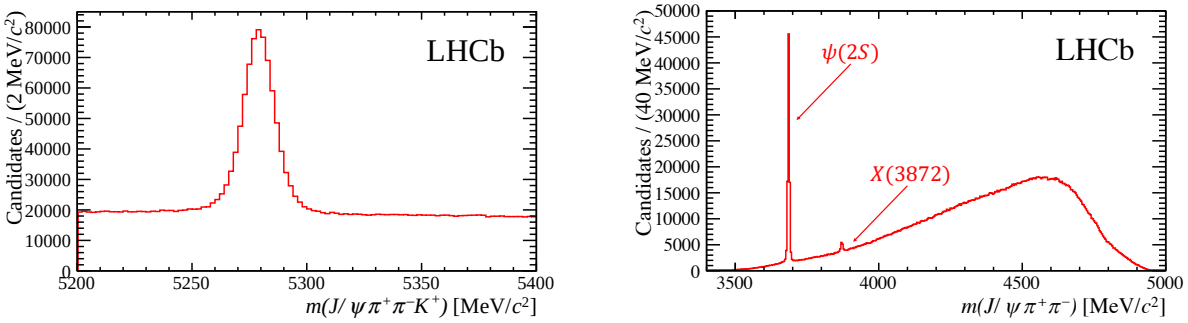
The data used in this experiment was a subset of the B meson decays recorded by LHCb in runs I (2011, 2012) and II (2015, 2016, 2017). Note that not all 2017 data has yet been included as some processing is still required. The luminosity [34] for each run can be seen in Table 3, this is roughly 85% of the available data.

LHC Annual Luminosity	
Years	L $[fb^-1]$
2011, 2012	3
2015, 2016	2
2017	1.5
Total	6.5

Table 3. Recorded LHC Luminosity per run included in this experiment's dataset.

To prepare the initial dataset, the DAVINCI framework [35] was used to select only the decays most likely to correspond to the channel of interest. A combination of considerations was made to construct this dataset. First and foremost, the assigned particle identity (PID) was considered: all events were required to contain a $\pi^+\pi^-$ pair, a K^+ and a $\mu^+\mu^-$ pair, originating from the decay of a parent B^+. In addition, it was required that the $\mu^+\mu^-$ pair have a combined mass within 50MeV/c^2 of the known J/ψ mass. An extensive set of other such requirements were imposed on various variables to further reduce background presence, including considerations of transverse momenta, confidence in the quality of particle reconstruction and so on. This process is standard in LHCb analyses and was performed prior to the beginning of the project so will not be discussed in detail here.

The dataset that results from this initial preparation is summarised in Figure 9. The left plot is a histogram of the invariant mass obtained when each candidate decay event's J/ψ, $\pi^\pm$ and K^+ (i.e. all daughter particles) four momenta are added together. This recombined mass therefore provides information on the reconstructed parent B^+ meson. The right plot is a histogram of the invariant mass obtained on addition of just the J/ψ and $\pi^\pm$ masses. Since these are the decay products we expect from the $X(3823)$, if a significant number of $B^+ \to X(3823)K^+$ decays occur then a resonance should be visible as an abundance of events around 3823 MeV/c^2 in this distribution.



(a) raw $m(J/\psi\pi^+\pi^-K^+)$ distribution

(b) raw $m(J/\psi\pi^+\pi^-)$ distribution

Figure 9. Raw dataset of $B^+ \to J/\psi\pi^+\pi^-K^+$ decays from LHCb representing a combined luminosity of 6.5 fb^{-1} (Tab.3). Evident in (a) is a strong peak around $m(B^+)$ and a significant combinatorial background. In (b) there are two peaks: one around $m(\psi(2S))$ and one around $m(X(3872))$, present as these resonances both have significant branching fractions for decay to $J/\psi\pi^+\pi^-$. There is a significant rising background evident. No clear resonance is yet seen in the $X(3823)$ region.

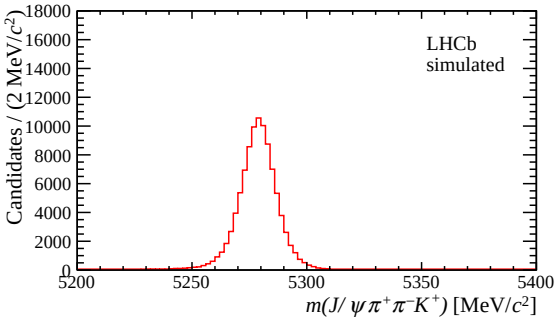
We see a strong peak around the mass of the B^+ meson (5279.32 MeV/c^2 [9]), indicating that the signal events do originate from a correctly reconstructed B^+ decay. We also see a significant combinatorial background presence. We see two clear peaks here in the $m(J/\psi\pi^+\pi^-)$ distribution, corresponding to the $\psi(2S)$ and $X(3872)$ as these states have significant branching fractions for decay to $J/\psi\pi^+\pi^-$. The suspected $X(3823)$ region (i.e. masses around 3823 MeV/c^2) is dominated by the rising background, so further treatment of the dataset was required before any attempt at observation could be made.

3.1.1 Monte-Carlo simulation

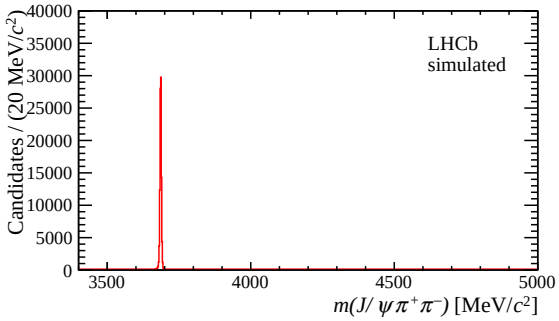
When performing analyses on a dataset containing a mix of signal events and background, it is vital to have a sense of what a signal event "should" look like so that we can separate the two. Signal and background events will generate different signatures in the detector, and we require a sense of what these signatures are before we can label each event. Some variables are obvious, for instance all signal events originate from a B^+ meson so will have a reconstructed $m(J/\psi\pi^+\pi^-K^+)$ close to 5280 MeV/c^2. More complex variables, particularly those encapsulating the reconstruction goodness-of-fit as discussed prior, are more difficult to intuitively identify.

As we understand very well how certain signal events are produced, and how their decay products will react with detector components to produce measurements, we can simulate these processes to exceptional accuracy. The method used most often in particle physics is Monte-Carlo (MC), and it is detailed for LHCb here: [36].

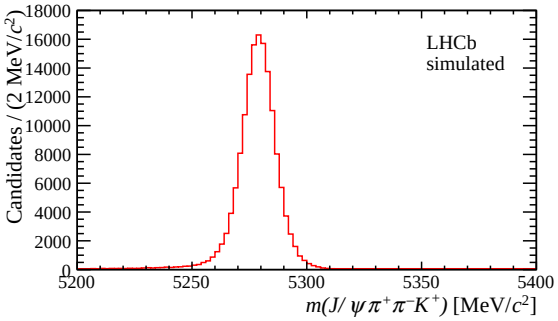
As the $\psi(2S)$ and $X(3872)$ resonances are well understood, and present in our dataset as clear signals for reference, the decision was made to use MC generated datasets of the decays



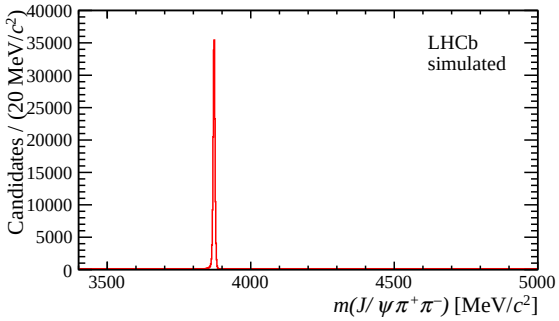
(a) MC $\psi(2S)$ $m(J/\psi\pi^+\pi^-K^+)$ distribution



(b) MC $\psi(2S)$ $m(J/\psi\pi^+\pi^-)$ distribution



(c) MC $X(3872)$ $m(J/\psi\pi^+\pi^-K^+)$ distribution



(d) MC $X(3872)$ $m(J/\psi\pi^+\pi^-)$ distribution

Figure 10. The Monte-Carlo (MC) datasets for the decays $B \to \psi(2S)K^+$, $\psi(2S) \to J/\psi\pi^+\pi^-$, $J/\psi \to \mu^+\mu^-$ (MC $\psi(2S)$) and $B \to X(3872)K^+$, $X(3872) \to J/\psi\pi^+\pi^-$, $J/\psi \to \mu^+\mu^-$ (MC $X(3872)$). Both show a peak around the B^+ mass in the reconstructed $m(J/\psi\pi^+\pi^-K^+)$ and at their corresponding resonances in $m(J/\psi\pi^+\pi^-)$. As the simulation is for signal-like events only, there is no background present. 98231 events were simulated for MC $\psi(2S)$ and 1590093 events for MC $X(3872)$.

$B \to \psi(2S)K^+$, $\psi(2S) \to J/\psi\pi^+\pi^-$, $J/\psi \to \mu^+\mu^-$ and $B \to X(3872)K^+$, $X(3872) \to J/\psi\pi^+\pi^-$, $J/\psi \to \mu^+\mu^-$. Such simulations are computationally expensive, and not yet available for the $X(3823)$, so a generalisation was required. These resonances are analogous in many ways to the $X(3823)$, and the variable distributions we are interested in are those which describe how well events represent a true signal, as opposed to background, and so are similar irrespective of exact parent resonance.

These datasets are summarised in Figure 10.

3.1.2 Fitting to reconstructed events

By fitting to the $m(J/\psi\pi^+\pi^-K^+)$ distribution, we can learn a number of important details about our dataset. Since all signals come from the same B^+ peak, we can use this to estimate the number of signal and background events we have throughout the analysis. This

is important as it allows us to keep track of the effects that the various selections we impose have. The more we cut out unwanted background events, the more we will invariably cut out some signal events due to coincidental overlap of variables. By estimating the number of signal events after each selection is applied, we can keep track of the efficiency of our methods on the signal events.

To do this we combine two fits: a signal component to account for the shape of the B^+ mass peak and an exponential component to account for the flat background. By normalising the integral of each fit function to unity, as the distributions are histograms counting the total number of events we can use the relative magnitude of each component in the total fit to tell us the number of events which account for each component of the fit. To clarify:

$$F_{\text{tot}} = N_{\text{sig}} f_{\text{sig}} + N_{\text{bgr}} f_{\text{bgr}} \tag{3}$$

Where F_{tot} is the total fit, N_{comp} is the number of events in the specified component and f_{comp} is the component-specific fit function, normalised to unity.

