

**Sensitivity Studies
for the Time-Dependent CP Violation
Measurement in $B^0 \rightarrow K_S^0 K_S^0 K_S^0$
at the Belle II-Experiment**

Paul Ferdinand Jäger

**A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE**

INSTITUT FÜR EXPERIMENTELLE KERNPHYSIK,
DEPARTMENT OF PHYSICS,
KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT)

Advisor: Prof. Dr. M. Feindt
Co-Advisor: Prof. Dr. T. Müller

JUNE 30TH, 2016

IEKP-KA/2016-10

Contents

1. Introduction	1
2. Theoretical background	3
2.1. Standard model	3
2.1.1. Hadrons	4
2.1.2. Mixed states	4
2.1.3. Mixing in neutral mesons	7
2.2. CP violation	8
2.2.1. CP Symmetry	8
2.2.2. Categories of CP violation	9
2.2.2.1. Direct CP violation	9
2.2.2.2. Indirect CP violation	9
2.2.2.3. Mixing-induced CP violation	10
2.2.3. CP violation in $B \rightarrow K_S^0 K_S^0 K_S^0$	11
3. Set up of the experiment	13
3.1. The SuperKEKB accelerator	13
3.2. The Belle II-Detector	14
3.3. Measuring time-dependent CP violation	14
4. Analysis setup & methods	17
4.1. Background components	18
4.2. Multivariate classification	18
4.2.1. Non-linear problems	18
4.2.2. Supervised learning	19
4.2.3. Generalization	19
4.2.4. Evaluation	20
4.2.5. Boosted Decision Tree	21
4.2.6. MVA in Belle II	22
4.3. Maximum-Likelihood fit	22
4.3.1. Fit method	22
4.3.2. Error estimation	24
4.4. Data simulation	24
4.5. Flavor-Tagging	26
4.5.1. Calibration	26
4.5.2. Assignment of tagging information	26
4.5.3. Flavor-Tagging quality	27
4.6. Analysis setup	27
4.7. Analysis Software	29
4.7.1. BASF2	29
4.7.2. Fit-Framework	30

5. Vertex Reconstruction	33
5.1. V0 Finder	33
5.2. B^0 vertex fit	35
5.2.1. Signal side	35
5.2.2. Tag side	35
6. Selection	37
6.1. K_s^0 training	37
6.2. B^0 selection	38
6.2.1. BBDT output	38
6.2.2. ΔE and M_{bc}	38
6.2.3. Continuum suppression	40
6.2.4. FSBDT output	42
6.2.5. Δt	44
6.2.6. List of pre-cuts	44
7. Fit models and methods	47
7.1. Signal model	47
7.2. Background model	49
7.3. Combined model	51
7.4. Optimization of the \mathcal{O}_B -cut value	52
8. Results and discussion	53
8.1. Signal MC validation	53
8.2. \mathcal{O}_B -cut optimization results	54
8.3. Comparison to previous studies	56
8.4. FSBDT optimization results	58
8.5. Full MC results	59
8.6. Reconstruction and selection efficiencies	61
8.7. V0 Finder performance	62
9. Conclusion	65
Appendix	69
A. Belle Systematics	70
B. BDT Training results	70
B.1. KSBDT Training results	71
B.2. BBDT Training results	74
B.3. CSBDT Training results	76
B.4. FSBDT Training results	78

1. Introduction

Throughout the construction of the Belle II-Experiment, a new software framework is being developed, the 'Belle II software framework'. This thesis constitutes a validation of the new software tools as well as a sensitivity study for a time-dependent CP violation measurement at Belle II based on simulated Monte Carlo (MC) events. While the initial focus was set on optimizing the K_S^0 selection, the following question emerged: Is it sensible to implement an analysis-specific cut on the selection output based on the uncertainty of the corresponding measurement rather than a general cut based on maximizing a Figure of Merit? Therefore, an analysis was set up for measuring the time-dependent CP violation parameters S_f and A_f . In this approach, the statistical uncertainty of the measurement is fed back to optimize the K_S^0 -selection. In order to ensure a high dependency of the result on the K_S^0 information, the channel $B \rightarrow K_S^0 K_S^0 K_S^0$ was chosen. Furthermore, this channel is estimated to represent an important channel for the observation of New Physics at the Belle II-Experiment [1]. A new handle on minimizing the uncertainty would hence denote a crucial improvement for physics analysis in this channel.

In Chapter 2 an overview of the theoretical background is given by introducing main concepts like the Standard Model and CP violation. The corresponding experimental setup is then introduced in Chapter 3. In Chapter 4 the work-flow of a time-dependent CP violation analysis is presented, where the main concepts of reconstruction and selection processes as well as the statistical tools, which are required for their accomplishment, are described. Finally the specifics of this study with respect to the general analysis setup are pointed out. In Chapter 5 the concept of vertex reconstruction is introduced focusing on a new tracking algorithm, which promises to yield unbiased 4-momenta of K_S^0 mesons. Subsequently all selection training outputs are presented in Chapter 6, thereby listing all pre-cuts applied to the data sample before performing the measurement. The underlying fit model for this measurement is introduced in Chapter 7 as well as the setup of the K_S^0 optimization. In Chapter 8 the obtained results are shown, discussed and compared to previous studies. Finally an evaluation for different reconstruction methods is performed.

2. Theoretical background

2.1. Standard model

The field of physics is a constant interaction between the fields of theory and experiment. If an experiment shows a new observation, theories, e.g. mathematical models, are developed in order to explain it. If a new theory is postulated on the other hand, it is tested by experimentalists until it is either confirmed or proven wrong. In particle physics there has been one certain theory withstanding all its test for several decades. It is called the Standard model and describes the different elementary particles as well as the way they interact with each other, thereby building the matter of the universe. Figure 2.1 shows how the particles of the Standard model can be grouped according to their features and functions. The Quarks and the Leptons are Spin- $\frac{1}{2}$ particles also referred to as *Fermions* and make up all of the visible matter in the universe. The *Bosons* are Spin-1 particles and act as force carriers of the four fundamental interactions. The most recent member of the Bosons group is the famous Higgs-Boson, which was discovered at the LHC in 2012 [2].

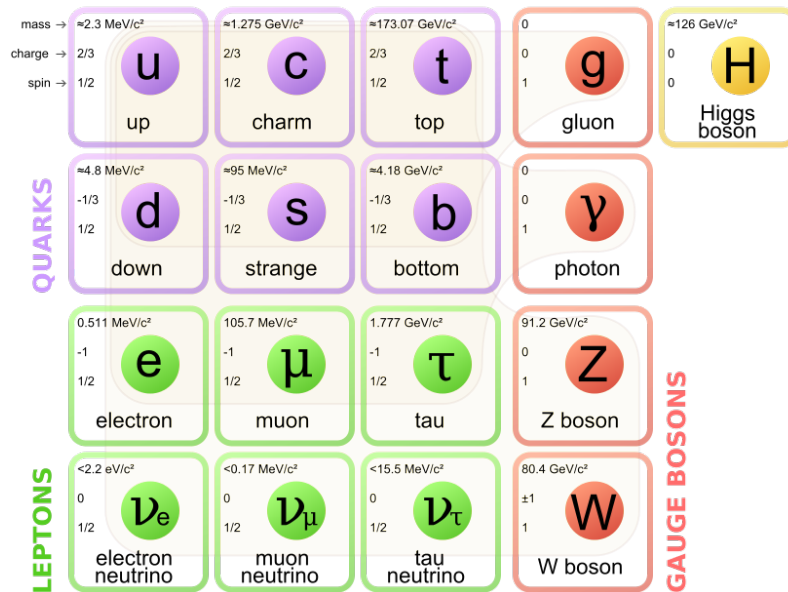


Figure 2.1.: All elementary particles of the Standard model, structured according to their subgroups and functionality. From [3].

2.1.1. Hadrons

Due to the phenomenon of quark confinement [4], quarks do not occur as independent particles, but as bound states called *hadrons*. The underlying interaction is the *strong force*, which is mediated by *gluons*. In the Standard model there are $qq\bar{q}$ states called *baryons* and $q\bar{q}$ states called *mesons*, where the bar denotes an anti-quark, i.e. a quark with opposite charge. Figure 2.2 shows the 16 mesons made up of all possible combinations of the four lightest quarks. Since the heaviest quark, i.e. the top quark, has a very short lifetime of about $0.5 \cdot 10^{-24} \text{s}$ it decays before being able to build a bound state and hence does not occur in hadrons. The hadrons investigated throughout this analysis are $B^0(\bar{B}^0)$ -mesons consisting of the quark combination $\bar{b}d(b\bar{d})$ and Kaons $K_0(\bar{K}_0)$ consisting of $\bar{s}d(s\bar{d})$.

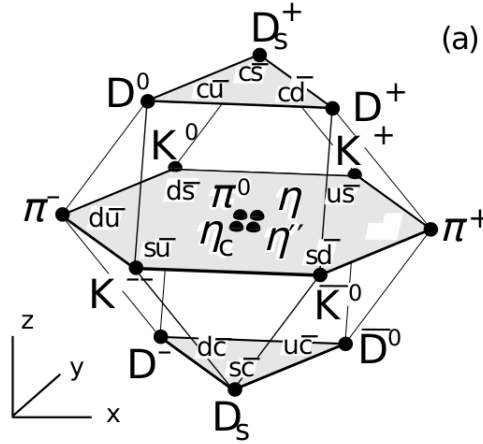


Figure 2.2.: Mesons for different combinations of the four lightest quarks. x-axis: d/u quarks, y-axis: s quarks, z-axis: c quarks. From [5]

2.1.2. Mixed states

One further fundamental interaction is the *weak force*. It describes flavor changing interactions via massive, charged *W-Bosons*, where the term 'flavor' categorizes the different types of fermions in the Standard model and gives them a unique property. The quarks appear in three generations shown in Equation 2.1. Classically the flavor change in a decay via weak interaction occurs within the generations.

$$\begin{pmatrix} u \\ d \end{pmatrix} \quad \begin{pmatrix} c \\ s \end{pmatrix} \quad \begin{pmatrix} t \\ b \end{pmatrix} \quad (2.1)$$

Another property of the particles is the mass. As for the quarks it is not possible to assign a certain mass to each flavor of the leptons, the *flavor eigenstates* do not equal the *mass eigenstates*. Hence the mass eigenstates are mixed states containing several flavor eigenstates to certain proportions. This phenomenon makes flavor changing between the generations possible. The probability of transitions into a certain flavor is described by the transition amplitude. These amplitudes are collected in the so called CKM-Matrix shown in Equation 2.2 [4]. Since its elements are a straight deduction from interaction theory, their precise measurement denotes a very popular test of the Standard model. These 'flavor physics' measurements have been conducted by experiments like Belle, BaBar and LHCb.

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix}_{\text{flavor}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \cdot \begin{pmatrix} d \\ s \\ b \end{pmatrix}_{\text{mass}} \quad (2.2)$$

Since the CKM-Matrix is unitary by definition, its number of free parameters reduces to four. These can be showed by expressing the Matrix as a three dimensional rotation between the three quark generations.

$$V_{CKM} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13}e^{-i\delta} \\ 0 & 1 & 0 \\ -s_{13}e^{i\delta} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (2.3)$$

$$= \begin{pmatrix} c_{12}c_{23} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix} \quad (2.4)$$

Here $s_{ij} = \sin(\theta_{ij})$ and $c_{ij} = \cos(\theta_{ij})$, where the angles θ_{ij} denote rotations between the generations i and j and make up three of the free parameters. The fourth parameter is the complex phase δ . It represents the only source of CP-Violation predicted by the Standard model, which is discussed further in Section 2.2.

Evaluating measurements on the CKM-Matrix elements has shown a strong hierarchical structure. Thus transitions within generations are highly favoured, i.e. $V_{ud}, V_{cs}, V_{tb} \approx 1$, whereas transitions between generations are highly suppressed. In order to express this structure in the matrix representation, the Wolfenstein parametrization using the parameters λ , A , ρ and η [6] is used, where

$$s_{12} = \lambda = \frac{|V_{us}|}{\sqrt{|V_{ud}|^2 + |V_{us}|^2}}, \quad s_{23} = A\lambda^2 = \lambda \left| \frac{V_{cb}}{V_{us}} \right|, \quad (2.5)$$

$$s_{13}e^{i\delta} = A\lambda^3(\rho + i\eta) = V_{ub}^*. \quad (2.6)$$

The Matrix can now be expanded in orders of $\lambda \approx 0.23$:

$$\begin{pmatrix} 1 - \frac{1}{2}\lambda^2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1 - \frac{1}{2}\lambda^2 & A\lambda^2 \\ A\lambda^3(\rho + i\eta) & -A\lambda^2 & 1 \end{pmatrix} = \mathcal{O}(\lambda^4). \quad (2.7)$$

This depiction visualizes the suppression in orders of λ (see also Table 2.1).

generation	orders of λ
within gen.	≈ 0
$1 \leftrightarrow 2$	λ
$2 \leftrightarrow 3$	λ^2
$1 \leftrightarrow 3$	λ^3

Table 2.1.: Suppression experienced for the transitions between different generations

Equation 2.7 shows also the occurrence of the CP violating phase in the elements V_{td} and V_{ub} . The latter causes CP-Violation to appear in several physical processes of the B meson such as the one investigated throughout this analysis. From the unitarity relations $V_{CKM} \cdot V_{CKM}^\dagger = V_{CKM}^\dagger \cdot V_{CKM} = 1$ six independent relations between the matrix elements can be deducted, where one of them looks like

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = \delta_{db} = 0. \quad (2.8)$$

These relations can be displayed as triangles in the complex plane, where the side lengths can be expressed again in orders of λ . Since the relation shown in Equation 2.8 is the only one providing side lengths of the same order, it is chosen for further investigation and called the *Unitarity Triangle* [7]. Figure 2.3 shows the triangle. Note that all sides are normalized by the experimentally best determined side $|V_{cd}V_{cb}^*|$.

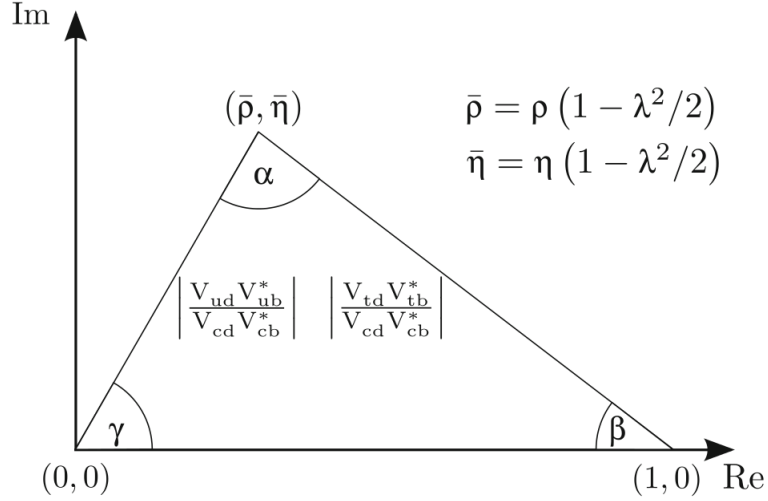


Figure 2.3.: Construction of the Unitarity Triangle in the complex plane. From [8]

The usual approach in flavor physics to testing the Standard model is to over-constrain the parameters of the Unitarity Triangle by many independent measurements. Conclusions can then be drawn by investigating the triangle's shape, e.g whether it is flat or closes. The angles of the Unitarity Triangle are defined as

$$\Phi_1 = \beta = \arg[-V_{cd}V_{cb}^*/V_{td}V_{tb}^*], \quad (2.9)$$

$$\Phi_2 = \alpha = \arg[-V_{td}V_{tb}^*/V_{ud}V_{ub}^*], \quad (2.10)$$

$$\Phi_3 = \gamma = \arg[-V_{ud}V_{ub}^*/V_{cd}V_{cb}^*], \quad (2.11)$$

where different notation conventions are used in the literature for these angles.

Figure 2.3 shows the current status of the determined constraints on the parameters.

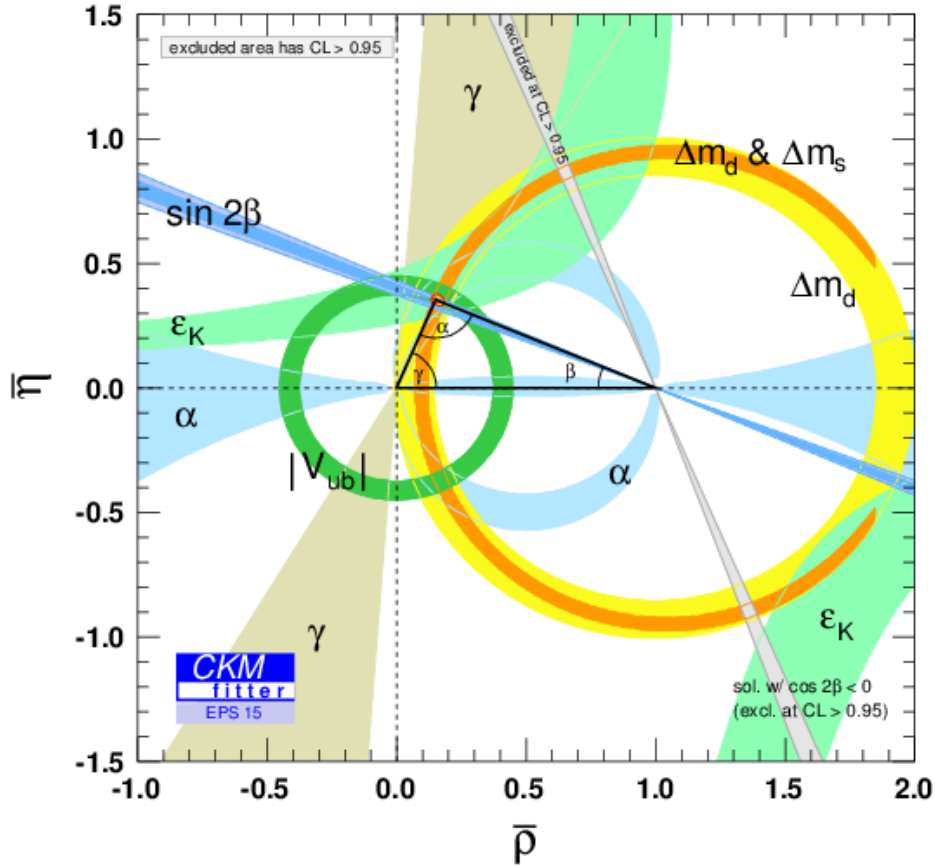


Figure 2.4.: Constraints on the Unitarity Triangle in the $\bar{\rho}$ - $\bar{\eta}$ -plane at current experimental status. The figure has been provided by the CKMfitter Group [9].

2.1.3. Mixing in neutral mesons

The phenomenon of mixing describes many physical processes of the Standard model related to the weak interaction. The B^0 mesons consist of the following quark combinations:

$$B^0 = d\bar{b}, \quad \bar{B}^0 = b\bar{d}. \quad (2.12)$$

As described in the previous section, these quark combinations only represent the mesons' flavor eigenstates and are different from their mass eigenstates. The latter are given by a linear combination leading to a heavier B meson B_H and a lighter B meson B_L :

$$|B_L\rangle = p|B^0\rangle + q|\bar{B}^0\rangle, \quad (2.13)$$

$$|B_H\rangle = p|B^0\rangle - q|\bar{B}^0\rangle, \quad (2.14)$$

with $p \approx q$. Since the B^0 mesons are produced as a $B\bar{B}$ pair, they constitute a quantum entangled mixed state with an oscillation according to Equation 2.13. Hence the probability of finding a B^0 or \bar{B}^0 in a state, which has been produced at $t = 0$ as a B^0 is given by

$$P_{B^0}(t) = \frac{1}{2}e^{-t/\tau_{B^0}}(1 + \cos(\Delta m t)), \quad (2.15)$$

$$P_{\bar{B}^0}(t) = \frac{1}{2}e^{-t/\tau_{B^0}}(1 - \cos(\Delta m t)), \quad (2.16)$$

where τ_{B^0} denotes the B^0 lifetime and Δm the mass difference of the two mass eigenstates. As shown in Figure 2.5 the mixing can be visualized by using Box diagrams.

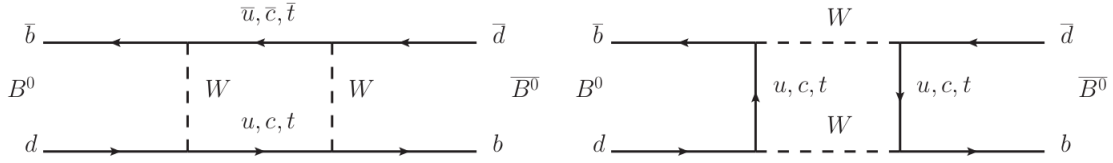


Figure 2.5.: Box diagrams for mixing of neutral B mesons.

2.2. CP violation

Despite the fact, that the Standard model has not been proven wrong yet, it is not able to explain all observations, that have been made by particle physics experiments. Popular examples are the hierarchy problem, the observed neutrino mass or missing dark matter candidates [4]. In order to cover these persisting discrepancies several suggestions for extensions to the classical model exist including new particles and interactions. These models are grouped under the term *physics beyond the Standard model*, or *New physics*. Any observation that would show a significant deviation from the Standard model would thus be a hint for New physics. The Belle II-Experiment focuses on investigating yet another phenomenon called *CP violation*, which causes the asymmetry between matter and anti-matter in the universe.

This section introduces the concept of CP symmetry before shedding light on the different ways of how it is broken. Finally the occurrence of CP violation in the decay channel of interest, $B^0 \rightarrow K_S^0 K_S^0 K_S^0$, is discussed.

2.2.1. CP Symmetry

When talking about CP violation, one refers to a breaking of the *CP symmetry*. The concept of symmetries plays an important role in particle physics and the Standard model: Conservation laws of variables like energy or momentum can be deduced from a phenomenon called *invariance* in space or time. Such invariance is present, if certain parameters (e.g the position in space or the direction of time) of a system can be changed without changing the underlying laws of physics.

CP symmetry represents a combination of charge symmetry (C) and parity symmetry (P). The former refers to an invariance under *charge conjugation*, which describes the replacement of all particles belonging to a certain decay by their anti-particles. The latter describes the invariance under *parity transformation*, where all particles get mirrored at the origin of the system, i.e. the signs of their spatial coordinates are flipped. While both symmetries have been known to be broken by the weak interaction since the 1950s, the breaking of their combination, i.e. the CP symmetry, describing the invariance under both described transformations applied at once, took roughly one more decade to be observed.

The Standard model predicts three sources for CP violation in the strong-, lepton-, and the weak-sector. While the first two have not been observed yet and are far from being done so, the latter has been the subject to many investigations and measurements. It is represented by the complex phase of the CKM-Matrix described in Equation 2.4. So far there have been no significant discrepancies between CP violation measurements and the Standard model prediction. The discrepancy lies however in the unexplained asymmetry observed in the universe, as

the predicted and measured CP violation is too small to account for it. Since the parameters measured in CP violation experiments are very sensitive, new measurements may reveal significant deviations from the Standard model, i.e. discover New Physics.

2.2.2. Categories of CP violation

There are three known forms of CP violation: direct, indirect, and mixing induced, with the latter being the one investigated by this analysis. All three are visualized by Figure 2.6 and shall be introduced briefly in the following.

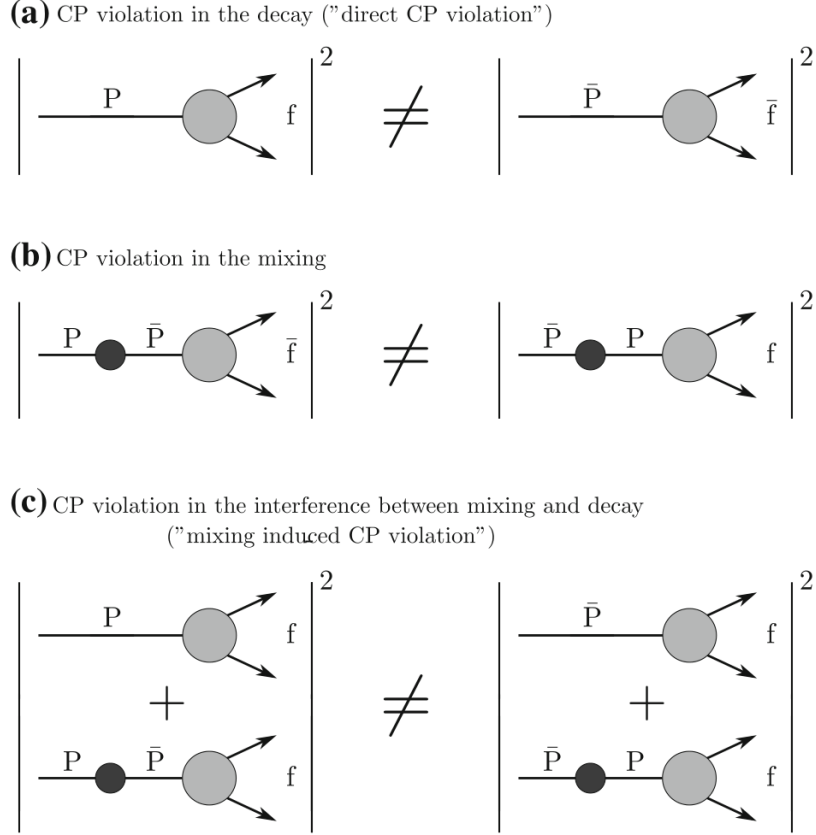


Figure 2.6.: Illustration of the three forms of CP violation.

2.2.2.1. Direct CP violation

This form of CP violation is also referred to as *CP violation in the decay*. It can occur, if the amplitudes, i.e. the transition probabilities, of two CP conjugated processes $P \rightarrow f$ and $\bar{P} \rightarrow \bar{f}$ have different magnitudes. Here P and f denote the decaying particle and its final state decay product respectively and the bar denotes CP conjugation. Direct CP violation occurs, if the corresponding amplitudes are not equal (see also Figure 2.6(a)):

$$\frac{A_f}{\bar{A}_{\bar{f}}} \neq 1. \quad (2.17)$$

2.2.2.2. Indirect CP violation

This form of CP violation is also referred to as *CP violation in the mixing*. Consider the flavor-specific decays from the previous subsection and no direct CP violation ($A_f = \bar{A}_{\bar{f}}$). The

forbidden decays $P \rightarrow \bar{f}$ and $\bar{P} \rightarrow f$ are only possible due to $P^0 - \bar{P}^0$ oscillations as described in Section 2.1.3. The CP violation then corresponds to unequal mixing probabilities (see also Figure 2.6(b)):

$$|\langle P^0 | \bar{P}(t) \rangle|^2 \neq |\langle \bar{P}^0 | P(t) \rangle|^2. \quad (2.18)$$

Note that the effect of indirect CP violation is negligible in the B meson system and hence not discussed further.

2.2.2.3. Mixing-induced CP violation

This form of CP violation is also referred to as *CP violation in the interference between mixing and decay*. Since it denotes the CP violation category to be studied throughout this analysis, it shall be explained in more detail here focusing on its occurrence in the B^0 meson system.

