Karlsruher Institut für Technologie

ETP-KA/2017-31

Machine learning algorithms for the Belle II experiment and their validation on Belle data

Thomas Keck

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

von der Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

M.Sc. Thomas Keck

aus Freiburg

Tag der mündlichen Prüfung:	15. Dezember 2017
Referent:	Prof. Dr. Michael Feindt
Korreferent:	Prof. Dr. Günter Quast

Contents

Chapter 1

Introduction

Physics deals with the development of quantitative models that describe nature. The most fundamental and precise models at present are: general relativity, which describes gravity as an interaction between the curvature of space-time and matter; and the Standard Model of particle physics (SM), which describes matter and the remaining electromagnetic, weak and strong interaction in the form of relativistic quantum fields.

The current high-energy physics (HEP) research focuses on the precise determination of the 19 free parameters of the SM and the search for new physics phenomena beyond the SM.

The Belle II experiment is part of this effort. Like its predecessor Belle, it is located at the SuperKEKB electron-positron collider in Tsukuba, Japan. It is designed to perform a wide range of high-precision measurements in all fields of heavy flavour physics, including: B meson decays; B_s^0 meson decays; charm physics; τ lepton physics; hadron spectroscopy; and pure electroweak measurements. These measurements will constrain the parameter space of the SM as well as some of its extensions.

This work focuses on the development of software, in particular machine learning algorithms, to advance scientific progress, and to enable and improve a wide range of physics measurements at Belle II. This thesis summarizes my contributions to the Belle experiment and its successor the Belle II experiment.

Four major topics are covered. The conversion of the data recorded by the Belle experiment into the new data-format used by Belle II in Chapter 2. The integration of state-of-the-art machine learning algorithms and novel data analysis techniques into the Belle II Software Framework (BASF2) in Chapter 3. The development of the `Full Event Interpretation` exclusive tagging algorithm, which is unique to the Belle II experiment in Chapter 4. And the validation of the entire analysis software stack using the benchmark measurement of the branching fraction of the rare decay $B \to \tau\nu_\tau$ in Chapter 5.

Chapter 2

From Belle to Belle II

During this thesis, the full $\Upsilon(4S)$ dataset of the Belle experiment and large amounts of the available Monte Carlo data were converted into the new data-format used by Belle II. Therefore, it is possible to evaluate the reliability and performance of the newly developed analysis methods, in particular data-driven techniques, before a comparable dataset from the Belle II experiment is available.

In the following I give a brief overview of the Belle experiment (Section 2.1), its successor the Belle II experiment (Section 2.2) and describe the technical aspects of converting the Belle dataset into the Belle II data-format (Section 2.3).

2.1. The Belle Experiment

From June 1$^{\text{st}}$ 1999 until June 30$^{\text{th}}$ 2010, the Belle experiment recorded 988 fb^{-1} of data at the KEKB asymmetric e^+e^- collider [1].

This summary of the KEKB accelerator and the Belle detector is based on [1] and [2]. Section 2.1.1 and Section 2.1.2 are adapted from my master's thesis [3].

2.1.1. KEKB Accelerator

The KEKB collider was dedicated to B physics and operated in the energy range of the Υ resonances. Its asymmetric beam energies induced a Lorentz boost $\beta\gamma = 0.42$ of the center of mass frame relative to the laboratory system, enabling the precise observation of the time evolution of B meson decays. During its runtime between 1998 and 2010 KEKB achieved the highest instantaneous luminosity of $2.1 \cdot 10^{34}\,\text{cm}^{-2}\text{s}^{-1}$ ever achieved by a collider [1, sec. 1.3]. The machine parameters of KEKB are summarized in Table 2.2.

2.1.2. Belle Detector

The sole interaction point (IP) was surrounded by the Belle detector to detect and identify particles produced by the collisions. Like the accelerator, the detector was specifically designed for the precise observation of B meson decays. This includes precise measurement of secondary vertices and good particle identification capabilities.

Figure 2.1 shows a schematic side view of the Belle detector.

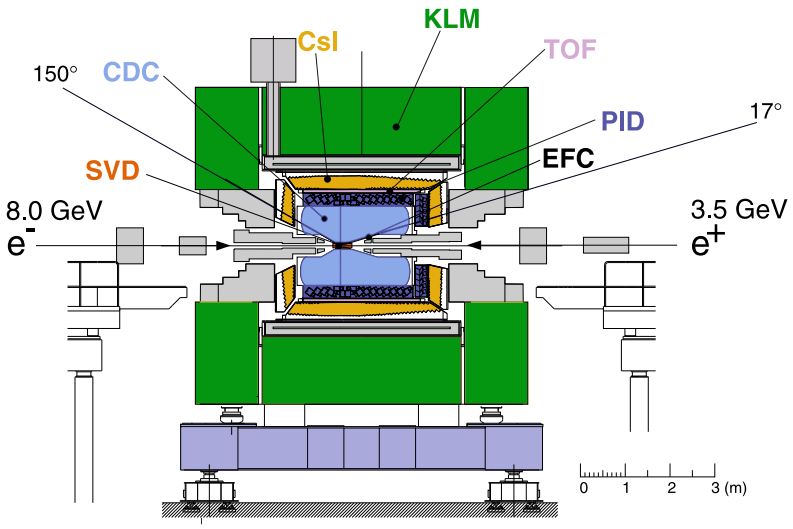


Figure 2.1.: Side view of the Belle detector. Adapted from [2].

Going outwards from the interaction point the Belle detector consisted of:

- a double-walled Beryllium beam pipe with a radius of 20 mm cooled by He gas;
- radiation-hard Bismuth Germanate crystals used as an extreme forward calorimeter (EFC) as well as a beam and luminosity monitor;
- four layers of double-sided Silicon strip detectors (SVD) for precise vertex detection;
- a central drift chamber (CDC), which measured momentum and energy loss of charged particles;
- an Aerogel Cherenkov counter (ACC) system for particle identification (PID);
- a time-of-flight (TOF) detector system with plastic scintillation counters;
- a segmented array of CsI(Tl) crystals with silicon photodiode readout for electromagnetic calorimetry (ECL);

Table 2.1.: Summary of the integrated luminosity collected by Belle, broken down by center-of-mass energy. Adapted from [1].

Resonance	On-resonance Luminosity (fb^{-1})	Off-resonance Luminosity (fb^{-1})
$\Upsilon(5S)$	121.4	1.7
$\Upsilon(4S)$ – SVD1	140.0	15.6
$\Upsilon(4S)$ – SVD2	571.0	73.8
$\Upsilon(3S)$	2.9	0.2
$\Upsilon(2S)$	24.9	1.7
$\Upsilon(1S)$	5.7	1.8
Scan > $\Upsilon(4S)$	n/a	27.6

- a superconducting solenoid which provided a homogeneous magnetic field of 1.5 T;
- and an iron support structure, which served as the return path of the magnetic flux and was instrumented with glass-electrode resistive plate counters for K_L and muon detection (KLM).

2.1.3. Recorded Dataset

The Belle dataset is grouped into 31 experiments[1]; each experiment marks the time period between two major shutdowns. Experiment 7 to 27 were recorded with the first Silicon Vertex Detector (SVD1). The remainder was recorded with the second (SVD2) detector and, in addition, was reprocessed in 2009 with an improved version of the reconstruction software. Each experiment is further subdivided into a number of runs.

Most of the data was recorded at the center-of-mass energy of the $\Upsilon(4S)$ resonance. In addition, data was also recorded at the $\Upsilon(1S)$, $\Upsilon(2S)$, $\Upsilon(3S)$ and $\Upsilon(5S)$ resonances. Apart from that, off-resonance data 60 MeV below the resonances was collected to estimate the continuum background from data instead of Monte Carlo simulation. Table 2.1 shows a summary of the integrated luminosity collected by Belle, broken down by the center-of-mass energy.

The raw data coming from the detector was calibrated, reconstructed and stored on tape using `PANTHER`-based[2] data summary tape (DST) files. After each experiment

[1] Enumerated from 7 to 73 using only odd numbers and skipping the numbers 29, 57 and 59 for reasons unknown to the author.

[2] The `PANTHER` format consists of tables, which are compressed by the zlib library. The table formats are defined by `ASCII` header files.

the calibration constants were recomputed by detector experts or computed directly from data, and stored in the Belle Condition Database (based on `PostgreSQL`). Finally, the data of the completed experiment was reprocessed and stored in a compact form called mDST files[3].

For each experiment, ten times the real integrated luminosity in $b\bar{b}$ events and six times that in continuum events were simulated using `EvtGen` and `GEANT3`, and reconstructed with the same software as was used for the detector data.

2.2. The Belle II Experiment

This summary of the SuperKEKB accelerator and the Belle II detector is based on the detailed description in the technical design report [4] and is an updated and condensed version of a summary presented in my master's thesis [3].

2.2.1. SuperKEKB Accelerator

The accelerator was shut down in June 2010 and upgraded to SuperKEKB to increase the instantaneous luminosity to $8 \cdot 10^{35}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, which is 40 times the peak instantaneous luminosity of KEKB [5, sec. 2].

The higher instantaneous luminosity is obtained by adopting the Nano-Beam scheme [5, sec. 2], which requires a larger crossing angle to fit the final focusing magnets [4, sec. 3.1]. Moreover, the beam current in both rings is doubled. The beam energy asymmetry was reduced to mitigate the shortened beam lifetime due to the Touschek effect[4]. In consequence the Lorentz boost is reduced to $\beta\gamma = 0.28$. The relevant machine parameters of KEKB and SuperKEKB are summarized in Table 2.2.

2.2.2. Belle II Detector

The original detector is currently upgraded to match the higher instantaneous luminosity of SuperKEKB. The most important objectives for the upgraded detector are higher physics and background rate tolerance, better physics performance despite smaller Lorentz boost, and improved radiation hardness [5].

Figure 2.2 shows the Belle II detector.

Going outwards from the interaction point the Belle II detector consists of:

[3] A reduced and compressed form of the data summary tape files.

[4] The Touschek effect is a loss mechanism due to large angle coulomb scattering inside a bunch.

Table 2.2.: Achieved parameters of KEKB and design parameters of SuperKEKB [4].

