Selection of $B_s^0 \to \psi(2S)K_S^0$ decays via multivariate analysis

Lukas Bertsch
lukas.bertsch@tu-dortmund.de

Tabea Hacheney
tabea.hacheney@tu-dortmund.de

Tom Troska
tom.troska@tu-dortmund.de

Start of course: 13th of June 2024

TU Dortmund University – Faculty of Physics

Contents

1. Abstract

This analysis is aimed at finding events coming from the decay $B_s^0 \to \psi(2S)K_S^0$ using a large data sample recorded by the LHCb experiment in Run 2. To identify and extract rare decay channels, an extensive analysis including machine learning methods is needed. The extracted number of signal events is $n_{\text{sig}} = 45$ and the number of background events is $n_{\text{bkg}} = 32$, with a significance of $m = 5.1$ in the signal region.

2. Introduction

The Standard Model (SM) of Particle Physics describes all fundamental interactions except for the gravitational force. However, there are many phenomena such as the gravitational force, the enormous matter-antimatter asymmetry observed in the universe and the presence of dark matter that cannot be properly explained. Hence, it is important to test the SM by means of precision measurements. In contrast to searches of New Physics (NP) through the direct production in high-energy collisions, they can also be produced virtually in low-energy processes such as in hadronic decays involving beauty or charm quarks. Due to Heisenberg's uncertainty principle, they can exist for a short period of time and contribute as quantum corrections, representing a complementary way of searching for NP indirectly. The Large Hadron Collider beauty (LHCb) experiment investigates the differences between matter and antimatter by studying mesons containing charm and beauty quarks.

This analysis is aimed at using machine learning methods to extract rare signal events from a large data sample including $B_s^0 \to \psi(2S)K_S^0$ candidates recorded during Run 2 of the LHCb experiment. The decay channel $B_s^0 \to \psi(2S)K_S^0$ is a very rare decay, resulting in the data being dominated by combinatorial background and more common decays, which have been recorded due to trigger criteria being wrongly fulfilled. To identify and extract these rare signal events, an extensive analysis including machine learning methods is needed.

The used data has been recorded uring Run 2 (2015-2018) of the LHCb experiment at CERN, proton bunches with a centre of mass energy of 13 TeV collided approximately 40×10^6 times per second [1]. The theoretical framework will be presented in section 3, followed by a description of the LHCb experiment in section 4. The analysis strategy is presented in section 5, followed by the study of the dataset and the training of the machine learning model in section 6. The final results are summarized in section 7.

3. Theory

Before diving into the analysis, one has to understand the underlying kinematics of the decay. The B_s^0 meson consists of a s and a $\bar{b}$ quark (next to other sea quarks and gluons) and hence has a neutral net electric charge. As for this decay channel, the b quark of the B_s^0 meson interacts with a W^+ gauge boson and results in a charm pair $c\bar{c}$ and a $\bar{d}$ quark in the final state. Together with the strange quark from the B_s^0 meson, the final state consists of a $\psi(2S)\,[c\bar{c}]$ meson and a K_S^0 ("k-short") being a superposition of $d\bar{s}$ and $\bar{d}s$. This decay can only occur via the weak force, since the quark flavour changes from $\bar{b}$ to $\bar{d}$. The Feynman diagram depicting this process is shown in Figure 1.

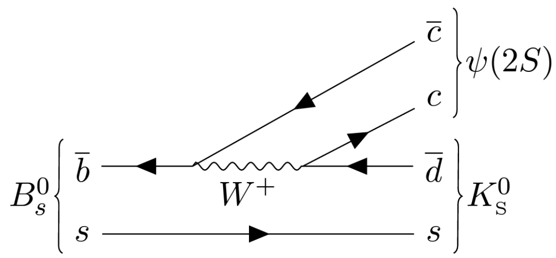


Figure 1: Feynman diagram of $B_s^0 \to \psi(2S)K_S^0$ [1].

The $\psi(2S)$ and K_S^0 are not directly detected in the detector. While $\psi(2S)$ mainly decays via strong interactions, it also decays via electromagnetic force into a lepton pair (l^+l^-). Since the detector has a good efficiency for detecting muons, only muon pairs will be regarded ($\mu^+\mu^-$). Due to the conservation of energy and momentum, the total mass of the muon pair is approximately equal to the mass of the $\psi(2S)$ meson. Due to energy conservation, the total mass of the muon is exactly equal due to the mass of the $\psi(2S)$ meson, but due to the limited detector resolution, the masses might differ to some extend. With cuts on the muon pair masses, the collected data is reduced to less signal candidates while rejecting charm resonances $c\bar{c}$.
The dominant decay channel of K_S^0 is a pair of pions ($\pi^+\pi^-$) via weak interaction. The K_L^0 meson also consists of the superposition of $d\bar{s}$ and $\bar{d}s$ and has the same decay channels. However, the branching fraction of that decay channel is two magnitudes smaller than for the K_S^0 and will therefore not interfere with the relevant signal candidates to a significant level. The mass and other kinematic variables of the K_S^0 can be reconstructed with the produced pion pairs.
The given recorded data consists of various variables including kinematic properties of the muons, pions as well as the reconstructed $\psi(2S)$ and K_S^0 mesons. The B_s^0 can be reconstructed using the properties of the $\psi(2S)$ and K_S^0 while using their true invariant masses, being $m_{\psi(2S)} = (3686.097 \pm 0.011)\,\text{MeV}$ and $m_{K_S^0} = (497.611 \pm 0.013)\,\text{MeV}$ [2].
The recorded data set does not only include $B_s^0 \to \psi(2S)K_S^0$ signal. Moreover, it primarily contains combinatorial background, consisting of random tracks mimicking the signal. The much more abundant decay $B^0 \to \psi(2S)K_S^0$ is also included in the recorded data.