If the final state of a decay is a CP eigenstate, i.e. $f = \bar{f}$ for both B^0 and \bar{B}^0 , an interference between the decay without mixing $B^0 \rightarrow f$ and the decay with mixing $B^0 \rightarrow \bar{B}^0 \rightarrow f$ emerges. A sketch of this effect is shown in Figure 2.6(c). Note that since the mentioned interference depends on the $B^0 \bar{B}^0$ oscillation and hence the time of propagation, this form of CP violation is also called *time-dependent CP violation*. The asymmetry between the two possible decays is defined by

$$A_{CP}(t) = \frac{\Gamma(\bar{B}^0(t) \rightarrow f_{CP}) - \Gamma(B^0(t) \rightarrow f_{CP})}{\Gamma(\bar{B}^0(t) \rightarrow f_{CP}) + \Gamma(B^0(t) \rightarrow f_{CP})}, \quad (2.19)$$

which denotes the normalized difference of their decay rates. This definition reveals how crucial the flavor information, i.e. whether the meson decaying into f_{CP} was a B^0 or \bar{B}^0 , is for studying this form of CP violation. However, since f represents a CP eigenstate, that is common to both originating B mesons independent of their flavor, it does not contain this information, which can hence not be extracted from the signal-side reconstruction. Instead, the effect of quantum entanglement is exploited here: Due to their coherent evolvement the first decaying B meson gets a determined flavor content at the time of its decay forcing the accompanying B meson to have the opposite flavor. Hence the required flavor information can be extracted by performing a *flavor tag* on the accompanying B meson, also referred to as B_{tag} . This flavor tag is performed by using different kinds of information from the corresponding *tag side* decay, e.g. the charge of high momentum leptons. Since the second B meson to decay can still exhibit oscillations and hence change its flavor after the first one decayed, the act of introducing the flavor tag information transfers the time dependence of $A_{CP}(t)$ into a dependence on the proper decay time difference $\Delta t = t_{CP} - t_{tag}$ between the two B mesons:

$$A_{CP}(t) = \frac{\Gamma[\Psi(t_{CP}, t_{tag}) \rightarrow f_{CP}, \bar{f}_{tag}] - \Gamma[\Psi(t_{CP}, t_{tag}) \rightarrow f_{CP}, f_{tag}]}{\Gamma[\Psi(t_{CP}, t_{tag}) \rightarrow f_{CP}, \bar{f}_{tag}] + \Gamma[\Psi(t_{CP}, t_{tag}) \rightarrow f_{CP}, f_{tag}]}. \quad (2.20)$$

Here $\Psi(t_{CP}, t_{tag})$ denotes the wave function of the coherent mixing state of the two B mesons depending on their respective decay times. Using the time evolution relations of the two mesons described in Section 2.1.3, the following expression can be deduced for the asymmetry:

$$A_{CP}(\Delta t) = S_f \sin(\Delta m \Delta t) - A_f \cos(\Delta m \Delta t). \quad (2.21)$$

Here Δm denotes the mass difference between the two B mesons (see Equation 2.15). S_f is a variable representing the mixing induced CP violation and can be related to the CKM Matrix elements by $S_f = -\xi \sin(2\beta)$, where ξ is either 1 or -1 for CP even or odd final states and β denotes an angle of the Unitarity Triangle (see Equation 2.9). A_f is a variable related to the direct CP violation. Both parameters S_f and A_f are referred to as the *time-dependent*

CP violation parameters. Using Equation 2.21, the probability of finding one B meson in a CP eigenstate and the second one decaying into flavor-specific state with a proper decay time difference Δt is found to be:

$$P(\Delta t) = \frac{e^{-|\Delta t|/\tau_{B^0}}}{4\tau_{B^0}} [1 + q(S_f \sin(\Delta m_{B^0} \Delta t) + A_f \cos(\Delta m_{B^0} \Delta t))]. \quad (2.22)$$

Here $q = +1(-1)$ denotes the b-flavor charge, if the accompanying B meson is tagged as a $B^0(\bar{B}^0)$. Since the flavor tagging information is not provided as per-event information, but as an overall wrong tag fraction w (due to reasons further discussed in Section 4.5), Equation 2.22 can be expressed as:

$$P_{sig}(\Delta t) = \frac{e^{-|\Delta t|/\tau_{B^0}}}{4\tau_{B^0}} [1 + q\Delta w + q(1 - 2w) (S_f \sin(\Delta m_{B^0} \Delta t) + A_f \cos(\Delta m_{B^0} \Delta t))]. \quad (2.23)$$

This equation describes the expected behaviour of B^0 mesons under mixing induced CP violation and will hence be used as the signal model for the measurement performed in this analysis. By containing the time-dependent CP parameters it gives experimental access to the investigation of this form of CP violation. Therefore a data set containing the proper decay time difference Δt and the flavor tagging information w and Δw for a large amount of B^0 mesons needs to be produced by an experiment. Subsequently an unbinned Maximum Likelihood fit can be performed on this data set with S_f and A_f being the only floating parameters.

2.2.3. CP violation in $B \rightarrow K_S^0 K_S^0 K_S^0$

The signal decay channel $B \rightarrow K_S^0 K_S^0 K_S^0$ proceeds via a $b \rightarrow s$ transition on the quark level. Note, that this flavor change does not include a charge transfer and can hence not simply be carried out by a charged W^\pm boson. Such transitions are called *Flavor changing neutral currents* (FCNC) and are highly suppressed in the Standard model. Due to their shape, the respective Feynman diagrams are also referred to as *penguin diagrams*. The mixing-induced CP violation in this decay is given by an interference between the decays shown in Figure 2.7. Note, that these Feynman diagrams correspond to the sketch in Figure 2.6(c) and also contain the $B^0 \bar{B}^0$ oscillations from Figure 2.5.

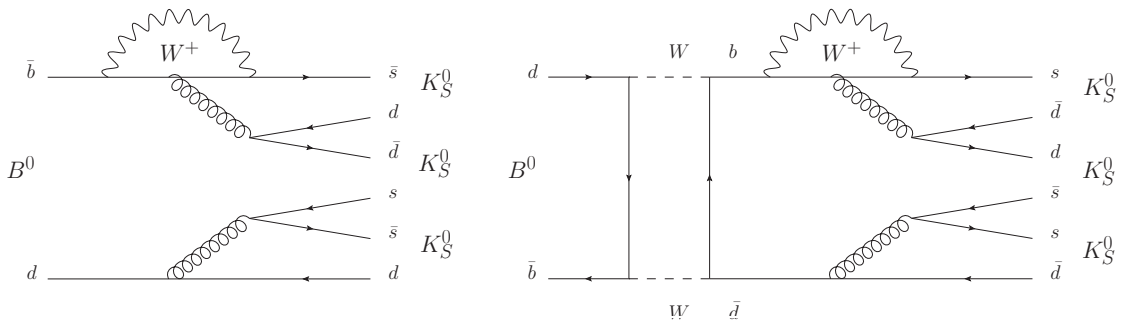


Figure 2.7.: Feynman diagrams of the signal decay channel without mixing (*left*) and with mixing (*right*). The interference of both decays causes the mixing induced CP violation.

3. Set up of the experiment

As described in the previous section, in order to study time dependent CP violation, the proper decay time difference Δt and the flavor tagging information w and Δw of a large amount of B^0 mesons need to be measured. For the production of B^0 meson pairs high energy particles are accelerated and collided inside a particle detector. This chapter introduces the corresponding accelerator and detector for this study, before giving an explicit description on how to set up a time-dependent CP violation measurement.

3.1. The SuperKEKB accelerator

The SuperKEKB accelerator is currently being built as an upgrade of its predecessor, the KEKB accelerator, and situated in Japan. It is constructed as a two-ring accelerator, where high energy beams of electrons and positrons are generated in opposite directions and subsequently crossed in order to cause a controlled collision [10]. When colliding, the beams are tuned to exhibit a centre-of-mass energy of

$$E_{CMS} = 2\sqrt{E_{HER}E_{LER}} = 10.58 \text{ GeV}, \quad (3.1)$$

where HER denotes the high energy ring and LER the low energy ring. The asymmetry of the energies in the two rings is a crucial feature of the collider, since it is necessary for performing time-dependent measurements (see section 3.3). The centre-of-mass energy of 10.58 GeV corresponds to the mass of the $Y(4S)$ resonance, which is an excited state of bottomium, i.e. a bound system consisting of a $b\bar{b}$ quark pair. Therefore it decays into an entangled $B\bar{B}$ pair with a branching fraction of more than 96%. Figure 3.1 shows the cross section of e^+e^- collisions. Since the collision energy equals the mass of the resonance, the corresponding $B\bar{B}$ pair is produced almost at rest. However, the described asymmetry of beam energies in the two rings induces a Lorentz Boost of the $B\bar{B}$ pair, which causes a drift along the z-axis, i.e. parallel to the beam pipe:

$$\beta\gamma = \frac{E_{HER} - E_{LER}}{E_{CMS}} = 0.42. \quad (3.2)$$

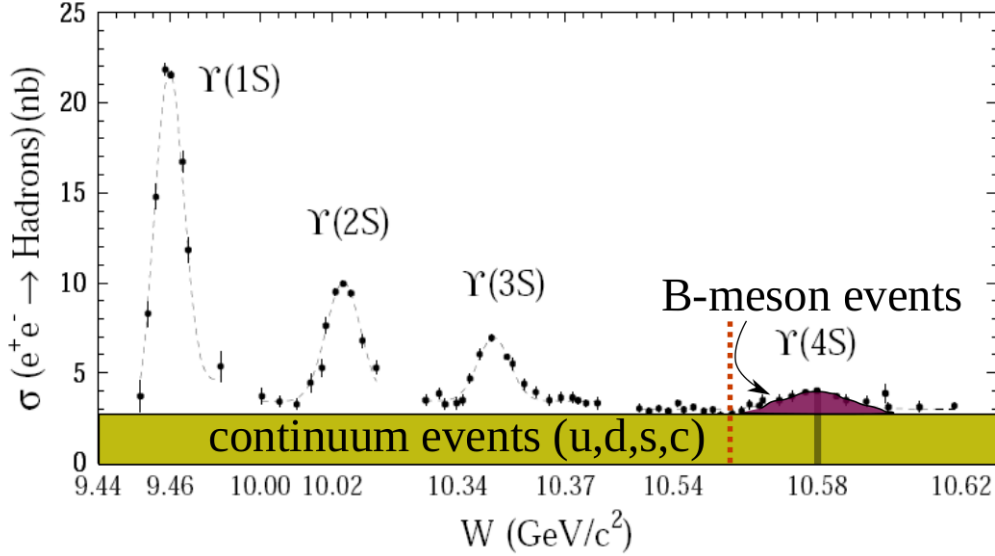


Figure 3.1.: The e^+e^- to hadrons cross section in the $Y(1S)$ - $Y(4S)$ region. The red dashed line marks the kinematic threshold for the production of a $B\bar{B}$ pair. From [11].

3.2. The Belle II-Detector

The Belle II-Experiment is a particle detector currently being built as an upgrade to its predecessor, the Belle detector. It will be placed along the SuperKEKB accelerator, more specifically at the crossing point of the two beams in order to analyze the collisions. It consists of different layers in the x-y plane around the beampipe, with each layer being built in order to extract different information from the collisions. The crucial layers for this study are the three closest to the beampipe: The two vertex detectors (PXD and SVD) are used for precisely reconstructing decay vertices (see Chapter 5) and for finding low momentum tracks. The Central Drift Chamber measures the momentum of particles by determining their curvature. A more detailed description of the detector can be found in [12].

3.3. Measuring time-dependent CP violation

This section describes how the information required for a time-dependent CP violation study is extracted from the entangled $B^0\bar{B}^0$ pair produced inside the Belle II detector. Figure 3.2 shows a sketch of the corresponding setup. As described in Section 3.1, the entangled pair is boosted alongside the z-axis. This way, the proper decay time difference Δt can be determined by measuring the difference of distances of the two corresponding decay vertices:

$$\Delta z = \gamma\beta c\Delta t. \quad (3.3)$$

Determining the two vertices with high resolution is crucial for the precision of this measurement and is explained further in Chapter 5. The flavor of the accompanying B meson, i.e. the B_{tag} , is identified by the Flavor tagging tool described in Section 4.5.

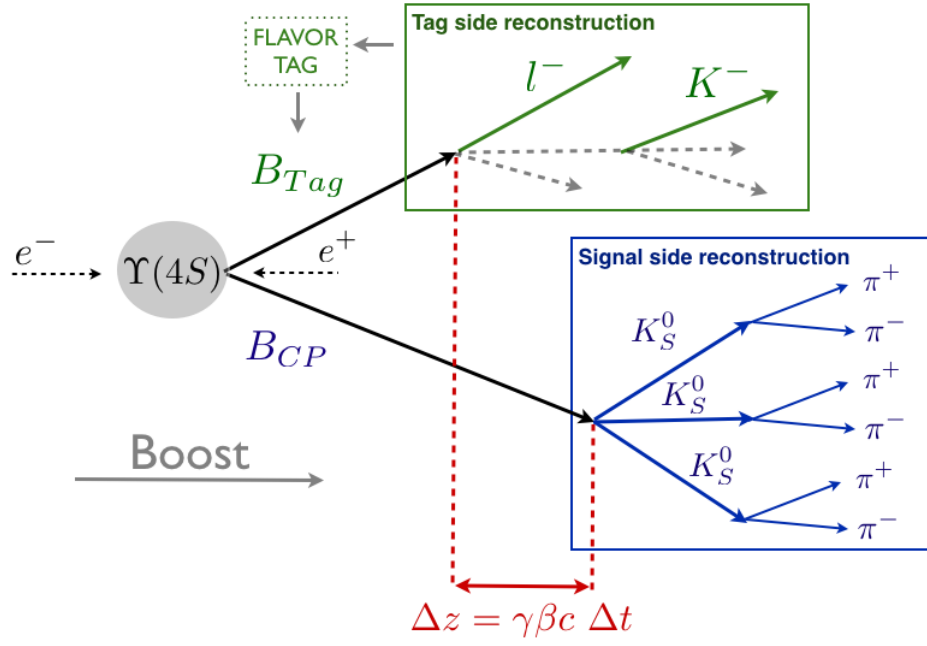


Figure 3.2.: Setup of a time-dependent measurement.

4. Analysis setup & methods

Every e^+e^- collision caused by the particle collider leaves several tracks and clusters in the detector layers. All traces belonging to the same collision are referred to as one *event*. If an event happens to include the decay channel of interest, $B^0 \rightarrow K_S^0 K_S^0 K_S^0$, this event is called a *signal event*. The traces left however, represent only the end of the decay chain (see Figure 3.2). Hence if one was to measure a physical process related to a particle further up in the decay chain, e.g. B^0 mesons, this particle would need to be reconstructed from the information given by the tracks and clusters. This procedure is thus referred to as *reconstruction* and constitutes a reverse engineering process: The information from tracks and clusters is used to form variables of interest, e.g. the energy or momentum of a particle. Depending on the decay channel investigated, these variables are then used as criteria for combining the tracks and clusters in order to form candidates of the particle, which represents the previous step in the decay chain. In doing so a list of candidates is produced for each such step while reconstructing the whole decay process. Finally a list of candidates for the initial B^0 meson is produced.

All candidates in the list that represent reconstructed B^0 meson coming from signal events are called *signal candidates*. Note that also wrong candidates can be formed in signal events, e.g. by wrong combinations, which are then referred to as mis-reconstructed signal. These candidates are ignored throughout this analysis however, since only a negligible amount has been found. The wrong candidates being reconstructed in background events are called *background candidates*. The different background components are presented in Section 4.1. The next step is to optimize the composition of the candidate list according to the requirements of this measurement. The quality of a measurement's result is among other things given by its statistical uncertainty and the extent to which the resulting value is biased. The first criteria is optimized by maximizing the number of contributing candidates. This optimization however causes a conflict of interest with the latter criteria, since wrong candidates can lead to a bias. Hence a method is needed that removes background candidates from a list while keeping the signal candidates. This task is referred to as the *selection* of an analysis. As the selection can also be described by tagging each candidate as either signal or background, it is often reduced to a standard problem of statistics called *classification*. The classification method used throughout this analysis is called *multivariate classification* and is introduced in Section 4.2. After the selection the actual measurement can be performed on the candidate list, in this case represented by an unbinned Maximum Likelihood fit of the time-dependent CP violation parameters S_f and A_f . The statistical background for this fit method is explained in Section 4.3. Subsequently, in Section 4.4, the concept of a 'blind analysis' is introduced and the usage of

different types of simulated data is discussed. In Section 4.5 the Flavor Tagger training output and calibration procedure is presented. An overview of the analysis work-flow regarding the selection, the fit, and the optimization is given in Section 4.6. Finally the software frameworks used throughout this analysis are presented in Section 4.7.

4.1. Background components

There are two dominant background processes:

- **Generic background:** This component describes B^0 mesons decaying via a $b \rightarrow u, d, s, c$ transitions. The corresponding final state particles can then be mistaken as being part of a signal decay chain and hence be combined to form wrong B^0 candidates, i.e. B^0 mesons not representing the signal decay channel. As described in Section 4.4, the few actual signal events occurring in this sample have been deselected.
- **Continuum Background:** This background component emerges from processes, where the colliding positron and electron do not generate a $Y(4S)$ resonance, but decay into a $q\bar{q}$ pair immediately: $e^+e^- \rightarrow q\bar{q}$, where $q = u, d, s, c$ denotes the contributing quark flavors (Figure 3.1 shows the continuum component as the underlying events, which do not belong to the peaks of the resonances). Hence no real B^0 mesons are produced in these decays, however, the corresponding final state particles can be combined to form fake B^0 candidates, i.e. other particle types being mistaken for B^0 mesons.

4.2. Multivariate classification

When performing a classification of objects, e.g. tagging particle detector events as either signal or background, the first step is to describe these objects by a set of n discriminating variables, i.e. features with a high separation power between signal and background. This n -dimensional *feature vector* spans a n -dimensional space, where objects are represented as points. Classically, separation is obtained by simply applying independent cuts on the single dimensions (i.e. features). The cut-based method however does not take into account correlations between the features, which gave rise to a newer method called *multivariate classification* or *multivariate analysis* (MVA). This method is able to handle correlations by introducing *multivariate classifiers* - tools which represent advanced Machine Learning algorithms such as Neural Networks, Boosted Decision Trees (BDT) or Support Vector Machines. A classifier reduces the n -dimensional feature vector to a single, new feature, which can be cut on for optimal separation. In feature space a classification task can be described by finding the best separating hyperplane, which then represents the decision rule applied. This process constitutes an optimization task and is hence referred to as *training* of the classifier.

4.2.1. Non-linear problems

Since the classical cut-based method does not look at correlations of features it yields only linear decision rules, i.e. straight hyperplanes in feature space. A multivariate classifier on the other hand can use the feature correlations for building non-linear decisions, i.e. curved hyperplanes. Figure 4.1 shows the non-linear XOR-problem solved by the dimensional reduction of a multivariate classifier.

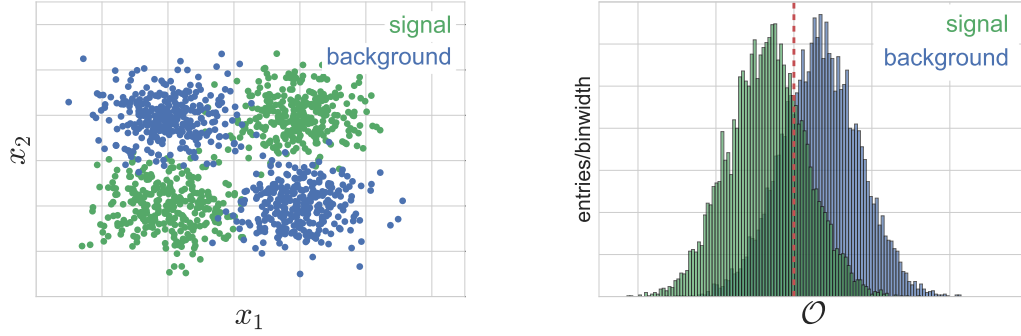


Figure 4.1.: *Left:* The non-linear *XOR-problem*. Labelled events in a two dimensional feature space $\vec{x} = (x_1, x_2)$ cannot be separated by a linear decision rule (i.e. line in the space). Hence this classification task can not be accomplished by a classical cut-based method. *Right:* Example for a classifier output. The new variable denotes a projection of x_1 and x_2 into the best separating dimension \mathcal{O} . The classification can now easily be performed by cutting on \mathcal{O} .

4.2.2. Supervised learning

The training method used throughout this analysis is called Supervised learning. Here a huge amount of training data (x, y) are fed into the classifier. These are labelled events, where x denotes the feature vector or *input* and y the provided class allocation (e.g. the colour of events in Figure 4.1) also referred to as *target*. The internal parameters are then adjusted towards optimal separation by minimizing an error function like the *mean-squared error function* [13]

$$Err = \frac{1}{n} \sum_{i=1}^n (\mathcal{O}(x_i) - y_i)^2, \quad (4.1)$$

where i denotes a sum over n events and \mathcal{O} denotes the output obtained from the classifier for an input x_i . Hence $\mathcal{O}(x_i)$ represents the suggested class allocation for event i . Intuitively small differences between the suggestions and the targets y_i yield correct classification.

4.2.3. Generalization

The outcome of the training is a set of internal parameters representing the decision rule of the classifier. They are fixed from then on in order to apply this optimized rule to new, unseen events. This second step of predicting an output using the decision rule learned during the training process is called *expert step* in the following. This crucial ability to learn the underlying patterns and correlations of a data set is called *generalization*. If a classifier does not generalize well, it may consist of too few internal parameters and is hence not able to handle the too complex problem. The case however, which is more complicated to get under control, is called *overfitting* and appears if the classifier consists of too many internal parameters. The training data might not be enough to optimize all of the parameters, and the classifier hence only learns the statistical fluctuations of this given data, or learns the training data by heart. The classification performance is then found to be worse on an unseen validation sample than on the training sample, which makes it possible to detect overfitting already during the training. Measures to avoid overfitting are, e.g. to enhance the data-to-parameters ratio, or penalize high and unstable values for parameters (*regularization*).

4.2.4. Evaluation

Due to the nature of the goal of a classification, its quality can be evaluated by calculating the efficiency ϵ and the purity p :

$$\epsilon = \frac{s_c}{s_{bc}}, \quad (4.2)$$

$$p = \frac{s_c}{b_c}. \quad (4.3)$$

Here s_{bc} and s_c denote the number of signal candidates before and after the applied cuts respectively and b_c the number of background candidates after the applied cuts.

In order to visualize the performance or properties of a classification output \mathcal{O} , different plots involving p and ϵ can be drawn:

- **ROC curve:** The Receiver Operating Characteristic (ROC) is obtained by plotting the purity p over the efficiency ϵ for different cuts on \mathcal{O} . Since each of these cuts denotes a point (working point) in the p - ϵ -space, by scanning over the whole range of \mathcal{O} , a curve can be drawn. Figure 4.2(*left*) shows an example of a ROC curve. As both values, p and ϵ , are tried to be maximized in a classification, the optimal working point of $p = 1, \epsilon = 1$ is positioned in the top-right corner of the plot. In other words, the better the classifier performance, the closer the corresponding ROC curve is found to this point. The task of deciding on a cut value on the output \mathcal{O} can then also be described as choosing a working point in the p - ϵ point along the ROC curve.
- **Purity per bin:** This plot is crucial in evaluating if the classifier output \mathcal{O} can be interpreted as a probability for each event, that it belongs to the target class at hand (e.g. is a signal event). In order for this interpretation to be valid, the value of each bin of \mathcal{O} needs to match the purity found in this bin: $p(\mathcal{O}) = \mathcal{O}$. If for instance one looks at the ground truth of all events in the bin containing $\mathcal{O} = 0.4$, 40% of them should belong to the target class ($p = 0.4$), in order for the following statement to be valid: 'An event with assigned $\mathcal{O} = 0.4$ belongs to the target class with a probability of 40%'. In order to visualize this property of an output, the *Purity per bin* can be plotted over the output. Figure 4.2(*right*) shows an example. Note that the rule $p(\mathcal{O}) = \mathcal{O}$ constitutes a diagonal line in the $p(\mathcal{O})$ - \mathcal{O} -space. Hence for the probability interpretation to be valid, all points in this plot should be placed on the diagonal line. This process can be described as *weighting* of a classifier output and is done during the training of a classifier and depends on the signal fraction sf given during the training:

$$\mathcal{O}_{weighted} = \frac{sf \cdot \mathcal{O}}{sf \cdot \mathcal{O} + (1 - sf) \cdot (1 - \mathcal{O})}. \quad (4.4)$$

If however, the signal fraction during the expert step differs from the one in the training ($sf_{train} \neq sf_{exp}$), the probability interpretation is not valid anymore, and the output needs to be *reweighted* according to sf_{exp} . The described procedure is used later throughout this analysis (see Section 6.2.4).

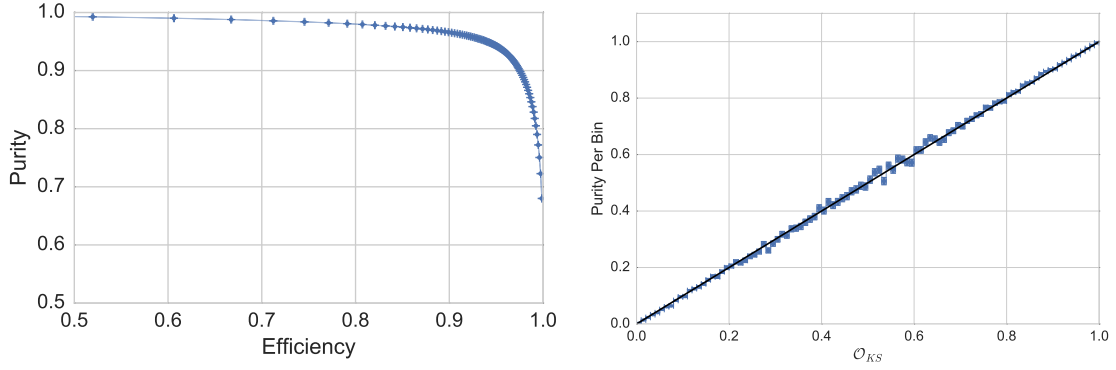


Figure 4.2.: *Left*: Example for an ROC plot. *Right*: Example for an Purity per bin plot.

4.2.5. Boosted Decision Tree

The classifier type used throughout this analysis is called Boosted Decision Tree (BDT). A trained BDT is a combination of if-then statements, which can be expressed as nodes in a tree-like structure. Each node represents a linear cut on a single variable in feature space. Events get classified by sorting them down the tree, i.e. enquiring the according statements, where the starting node is called *root node* and the ending nodes are called *leaf nodes*. The leaf nodes are labelled as either signal or background and hence represent the classification-decision for each event. Figure 4.3 shows an example for a trained BDT. The tree structure

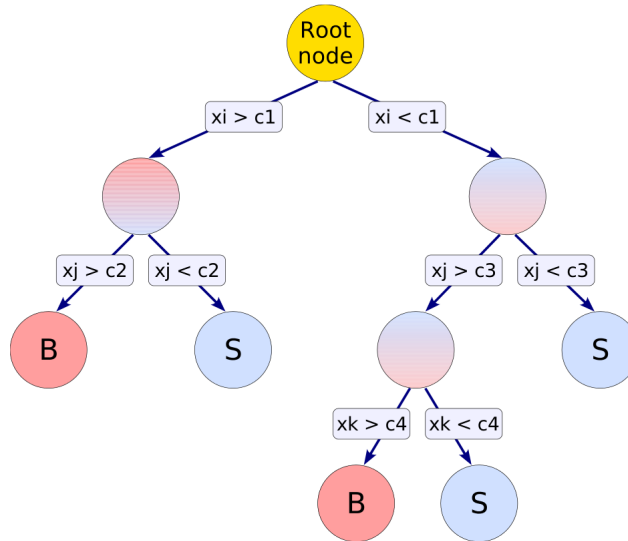


Figure 4.3.: Example for the structure of a BDT. From [14]

is built during the training process (or growing process), where for each new node all possible variables and cuts are tried for optimal separation according to Equation 4.1. This is done until a node fulfils the stopping criterion. One example for this criterion is the so called *Gini index* defined as

$$p \cdot (1 - p), \quad (4.5)$$

where p denotes the signal purity in the node [14]. Figure 4.4 shows how the Gini index constitutes a symmetric separation function, since a small value is obtained for a high purity of either signal or background. If the index goes below a defined threshold, the node gets declared

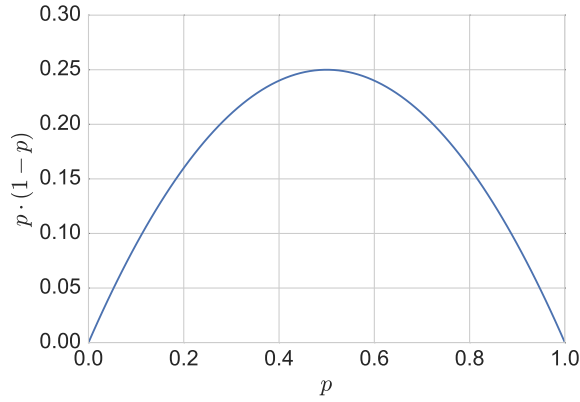


Figure 4.4.: Gini Index.

as a leaf node and the training in this branch is finished. In order for a decision tree to be a powerful classifier it would require to consist of many nodes and layers of nodes, hence a high number of parameters. As this kind of big structure favours overfitting, a different method called *adaptive boosting* is applied [15]. A boosted decision tree is a collection of many single decision trees or *weak learners*, where the final classification decision is made via a majority voting of the single classifiers. Apart from the feature of avoiding big tree structures, there is another enhancement to this method: By assigning weights to each event, a weak learner can be forced to focus on the events that have been misclassified by its predecessor. In doing so the easy events can be eliminated from the training by downgrading their weight in order to speed up the process and learn less statistical fluctuations of these events. Hence a well-regularized and powerful classifier can be trained.