Machine Parameter	KEKB		SuperKEKB	
	e^-	e^+	e^-	e^+
Beam current (A)	1.64	1.19	3.60	2.61
Energy (GeV) (E_{HER}/E_{LER})	8.0	3.5	7.0	4.0
β_y^* (mm)	5.9	5.9	0.27	0.41
Crossing angle (mrad)	22		83	
Beam lifetime (min)	200	150	10	10
Luminosity ($10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$)	2.11		80	

- a double-walled Beryllium beam pipe with a radius of 12 mm cooled by paraffin;
- a pixel detector based on the DEPFET[5] technology (PXD) for precise vertex detection;
- four layers of double-sided silicon strip detectors covering the full $17° - 150°$ acceptance of the Belle II detector (SVD) to extrapolate the tracks reconstructed in the CDC to the PXD and to reconstruct low-momentum tracks;
- a central drift chamber (CDC), which measures momentum and energy loss of charged particles, and provides a fast trigger signal for the Level-1 (L1) trigger system;
- a proximity-focusing[6] Aerogel Ring-Imaging Cherenkov detector (ARICH) for particle identification (PID) in the forward end-cap;
- a time-of-propagation counter (TOP) in the barrel region using an array of 16 quartz bars between the outer CDC cover and the calorimeter's inner surface providing PID information and a timing signal with a resolution of a few nanoseconds to the trigger system;
- a segmented array of CsI(Tl) crystals in the barrel region and pure CsI crystals in the end-caps readout photo-diodes for electromagnetic calorimetry (ECL);
- a superconducting solenoid which provides a homogeneous magnetic field of 1.5 T;
- and an iron support structure, which serves as the return path of the magnetic flux, and is instrumented with glass-electrode resistive plate counters (RPC) in the barrel region and scintillator strip in the end-caps for K_L and muon detection (KLM).

[5] DEPleted Field Effect Transistor.

[6] An increasing refractive index is used to reduce the spread of the ring image due to emission point uncertainty.

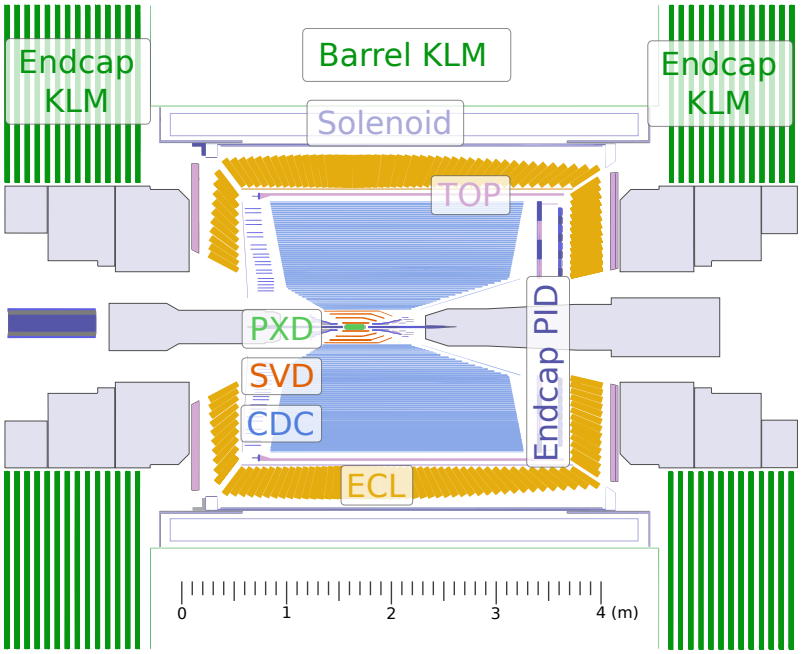


Figure 2.2.: Side view of the upgraded Belle II detector [6].

2.2.3. Anticipated Dataset

By 2025, Belle II will record 50 ab^{-1} of data, which corresponds to 50 times the integrated luminosity of Belle. The current (March 17th 2017) luminosity projection is shown in Figure 2.3.

2.2.3.1. Data Acquisition

Belle II uses a two-level trigger system, with an FPGA-based Level-1 (L1) trigger decision and a high level trigger (HLT) farm.

At the design luminosity the nominal average L1 trigger rate is expected to be up to 30 kHz. For each trigger signal the data is read out by the data acquisition system from the detector front-end electronics. The high level trigger farm reconstructs the event and performs a physics-level event selection using the event data from all sub-detectors. The final rate of raw events written to the storage is expected to be between 6 kHz and 10 kHz. The HLT consists of multiple Linux-based PC clusters and runs the `Belle II Analysis Software Framework` (`BASF2`) [8].

2.2.3.2. Offline Reconstruction and Monte Carlo Production

The same software framework is used in offline reconstruction, Monte Carlo production, and physics analysis. After machine-dependent calibration parameters are

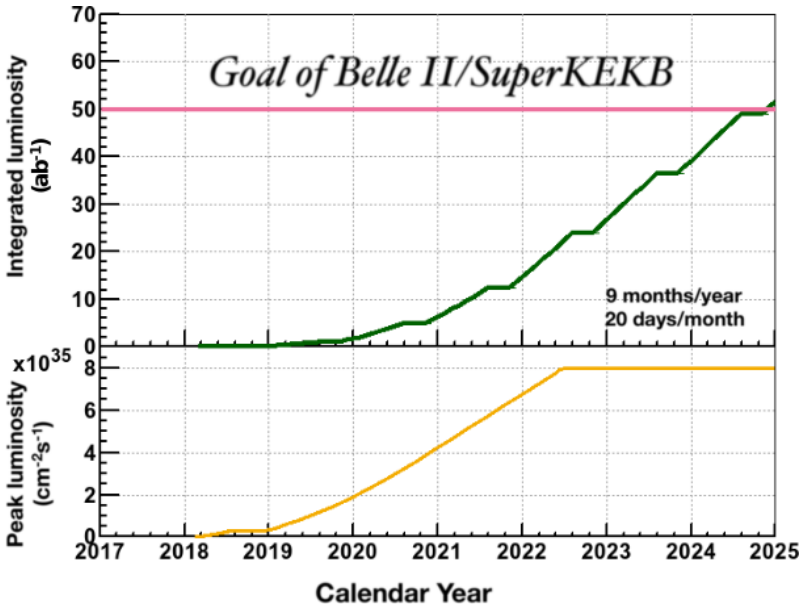


Figure 2.3.: SuperKEKB integrated and instantaneous luminosity projection [7] (March 17$^{\text{th}}$ 2017).

determined, the raw data is reconstructed and stored at the KEK computing center. Monte Carlo production and reconstruction will be distributed to data centers around the world. The reconstructed information is stored in `ROOT`-based mDST files.

2.3. Belle to Belle II dataset

In the above discussion of the recorded Belle and anticipated Belle II dataset, four levels of data processing can be distinguished:

1. **online reconstruction** – the read-out of the detector and the trigger system, producing the **raw-data** (DST files);
2. **offline reconstruction** – cluster reconstruction, track finding and fitting, producing the **mDST data**;
3. **mDST analysis** – creation of final state particle hypotheses, reconstruction of intermediate particle candidates and vertex fitting, producing **flat n-tuples**;
4. and **n-tuple analysis** – fit to theoretical predictions in order to extract interesting observables, producing **scientific papers**.

Converting the raw-data is in principle possible, but the differences between the Belle and Belle II detector render this a difficult and ill-defined task. While this would

allow for the validation of the Belle II reconstruction software (e.g. the track finding and fitting algorithms) on Belle data, this would be only of limited use due to the vastly different expected background and the availability of data from cosmic runs.

The Belle to Belle II dataset conversion converts the Belle mDST data, which contains mostly detector independent objects like tracks and energy clusters, into the new mDST format used by `BASF2`. This enables the validation of the Belle II analysis software, and (re-)production of Belle measurements using the improved software. Figure 2.4 displays an overview of `BASF2` including the conversion path presented in this work.

By comparing the original Belle results, the Belle results obtained from converted data in `BASF2`, and Belle II sensitivity studies on Belle II Monte Carlo, it is possible to assign improvements in the sensitivity and occurring inconsistencies to the analysis and reconstruction algorithms, separately.

2.3.1. Overview

The software responsible for reading in the old `Belle AnalySis Framework` (`BASF`) [9] data-format and representing the data in memory was isolated, cleaned up and compiled into a new library named `belle legacy`. A new package was introduced in `BASF2` called `b2bii` (Belle to Belle II). It contains three `BASF2` modules developed with the help of the `belle_legacy` library. The conversion process is visualized in Figure 2.5.

The `B2BIIMdstInput` module opens the `PANTHER`-based Belle mDST files and reads the data event-by-event into the main memory. The data of the current event is represented in the memory by a series of `PANTHER` tables.

The `B2BIIFixMdst` module applies various calibration factors onto the `PANTHER` tables, for instance on the beam-energy, the momenta and error matrices of the fitted tracks, the energy deposition in the ECL, and the particle identification information of the CDC and TOF. It also performs standard cuts to ensure that the selection of the detector data and Monte Carlo events is identical. Finally, π^0 candidates are reconstructed from the γ particle objects and the corrected ECL clusters.

The `B2BIIConvertMdst` module converts the information stored in the Belle `PANTHER` tables and writes it to the Belle II `DataStore`. The beam-energy and IP-profile is collected in the `BASF2` `BeamParameters` object and stored in the condition Database of Belle II.

