LIP-STUDENTS-25-26

B mesons studies in LHC Run 3

Ricardo Ribeiro[1,a], Mariana Soares[2,b], Pedro Chan[2,c] and Lekai Yao[3,d]

[1] *Instituto Superior Técnico, Lisboa, Portugal*
[2] *Faculdade de Ciências, Lisboa, Portugal*
[3] *Tsinghua University, Beijing, China*

Project supervisors: Henrique Legoinha, Nuno Leonardo *October 4, 2025*

Abstract. In the early stages of heavy-ion collisions, bottom quarks are produced, then they traverse and interact with Quark-Gluon Plasma (QGP) medium, later they hadronize by pairing with another quark, forming a B meson, which preserves information about the QGP effect on its proprieties. The production of B mesons in proton-proton collision was studied at a center-of-mass energy of $\sqrt{s} = 5.36\ TeV$. The B^+, B^0, B_s mesons were reconstructed through the decay chains: $B^+ \to J/\psi\ K^+ \to \mu^+\ \mu^-\ K^+$, $B^0 \to J/\psi\ K^*(892)^0 \to \mu^+\ \mu^-\ K^+\ \pi^-$, $B_s^0 \to J/\psi\ \phi(1020) \to \mu^+\ \mu^-\ K^+\ K^-$. The data were collected with the CMS detector in 2024 during LHC Run3 and correspond to an integrated luminosity of 455 pb^{-1}. Signal extraction was done by using Cut-based technique, followed by an Extended Unbinned Maximum Likelihood fit with RooFit. The sPlot method and Sideband Subtraction method were used to perform MC Validation. This work is a baseline to compare with future multivariate (MVA) analysis in pp and Heavy-ion (PbPb) collision data. This shall allow the study of the QGP effect on B meson production by measuring the Nuclear Modification Factor R_{AA}.

KEYWORDS: LHC, CMS, b quark, B mesons, QGP

Contents

[a] e-mail: ricardo.silva.ribeiro@tecnico.ulisboa.pt
[b] e-mail: fc57141@alunos.ciencias.ulisboa.pt
[c] e-mail: fc60014@alunos.ciencias.ulisboa.pt
[d] e-mail: lekai.yao@cern.ch

1 Introduction

1.1 Motivation

On the early stages of Heavy-ion collisions, bottom quarks are produced, then they traverse and interact with the Quark-Gluon Plasma medium (produced in Heavy-ion collisions at the LHC), and in end hadronize forming the B hadrons. Measuring the B hadron properties allows us to study properties of the QGP.

The study of different B meson species through the analysis of pp and PbPb collision data collected at the LHC by the CMS experiment, allows to study underlying mechanisms that lead to effects such as supression or enhancement, by measuring the Nuclear Modification Factors R_{AA}.

As pp collisions are expected to produce a negligible amount of QGP, they pose as a baseline for heavy ion studies. The B mesons have long lifetimes, leading to a significant and measurable flight length, which makes it particularly interesting to study.

In this internship, the goal is to analyze pp collision data collected by CMS at the LHC, to study B^+, B^0, B_s mesons using Cut-Based Techniques and Single-Variable Optimization. This further serves as baseline to future study employing Machine Learning methods, in pp but also PbPb collision data.

1.2 The CMS detector

The CMS detector (Compact Muon Solenoid) is one of the general purpose detectors at the LHC, in CERN. It has a cylindrical shape and consists of a central region where the collisions occur. It is layered into several sub-detectors, of which the most relevant for this project are the silicon trackers and muon chambers. Particles are reconstructed using algorithms that combine the signals provided by the sub-detectors. There is also a Trigger system that decides in real time whether to discard or keep events (a full description can be found in [1]).

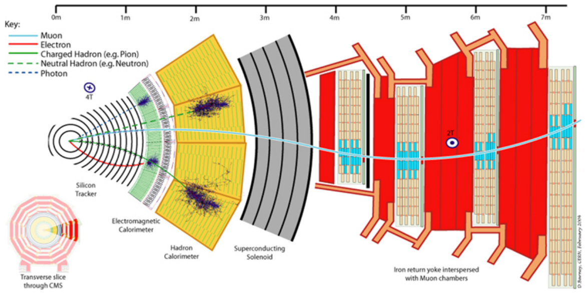

Figure 1: Schematic transverse view of the CMS detector

1.3 Relevant variables

Here is a list of the main variables that can be measured with the detector and were used in the selection:

- Normalized Flight Length: 3D distance between the beam spot and the vertex where B meson decayed, normalized by its uncertainty;
- Normalized Flight Length in transversal plane: distance between the beam spot and the vertex where B meson decays in transversal plane, normalized by its uncertainty;
- Pointing Angle (α): angle between the flight direction and the reconstructed B meson momentum;
- Pointing Angle in transverse plane (θ): pointing angle between the flight direction and the reconstructed B meson momentum in transverse plane;
- dR: Angle between the J/ψ and hadron track;
- tktkmass: Sum of invariant masses of two hadron tracks.

2 Datasets

2.1 Dataset information

In this study, the pp collision data at the center-of-mass energy $\sqrt{s} = 5.36\, TeV$ were analyzed. The data were collected with the CMS Detector in 2024 during LHC Run 3 and correspond to a luminosity of 455 pb^{-1}.

The B mesons are reconstructed as:

$B^+ \rightarrow J/\psi\ K^+ \rightarrow \mu^+\ \mu^-\ K^+$,
$B^0 \rightarrow J/\psi\ K^*(892)^0 \rightarrow \mu^+\ \mu^-\ K^+\ \pi^-$,
$B_s^0 \rightarrow J/\psi\ \phi(1020) \rightarrow \mu^+\ \mu^-\ K^+\ K^-$.

2.2 Preselection

The goal is to visualize and extract the B meson Signals from the data. Without any selection applied, the signal peaks are not visible in the invariant mass spectra. This is because, besides the intended signals, the data contains many other processes that form Backgrounds, mimicking and hiding the signals. Preliminary selection cuts are thus applied to the data, with the goal of reducing the level of background contamination, and revealing the signals.

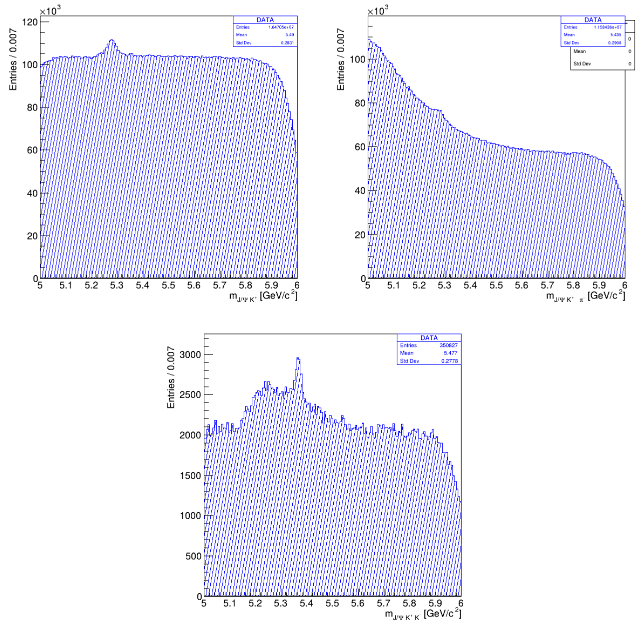

Figure 2: B^+, B_0 and B_s^0 Mass Spectrum after Preselection

2.2.1 Muon & J/ψ selection

The muon candidates are selected according to the *soft-muon* criteria. All muon candidates must fulfill the following acceptance selections:

$$
\begin{aligned}
p_T^\mu &> 3.5\,\mathrm{GeV}/c && \text{for } |\eta^\mu| < 1.2 \\
p_T^\mu &> (5.47 - 1.89\,|\eta^\mu|)\,\mathrm{GeV}/c && \text{for } 1.2 \le |\eta^\mu| < 2.1 \\
p_T^\mu &> 1.5\,\mathrm{GeV}/c && \text{for } 2.1 \le |\eta^\mu| < 2.4
\end{aligned} \quad (1)
$$

Finally, the surviving muon candidates are paired to form J/ψ candidates according to the following requirements:

- the two muons must feature opposite signs;
- the dimuon invariant mass has to be within 0.15 GeV from the PDG J/ψ mass;
- probability for the two muon tracks to originate from the same decay vertex > 1%.