To differentiate between background and signal, an extensive analysis including training a classifier is needed.
To evaluate and optimize the classifier (maximizing the number of signal candidates while keeping the background low), a loss function is to be minimized. The inverse of such loss functions is called figure of merit (FOM). In this analysis, the Punzi figure of merit for optimizing the signal efficiency

$$\mathrm{FOM} = \frac{\epsilon_{\mathrm{sig}}}{5/2 + \sqrt{N_{\mathrm{bkg}}}} \tag{1}$$

is to be maximized with N_{bkg} being the background candidates in the signal region (combinatorial background + $B^0 \to \psi(2S)K^0_S$ events) and ϵ_{sig} being the efficiency of classifying the signal. The signal efficiency is calculated by using a Monte Carlo simulation of $B^0_s \to \psi(2S)K^0_S$. The BDT cannot discriminate between B^0_s and B^0_d events, due to the similarity of kinematic properties. However, the resolution of the LHCb detector is well enough, to distinguish between the two decays in the mass spectrum after removing most background candidates.
To train an efficient and well performing classifier, the right choice of variables is needed. One has to look for those variables, which differ the most for background and signal. To find these, one can caluclate the largest distance between the cumulative probability distributions F^i of these variables:

$$\sup_n |F^1_n - F^2_n|, \tag{2}$$

with the index n running over all bins of the distributions.
The final result can be evaluated using the significance, being the number of signal candidates devided by the total number of events in the signal region, including background and signal. It can be calculated by

$$m = \frac{N_{\mathrm{sig}}}{\sqrt{N_{\mathrm{sig}} + N_{\mathrm{bkg}}}}, \tag{3}$$

using the given number of signal and background candidates by the classifier. The significance is a quantity to measure of how likely the result is produced by statistical fluctuations.

4. The LHCb detector

Located at CERN, the LHC is the largest particle collider in the world and houses a number of experiments. Opposing protons beams are collided at four different interaction points. At one of these, the LHCb experiment is installed. This detector is constructed as a single-arm forward spectrometer which allows for an optimized detection in the pseudorapidity range of $2 \leq \eta \leq 5$. With the help of this design, the LHCb detector is well suited for the detection of decays that include b and c quarks because the production

of these is favored for small angles along the beam axis.
The LHCb experiment consists of several sub-detectors, each with a different special purpose. A cross-section of the detector used during the second data-taking run is depicted in Figure 2. In the immediate proximity of the interaction point, the Vertex

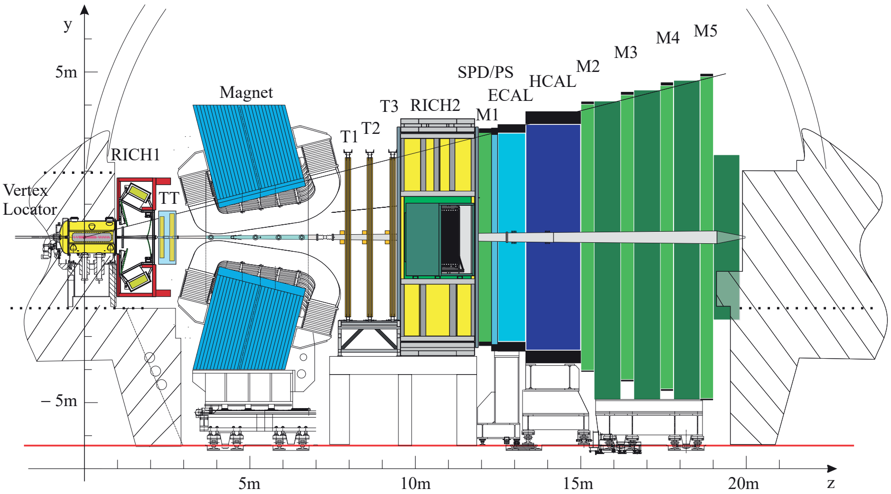


Figure 2: Cross-section of the LHCb detector. A right-handed coordinate system with the z-axis parallel to the beam pipe and the y-axis orientated to the top is used to describe the detector.The different sub-detectors and their purpose and functionality are discussed in this section [3].

Locator (VELO) is located. Primary and secondary vertices are identified by the 52 silicon pixel modules. The detector is divided into two halves that are retractable. During the injection phase, the VELO is sitting in its retracted position because the beams are not as focused as during the stable beam phase. It is imperative to prevent the beams from colliding with the detector material due to their high energy, which could cause significant damage. Due to the closeness of the VELO to the interaction point, the primary veritices can be reconstructed with great accuracy.
One of the two Ring Imaging Cherenkov detectors (RICH) is located right before the 1.4 T dipole magnet while the other RICH detector is situated behind the tracking stations. Two materials with different refractive indices are used in the two RICH detectors to determine the velocity of the particles via the Cherenkov effect. The implementation of two detectors allows for a larger momentum range reconstruction.
The position of the particles are detected by the tracking stations, the Tracker Turicensis (TT) and T1-T2. With this information and the previously measured velocity, the momentum of the particles can be calculated.
For the determination of the energy of the particles, the electromagnetic and hadronic calorimeters (ECAL and HCAL) are employed. To accomplish this task for the ECAL, the shashlik calorimeter technology is used. Here, an absorber material and a detector layer of a scintillating material are stacked alternately. The HCAL is constructed similarly but with the scintillating tiles running parallel to the beam axis.