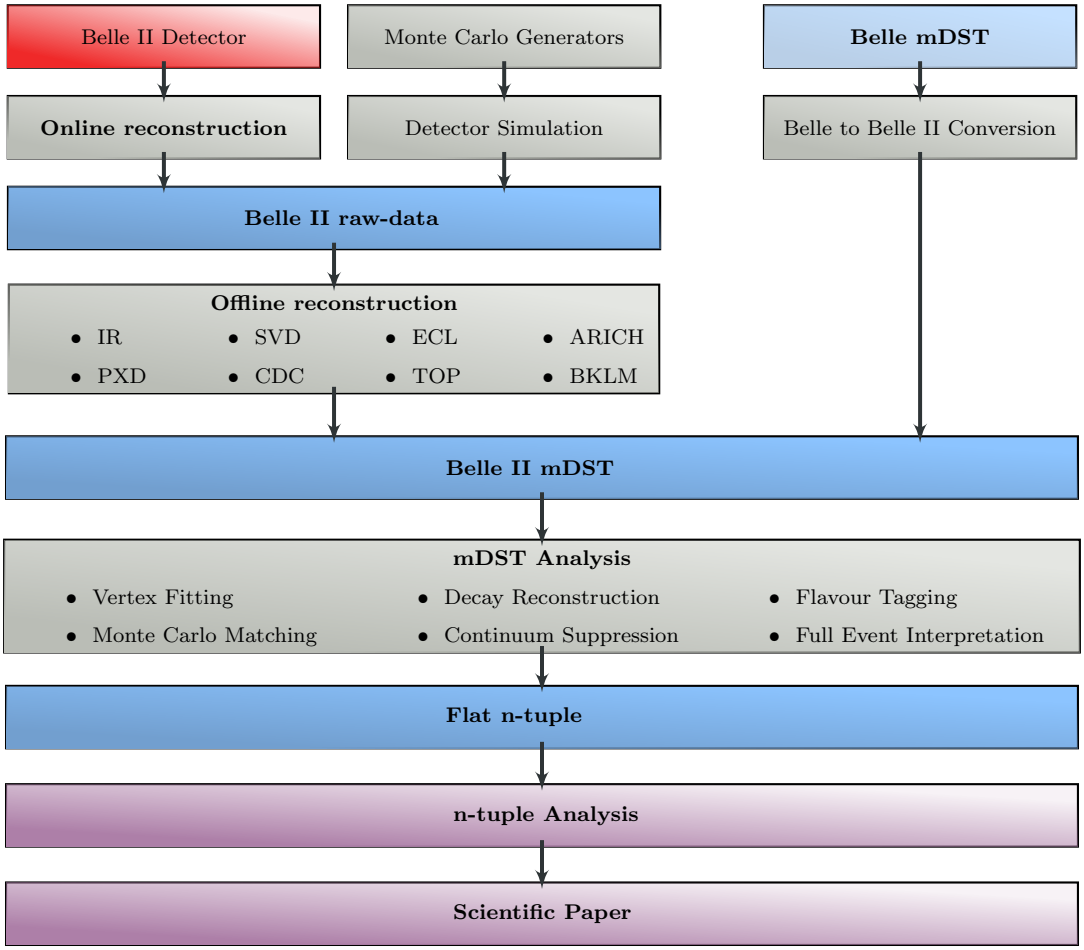


Figure 2.4.: Schematic overview of the data-flow in the Belle II environment. Data is provided by the Belle II detector (red); the MC generators; or Belle mDST files. `BASF2` is responsible for MC generation, detector simulation, online reconstruction, offline reconstruction, mDST analysis and the Belle to Belle II conversion (gray), as well as writing and reading the different data-formats (blue). Analysis-specific user-code is only required during the n-tuple analysis, which extracts the desired physics observables (purple).

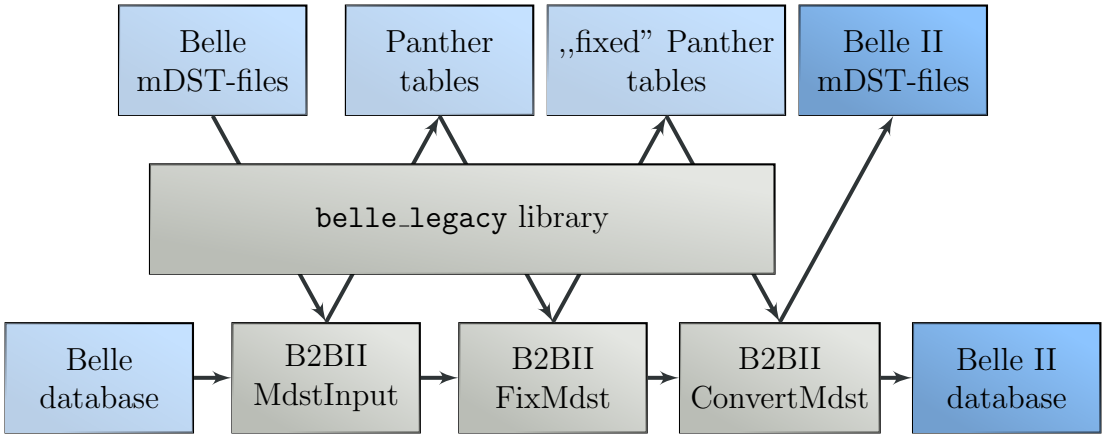


Figure 2.5.: Schematic view of the conversion process of Belle (light blue) to Belle II (blue) mDST files using the `BASF2` modules (gray) provided by the `b2bii` package and the original Belle software provided by the `belle_legacy` library (gray).

2.3.2. Implementation Details

The detailed matching between `PANTHER` tables and corresponding `BASF2` data-objects is shown in Figure 2.6. In the following I describe the conversion process in detail for future reference.

2.3.2.1. Event Information

Event information like the beam energy and position of the interaction point are loaded from the Belle condition database and stored in `BeamParameters` objects that can be uploaded to the Belle II condition database.

`BASF` assumed a constant magnetic field of 1.5 T even outside the detector region. The conversion accounts for this as well by adopting the same convention in `BASF2`, where otherwise a 3D map of the magnetic field is used.

2.3.2.2. Monte Carlo

The Monte Carlo information stored in the `Gen_hepevt` table is converted into an `MCParticleGraph`, hence the fine-grained unified Monte Carlo matching algorithm of `BASF2` can be used, and problems contained in algorithms used by `BASF` are avoided [6, sec. 4.3].

The `Gen_hepevt` table includes special entries for a common mother of beam-background particles (PDG code 911) and for virtual photons (PDG code 0). These

entries are ignored during the conversion, because there are no corresponding concepts in Belle II. For instance, in `BASF2` beam-background is indicated by a motherless Monte Carlo particle.

The original Belle software does not provide Monte Carlo information for `KLMClusters`, following the approach of [10, sec. 5.2] reconstructed K_L^0 are matched to the closest true Monte Carlo K_L^0 within ± 15 degrees in both θ and ϕ.

Furthermore, unlike Belle II Monte Carlo, the Belle Monte Carlo does not provide information on the differentiation between photons generated directly by the fundamental matrix-element calculated by the Monte Carlo generator `evtgen` [11] (hereinafter referred to as gamma) and photons generated afterwards for instance by `PHOTOS` [12] or the simulation [13] (hereinafter referred to as final state radiation) (see [14, Appendix C]). Often a reconstructed particle which misses final state radiation is considered a signal, whereas a missing gamma is considered as a wrong reconstruction. A simple heuristic is applied to distinguish the two cases: Photons from a decay $M \to AB...\gamma$ are flagged as final state radiation, and photons from a decay $M \to A\gamma$ are flagged as gammas. In particular photons from $\pi^0 \to \gamma\gamma$ and $D^* \to D\gamma$ are considered gammas. Other cases like $B \to \mu\nu\gamma$ are regarded by the heuristic as final state radiation and have to be treated by the analyst herself[7].

The official Belle Monte Carlo campaigns produced ten times the real integrated luminosity in $b\overline{b}$ events and six times that in continuum events, however some inconsistencies were encountered during the development of the conversion software, which were fixed if possible: The Monte Carlo campaign deleted the 8 left-most bits of the 32 bit long PDG codes during the Monte Carlo simulation[8]. During the conversion these corrupted PDG codes are restored by matching their lower 24 bit to known PDG codes. In the official Belle Monte Carlo campaign from 2010 for $B \to u\ell\nu$ and other rare B decays, the mass of almost all MC particles is set to zero, which can lead to wrong results if this quantity is used during the analysis. However, this information is redundant since the correct mass of the MC particles can be calculated using either the PDG values or the MC four-momenta.

2.3.2.3. Tracks

The tracking output of `BASF` is transformed and stored into `Track` and associated `TrackFitResult` objects. The transformation is unique but non-trivial because the

[7] In this case, photons from initial and final state radiation are physically indistinguishable, since the corresponding amplitudes interfere. Actually, there is no correct answer to the question of whether the photon is final state radiation or not. Hence, the behavior of the heuristic is not wrong, but probably unexpected by the analyst, because the initial state radiation amplitude dominates in this decay.

[8] `BASF` already implemented a function for recovering the lost bits, but it was apparently not applied.

two experiments employ different track parameterizations and conventions for the reference point of the track.

2.3.2.4. ECL Clusters

The ECL information is stored in the `ECLCluster` object and two `ParticleLists` are filled containing the γ and π^0 candidates created by `B2BIIFixMdst` earlier. The lists are named `gamma:mdst` and `pi0:mdst`, respectively.

2.3.2.5. KLM Clusters

The KLM information is stored in the `KLMCluster` object and a `ParticleList` is filled containing K_L^0 candidates. The list is named `K_L0:mdst`.

2.3.2.6. V0 Objects

The output of the `V0 Finder` is directly transformed into `ParticleList` objects containing candidates for γ, K_S^0 and Λ. The lists are named `gamma:v0mdst`, `K_S0:mdst` and `Lambda0:mdst`, respectively. Additional quality information is stored in the `ExtraInfo` field of the `Particle` objects under the keys `goodKs`, `ksnbVLike`, `ksnbNoLam` and `ksnbStandard`.

2.3.2.7. PID Information

The PID information provided by the different Belle sub-detectors is mapped to similar Belle II sub-detectors, so that the physical meaning of the information is partially preserved. In particular the Belle time-of-flight (TOF) and Aerogel Cherenkov counter (ACC) detectors are mapped to the Belle II time-of-propagation (TOP) and Aerogel ring imaging Cherenkov (ARICH) detectors, respectively.

2.3.2.8. Relations

Finally, some of the created data-objects are related to one another (see 2.6). Hence, `BASF2` relations are created: from the `ECLCluster` to the `MCParticle` and `Track` which are responsible for the creation of the cluster; similarly from the `KLMCluster` to the `ECLCluster` and `Track`; from the `Track` to the `MCParticle` that created it; and from the `Track` to the corresponding `PIDLikelihoods`. Additional relations are created between the `Particle` objects in the created `ParticleLists` and the corresponding `MCParticle` and `PIDLikelihoods`. The links between `TrackFitResults` to `Tracks`; `Tracks` to `Particles`; and `Clusters` to `Particles` are not represented by relations in `BASF2`.