4.2.6. MVA in Belle II

In particle physics simulated MC events are used as training data, in order to obtain a set of optimized selection cuts, which is later applied (generalized) on unlabelled data obtained from the detector. This method includes a high reliance on the correctness of the MC simulation. Also the process of learning of a classifier is not as transparent as the application of simple manual cuts, which causes MVA-methods to have the reputation of being a ‘black box’. Such downsides initially led to controversial discussions about the usage of MVA-methods in particle physics analyses, until significant improvements w.r.t the cut-based method were proven. In the Belle II software framework the TMVA-interface [14] makes it possible to easily choose from a set of classifier types and call them in a python steering file.

4.3. Maximum-Likelihood fit

4.3.1. Fit method

If a probability density $f(x|a)$ depends on an observable x and an unknown parameter a , this parameter can be estimated by performing a Maximum-Likelihood fit [16]. Measuring the observable n times yields a series of resulting values $x_1, x_2, x_3, \dots, x_n$, which shall be referred to as *data sample* in the following. During the fit-procedure the value \hat{a} is determined, for which the model $f(x|\hat{a})$ describes the data sample in the most consistent way. If the data is provided as single measured events the described procedure is called an *unbinned* Maximum Likelihood fit. If the data is put into a histogram first however, i.e. provided as bin-entry values, the procedure is referred to as a *binned* Maximum Likelihood fit.

When a CP-fit is performed, the fit model depends not only on one observable x , but other per-event information (e.g. the wrong tag information). Hence if one were to bin the data before fitting, this additional information would get lost or need to be projected into the remaining dimension. For this reason only *unbinned* fits are performed throughout this analysis.

For the optimization the Likelihood function shown in Equation 4.6 is defined. It represents the quality with which a model describes the given data sample, i.e. the *Likelihood* of describing the data sample correctly.

$$L(a) = f(x_1|a) \cdot f(x_2|a) \dots f(x_n|a) = \prod_{i=1}^n f(x_i|a) \quad (4.6)$$

Figure 4.5 shows an example, where $L(a)$ is calculated for two values of a on the same data sample. Since $f(x|a_1)$ describes the data sample much better than $f(x|a_2)$, higher function values are obtained yielding a higher Likelihood $L(a_1) > L(a_2)$. This example intends to visualize the main concept of a Maximum-Likelihood fit: By maximizing the Likelihood function $L(a)$ the best fitting parameter \hat{a} is determined. Since the fitting algorithm is a numeric

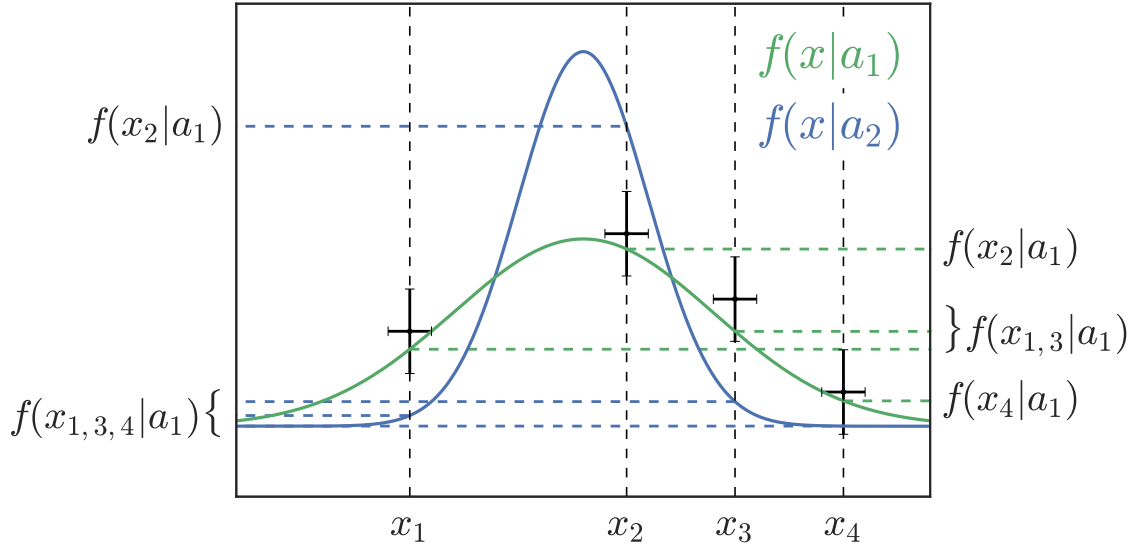


Figure 4.5.: Example of an evaluation of the Likelihood $L(a)$ for two different parameters a_1 and a_2 .

procedure, an instability emerges that has to be taken care of. Since f represents a probability its value is per definition $f < 1$. Hence $L(a)$ being the product of many probabilities denotes a very small number making it harder for the fitting algorithm to find a stable convergence. In order to avoid this effect, the negative-logarithmic-Likelihood $F(a)$ is given by

$$F(a) = -\ln(L(a)) = -\sum_{i=1}^n \ln(f(x_i|a)). \quad (4.7)$$

The applied logarithmic function transfers the product over the probabilities into a sum. This yields a fitting function with higher values and hence a much more stable optimization. By additionally flipping the sign of the likelihood function this procedure is transferred into a *minimization* problem. This is done only as a convention following the majority of optimization problems and algorithms.

4.3.2. Error estimation

As stated before, the statistical uncertainty of the fit $\sigma(\hat{a})$ is a crucial observable throughout this analysis. It can be obtained by making a simple assumption: In most cases $L(a)$ can be approximated by a Gauss-Function around its maximum \hat{a} . But how can the standard deviation be extracted from the actual fit-function $F(a)$? The Gauss-shape of $L(a)$ yields a parabolic shape of $F(a)$, which allows a Taylor-expansion around its minimum \hat{a} :

$$F(a) = F(\hat{a}) + \frac{1}{2} \frac{d^2 F}{da^2} (a - \hat{a})^2 + \dots, \quad (4.8)$$

where $\frac{dF}{da} = 0$ at $a = \hat{a}$ by definition. Further the Likelihood function can be written as

$$L(a) \cong \exp(-F(a)) = \text{const} \cdot \exp\left(-\frac{1}{2} \frac{d^2 F}{da^2} (a - \hat{a})^2\right) \quad (4.9)$$

$$\stackrel{!}{=} \text{const} \cdot \exp\left(-\frac{(a - \hat{a})^2}{2\sigma^2}\right). \quad (4.10)$$

Here the last expression represents the assumed Gauss shape of $L(a)$. From this equation an expression for the uncertainty σ can be extracted:

$$\sigma(\hat{a}) = \left(\frac{d^2 F}{da^2} \Big|_{\hat{a}} \right)^{-1/2}. \quad (4.11)$$

Hence the statistical uncertainty of the fit depends on the curvature of $F(a)$ only, which allows a fairly intuitive interpretation: For a low uncertainty, i.e. a steep parabolic shape of $F(a)$ a variation in a has a high impact on the value of $F(a)$, which in return constitutes a very stable minimum. In other words it is a sign for a precise and confident estimation of \hat{a} . Figure 4.6 shows how $\sigma(\hat{a})$ can be determined in multipliers of r directly from the plot of $F(a)$ over a . This is done by increasing the function value to $F(\hat{a} \pm r\sigma)$. If put into Equation 4.8 this yields

$$F(\hat{a} \pm r\sigma) = F(\hat{a}) + \frac{1}{2} \frac{d^2 F}{da^2} (r\sigma)^2 \stackrel{(4.11)}{=} F(\hat{a}) + \frac{1}{2} r^2, \quad (4.12)$$

where the distance d denotes the required increase of $F(a)$ for determining the uncertainty $r\sigma$ and is defined as

$$d = F(\hat{a} \pm r\sigma) - F(\hat{a}) = \frac{1}{2} r^2. \quad (4.13)$$

The uncertainty $r\sigma$ can then simply be read off from the x-axis of the plot. Throughout this analysis only the standard deviation 1σ ($r = 1$) is investigated, which yields $d = 0.5$.

4.4. Data simulation

In a real analysis simulated MC events are used for optimizing selection cuts, training multivariate classifiers, calibrating analysis tools, and many other tasks. Only when these preparatory steps are completed, i.e. the entire analysis process is set up, can the measured data be looked at. The idea of this concept called 'blind analysis' is to avoid a prior of the respective physicist performing the measurement. This prior may come from theoretical expectations concerning the measurement or results of previous measurements. Hence if the measured data is known while setting up an analysis, there is a high danger to influence this analysis in order to obtain the desired or expected results. The described bias may not even be intended by the researcher but emerge as a subconscious effect. Hence it is crucial for the analysis set up to not be changed as soon as the data is un-blinded. Since this project constitutes a simulated analysis, not only the calibration and training samples, but also the 'measured

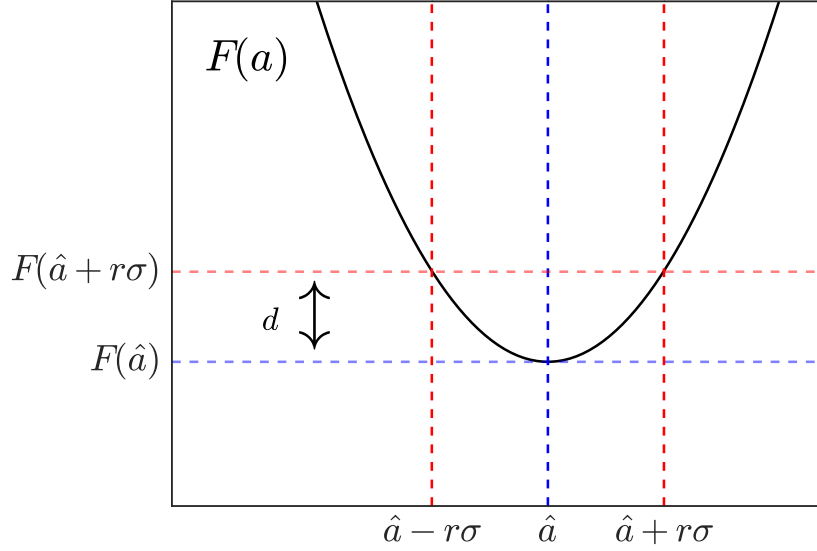


Figure 4.6.: Error estimation using the parabola of the negative logarithmic Likelihood $F(a)$.

data' are produced as simulated MC events. Hence it is important for this analysis to strictly distinguish between the two, i.e. make clear for each task which MC samples are used and what they are intended to represent. The crucial point is to treat the 'measured data' as such, i.e. not use any MC-information that would not be available for data coming from the detector.

The physical processes expected to happen in the detector are modelled throughout the MC simulation by different samples for generic and continuum events (see section 4.1). Note, that signal events are not taken from the centrally produced MC events (provided by the collaboration), since the inclusion of CP violation effects and branching fractions can not be confirmed on them. Instead 10^7 signal MC events have been generated separately for this analysis, while ensuring the correct production parameters. Furthermore, two categories of sample compositions shall be distinguished here:

- **high statistics samples:** When using MC events for calibration, training or pre-fits, i.e. steps of preparing a measurement, it is sensible to use high statistic samples. By doing so one uses as many MC events as available in order to minimize the statistical uncertainties on the determined parameters.
- **scaled samples:** On the other hand, when performing the actual measurement, the MC samples are supposed to imitate the data coming from the detector and need to be scaled to the according branching fractions given by the signal decay channel. As for the background samples only an equivalent of $1ab^{-1}$ of data is available, this amount is used for the high statistic applications as well as the measurement. Hence only the signal sample needs to be scaled down for the measurement. The according branching fraction(bf) for the decay $B^0 \rightarrow K_S^0 K_S^0 K_S^0$ is given by $bf = 3.0 \times 10^{-6}$ [5]. In order to determine the number of signal events expected to occur in $1ab^{-1}$ of data N_s , the following calculation needs to be done:

$$N_s = N_{BB} \cdot r_B \cdot bf = 3152. \quad (4.14)$$

Here N_{BB} denotes the total number of produced $B\bar{B}$ pairs. This number is taken from [17] and scaled to the according amount of data: $N_{BB} = 1081 \cdot 10^6$. As the signal

decay occurs only in neutral B mesons, the production ratio $r_B = 0.486$ [5] of neutral and charged B mesons is applied as a further scaling factor. Note, that N_s denotes the number of expected events *before* reconstruction, i.e. the ones occurring in the detector.

4.5. Flavor-Tagging

The process of Flavor Tagging is explained in Section 3.3. Throughout this analysis three BDTs are trained (Track level, Combiner lever, Particle level [19]) with three data sets of 5×10^5 signal events each. Figure 4.7 shows the output of the Flavor-Tagger, which is provided as the variable $qrCombined$. Here q denotes the Flavor Tagger decision, where $q = 1$ represents a B^0 -tag (the tag side B meson is a B^0), and $q = -1$ represents a \bar{B}^0 -tag (the tag side B meson is a \bar{B}^0). The so called 'dilution factor' r can be seen as the confidence with which the decision is made. Thus $r \approx 1$ describes a confident flavor-assignment, and for $r \approx 0$ no useful decision can be made.

4.5.1. Calibration

The fact that the Flavor-Tagging output $qrCombined$ provides a per-event information makes the validation of this MC based information on real data difficult. Hence, for a better comparison the original output $qrCombined$ needs to be binned. This calibration step is performed on a high-statistics signal MC sample, by dividing $qrCombined$ in 6 bins of r , where the bin edges are taken from the official Belle binning [18]. For each bin the wrong tag fraction w and its differences in flavors Δw are then determined as

$$w_+ = \frac{B_w^0}{B_w^0 + B_c^0}, \quad w_- = \frac{\bar{B}_w^0}{\bar{B}_w^0 + \bar{B}_c^0}, \quad (4.15)$$

$$w = \frac{w_+ + w_-}{2}, \quad \Delta w = \frac{w_+ - w_-}{2}, \quad (4.16)$$

where $B_w^0(B_c^0)$ are the number of wrong (correct) assigned events in each bin [8]. Table 4.1 shows the described binning as well as the results for w and Δw .

bin	interval in r	w	Δw
1	[0 – 0.125]	0.471	–0.0440
2	[0.125 – 0.25]	0.411	–0.0089
3	[0.25 – 0.5]	0.319	–0.0379
4	[0.5 – 0.625]	0.230	0.0512
5	[0.625 – 0.75]	0.171	–0.0080
6	[0.75 – 0.875]	0.105	–0.0399
7	[0.875 – 1]	0.027	0.00057

Table 4.1.: Determined values for the wrong tag fraction w and the wrongtag fraction difference Δw in bins of r .

4.5.2. Assignment of tagging information

In this subsequent step each event ('real data') gets assigned a new binned (calibrated) tagging information

$$qr_{new} = q(1 - 2w(qr_{old})). \quad (4.17)$$

Here qr_{old} denotes the original per-event flavor tagging output, which is then used to determine the associated wrong tag fraction w according to the MC calibration in Table 4.1. Note, that qr_{new} now denotes binned information rather than per-event information and can hence be validated on control samples. Also the mistag fraction difference information Δw is assigned to each event.

4.5.3. Flavor-Tagging quality

For evaluation of the Flavor-Tagging quality the *effective tagging efficiency* is defined as

$$Q = \sum_i^n \epsilon_i (1 - 2w_i)^2. \quad (4.18)$$

Here ϵ_i denotes the *tagging efficiency* in each bin. Since the number of events, which have not been able to be tagged by the Flavor tagger are negligible ($\epsilon \approx 1$), ϵ_i simply corresponds to the fraction of events per bin. In this analysis the value $Q = 32.6\%$ was determined, which compares well to the results of previous Flavor-Tagging studies at Belle II [19]. Note that despite the fact that the decays of the two B mesons in one event are completely uncorrelated, the cleanliness of the B_{sig} decay does have an impact on the quality of flavor tagging decision on B_{tag} . This means the data samples with tight selection cuts on the reconstruction quality, etc., are more likely to represent overall cleaner events providing an easier setup for the flavor tagging algorithm to work. This causes the value of Q to be highly dependent on the pre-cuts applied to the sample, that is used for its determination. In conclusion, Q should only serve as a benchmark variable between analyses, flavor tagging techniques, and even experiments, if it is provided with the corresponding information about the properties of its determination.

4.6. Analysis setup

Figure 3.2 was used in Section 3.3 for visualizing the setup of a time-dependent CP violation measurement. It also shows a sketch of the decay chain of the signal decay process. Thus in order to reconstruct B^0 mesons, first K_S^0 mesons have to be reconstructed from combinations of two tracks of π^+ and π^- . Subsequently three K_S^0 candidates in one event can be combined in order to form a B^0 candidate. Previous analyses in this channel applied a cut-based K_S^0 selection directly on the corresponding candidate list. The MVA-based equivalent to this method would be to train a BDT using several kinematic variables of the K_S^0 mesons (KSBDT). The KSBDT's output \mathcal{O}_{KS} could then be interpreted as individual score for each candidate describing the likelihood to represent a correctly reconstructed K_S^0 . The selection could then be performed by cutting on the K_S^0 candidate list according to this score. The cutting point would usually be chosen at the maximum of the *Figure of Merit*(FOM):

$$FOM = \frac{s}{\sqrt{s+b}}. \quad (4.19)$$

Here $s(b)$ denotes the number of signal(background) candidates in the list after the corresponding cut. The FOM denotes a modification of the signal to background ratio, that evaluates efficiency and purity of a selection at the same time. In order to distinguish this FOM from others, it is referred to as FOM_{KS} in the following.

Since a B^0 candidate has three associated K_S^0 daughters, correlations between the daughters would not be taken into account by cutting on each daughter individually. These correlations however, are crucial for the B^0 selection in this decay channel. Hence this analysis takes a different approach: Before applying any K_S^0 selection, the B^0 mesons are reconstructed. Subsequently, the three output scores \mathcal{O}_{KS} of the three daughters associated to a B^0 candidate

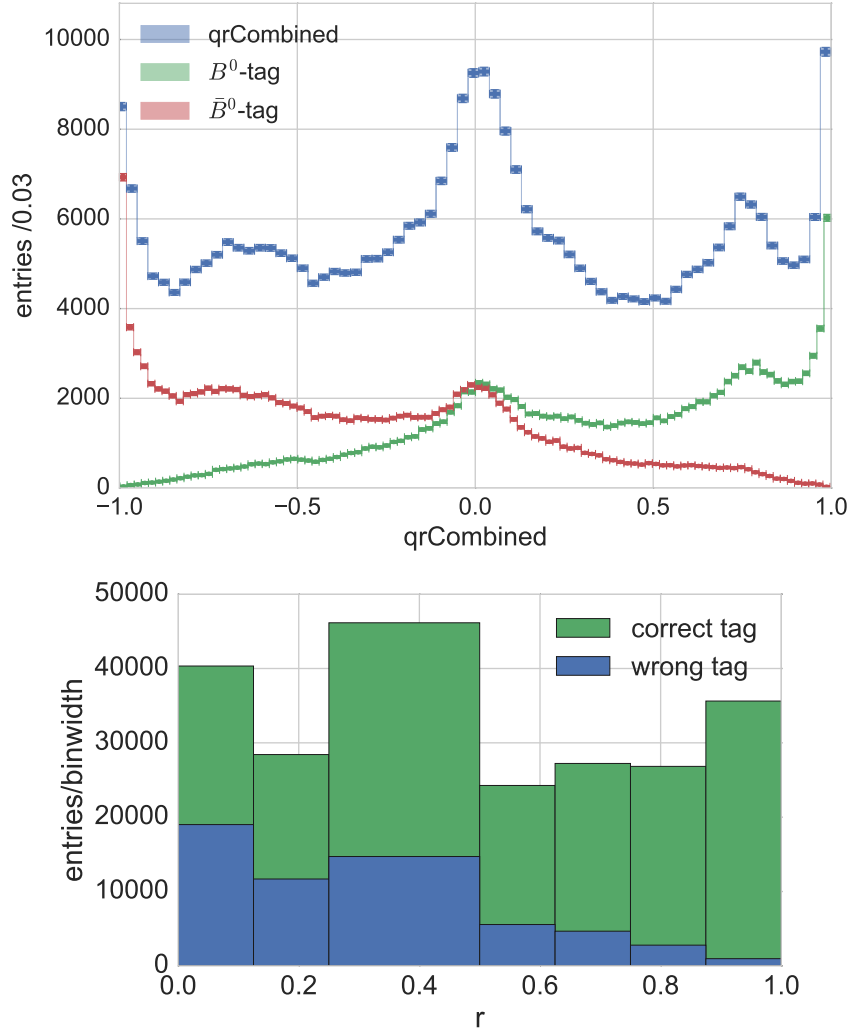


Figure 4.7.: *Top*: Flavor-tagger output for the signal MC sample. The slopes of the B^0 -tag distribution (green) and \bar{B}^0 -tag distribution (red) towards their correct values shows the separation power of the Flavor Tagger. The peak at 0 shows undecided events. *Bottom*: Wrong tag fractions w in bins or r .

are taken as input variables for a new B^0 -BDT (BBDT). The output of the BBDT, \mathcal{O}_B , can then be used as an overall correlated K_S^0 selection score. Here again the question emerges where to cut on the output \mathcal{O}_B for an optimal selection. While the Figure of Merit FOM_B of the output \mathcal{O}_B constitutes one possible choice, one should keep in mind that the goal of a good selection is to enable a significant and unbiased measurement. Hence a new cut-criteria is introduced: The optimal cut value of \mathcal{O}_B is that which minimizes the uncertainty of the time-dependent CP violation parameters S_f and A_f , i.e. $\sigma(S_f)$ and $\sigma(A_f)$. This minimum value $\mathcal{O}_{B,min}$ is determined by performing the unbinned Maximum Likelihood fit (see Section 4.3) at 55 values of \mathcal{O}_B , which denote edges of equally distributed intervals. The minimum cut value $\mathcal{O}_{B,min}$ is then fed back to serve as the selection cut for the actual measurement. This method shall be evaluated in comparison to the FOM_B -based selection cut and the FOM_{KS} -based selection cut.

After the selection is done, the time-dependent CP violation measurement is performed. In order for the parameters S_f and A_f to be the only two floating parameters of the fit, some

preparatory steps have to be performed in advance: Besides determining other free parameters of the fit model (shape parameters) in so called *pre-fits*, another BDT is trained using kinematic variables of the B^0 mesons for separation. By doing so, a per-event signal probability score f_{sig} is generated and used later in the fit model (see Equation 7.11) for weighting the B^0 candidates. The corresponding BDT is referred to as FSBDT (see 6.2.4 for further explanation).

The entire analysis setup described above is shown in Figure 4.8. It also displays how the data-flow in a real analysis would look, when using this setup. While one set of MC events is used for calibration, training of the BDTs and optimization of the K_S^0 selection cut, the actual data coming from the detector is directed only through the expert steps of the BDTs and the final fit performing the determination of only S_f and A_f . The described separation of data flows is being emphasized at this point, because in the latter no MC information must be used. This fact might sound obvious, but is still important to remember when using MC simulated data for both, the calibration and the imitation of real data (see Section 4.4).

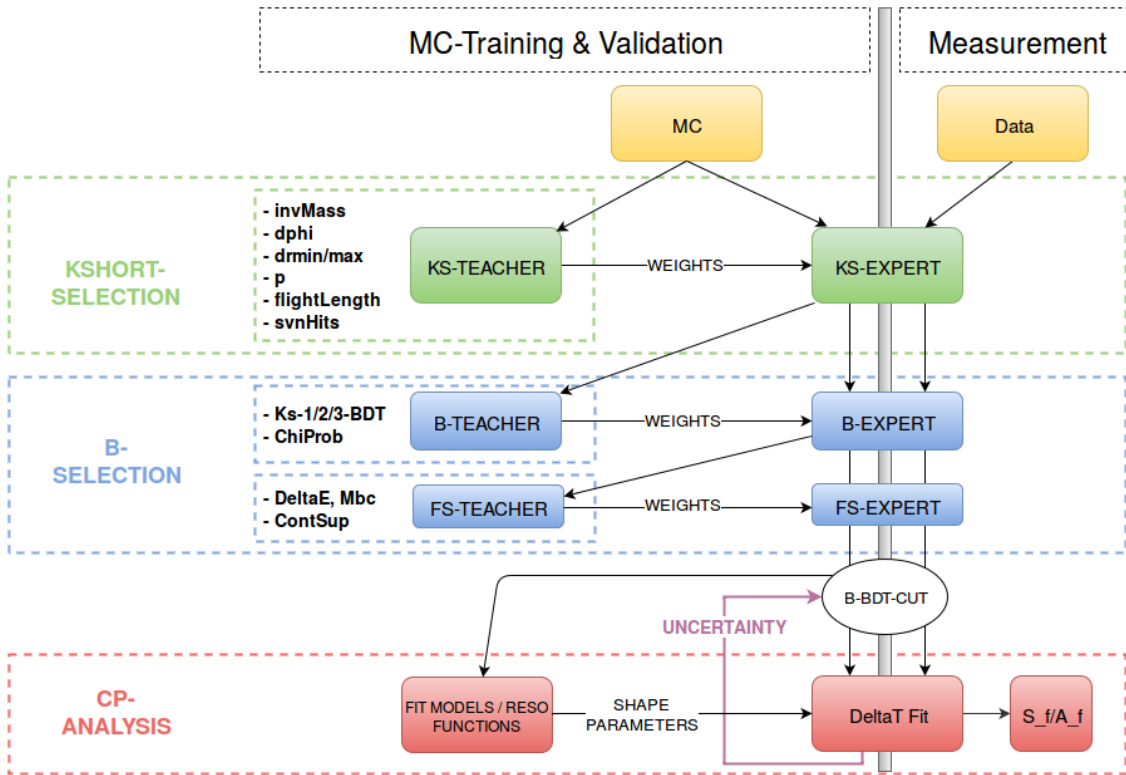


Figure 4.8.: Analysis work-flow. The term *weights* refers to the tree structure and cuts developed during the training process, which are subsequently provided for the expert step.

4.7. Analysis Software

4.7.1. BASF2

The software framework for the Belle II Experiment (BASF2) is currently being written from scratch in order to obtain better performance and usability with respect to Belle [12]. The framework includes several steps of data processing provided in the form of packages:

- MC event generation

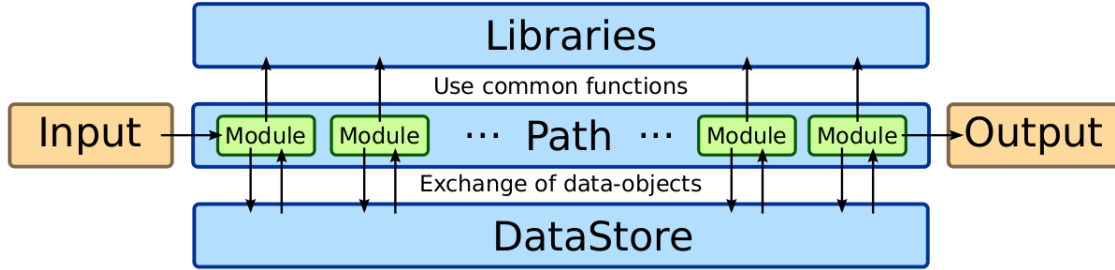


Figure 4.9.: Structure of a process path in BASF2. From [12].

- Detector simulation
- Track reconstruction
- Physics analysis

The framework is written in C++ and interfaced with python as a steering language. The concept of BASF2 is to allow for the user to process large amounts of data in a flexible way. Hence small processing blocks, called *modules*, are introduced, containing configurable tasks, which they process in an event-by-event manner. The user is able to place several modules together, thereby creating his individual processing *path*. In each package the modules work on top of libraries. Information to be shared amongst the modules, e.g. the required input or created output, is stored as *dataobjects* in the *DataStore*. The output after reconstruction is written out in form of a ROOT Tree [20]. Figure 4.9 shows the work-flow of a processing path in BASF2.