Muon stations at the end of the detector identify the muons that are passing through the detector. This is accomplished by four multi-wire proportional chambers.
Because of the large number of potential events, a trigger system is used to identify interesting decays. The data of these decays are then saved for further offline analysis.

5. Analysis strategy

The data used in this analysis consists of three different data samples. One is the actual measured data, which contains reconstructed $B^0_{(s)} \to \psi(2S)K^0_S$ candidates and is dominated by combinatorial background. The two other datasets are Monte Carlo simulation samples of the signal decay $B^0_s \to \psi(2S)K^0_S$ and the control channel $B^0_s \to \psi(2S)K^0_S$.
First, the training samples for the background and the signal need to be defined. In the case for the data of $B^0_s \to \psi(2S)K^0_S$ decays, the background is predominatly combinatorial. The recorded LHCb data are used for the training of this classifier. It is to note that only the region where the background is present is used for the search.
The training samples for the signal classification cannot be sourced from the recorded data but must be simulated with Monte Carlo simulations. Imperfect simulation data that do not match the true data of the decay can cause problems. This is due to not perfect theoretical models and computing constraints. The challenges of this problem can at least be partially overcome by introducing computing weights. The here provided simulation data contain a variable called `kinematic_weights` that helps to correct for the mismatches between simulation and recorded data.
Feature selection is another important aspect of this analysis. The classifier uses a number of features to decide whether the data are background or signal. Features, that can be used here, are extracted from the control channel $B^0 \to \psi(2S)K^0_S$. This decay is similar to the decay $B^0_s \to \psi(2S)K^0_S$ and therefore its features can be fed to the classifier. In principle, a classifier can use any number of features, but the runtime increases with the number of features. Having too many features can cause overfitting, where the classifier learns the training data and noise too well, resulting in poorer performance on new, unseen data. Some of the here provided features are also redundant and do not include additional information.
For the purpose of ascertaining the best classification threshold, the FOM (Equation 1) is determined by calculating the signal efficiency and number of background events in the signal region.
In the final step, the trained and optimized classifier is used to classify the recorded LHCb data. This leads to the removal of the combinatorial background so that the B^0 and B^0_s peaks are clearly recognizable. The number of signal events can then be extracted by fitting a background and signal model to the data.

6. Analysis

At first, the invariant mass distributions of the reconstructed $B_s^0 \to \psi(2S)K_S^0$are inspected for the simulation and real data. In Figure 3, the invariant mass distribution of the reconstructed B_s^0 events in the signal simulation is shown. As expected, a clear peak at the B_s^0 mass can be seen. The mass distribution of the reconstructed B^0 candidates in

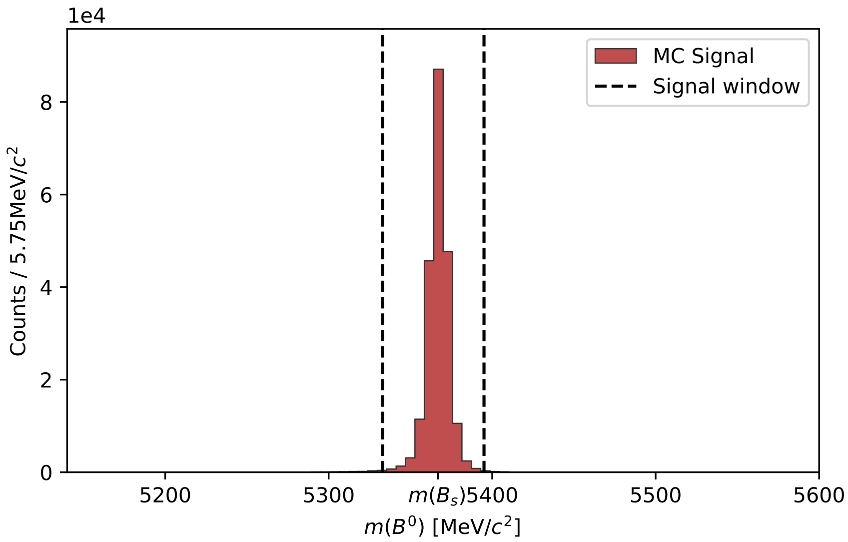


Figure 3: Invariant mass distribution of the B_s^0 candidates for the signal channel simulation data.

real data can be seen in Figure 4. Here, a peak at the nominal B^0 mass can be seen. However, due to dominating combinatorial background, no peak at the B_s^0 mass is visible. The dataset also contains *sWeights* which can be used to extract the contribution of the control channel in the dataset. By weighting the mass histogram with the sWeights, only the control channel contribution remains in the plot, as can be seen in Figure 5.