The distribution of events about the B^+ meson mass clearly differs significantly from the Relativistic Breit-Wigner shape of the $\psi(2S)$ (Fig.2). The reason for this is twofold. Firstly the B^+ is relatively long-lived ($1.636 \times 10^{-12} s$ [9]), making it a very narrow state: at this mass, the width is narrower than the resolution of the detector so does not factor in to the distribution. Secondly and more importantly, electronic noise, physical construction and so on all introduce a degree of uncertainty into the measurements made by any detector. This has the effect of smearing the mass measurement from a narrow resonance to some broader resolution function. Most often these are modelled with a combination of Gaussian functions or Crystal Balls [37] (a Gaussian of width σ and mean $\bar{x}$, with a power law tail of exponent n beyond a threshold $\frac{x-\bar{x}}{\sigma} > \alpha$).

Fitting was performed using the ROOT [38] libraries of C++ [39], specifically RooFit [40]. The fit functions are complex, often constructed from a combination of resolution and resonance functions. As such, they can contain many parameters and so the optimisation procedure can be slow. It is useful to first perform fits to the cleaner signal-only MC distributions to provide a good starting point before then applying the procedure to the recorded data. In the case of the reconstructed $J/\psi\pi^+\pi^-K^+$ mass, the optimal fit is a two-tailed Crystal Ball function - modelled by fixing two Crystal Balls to have the same mean and amplitude - plus an additional Gaussian to represent the signal component, and an exponential background to account for the background. In simple summary:

$$f_{\text{sig}} = r\Big(\frac{1}{2} f_{\text{CB1}}\big(m; \bar{m}, \alpha_1, n_1, \sigma_1\big) + \frac{1}{2} f_{\text{CB2}}\big(m; \bar{m}, \alpha_2, n_2, \sigma_1 \times \frac{\sigma_2}{\sigma_1}\big)\Big) + (1-r) f_{\text{g}}\big(m; \bar{m}, \sigma_{\text{g}}\big) \tag{4}$$

$$f_{\text{bgr}} = e^{cm} \tag{5}$$

$m(J/\psi\pi^+\pi^-K^+)$ Fit Parameters			
Parameter	Description	MC Value	Data Value
$\bar{m}$ [MeV/c^2]	mean mass	5278.97(2)	5279.07(2)
α_1	f_{CB1} threshold	1.47(4)	fixed
α_2	f_{CB2} threshold	-2.20(5)	fixed
n_1	f_{CB1} power law exponent	1.55(9)	fixed
n_2	f_{CB2} power law exponent	1.97(16)	fixed
σ_1 [MeV/c^2]	f_{CB1} core width	9.8(3)	9.46(5)
$\frac{\sigma_2}{\sigma_1}$	$f_{\mathrm{CB}i}$ core width ratio	0.96(4)	fixed
r	f_{g}, $f_{\mathrm{CB}i}$ relative amplitude	0.52(4)	fixed
σ_g [MeV/c^2]	f_{g} width	5.68(12)	fixed
c [$\times 10^{-4}$ c^2/MeV]	Background exponent	n/a	-4.12(13)
N_{sig} [$\times 10^5$]	signal yield	n/a	5.458(12)
N_{bgr} [$\times 10^6$]	background yield	n/a	1.8646(17)

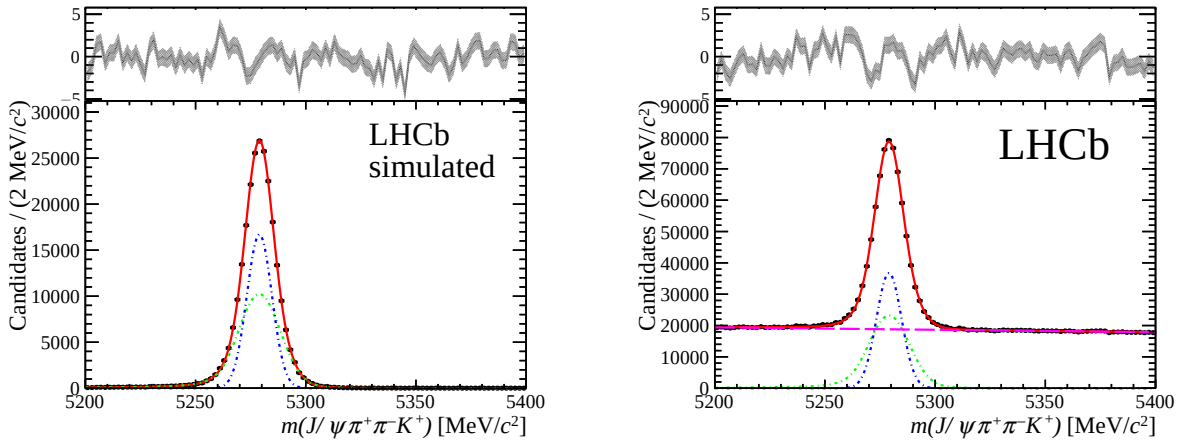
Table 4. Parameters for the two-tailed Crystal Ball plus Gaussian fit to the combined $\psi(2S)$ and $X(3872)$ Monte-Carlo $m(J/\psi\pi^+\pi^-K^+)$ distribution, and the two-tailed Crystal Ball plus Gaussian plus exponential background fit to the same distribution in data.

Where m is the mass at which the value of the function is evaluated, $f_{\mathrm{CB}i}$ is the ith Crystal Ball function, f_{g} is the Gaussian and the remainder of the parameters are defined in Table 4.

This leads to a fit with 12 free parameters, many of which will be (at least somewhat) covariant. As the shape of the signal distribution in data and MC was (by design) very similar, the MC dataset was used to find the initial values for the signal component parameters: most of which were then fixed. Figure 11 summarises these fits.

From this we observe, in a dataset with only preparatory selections made, a signal yield of just over half-a-million events and a background yield of just under two million. A key goal in the analysis that follows is to keep the former number as high as possible whilst reducing the latter.

Another factor that can be determined from these fits is the sWeight [41], a value that can be assigned to each entry in the dataset to quantify how likely it is to correspond to a signal event. These are generated by comparing the likelihood of a given event to correspond to the signal or background component of the fit to data. Their utility lies in enabling a simple background reduction method: when producing a histogram to describe the distribution of a given variable, if the contribution from each event is weighted by its sWeight, the resultant histogram will be equivalent to that produced when a histogram containing events from a region of predominantly background is subtracted from a histogram containing events from a region of significant signal contribution. This is useful when trying to estimate how variables are distributed for signal events.



(a) Fitted Monte-Carlo $m(J/\psi\pi^+\pi^-K^+)$ distribution

(b) Fitted data $m(J/\psi\pi^+\pi^-K^+)$ distribution

Figure 11. (a) shows a two-tailed Crystal Ball (green) plus Gaussian (blue) fit to the $m(J/\psi\pi^+\pi^-K^+)$ distribution in the combined $\psi(2S)$ and $X(3872)$ Monte-Carlo datasets. (b) shows both of these components (same colours) plus an exponential background (magenta) fitted to the $m(J/\psi\pi^+\pi^-K^+)$ distribution in data. The normalised residuals (top) show clear structure around either tail of the B^+ mass peak, but for the purposes of these fits precision is not vital.

3.1.3 Reweighting

It is vital that MC datasets closely resemble signal events. Although the events and detector are well understood and accurately simulated, since some approximations must be made, it is to be expected that some deviation will occur.

It was found that most variables closely matched their corresponding data distributions, but deviation was observed (Fig.12). In order to correct for this, a reweighting was performed for the two MC datasets. This is a procedure which assigns a "weight", or relative importance, to each event which can be carried throughout the analysis.

Since this is typically performed to have one distribution match another, it is most simply done via a technique known as histogram division. For each histogram bin, a weight is determined by:

$$w_{\text{bin}} = \frac{N_{\text{bin, data}}}{N_{\text{bin, MC}}} \tag{6}$$

Where $N_{\text{bin, data/MC}}$ is the count of data or MC events in a given bin. This weight for each bin is then assigned to every event in that bin. When considering the reweighted data, instead of then considering a "count" of 1 for each event, we consider the event weight instead: both in plotting and in statistical treatment.

For a single variable, this leads to perfect agreement between data and MC, but we consider many. Histogram reweighting would therefore bias our distributions towards a match for a

single variable. Instead, a more complex technique which spanned multiple variable distributions (Tab.5) using Gradient Boosted Decision trees was used: GBReweighter [42, 43], part of the hep_ml library [44] implemented in Python2 [45].

Reweighter Variables	
Variable	Description
$\log(\chi^2_{\mathrm{DTF}})$	Goodness-of-fit to reconstructed event model [46]
N_{tracks}	Number of charged tracks in detector at time of collision
B^+_p	Momentum of parent B^+
$B^+_{p_T}$	Transverse momentum of parent B^+

Table 5. Summary of variables used to reweight the $\psi(2S)$ and $X(3872)$ Monte-Carlo simulated datasets.

The algorithm requires data and sWeights for producing a reweighting. The sWeights act as a method for removing the contribution of background events to the variable distributions, since it is the signal only that we expect Monte-Carlo to resemble. Such datasets were prepared separately for the for the $\psi(2S)$ and $X(3872)$ regions: as these are different resonances we would expect their variable distributions to differ slightly. For reweighting the $\psi(2S)$ MC, all data events satisfying $3666\mathrm{MeV/c^2} \leq m(J/\psi\pi^+\pi^-) \leq 3706\mathrm{MeV/c^2}$ were selected. This is a window of $\pm 20\mathrm{MeV/c^2}$ about the $\psi(2S)$ resonance mean. Similarly for the $X(3872)$ a requirement of $3852\mathrm{MeV/c^2} \leq m(J/\psi\pi^+\pi^-) \leq 3892\mathrm{MeV/c^2}$ was imposed. Fitting to the B^+ mass peak in each of these resonances, sWeights were generated. Each MC dataset was then reweighted separately.