4.7.2. Fit-Framework

After obtaining the reconstructed B^0 candidate list, including all BDT scores needed for the selection, they are stored locally in order to perform the \mathcal{O}_B -cut optimization and the time-dependent CP fit. Therefore a fit-framework has been written from scratch as part of this thesis using python. This framework implements all necessary steps to construct the fit model, e.g., performing normalizations of all probability density functions (PDFs). This procedure was chosen in contrast to previous analyses, which simply relied on predefined ROOT functions, for two reasons: First, the ROOT package developed for time-dependent CP fits (R00RARFIT) exhibits a very untransparent design, with several layers of nested code and little documentation. Additionally, this thesis represents the first time-dependent study among the Belle II collaboration and no certain fit-framework has been agreed on yet by the working group. Hence the usage of the ROOT fit-framework was feasible, but the corresponding gain of expertise expected to be smaller as opposed to a self-written framework. Secondly, software tools like the PANDAS library [21] and the JUPYTER Notebook [22] have been presented among the Belle II collaboration, as being part of a new movement in analysis software design. Implementing a fit-framework using these tools hence seemed to be a good investment for future projects in the experiment.

The PANDAS library provides the DataFrame object, which allows for fast and intuitive data-analysis processes. Hence this analysis prefers to handle the data structured in DataFrames rather than ROOT Trees. The JUPYTER Notebook can be used as an editor for Ipython, with the speciality of loading all input data into the cache memory. Therefore the user is able to compile different parts of the code individually without having to reload the input data from a hard drive. The editor design of the JUPYTER Notebook adapts to this functionality

by providing a cell structure for the code to be written in. Each cell denotes one element of code to be compiled at once. Hence this tool allows for a fast and nicely structured work-flow, when performing a physics analysis.

5. Vertex Reconstruction

When combining decay products to form particle candidates, the estimated position of this candidate (*vertex*) as well as its momentum and invariant mass are determined by performing a kinematic fit [23]. Thereby a χ^2 term is minimized, which contains the 4-momentum information of the corresponding decay products. The resulting χ^2 value can then be used to construct a probability *chiP* for estimating the quality of the fit [23] and hence as an exclusion criteria during the selection process.

This chapter describes the vertex reconstruction for K_S^0 and B^0 mesons. Note that, as this thesis only covers Kaons decaying to charged pions ($K^0 \rightarrow \pi^+\pi^-$), only the fit regarding charged tracks are described throughout this section.

5.1. V0 Finder

As previously mentioned, K_S^0 candidates are built by combining two oppositely charged tracks, which are assumed to be pions. The trajectory of a track can be described by a helix and hence can be parametrized by the 5 helix parameters [23]. In order to save storage space, only the helix parameters are stored by the tracking algorithm, whereas the individual hits belonging to a trajectory are thrown away. When travelling through the detector, particles suffer an energy loss $\propto \frac{dE}{dx}$ due to material effects. This is accounted for by extending all track fits to the IP, while applying an energy correction to the stored parameters, before storing the information as a helix. Hereby all fitted tracks are expected to originate inside the beampipe. A problem emerges, if a particle decays outside the beampipe, since the generally applied correction overestimates the distance covered by the particle and hence the loss of energy it experienced. As the estimated energy loss gets added to the 4-momenta of the tracks by the correction, their associated energy is expected to suffer a bias due to the overestimation. This issue is handled by a recently implemented extension to the tracking algorithm, the *V0 Finder* module. When using the V0 Finder, a vertex fit to the two tracks is performed before the parametrization. With the information about the vertex position being provided now by the V0 Finder, the energy correction can be applied only until the *point of closest approach* to this position without overestimating the covered distance of the tracks. Subsequently the helix is extended to the IP, thereby determining the helix parameters. As the 4-momenta of the tracks are being extracted from the helix parameters, the ones provided by the V0 Finder are expected to contain no bias as opposed to the ones with overestimated energy correction. In order to compare the two methods described, the classical (biased) one is referred to as the *reconstruct*

decay method (RD) in the following. Figure 5.1 shows a sketch of the described process. When performing the vertex reconstruction of the K_S^0 , its invariant mass is determined using

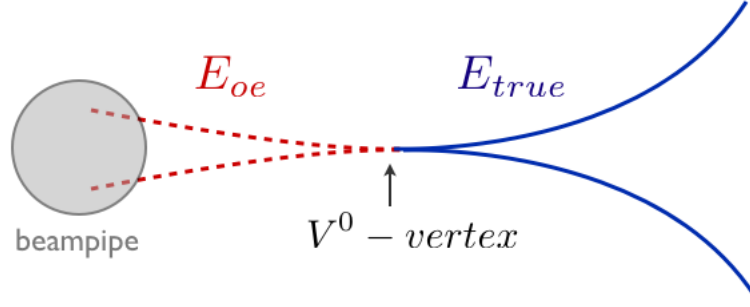


Figure 5.1.: Sketch of a V0 track fitting process. The red dashed line denotes the distance, which is not covered by the particle and hence causes overestimation of energy E_{oe} . The blue line denotes the helices fit to the vertex point, like its done by the V0 Finder Module, thereby allowing for the true correction to be made (E_{true}).

the 4-momentum information of the two associated daughter tracks:

$$M^2 = (E_1 + E_2)^2 - \|\mathbf{p}_1 + \mathbf{p}_2\|^2 = m_1^2 + m_2^2 + 2(E_1 E_2 - \mathbf{p}_1 \cdot \mathbf{p}_2). \quad (5.1)$$

When plotting the distribution of the invariant mass, a shift towards positive masses is expected for the biased RD tracks. Figure 5.2 shows the distributions for both methods. As expected the V0 Finder provides an unbiased peak at the nominal K_S^0 mass $m_{KS} = 0.498 \pm 0.024 \text{ GeV}$, whereas the RD causes the K_S^0 candidates to exhibit a shift towards positive masses.

However, investigating the V0 Finder module, a loss in efficiency has been observed. Subsequently, several 'hidden' cuts in the V0 Finder module (at track-reconstruction stage) as well as in the corresponding analysis steering files have been detected as default settings:

- **track-reconstruction stage**
 - $\chi^2 < 50$ cut on the V0 vertex quality.
 - rejectCandidate function, intended to reject curling tracks.
 - 60 MeV window on the invariant mass for V0s inside the beampipe ($r < 1 \text{ cm}$).
- **analysis stage**
 - 400 MeV mass window before vertex fit.
 - Vertex fit confidence level > 0 .
 - 40 MeV mass window on the invariant mass after the (analysis-)vertex fit.

The cuts at the track-reconstruction stage present a problem, since all MC events are provided after track-reconstruction. Therefore no possibilities are given to the user in order to modify these cuts (the cuts at analysis stage, however, have been removed throughout this analysis). Hence the efficiency loss can only be overcome by combining the V0 Finder information with the RD information. This is accomplished by the *V0 Combiner* module, which has been developed as part of this thesis. When using the V0 Combiner, all K_S^0 candidates found by the V0 Finder are saved into a separate list. Subsequently, all K_S^0 from RD are checked to be copies of the V0 Finder's candidates, i.e. constitute combinations of the same tracks. If no overlap is found for an RD candidate, the V0 Combiner adds it to the list. By doing so the

high quality candidates from the V0 Finder are prioritized but filled up with RD candidates in order to account for the efficiency loss.

As the V0 Finder module has not been used in an analysis before, this thesis aims to evaluate its performance. Therefore several parts of the measurement throughout this study is performed using the three methods (RD, V0 Finder and V0 Combiner) and comparing the results subsequently.

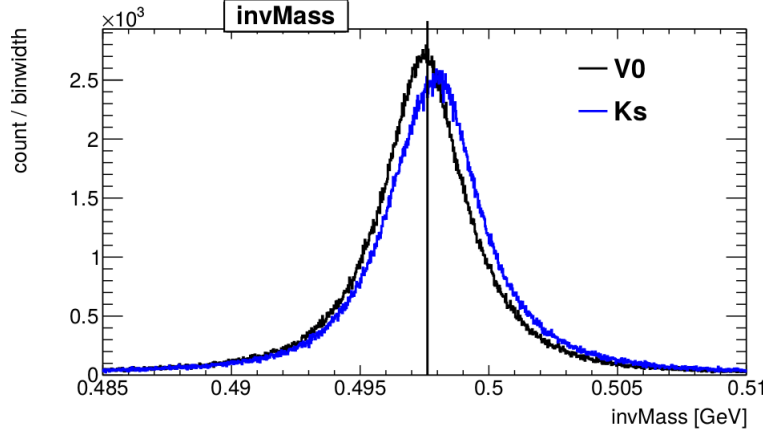


Figure 5.2.: m_{KS} distribution for events reconstructed by the V0 Finder (V0, *black*) and the RD method respectively (Ks, *blue*).

5.2. B^0 vertex fit

As illustrated in Figure 3.2, for a time-dependent CP measurement to be performed, the B^0 meson decaying via $B^0 \rightarrow K_S^0 K_S^0 K_S^0$, i.e. the *signal-side* B^0 , needs to be reconstructed as well as the accompanying B^0 meson used for flavor tagging (*tag-side*). Both are reconstructed via vertex fits of their associated decay products, which are described in this section.

5.2.1. Signal side

For the signal-side vertex fit the information of all three associated K_S^0 daughters is used. The corresponding trajectories are obtained from extensions of the momentum vectors of the daughters. Since the B^0 mesons undergoes a boost in z-direction (see Section 3.3), it is likely to decay inside an area alongside the z-axis around the IP, which exhibits the shape of a tube. Hence an additional constrain on the vertex fit can be applied for better resolution. This *IP Tube Constraint* is implemented by extending the χ^2 term, which is being minimized throughout the fit, by an extra addend:

$$\chi_{IP}^2 = \left(\frac{x}{\sigma_x} \right)^2 + \left(\frac{y}{\sigma_y} \right)^2. \quad (5.2)$$

Here σ_x (σ_y) denotes the extent of the tube in x-direction(y-direction).

5.2.2. Tag side

The vertex of the tag B^0 is determined by an inclusive kinematic fit, since no explicit knowledge about its decay is provided. Therefore, all tracks which have are not part of the signal side reconstruction are fitted together forming the *tag vertex*. Since not all remaining tracks have to

be direct B^0 decay products, but could as well originate from the decay of long-living secondary particles (e.g. D mesons), the track with the highest impact on the χ^2 value is removed, before performing the fit again. This procedure is repeated until the χ^2 value reduces below a chosen threshold [8]. Note, that the IP Tube Constraint is applied on the tag-side as well.

6. Selection

In this chapter the variables used for selection throughout the analysis are presented. Note that some of the variables are given in the form of classifier outputs (see Section 4.6). In these cases the expert output, i.e. the output for the scaled sample relevant for the measurement (see Figure 4.8) are shown. In order for the visualization of the distributions to not be dependent on the low statistics of the scaled sample, the averaged MC sample is constructed by averaging over 1000 scaled signal samples (for lack of more MC events this procedure cannot be conducted for the background samples). Subsequently, the pre-cuts, i.e. cuts applied before the actual measurement in order to speed up the offline analysis processes, on these variables are presented and listed. Hereby this chapter intends to successively describe the composition of the final B^0 candidate list, as it is used in the CP fit.

6.1. K_S^0 training

As described in Section 5.1, the K_S^0 candidates are formed by combining two associated charged pion tracks and performing a vertex fit. Hence the separating variables for K_S^0 selection constitute either information about the kinematics of the daughter tracks or about properties of the determined vertex. Only K_S^0 , which represent the corresponding decay products of $B^0 \rightarrow K_S^0 K_S^0 K_S^0$, are defined as signal throughout this training, since the KSBDT is supposed to adapt to the signal decay and not to select K_S^0 in general. The set of variables used throughout this analysis is based on previous K_S^0 selection studies at Belle [24], but has been enhanced by new variables or combinations of variables:

- **cos(dphi):** This is the cosine of the azimuthal angle between the *momentum vector* and the *decay vertex vector* of a K_S^0 candidate. The momentum vector denotes a straight line drawn from the IP towards the determined vertex position, whereas the decay vertex vector is the result of combining the two 4-momenta of the associated daughter tracks. Hence for signal events the two vectors are expected to be parallel to each other ($\cos(dphi) \approx 0$).
- **drmin(drmax):** This is the shorter(longer) distance from the IP to the two tracks in x-y plane.
- **significance of flight length:** The flight length is given by the x-y position r of the determined vertex. Due to the different detector layers, the resolution of r is a function of r and needs to be taken into account. This is done by relating r to the correlation

matrix V_{ij} associated to the corresponding vertex fit:

$$s = \frac{r}{\sqrt{\sum_{ij} \frac{\partial r}{\partial x_i} V_{ij} \frac{\partial r}{\partial x_j}}} = \frac{r^2}{\sqrt{\vec{x} V \vec{x}}} r = \sqrt{\vec{x} \cdot \vec{x}}. \quad (6.1)$$

Since K_S^0 have a lifetime of $\tau_{KS} = 0.895 \times 10^{-10} s$ at a measurable scale, the s -distribution of signal candidates is expected to exhibit a longer tail corresponding to secondary vertices outside the beampipe as opposed to background candidates, which mostly decay inside the beampipe.

- **chiP**: This is the χ^2 probability taken from the vertex fit. As no pre-cuts on this value have been applied during the fit, background candidates are expected to exhibit a peak at $chiP = 0$ for low-quality vertex fits.
- **M**: This is the invariant mass of the K_S^0 candidate determined from the two 4-vectors of the associated daughter tracks (see Equation 5.1). Signal candidates are expected to exhibit a peak at the nominal K_S^0 mass $m_{KS} = 0.498 \pm 0.024 \text{ GeV}$ [5].
- **p**: This is the momentum of the K_S^0 candidates determined by the addition of the two 4-Vectors of the associated daughter tracks. Signal candidates are expected to have high momentum, since they represent direct decay products of a B^0 meson (The large energy of the B^0 is only shared by three decay products).
- **svnHits**: This variable counts the number of hits in the SVN (see Section 3.2) of the two associated daughter tracks. Since K_S^0 can decay outside the beampipe and partly inside the SVN, the corresponding daughter tracks may not travel through the entire SVN and hence cause less hits.

The KSBBDT training output including evaluation and correlation plots, as well as plots of all input variables can be found in Appendix B.1.

6.2. B^0 selection

6.2.1. BBDT output

As the purpose of the BBDT is explained in detail in Section 4.6, here only the input variables shall be listed:

- **Ks1**: KSBBDT output \mathcal{O}_{KS} of the *first* associated daughter to the B^0 candidate.
- **Ks2**: KSBBDT output \mathcal{O}_{KS} of the *second* associated daughter to the B^0 candidate.
- **Ks3**: KSBBDT output \mathcal{O}_{KS} of the *third* associated daughter to the B^0 candidate.
- **chiPB**: χ^2 -probability of the B^0 vertex fit (see Chapter 5).

The BBDT training output including evaluation and correlation plots, as well as plots of all input variables can be found in Appendix B.2. The expert output is shown in Figure 6.1. Since the output \mathcal{O}_B denotes the variable the optimization is performed on (see Section 7.4), only a loose pre-cut $\mathcal{O}_B > 0.1$ is applied.

6.2.2. ΔE and Mbc

ΔE and Mbc are two important variables for describing kinematic features of B^0 mesons. Their definition as well as their use for the selection process shall be explained here:

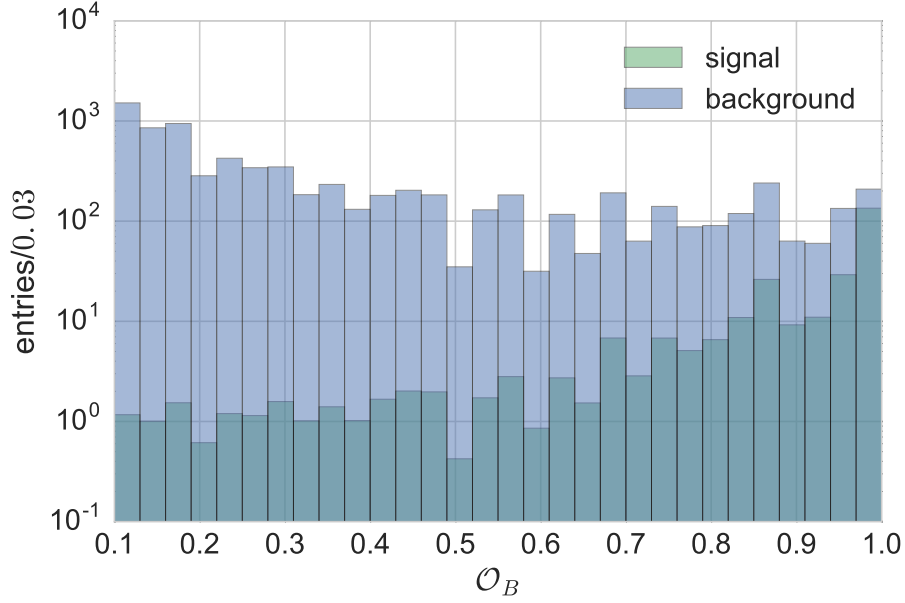


Figure 6.1.: BBDT expert output for an averaged sample after application of all pre-cuts in logarithmic scaling. The background includes the generic and the continuum components.

- ΔE : The difference of nominal and measured energy is defined as

$$\Delta E = E_B^{CMS} - E_{beam}^{CMS}. \quad (6.2)$$

Here E_{beam}^{CMS} is half the beam energy in the centre-of-mass frame. In the signal case $e^+ + e^- \rightarrow Y(4S) \rightarrow B^0 + \bar{B}^0$ the beam energy is (almost) completely transformed into the energy of the $B^0 \bar{B}^0$ pair. This effect is measurable with a resolution of some 100 MeV. Hence the distribution of signal events is expected to exhibit a peak at $\Delta E = 0$ [25].

- Mbc : The beam energy constraint mass is defined as

$$Mbc = \sqrt{(E_{beam}^{CMS})^2 - (\vec{p}_B^{CMS})^2}, \quad (6.3)$$

where \vec{p}_B^{CMS} is the momentum vector in the centre-of-mass frame. The nominal value for a B^0 meson is $Mbc = 5.2795 \text{ GeV}$.

Figure 6.2 shows the obtained distribution for ΔE and Mbc . Instead of using the variables for selection cuts, they are taken as input variables for the FSBBDT training, which in return contributes crucial information to the CP fit. Hence only loose pre-cuts are applied on the variables:

$$|\Delta E| < 0.075 \text{ GeV}, \quad (6.4)$$

$$5.27 \text{ GeV} < Mbc < 5.29 \text{ GeV}. \quad (6.5)$$

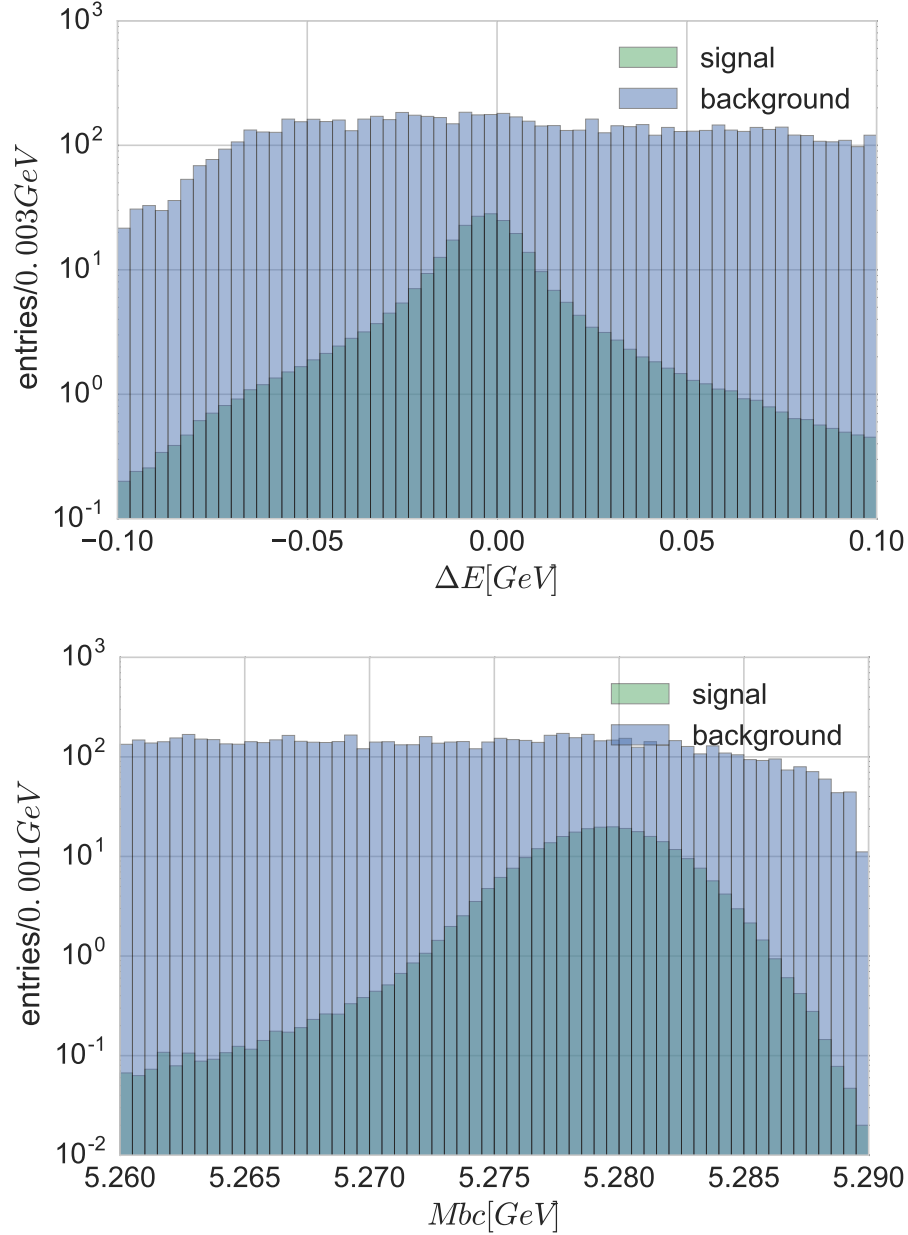


Figure 6.2.: ΔE and Mbc distributions of averaged samples after application of all pre-cuts in logarithmic scaling. The background includes the generic and the continuum components.

6.2.3. Continuum suppression

Continuum suppression is a selection process, which uses a set of separating variables in order to suppress continuum background (see Section 4.1). These variables are constructed by taking advantage of geometric differences between the decay of B^0 mesons and the prompt decay of continuum events: As the B^0 pair is the product of the $Y(4S)$ resonance decay, it is produced approximately at rest, causing the further decay to occur isotropically. The $q\bar{q}$ pair of a continuum event on the other hand, is generated with high transverse momenta in opposite directions. Due to momentum conservation these momenta are transferred to their corresponding hadronization products, thereby forming a jet-like decay structure in the centre-of-mass frame. Figure 6.3 shows the topological difference, from which several shape variables

can be deduced providing a high separation power between real B^0 events and continuum events. For an enhanced selection performance these discriminating variables are used as input variables for a continuum suppression BDT training (CSBDT). Figure 6.4 shows the expert output of the CSBDT. The corresponding training output including evaluation and correlation plots, as well as plots of all input variables can be found in Appendix B.3. Since the output variable \mathcal{O}_{CS} is used as input for the FSBDT training, as done for ΔE and Mbc , only a loose pre-cut $\mathcal{O}_{cs} > 0.1$ is applied.

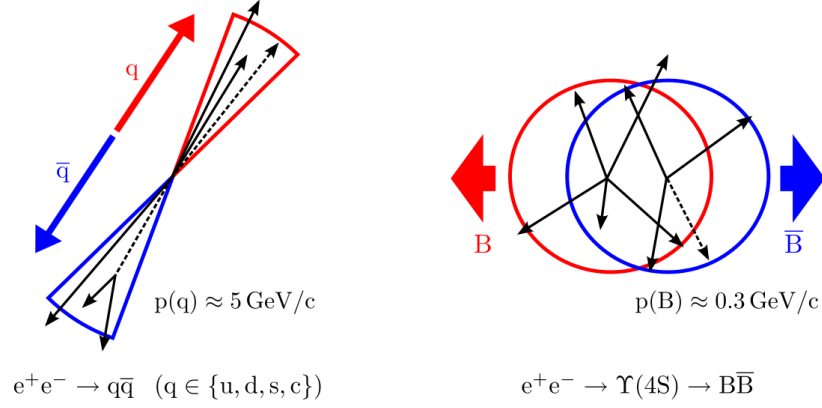


Figure 6.3.: Topological characteristics of continuum events: The $q\bar{q}$ pair on the left hand side comes with back-to-back momentum forming a jet-like event structure. The B^0 meson pair on the right hand side has negligible momentum forming a spherical event shape [8]

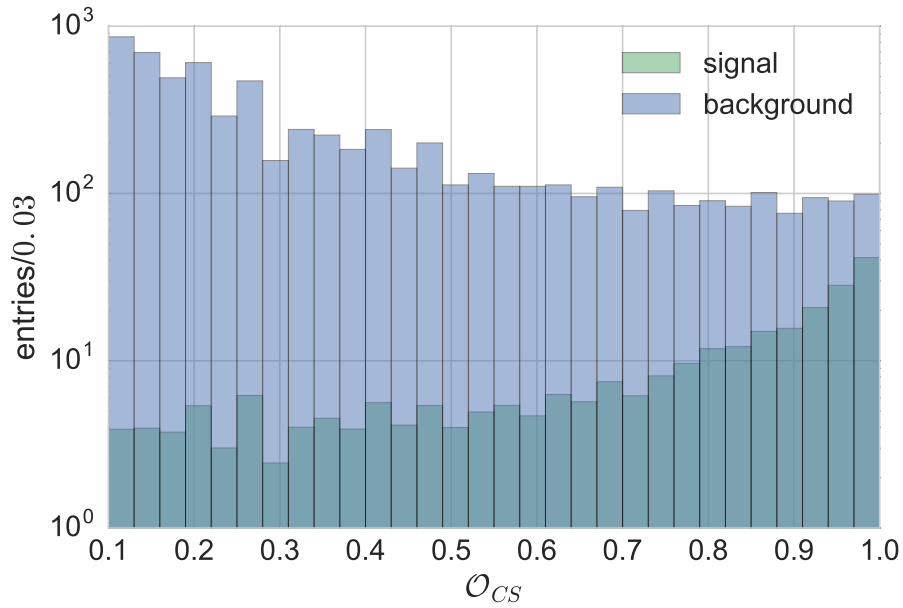


Figure 6.4.: CSBDT expert output for an averaged sample after application of all pre-cuts in logarithmic scaling. Note that background here refers only to the continuum component of the background. Hence signal refers to all non-continuum events.

6.2.4. FSBDT output

As described in Section 4.6, the purpose of the FSBDT is to provide an output \mathcal{O}_{FS} , that can be interpreted as a per-event signal probability. The following variables are taken as input for the FSBDT training (all input variables are described in detail in previous sections of this chapter):

- Mbc
- ΔE
- \mathcal{O}_{CS}
- $|qrCombined|$

The FSBDT training output, including evaluation and correlation plots, as well as plots of all input variables, can be found in Appendix B.4. The expert output is shown in Figure 6.5. In previous time-dependent CP violation measurements, the signal probability f_{sig} is obtained by performing a three-dimensional fit to the PDFs of $Mbc, \Delta E$, and \mathcal{O}_{CS} [17]. The absolute value of the flavor tagging information $|qrCombined|$ is then used to further divide the PDFs in seven bins of r (in Section 4.5). Correlations between the variables however, are not taken into account this way (see Section 4.2). The FSBDT on the other hand, by its nature is able to consider these correlations and is hence developed as a new method throughout this analysis. Additionally no dividing in bins of r is necessary, since the flavor tagging information can simply be added to the training as an input variable. However, one has to be careful interpreting its output directly as a probability (see Section 4.2), since the output-weighting was performed during the training with $sf_{train} = 0.5$ according to Equation 4.4. This signal fraction differs from the one given during the expert step by the scaled sample $sf_{exp} = 0.208$ (see Equation 8.7), when simulating the actual measurement. One possible solution to this problem is to just reweight the output \mathcal{O}_{FS} according to sf_{exp} using Equation 4.4. This signal fraction however, denotes a MC-information not known on real data and would hence have to be determined via an extra signal yield fit. Additionally, not only the change in signal to background ratio needs to be taken into account, but also the ratios of the different background samples to each other. As all ratios together constitute a high amount of MC information to be determined from extra fits, a different method for reweighting \mathcal{O}_{FS} is chosen here: The \mathcal{O}_{FS} distribution is histogrammed in 30 bins (see Figure 6.5) and the purity in each of these bins is subsequently determined. This is done on the averaged MC sample in order to lower the statistical dependency. In the next step, each 'real data' event gets assigned a probability f_{sig} according to the previously calibrated purity according to the corresponding bin of \mathcal{O}_{FS} . This can be thought of as a method analogous to the Flavor Tagging calibration in Section 4.5. Figure 6.6 shows the obtained purities in bins of $\mathcal{O}_{FS,original}$ and $\mathcal{O}_{FS,reweighted}$. Since the latter are all found to be close to or on the diagonal line, the probability interpretation is restored (see Section 4.2).