6.1. Definition of a signal window

The mass distribution of the signal decay peaks only in a short window of the whole mass range given in the dataset. In order to know, where signal is expected, a signal window has to be defined. This is done by calculating every interval containing 99 % of data in the signal simulation and choosing the shortest interval. Here, this interval follows as 5333.4 to 5394.6 MeV/c^2, defining the signal window of 5333 to 5395 MeV/c^2 which can also be seen in the aforementioned plots (3, 4). Subsequently, the upper sideband ('*USB*'), containing mostly combinatorial background, is defined as the area with reconstructed mass > 5400 MeV/c^2.

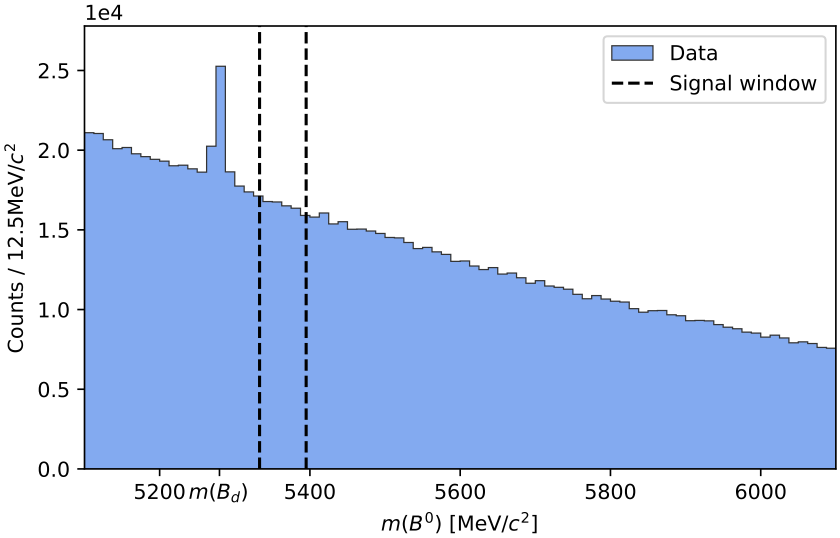


Figure 4: Invariant mass distribution of the B^0 candidates for the recorded LHCb data.

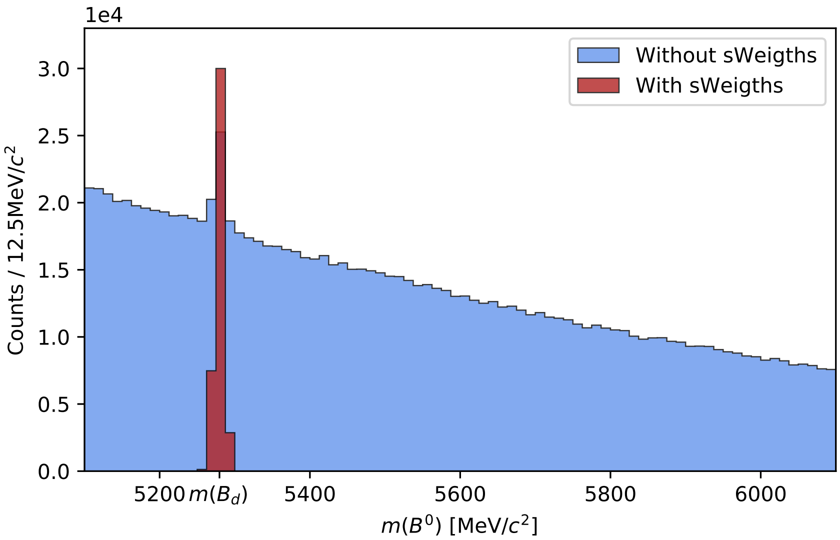


Figure 5: Invariant mass distribution of the B^0 candidates for the recorded LHCb data with and without the sWeights.

6.2. Feature selection

In order to train a multivariate classifier capable of separating signal from background, meaningful features from all available variables in the dataset have to be selected. In total, 863 variables are listed in the dataset. After removing event, utility, trigger and spatial coordinate variables, 398 variables remain. For these variables, the correlation to the invariant mass is calculated and variables having a correlation coefficient of 0.3 or higher are excluded. Variables with too high correlation to the invariant mass would introduce a bias in training the classifier and could not be used in similarity checks between simulation and data, because the sWeights are based on the B^0 candidate mass. The remaining 390 variables are checked to be correctly modelled by simulation and have significantly different distributions for signal and background. This is done using the Kolmogorov Smirnov test statistic defined in Equation 2 for weighted distributions as a measure of similarity. To check agreement between simulation and data, the distributions of sWeighted data are compared to the control channel distributions and only variables with a test statistic of $d < 0.05$ are kept. 190 variables with a higher test statistic are removed. The variables are also required to have a test statistic of $d > 0.2$ when comparing signal (simulation) and background (USB), leaving 94 variables. After further removing variables that are highly correlated (correlation > 0.9) to others or dublicates of another variable, 18 variables are left that are used for the training of the multivariate classifier. The distributions of four of these variables for signal and background, as well as