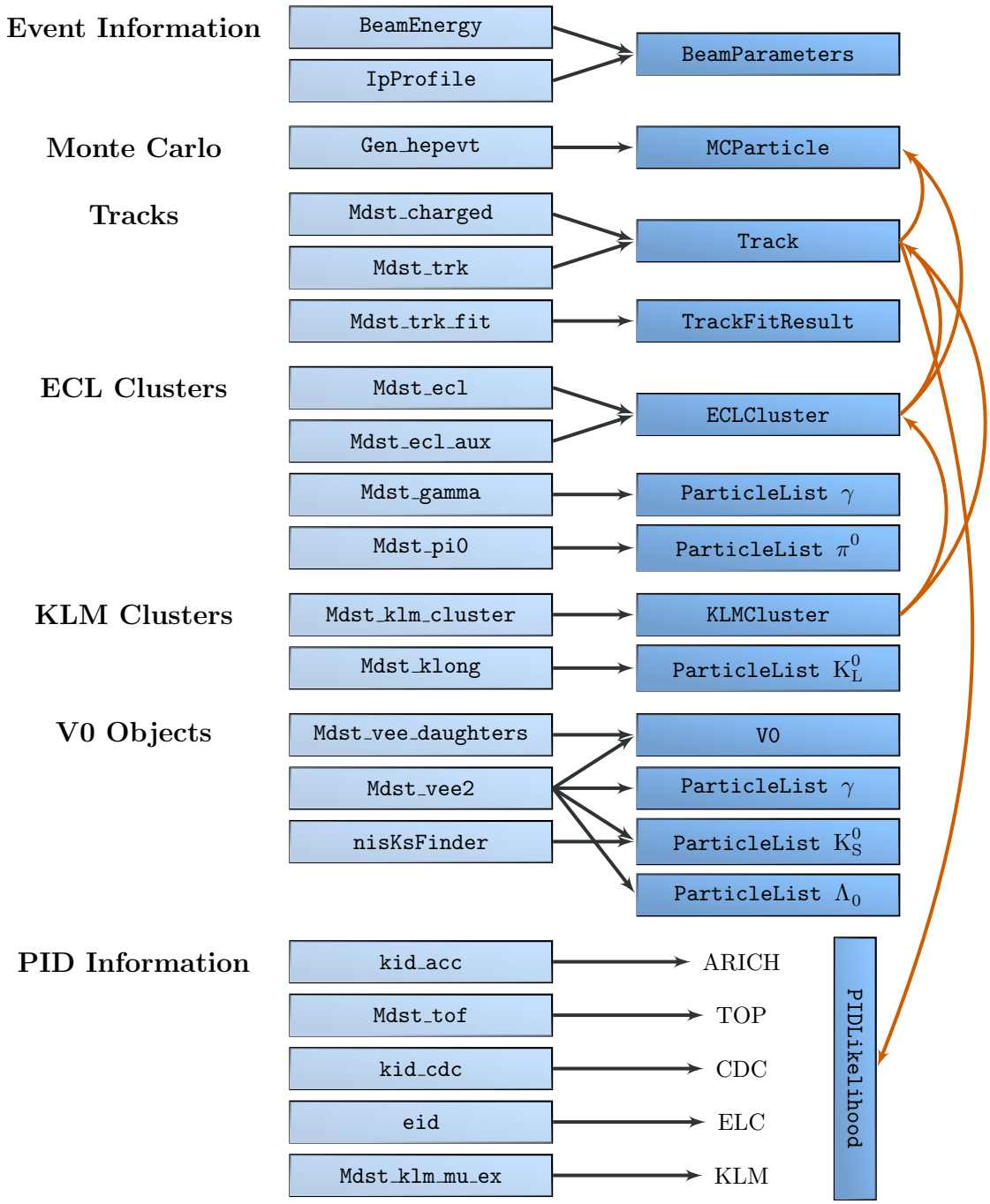


Figure 2.6.: Matching of the Belle PANTHER Tables (light blue) to the Belle II DataStore-objects (blue) and DataStore-relations (orange).

2.3.3. Validation

In order to ensure the correctness of the conversion, a study was performed with Monte Carlo simulated Belle events. For each of the following six physical processes 100000 simulated events were investigated, partitioned according to the relative integrated luminosity of all relevant experiments $7 - 65$:

- generic $\Upsilon(4S)$ events:
  - $\Upsilon(4S) \to B^+B^-$ (evtgen-charged),
  - $\Upsilon(4S) \to B^0\overline{B}^0$ (evtgen-mixed);
- continuum events:
  - $e^+e^- \to c\bar{c}$ (evtgen-charm),
  - $e^+e^- \to q\bar{q}$ (uds) (evtgen-uds),
  - $e^+e^- \to \tau^+\tau^-$ (qed-tautau);
- and signal events $B^+ \to \tau^+\nu_\tau$ (signal).

Furthermore, for each of the following beam conditions 100000 events recorded by the Belle experiment were investigated, partitioned according to the relative integrated luminosity of all relevant experiments $7 - 65$:

- on-resonance events recorded at the center of mass energy of the $\Upsilon(4S)$ resonance;
- and off-resonance events recorded 60 MeV below.

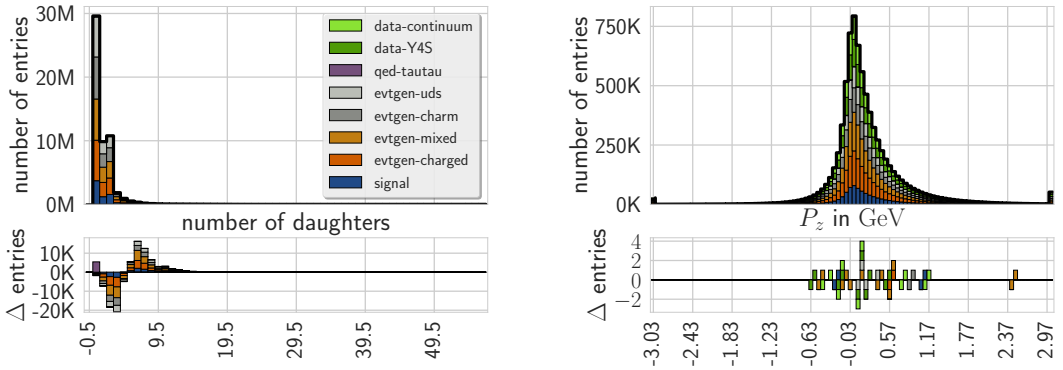
The events were processed with the old `BASF` framework and more than 360 quantities[9] were extracted from the `PANTHER` tables shown in Figure 2.6. The complete list of extracted quantities can be found in Section A.1. Afterwards the events were processed a second time with the new `BASF2` software using the `b2bii` conversion, and same quantities were extracted.

2.3.3.1. Observed Differences

Most quantities do not differ.

The observed differences between `BASF` and `b2bii` were further investigated and either corrected or classified as harmless. Minor differences occur due to small shifts caused by numerical imprecision leading to the migration of events between

[9] For instance: Kinematic quantities like four-momenta, Monte Carlo information, PID information and beam-parameters.



(a) Example of major differences: The number of daughters of the Monte Carlo particle objects is shifted to smaller values because virtual photons are ignored during the conversion.

(b) Example of minor differences: Momentum of Tracks in z direction exhibiting migration during the conversion due to numerical imprecision and special floating point values.

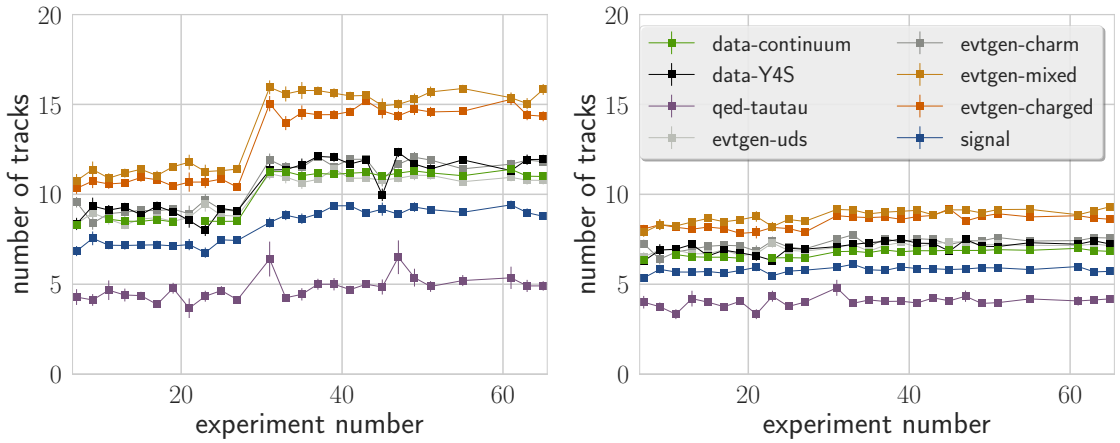
Figure 2.7.: Comparison of `BASF` (Belle) and `b2bii` (Belle II). The leftmost (rightmost) bin represents the underflow (overflow) bin. The upper plots show the superimposed Belle (each component is shown individually) and Belle II Monitoring Histograms (the total number of entries is shown as a black line). The lower plots show the differences between Belle and Belle II, hence a positive (negative) difference means there are less (more) entries in the Belle II Monitoring Histogram.

adjacent bins, especially for values near zero, and differences in the treatment of special floating point values such as infinity and NaN (Not a Number) leading to migration from the overflow/underflow bin to the bin including zero in rare cases (see Figure 2.7b).

Further differences are found: in the PDG codes of `MCParticles` due to the recovery of the full 32 bit as mentioned above; the number of daughters of the `MCParticles` due to the unconverted virtual photons occurring in nuclear interactions between the hadronic final state particles and the detector material (see Figure 2.7a); and in all kinematic quantities of $V0$ and π^0 objects after the mass-constrained vertex fit caused by different software employed to fit the vertices.

2.3.3.2. Experiment Dependency

The changes in hardware and software between the 31 Belle experiments influence the quantity and quality of the recorded data. In order to carry out physics analyses and compare them with studies on Belle II Monte Carlo events, quantities should be used, which are robust against differences between the Belle experiments. The beam-induced background in Belle II is expected to be increased by a factor $20 - 30$. Quantities that are already fluctuating between Belle experiments are likely to be



(a) All tracks found by the tracking software.

(b) Only good tracks: Impact parameters less than 2 cm and 4 cm in transverse and beam axis direction, respectively.

Figure 2.8.: The mean number of reconstructed tracks over the Belle experiments. The error bars show the uncertainty on the mean value, however they are mostly too small to be visible. The increase in the number of tracks after experiment 27 is caused by the switch to the improved reconstruction software and the replacement of the SVD.

