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. Theorie

[1]

2. Durchführung

3. Analysis

The data used in this analysis consists of three different data samples. One is the actual measured data, which contains reconstructed $B^0 \to \psi(2S)K^0_S$ candidates and is dominated by combinatorial background and a B^0_d peak. 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 decay $B^0_d \to \psi(2S)K^0_S$. In Figure 1, the invariant mass distribution of the reconstructed B^0_s events in the signal simulation is shown. As expected, a clear peak at the B^0_s mass can be seen. The mass distribution of the reconstructed B^0 candidates in

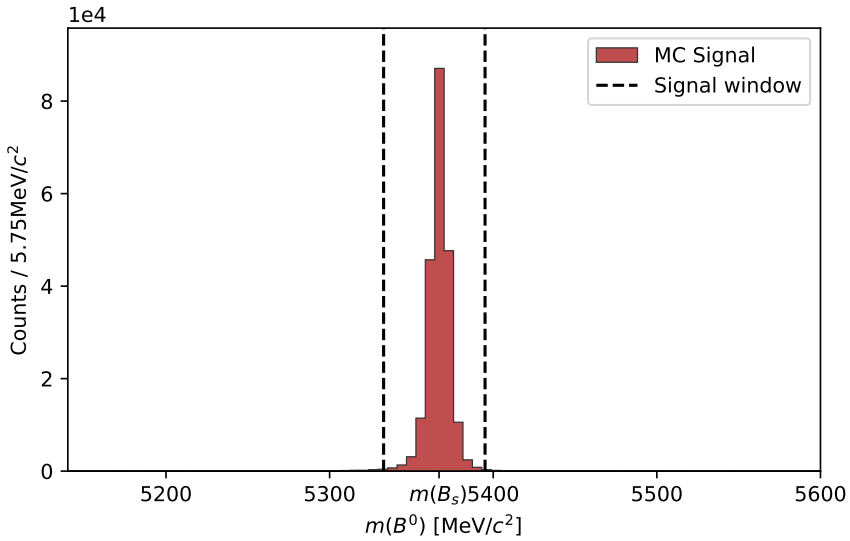

Figure 1: Invariant mass distribution of the B^0_s candidates for the signal channel simulation data.

real data can be seen in Figure 2. Here, a peak at the nominal B^0_d mass can be seen. However, due to dominating combinatorial background, no peak at the B^0_s 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 3.

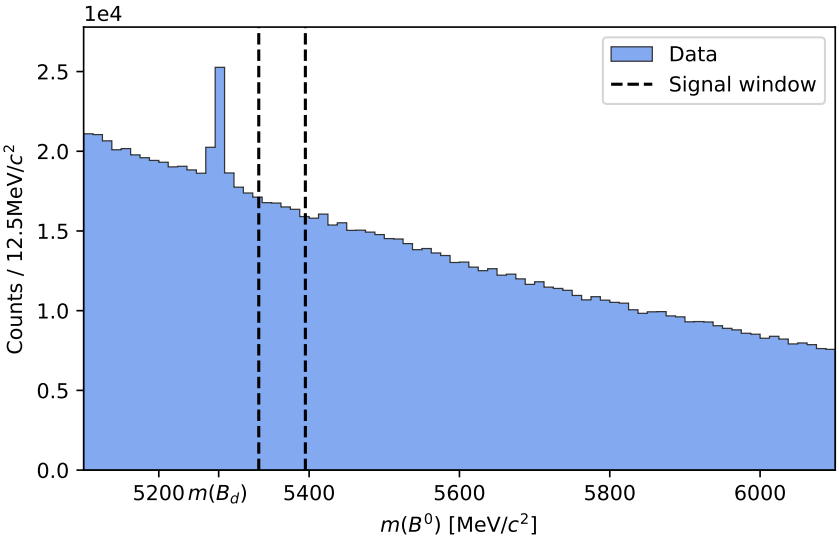

Figure 2: Invariant mass distribution of the B^0 candidates for the recorded LHCb data.