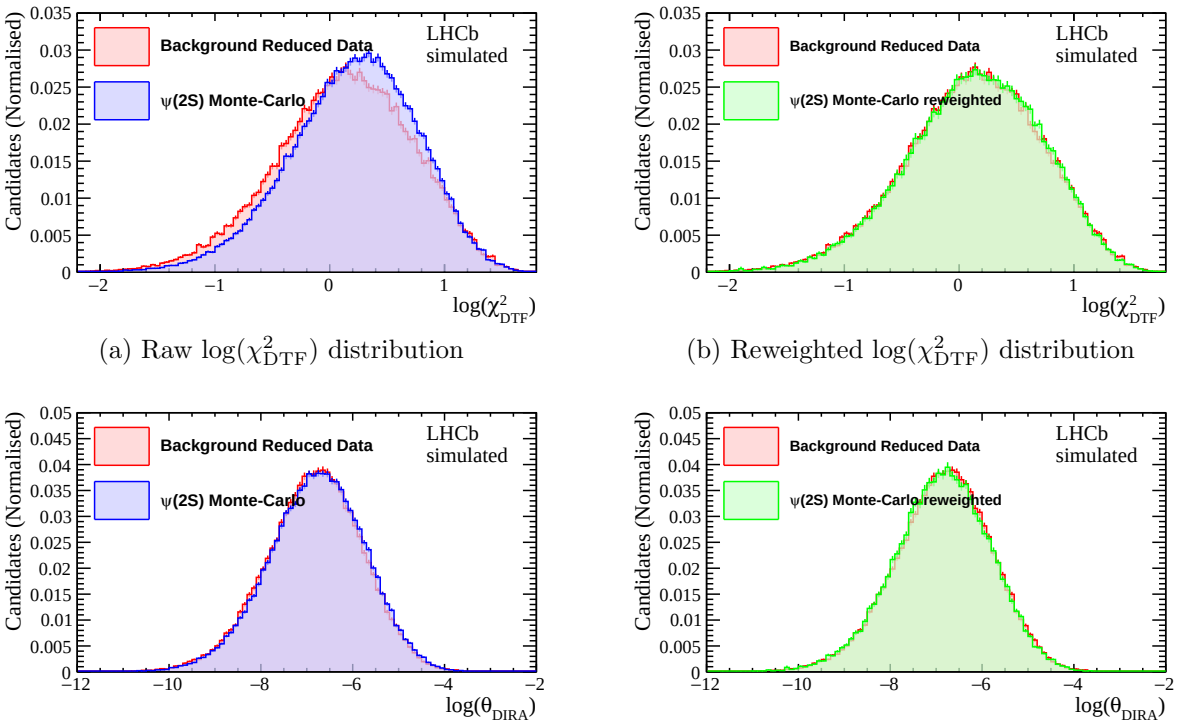
Figure 12 demonstrates how this enabled a correction to be applied to the $\log(\chi^2_{\mathrm{DTF}})$ distribution in $\psi(2S)$ without compromising the $\log(\theta_{\mathrm{DIRA}})$ distribution. Similarly successful results were observed for $X(3872)$.

3.2 Initial selections

It is worth noting that when each of these steps which follow were performed, it was not until the multivariate selection model was completely optimised that the $X(3823)$ region was studied in any detail. The decision to remain blind to the region in this sense was made so as not to bias the methods towards a specific result.

Before proceeding to the complex application of a multivariate selection model, the presence of background in the data was further reduced by applying some simple selection criteria.

As can be seen in the summary plots of Figure 9, there is a rising background component for $m(J/\psi\pi^+\pi^-)$ larger than any of the resonances we are interested in; namely $\psi(2S)$, $X(3823)$ and $X(3872)$. The Q value [47], which measures how much energy is released by a decay, was used to eliminate much of this background. For a general decay $B^+ \to (X \to J/\psi\pi^+\pi^-)K^+$



(a) Raw $\log(\chi^2_{\mathrm{DTF}})$ distribution

(b) Reweighted $\log(\chi^2_{\mathrm{DTF}})$ distribution

(c) Raw $\log(\theta_{\mathrm{DIRA}})$ distribution

(d) Reweighted $\log(\theta_{\mathrm{DIRA}})$ distribution

Figure 12. Using the GBReweighter algorithm, the resemblance of $\log(\chi^2_{\mathrm{DTF}})$ in Monte-Carlo Simulation to the background reduced (i.e. signal only) data was improved (top), whilst preserving the accuracy of the $\log(\theta_{\mathrm{DIRA}})$ distribution (bottom).

where X is any charmonium-like resonance (in our case $\psi(2S)$, $X(3823)$ or $X(3872)$), for the X-specific portion of the decay this is defined as:

$$Q \equiv [m(X) - m(J/\psi) - m(\pi^+\pi^-)]c^2 \tag{7}$$

Where we pair $\pi^\pm$ (dimuon system) as the exact mechanics of the $X(3823)$ decay are not known. We do not know the exact Q value for the $X(3823)$ decay, but we can place a limit on it by considering the maximum energy that can be released from this decay: where the $X(3823)$ decays to the J/ψ, π^+ and π^-, all as independent particles in their ground states:

$$Q_{\max} = [m(X(3823)) - m(J/\psi) - m(\pi^+) - m(\pi^-)]c^2 \tag{8}$$

Substituting rough values from the PDG [9]:

$$Q_{\max} = [3823 - 3096 - 139 - 139]\text{MeV} \simeq 500 \text{ MeV} \tag{9}$$

In practice, this is applied to the data by imposing, for each event:

$$m(J/\psi\pi^+\pi^-) - 3096 - m(\pi^+\pi^-) < 500 \tag{10}$$

Where $m(J/\psi\pi^+\pi^-)$ is - as before - the reconstructed mass of the $X(3823)$ candidates' decay products, 3096 is the (rough) PDG [9] value for the mass of the J/ψ and $m(\pi^+\pi^-)$ is the reconstructed mass of the event's dimuon system.

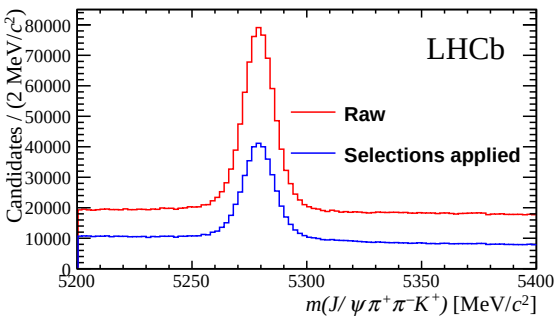
As the $X(3872)$ resonance is at a higher mass than the $X(3823)$ we should be wary of this Q value selection removing many of these events. Due to the resonant structure of the dimuon system in the $X(3872) \to J/\psi\pi^+\pi^-$ decay, Q value limits as low as 250 MeV do not significantly impact the number of signal events here [48].

These simple selections are also applied to the MC datasets for completeness, though they have little effect as they are designed to remove background events.

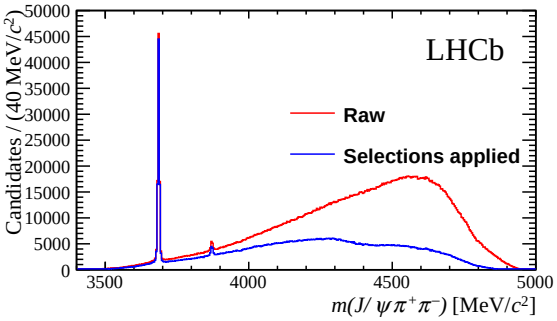
Figure 13 summarises the effect of these selections on the raw data. Of note is a large reduction across the whole distribution of $m(J/\psi\pi^+\pi^-K^+)$, including a significant reduction in the number of signal events. Looking at the $m(J/\psi\pi^+\pi^-)$ distribution however, we see that the bulk of the reduction in events comes from the rising background component: leaving most of the $\psi(2S)$ and $X(3872)$ events unaffected. There is a slight reduction in background in the $X(3823)$ region, but no resonance yet evident.

3.3 Machine learning

These initial selections are useful in reducing much of the obvious background, but their application is limited: they are each effective in only one dimension. For any given variable, they allow us to reject all events outwith a certain range as background and accept the remainder as signal. This is fine for Q-value which we know non-physical mis-reconstructed background events will be effectively removed by, but as Figure 14 shows, there are many



(a) $m(J/\psi\pi^+\pi^-K^+)$ distribution with initial selections



(b) $m(J/\psi\pi^+\pi^-)$ distribution with initial selections

Figure 13. Dataset of $B^+ \to J/\psi\pi^+\pi^-K^+$ from LHCb representing a combined luminosity of 6.5 fb^{-1} (Tab. 3) with initial selections applied. Evident in (a) is a significant reduction in the number of events compared to the raw dataset, including around the B^+ resonance. In (b) there is a significant reduction in the rising background, but the $\psi(2S)$ and $X(3872)$ resonances are mostly unaffected. There is still no clear resonance in the $X(3823)$ region.

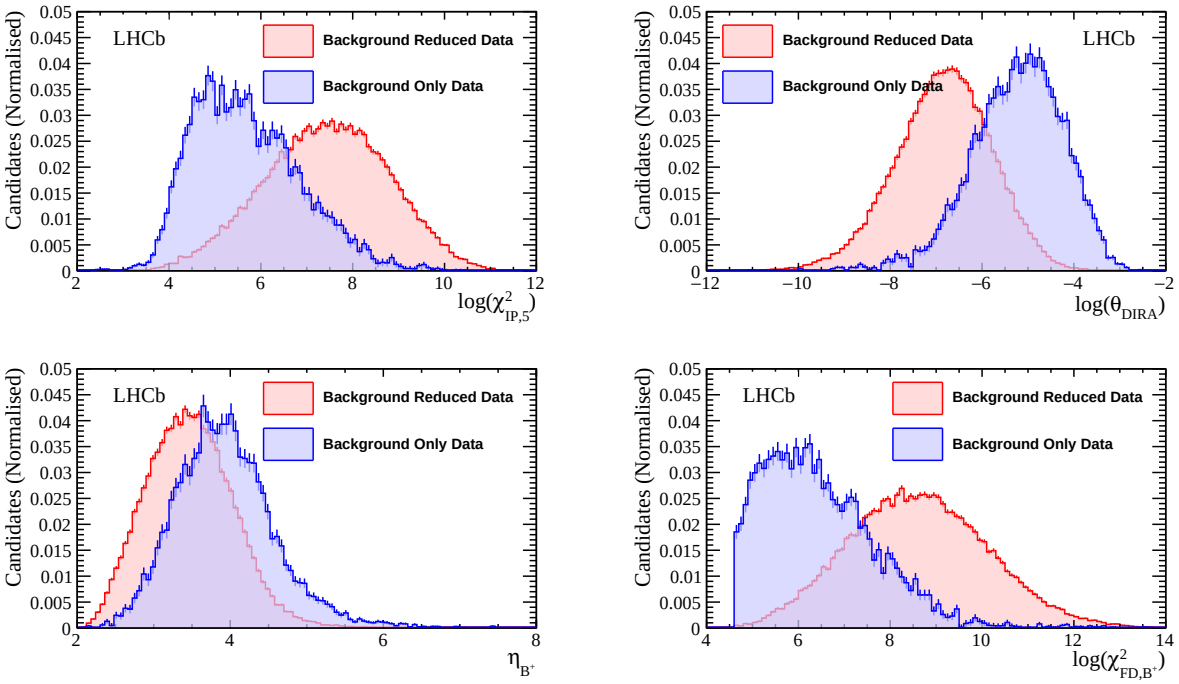


Figure 14. Distribution of various variables passed to Multivariate Selection Model for events in data satisfying $3666\text{MeV/c}^2 \leq m(J/\psi\pi^+\pi^-) \leq 3706\text{MeV/c}^2$, i.e. in the $\psi(2S)$ region.