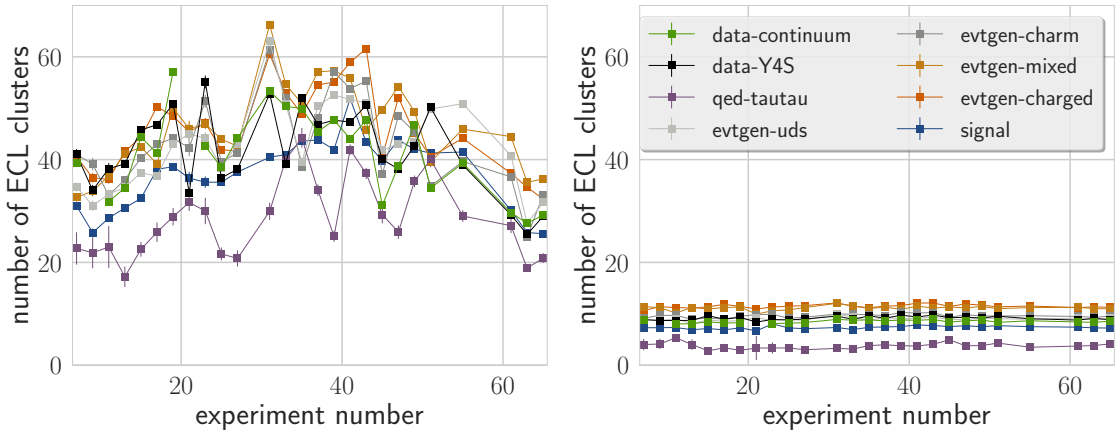
highly dependent on the assumed background conditions in Belle II. One of the most basic properties of an event is the number of fitted tracks. As can be seen in Figure 2.8 there are large differences between the Belle experiments for this quantity.

Therefore, the following selections are applied to Belle events in the remainder of this work: **good tracks** are selected with impact parameters less than 2 cm and 4 cm in transverse and beam axis direction, respectively (see Figure 2.8); and **good clusters** are selected without an associated track and with an energy of 50 MeV, 100 MeV, 150 MeV in the barrel, forward and backward region, respectively (see Figure 2.9). These are the same cuts which were usually applied in Belle analyses.

Some differences between the experiments remain after the cuts. In particular due to the switch to the improved reconstruction software and the replacement of the SVD after experiment 27.

2.3.3.3. Monte Carlo simulation and detector data differences

This thesis uses the last official Belle Monte Carlo campaign (prefixed with `evtgen-`). In addition the latest special Monte Carlo campaign for rare B decays is used (prefixed with `special-`). And finally some dedicated Monte Carlo samples for QED



(a) All clusters found by the reconstruction software.

(b) Only good clusters: No associated tracks and E $<$ 50, 100, 150 MeV in the barrel, forward and backward region, respectively.

Figure 2.9.: The mean number of reconstructed clusters over the Belle experiments. The error bars show the uncertainty on the mean value, however they are mostly too small to be visible. The large differences between the experiments are due to different beam conditions.

background were investigated (prefixed with `qed-`). Those were usually not used in Belle B meson analyses, because the `hadronBJ` skim suppresses them very efficiently.

Table 2.3 shows the extracted fraction and the corresponding fractions in the 7$^{\text{th}}$ Belle II Monte Carlo campaign for comparison.

The simulation of continuum events, that is events which do not contain a $\Upsilon(4S)$, is challenging due to the required knowledge of the correct QCD hadronization factors and branching fractions. In addition, the exact composition of the continuum background is poorly documented and not all components are part of the available official (or even unofficial) Monte Carlo campaign, e.g. $e^-e^+ \to \gamma\gamma$, is not available.

In consequence, the continuum description of Belle is not in agreement with the observed data. The off-resonance data recorded by the Belle detector can be used as a cross-check and sometimes replacement for Monte Carlo simulated continuum events. In a physics analysis the observables are usually extracted using a model for the signal and background components, the relative fractions are usually not fixed and determined directly from data. Nevertheless, the continuum component description is often a leading systematic uncertainty in Belle analyses [15].

The continuum composition was investigated using 10000 simulated and recorded events for each component, scaled according to the fractions stated in Table 2.3.

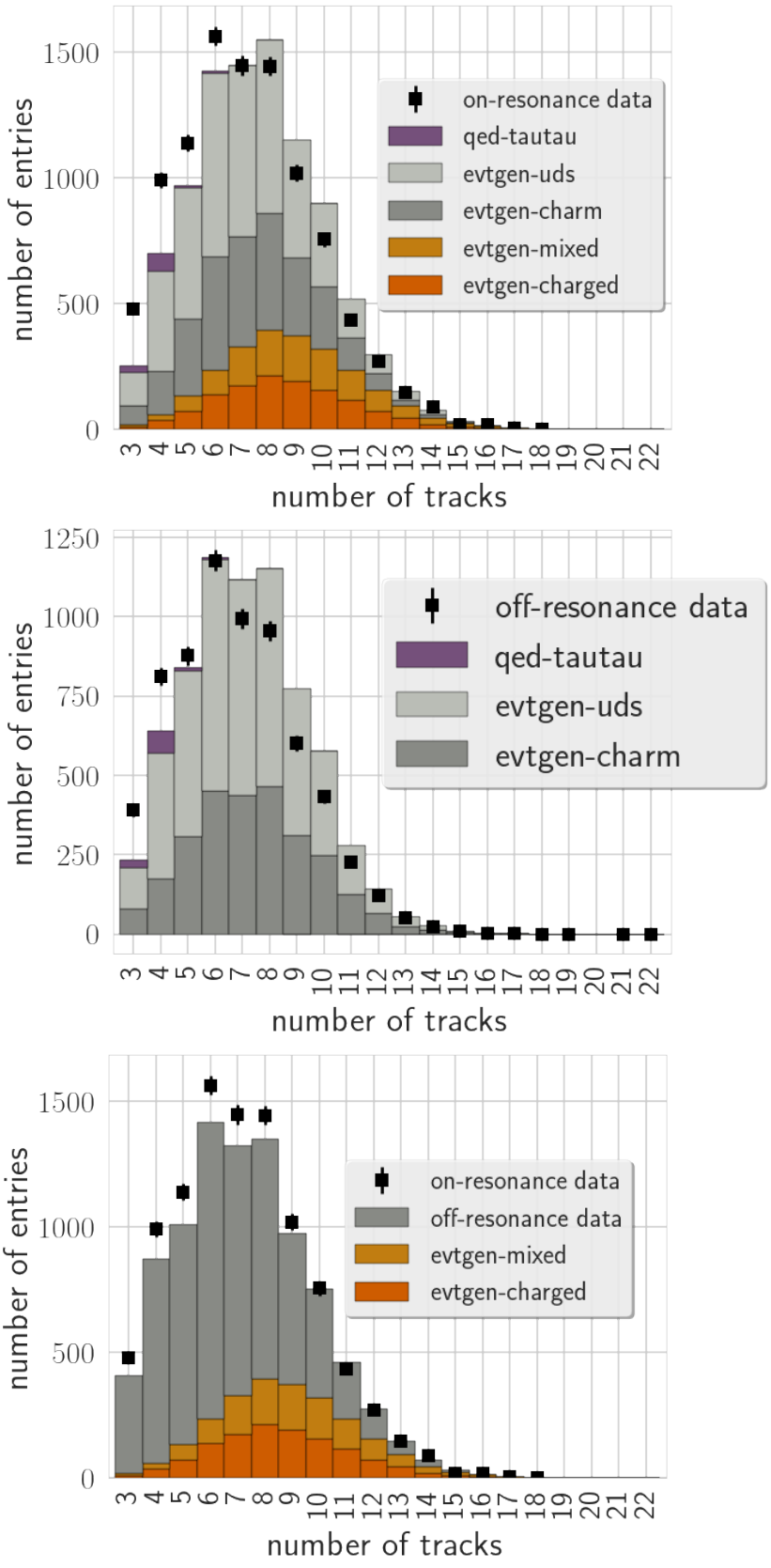


Figure 2.10.: The distribution of the number of good tracks in each event.

Table 2.3.: Number of simulated events corresponding to the recorded luminosity of $0.711\,\mathrm{ab}^{-1}$ normalized to the number of the recorded events at the $\Upsilon(4S)$ resonance. Belle simulated rare and $b \to u\ell\nu$ decays separately, while Belle II includes them in the evtgen-charged and evtgen-mixed samples. In contrast to the official Belle MC campaign, the relative ratio of $B^+B^-/B^0\overline{B}^0 = 0.514/0.486$ was taken into account.

Type	Belle	Belle II
evtgen-charged	0.1264	0.1310
evtgen-mixed	0.1195	0.1238
evtgen-charm	0.2933	0.3078
evtgen-uds	0.4716	0.5534
qed-tautau	0.1996	0.2052
qed-mumu	0.2395	–
qed-eemm	0.2539	–
special-mixedrare	0.00102	–
special-chargedrare	0.000824	–
special-mixedulnu	0.000859	–
special-chargedulnu	0.000923	–

Of the considered QED backgrounds only $e^-e^+ \to \tau^-\tau^+$ contributes. The other components are suppressed by the so-called `hadronBJ` cut, which is applied per default for all B analyses. As can be seen from Figure 2.10 the Monte Carlo continuum description is incomplete, it underestimates (overestimates) the events in the low-multiplicity (high-multiplicity) region. In contrast, the $\Upsilon(4S)$ Monte Carlo simulation combined with the off-resonance data fits the data better.

The over-estimation of continuum background, most-likely the $c\overline{c}$ component, leads to an increase of D meson candidates and finally to more combinatorial background for B mesons. An example is shown in Figure 3.16, where D mesons were reconstructed on Monte Carlo simulated events, on-resonance data and off-resonance data.

Throughout this thesis, these expected differences between Monte Carlo and data are observed in several places. Section 3.3.3 describes techniques to mitigate the influence of these differences in the context of multivariate classification. Section 4.3.2 investigates the continuum component in the context of exclusive tagging. In the benchmark analysis in Chapter 5, the continuum component is a large background in the hadronic τ decay-channels. Hence, the continuum description was investigated in detail and compared to off-resonance data.

2.4. Conclusion

The Belle to Belle II Conversion enables Belle II physicists to analyze the dataset recorded by Belle using `BASF2`. The conversion process was validated on a basic level by ensuring the same output for a large number of quantities. Differences which emerged were studied and explained.

In order to validate `BASF2` on a global level, physics analyses have to be performed and compared to results published by the Belle collaboration. In this thesis the decay $B \to \tau\nu_\tau$ is investigated and compared to [16] and [15].

Other measurements using the `b2bii` conversion are in preparation:

- the branching ratio of $B^+ \to \ell^+\nu\gamma$ with hadronic tagging [17],
- the branching ratio of $B^+ \to \ell^+\nu\gamma$ with semileptonic tagging [18],
- the branching ratio of $B^+ \to \ell^+\nu$ with inclusive tagging,
- and the search for $B_s^0 \to \phi\pi^0$ [19].