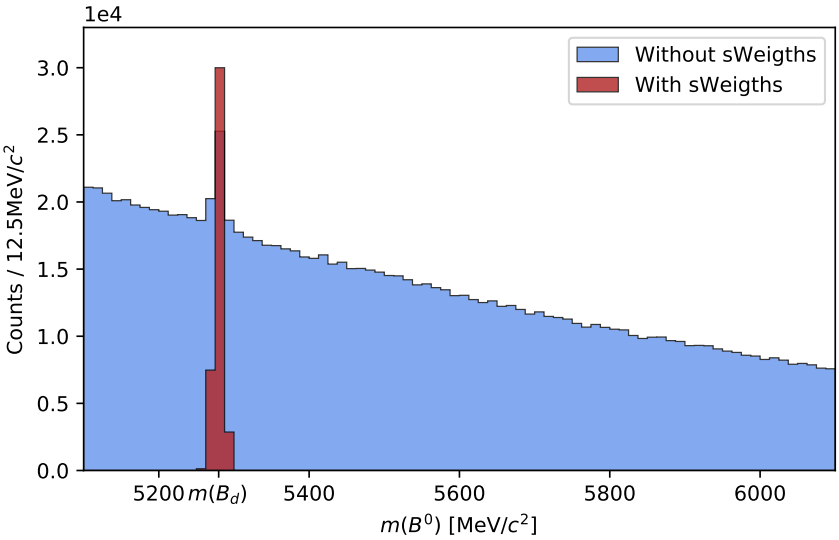

Figure 3: Invariant mass distribution of the B^0 candidates for the recorded LHCb data with and without the sWeights.

3.1. 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 $\mathrm{MeV/c^2}$, defining the signal window of 5333 to 5395 $\mathrm{MeV/c^2}$ which can also be seen in the aforementioned plots (1, 2). Subsequently, the upper sideband ('*USB*'), containing mostly combinatorial background, is defined as the area with reconstructed mass $> 5400\,\mathrm{MeV/c^2}$.

3.2. 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 extracted. 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 **cite theory?** 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 190 variables with a test statistic of $d > 0.05$ 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 control channel data and simulation are shown in Figure 4. Distributions for all variables and the correlation matrix can be found in the appendix A.1.

3.3. 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` [2] 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 corss

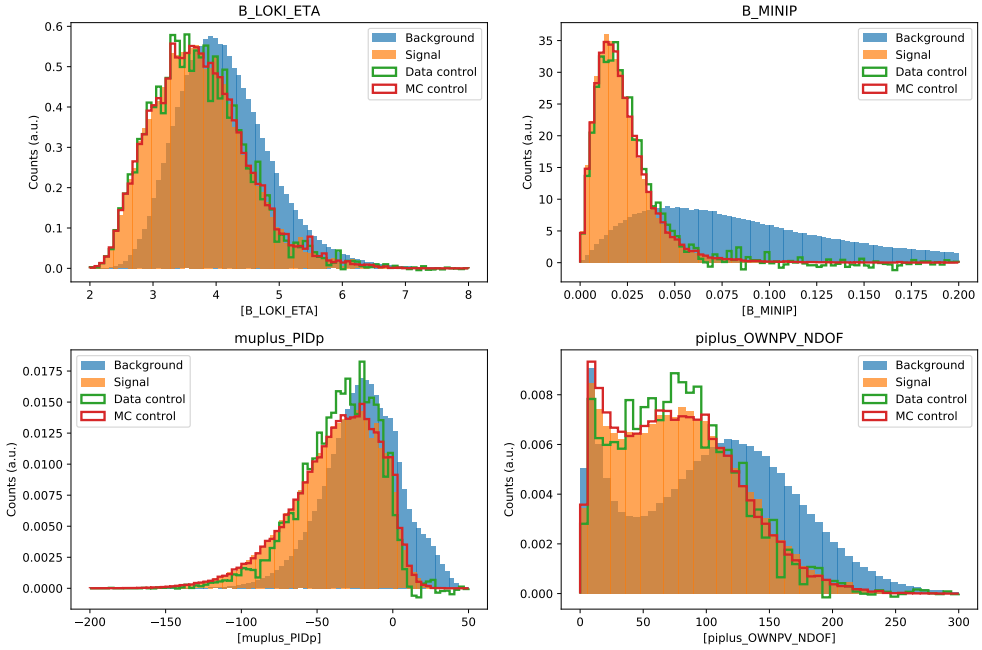

Figure 4: Distributions of four selected variables used in the MVA for simulation, reweighted data and background.