2.2.2 Track selection

Hadron Track candidates, $K^\pm$ and $\pi^\pm$ in the context of this study, were selected according to the following acceptance criteria:

- transverse momentum $p_\mathrm{T} > 0.5\,\mathrm{GeV}/c$,
- pseudorapidity $|\eta| < 2.4$,

and quality criteria:

- normalized uncertainty on the track p_T, $(\sigma_{p_T}/p_T) < 0.1$;
- sum of the numbers of Pixel and Strip hits $N_\mathrm{hits} > 10$;
- χ^2/NDF probability divided by N_hits, < 0.18.

2.2.3 Chi-squared confidence-level cut

By zooming into the [0, 0.005] region of chi-squared confidence-level variable for MC Signal and Background distributions, a loose cut (Chi2cl>0.003) was applied, which removed plenty of Background Noise that accumulated near Chi2cl=0.

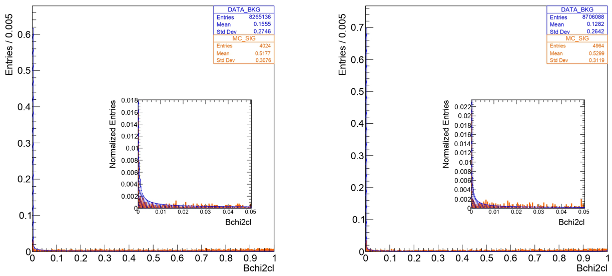

Figure 3: Chi Squared Confidence Level Cut Study

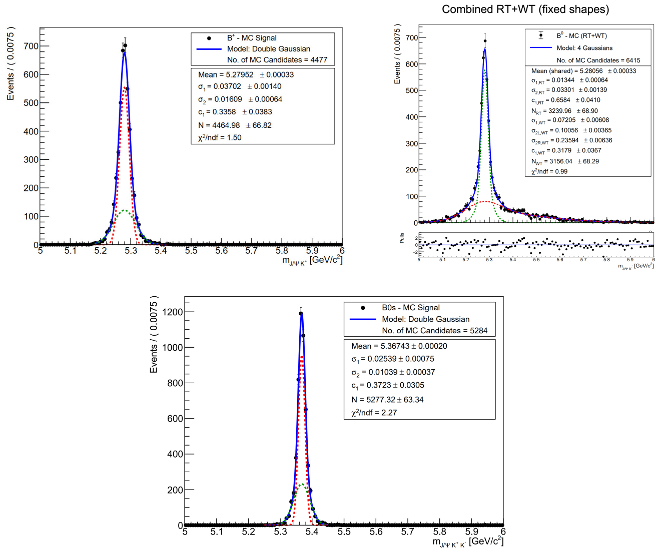

Figure 4: MC Signal Fit for B^+, B_0 and B_s^0

2.2.4 $K^*(892)^0$ & ϕ mass windows

To ensure both hadron tracks come from the same parent particle ($K^*(892)^0$ for B^0, and $\phi(1020)$ for B_s^0), a $K^*(892)^0$ mass cut was applied to B^0 data and a $\phi(1020)$ mass cut was applied to B_s^0 Data:

- $K^*(892)^0$ window: Abs(Btktkmass - 0.89594) < 0.25 (GeV) ;
- $\phi(1020)$ window: Abs(Btktkmass - 1.019455) < 0.015 (GeV) .

3 Selection

Due to the extensive amount of Combinatorial Background that remained after the Preselection cuts, more cuts are necessary to get a clear Signal peak.

3.1 Signal & sideband regions

To study on which variables to cut and where to cut, we first need to remove Background from the data. By fitting the MC Signal, the Signal Region can be estimated; after that, any events outside of the Signal Region are considered as Background.

Using a $4\sigma_{\text{effective}}$ window, the Signal Region is defined.

For B^+, we consider only the right sideband in the selection process, as Partially Reconstructed background is present on the left sideband.

For B^0, due to a tagging problem, the signal region was chosen by eye, [5.1, 5.6] GeV.

For B_s^0, a $3\sigma_{\text{effective}}$ window was used for the signal region.

The fits to the MC invariant mass spectra are shown in Fig.4.

3.2 Variable distribution

With the Signal and Background regions defined, MC Signal and Combinatorial Background distributions were produced, and compared as shown in Fig. 5.

The MC distributions are expected to accurately describe the Data Signal. Later a MC Validation study is performed, to verify how well simulation describes data (Fig. 10).