Machine Learning Variables	
Variable	Motivation
$\log(\chi^2_{\mathrm{DTF}})$	Normalised goodness of fit of decay being modelled as a $B^+ \to (X \to J/\psi\pi^+\pi^-)K^+$ [46] - true signal events will be a good fit and so have a value close to unity, misidentified backgrounds will fit this model poorly
η_{B^+}	Pseuduorapidity [49] of the parent B^+ - expected to be higher for signal events
$\log(\theta_{\mathrm{DIRA}})$	Cosine of angle between momentum of parent B^+ and combined daughters, angle should approach zero for signal events due to conservation of momentum
$\log(\chi^2_{\mathrm{IP},B^+})$	The difference in the vertex-fit χ^2 of a given PV reconstructed with and without the track (or composite particle) under consideration, should be low for signal parents and high for signal daughters
$\log(\chi^2_{\mathrm{IP},1})$	*as above*, 1 denotes this is for the daughter particle with the lowest value
$\log(\chi^2_{\mathrm{IP},2})$	*as above*
$\log(\chi^2_{\mathrm{IP},3})$	*as above*
$\log(\chi^2_{\mathrm{IP},4})$	*as above*
$\log(\chi^2_{\mathrm{IP},5})$	*as above*, 5 denotes this is for the daughter particle with the highest value
$\log(\chi^2_{\mathrm{FD},B^+})$	The difference in χ^2 for a reconstructed PV and decay vertex separation with and without the parent B^+ under consideration, since this parent should be responsible for the traversed distance this should be high for true signal events
$p_T^{\mu,max}$	Maximum transverse momentum of $\mu^\pm$ used to reconstruct J/ψ - expected to be higher for signal events
$p_T^{\mu,min}$	Minimum transverse momentum of $\mu^\pm$ used to reconstruct J/ψ - expected to be higher for signal events

Table 6. Summary of variables passed to the Machine Learning Models.

variables for which signal and background are clearly differently distributed but are not separable by a single cut. The goal of Machine Learning in this experiment was, therefore, to combine these discriminating variables in such a way so as to provide a single metric which could be used to separate signal events from background.

In total, 12 such variables were considered. These are summarised in Table 6. In most cases, the logarithm of each variable was considered as many have a skewed distribution: for instance, variables corresponding to a normalised χ^2 parameter have a value around unity for the bulk of events, with relatively few events taking higher values.

There are two key phases in preparing a Machine Learning model: training and application. In the first, we "teach" our model what a signal and background event should "look like": in effect what this means is it "learns" what patterns to look for in the discriminating variables so that when it is provided with data in future, it can assign a probability of each event in this new data corresponding to signal or background. For these patterns to be "learned" and associated with either signal or background, we require an initial labelled training set. For

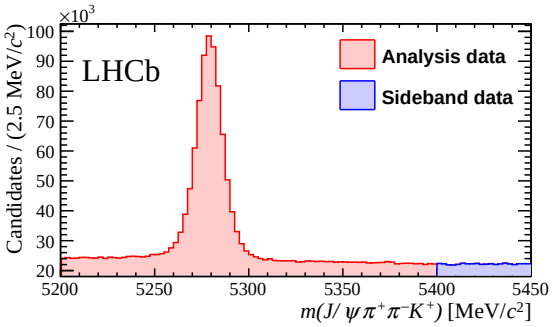


Figure 15. Reconstructed $m(J/\psi\pi^+\pi^-K^+)$ distribution for all events in the collected dataset. Events used as data in the analysis ($5200\text{MeV/c}^2 \leq m(J/\psi\pi^+\pi^-K^+) \leq 5400\text{MeV/c}^2$) are shown in red. The upper sideband ($m(J/\psi\pi^+\pi^-K^+) > 5400\text{MeV/c}^2$) is shown in blue, this was used as the labelled background training set for developing Machine Learning models. A lower sideband was not considered as the gap beside the B^+ mass peak was too narrow.

the signal events this is accomplished with the Monte-Carlo sets: by construction they only contain signal events and we have ensured they closely resemble the distribution of signal events in the true data. To produce a labelled background set, we consider a region known as the "upper sideband" (Fig.15): recorded events with a reconstructed $m(J/\psi\pi^+\pi^-K^+)$ sufficiently far from the B^+ peak that we can be sure they are all background.

3.3.1 Selection of model

There are various techniques for performing Machine Learning common to particle physics, including Artificial Neural Networks (ANNs) [50] and Boosted Decision Trees (BDTs) [51]. Many are implemented in the TMVA library [52] of ROOT, which was used to initially explore the performance of various methods on the Training datasets (Monte-Carlo simulation plus sideband). Table 7 summarises the performance of various Machine Learning techniques implemented in TMVA. Each performs to a remarkably similar level, but the BDT methods are significantly faster.

With this in mind, a BDT-based method was deemed most appropriate for this analysis. Given the ease of use Python2 [45] and the efficiency of its data analysis packages Numpy [57] and Pandas [58], the decision was made to seek an implementation of BDT classification in Python. Extreme Gradient Boosting (XGBoost) [59] is an open source library which provides methods of performing Machine Learning classification tasks through Gradient Boosted Decision trees in Python, and has proven itself to be effective in particle physics data analysis [60].

3.3.2 Hyperparameters

XGBoost allows tuning of a number of hyperparameters - factors which affect construction of the model itself. These can have a significant effect on performance, but can also lead to

TMVA Machine Learning Method Summary			
TMVA Method	Description	ROC Integral	Training Time (s)
MLP	Traditional ANN	0.986	1.11×10^3
MLPBFGS	Traditional ANN with BFGS optimization method [53]	0.987	1.8×10^4
BDT	Decision Tree Ensemble with Adaptive Boosting [54]	0.986	127
BDTG	Decision Tree Ensemble with Gradient Boosting [55]	0.986	156

Table 7. Summary of the performance of various Artificial Neural Network (ANN) and Boosted Decision Tree (BDT) methods implemented in TMVA on training data. ROC Integral is a common Machine Learning performance metric, with 1 representing a perfect classifier, and 0.5 representing random classification [56].

overfitting, so a careful trade-off was required. Most were left at their default values, but the effect of changing the most significant was studied. These are summarised in Table 8.

XGBoost Hyperparameters		
Parameter	Value	Description
Learning Rate	0.1	How fine a step is taken in exploring the objective function
N Estimators	600	How large an ensemble of trees is considered
Max Depth	6	Maximum depth of tree, i.e. how many variables are considered per complete branch

Table 8. Hyperparameters tuned in the XGBoost Classifier.

To find the optimal hyperparameters, with regards to both performance and overfitting, the Logarithmic-Loss (or cross-entropy [61]) metric was used. This is defined as:

$$-\log\left(p(\vec{y_T}|\vec{y_P})\right) = -\sum_{n=1}^{N_{\text{events}}}\left(y_{T,n}\log(y_{P,n}) + (1-y_{T,n})\log(1-y_{P,n})\right) \tag{11}$$

Where, for a labelled training set $X_n, n \in [1, N_{\text{events}}]$ events, $\vec{y_T}$ is a vector of true labels with $y_{T,n} = 1$ if X_n is a signal event and $y_{T,n} = 0$ if X_n is background. $\vec{y_P}$ is a vector of probabilities assigned by the Machine Learning model to each training event, $y_P = p(X_n = \text{signal})$, i.e. 1 is absolute confidence that a given event is signal and 0 is absolute confidence that a given event is background. $p(\vec{y_T}|\vec{y_P})$ thus quantifies the quality of the model by measuring how likely it would be for us to observe the true labels if they were generated from our model's predictions.

Logarithmic-loss penalises each prediction for its distance from the true label, so in minimising it we produce a model which gives us the best split in probability distributions between signal and background.

To perform hyperparameter optimisation whilst checking for overfitting, the labelled training dataset was randomly shuffled then split in to two: with 80% of events forming the optimisation dataset and the remaining 20% forming a validation set. For a given set of hyperparameters, a weighted XGBoost model was trained on the optimisation set, then applied to the same dataset to produce a set of probabilities $\vec{y}_{P,\text{opt}}$ from which a log-loss score was determined. The model was then also applied to the validation set (without retraining) to assess its performance on unseen data, from which another log-loss score was determined. As the data to which we would then apply the trained XGBoost model are unlabelled, the important performance metric was the log-loss score on unseen data. The optimal set of hyperparameters were therefore those which minimised log-loss on the validation set.

As training an XGBoost model for each set of hyperparameters is a slow process, a naive grid search across the possible space of parameters would have been too time-intensive. Instead, the Scikit-Optimize [62] Python package was used, which implements a method for performing Bayesian optimisation by approximating the objective function (log-loss of the validation set, in our case) to be distributed as a multivariate Gaussian, with dimensions corresponding to the set of possible hyperparameter values [63]. The values obtained from this process are shown in Table 8.

3.3.3 Applying the decision tree

With the optimal hyperparameter values obtained, the full multivariate analysis was then performed. No validation set was partitioned when training the XGBoost model to be applied to data as the hyperparameter optimisation process had already accounted for overfitting. Instead, the XGBoost model with hyperparameter values as specified in Table 8 was trained on the full combined Monte-Carlo and sideband datasets, with the Monte-Carlo weights generated prior included in the training procedure.

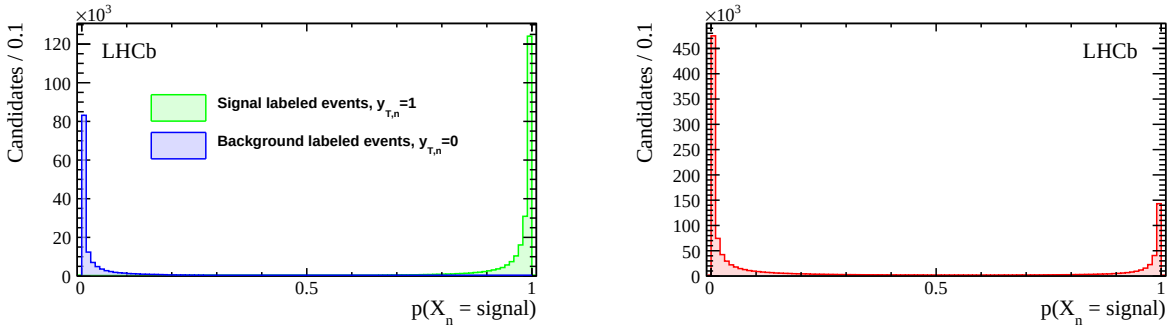
The full model was then applied to both the training set and to dataset containing the reconstructed B^+ candidate decays to generate $p(X_n = \mathit{signal})$ for each event. The output of the trained XGBoost classifier is summarised in Figure 16.

3.3.4 Threshold probability optimisation

Given the output taken from the XGBoost classifier is a probability for each event to be signal, we have to define some threshold value at which to label a given event to be signal or background, i.e.:

$$y_{L,n} = \begin{cases} 1, & \text{if } y_{P,n} \geq p_t \\ 0, & \text{if } y_{P,n} < p_t \end{cases} \tag{12}$$

Where, for a given event X_n, $y_{L,n}$ is the assigned label (1 for signal, 0 for background), $y_{P,n}$ is the assigned XGBoost probability and p_t is the threshold probability. Naively, we could set $p_t = 0.5$, and thus give each event the label to which it is most likely to correspond -



(a) XGBoost predicted probabilities on Training Set

(b) XGBoost predicted probabilities on data

Figure 16. Distribution of probabilities assigned to each training (a) and data (b) event by trained XGBoost classifier. Evident are clear peaks at the signal ($y_{P,n} = 1$) and background ($y_{P,n} = 0$) ends of the scale, with a small distribution in-between - particularly in data.

and looking at Figure 16.(a) this would seem to separate the data well. Playing with this threshold, however, allows us to gain a further degree of freedom in optimising the XGBoost classifier prediction.