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 5. Additionally, the feature importance of the BDT variables is checked for imbalances. As can be seen in Figure 6, all variables are of similar importance.

3.4. Optimization of the classification threshold

After applying all BDT's to the data, a cut on the classifers response has to be made. Therefore the mean classifer response between all 5 BDT's is calculated. The classification threshold separating signal and background is optimized using the Punzi figure of merit

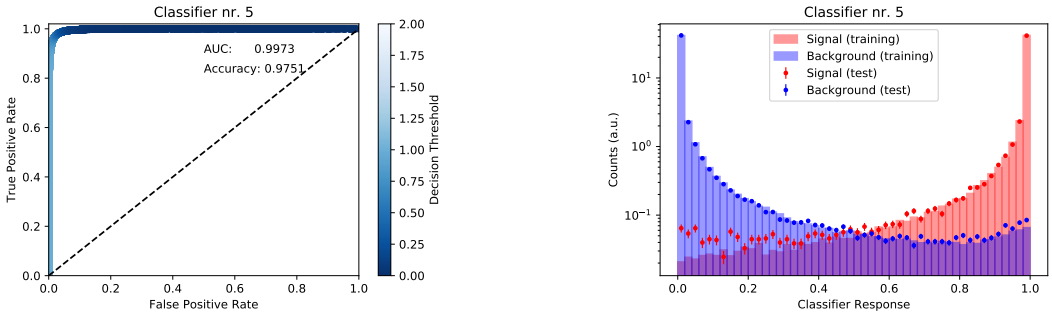

Figure 5: ROC curve (left) and response on training and test dataset (right) for one of the trained classifiers.

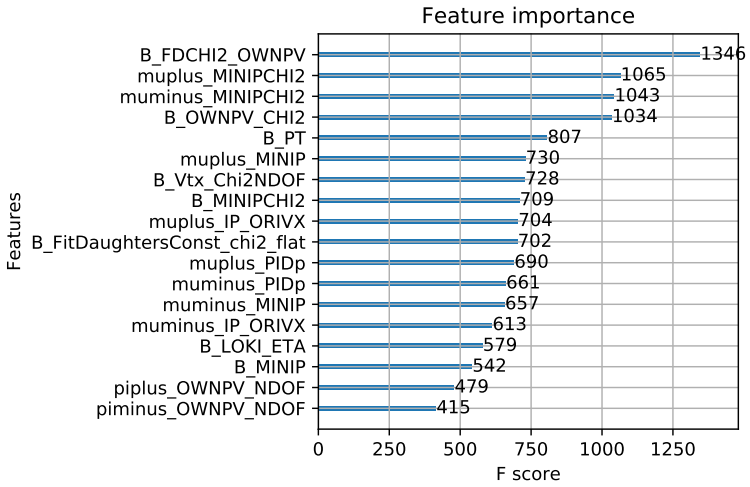

Figure 6: Feature importance of the BDT training variables.

cite theory here. 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 selectrion threshold, and extrapolated to the signal window. For the fitting, the python library `iminuit` [3] is used. The resulting values of the figure of merit are plotted against different thresholds in Figure 7 The optimal classification threshold is