3.3 Figure of merit

In order to minimize the Combinatorial Background in the Signal region, the statistical significance (Signal uncertainty from Poisson Distribution) is maximized. It acts as the figure of merit (FOM) employed in the optimization procedure, being re-calculated after applying a cut.

The computation of the FOM is performed as

$$FOM = \frac{S_{MC} \cdot f_s}{\sqrt{S_{MC} \cdot f_s + B_{SB} \cdot f_b}} = \frac{S_{Data}}{\sqrt{S_{Data} + B_{SR}}}$$

where S and B are the numbers of Signal and Background events. Due to the arbitrariness of the size of the MC samples, two scaling factors (f_s and f_B) are required to scale the number of MC Signal events to that expected in data, and to scale the number of background events between the Sideband and the Signal regions:

$$f_s = \frac{S_{Data}}{S_{MC}} \quad \text{and} \quad f_b = \frac{B_{SR}}{B_{SB}} .$$

3.4 Pre-Cut

To obtain the scaling factors f_s and f_b, the invariant mass distribution in data was fitted with a model. However, the abundance of Combinatorial Background posed a difficult challenge for fitting. To overcome this, the MC Signal/Background comparisons were analyzed, and a cut was

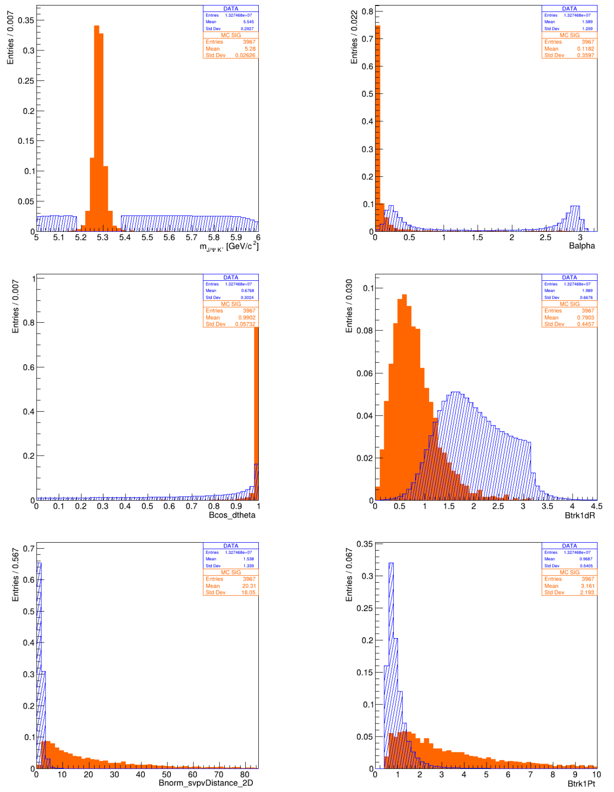

Figure 5: B^+ Normalized MC Signal and Data Combinatorial Background Distributions (Right sideband)

chosen by eye, from which a cleaner Signal peak was obtained in data, as shown in Fig. 6.

Following the data fit, and the extraction of the scaling factors, the above cut is not used further.

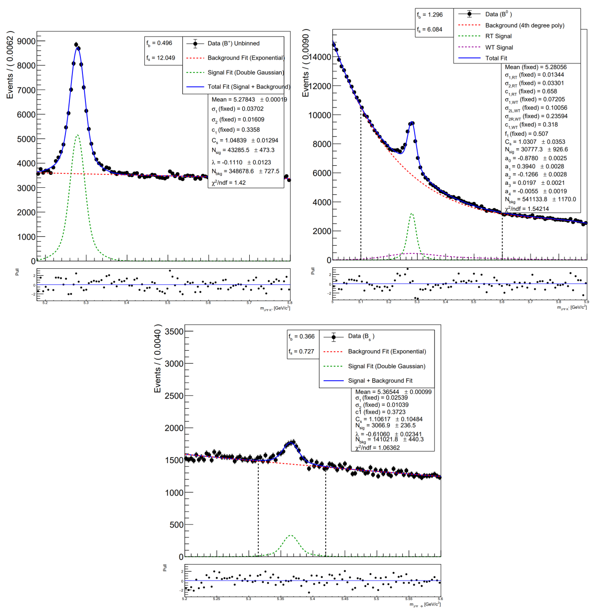

Figure 6: Pre-cut Data Fit for B^+, B_0 and B_s^0

3.5 Optimization

With the scaling factors obtained from Pre-Cut Fit, the Significance for each binned cut was computed, by simply counting MC Signal and Sideband Background events before and after cuts. Significance Curve graphs were drawn. The value of the cut that maximizes the Significance is considered the optimal cut.