A common metric to optimise in particle physics is the ratio of signal events to the (Poisson estimated) variance of the total number of events, defined as:

$$f(p_t) = \frac{S(p_t)}{\sqrt{S(p_t) + B(p_t)}} \tag{13}$$

Explicitly, we seek the threshold probability p_t which gives us the maximal value of $f(p_t)$, where $S(p_t)$ is the number of signal events in our region of interest (i.e. about the expected location of the $X(3823)$ resonance) satisfying $y_{p,n} > p_t$ and $B(p_t)$ is the number of background events in the same region satisfying the same condition (i.e. misclassified as signal). The problem with this approach is that as we are still blinded to the region we cannot directly obtain S or B, so must apply an indirect approach.

Signal efficiency, $\epsilon(p_t)$, is defined $S(p_t) = \epsilon(p_t)S_0$, where S_0 is the number of signal events prior to any cut. As we cannot obtain these for the $X(3823)$ directly, we can instead use the properties of the $\psi(2S)$ and $X(3872)$ signals in our dataset to estimate them. As the variables passed to XGBoost are designed to discern signal from background without saying much about the event itself (i.e. we study the quality of an event being fitted as a signal, rather than properties such as reconstructed mass), the classifier's performance should generalise well between $\psi(2S)$, $X(3872)$ and $X(3823)$. For Monte-Carlo, it is easy to determine efficiency by taking S_0 to be the number of training events labelled as signal prior to any threshold probability being applied, and by taking $S(p_t)$ to be the number of signal events correctly labelled as signal by the classifier for a given threshold probability. For efficiency in data, the fits to the B^+ mass peak as described in Sec.3.1.2 give us an estimate of the number of signal and background events, so can be used to predict the efficiency by considering fits in the $\psi(2S)$ and $X(3872)$ regions in turn prior to and following the application of

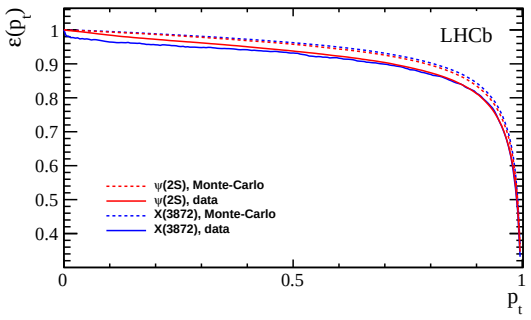


Figure 17. Signal efficiencies for a range of threshold probabilities. $\psi(2S)$ and $X(3872)$ converge in data and Monte-Carlo respectively, but there is a clear discrepancy between data and Monte-Carlo.

a threshold probability. These regions are defined to be the set of all events which satisfy simultaneously a window of $m(B^+)\pm 20$ MeV/c^2 in $m(J/\psi\pi^+\pi^-K^+)$ and $m(X)\pm 20$ MeV/c^2 in $m(J/\psi\pi^+\pi^-)$, where X is the relevant resonance.

Figure 17 shows how signal efficiency changes with threshold probability for a range of values. $\psi(2S)$ and $X(3872)$ converge in both Monte-Carlo and data respectively as the threshold increases, but there is a clear discrepancy between the simulated and real situations. This provides confidence that efficiencies obtained for either of these resonances in the data will generalise well to the $X(3823)$ region, but indicates we must be wary of differences between data and Monte-Carlo. As such, the decision was made to estimate the $X(3823)$ signal efficiency to be the same as $\psi(2S)$ in data. The $\psi(2S)$ was selected over the $X(3872)$ for logistical reasons: its stronger peak is faster to fit to.

An estimate of S_0 was also required. Assuming a factor 20 suppression of the B^+ decay to $X(3823)$ compared to $X(3872)$ - analogous to the relationship between decays to χ_{c1} and χ_{c2} [9]- the estimation $S_{0,X(3823)} = S_{0,X(3872)}/20$ was made. Summarising:

$$S(p_t)_{X(3823)} \simeq \epsilon(p_t)_{\psi(2S)} \times S_{0,X(3872)}/20 \tag{14}$$

As the sideband is by definition entirely background events, $B(p_t)$ was assumed to be equivalent to the number of sideband events in the region 3803 MeV/c^2 $\leq m(J/\psi\pi^+\pi^-) \leq$ 3843 MeV/c^2 satisfying $y_{P,n} > p_t$ - i.e. the number of background events incorrectly identified as signal. A roughly constant background presence is assumed throughout the $m(J/\psi\pi^+\pi^-K^+)$ distribution, so using the sideband is a valid estimation. We then scale this to be of a window ± 20 MeV/c^2 about the $m(B^+)$ peak, which is performed by a simple multiplication of a factor 0.8 to account for the 50 MeV/c^2 width of the sideband.

To optimise the metric, a scan of 200 points was made in the range $0 \leq p_t < 1$ and $f(p_t)$ determined at each point using S and B as defined prior. The optimum was found at $p_{t,\text{opt}} = 0.945$ (Fig.18). The effect of imposing the requirement $y_{P,n} > p_{t,\text{opt}}$ on all data is summarised in Figure 19.

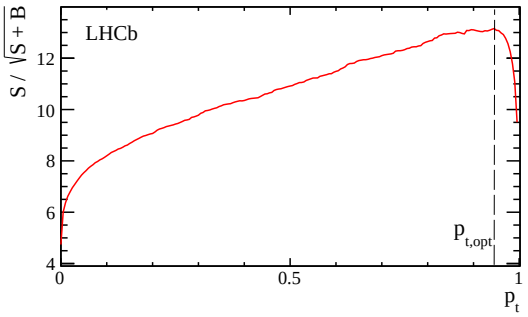
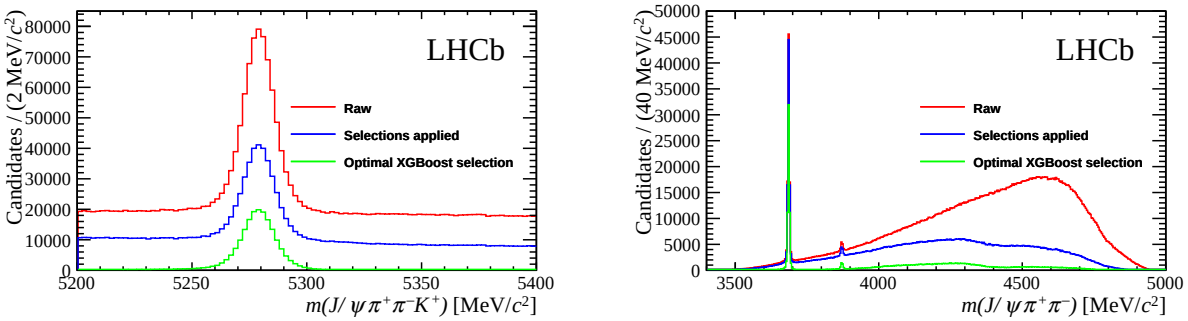


Figure 18. Estimated $X(3823)$ signal efficiency in data for a range of threshold probabilities, peaking at $p_t = 0.945$.



(a) $m(J/\psi\pi^+\pi^-K^+)$ distribution with XGBoost threshold applied

(b) $m(J/\psi\pi^+\pi^-)$ distribution with XGBoost threshold applied

Figure 19. Dataset of $B^+ \to J/\psi\pi^+\pi^-K^+$ from LHCb representing a combined luminosity of 6.5 fb^{-1} (Tab. 3) with initial selections and optimal XGBoost threshold probability applied. Evident in (a) is an almost complete reduction in the number of events distant from the B^+ resonance. In (b) there is a significant reduction in the rising background, and a slight reduction of the $\psi(2S)$ and $X(3872)$ resonances.

3.4 Significance estimation

With all selections thus applied, the focus was shifted to searching for a resonance in the $X(3823)$ region. The experiment remained blinded until a model for fitting to the region was fully developed.

3.4.1 Resolution model

As the resonance is in the $m(J/\psi\pi^+\pi^-)$ distribution, a new fit model was required. At this stage in the experiment, the $X(3823)$ region remained blinded so this model was developed on the $\psi(2S)$ and $X(3872)$ resonances, then interpolated for application to the $X(3823)$. As we sought the mass and branching fraction of the $X(3823)$, the fit models developed for the $\psi(2S)$ and $X(3872)$ resonances were used to fix as many parameters as possible: leaving free only signal yield, background yield and mean mass.

The fits to the $\psi(2S)$ and $X(3872)$ now required two components to consider: at the mass range in which $m(J/\psi\pi^+\pi^-)$ for both resonances sit, the detector has sufficient resolution to observe the relativistic Breit-Wigner shape of each resonance. As such, the total distribution to fit consisted of both the intrinsic Breit-Wigner shape of each resonance plus the detector effects. To account for detector effects, two Crystal Balls were used: one with the exponent of its power law tail component fixed to $n = 1$ to account for QED effects, and the other with its exponent of the power law tail free. For both crystal balls, the power law tail threshold values were parameterised as:

$$\alpha = 2.48 \times \frac{59.5\sigma}{1 + 59.5\sigma} \tag{15}$$

Where σ is the width of the associated Crystal Ball function. The was done to reduce the number of free parameters required when fitting.