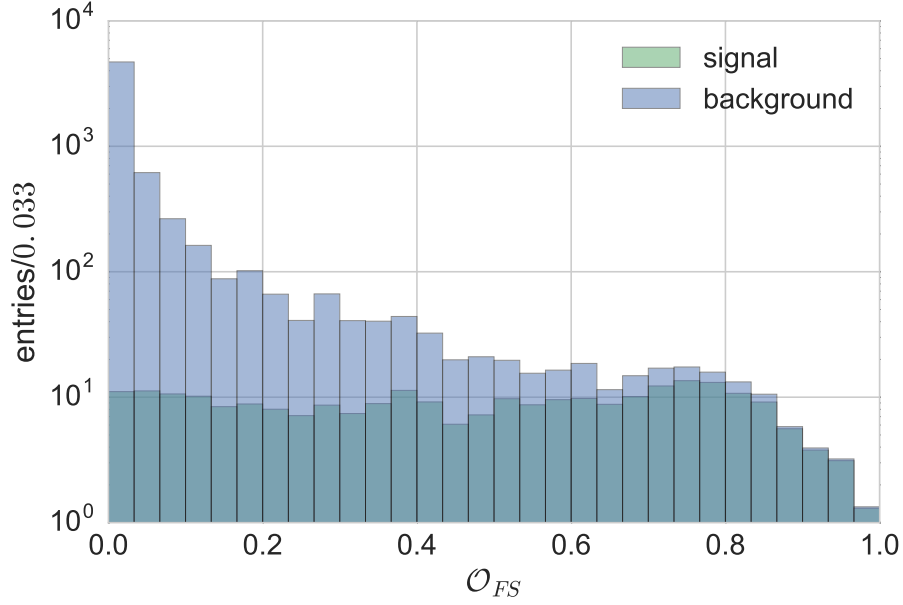


Figure 6.5.: FSBDT expert output for an averaged sample in logarithmic scaling. The background includes the generic and the continuum components.

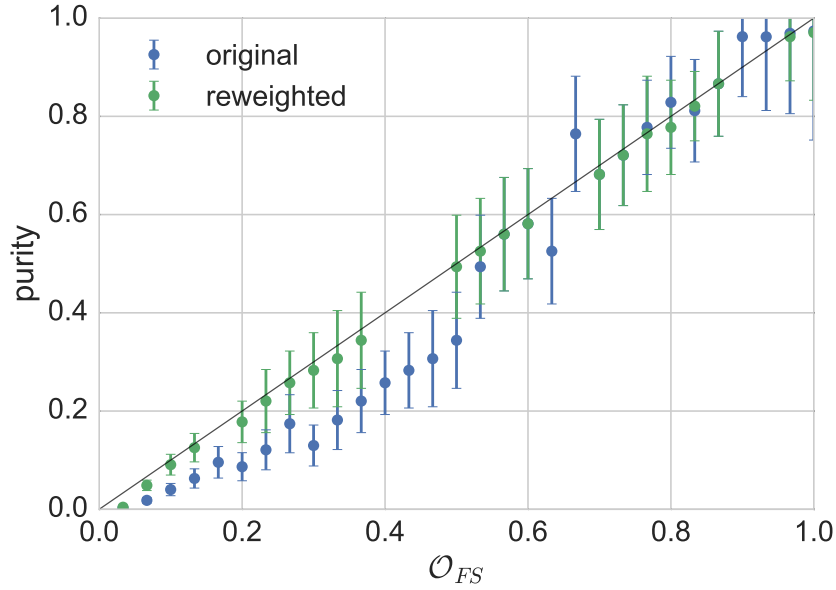


Figure 6.6.: Purity per bin for the original and reweighted output of the FSBDT.

Despite the benefit of handling the correlations correctly, it is a controversial decision to replace a data-driven method like the three-dimensional fit by an MC-driven method like the FSBDT, since the overall MC dependency of the analysis is increased. A combined solution using the benefits of both methods could be the *advanced s-plot method* [26], which makes it possible to train a multivariate classifier directly on real data. Fortunately this method will be implemented in BASF2 in the near future and can hence be considered as an option for subsequent analyses.

6.2.5. Δt

The determination of the proper decay time difference Δt is explained in detail in Section 3.3. Since it contains the information of the B^0 mesons' time dependence, it denotes the variable on which the CP fit is performed (see Chapter 7). The shape of its distribution cannot be distorted, in order to avoid biases in the fit. Hence it is not used as an input variable for any of the BDTs used throughout the selection process. However a pre-cut $|\Delta t| < 20 \text{ ps}$ is applied. This cut value is optimized according to the statistical uncertainty $\sigma(S_f)$ of the CP fit and is found tighter than in comparable Belle analyses [17]. However, since the expected lifetime of a B^0 meson is given as $\tau_{B^0} = 1.519 \text{ ps}$ [5], the applied cut already retains events with more than a factor of $10\times$ higher lifetimes than the expected one. The tight cut makes an extension of the fit model by a so called *outlier PDF*, which is used in order to cover events with $|\Delta t| > 40 \text{ ps}$, unnecessary. Figure 6.7(*top*) shows the distribution for Δt .

As shown in Figure 6.7(*bottom*), no separation between signal and background can be observed on the error of the proper decay time difference $\sigma(\Delta t)$. Hence no further pre-cuts are applied to this variable.

6.2.6. List of pre-cuts

Table 6.1 shows a list of all pre-cuts applied to the B^0 candidate list before performing the CP fit. In summary, relatively loose pre-cuts are chosen in order to postpone the bulk of the selection towards the CP fit itself. This is done because the events are yet to be weighted by the fit model according to the signal probability f_{sig} and the cut on \mathcal{O}_B is yet to be optimized directly on the outcome of the measurement (see Section 7.4).

pre-cut	unit
$0.3 < m_{K_S^0} < 0.7$	GeV
$5.26 < Mbc < 5.29$	GeV
$ \Delta E < 0.1$	GeV
$ \Delta t < 20$	ps
$\mathcal{O}_B > 0.1$	—
$\mathcal{O}_{CS} > 0.1$	—

Table 6.1.: Summary of all pre-cuts applied to the B^0 candidate list.

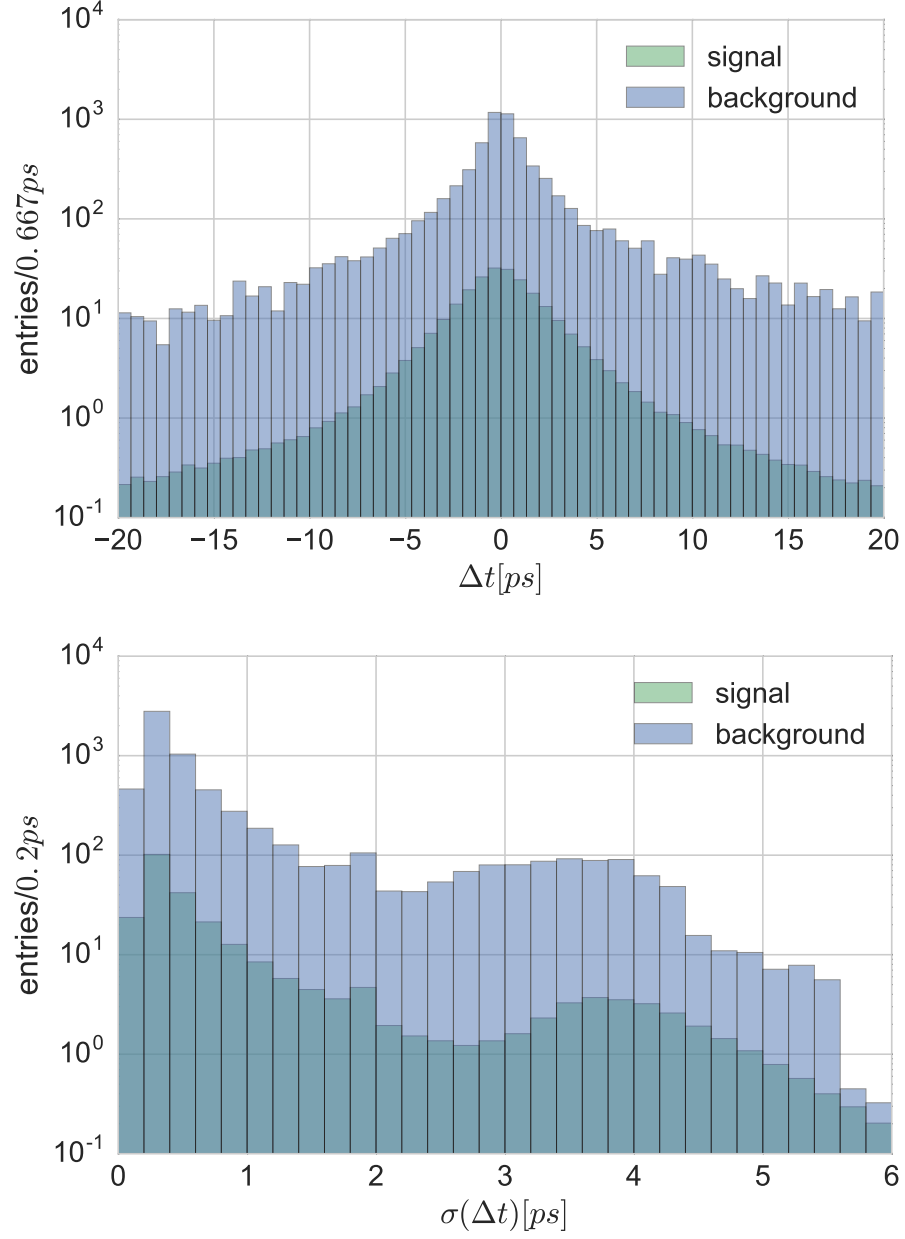


Figure 6.7.: Δt distribution of an averaged sample after application of all pre-cuts in logarithmic scaling *top*. $\sigma(\Delta t)$ distribution of an averaged sample in logarithmic scaling *bottom*. The background includes the generic and the continuum components.

7. Fit models and methods

This chapter introduces the model for the time dependent CP fit. It represents the expected behaviour of the particles according to the Standard Model. Since S_f and A_f shall be the only floating parameters in the fit, several pre-fits are performed in order to fix the other parameters. First the signal model is introduced and convolved with the according resolution function in Section 7.1. Subsequently, the same is done for the background components in Section 7.2. The combined model is introduced in Section 7.3, thereby introducing the functionality of the signal probability f_{sig} . Finally the setup of the \mathcal{O}_B -cut optimization is presented in Section 7.4.

7.1. Signal model

The signal model (Equation 2.23) is deduced in Section 2.2 and shall be presented here again in order to discuss its functionality throughout the CP fit. Note, that the B^0 lifetime τ_{B^0} and the mass difference between the two B^0 mass eigenstates Δm_d are fixed to their world average values ($\tau_{B^0} = 1.519 \times 10^{-12}$ s, $\Delta m_{B^0} = 0.510 \times 10^{12}$ \hbar s $^{-1}$) [5].

$$P_{sig}(\Delta t) = \frac{e^{-|\Delta t|/\tau_{B^0}}}{4\tau_{B^0}} [1 + q\Delta w + q(1 - 2w) (S_f \sin(\Delta m_{B^0}\Delta t) + A_f \cos(\Delta m_{B^0}\Delta t))]. \quad (7.1)$$

Here w , q and Δw denote the flavor-tagging information determined in Section 4.5. The wrong tag fraction w can be seen here as a per-event weight. Events with a high confidence on the tagging decision, i.e. a low assigned w , shall affect the fit more than ones with low confidence. Since the detector comes with a finite position resolution, the determined vertex positions suffer a smearing around their true value. In order to take this effect into account, the model 7.1 has to be forward convolved with a resolution function R_{sig} . This function models the signal resolution, which is defined as $\Delta t - \Delta t'$, where $\Delta t'$ is the true, i.e. un-smearred vertex position. For signal events, $\Delta t'$ is available as the MC-Truth information of each candidate. Hence the resolution can be obtained by fitting R_{sig} to the distribution of $\Delta t - \Delta t'$, where R_{sig} is modelled by a triple Gaussian:

$$R_{sig}(\Delta t - \Delta t') = f_{main}^{sig} G(\Delta t - \Delta t', \mu, \sigma_{main}) + (1 - f_{main}^{sig}) + (f_{tail}^{sig} G(\Delta t - \Delta t', \mu, \sigma_{tail}) + (1 - f_{tail}^{sig}) G(\Delta t - \Delta t', \mu, \sigma_{ol})). \quad (7.2)$$

The shape parameters μ , f_{main}^{sig} , f_{tail}^{sig} , σ_{main} , σ_{tail} and σ_{ol} are determined in a separate step, where R_{sig} is fit to a MC sample of 2×10^6 correctly reconstructed signal events. Figure 7.1

shows the plot and Table 7.1 shows the resulting values (with all amplitudes normalized). Since R_{sig} is convolved with the signal model, it needs to be normalized in order to ensure normalization for the resulting PDF of the convolution. The convolution integral is defined

μ	-0.0477 ± 0.0115
f_{main}^{sig}	0.326 ± 0.0098
f_{tail}^{sig}	0.588 ± 0.0064
σ_{main}	0.454 ± 0.0082
σ_{tail}	1.112 ± 0.0336
σ_{ol}	4.156 ± 0.0673

Table 7.1.: Shape parameters of R_{sig} .

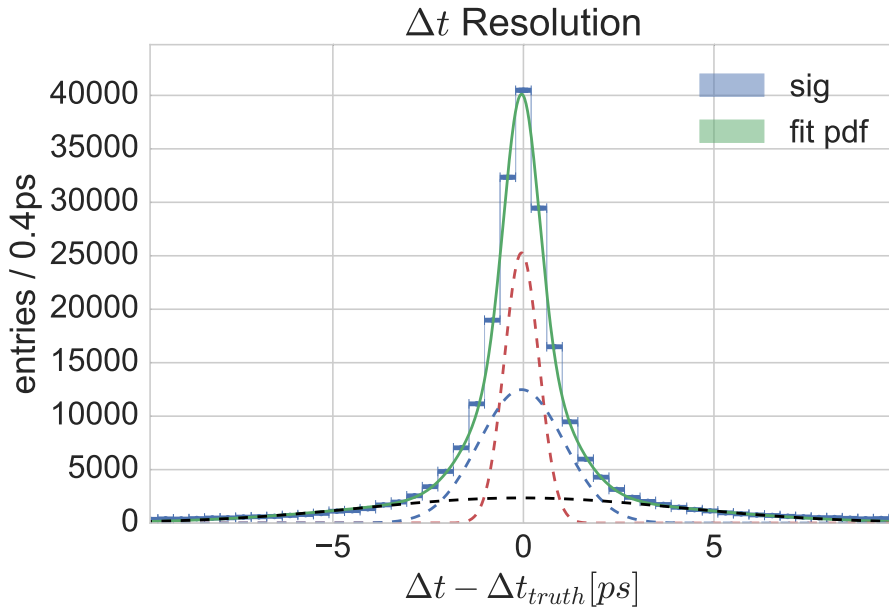


Figure 7.1.: Triple-Gauss fit to R_{sig} . The dashed lines show the three single components of the model: σ_{main} (red), σ_{tail} (blue) and the outlier component σ_{ol} (black).

as

$$\mathcal{P}_{sig,i} = \int [P_{sig}(\Delta t') R_{sig}(\Delta t_i - \Delta t')] d(\Delta t'). \quad (7.3)$$

It illustrates in a nice way how R_{sig} can be seen as a weighting factor of the model. The measured values Δt_i are only fed into R_{sig} , where they get evaluated according to their distance to the corresponding value of the integration constant $\Delta t'$. As both distributions contributing to the convolution are normalized PDFs, the outcome $\mathcal{P}_{sig,i}$ can be interpreted as a per-event likelihood. The unbinned handling of events is necessary at this point, since the signal model (see Equation 7.1) is a conditional PDF depending on the per-event flavor tagging information.

The CP fit is performed by maximizing the likelihood function $L = \prod_i \mathcal{P}_{sig,i}(\Delta t_i; S_f, A_f)$, or minimizing the negative logarithmic likelihood, respectively (see Section 4.3). Note that for each step of the minimization process, i.e. each new set of parameters S_f and A_f , the convolution (Equation 7.3) is performed again. As this is being done numerically, the integral

from Equation 7.3 is transferred into a sum:

$$\int_a^b f(x) dx \approx \left(\frac{b-a}{n}\right) \sum_{i=1}^n f(x_i) \quad (7.4)$$

This transformation yields instant normalization for the convolved model.

7.2. Background model

The background model represents a theoretical description of the expected behaviour of the background components introduced in Section 4.1. The B^0 mesons of the first component, namely the generic background, which undergo a $b \rightarrow c$ transitions are expected to exhibit different lifetimes as opposed to the signal decays due to a different phase space. Hence the generic background is modelled by an exponential function with a lifetime τ_{bg} . The second component, i.e. the continuum background, is modelled by a δ function. This is due to the fact that the resulting quarks of the process $e^+e^- \rightarrow q\bar{q}$ occur on a very small time-scale and the corresponding time dependence is hence assumed to be a prompt distribution [8]. The background model is given by

$$P_{bg}(\Delta t) = f_\delta \delta(\Delta t - \mu_\delta) + (1 - f_\delta) e^{\frac{-|\Delta t - \mu_{bg}|}{\tau_{bg}}}, \quad (7.5)$$

This model will be combined with the signal model (see Equation 7.1) later to form the combined model, which then represents a complete description of the data obtained from the detector.

Since the background events also suffer a smearing effect by the finite vertex resolution, a convolution with a double-Gaussian resolution function R_{bg} is performed according to Equation 7.3. However, as the continuum component does not consist of real B^0 candidates, there is no truth information for Δt from MC. Hence a resolution definition as used for signal events is not valid anymore. Instead a conditional function is introduced by

$$R_{bg}(\Delta t; \sigma_{\Delta t}) = f_{bg}^{main} G(\Delta t, 0, s_{bg}^{main} \sigma_{\Delta t}) + (1 - f_{bg}^{main}) G(\Delta t, 0, s_{bg}^{tail} \sigma_{\Delta t}), \quad (7.6)$$

which depends on the per-event error $\sigma_{\Delta t}$. This error is calculated via

$$\sigma_{\Delta t} = \sqrt{\sigma_{recV}^2 + \sigma_{tagv}^2} / (\gamma\beta c), \quad (7.7)$$

as a propagation of the errors on the two B^0 -Vertex fits (see Section 5.2). Note that the factor $\gamma\beta c$ corresponds to the Lorentz Boost described in Equation 3.2. As a replacement for the shape parameters in the Gauss-function, the scaling factors s_{bg}^{main} and s_{bg}^{tail} are introduced. The convolution in the background case is given by

$$\mathcal{P}_{bg,i} = \int [P_{bg}(\Delta t') R_{bg}(\Delta t_i; \sigma_{\Delta t,i})] d(\Delta t'), \quad (7.8)$$

where the prompt component is integrated analytically like

$$\int \delta(x - x_0) f(x) dx = f(x - x_0). \quad (7.9)$$

Similar to the pre-fit of the signal resolution function R_{sig} , a separate unbinned Maximum-Likelihood fit of $\mathcal{P}_{bg,i}$ is performed on the background MC samples in order to determine the shape parameters $\mu_{\tau_{bg}}, \tau_{bg}, s_{bg}^{main}, s_{bg}^{tail}, f_{bg}^{main}, f_\delta$ and μ_δ . Table 7.2 and Figure 7.2 show the results. Since $\mathcal{P}_{bg,i}(\Delta t, \sigma_{\Delta t})$ is a conditional PDF depending on the per-event information

$\mu_{\tau_{bg}}$	0.600 ± 0.404
τ_{bg}	5.394 ± 3.184
s_{bg}^{main}	0.786 ± 0.050
s_{bg}^{tail}	2.148 ± 0.116
f_{bg}^{main}	6.574 ± 4.620
f_{δ}	0.102 ± 0.1996
μ_{δ}	0.0004 ± 0.003

Table 7.2.: Shape parameters of the background model P_{bg} .

$\sigma_{\Delta t}$, this dependence needs to be marginalized in order for it to be visible in the Δt dimension. This projection is done by building a PDF for each value $\sigma_{\Delta t,i}$ in the data set and constructing an average curve C from these [23]:

$$C(\Delta t) = \frac{1}{n} \sum_{i=0}^n \mathcal{P}_{bg,i}(\Delta t; \sigma_{\Delta t,i}). \quad (7.10)$$

Note that this task is automatically done, when using high-level ROOT functions, which is why the per-event errors like in Figure 7.2 are generally not displayed.

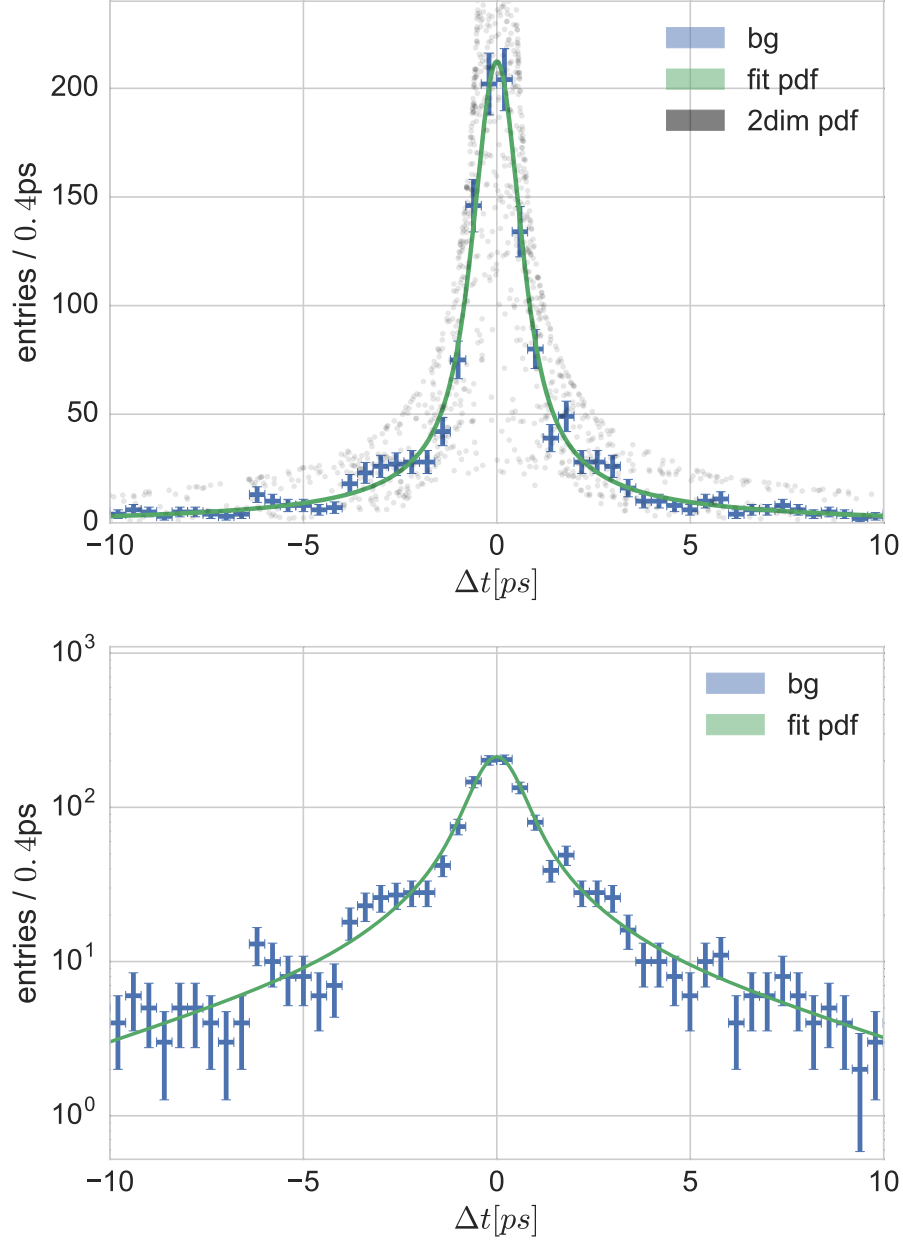


Figure 7.2.: Background model fit to the Background MC sample with the shape parameters from Table 8.4. *Top*: Averaged PDF in non-logarithmic scaling. The dots are obtained by assigning the likelihood $\mathcal{P}_{bg,i}$ to each event. This illustrates the $\sigma_{\Delta t}$ dependency, when not marginalized beforehand. *Bottom*: Averaged PDF from Equation 7.10 displayed in logarithmic scaling.

7.3. Combined model

The combined model shown in Equation 7.11 is given by a linear combination of the signal model and the background model. Note that S_f and A_f are the only floating parameters left in this model, since all shape parameters have been determined via pre-fits as described in the previous sections.

$$\mathcal{P}_{comb}(\Delta t) = f_{sig} \mathcal{P}_{sig} + (1 - f_{sig}) \mathcal{P}_{bg} \quad (7.11)$$

Here the scaling factor f_{sig} weights each event according to its probability of being a signal-like event. The underlying idea is to feed signal-like events into the signal component of the model \mathcal{P}_{sig} and background-like events into the background component \mathcal{P}_{bg} . As only \mathcal{P}_{sig} includes the CP-parameters S_f and A_f the aim is for background-like events to not affect their determination. Hence f_{sig} is referred to as *signal probability* and denotes per-event information, which is provided as the reweighted FSBTD output \mathcal{O}_{FS} (see Section 6.2.4). The functionality of f_{sig} in the model is the justification for not applying any further cuts on \mathcal{O}_{FS} , but rather taking more events into the fit and letting f_{sig} do the selection by weighting down the background-like events. This method is evaluated in Section 8.4 by determining the uncertainties $\sigma(S_f)$ and $\sigma(A_f)$ for different cut values on \mathcal{O}_{FS} .

7.4. Optimization of the \mathcal{O}_B -cut value

As described in Section 4.6, one of the goals of this thesis is an analysis-specific optimization of the K_s^0 -selection. Hence one cut-value of \mathcal{O}_B is determined, which minimizes the statistical uncertainties $\sigma(S_f)$ and $\sigma(A_f)$.