Depending on the case, one or a maximum of two optimized cuts were applied to the data. The results are shown in Fig. 7.

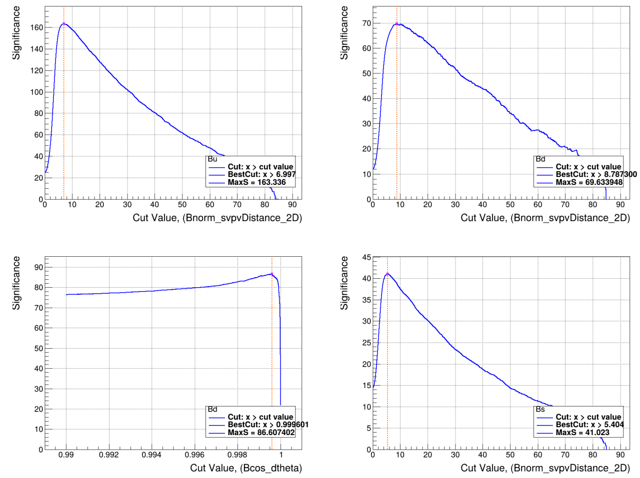

Figure 7: Significance Curves for B^+, B_0 and B_s^0

4 Signal Yield

4.1 Extended Unbinned Maximum Likelihood Method

We are interested in measuring the signal yields. With this objective, the B mesons' invariant mass spectrum is fitted using the Extended Unbinned Maximum Likelihood (EUML) method [2]. The EUML fit is implemented using the RooFit package [3].

The parameters that best describe the data, $\vec{\lambda}$, can be obtained by maximizing the likelihood function, $\mathcal{L}(\vec{\lambda})$:

$$\mathcal{L}(\vec{\lambda}) = \frac{N^{N_{obs}} e^{-N}}{N_{obs}!} \prod_{i=1}^{N_{obs}} l(m_i, \vec{\lambda}) \,, \qquad (2)$$

where N_{obs} is the total number of observed events, N is the estimated total number, m_i is the invariant mass of the i-th event, and l is the model, i.e. a probability distribution function (PDF) describing the shape of the data.

4.2 Nominal fit model for B^+ meson

In the nominal fit to the B^+ meson, the signal component is modeled by a sum of two Gaussian functions, with the same mean, and widths given by the Monte Carlo simulation, σ_1^{MC} and σ_2^{MC}, scaled by a common resolution factor

C_s. The relative fraction of the first Gaussian is fixed to c_1^{MC}. The relative fraction and the widths are fixed to preserve the signal shape from the Monte Carlo simulation.

$$\begin{aligned}\mathcal{P}_S(m \mid \mu, C_s) = c_1^{MC} \cdot Gauss(m; \mu, C_s \cdot \sigma_1^{MC}) \\ + (1 - c_1^{MC}) \cdot Gauss(m; \mu, C_s \cdot \sigma_2^{MC})\end{aligned} \tag{3}$$

where the Gaussian function is defined as:

$$Gauss(m; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(m-\mu)^2}{2\sigma^2}\right]. \tag{4}$$

The Combinatorial Background is described by an exponential function:

$$\mathcal{P}_{\exp}(m \mid \lambda) = e^{\lambda m} \tag{5}$$

For the B^+ mesons, there is an additional background component, which corresponds to the partially reconstructed B meson decays (e.g. $B_s^0 \to J/\psi K^+ K^-$, where it can be misreconstructed as a B^+ candidate, due to the presence of a positive hadron track combined with two oppositely charged muons). This component is called the Partially Reconstructed Background, and it is described by a complementary error function,

$$\mathcal{P}_{\mathrm{erfc}}(m \mid c_{sf}, c_{sc}) = \mathrm{erfc}\left(\frac{m - c_{sf}}{c_{sc}}\right). \tag{6}$$

The combined extended model for the B^+ meson candidates is given by:

$$\begin{aligned}\mathcal{P}_{B^+}(m) = N_{\mathrm{sig}}\,\mathcal{P}_S(m \mid \mu, C_s) + N_{\mathrm{bkg}}\,\mathcal{P}_{\exp}(m \mid \lambda) \\ + N_{\mathrm{erfc}}\,\mathcal{P}_{\mathrm{erfc}}(m \mid c_{sf}, c_{sc}).\end{aligned} \tag{7}$$

4.3 Nominal fit model for B_s^0 meson

In the nominal fit to the B_s^0 meson, the signal component is similar to the B^+ meson case, following equations (3) and (4).