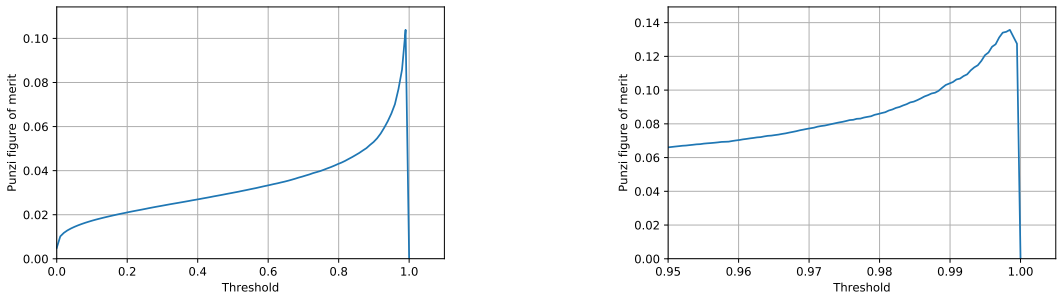

Figure 7: The Punzi figure of merit for the mean classifier response in different intervals of the threshold.

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

3.5. 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 8. A peak of the signal decay $B_s^0 \to \psi(2S)K_S^0$ can be seen.
In order to determine the signal yield, a fit to the mass spectrum of the B^0 candidates is applied. The signal peaks are modeled with two gaussian distributions, which start values for the mean μ and width σ are defined by fits to simulation. The values itself 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 backgorund components. The resulting graph is shown in Figure 9. The fit parameters follow as

$$
\begin{aligned}
s_{B_s} &= 43 \pm 8 & s_{B_d} &= 5010 \pm 71 \\
\mu_{B_s} &= (5366.9 \pm 0.9)\,\frac{\text{MeV}}{\text{c}^2} & \mu_{B_d} &= (5279.9 \pm 0.9)\,\frac{\text{MeV}}{\text{c}^2} \\
\sigma_{B_s} &= (4.6 \pm 0.8)\,\frac{\text{MeV}}{\text{c}^2} & \sigma_{B_s} &= (6.22 \pm 0.07)\,\frac{\text{MeV}}{\text{c}^2} \\
b &= 492 \pm 24 & \tau &= 124 \pm 6.
\end{aligned}
$$

From these, a significance proxy **cite theory?** of the observation of the signal can be calculated. Therefore, the signal and background events in the signal window are

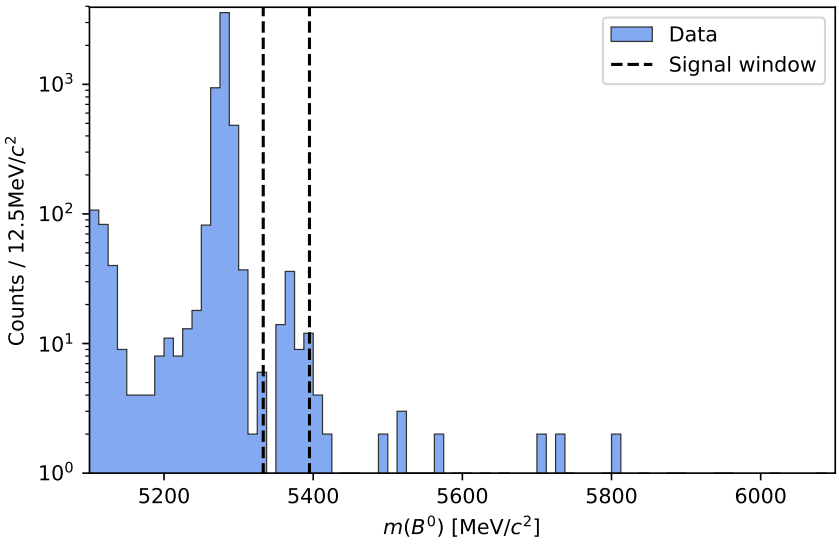

Figure 8: Semi logarithmic invariant mass distribution of the B^0 candidates in data, after the cut on the classifier response is applied.