When scanning over the \mathcal{O}_B -cut value, the flavor tagging calibration, the signal probability calibration as well as the signal resolution shape parameters need to be determined for each cut-value, because these denote information which heavily depends on the pre-selection settings. Further, an averaging over 300 signal samples is performed for each \mathcal{O}_B -cut value, in order to increase the statistics throughout the optimization. Note however, that due to a lack of additional MC events, no averaging of background samples is performed. This could lead to potential biases or wrongly estimated uncertainties. The averaging is done by performing the CP-fit for each of the 300 scaled MC samples and subsequently determine the mean and the standard deviation of the 300 obtained independent results. Figure 7.3 shows this process. The results, i.e. values to be recorded during the optimization, are the uncertainties of

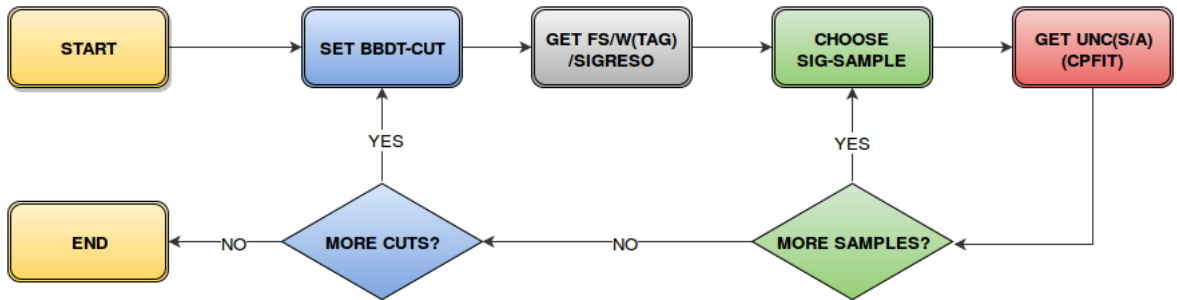


Figure 7.3.: Flowchart of the optimization-algorithm.

the CP violation parameters $\sigma(S_f)$ and $\sigma(A_f)$, which are taken directly from the fit result. Furthermore, a *pull* value is computed from each fit result defined by

$$pull(S_f) = \frac{S_{f,i} - S_{f,true}}{\sigma(S_f)} . \quad (7.12)$$

Here $S_{f,i}$ denotes the resulting value for sample i and $S_{f,true} = 0.7$ is taken from the truth generated in MC (see Table

8. Results and discussion

8.1. Signal MC validation

In this section the signal model is validated by fitting $\mathcal{P}_{sig,i}$ to the Signal MC sample from Section 7.1. Here most of the implementations which have been developed up to this point in the analysis are evaluated. For a positive evaluation an unbiased fit result is expected, i.e. the values of S_f and A_f should be obtained as they have been generated in the MC sample. Table 8.1 shows the results. Both values of S_f and A_f deviate less from their generated values than the corresponding standard deviation. Figure 8.1(*top*) shows the (signal MC) CP fit result, which is split up into B^0 -tagged events ($r > 0.9$) and \bar{B}^0 -tagged events ($r < -0.9$). This splitting is done for better visualization of the CP asymmetry, which manifests itself in the shifting between the two distributions. This effect is illustrated more obviously by the so-called asymmetry-plot shown in Figure 8.1(*bottom*). Here the difference between the two distributions is calculated for each bin and normalized:

$$(N_{B^0} - N_{\bar{B}^0}) / (N_{B^0} + N_{\bar{B}^0}). \quad (8.1)$$

If there was no CP asymmetry ($S_f = A_f = 0$) only a straight horizontal line would be observed. Hence this method denotes a nice way of visualizing the intensity and properties of the asymmetry.

Table 8.1.: Resulting values of the CP-fit to the Signal MC sample.

	S_f	A_f
generated	-0.7	0
Sig-MC-result	-0.704 ± 0.010	0.0016 ± 0.0056

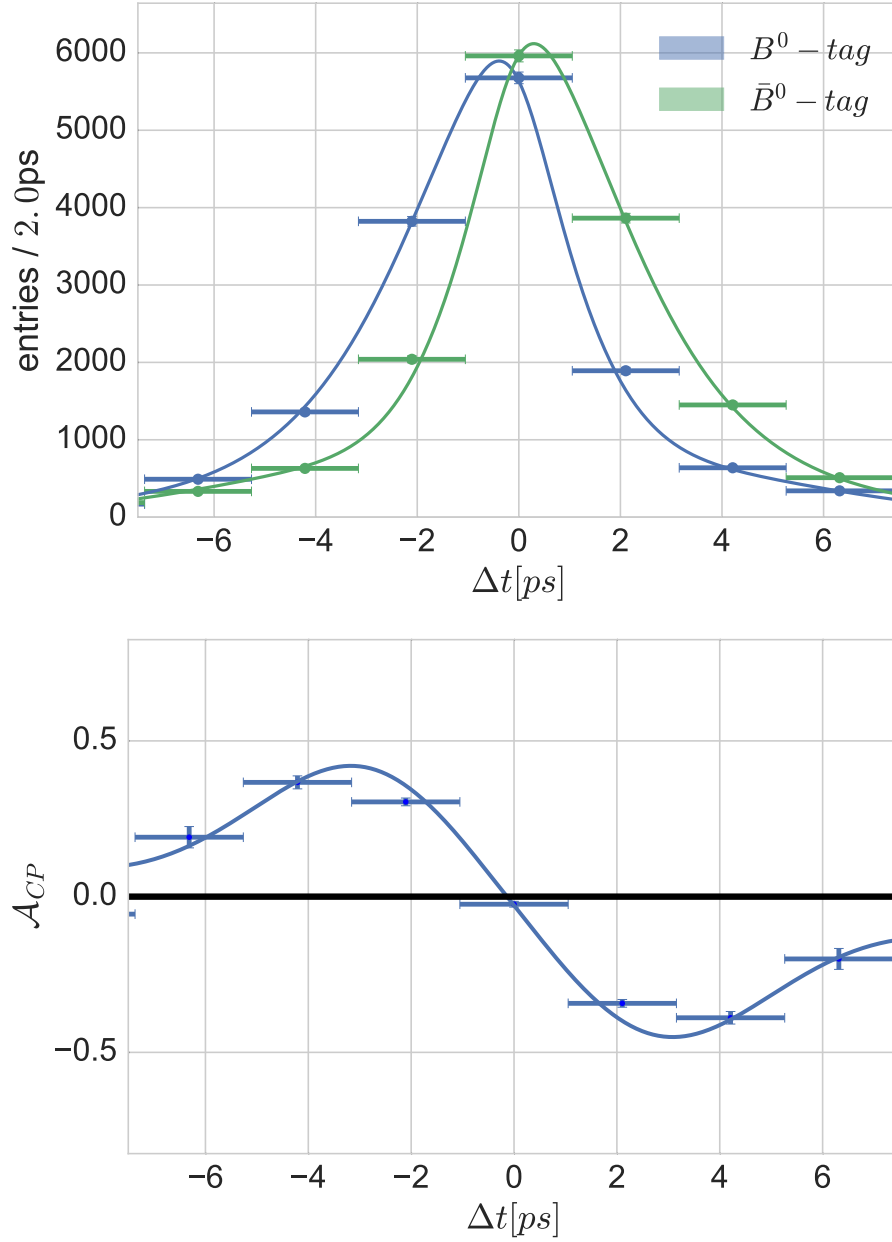


Figure 8.1.: CP Fit on the Signal MC sample *top*. Asymmetry plot of the signal MC sample *bottom*.

8.2. \mathcal{O}_B -cut optimization results

Figure 8.2 shows the optimization results plotted over the cut value of \mathcal{O}_B . Note that each marker denotes the mean of the 300 results for the corresponding cut value, while the error-bar denotes the standard deviation.

The minimum value of the uncertainty $\sigma(S_f)$ is obtained in the fit with $\mathcal{O}_B > 0.25$. This value represents the suggested selection cut for optimizing the determination of S_f . Therefore, one obtains $\sigma(S_f) = 0.294$ and $\sigma(A_f) = 0.193$ at this minimum. When looking at the corresponding pull distribution however, a deviation from $\mu = 0$ for low \mathcal{O}_B -cut values can be observed. This bias may be a result of only having performed the optimization using the one available background sample (see Section 7.4). For higher cut values the back-

ground contribution decreases, and hence the corresponding bias decreases. This explanation is strengthened, since no bias has been observed on the signal MC validation (see Section 8.1). Furthermore, for the standard deviation a value of $\sigma(\text{pull}(S_f)) < 1$ is observed. Looking at Equation 7.12 this can be due to either $\sigma(S_f)$ being overestimated by the fit or the observed deviation $S_{f,i} - S_{f,true}$ being smaller than expected. The latter is the more likely case, since again only having one background sample at hand introduces a lack of fluctuation or deviation.

The results for parameter A_f are found to be more consistent with expectation, i.e. not found to be as affected by the background-introduced bias. The uncertainty $\sigma(A_f)$ converges at a cut value $\mathcal{O}_B > 0.72$ and the corresponding pull distribution shows a mean of $\mu \approx 0$ and a standard deviation of $\sigma \approx 1$. At this minimum one obtains $\sigma(S_f) = 0.307$ and $\sigma(A_f) = 0.184$

The Figure of Merit FOM_B shows a maximum at a cut value of $\mathcal{O}_B > 0.97$. Since the scan over \mathcal{O}_B does not cover cut values > 0.97 , the convergence is not visible in this plot. Figure 8.6(bottom) however, shows a clear maximum at this value. Nevertheless, the uncertainties $\sigma(S_f)$ and $\sigma(A_f)$ are both found to increase heavily at $BBDT = 0.97$ as opposed to their suggested cut values. Hence FOM_B does not denote a sensible cutting criteria, which is not surprising given its high dependence on the signal to background ratio and the $B_{f_{sig}}$ cut respectively.

If one were to decide between cutting at the values suggested by $\sigma(S_f)$ or $\sigma(A_f)$, which corresponds to $\mathcal{O}_B > 0.25$ or $\mathcal{O}_B > 0.72$, one would prefer the tighter cut. Given that both parameters are equally important to determine, a tighter cut usually promises a cleaner data sample and hence less systematic biases. The same is observed here by looking again at the pull distribution of S_f , where at $\mathcal{O}_B > 0.72$ the described background-introduced bias has vanished. The most obvious approach to a K_S^0 selection cut, however, is to simply cut on the BDT_{KS} output before even reconstructing B^0 candidates (see Section 4.6). From the KSBDT training output (see Appendix B.1) a cut value of $\mathcal{O}_{KS} > 0.49$ is suggested, which refers to the maximum of the according Figure of Merit FOM_{KS} . In order to compare this method to the analysis-specific one suggested by this study, the same 300 samples, which have been used for the optimization before, are cut at $\mathcal{O}_{KS} > 0.49$, and the analysis is performed without using a BBDT. By doing so, the results $\sigma(S_f) = 0.356$ and $\sigma(A_f) = 0.212$ are obtained. This corresponds to values more than 10% higher than the ones obtained by minimizing $\sigma(A_f)$, which shows a clear improvement by the latter method. Table 8.2 shows all obtained results for comparison.

opt. parameter	\mathcal{O}_B -cut	$\sigma(S_f)$	$\sigma(A_f)$
$\sigma(S_f)$	0.25	0.294	0.193
$\sigma(A_f)$	0.72	0.307	0.184
FOM_B	0.97	0.393	0.226
FOM_{KS}	—	0.356	0.212

Table 8.2.: Resulting uncertainties for different selection methods. The left column shows the corresponding parameter on which the selection is optimized.

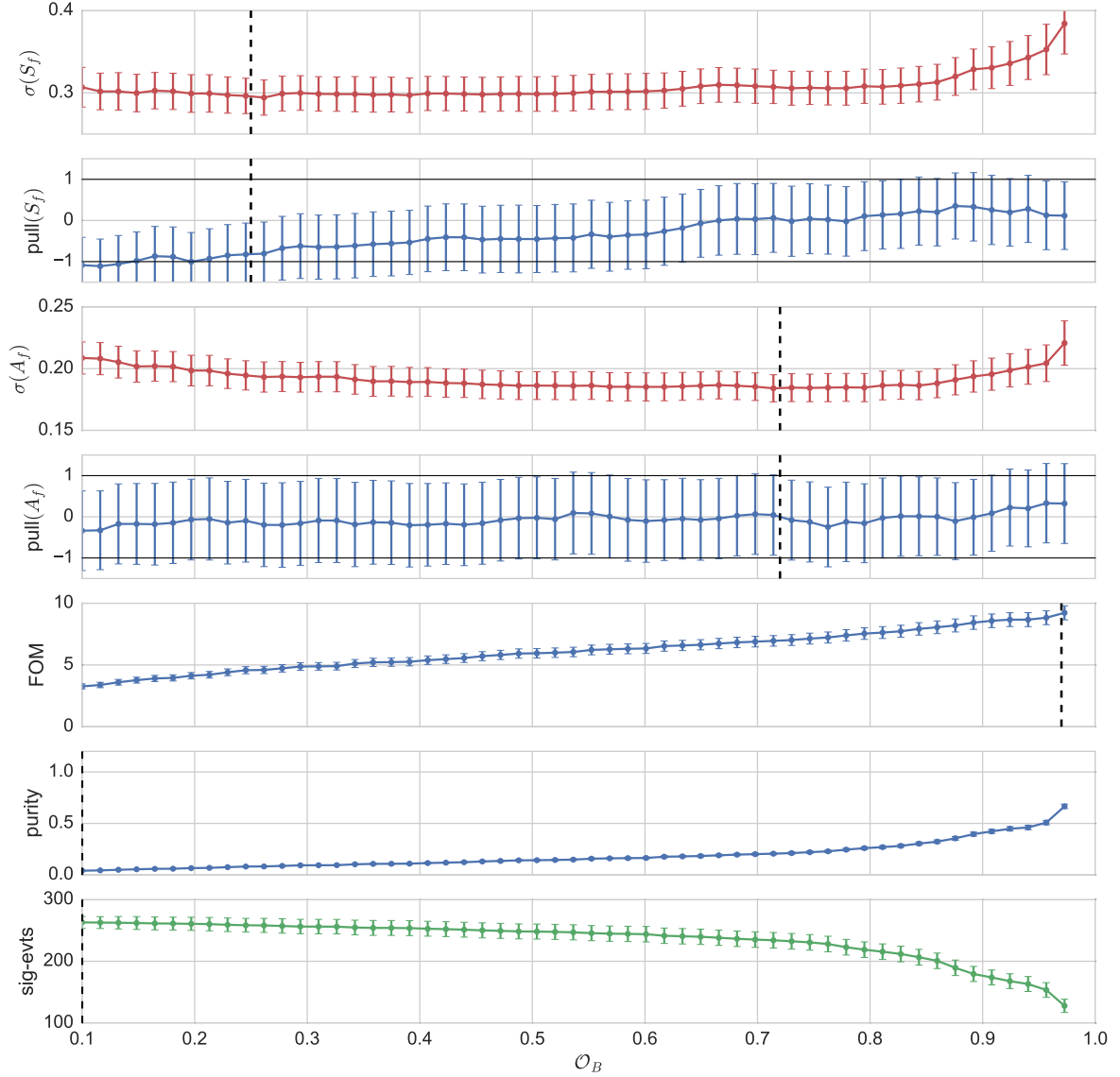


Figure 8.2.: Resulting values recorded during \mathcal{O}_B -cut value optimization. The dashed lines represent cut values suggested for optimization of the corresponding parameter.

8.3. Comparison to previous studies

Figure 8.3 shows an extrapolation of the uncertainty $\sigma(S_f)$ for different amounts of data, which was made from Belle's $B \rightarrow K_S^0 K_S^0 K_S^0$ measurement. As this analysis uses $1ab^{-1}$ equivalent of data, the corresponding expectation yields $\sigma(S_f) \approx 0.24$. Comparing this value with the ones shown in Table 8.2, this study is not able to match the mentioned expectations, which is not surprising due to the following reasons:

- The branching fractions of a decaying K_S^0 are given by [5]:

$$\begin{aligned} bf(K_S^0 \rightarrow \pi^0 \pi^0) &= 0.307, \\ bf(K_S^0 \rightarrow \pi^+ \pi^-) &= 0.692. \end{aligned} \quad (8.2)$$

The branching fractions of the signal B^0 then yield:

$$\begin{aligned} bf(B^0 \rightarrow K^\pm K^\pm K^\pm) &= 0.331, \\ bf(B^0 \rightarrow K^{00} K^\pm K^\pm) &= 0.441, \\ bf(B^0 \rightarrow K^{00} K^{00} K^\pm) &= 0.196. \end{aligned} \quad (8.3)$$

Here K^{00} denotes a K_S^0 decaying via $K_S^0 \rightarrow \pi^0 \pi^0$ and K^\pm via $K_S^0 \rightarrow \pi^+ \pi^-$ respectively. The K^{00} are not included in this analysis, because the Belle II vertex fitting tool is not yet able to process K_S^0 coming from different reconstruction paths. Hence only the first line of Equation 8.3 contributes to this analysis, i.e. roughly 67% of the signal decay events are not included.

- Tracking algorithms are still under construction, e.g., information from the CDC is not yet extrapolated and hence not combined with hits from the inner vertex detectors. This heavily affects the vertex information provided for long living K_S^0 mesons.
- Using a Triple-Gauss function for the Δt signal resolution R_{sig} only denotes a simplified approach as opposed to a very elaborated resolution treatment at Belle [27].

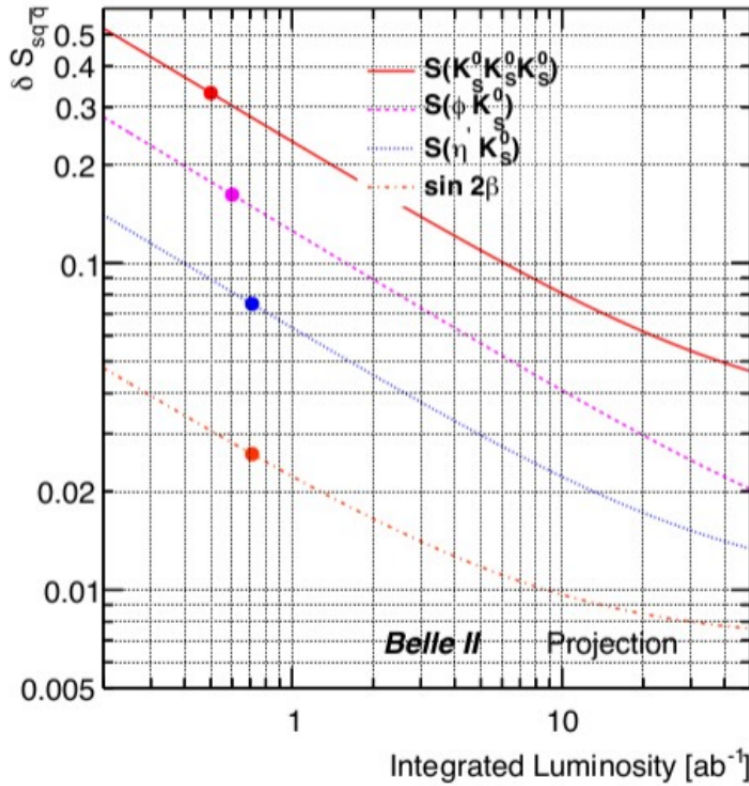
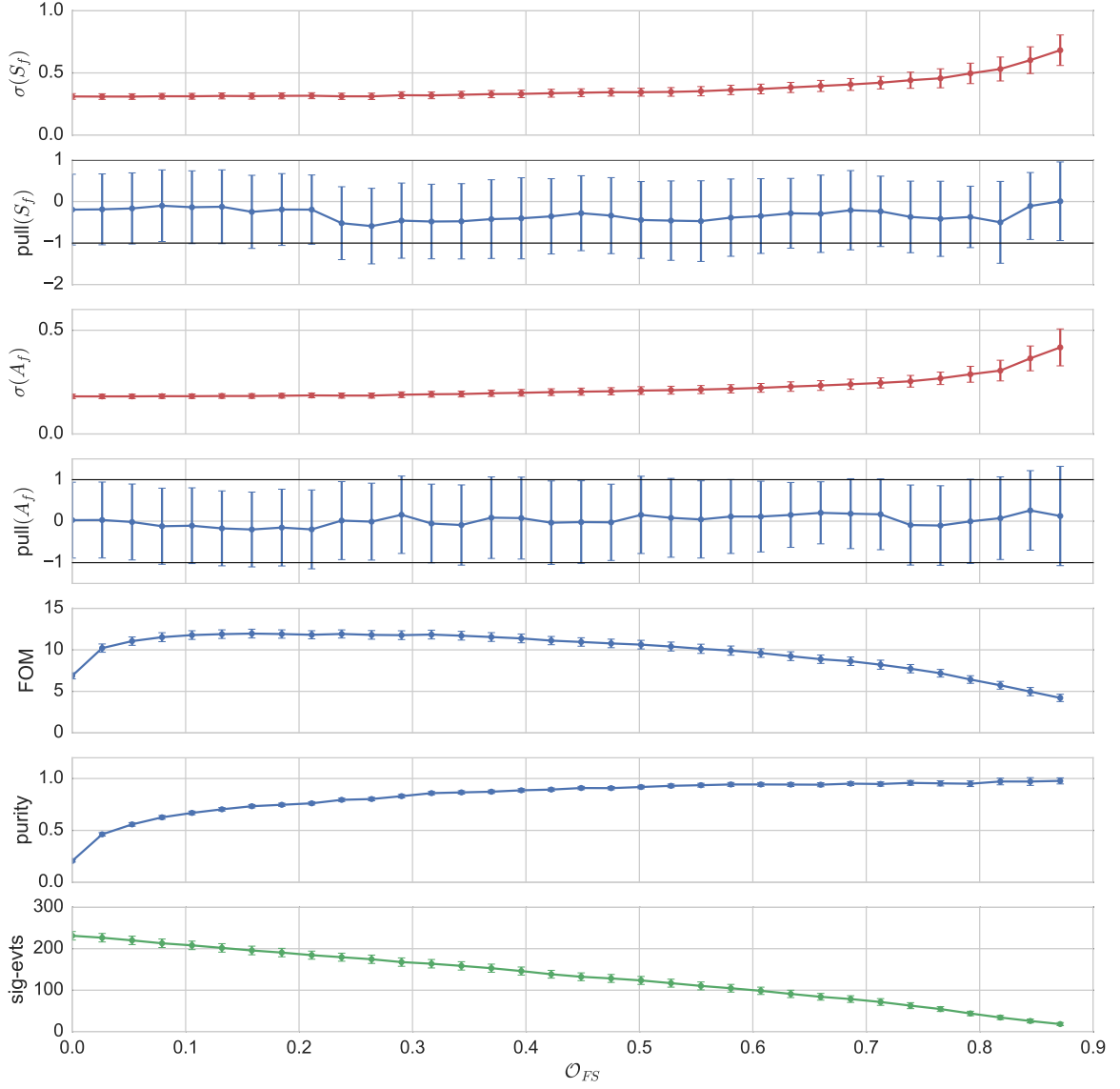


Figure 8.3.: Extrapolation of $\sigma(S_f)$ as expected for different amounts of data. The dots represent the results of the most recent Belle analyses. From [28].

8.4. FSBDT optimization results

In order to evaluate the signal probability f_{sig} used throughout this analysis, the optimization described in Section 7.4 is also performed by scanning over the cut value of the $FSBDT$, \mathcal{O}_{FS} . Note that, since optimization is a very time-consuming task, it is only performed by averaging over 40 samples here instead of 300. However, the reduced number of samples is sufficient for performing this evaluation. At the BBDT, the previously determined value of $\mathcal{O}_B > 0.72$ is now applied. Figure 8.4 shows the results. As expected, $\sigma(S_f)$ and $\sigma(A_f)$ both exhibit a minimum at $\mathcal{O}_{FS} = 0$, which denotes a confirmation to not apply any selection cuts on the $FSBDT$ throughout this analysis. It further shows, how the f_{sig} method in the fit-model (Equation 7.11) works nicely by down-weighting background events and hereby preventing them from affecting the determination of S_f and A_f , instead of cutting them out. Thus, again, the Figure of Merit, which here suggests a cut at $\mathcal{O}_{FS} = 0.25$, is not determined to be a good cut criteria.

Figure 8.4.: Resulting values recorded during \mathcal{O}_{FS} -cut value evaluation.

8.5. Full MC results

In this section, the actual time-dependent measurement is simulated, as would be performed using real data coming from the detector. The scaled Full MC sample is obtained by randomly choosing one of the 300 signal samples used for the \mathcal{O}_B optimization. Subsequently the previously optimized cut $\mathcal{O}_B > 0.72$ is applied and the CP-fit is performed. Note that the Full-MC study imitates the un-blinding step of a real analysis, where no more MC information is used and all scalings are performed correctly according to the branching fractions (see Section 4.4). This corresponds to the data-flow on the right hand side of Figure 4.8. The results are shown in Table 8.3. The surprisingly low values of $\sigma(S_f)$ and $\sigma(A_f)$ denote only a lucky fluctuation from the averaged ones shown in Table 8.2. Figure 8.5 shows the corresponding plots, where

for a cleaner visualization the lines represent only the signal model (background subtracted plot [29]). The dataset is shown in a signal-enhanced plot, i.e. $r > 0.5$ and $\mathcal{O}_{\mathcal{FS}} > 0.5$.

Table 8.3.: Resulting values of the CP fit to the Full MC sample.

	S_f	A_f
generated	-0.7	0
Full-MC-result	-0.687 ± 0.297	0.074 ± 0.178

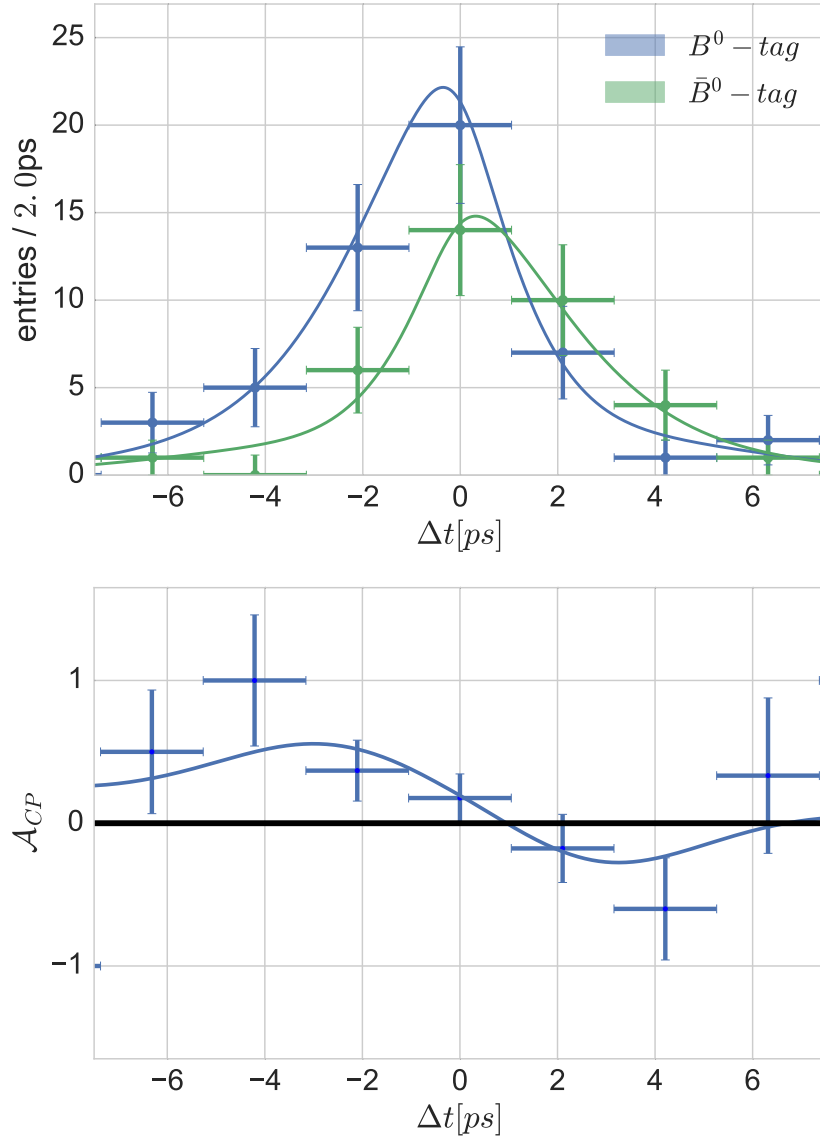


Figure 8.5.: *Top:* CP Fit on the Full MC sample. *Bottom:* Asymmetry plot of the Full MC sample.

Note, that the study of systematic errors is not covered by the scope of this thesis. From previous studies however (see Appendix A), we can extract an expectation for the scale of the systematic errors, which is much smaller than the statistics investigated here. Hence the latter

is the main source of uncertainty throughout this analysis.

8.6. Reconstruction and selection efficiencies

Table 8.4 shows the number of events which remain before (second column) and after (third column) the application of the selection cut on \mathcal{O}_B . Hence the second column represents information about reconstruction efficiencies, whereas the third column shows the selection efficiency of the \mathcal{O}_B cut.

sample	before \mathcal{O}_B -cut	after \mathcal{O}_B -cut
<i>signal</i>	264	234
<i>misrec.signal</i>	82	18
<i>mixed</i>	490	83
<i>charged</i>	164	13
<i>uubar</i>	529	80
<i>ddbar</i>	234	47
<i>ssbar</i>	1364	268
<i>ccbar</i>	3628	380

Table 8.4.: Number of remaining events per sample after pre-selection.