Furthermore, `b2bii` can be used to study the performance differences between the `Belle` and `Belle II` experiment, to optimize the latter as soon as first data has been collected.

Finally, the conversion ensures the preservation of the legacy of the Belle experiment: The full recorded dataset of nearly $1\,\mathrm{ab}^{-1}$ of data, which led mankind to the verification of CKM mechanism and the observation of tetra-quarks.

Chapter 3

Multivariate Analysis Algorithms

In recent years, the field of multivariate analysis and machine learning evolved rapidly, and provided powerful techniques, which are currently adopted in all fields of science. Prominent use-cases include: image and speech recognition, stock market trading, fraud detection, and medical diagnosis.

In high energy physics (HEP) multivariate analysis (MVA) methods are extensively used: to identify interesting collision in the trigger system; reject beam-induced hits in the drift chamber during track finding; infer the deposited energy in the calorimeter; provide particle identification information; reject particle candidates from combinatoric and physics background in an analysis.

The Belle II collaboration decided to provide a multivariate analysis package named `mva` as part of `BASF2` to encourage physicist to take advantage of existing methods and to keep up with the recent developments in the field. The package provides: a common code base for `BASF2`; support for all major MVA frameworks; transparent run-dependent loading of the fitted models from the `Belle II Conditions Database`; automatable tools for fitting, inference and evaluation of models.

The `mva` package and many algorithms based on it were implemented during this thesis. In the following chapter I outline the theoretical foundation of multivariate analysis and machine learning (Section 3.1), describe technical aspects of the `mva` package (Section 3.2), and present applications based on the package (Section 3.3).

3.1. Theory

Multivariate Analysis (MVA) algorithms are used to describe the distribution of observed data $\vec{x}$ and to deduce properties of the underlying process generating the data; the algorithms are based on multivariate statistics. Common tasks include: the approximation of a function $f(\vec{x})$ (regression Section 3.1.1); the separation of data-points into signal and background (classification Section 3.1.2); and grouping of similar data-points (clustering).

Machine learning (ML) provides an effective way to automatically learn the required statistical model using an appropriate domain-specific dataset, i.e. knowledge is extracted from experience, which can be viewed as a simple form of artificial intelligence. Usually, machine learning algorithms include a fitting-phase during which a statistical model is learned from the provided (training) dataset, and an inference-phase during which the statistical model is used to infer the desired (target) information for a new independent (test) dataset.

Depending on the available feedback during the fitting-phase one can distinguish between three different types of learning: the target information is provided during the fitting-phase (supervised learning); no additional information is provided during the fitting-phase (unsupervised learning); rewards and punishments are provided in a dynamic environment (reinforcement learning).

In HEP the target information is usually known from Monte Carlo simulated events, or can be inferred on data via data-driven techniques (see Section 3.3.3). Therefore, most tasks can be solved using supervised learning. In supervised learning the target information $y \sim f(\vec{x}) + \epsilon$ is predicted by a statistical model $\widehat{f}$ using a feature vector $\vec{x}$. The statistical model has internal degrees of freedom (so-called weights $\vec{w}$), which can be adapted by an algorithm to minimize the discrepancy of the true value y from the predicted value $\widehat{y} = \widehat{f}(\vec{x}, \vec{w})$, where the discrepancy is defined by a loss-function $\mathcal{L}(y, \widehat{y}, \vec{w})$.

Detailed introductions into the topic can be found in [20] and [21]. The relevant concepts are summarized in the remainder of this section.

3.1.1. Regression

A statistical model $\widehat{f}(\vec{x}, \vec{w})$, which estimates the value of a stochastic function $y = f(\vec{x}) + \epsilon$ is called a regressor. It can be interpreted as the function which minimizes the risk functional

$$R(\vec{w}) = \int \mathcal{L}(y, \widehat{f}(\vec{x}, \vec{w}), \vec{w}) \mathrm{d}P(\vec{x}, y), \tag{3.1}$$

where $P(\vec{x}, y)$ is the unknown joint probability distribution [22]. In the framework of empirical risk minimization (ERM) the risk functional is approximated by the empirical risk using a training dataset $(\vec{x}_i, y_i)$

$$R_{\text{emp}}(\vec{w}) = \sum_i \mathcal{L}(y_i, \widehat{f}(\vec{x}_i, \vec{w}), \vec{w}). \tag{3.2}$$

The weights $\vec{w}$ of the statistical model are chosen so that they minimize the empirical risk. This is known as the principle of empirical risk minimization [22].

A suitable loss function can be chosen using the ERM framework [23] or based on the maximum likelihood principle (see Section B.1). Typical examples are: the squared loss $(\widehat{y} - y)^2$; and the absolute loss $|\widehat{y} - y|$. Often additional terms are added to the loss function which depends on the weights $\vec{w}$ of the model, to impose certain smoothness and regularization assumptions on $\widehat{f}$.

A well-known example for the application of the principle of ERM is the linear regression model $y_i = \vec{x}_i \cdot \vec{w}$ with a squared loss, which leads to the well known least-square solution for the weights of the linear regression model $\vec{w} = (X^T X)^{-1} X^T \vec{y}$, where $X = (\vec{x}_2, \vec{x}_2, \ldots, \vec{x}_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$.

Various methods for regressions have been developed. One can distinguish between:

- **Parametric methods** with a fixed parametrized model structure and a finite number of unknown parameters that are estimated from the training dataset. Typical examples include: Generalized Linear Models [24] and Artificial Neural Networks [20, Chapter 5], [21, Chapter 11].
- **Nonparametric methods** estimate an appropriate model structure from the training dataset. Typical examples include: Kernel Density Estimator [25], [21, Chapter 6] and Decision Trees [21, Chapter 9].

As stated above, the estimated function f is stochastic. Therefore, its value y is a stochastic variable, which is distributed according to a conditional probability density function $P(y|\vec{x})$. In consequence, regression can be generalized to predict location, scale and shape parameters using a distribution assumption [26]; or the conditional density function itself without assuming a particular distribution using quantile regression [27] or mixture density networks [20, Chapter 5.6].

3.1.2. Classification

A statistical model $\widehat{f}$, which distinguishes signal $(y = 1)$ from background $(y = 0)$ is called a classifier. Classification can be viewed as a particular form of regression. The classifier is a regressor for the signal probability $P(y = 1|\vec{x})$. At a certain threshold (working point) C, data-points with $\widehat{f} > C$ are classified as signal, and background

otherwise. The choice of the working point depends on the expected cost of signal classified as background (type I error) and background classified as signal (type II error), or equivalently on the desired signal efficiency and purity (which are usually more closely related to the figure of merit used in physics analyses).

There is a close connection to the theory of hypothesis testing, the statistical model can be interpreted as a test-statistic $T = \widehat{f}(\vec{x}, \vec{w})$. The most efficient test statistic at a given significance level to distinguish between two simple hypotheses is given by the Neyman-Pearson Lemma

$$T_{\mathrm{NP}}(\vec{x}) = \frac{\mathrm{PDF}_{\mathrm{S}}(\vec{x})}{\mathrm{PDF}_{\mathrm{B}}(\tilde{\mathrm{x}})}, \tag{3.3}$$

where PDF denotes the probability density functions (PDF) of signal and background data-points. A sound introduction to hypothesis testing and the Neyman-Pearson Lemma can be found in the original paper [28].

The PDFs are usually unknown and cannot be sampled in a high dimensional feature space $\vec{x}$, due to the curse of dimensionality (see Figure 3.1), in consequence T_{NP} is unknown and has to be approximated with analytical (using distribution assumptions) or numerical (using machine learning) methods. Two approaches can be distinguished:

- **Generative Methods** like Fisher's discriminant [29], Kernel Density Estimators [25], [21, Chapter 6] or Gaussian mixture models [20, Chapter 9]; which approximate the probability density functions for signal $\widehat{f_S} \approx \mathrm{PDF}_{\mathrm{S}}(\vec{x})$ and background $\widehat{f_B} \approx \mathrm{PDF}_{\mathrm{B}}(\vec{x})$ separately, and can therefore be used to generate new data-points.
- **Discriminative Methods** like Boosted Decision Trees (BDT) [30], [21, Chapter 10], Support Vector Machines (SVM) [20, Chapter 7], [21, Chapter 12] or Artificial Neural Networks (ANN) [20, Chapter 5], [21, Chapter 11]; which directly approximate $\widehat{f}(\vec{x}) \approx T_{\mathrm{NP}}$, usually under the assumption that the dimension d of the hyper-plane separating (discriminating) signal and background in the feature space is much smaller than the dimension D of the feature-space itself $d \ll D$.

The methods differ in: the convergence rate against the optimal solution with increasing statistics, their robustness against outliers and measurement errors; the bias and variance of the obtained model (see Section 3.1.3).

Typical loss functions used for classification are: the logistic loss [23]; the hinge loss [23]; and the cross entropy [20, Chapter 4.3.2]. The loss-functions differ in their robustness against outliers and their convergence rate (see [23]). A motivation for the cross-entropy loss-function, which is extensively used in the field of deep-learning, based on the maximum likelihood principle is given in Section B.1.

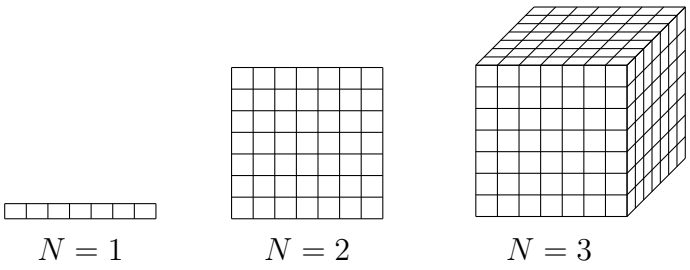


Figure 3.1.: Visualization of the curse of dimensionality. The phase-space of an N dimensional function grows exponentially, thus the number of events necessary to sample a high-dimensional probability density function grows exponentially as well and is unviable.

3.1.3. Model complexity

The complexity or number of degrees of freedom (NDF) of the statistical model strongly influences the prediction error of the model on an independent test dataset.

3.1.3.1. Bias-Variance Dilemma

There are three possible sources of errors in any statistical model $\widehat{f}(\vec{x})$ describing a multivariate distribution $y \sim f(\vec{x}) + \epsilon$.