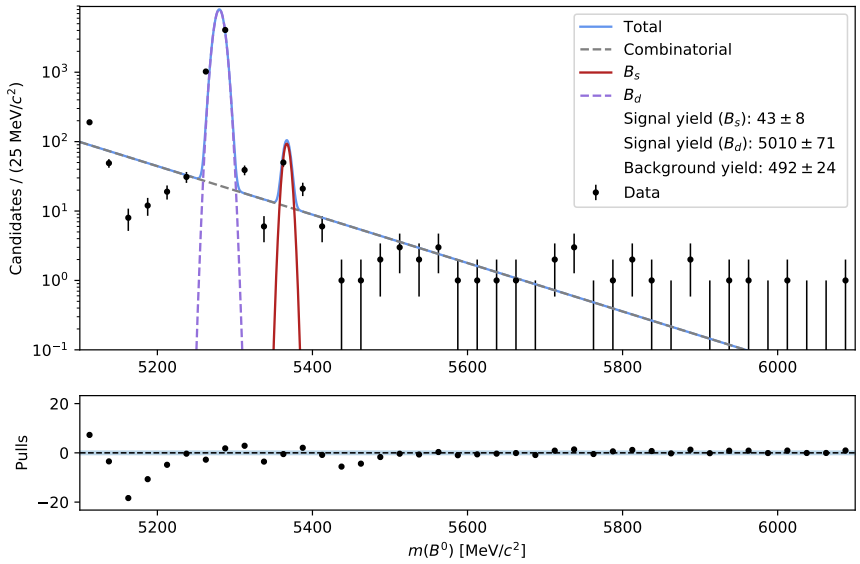

Figure 9: Fit to the invariant mass spectrum of the data in semi logarithmic depiction.

interpolated from the fit results as $n_{\mathrm{sig}} = 43$ and $n_{\mathrm{bkg}} = 30$. The significance proxy then reads $m = 5.07$.

4. Diskussion

References

[1] *Versuch zum Literaturverzeichnis*. TU Dortmund, Fakultät Physik. 2022.

[2] 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.