For the B^0 signal reconstruction efficiency, one finds using Equation 4.14:

$$\epsilon_{S,B^0} = \frac{N_{rec}}{N_{gen}} = \frac{264}{3152} = 8.37\%. \quad (8.4)$$

This in return yields a K_S^0 reconstruction efficiency of

$$\epsilon_{S,K_S^0} = \frac{\sqrt[3]{\epsilon_{S,B^0}}}{bf(K_S^0 \rightarrow \pi^+\pi^-)} = \frac{0.203}{0.69} = 63.40\%. \quad (8.5)$$

Note, that this study only covers K_S^0 decaying to charged pions. The tracking efficiency of π^{+-} can hence be determined to be

$$\epsilon_{S,\pi^\pm} = \sqrt{\epsilon_{S,K_S^0}} = 79.62\%. \quad (8.6)$$

These numbers are meant to serve future Belle II analyses for comparison. Note however, that MC events have been generated without beam background throughout this study, which would cause the tracking efficiency to decrease significantly [23].

The third column represents the selection efficiency, since it shows the remaining events after cutting on \mathcal{O}_B . Note that these numbers represent the sample size, as they are used throughout the Full MC study. The corresponding signal fraction can be calculated as

$$sf_{exp} = 0.208, \quad (8.7)$$

where sf_{exp} here corresponds to the expert step as opposed to the signal fraction sf_{train} during the BDT training. One should keep in mind that the total number of remaining background events and hence sf_{exp} has a questionable significance due to the fact that most of them are detected as background and down-weighted accordingly by f_{sig} throughout the fit.

8.7. V0 Finder performance

As described in Section 5.1, this analysis implemented the V0 Combiner module for K_S^0 reconstruction. The performance of the three different methods V0 Finder, RD and V0 Combiner are compared in this section by evaluating their ROC curves and their corresponding Figures of Merit. Both plots are obtained by scanning over \mathcal{O}_B , hence all ROC curves correspond to the BBDT selection. Figure 8.6 shows the results.

Looking at the ROC plot (*top figure*), the V0 Finder can be observed to have the lowest reconstruction efficiencies. This can be explained by the tight online cuts at the track-reconstruction stage. The RD method in contrast, shows better efficiency which decreases rapidly with higher purity. This indicates a low separation power between signal and background events. As expected, the V0 Combiner module combines the qualities of its two predecessors, i.e. the efficiency of RD and the separation power of the V0 Finder. Interestingly, the V0 Finder seems not to find many new K_S^0 , i.e. ones overseen by the RD method, instead it shows a more powerful separation for the existing ones. This is exactly what one would expect from the energy correction described in Section 5.1, which yields a shift of the invariant masses of the K_S^0 towards the nominal mass.

Additionally, the suggested selection cuts from Table 8.2 are displayed in the plot as working points, in order to compare the performances of their corresponding methods. Note, that all selection methods use the V0 Combiner reconstruction for this comparison. It can be observed, that the FOM_{KS} based method does not only yield worse uncertainty values for $\sigma(S_f)$ and $\sigma(A_f)$ but is also outperformed by the $BBDT$ method in terms of selection performance. Hence its working point does not reach the ROC curve representing the $BBDT$ selection. The Figure of Merit plot (*bottom figure*) shows similar effects in slightly different visualization. Here, again, the V0 Combiner outperforms the other reconstruction methods, while the $BBDT$ selection outperforms the FOM_{KS} based selection.

A further measure of evaluation is provided by comparing the proper decay time difference resolution for each reconstruction method. This resolution value is determined by calculating the weighted sum σ_w of the standard deviations $\sigma_{main}, \sigma_{tail}$ and σ_{ol} from the triple Gaussian fit to the signal resolution in Section 7.1:

$$\sigma_w = f_{main} \sigma_{main} + (1 - f_{main})[f_{tail} \sigma_{tail} + (1 - f_{tail}) \sigma_{ol}]. \quad (8.8)$$

The resulting values of the Δt resolution for RD, V0 Finder and V0 Combiner are shown in Table 8.5. As expected the V0 Finder outperforms the RD method and the V0 Combiner result is placed between the two. This result can be interpreted as a validation of the improved vertex fitting by the V0 Finder method (see Chapter 5).

	reco. method	σ_w
V0 Finder		1.65
RD		1.78
V0 Combiner		1.75

Table 8.5.: Δt resolution results for different reconstruction methods.

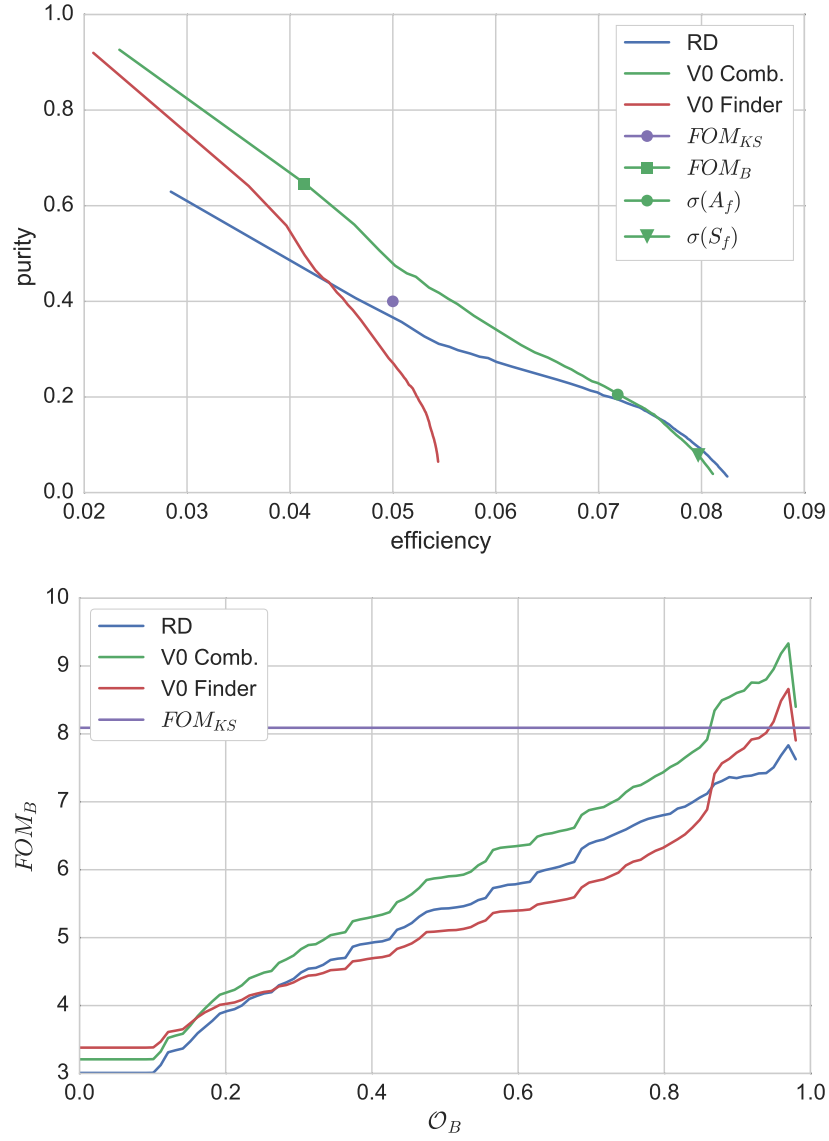


Figure 8.6.: *Top*: ROC curves from the BBDT selection for different reconstruction methods and selection working points. *Bottom*: FOM_B for different reconstruction methods.

9. Conclusion

This study presents a new handle on the minimization of the statistical uncertainty of a time-dependent CP violation fit by optimizing the K_S^0 selection. The new technique outperforms the classic FOM-based method by decreasing the uncertainty on both parameters S_f and A_f by more than 10%. This is done by forwarding the classifier output of the K_S^0 selection into a new classifier trained on reconstructed B^0 candidates. By combining the K_S^0 score of all three daughters, their correlations can be taken into account, which in return causes the enhancement. Hence this study can be used as a helpful basis for future analyses, which include any type of K_S^0 selection.

As stated in Section 8.3, the resulting uncertainties do not match the expectations obtained by extrapolating Belle’s analysis, since several components of BASF2 are still under development. It is emphasized that the goal of this study was not to provide an end-to-end analysis in order to compete with previous ones, but rather to shed light on new methods and possibilities with respect to the fast development in the field of data analysis. Once the current limitations in Belle II software are overcome, this analysis should be revisited, as the suggested methods show significant improvement in study-internal comparison to classic methods.

Furthermore, a strong dependency of the uncertainties $\sigma(S_f)$ and $\sigma(A_f)$ on the wrong tag fraction w of the Flavor Tagger are revealed throughout this study hinting upon how important the ongoing work on improving the Flavor Tagger performance is. With the future perspective of a Deep Learning approach to the Flavor Tagger, that increases the Tagging power up to $Q = 40\%$ and is currently being implemented in BASF2, a significant decreasing of the statistical uncertainties can be expected [30].

Being the first analysis to use the V0 Finder for K_S^0 reconstruction, it was possible to detect hidden cuts in the module, which cause a significant efficiency loss. With the centrally produced MC events being provided as already reconstructed tracks, some of these cuts cannot be modified by the user. This issue has been addressed in collaboration meetings and conferences and is currently under discussion. However, as part of this thesis, a solution has been developed in the form of the V0 Combiner module, which combines the benefits of the classic method and the V0 Finder and thereby outperforms the former two with respect to the ROC curve.

A special focus of this study is also set on fine-tuning the K^0 selection variables. As described in Section 6.1, missing variables are added to the framework and several entirely new

variables are added to the selection. With the recently implemented database in BASF2, this list of variables can now be used in order to set up a generic K_S^0 training. By storing the weight files in a database this training could be accessible to every user in the future.

Being one of the first analysis sensitivity studies to be carried out in BASF2, which is still under development, this work denotes an important object of evaluation coming from a user's perspective. Working together closely with the developers, many tools and modules have been tested, fixed, modified, or in case of the V0 Combiner developed, in order to handle emerging issues.

Furthermore, this analysis is the first one to perform CP fits by entirely using the JUPYTER Notebook software (see Section 4.7). As opposed to previous Belle analyses, no high level ROOT functions are used; instead a new fit framework was written from scratch using python. This framework is not meant to compete with existing methods in terms of performance, generality, or coverage of use cases. Since it is written in a high-level language however, it provides an easy to read approach making the processes and functions at hand more transparent. The idea is, that future analysts, which perform time-dependent CP violation studies, have available a user friendly code basis to work with. Additionally, by developing the framework, more knowledge and understanding could be gained for myself.

Bibliography

- [1] P. Urquijo, "Physics prospects at the Belle II experiment," *doi:10.1016/j.nuclphysbps.2015.04.004*, 2015.
- [2] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. A. Khalek, A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins *et al.*, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the LHC," *Physics Letters B*, vol. 716, no. 1, pp. 1–29, 2012.
- [3] 03.06.2016. [Online]. Available: https://en.wikipedia.org/wiki/Standard_Model#/media/File:Standard_Model_of_Elementary_Particles.svg
- [4] B. R. Martin and G. Shaw, *Particle physics*. John Wiley & Sons, 2013.
- [5] K. A. Olive *et al.*, "Review of Particle Physics," *Chin. Phys.*, vol. C38, p. 090001, 2014.
- [6] L. Wolfenstein, "Parametrization of the Kobayashi-Maskawa matrix," *Physical Review Letters*, vol. 51, no. 21, p. 1945, 1983.
- [7] M. Battaglia, A. Buras, P. Gambino, A. Stocchi, D. Abbaneo, A. Ali, P. Amaral, V. Andreev, M. Artuso, E. Barberio *et al.*, "The CKM matrix and the unitarity triangle," *arXiv preprint hep-ph/0304132*, 2003.
- [8] M. Röhrken, "Time-dependent CP-violation measurements," 2014.
- [9] 03.06.2016. [Online]. Available: http://ckmfitter.in2p3.fr/www/results/plots_moriond12/png/rhoeta_large.png
- [10] T. Abe, I. Adachi, K. Adamczyk, S. Ahn, H. Aihara, K. Akai, M. Aloï, L. Andricek, K. Aoki, Y. Arai *et al.*, "Belle II technical design report," *arXiv preprint arXiv:1011.0352*, 2010.
- [11] S. Neubauer, "Search for $B \rightarrow K^{(*)} \nu \bar{\nu}$ Decays Using a New Probabilistic Full Reconstruction Method," 2011. [Online]. Available: <https://ekp-invenio.physik.uni-karlsruhe.de/record/48215>
- [12] T. Keck, "The full event interpretation for Belle II," 2014, Karlsruher Institut für Technologie (KIT), Masterarbeit, 2014. [Online]. Available: <https://ekp-invenio.physik.uni-karlsruhe.de/record/48602>
- [13] T. M. Mitchell, "Machine learning," 1997.
- [14] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," *PoS*, vol. ACAT, p. 040, 2007.
- [15] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.
- [16] V. Blobel and E. Lohrmann, "Statistische und numerische Methoden der Datenanalyse," 1998. [Online]. Available: <https://books.google.de/books?id=NclxajwHOZIC>

-
- [17] K.-F. Chen *et al.*, “Time Dependent CP Asymmetries in $B^0 \rightarrow \eta' K^0, \phi K^0, K_S^0 K_S^0 K_S^0$ and $B^0 \rightarrow J/\psi K^0$ decays with 532 Million $B\bar{B}$ Pairs,” *Internal Belle Note 929*, 2006.
 - [18] H. Kakuno, K. Hara, B. Casey, K.-F. Chen, H. Hamasaki, M. Hazumi, T. Iijima, N. Katayama, T. Okabe, Y. Sakai *et al.*, “Neutral b flavor tagging for the measurement of mixing-induced cp violation at belle,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 533, no. 3, pp. 516–531, 2004.
 - [19] M. Gelb, “Neutral B Meson Flavor Tagging for Belle II,” 2015.
 - [20] 20.06.2016. [Online]. Available: <https://root.cern.ch/>
 - [21] 20.06.2016. [Online]. Available: <http://pandas.pydata.org/>
 - [22] 20.06.2016. [Online]. Available: <http://jupyter.org/>
 - [23] A. Bevan, B. Golob, T. Mannel, S. Prell, B. Yabsley, H. Aihara, F. Anulli, N. Arnaud, T. Aushev, M. Beneke *et al.*, “The physics of the b factories,” *The European Physical Journal C*, vol. 74, no. 11, pp. 1–928, 2014.
 - [24] K. S. Hiroshi Nakano, Akimasa Ishikawa, “Ks selection with neurobayes and nisksfinder class,” *Internal Belle Note 1253*, 2012.
 - [25] B. Kronenbitter, “Measurement of Polarization and CP violation in $B^0 \rightarrow D^{*+} D^{*-}$ Decays with the Belle detector,” 2011.
 - [26] M. Pivk and F. Le Diberder, “splot: a statistical tool to unfold data distributions, 424 nucl. instrum. meth. a555 (2005) 356,” *arXiv preprint physics/0402083*, vol. 425.
 - [27] H. Tajima, H. Aihara, T. Higuchi, H. Kawai, T. Nakadaira, J. Tanaka, T. Tomura, M. Yokoyama, M. Hazumi, Y. Sakai *et al.*, “Proper-time resolution function for measurement of time evolution of B mesons at the KEK B-Factory,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 533, no. 3, pp. 370–386, 2004.
 - [28] P. Urquijo, “Flavor Physics lecture slides,” 2014.
 - [29] L. Šantelj, Y. Yusa, A. Abdesselam, I. Adachi, H. Aihara, S. Al Said, D. Asner, V. Aulchenko, T. Aushev, R. Ayad *et al.*, “Measurement of time-dependent CP violation in $B \rightarrow \eta' K_S^0$ decays,” *Journal of High Energy Physics*, vol. 2014, no. 10, pp. 1–20, 2014.
 - [30] 20.06.2016. [Online]. Available: <https://kds.kek.jp/indico/event/21086/contribution/1/material/slides/0.pdf>

Appendix

A. Belle Systematics

		S	A
wrong_tag	...	0.012180	0.005882
physics	...	0.001084	0.000760
resolution	...	0.049342	0.015523
background_shape	...	0.006984	0.009726
signal_fraction	...	0.057060	0.048500
fit bias	...	0.024437	0.013244
vertex recon	...	0.010444	0.020261
tag side int	...	0.001000	0.042000
TOTAL	...	0.081216	0.071224

Figure A.1.: Systematic Uncertainties on the CP-parameters determined by the most recent Belle I analysis [17].

B. BDT Training results

All trainings have been performed using a BDT with 1000 trees and 3 layers. The signal ratio was chosen to be $sf = 0.5$ during the trainings. The number of events used for the training was limited by the number of background events passing the corresponding pre-cuts. If the number of signal and background events was not equal, the *signalRatio* tool provided by the TMVA package automatically scaled down the exceeding events in order to ensure $sf = 0.5$.

B.1. KSBDT Training results

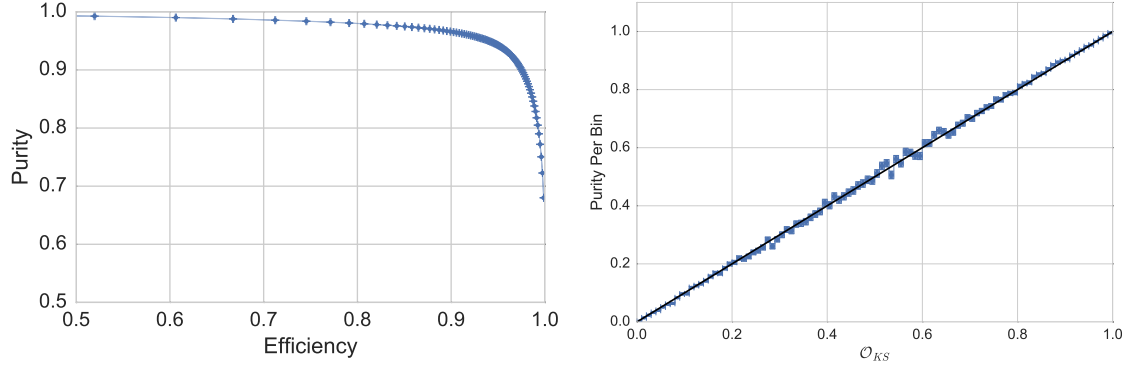


Figure B.2.: *Left*: ROC curve. *Right*: Purity per bin plot.

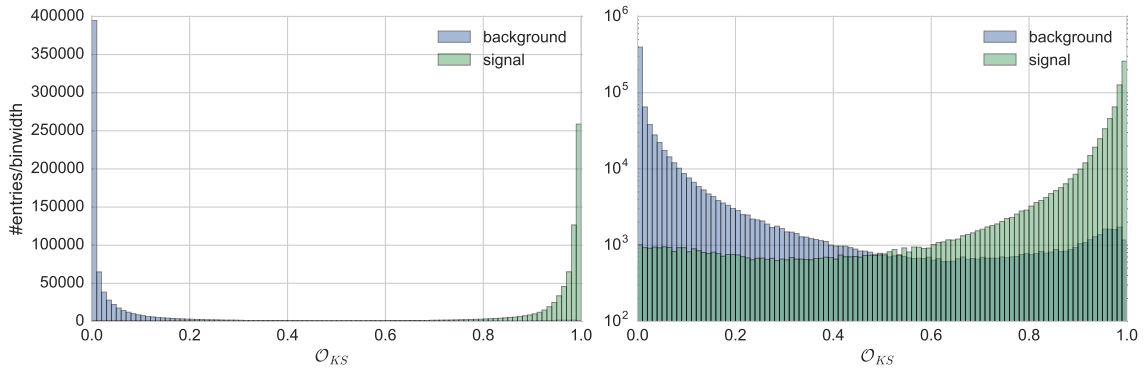


Figure B.3.: *Left*: BDT output. *Right*: BDT output in logarithmic scaling.

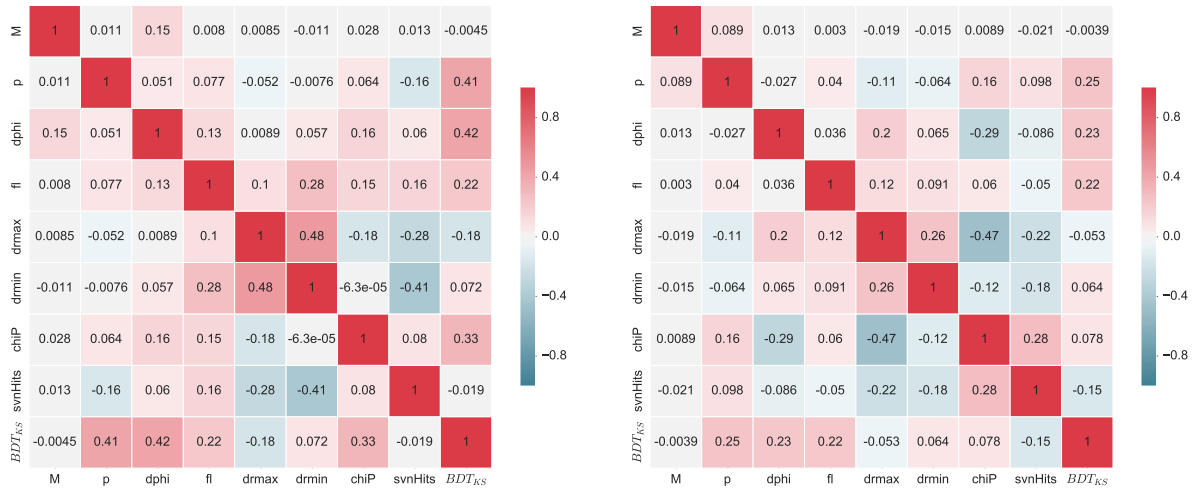
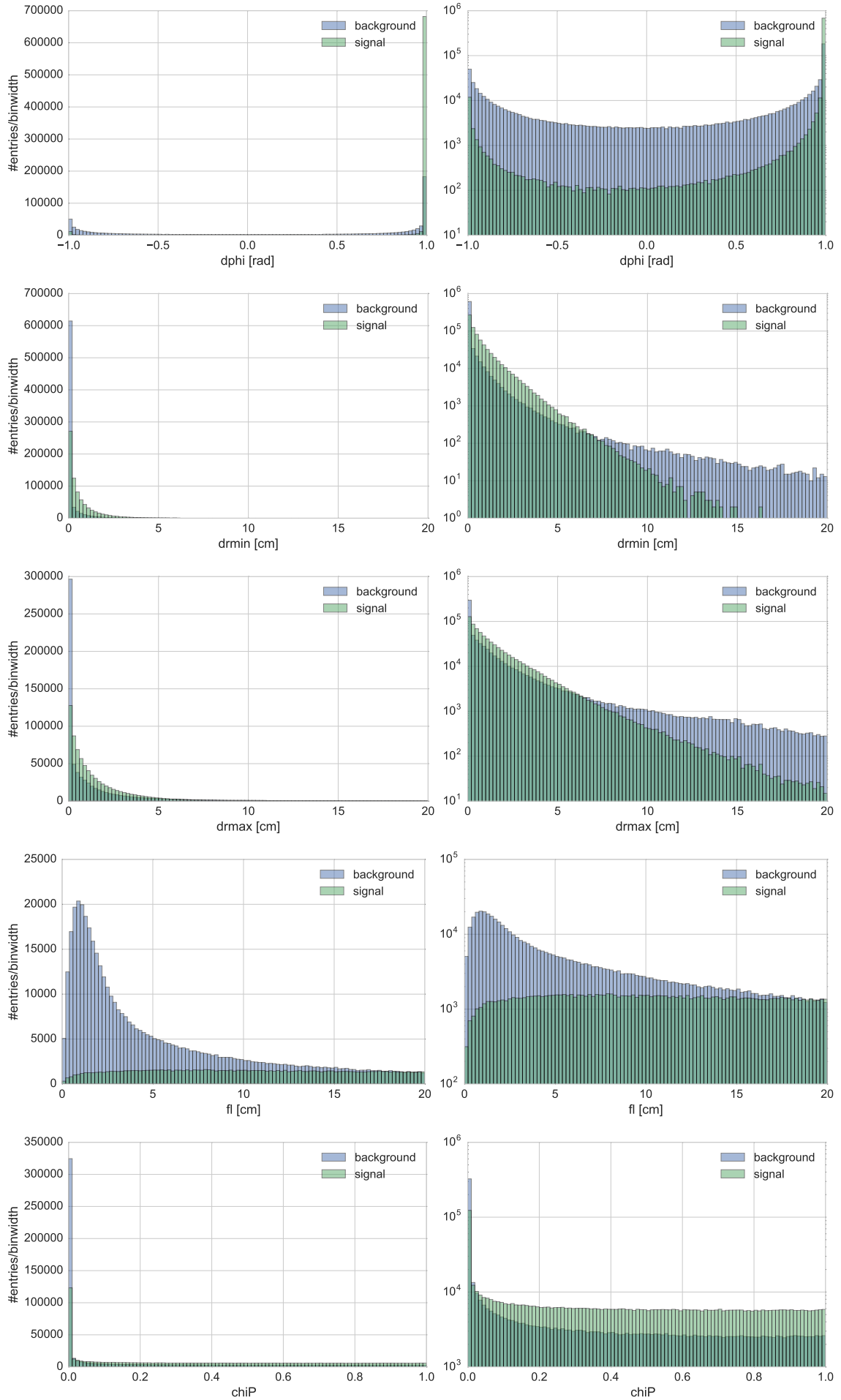


Figure B.4.: *Left*: Correlation heat map for signal events. *Right*: Correlation heat map for background events.