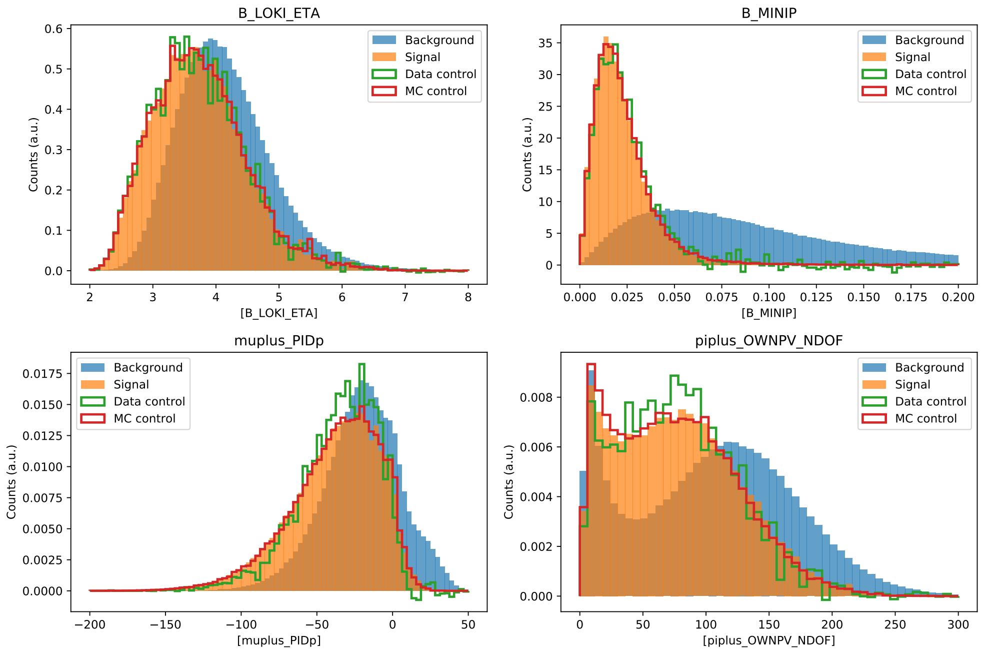


Figure 6: Distributions of four selected variables used in the MVA for simulation, reweighted data and background.

control channel data and simulation are shown in Figure 6. Distributions for all variables and the correlation matrix can be found in the appendix A.1.

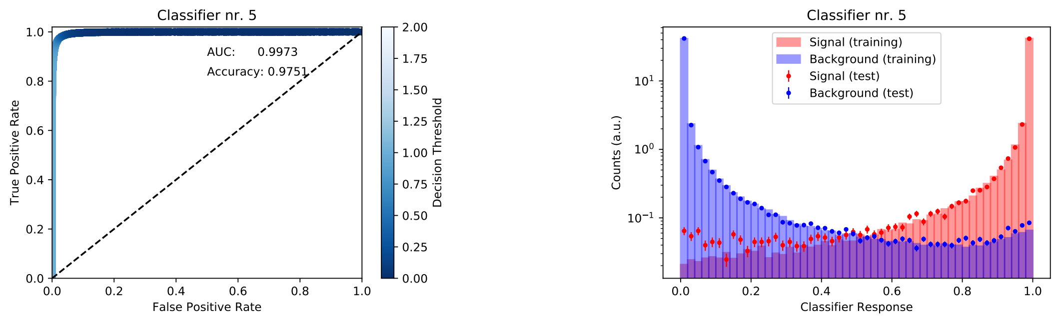
6.3. Training of a multivariate classifier

With the now selected variables, a multivariate classifier can be trained. For this purpose, a boosted decision tree as implemented in the package `XGBoost` [4] is used. The training data consists of 637 410 background events and kinematically reweighted signal simulation, corresponding to 155 805 events. Because of this inbalance, each background event is weighted with a factor of 155805/637410. The hyperparameters of the classifier are optimized via a random search followed by a grid search using cross validation and a subset of 40 000 training samples. The resulting parameter values can be read from Table 1. For the classification, five individual BDT's are trained using 5-fold cross

Table 1: Hyperparameter values of the trained classifiers determined by grid search.

Parameter	Value
`n_estimators`	1000
`learning_rate`	0.1
`max_depth`	4
`reg_lambda`	1
`n_iter_no_change`	5

validation and the hyperparameters from Table 1. To evaluate the classifiers performance, the ROC curve is viewed and the area under the curve, as well as the accuracy are calculated. To check for overtraining, the response of the classifier is compared for training and simulation data in a logarithmic plot. The ROC curve and the train-test comparison of the fifth trained classifier can be seen in Figure 7. Additionally, the feature



(a) ROC curve.

(b) Performance on test vs. training dataset.

Figure 7: ROC curve (left) and response on training and test dataset (right) for one of the trained classifiers.

importance of the BDT variables is checked for imbalances. As can be seen in Figure 8, all variables are of similar importance.

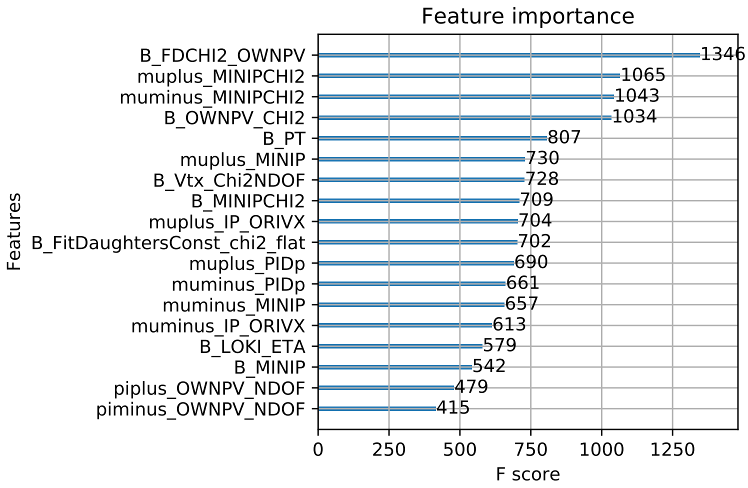


Figure 8: Feature importance of the BDT training variables.