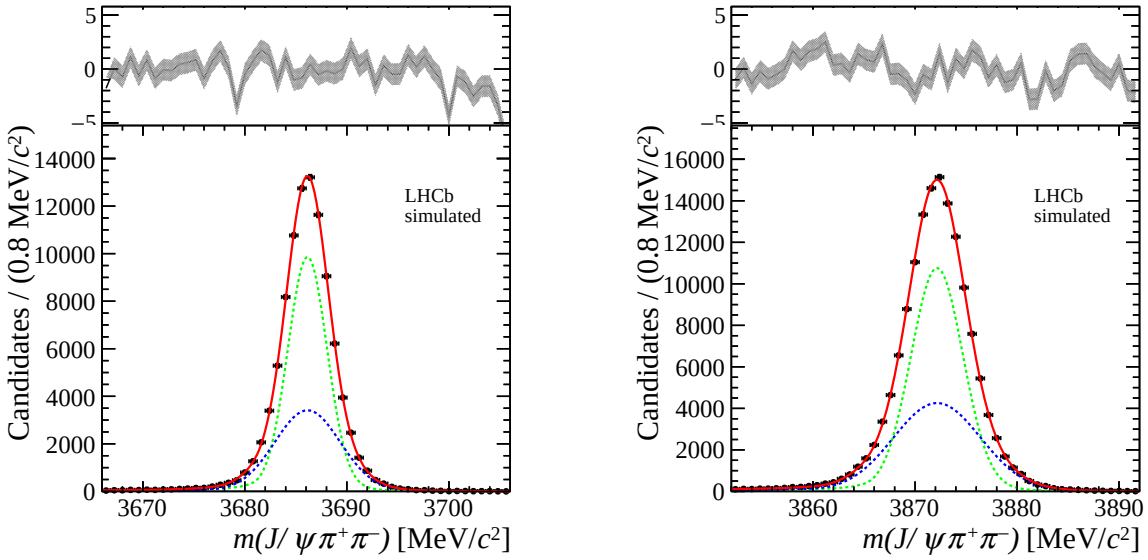
Each Crystal Ball was then separately convolved with a relativistic Breit-Wigner, with values for the mean mass and width for the $\psi(2S)$ and $X(3872)$ taken from the PDG [9]. For each Crystal Ball function f_{CBi}, the complete signal function is thus given by:

$$f_{\mathrm{sig,i}} = \int_{m_{\min}}^{m_{\max}} B(m'; M, \Gamma) f_{\mathrm{CBi}}(m - m'; \bar{m}, \sigma_i, n_i) dm' \tag{16}$$

Where B is the relativistic Breit-Wigner function (with M and Γ fixed separately for the two resonances) and $m_{\max}$ and $m_{\min}$ accounting for limits applied to the Breit-Wigner distribution applied when generating MC data for computational efficiency. The convolution process maintains the unit normalisation of the Crystal Ball functions, so to combine the contribution of both a free coefficient r (as opposed to the equal contribution of each Crystal Ball in Sec.3.1.2) was incorporated in the final signal model through:

$$f_{\mathrm{sig}} = r f_{\mathrm{sig},1}(m; \bar{m}, n_1 = 1, \sigma_1) + (1 - r) f_{\mathrm{sig},2}(m; \bar{m}, n_2, \sigma_2) \tag{17}$$

This fit model was sufficient for the Monte-Carlo datasets, which contained only signal events by definition. As in the case of the B^+ mass fits, these were used to obtain initial values for



(a) $m(J/\psi\pi^+\pi^+)$ distribution for $\mathrm{MC}^{\psi(2S)}$

(b) $m(J/\psi\pi^+\pi^+)$ distribution for $\mathrm{MC}^{X(3872)}$

Figure 20. Double Crystal Ball convolved with Relativistic Breit-Wigner fits to Monte-Carlo generated $m(J/\psi\pi^+\pi^+)$ distributions for $\psi(2S)$ and $X(3872)$ resonances. $f_{\mathrm{sig},1}$ is shown in green and $f_{\mathrm{sig},2}$ in blue. Fit parameters are summarised in Table 9.

many of the fit parameters to make optimising in data more feasible. These Monte-Carlo fits are shown in Figure 20, and the fit parameters summarised in Table 9.

To account for discrepancies between data and MC, the same fit was also performed for data in the $\psi(2S)$ region, with all parameters fixed to their MC value except the widths. These were freed, with their relative size maintained, by multiplying by a floated scaling factor γ (i.e. $\sigma_{\mathrm{data},i} = \gamma\sigma_{\mathrm{MC},i}$). In addition, an exponential fit to the background was made, with signal and background yield parameters included as in the fits to B^+ mass. This fit is shown in Figure 21, with parameters summarised in Table 9.

With these fit models, the $X(3823)$ resolution function could then be interpolated. It is known that the detector resolution scales with energy (and therefore mass) via the relation:

$$\sigma(m) \propto m^{1/2} \tag{18}$$

With this functional knowledge, an interpolation was performed separately for the (scaled) width of each Crystal Ball to best infer the resolution at the $X(3823)$ mass. This was done assuming a resolution of zero at a threshold mass $m_{\mathrm{th}} = m_{J/\psi} + 2m_\pi = 3376$ MeV/c^2, as in [64]. The best estimates obtained were $\sigma_1 = 2.59$ and $\sigma_2 = 4.39$ (Fig.22).

3.4.2 Background model

Within the $X(3823)$ region, there is an irreducible component of the background originating from the decay of $B^+ \to J/\psi(K_1(1270) \to K^+\pi^+\pi^-)$ which has all the same daughter

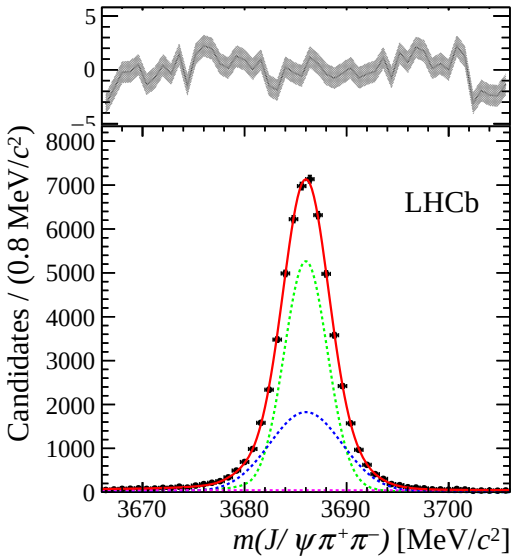


Figure 21. Double Crystal Ball convolved with Relativistic Breit-Wigner fit to $m(J/\psi\pi^+\pi^+)$ distribution for data in the $\psi(2S)$ region. $f_{\rm sig,1}$ is shown in green, $f_{\rm sig,2}$ in blue and an exponential background component in magenta. Fit parameters are summarised in Table 9.

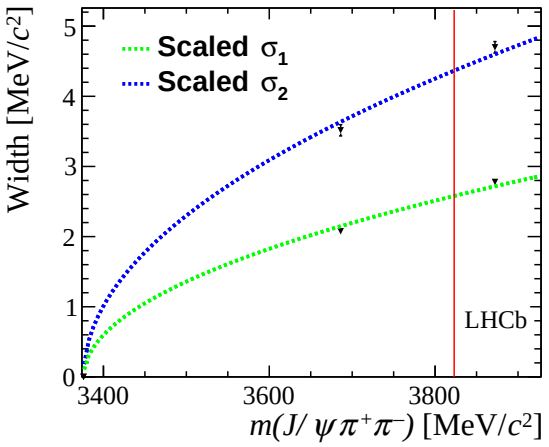


Figure 22. Interpolation of the scaled resolution parameter for the Double Crystal Ball resolution model to be applied to the $X(3823)$ region. The red line denotes the expected $X(3823)$ mass, returning a best estimate of $\sigma_1 = 2.59$ and $\sigma_2 = 4.39$.

$m(J/\psi\pi^+\pi^-)$ Fit Parameters			
Parameter	$\psi(2S)$	$X(3872)$	X(3823)
$m_{\min}$ [MeV/c^2]	3685.25	3977.0	n/a
$m_{\max}$ [MeV/c^2]	3690.65	3896.3	n/a
M [MeV/c^2]	3686.11	3871.5	n/a
Γ [MeV/c^2]	0.304	0.317	n/a
$\bar{m}$ [MeV/c^2]	3686.14(1)	3872.16(1)	3823.7(13)
n_1	1	1	1
n_2	1.10(11)	0.46(5)	0.46
σ_1 [MeV/c^2]	1.86(3)	2.49(4)	2.59
σ_2 [MeV/c^2]	3.15(7)	4.22(7)	4.39
r	0.64(3)	0.60(2)	0.60
γ	1.107(5)	n/a	n/a
c [$\times 10^{-3}$ c^2/MeV]	-3(2)	n/a	4.9(5)
N_{sig}	57100(300)	n/a	60(20)
N_{bgr} [$\times 10^3$]	2.22(9)	n/a	4.19(7)

Table 9. Parameters for the double Crystal Ball fit to $\psi(2S)$ and $X(3872)$ Monte-Carlo $m(J/\psi\pi^+\pi^-)$ distributions, and the double Crystal Ball plus exponential background fit to the $\psi(2S)$ and $X(3823)$ distributions in data. Values shown without error are fixed prior to fitting.

particles as the decay we are interested in and will therefore commonly be included in our dataset. Since we train the decision tree to separate true signal events from coincidentally reconstructed ones, it will be poor at removing these decays from the region. To account for their presence, we can instead determine what their distribution will look like and thus produce a fit component on top of which we can add the interpolated $X(3823)$ shape. To determine a good model to apply to this background, a simulation of 600,000 $B^+ \to J/\psi(K_1(1270) \to K^+\pi^+\pi^-)$ decays was prepared using RapidSim [65]. The performance of three simple models - exponential, 1st order Chebychev polynomial and 2nd order Chebychev polynomial [66] - is virtually identical in the region where we search for the $X(3823)$ (Fig.23), so the simplest method of fitting an exponential function to account for the background was selected.

3.4.3 Fitting to the $X(3823)$ region

With the background and resolution models developed, we could then apply the combined fit model to the $X(3823)$ region of the data. As the exact mass is unknown, a broad range of 3765MeV/c^2 $\leq m(J/\psi\pi^+\pi^-) \leq$ 3845MeV/c^2 was considered.

All of the parameters for the resolution function determined for the $X(3872)$ were used due to the relative proximity of this resonance, with the interpolated widths substituted. The only parameter floated in this portion of the fit was the mean mass (i.e. the x-position), left

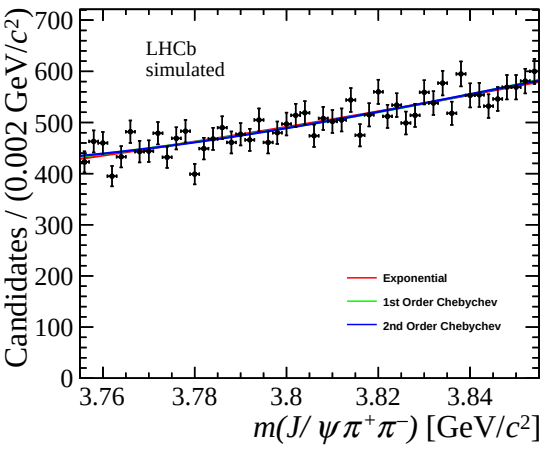


Figure 23. The fit of three different functions to the $X(3823)$ region of 600,000 simulated $B^+ \to J/\psi(K_1(1270) \to K^+\pi^+\pi^-)$ decays: an irreducible background component present in the analysis dataset. The performance of each model is virtually identical so the simplest exponential function was selected.

to vary freely in the entire mass range considered. This was combined with the exponential background as described prior. The relative normalisations of the two components were again achieved using yields, to give a combined fit function of:

$$F_{\text{tot}} = N_{\text{sig}} f_{\text{sig}}(m; \bar{m}) + N_{\text{bgr}} e^{cm} \tag{19}$$

The Breit-Wigner component of the $X(3823)$ was neglected for this initial fit, as the primary goal of the analysis was to determine only the presence of a signal. The fit parameters are summarised in Table 9; of particular note is the placement of the optimal signal component at a mass of 3823.7(13) MeV/c^2 and a not-insignificant signal yield of around 60 events. Figure 24 shows the fit obtained.