The Combinatorial Background is described by a linear combination of Chebyshev polynomial functions up to the 4^{th} order:

$$\begin{aligned}\mathcal{P}_{\mathrm{bkg}}(m \mid a_0, \ldots, a_4) = a_0 + a_1 x + a_2(2x^2 - 1) \\ + a_3(4x^3 - 3x) \\ + a_4(8x^4 - 8x^2 + 1)\end{aligned} \tag{8}$$

where x is the rescaled mass variable that maps the fit range $[m_{\min}, m_{\max}]$ onto the Chebyshev domain $[-1, 1]$:

$$x = \frac{2m - (m_{\max} + m_{\min})}{m_{\max} - m_{\min}}. \tag{9}$$

The combined extended model for the B_s^0 meson candidates is given by:

$$\mathcal{P}_{B_s^0}(m) = N_{\mathrm{sig}}\,\mathcal{P}_S(m \mid \mu, C_s) + N_{\mathrm{bkg}}\,\mathcal{P}_{\mathrm{bkg}}(m \mid a_0, \ldots, a_4). \tag{10}$$

4.4 Nominal fit model for B^0 meson

For the nominal fit of the B^0 meson, the signal is split into a *Right–Tag (RT)* and a *Wrong–Tag (WT)* component. Without dedicated particle identification (PID), the detector cannot tell which of the two opposite-charge tracks is the Kaon and which is the Pion, so both mass–assignment hypotheses must be considered.

While candidates produced with the correct mass assignment (RT) result in a peak resonance, those formed with the incorrect assignments (WT) produce a broader, asymmetric mass response. The two shapes were studied in MC simulated samples. The nominal fitting method was performed separately while sharing the same mean.

The RT signal component is also similar to the B^+ case, using Monte Carlo RT simulation and following equations (3) and (4).

The WT signal component is modeled by a sum of a Gaussian function with an Asymmetric Gaussian function, where $\sigma_{2L}^{\mathrm{MC}}$ represent the left side width, and $\sigma_{2R}^{\mathrm{MC}}$ represent the right side width:

$$\begin{aligned}\mathcal{P}_{\mathrm{WT}}(m \mid \mu, C_s) = c_1^{\mathrm{MC}}\, Gauss\,(m; \mu,\, C_s \cdot \sigma_1^{\mathrm{MC}}) \\ + (1 - c_1^{\mathrm{MC}}) \cdot \\ AsymGauss\,(m; \mu,\, C_s \cdot \sigma_{2L}^{\mathrm{MC}},\, C_s \cdot \sigma_{2R}^{\mathrm{MC}})\end{aligned} \tag{11}$$

where *Gauss* is the same as in equation (4), and

$$AsymGauss\,(m; \mu, \sigma_L, \sigma_R) = \frac{\sqrt{2/\pi}}{\sigma_L + \sigma_R} \begin{cases} \exp\left[-\dfrac{(m-\mu)^2}{2 \cdot \sigma_L^2}\right], & m < \mu \\ \exp\left[-\dfrac{(m-\mu)^2}{2 \cdot \sigma_R^2}\right], & m \geq \mu \end{cases} \tag{12}$$

The Combinatorial Background is similar to the B_s^0 case, following equations (8) and (9).

Therefore, the combined extended model for the B^0 meson candidates is given by:

$$\begin{aligned}\mathcal{P}_{B^0}(m) = N_{\mathrm{RT}}\,\mathcal{P}_S(m \mid \mu, C_s) + N_{\mathrm{WT}}\,\mathcal{P}_{\mathrm{WT}}(m \mid \mu, C_s) \\ + N_{\mathrm{bkg}}\,\mathcal{P}_{\mathrm{bkg}}(m \mid a_0, \ldots, a_4).\end{aligned} \tag{13}$$

4.5 Systematic uncertainties

The above nominal models, for each B meson, were chosen because of their good capability to describe the data and due to their simplicity (in accordance with the Occam's razor principle).