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

A. Anhang

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

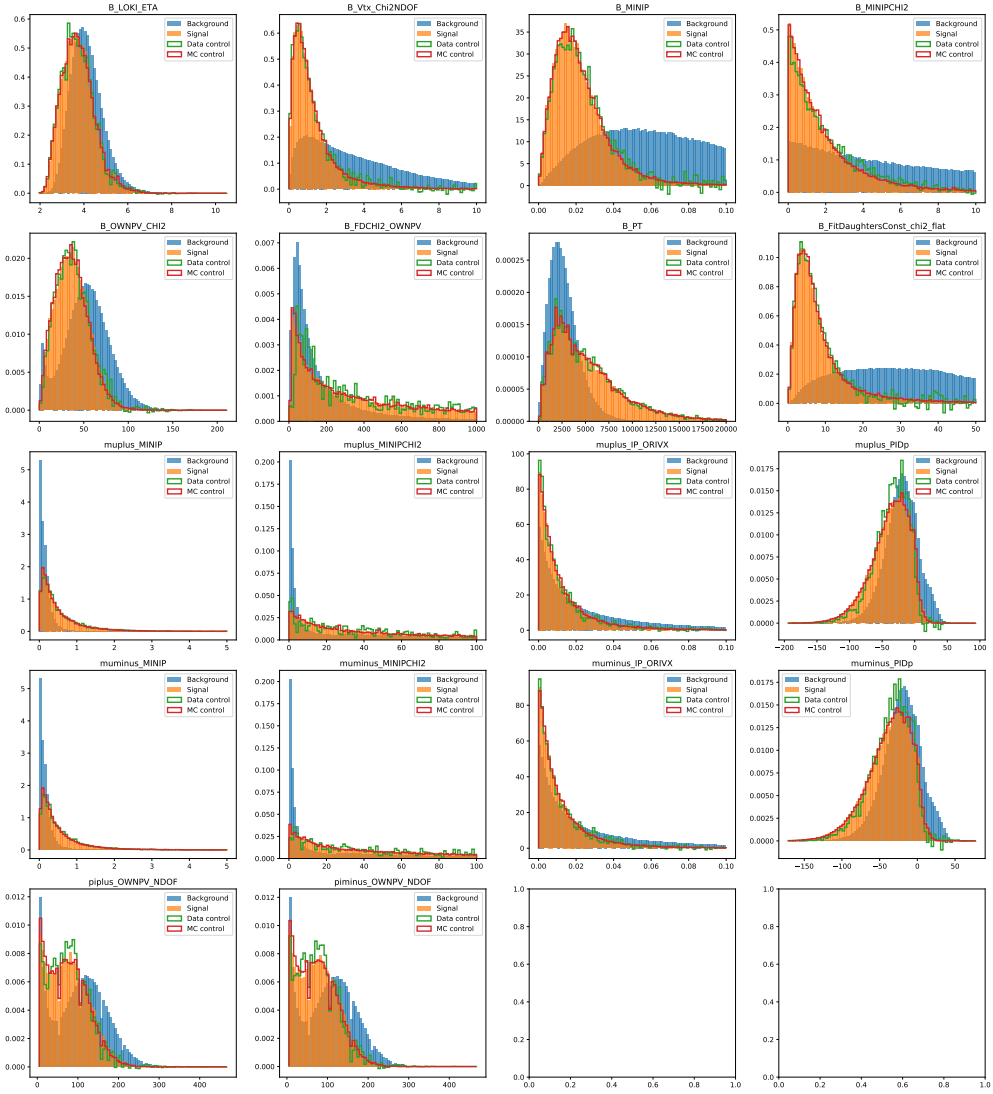

Figure 10: Distributions of the variables used in the MVA for simulation, reweighted data and background.

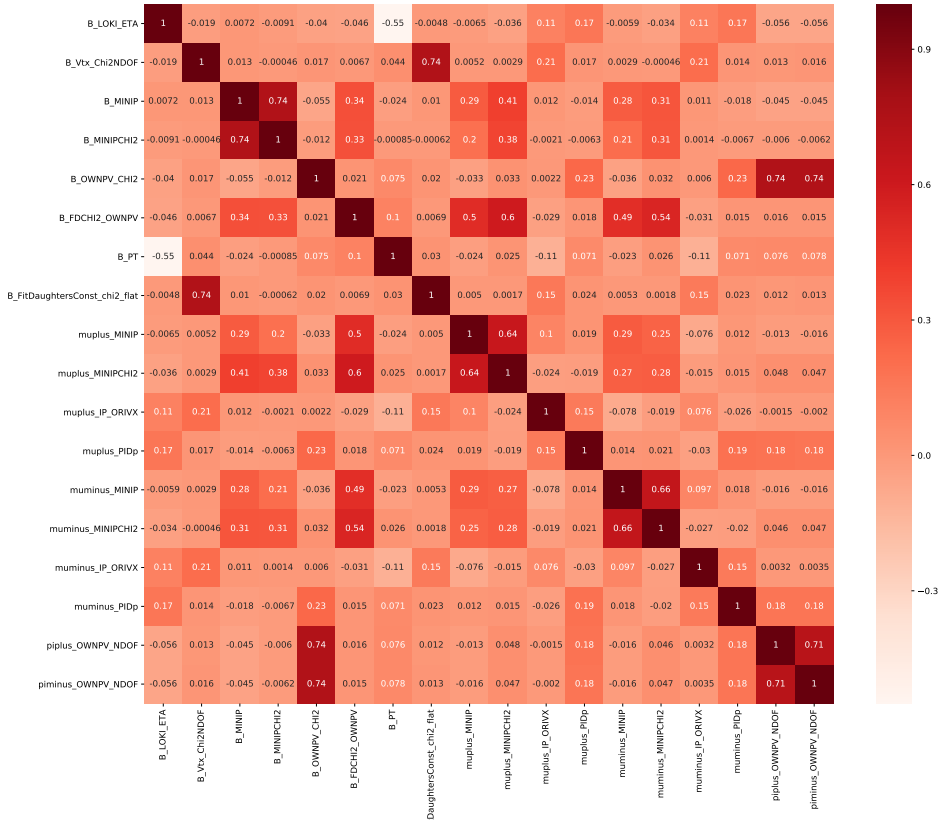

Figure 11: Correlations between the selected variables.