4 Results

4.1 Significance of $X(3823)$ signal fit

The result from fitting to the $X(3823)$ is encouraging in its placement of the optimal signal component at a mass of 3823.7(13) MeV/c^2, however we require a statistical significance of 3σ to quote "evidence" of the resonance and 5σ to quote "observation". RooFit returns the likelihood for a given model and dataset, so a simple hypothesis test was performed to determine the significance of the fit model obtained. With a null hypothesis corresponding to the absence of any signal (i.e. N_{sig} fixed to zero), Wilks' theorem [67] was used to determine a significance of 2.4σ.

This value is not sufficient to constitute evidence, but the fit does provide encouragement as it naturally falls near the expected mass.

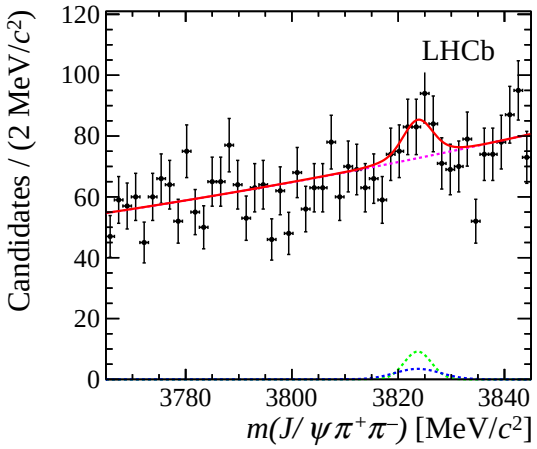


Figure 24. The fit of the two interpolated Crystal Ball functions (f_{CB1} green, f_{CB2} blue) plus an exponential background (magenta). The optimal fit places the signal component at a mass of 3823.7(13) MeV/c^2: encouragingly close to the expected value for the $X(3823)$.

4.2 Limit on $X(3823)$ production

The significance does not therefore allow us to draw any conclusions on the presence of the $X(3823)$ resonance in $B^+ \to J/\psi\pi^+\pi^-K^+$ decays, or on its properties. Another key parameter though in particle physics is the branching fraction: how likely a given particle is to follow a specific decay route. We can use this analysis to provide an estimate of:

$$\mathcal{B}r\big(B^+ \to X(3823)K^+\big) \;\times\; \mathcal{B}r\big(X(3823) \to J/\psi\pi^+\pi^-\big) \tag{20}$$

This was done by replacing the signal yield component of the fit to the $X(3823)$ region by a function of the branching fraction; the new parameter free to fit:

$$N_{\mathrm{sig}} = \frac{\mathcal{B}r\big(B^+ \to X(3823)K^+\big) \;\times\; \mathcal{B}r\big(X(3823) \to J/\psi\pi^+\pi^-\big)}{f_{\mathrm{norm}}} \tag{21}$$

Where f_{norm} is a normalisation factor that allows us to determine the branching fraction for the $X(3823)$ by comparing to a known value. If this comparison is performed relative to the $\psi(2S)$ - the most prevalent signal in our dataset - this is defined:

$$f_{\mathrm{norm}} = \frac{\mathcal{B}r\big(B^+ \to \psi(2S)K^+\big) \times \mathcal{B}r\big(\psi(2S) \to J/\psi\pi^+\pi^-\big)}{N(\psi(2S))} \;\times\; \frac{\epsilon_{\psi(2S)}}{\epsilon_{X(3823)}} \;\times\; \frac{\mathcal{L}_{\psi(2S)}}{\mathcal{L}_{X(3823)}} \tag{22}$$

Since both resonances are in the same datasets the luminosities ($\mathcal{L}_X$) cancel, and as we have assumed efficiencies (ϵ_X) generalise across resonances in data (Fig.17), these cancel too. $N_{\psi(2S)}$ is taken from the fit to the B^+ mass peak in the $\psi(2S)$ region as described earlier, and the branching fractions are known [9].

An iterative likelihood ratio method to generate a 90% confidence interval on the branching fraction was then applied. A scan across 400 mass values in the interval 3765 MeV/c^2 $\leq$

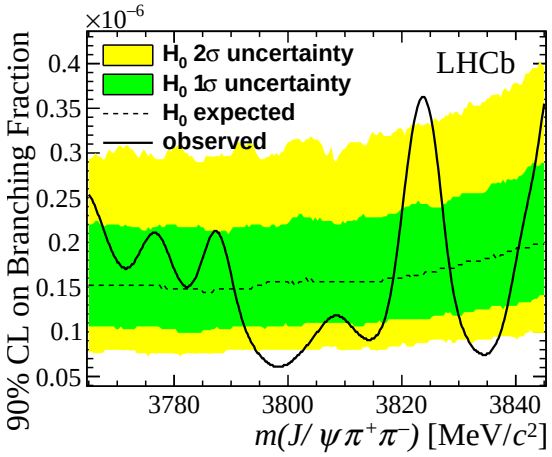


Figure 25. Brazil-band [68] plot for the 90% confidence limit on $\mathcal{B}r\big(B^+ \to X(3823)K^+\big) \times \mathcal{B}r\big(X(3823) \to J/\psi\pi^+\pi^-\big)$. The dotted line represents the expected branching fraction limit for the null hypothesis (i.e. no signal present). The coloured bands represent the uncertainty on this expectation, determined from 1000 simulated datasets. We see an excess in limit around the expected $X(3823)$ mass.

$m(J/\psi\pi^+\pi^-) \leq 3845$ MeV/c^2 was performed, with the likelihood estimation made at each. At each scan point m', $\bar{m}$ was fixed to the value m', then the value of the branching fraction in the model iteratively stepped 300 times in the range $0 \leq \mathcal{B}r \leq 150 f_{\rm norm}$. At each point the likelihood of the produced model was determined, with the 90% confidence interval on branching fraction for a given m' being the value which enclosed 90% of the total integrated likelihood.

To get a sense of how likely these values were to correspond to a fluctuation in noise only, the same procedure was repeated 1000 times using"toy" datasets generated from a probability density function representative of the background component of the fit to the $X(3823)$ region. For each dataset, 4200 events (roughly the fit background yield) were generated from the pdf:

$$p(m) = e^{0.0049m} \tag{23}$$

In the same range 3765 MeV/c^2 $\leq m(J/\psi\pi^+\pi^-) \leq$ 3845 MeV/c^2. The comparison of the distribution of these noise-only fits to the true data can be seen in Figure 25. Encouragingly, we again see a deviation above the background about the expected mass of the $X(3823)$. Taking the optimal fit (i.e. the one that maximises likelihood), we can therefore place a confidence limit on the branching fraction $\mathcal{B}r\big(B^+ \to X(3823)K^+\big) \times \mathcal{B}r\big(X(3823) \to J/\psi\pi^+\pi^-\big) < 3.6 \times 10^{-7}$.

Considering the optimal fit gives a value of $\mathcal{B}r\big(B^+ \to X(3823)K^+\big) \times \mathcal{B}r\big(X(3823) \to J/\psi\pi^+\pi^-\big) = 2.3(9) \times 10^{-7}$, where the error is statistical only: see the following section for a discussion of systematics.

4.3 Systematic uncertainties

There are various sources of systematic uncertainty throughout the analysis where variations in method have the potential to affect the result.

Changing the B^+ fit model from a two-tailed Crystal Ball plus Gaussian to the double Crystal Ball shape applied to the resonances in $m(J/\psi\pi^+\pi^-)$ fit produces a clearly inferior distribution so its effects are not considered.

Fixing the $X(3823)$ resolution model to the parameters obtained by fitting to the $\psi(2S)$, as opposed the $X(3872)$ results in a change to the obtained branching fraction of $\Delta\mathcal{B}r = 0.1 \times 10^{-7}$.

Estimation methods of the optimal decision tree probability threshold (e.g. using the Punzi condition [69], optimising with MC efficiencies, changing the efficiency scaling factor) result in a small fluctuation of the threshold value between $0.925 \leq p_t \leq 0.975$. Values around the lower end of this range have no effect on the obtained value, but larger values do manifest as a reduction of up to $\Delta\mathcal{B}r = 0.3 \times 10^{-7}$.

By far the largest effect on the observed branching fraction is produced when the resolution used to normalise the function is the $X(3872)$: the branching fraction is a factor of two smaller, suggesting a deeper issue. We believe this is as the value of N_X used in the formula is determined from fits to the B^+ mass: much of the background in the $m(J/\psi\pi^+\pi^-)$ distribution is known to be due to the irreducible presence of $B^+ \to J/\psi(K_1(1270) \to K^+\pi^+\pi^-)$ decays. As these are true signal events, their $m(J/\psi\pi^+\pi^-K^+)$ distribution will be within the B^+ peak also, so will result in an increased number of suspected events. As this is a rising background, the effect of this will be much more significant for the B^+ mass fit of $X(3872)$ region events (Fig.21 shows a nominal background). This would suggest that N_{sig} would be better estimated from the $m(J/\psi\pi^+\pi^-)$ distributions when specific resonances are concerned. This could also explain the deviation from efficiencies obtained on MC datasets - they do not have the irreducible background component included. Given the relative scale of the issue for the $X(3872)$ resonance, this normalisation factor is not considered. Determining f_{norm} using $N_{\psi(2S)}$ determined this way has little effect on the determined branching fraction, expected as the change in number of events here is nominal.

Combining these effects in quadrature, we therefore conclude a systematic uncertainty of $\Delta\mathcal{B}r = 0.3 \times 10^{-7}$.

5 Conclusions

We can therefore conclude this analysis has produced optimistic signs of an $X(3823)$ resonance: a signal component of statistical significance 2.4σ is determined at a mass of 3823.7(13) MeV/c^2. This is not sufficiently significant to constitute evidence, but given it sits at a mass in agreement with the expected value, it is encouraging.

Using the same fit model, we are able to place a confidence limit on the branching fraction of $\mathcal{B}r\big(B^+ \to X(3823)K^+\big) \times \mathcal{B}r\big(X(3823) \to J/\psi\pi^+\pi^-\big) < 3.6 \times 10^{-7}$ at $CL = 90\%$. In addition, we find the optimally fitted branching fraction value to be $\mathcal{B}r\big(B^+ \to X(3823)K^+\big) \times \mathcal{B}r\big(X(3823) \to J/\psi\pi^+\pi^-\big) = (2.3 \pm 0.9 \pm 0.3) \times 10^{-7}$, where the first error is statistical and the second systematic. These values are aligned with what we would expect for a (rough) factor 20 suppression compared with B^+ meson decays to the $X(3872)$[9].

There are a few obvious areas for development. Firstly, it is difficult to quantify exactly how the issue in determining N_X from the B^+ peak only propagates back through analysis: this should be addressed before any further steps are taken.

The key open area in the analysis is the statistical significance of the fit. The optimal mass for the fit and the proximity of the significance to the value required for evidence are encouraging signs. Efforts should be made to try to improve this significance. As the remainder of the 2017 LHCb data is processed, it can be added to the analysis without modification. In addition, a dataset of $B \to J/\psi\pi^+\pi^-\pi^+$ decays (i.e. the final K^+ is substituted for a π^+) is available, and could be added to the analysis with only minor modification. Perhaps with additional events more $X(3823)$ decays will be present.