However, there is a large number of alternative fitting models that could be used to describe the data. As such, a few possible fitting models were selected, which are referred to as systematic variations. Their purpose was to estimate the systematic uncertainty by comparing the signal yields obtained with the nominal model and these alternative models, according to the following equation:

$$\delta_{sys.} = \frac{|N_S\,(\mathrm{nominal}) - N_S\,(\mathrm{variation})|}{N_S\,(\mathrm{nominal})} \cdot 100\%. \tag{14}$$

For the B^+ meson, three different signal models and three background models were considered. The signal

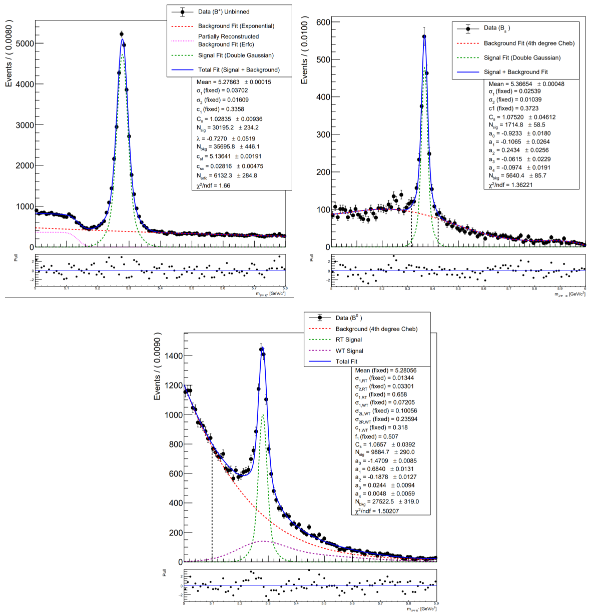

Figure 8: Data Nominal Fit after Optimized Cuts for B^+, B_s^0 and B_0

model variations were the Crystal Ball + Gaussian, the Triple Gaussian and the Fixed Mean[1]. The background model variations were the first-order polynomial (linear function), the second-order polynomial, and a different mass range[2].

Using equation (14), the systematic uncertainties for each model were calculated, and shown in Fig. 9.

The total systematic uncertainty is obtained using equation (15)

$$\delta_{Total\ sys.} = \sqrt{max(\delta_{sig.})^2 + max(\delta_{bkg.})^2} \qquad (15)$$

while the total statistical uncertainty is calculated using the equation (16)

$$\delta_{\text{Stat.}} = \frac{\sigma_S \text{ (nominal)}}{N_S \text{ (nominal)}} \cdot 100\% \,. \qquad (16)$$

5 MC validation

In general, simulation is not guaranteed to describe data with all accuracy. We would like to verify the level of agreement between simulation and data, for the various variables employed in selection.

However, even in the signal region, a huge amount of background is present. For addressing this, two methods were used to extract signal distributions from data:

- sPlot, and

[1]In the data fit, the mean parameter is fixed to the value obtained from the MC fit, making the model equivalent to the nominal one apart from this constraint.

[2]In the data fit, the mass range used did not include the left sideband region, effectively removing the Partially Reconstructed Background, and with it the need for the complementary error function. Apart from this constraint, the model is similar to the nominal one.

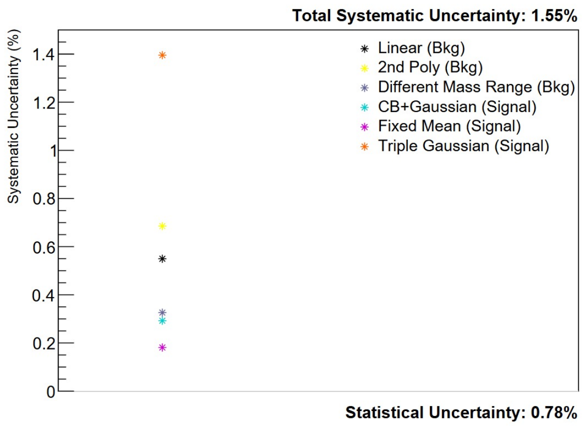

Figure 9: B^+ Systematics and Statistical Uncertainties

- Sideband Subtraction.

The use of two methods further allow for cross validation.