6.4. Optimization of the classification threshold

After applying all BDT's to the data, a cut on the classifiers response has to be made. Therefore the mean classifier response between all 5 BDT's is calculated. The classification threshold separating signal and background is optimized using the Punzi figure of merit Equation 1. The signal efficiency ε is therefore calculated for each threshold as the selection efficiency on the signal simulation. The background yield in the signal region B is estimated via a fit of an exponential function to the upper sideband of the data after applying the selection threshold, and extrapolated to the signal window. For the fitting, the python library `iminuit` [5] is used. The resulting values of the figure of merit are plotted against different thresholds in Figure 9. The optimal classification threshold is

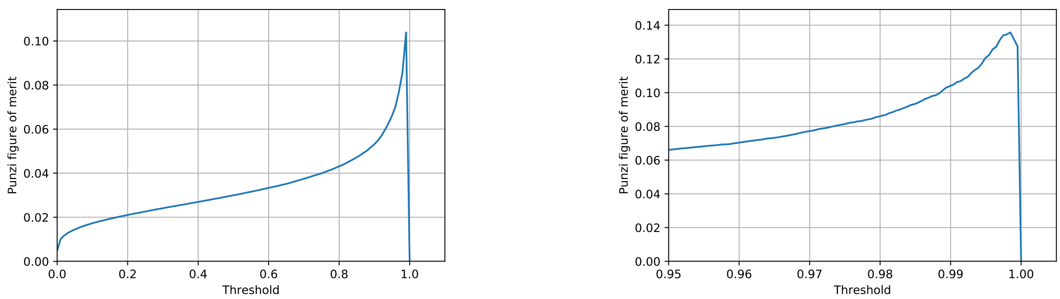


Figure 9: The Punzi figure of merit for the mean classifier response in different intervals of the threshold.

the maximum value of the Punzi figure of merit and reads as $t = 0.998$.

6.5. Evaluation of the signal yield

The invariant B^0 mass distribution of the data after applying the cut on the mean BDT response can be seen in a semi-logarithmic plot in Figure 10. In the mass region below

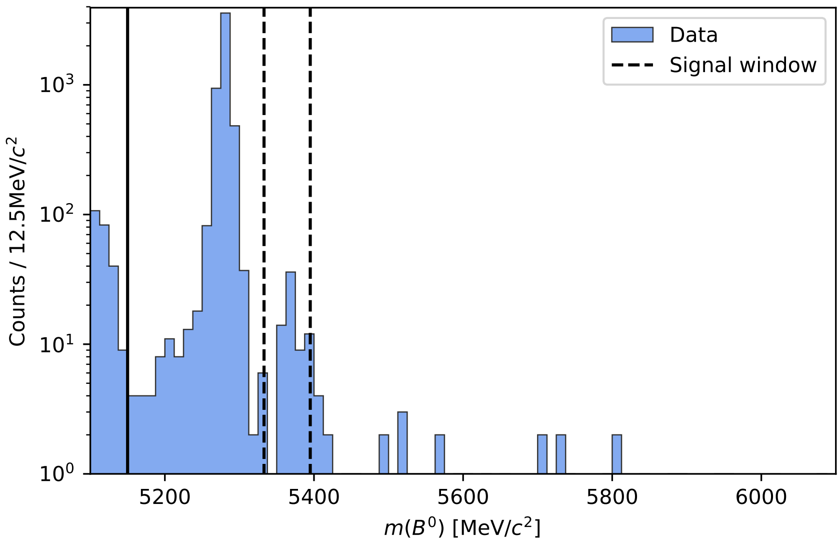


Figure 10: Semi logarithmic invariant mass distribution of the B^0 candidates in data, after the cut on the classifier response is applied.

5150 MeV, a prominent background band of partially reconstructed decays can be seen. In order to determine the signal yield, a fit to the mass spectrum of the B^0 candidates is applied. The fit range is defined as $5200\,\mathrm{MeV} < m(B^0) < 6100\,\mathrm{MeV}$ to not include the partially reconstructed background. The signal peaks are modeled with two gaussian distributions with a fixed width ratio α which is determined from simulation. Using this definition, the width of the B^0 peak is given by σ_0 and the width of the B_s^0 peak by $\sigma_0 \cdot \alpha$. The values of the width σ_0 and the mean values μ_{B_d}, μ_{B_s} of the gaussian distributions are allowed to vary freely to account for mass resolution differences in data and simulation. The background is again modeled by an exponential function with decay constant τ. An extended, unbinned negative Log-Likelihood fit is performed, including the fractions s_{B_d}, s_{B_s} and b of signal, control channel and background components. The resulting graph is shown in Figure 11. The fit parameters follow as

$$
\begin{aligned}
s_{B_s} &= 45 \pm 8 & s_{B_d} &= 5010 \pm 70 \\
\mu_{B_s} &= (5367.3 \pm 1.2)\,\frac{\mathrm{MeV}}{\mathrm{c}^2} & \mu_{B_d} &= (5279.9 \pm 0.9)\,\frac{\mathrm{MeV}}{\mathrm{c}^2} \\
\sigma_0 &= (6.21 \pm 0.08)\,\frac{\mathrm{MeV}}{\mathrm{c}^2} & \alpha &= 1.056 \pm 0.011 \\
b &= 233 \pm 19 & \tau &= 140 \pm 10.
\end{aligned}
$$