1. (Bias) The model is not complex enough to describe all relevant aspects of the data – the model is under-fitted. This leads to a biased model $\text{Bias}(\widehat{f}) = \text{E}(\widehat{f} - f)$.
2. (Variance) The model is too complex and is dominated by the statistical fluctuations of the training data – the model is over-fitted. This leads to a model with a large variance $\text{Var}(\widehat{f}) = \text{E}(\widehat{f}^2) - \text{E}(\widehat{f})^2$.
3. (Irreducible) The intrinsic noise ϵ of the underlying distribution introduces an irreducible error $\text{Var}(y) = \text{Var}(\epsilon)$.

Mathematical we can expand the expected quadratic deviation of the statistical model $\widehat{f}$ from the true value y into these three components

$$\text{E}\left[(y - \widehat{f}(\vec{x}))^2\right] = \text{Bias}\left[\widehat{f}(\vec{x})\right]^2 + \text{Var}\left[\widehat{f}(\vec{x})\right] + \text{Var}\left[y\right].$$

Figure 3.2 visualizes the trade-off between bias and variance as a function of the model complexity. In the literature this fact is known as the Bias-Variance Dilemma [20, Chapter 3.2], [21, Chapter 2.9]. Since the over-fitting effect cannot be seen on the training data, it is crucial to always test the model on an independent validation dataset. Choosing the optimal model complexity is the key to the successful employment of MVA methods.

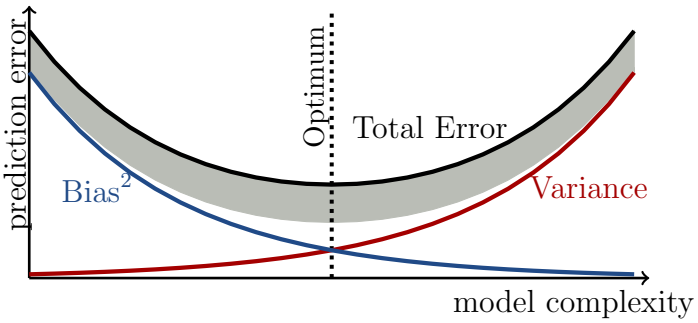


Figure 3.2.: Visualization of the trade-off between bias and variance of a MVA method depending on the model complexity.

3.1.3.2. Controlling the model complexity

The size of the training dataset at hand defines an upper bound for the feasible model complexity. The model cannot resolve the underlying distribution f below the threshold given by statistical fluctuations, i.e. the bigger the training dataset, the more degrees of freedom of a model can be constrained.

MVA algorithms usually provide several hyper-parameters Θ (in contrast to the parameters/weights of the statistical model itself) to control the complexity of the underlying model. The theoretical foundation is provided by the principle of structural risk minimization (SRM) [22]. Following [22], one can distinguish three different concepts of SRM:

- **Structure given by the architecture** – Hyper-parameters which define the architecture of the model directly control the global NDF of the model. Typical examples are: the number of hidden layers and neurons in an artificial neural network; the number of trees and their depth in a boosted decision tree.
- **Structure given by the learning procedure** – Hyper-parameters which control the effective NDF in different regions of the feature-space during the fitting-phase. Typical examples are: weight-decay in an artificial neural network; pruning algorithms for a decision tree; and boosting algorithm in ensemble methods.
- **Structure given by preprocessing** – Hyper-parameters which control the degree of degeneracy in the input data by applying a transformation to the input features prior to the fitting-phase. Typical examples are: binning and smoothing of the input features; dimensionality reduction using principal component analysis.

SRM provides another interpretation of the Bias-Variance Dilemma introduced above, where the bias is given by the empirical risk and the variance is introduced

by the generalization error due to the finite sample size. Using the hyper-parameters Θ, a series of models $\widehat{f}_1, \widehat{f}_2, \ldots, \widehat{f}_n$ with increasing complexity (defined by their Vapnik–Chervonenkis (VC) dimension h [21, Chapter 7.9]) is defined. The empirical risk is minimized for each model (meaning the model is fitted). Using the framework of SRM, an upper bound $C(h, l)$ for the generalization error introduced by the finite training sample size can be calculated, i.e, an upper bound for the difference $|R - R_{\text{emp}}|$ between the actual risk R and the empirical risk R_{emp}. The principal of structural risk minimization defines the optimal model as the one with the minimal guaranteed risk $R_{\text{emp}} + C$.

However, the upper bound C can only be calculated for certain models, and is usually only a weak bound. In practice, the optimal model is selected by optimizing the associated hyper-parameters, e.g. using grid or random-search [31]; gradient descent algorithms [32]; or Bayesian optimization algorithms [33].

3.1.4. Data analysis in High Energy Physics

Data analysis plays a crucial role in HEP experiments like Belle II. There are unique characteristics and associated challenges.

- The inherent stochastic nature of quantum field theory rules out a ground truth for individual data-points. Therefore, supervised machine learning and data analysis have to rely on Monte Carlo simulation or on advanced data-driven techniques to establish a ground truth.
- The investigated phenomena are usually very subtle and cause only small effects. Therefore, large amounts of events (often in the order of billions) have to be analyzed in a high-dimensional feature-space.
- The statistical and systematic uncertainties have to be treated very carefully to produce sound scientific results.

Consequently, the HEP community developed data analysis strategies suited for these requirements. Following the history of data analysis in HEP, I describe the advantages and disadvantages of the traditional cut-based (Section 3.1.4.1), the current multivariate (Section 3.1.4.2) and the prospective deep learning (Section 3.1.4.3) approach. An in-depth introduction can be found in [34].

3.1.4.1. Traditional Cut-based Analyses

In the traditional cut-based approach, the physicist determines a series of one-dimensional cuts on the dataset, to incrementally increase the signal-to-noise ratio in the signal-region.

For instance, one could cut on the K^+ PID information, the invariant mass of the π^0, and the χ^2 probability of the vertex fit to increase the signal-to-noise ratio in the reconstruction of the decay $D^0 \to K^-\pi^+\pi^0$.

The cuts are determined on Monte Carlo simulated events or in control regions[1]. The final signal region is only unblinded once the whole analysis strategy is fixed.

Physics knowledge can be easily incorporated into this process: phase-space with large systematic uncertainties (for instance due to unreliable Monte Carlo simulation or threshold effects in reconstruction algorithms) can be excluded; the dataset is preprocessed and reduced to a suitable subset for this analysis (for instance intermediate particles are reconstructed and irrelevant features are dropped); new physically motivated features can be constructed (for instance angles in particular reference frames); and erroneous assumptions or software can be spotted by experienced analysts.

It is relatively easy to estimate the systematic uncertainty introduced by the cuts, by varying the cut in a certain range. Although, this common procedure has been criticized (see [34, Chapter 8.4.3]). Moreover, the agreement between Monte Carlo simulation and data can be verified after each cut.

The fixed analysis procedure is usually simple enough to be included in recasting frameworks, which recast existing analysis to provide bounds for new physics scenarios, by automatically analyzing custom signal Monte Carlo simulated events [35].

On the other hand, multivariate correlations are not (or only to a small degree) taken into account, which leads to reduced signal efficiency. Possible correlation between the cuts are often neglected. There is no obvious choice for the range in which the cuts are varied to estimate the systematic uncertainties. Finally, the process of manual cut determination and feature engineering is time-consuming and error-prone in itself.

3.1.4.2. Multivariate Analysis

In a multivariate analysis approach, the physicist uses MVA methods like machine learning to create a statistical model of the dataset. Either this can be used to infer the signal probability of each event, thereby increasing the signal-to-noise ratio in the signal-region, or to directly model and extract the signal component in the dataset. The statistical model is determined on Monte Carlo simulated events, in control regions or using other data-driven methods. Sometimes it is possible to calibrate the model on data (see Section 3.3.1.2).

[1] A dataset without signal, but comparable background properties.

The multivariate correlations between the input features are taken into account, frequently this leads to an improved signal efficiency by a factor two or more (e.g. [36]). The available dataset is therefore exploited more efficiently and the statistical uncertainty of the final physics result is reduced.

The preprocessing of the dataset and the creation of useful physically motivated high-level features is still handled by the physicist. The subsequent model building is automatized and can be applied to similar classes of analyses without further human intervention (see Section 3.3.1.1).

On the other hand, the validity of the statistical model has to be carefully verified using control regions and channels, to avoid possible over-fitting and mis-modeling effects. Physics knowledge must still be incorporated into this process to exclude problematic phase-spaces. Due to the increased complexity compared to the cut-based approach, the method itself is seen as a black box by some analysts, leading to distrust and increased systematic uncertainty of the final physics results.

Today, MVA is an established tool in HEP with clear advantages in analyses with dominating statistical uncertainty.

3.1.4.3. Deep Learning

The deep-learning approach is a multivariate analysis method which includes parts of the preprocessing and the feature engineering into the statistical model. It shares the same traits as the general multivariate analysis approach. Deep learning uses deep artificial neural networks and takes advantage of the large available training datasets and the massively parallel computing power provided by modern hardware.

It is long known that neural networks with one hidden layer and an arbitrary, bounded, non-constant activation function can approximate arbitrary functions in L^{p2} [37]. However, as was shown recently in [38], deep neural networks with more than one hidden layer can approximate functions of practical interest (e.g. $f(x, y) = x \cdot y$) with exponentially fewer parameters than shallow networks.

Low-level features like the measured four-momenta and calorimeter entries are directly fed to a deep artificial neural network. The deep neural network is intended to learn suitable high-level features automatically. Recent research [39] indicates that deep neural networks can extract more information from low-level features in comparison to high-level features created by experienced human analysts. Consequently, deep-learning approaches outperform established algorithms in many fields, as was reviewed in [40].

[2] The Lebesgue space of p-integrable functions

Physical knowledge can be incorporated into the architecture of the deep neural network. In particular, the explicit consideration of invariance properties (e.g., the invariance against rotations around the beam-line) is important.

The application of convolutional and recurrent neural networks in HEP is investigated. Adversarial networks [41] allow to selectively exclude features from the generated statistical model, to enforce a uniform signal selection efficiency in this feature [42]. This is discussed in greater detail in Section 3.3.2.

In the theory of representation learning [43], the improved performance is explained by the transformation of the low-level input feature space into a representation, which allows to easily solve the tasks at hand.

An illustrative example is introduced in [44]. In this work a deep neural network is presented, which can create textual descriptions of an image. The pixels of the image are transformed to an intermediate representation of the objects contained in the image by a convolutional neural network. Subsequently, the intermediate representation is transformed into a textual description of the image by a recurrent neural network.

The key concepts of representation learning are (as described in [43]):

- **distributed representations** – the number of input regions which can be distinguished grows exponentially $\mathcal{O}(2^N)$ in the numbers of parameters N;
- **depth** – the ways to re-use learned features grow exponentially with the depth of the network;
- **abstraction** – the learned features in the deeper layers are increasingly invariant to most local changes of the input;
- and **disentangling factors of variation** – the learned features represent independent properties of the input data.