5.1 sPlot

The sPlot method uses the information (Covariance Matrix, Fitted normalized PDFs and yields) obtained from the mass spectrum Extended Unbinned Maximum Likelihood Fit to calculate Data Signal sWeight. Then the data signal distribution in other variables can be plotted (e.g. transverse momentum), by associating its signal sWeight for each event instead of 1. The sWeight for event i is calculated as:

$$\omega_i^{(s)} = \sum_j \frac{V_{sj} \cdot \mathcal{P}_j(m_i)}{\sum_k N_k \mathcal{P}_k(m_i)} \,,$$

where V_{sj} is the element of Covariance Matrix, $\mathcal{P}(m_i)$ the normalized PDF component of Signal for an event of mass i, N_k the Combinatorial Background and Signal Yields.

5.2 Sideband subtraction

In order to remove the background present in the signal region, in alternative, the invariant mass distribution was fitted through a model, and the Background yields in both the signal and sideband regions were extracted:

$$N_{signal} = N_{signal\ region} - \alpha \cdot N_{sideband\ region}$$

where α is a scaling factor

$$\alpha = \frac{\text{background yield in signal region}}{\text{background yield in sideband region}} \,.$$

To determine α, the Background yields obtained previously were used in formula above.

5.3 Comparison

The application of the two methods gave consistent results, cross-validating the implementations.

The methods were applied to several variables of interest. In general a good level of agreement was found between data and simulation, as shown in Fig. 10.

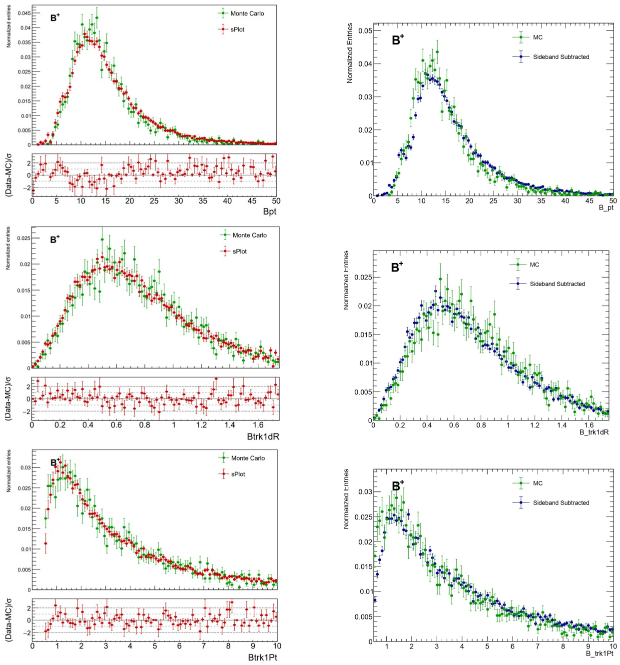

Figure 10: Signal distributions extracted from data with sPlot and Sideband-subtraction methods, and comparison to simulation, shown for the specified variables.

6 Conclusions

We have analyzed recent LHC Run 3 proton-proton collision datasets collected by the CMS experiment during 2024. We have extracted first signals for B^{+}, B^{0}, and B_s mesons, through the implementation of a selection optimization procedure, based on a set of discriminating variables. Data-driven techniques were employed for validating Monte Carlo simulation. These studies form the basis for ensuing measurements of b-quark production and hadronization, in pp and PbPb collisions, and the exploration of QGP properties using B mesons as precise probes with LHC Run 3.

Acknowledgements

Special thanks to Henrique Legoinha and Professor Nuno Leonardo for making this summer internship possible and for their dedication to the project. This internship offered valuable insights and opportunities not typically available to undergraduate students, while also improving our programming skills in C and Python and our understanding of the mathematical foundations of fitting methods.

References

[1] CMS Collaboration, JINST **3**, S08004 (2008)

[2] J. Ocariz, *Probability and statistics for particle physicists*, https://cds.cern.ch/record/1701936/files/p253.pdf (2013)

[3] R. Brun, F. Rademakers, Nucl. Instrum. Meth. A **389**, 81 (1997), proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996. See also http://root.cern.ch/

[4] H. Legoinha, *Probing the Quark Gluon Plasma with B_s^0 and B+ Mesons* (2023), CERN-THESIS-2023-064, `https://cds.cern.ch/record/2860807`

[5] M. Pivk, F. Diberder, *sPlot: a statistical tool to unfold data distributions*, https://arxiv.org/pdf/physics/0402083 (2005)