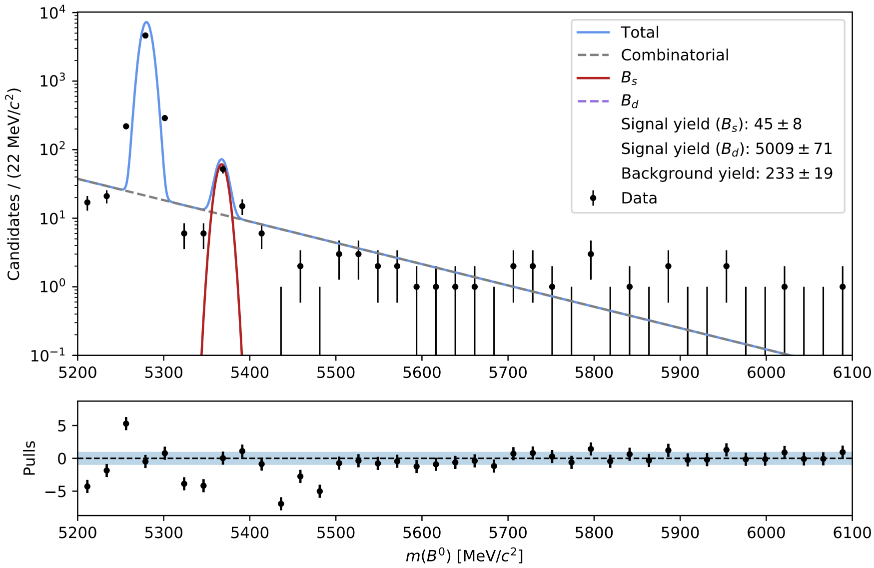


Figure 11: Fit to the invariant mass spectrum of the data in semi logarithmic depiction.

From these, the significance as defined in Equation 3 of the observation of the signal can be calculated. Therefore, the signal and background events in the signal window are interpolated from the fit results as $n_{\text{sig}} = 45$ and $n_{\text{bkg}} = 32$. The significance proxy then reads $m = 5.1$.

7. Discussion

The analysis of the data used for the study of the decay $B_s^0 \to \psi(2S)K_S^0$ yields a number of signal events of $n_{\text{sig}} = 45$ and a number of background events of $n_{\text{bkg}} = 32$ with a significance of $m = 5.1$. Several aspects need to be considered for the validity of these numbers.

First, the BDT seems to be slightly overfitted as can be seen in Figure 7b. Here, marginally more signal events are classified in the test data as in the training data. This could be overcome with further training of the BDT by implementing and optimizing different hyperparameters. Additionally, the signal of the B_s^0 decay is substantially smaller than the signal of the B_d^0 decay.

A further aspect, that could be improved, is the fit model of the signal peaks. In general, these peaks are not symmetric but in this analysis a double gaussian fit is performed. A more complex model has the potential to optimize the fit.

For the significance of about five sigma, it is to note that uncertainties were not considered for the calculation of this value so that the results overestimates the significance.

References

[1] *Selection of B0s → (2S)K0 S events*. TU Dortmund, Faculty of Physics, AG Albrecht. 2020. URL: `https://moodle.tu-dortmund.de/mod/resource/view.php?id=1598013`.

[2] *The Review of Particle Physics (2024)*. Particle Data Group. URL: `https://pdglive.lbl.gov/Viewer.action` (visited on 12/06/2024).

[3] A. Augusto Alves Jr. et al. 'The LHCb Detector at the LHC'. In: *JINST* 3 (2008), S08005. DOI: `10.1088/1748-0221/3/08/S08005`.

[4] T. Chen and C. Guestrin. 'XGBoost: A Scalable Tree Boosting System'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. DOI: `10.1145/2939672.2939785`.

[5] H. Dembinski and P. Ongmongkolkul et al. 'scikit-hep/iminuit'. In: (Dec. 2020). DOI: `10.5281/zenodo.3949207`.