No consideration of the reconstructed $m(J/\psi\pi^+\pi^-K^+)$ is made when making the final fit to the $X(3823)$ region. By considering events within a narrow 40MeV/c^2 window about the B^+ peak only, the significance of the $X(3823)$ fit can be improved to 3.2σ as a portion of the background is further reduced. This simple cut lacks robustness so is not a valid method, but provides encouragement that perhaps consideration of the sWeights produced by fitting to the final $m(J/\psi\pi^+\pi^-K^+)$ distribution could further reduce background sufficiently to obtain a statistically significant result.

There is something of a dark art to producing Machine Learning models, it is always possible that further feature engineering, optimisation and so on could result in a multivariate selection model with improved signal-background separation.

If a statistically significant result could be obtained, the next step would be to gather a dataset large enough to accurately determine the quantum numbers of the $X(3823)$ resonance to cement our understanding of its place within the charmonium sector. For an example of such an analysis, see: [16].

References

[1] Murray Gell-Mann. A schematic model of baryons and mesons. In *Murray Gell-Mann: Selected Papers*, pages 151–152. World Scientific, 2010.

[2] Vincent Mathieu, Nikolai Kochelev, and Vicente Vento. The physics of glueballs. *International Journal of Modern Physics E*, 18(01):1–49, 2009.

[3] Bernhard Ketzer. Hybrid mesons. *arXiv preprint arXiv:1208.5125*, 2012.

[4] John Weinstein and Nathan Isgur. $K\bar{K}$ molecules. *Physical Review D*, 41(7):2236, 1990.

[5] Richard F Lebed, Ryan E Mitchell, and Eric S Swanson. Heavy-quark QCD exotica. *Progress in Particle and Nuclear Physics*, 93:143–194, 2017.

[6] V. Bhardwaj et al. Evidence of a new narrow resonance decaying to $\chi_{c1}\gamma$ in $B \to \chi_{c1}\gamma K$. *Phys. Rev. Lett.*, 111(3):032001, 2013.

[7] E Eichten, K Gottfried, T Kinoshita, KD Lane, and T-M Yan. Charmonium: the model. *Physical Review D*, 17(11):3090, 1978.

[8] T Barnes, S Godfrey, and ES Swanson. Higher charmonia. *Physical Review D*, 72(5):054026, 2005.

[9] C. Patrignani et al. Review of Particle Physics. *Chin. Phys.*, C40(10):100001, 2016.

[10] Gregory Breit and Eugene Wigner. Capture of slow neutrons. *Physical review*, 49(7):519, 1936.

[11] J. E. Augustin, A. M. Boyarski, M. Breidenbach, F. Bulos, J. T. Dakin, G. J. Feldman, G. E. Fischer, D. Fryberger, G. Hanson, B. Jean-Marie, R. R. Larsen, V. Lüth, H. L. Lynch, D. Lyon, C. C. Morehouse, J. M. Paterson, M. L. Perl, B. Richter, P. Rapidis, R. F. Schwitters, W. M. Tanenbaum, F. Vannucci, G. S. Abrams, D. Briggs, W. Chinowsky, C. E. Friedberg, G. Goldhaber, R. J. Hollebeek, J. A. Kadyk, B. Lulu, F. Pierre, G. H. Trilling, J. S. Whitaker, J. Wiss, and J. E. Zipse. Discovery of a narrow resonance in e^+e^- annihilation. *Phys. Rev. Lett.*, 33:1406–1408, Dec 1974.

[12] J. J. Aubert, U. Becker, P. J. Biggs, J. Burger, M. Chen, G. Everhart, P. Goldhagen, J. Leong, T. McCorriston, T. G. Rhoades, M. Rohde, Samuel C. C. Ting, Sau Lan Wu, and Y. Y. Lee. Experimental observation of a heavy particle J. *Phys. Rev. Lett.*, 33:1404–1406, Dec 1974.

[13] V Ruuskanen and Nils A Toernqvist. The Okubo-Zweig-Iizuka rule and unitarity. Technical report, Helsinki Univ.(Finland). Research Inst. for Theoretical Physics, 1977.

[14] M Ablikim, MN Achasov, XC Ai, O Albayrak, M Albrecht, DJ Ambrose, A Amoroso, FF An, Q An, JZ Bai, et al. Observation of the ψ 1^3D_2 state in $e^+e^- \to \pi^+\pi^-\gamma\chi_{c1}$ at BESIII. *Physical review letters*, 115(1):011803, 2015.

[15] S-K Choi, SL Olsen, K Abe, T Abe, I Adachi, Byoung Sup Ahn, H Aihara, K Akai, M Akatsu, M Akemoto, et al. Observation of a narrow charmoniumlike state in exclusive $B^{\pm} \to K^{\pm}\pi^{+}\pi^{-}j/\psi$ decays. *Physical review letters*, 91(26):262001, 2003.

[16] R Aaij, C Abellan Beteta, B Adeva, M Adinolfi, C Adrover, A Affolder, Z Ajaltouni, J Albrecht, F Alessio, M Alexander, et al. Determination of the $X(3872)$ meson quantum numbers. *Physical review letters*, 110(22):222001, 2013.

[17] LHCb Public. http://lhcb-public.web.cern.ch/lhcb-public/.

[18] CERN. https://home.cern/topics/large-hadron-collider.

[19] CERN. https://home.cern/.

[20] ATLAS Collaboration. A particle consistent with the higgs boson observed with the atlas detector at the large hadron collider. *Science*, 338(6114):1576–1582, 2012.

[21] LHCb Collaboration. Lhcb detector performance. *International Journal of Modern Physics A*, 30(07):1530022, 2015.

[22] Daedalus. https://www.nevis.columbia.edu/daedalus/motiv/cp.html.

[23] R Aaij, C Abellan Beteta, A Adametz, B Adeva, M Adinolfi, C Adrover, A Affolder, Z Ajaltouni, J Albrecht, F Alessio, et al. First evidence for the decay $B_s^0 \to \mu^+\mu^-$. *Physical review letters*, 110(2):021801, 2013.

[24] R Aaij, C Abellan Beteta, B Adeva, M Adinolfi, C Adrover, A Affolder, Z Ajaltouni, J Albrecht, F Alessio, M Alexander, et al. Differential branching fraction and angular analysis of the decay $B^0 \to K^0\mu^+\mu^-$. *Journal of High Energy Physics*, 2013(8):131, 2013.

[25] R Aaij, C Abellan Beteta, B Adeva, M Adinolfi, C Adrover, A Affolder, Z Ajaltouni, J Albrecht, F Alessio, M Alexander, et al. Measurement of the CKM angle γ from a combination of $B^{\pm} \to D^{\pm}h$ analyses. *Physics Letters B*, 726(1-3):151–163, 2013.

[26] BaBar. http://www-public.slac.stanford.edu/babar/.

[27] Belle2. https://www.belle2.org/.

[28] R Antunes Nobrega, A Franca Barbosa, I Bediaga, G Cernicchiaro, E Correade Oliveira, J Magnin, J Marquesde Miranda, A Massafferri, E Polycarpo, A Reis, et al. LHCb reoptimized detector design and performance: Technical design report. 2003.

[29] Mark Thomson. *Modern particle physics*. Cambridge University Press, 2014.

[30] LHCb. http://lhcb-reconstruction.web.cern.ch/lhcb-reconstruction/Panoramix/XYZStates/.

[31] Victor Lavrenko. Decision trees. *Introductory Applied Machine Learning, INFR10069, The University of Edinburgh*, 2014.

[32] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[33] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[34] James Gillies. https://home.cern/cern-people/opinion/2011/03/luminosity-why-dont-we-just-say-collision-rate.

[35] LHCb. http://lhcbdoc.web.cern.ch/lhcbdoc/davinci/.

[36] G Corti. Overview of monte carlo simulation(s) in LHCb, 2009.

[37] Tomasz Skwarnicki. *A study of the radiative cascade transitions between the Upsilon-prime and Upsilon resonances*. PhD thesis, Institute of Nuclear Physics, Krakow, 1986. DESY-F31-86-02.

[38] Rene Brun and Fons Rademakers. Rootan object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1-2):81–86, 1997.

[39] Bjarne Stroustrup. *The C++ Programming Language*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition, 2000.

[40] Wouter Verkerke and David P. Kirkby. The RooFit toolkit for data modeling. *eConf*, C0303241:MOLT007, 2003. [,186(2003)].

[41] M Pivka and FR Le Diberderb. a statistical tool to unfold data distributions. *arXiv preprint physics/0402083*.

[42] Alex Rogozhnikov. Reweighting with boosted decision trees. *Journal of Physics: Conference Series*, 762:012036, 2016.

[43] https://arogozhnikov.github.io/hep_ml/reweight.html.

[44] Alex Rogozhnikov. https://github.com/arogozhnikov/hep_ml.

[45] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[46] Wouter D Hulsbergen. Decay chain fitting with a Kalman filter. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 552(3):566–575, 2005.

[47] Brian R Martin and Graham Shaw. *Particle physics*. Wiley, 2007.

[48] LHCb Collaboration. Quantum numbers of the $X(3872)$ state and orbital angular momentum in its $\rho^0 j/\psi$ decay. *Phys. Rev. D*, 92:011102, Jul 2015.

[49] Cheuk-Yin Wong. *Introduction to high-energy heavy-ion reactions*. World Scientific, 1994.

[50] Liliana Teodorescu. Artificial neural networks in high-energy physics. 2008.

[51] Byron P Roe, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2-3):577–584, 2005.

[52] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[53] Adrian S Lewis and Michael L Overton. Nonsmooth optimization via BFGS. *Submitted to SIAM J. Optimiz*, pages 1–35, 2009.

[54] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[55] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[56] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[57] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[58] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. SciPy Austin, TX, 2010.

[59] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[60] Tianqi Chen and Tong He. Higgs boson discovery with boosted trees. In *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, pages 69–80, 2015.

[61] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[62] Tim Head et al. https://scikit-optimize.github.io/.

[63] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[64] R. Aaij et al. Observation of $X(3872)$ production in pp collisions at $\sqrt{s} = 7$ TeV. *Eur. Phys. J.*, C72:1972, 2012.

[65] Greig A Cowan, DC Craik, and MD Needham. Rapidsim: An application for the fast simulation of heavy-quark hadron decays. *Computer Physics Communications*, 214:239–246, 2017.

[66] Pafnuti Lvovitch Tchebychev. *Théorie des mécanismes connus sous le nom de parallélogrammes*. Imprimerie de l'Académie impériale des sciences, 1853.

[67] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[68] Chris Wymant. Brazil-band plots for dummies. *lapth. cnrs. fr/pg-nomin/wymant/BrazilBandPlot. pdf*, 2012.

[69] Giovanni Punzi. Sensitivity of searches for new signals and its optimization. *arXiv preprint physics/0308063*, 2003.