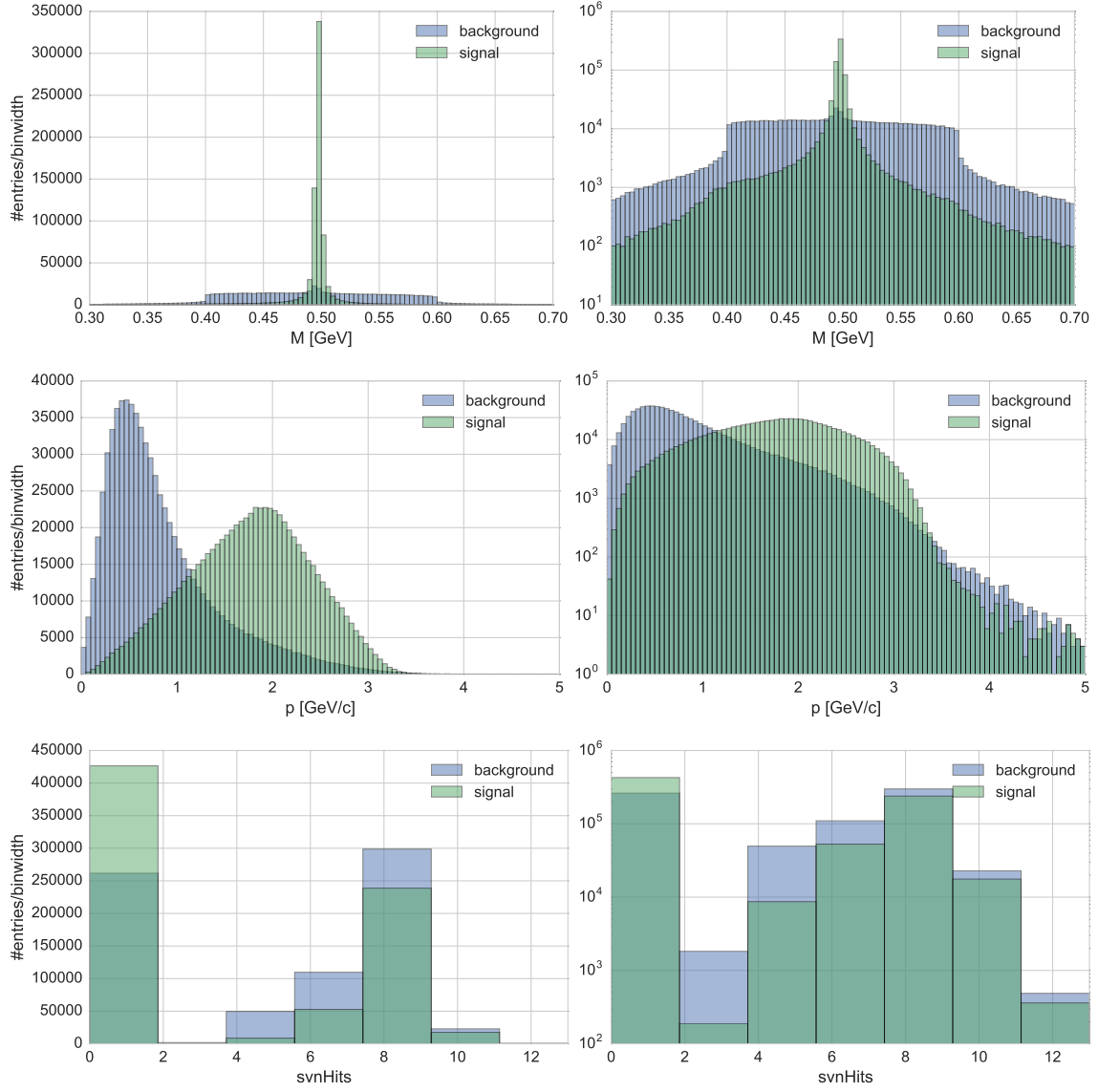
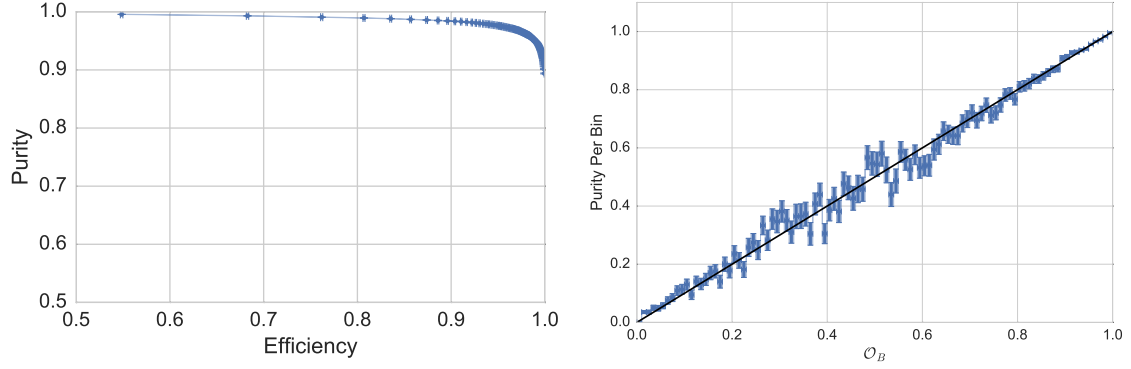
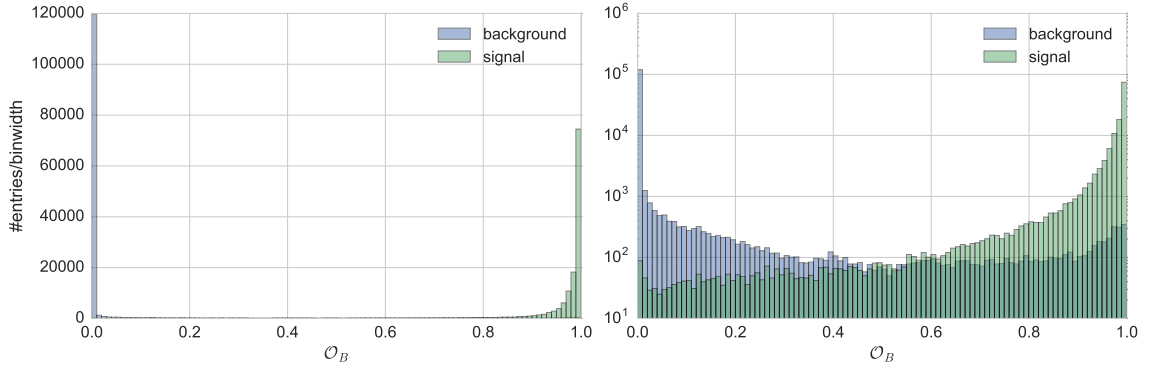
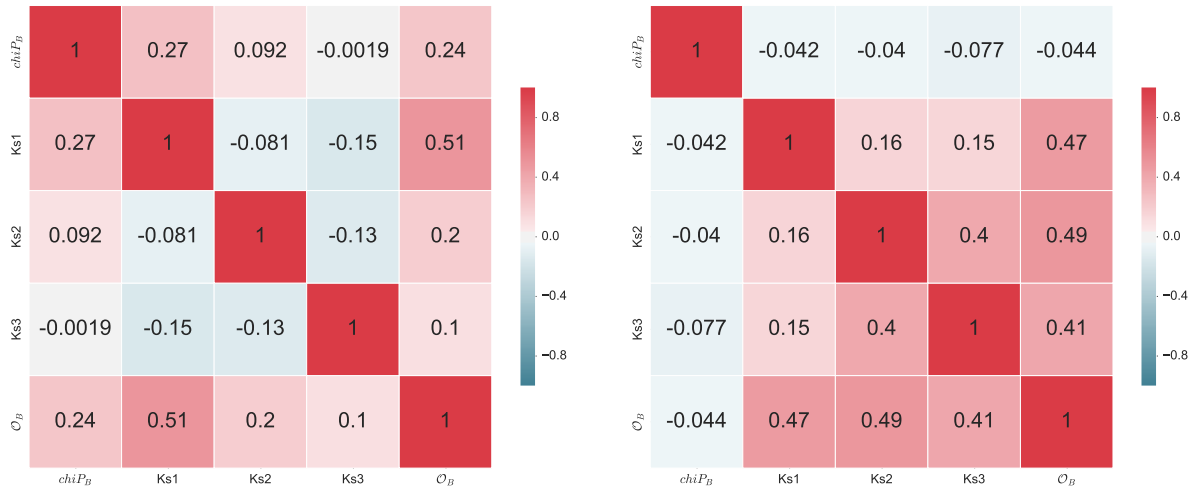


Figure B.5.: *Left:* Input variables. *Right:* Input variables in logarithmic scaling.

B.2. BBDT Training results

Figure B.6.: *Left*: ROC curve. *Right*: Purity per bin plot.Figure B.7.: *Left*: BDT output. *Right*: BDT output in logarithmic scaling.Figure B.8.: *Left*: Correlation heat map for signal events. *Right*: Correlation heat map for background events.

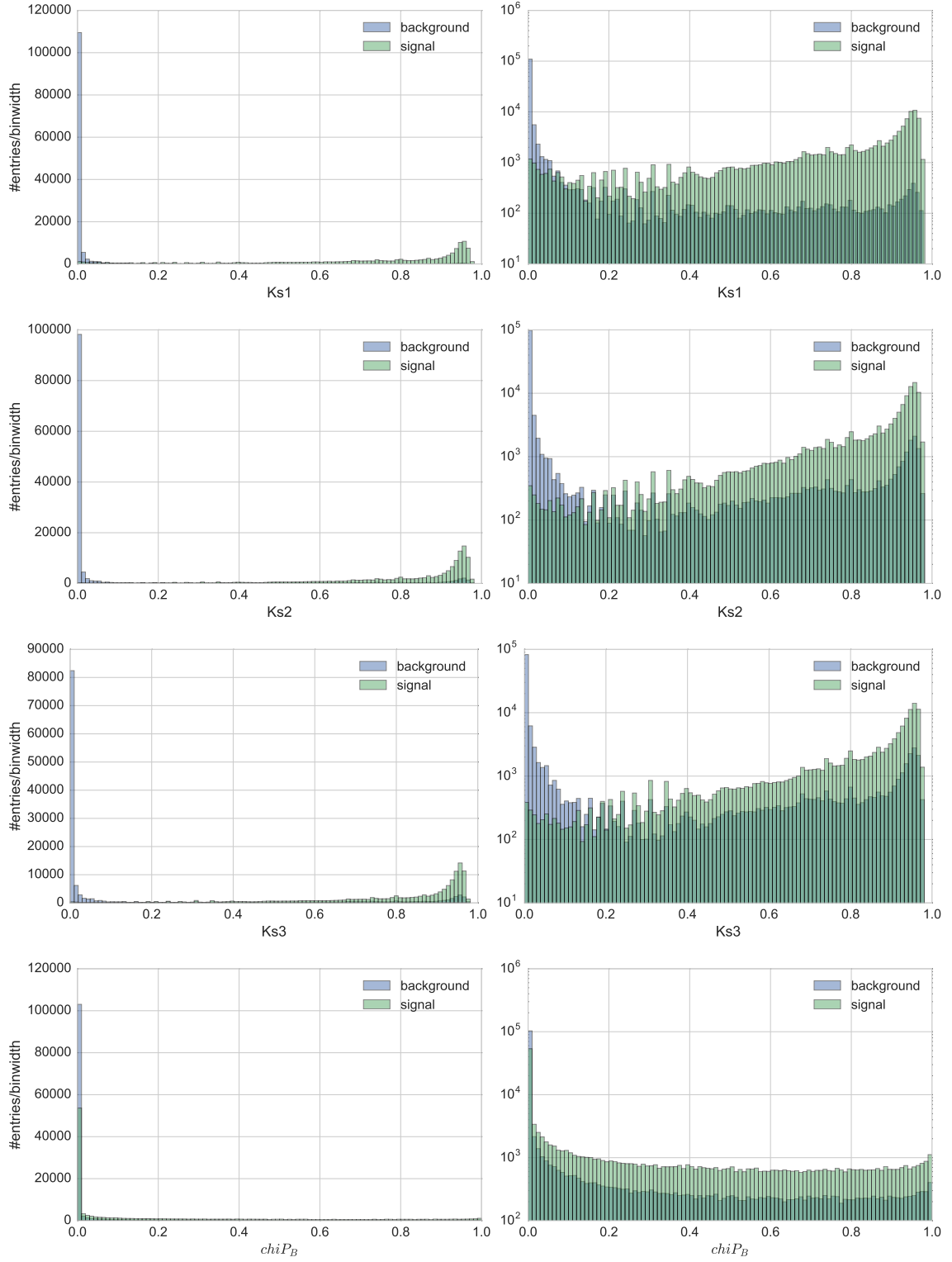
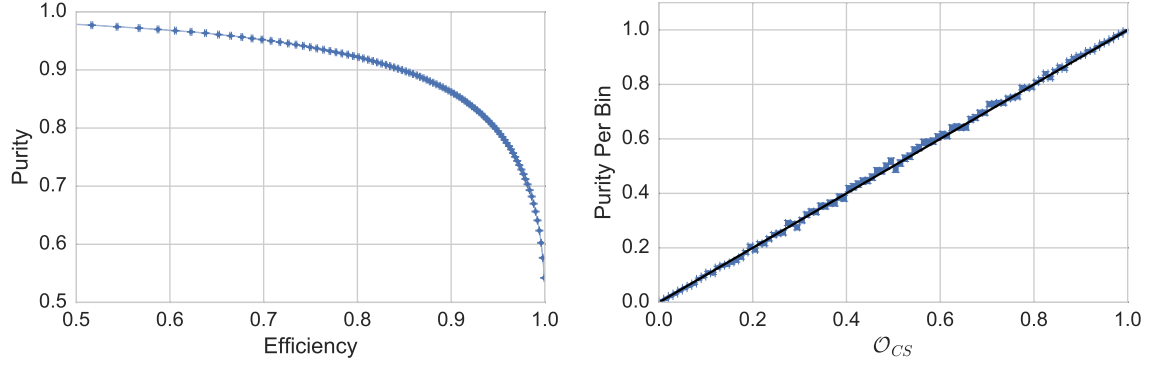
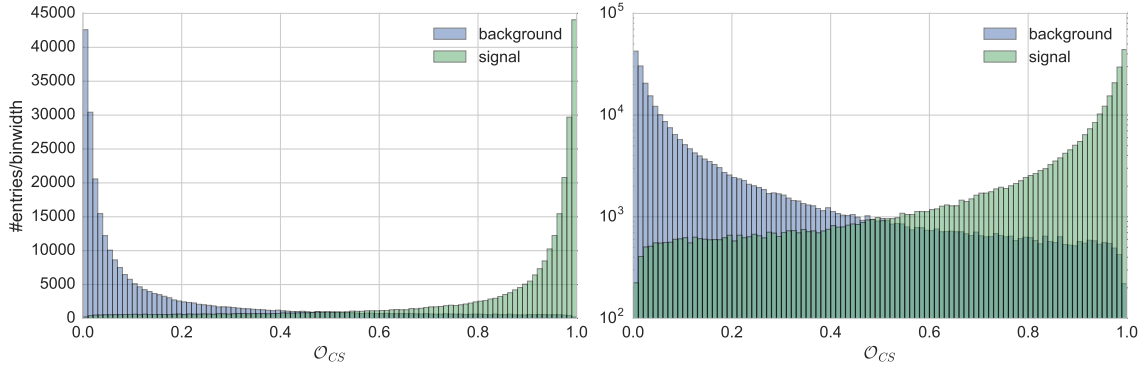
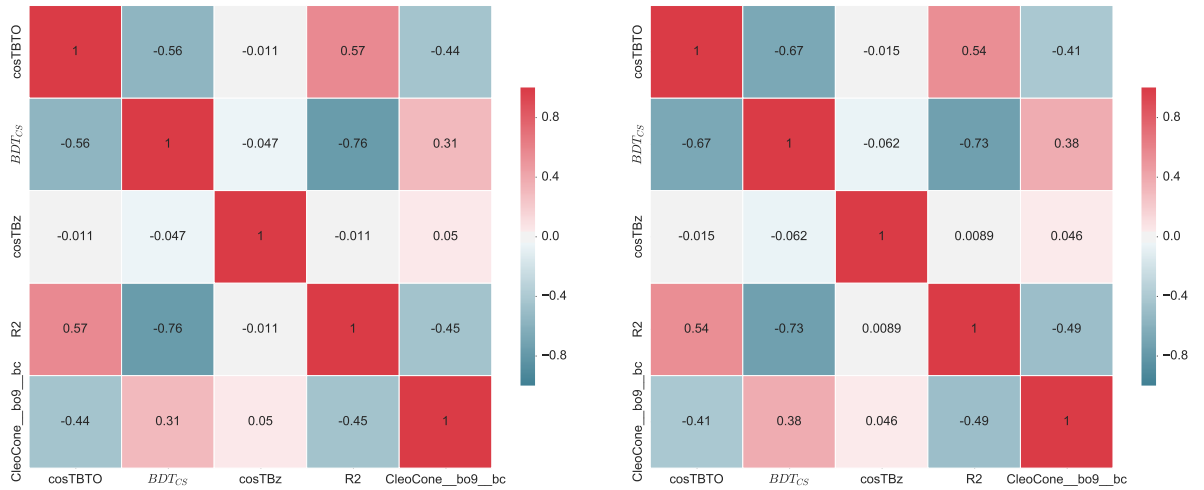


Figure B.9.: *Left*: Input variables. *Right*: Input variables in logarithmic scaling.

B.3. CSBDT Training results

Figure B.10.: *Left*: ROC curve. *Right*: Purity per bin plot.Figure B.11.: *Left*: BDT output. *Right*: BDT output in logarithmic scaling.Figure B.12.: *Left*: Correlation heat map for signal events. *Right*: Correlation heat map for background events.

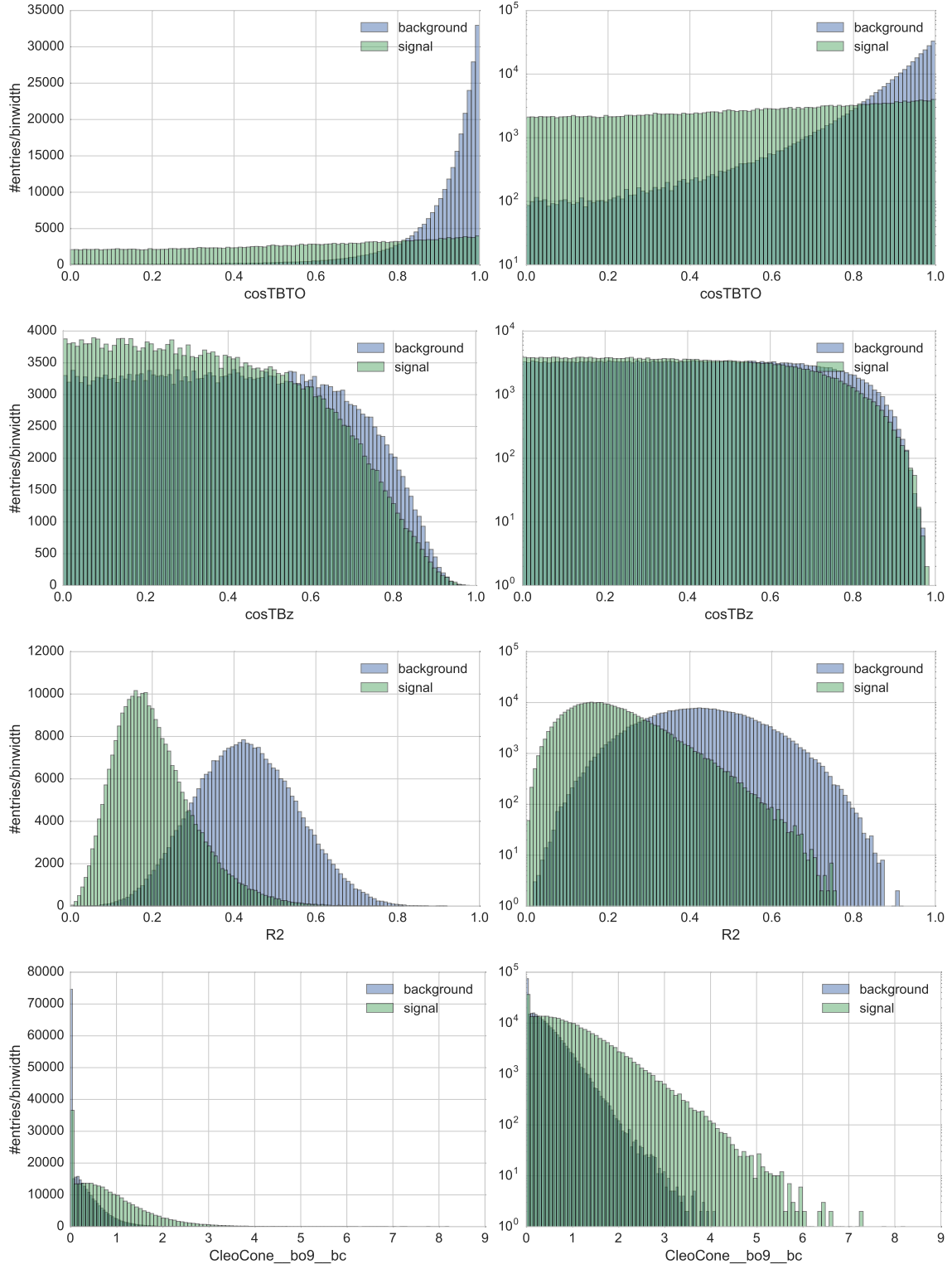


Figure B.13.: Note that the Continuum Suppression training is performed using 30 input variables, but only a selection of 4 variables is shown here. *Left:* Input variables. *Right:* Input variables in logarithmic scaling.

B.4. FSBDT Training results

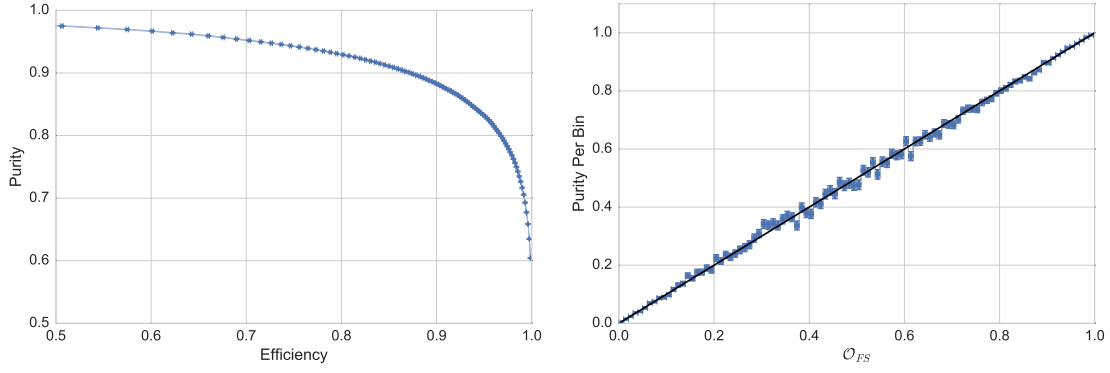


Figure B.14.: *Left*: ROC curve. *Right*: Purity per bin plot.

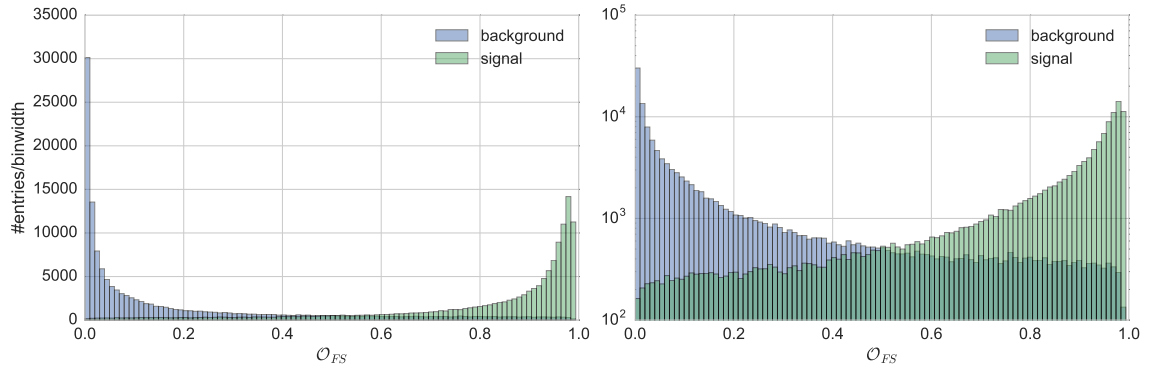


Figure B.15.: *Left*: BDT output. *Right*: BDT output in logarithmic scaling.

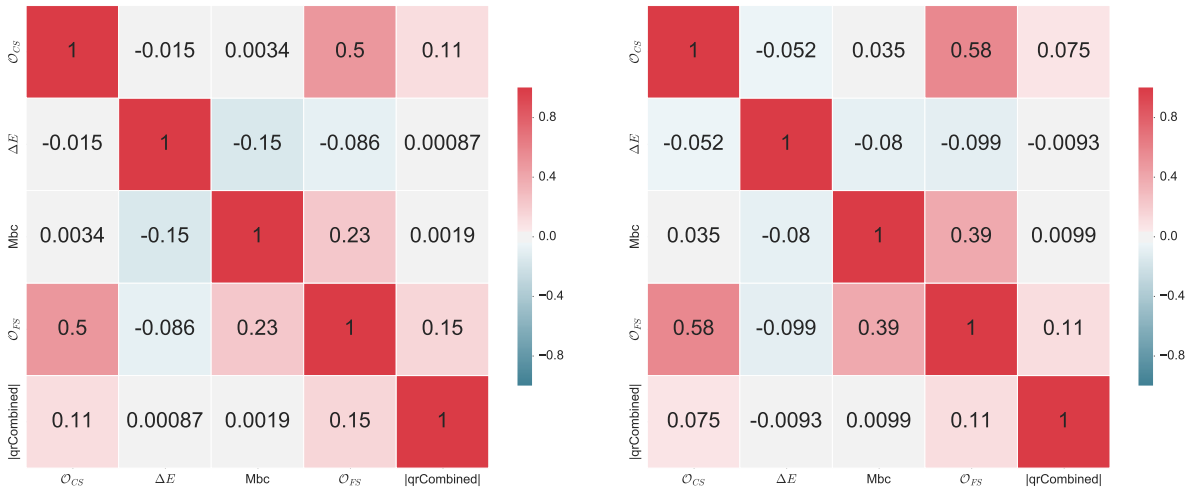


Figure B.16.: *Left*: Correlation heat map for signal events. *Right*: Correlation heat map for background events.

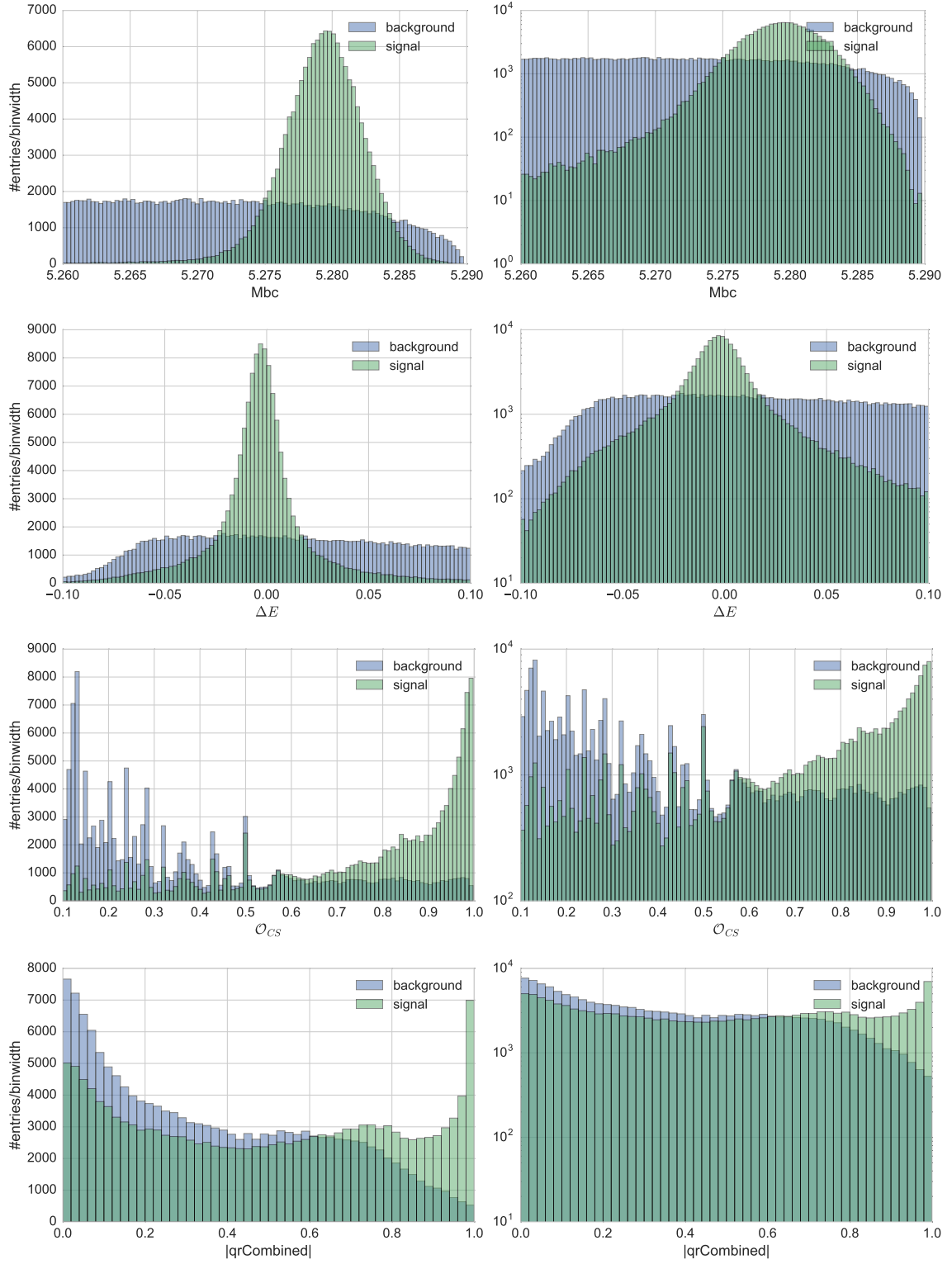


Figure B.17.: *Left:* Input variables. *Right:* Input variables in logarithmic scaling.

Danksagung

Hiermit danke ich Herrn Prof. Dr. Michael Feindt für die Betreuung dieser Arbeit. Mit seiner inspirierenden Art konnte er mich stets für Themen im Zusammenhang mit der Arbeit begeistern.

Weiter möchte ich Herrn Prof. Dr. Thomas Müller für die bereitwillige Übernahme des Korreferats danken.

Besonderer Dank gilt auch meinen Betreuern Dr. Pablo Goldenzweig und Dr. Martin Heck. Ohne sie wäre diese Arbeit nicht möglich gewesen.

Mein Dank gilt ebenso der gesamten Belle Gruppe des EKP für die angenehme Arbeitsatmosphäre. Bei Fragen und Problemen konnte ich stets auf die Hilfsbereitschaft meiner Kollegen zählen.

Eigenständigkeitserklärung

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und nur die angegebenen Hilfsmittel verwendet zu haben.

Paul Jäger

Karlsruhe, den 30.06.2016