A. Appendix

A.1. Correlations and distributions of the variables used for the MVA

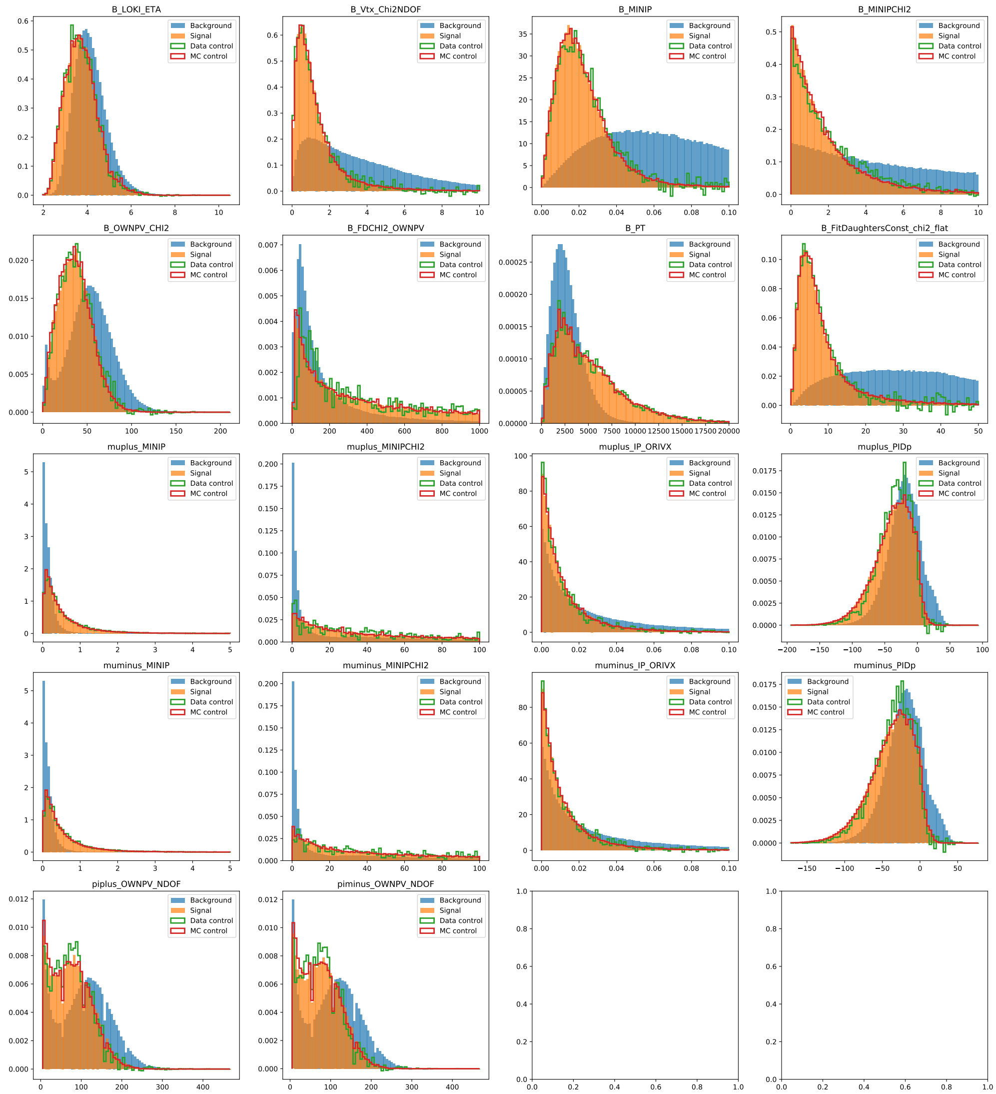


Figure 12: Distributions of the variables used in the MVA for simulation, reweighted data and background.

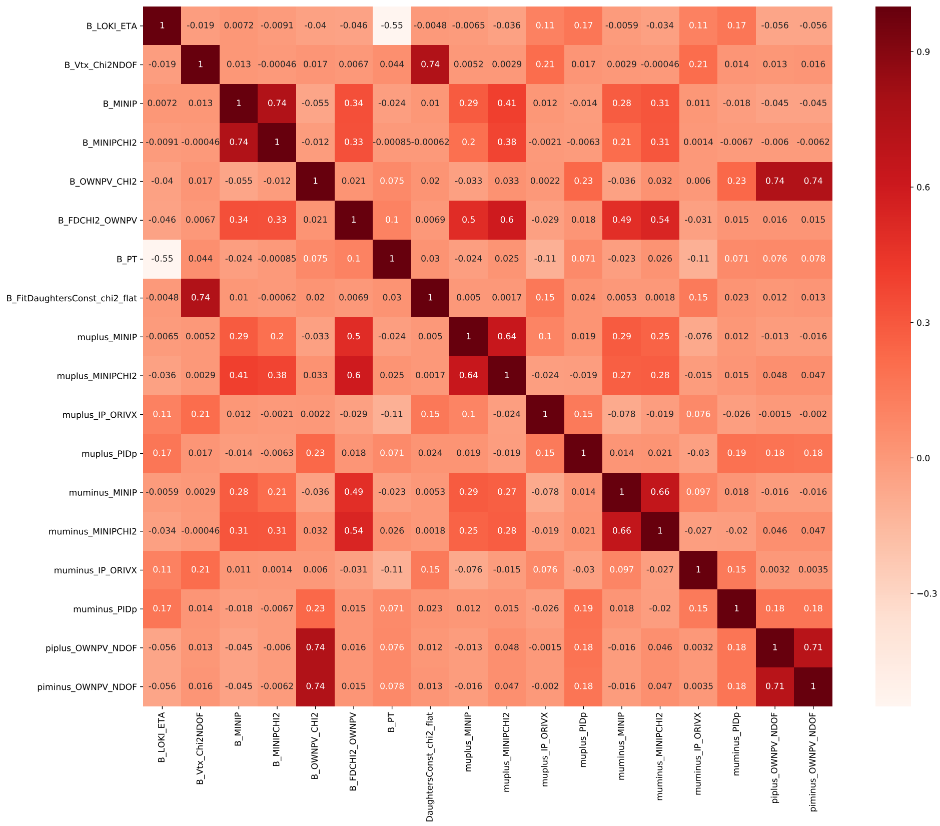


Figure 13: Correlations between the selected variables.