On the other hand, due to the direct usage of many low-level features it is unclear how to define suitable control regions and channels. At the same time, the sensitivity to mis-modeling in Monte Carlo simulation is further increased in comparison with ordinary MVA methods, due to the reliance on a multitude of low-level features. Therefore, it is more difficult to estimate systematic uncertainties of deep neural networks in absence of calibration techniques on data. This problem can be mitigated with data-driven techniques like event-based re-weighting (see Section 3.3.3).

Instead of feature engineering the time is spent in architecture engineering. Finally, the field evolves quickly and there are no established and commonly accepted fields of application in HEP yet.

The usage of deep learning revolutionized many fields in recent years. Applications with a learnable distributed representation, are in particular suitable for deep-learning.

3.2. Implementation

The `mva` package provides a common code base for MVA related algorithms in `BASF2`. Figure 3.3 shows a schematic overview of the package. It is used in: the online reconstruction by the `FastReco` path of the trigger package; the offline reconstruction by the `tracking`, `ecl` and `klm` packages; and the mDST analysis by all MVA-based algorithms like `Continuum Suppression`, `Flavour Tagging`, and the `Full Event Interpretation`.

The `mva` package can use most of the popular MVA frameworks as backend, but it is not a wrapper[3] around existing MVA frameworks. Rather, it provides glue-code, which integrates the frameworks into `BASF2`, so that the user can focus on dealing with the framework of his choice instead of the merits of `BASF2` and `ROOT`.

For the fitting-phase, the `mva` package: converts the training data from the `ROOT`-format used by Belle II into the appropriate backend-specific format; streams the data to the backend (conserving the potential out-of-core capability of the backend); bundles the generated backend-specific `WeightFiles` (serialized form of the underlying statistical models) and all information necessary to reproduce the fitting into a Belle II `WeightFile`; and uploads the Belle II `WeightFile` into the `Belle II Conditions Database` with an optional interval of validity (a range of experiments and runs for which the `WeightFile` should be used). In addition, basic examples for all backends are provided, which the user can use as a starting-point; and advanced examples explaining state-of-the art machine learning techniques like Bayesian hyper-parameter optimization [33]; adversarial neural networks [42]; feature importance calculation and sPlot [45] [46].

For the inference-phase, the `mva` package: (re-)loads the appropriate Belle II `WeightFile` from the `Belle II Conditions Database` for the current experiment and run; loads the backend-specific model; extracts the required features from the `DataStore`[4]; calculates the response using the model and stores the response in the `DataStore`.

For the evaluation, the `mva` package provides: plotting primitives for features (distribution, correlations); plotting primitives for the method response (distribution, ROC curves, over-fitting checks); and calculation of feature importances. The evaluation tools are backend-agnostic and can be used to compare different backends.

Finally, the package was designed with a large degree of automation in mind. The provided tools can be invoked from `C++`, `Python` and `bash` with similar interfaces. Fitting, inference and evaluation do not require human interaction.

[3] A wrapper would provide a backend-agnostic interface and in consequence the lowest common denominator of all backends.

[4] A memory-representation of the current event.

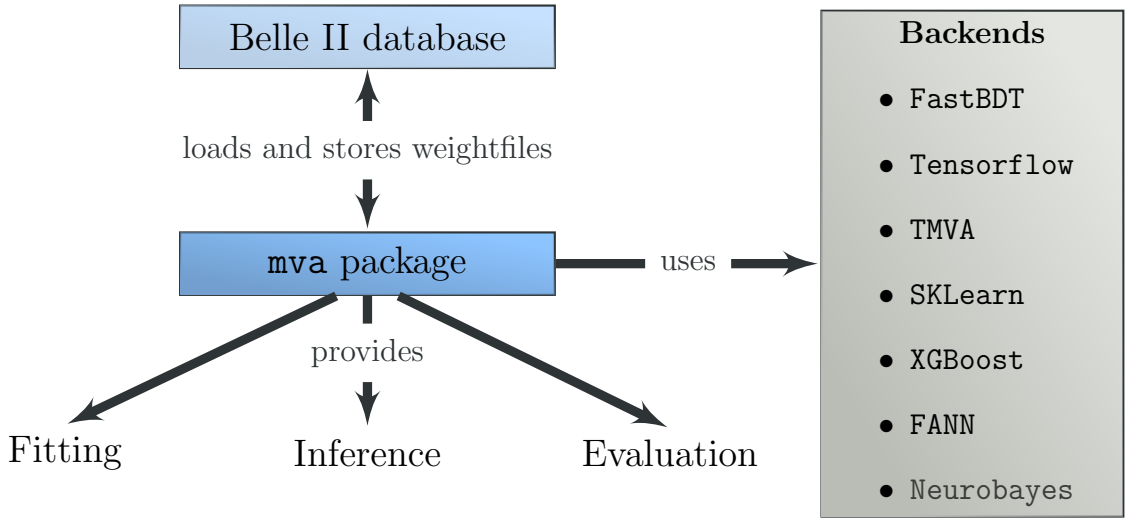


Figure 3.3.: Schematic overview of the `mva` package. It provides: an automatable fitting, inference and evaluation interface to the user; transparently loads and stores the `WeightFiles` (serialized form of the underlying statistical models) in the Belle II database; and can use the most popular MVA frameworks as backend.

3.2.1. Methods

In the following sections I briefly describe key ideas and concepts of the supported backends. However, the machine learning field evolves quickly and most of the projects will have published updated versions since the publication of this thesis. Besides the frameworks mentioned below, the `mva` package can be easily extended to new backends upon user request.

3.2.1.1. FastBDT

Belle II required a default multivariate classification algorithm which is: fast during fitting and application; robust enough to be trained in an automated environment; can be reliably used by non-experts; preferably generates an interpretable model and exhibits a good out-of-the box performance.

`FastBDT` [47] fulfills these requirements and is the current default classification method of the `mva` package. It implements a speed-optimized and cache-friendly implementation of the widely employed Stochastic Gradient-Boosted Decision Tree (BDT) algorithm [30], which exhibits a good out-of-the-box performance and generates an interpretable model. Furthermore, `FastBDT` supports: preprocessing with equal-frequency binning and purity-transformation; correct handling of missing data; and boosting to uniformity (see Section 3.3.2.3).

FastBDT was originally developed during this thesis to speed up the fitting and inference-phase of the Full Event Interpretation (see Chapter 4). The source-code is licensed under GPLv3 and available on github [48]. FastBDT is shipped with BASF2 and is available on all computing sites [47].

3.2.1.2. TMVA

The TMVA library is provided by the ROOT framework. TMVA [49] is traditionally used in HEP and implements a wide range of methods: analytical methods like the Fisher discriminant; decision tree based like gradient-boosted decision trees; different forms of neural networks; probability density estimators; and rule-based approaches. It supports classification, multi-class classification and regression. Furthermore, preprocessing steps like normalization, decorrelation and Gaussianization can be performed on all or a subset of features.

TMVA is actively developed by the HEP community, and started to adopt recent developments like deep learning. Since it is shipped with the ROOT framework, it is available on all computing sites.

3.2.1.3. FANN

The Fast Artificial Neural Network (FANN) library [50] is an open-source neural network library, which implements a multi-layer neural network. Input features and the output of the network can be automatically scaled to sensible ranges. Furthermore, it supports fitting and inference on multiple CPU cores. It is shipped with BASF2 and is available on all computing sites.

3.2.1.4. NeuroBayes

NeuroBayes [51] is a closed-source, commercial implementation of a neural network. It provides an advanced feature preprocessing (equal-frequency binning, purity-transformation, b-spline fits); Bayesian regularization methods like ARD (automatic relevance determination); and evaluation capabilities including feature importance estimations. It was extensively used by the Belle experiment.

As of the time of writing it is not further developed, but legacy support is available. Due to license restrictions it is only available on some computing sites (including the KEK computing center, DESY and KIT IETP). Belle II decided to use NeuroBayes only for belle legacy (for example in the b2bii package) code.

3.2.1.5. Python-based

The `mva` package can support all `Python`-based MVA frameworks. The technical details are discussed in Section B.2.

Officially supported (i.e. examples and test-code is provided by `BASF2`) are:

- `SKLearn` [52], is part of `SciPy`[5], it supports classification, regression, clustering, dimensionality reduction, model selection and preprocessing;
- `XGBoost` [53], is an open-source, scalable, portable and distributed Gradient Boosting library;
- `hep_ml`, an open-source machine learning library dedicated to algorithms used in HEP;
- `Tensorflow` [54], is an open-source library for (deep) neural networks using CPU and GPUs, with wide adoption by technology companies around the world;
- `Keras` [55], is an open-source, high-level interface to neural network frameworks like Tensorflow, it was developed for easy and fast prototyping of deep learning algorithms.

Other `Python`-based frameworks like `pylearn2` [56] and `theano` [57] are supported unofficially. The individual `Python`-based methods are not available on the computing sites, but can be installed by the user.

3.2.2. Benchmark: $D^0 \to K^-\pi^+\pi^0$

Separating correctly reconstructed particles from background is a frequent classification task during **mDST and n-tuple analysis**. The reconstruction of the decay $D^0 \to K^-\pi^+\pi^0[\to \gamma\gamma]$ is used as a benchmark classification in the sections below. This decay contains the most abundant final-state particles in the detector K, π and γ, and has a rich phase-space structure with various intermediate resonances like ρ and K^*. Furthermore, with a branching fraction of $(14.3 \pm 0.8)\%$ [58], it is also the most common decay of the D^0.

The decay is reconstructed from a K^- candidates with a kaon probability above 0.5, a π^+ candidate with a pion probability above 0.5, and a $\pi^0 \to \gamma\gamma$ candidate with χ^2 probability of a mass-constrained vertex fit above 0.1. Further selection criteria are applied on the γ candidates. The invariant mass of the final D^0 meson has to be between 1.7 GeV and 1.9 GeV. A vertex fit is performed, and the χ^2 probability

[5] A Python-based ecosystem, of open-source software for mathematics, science, and engineering.

of the fit has to be above 0.1. The following input features are used if not stated otherwise: the invariant mass of the D^0 meson; the pairwise invariant masses of the decay products $M_{K^-\pi^+}$, $M_{K^-\pi^0}$, and $M_{K^+\pi^0}$; particle identification information of K^- and π^+; D^0, K^-, π^+ and π^0 momenta; D^0, K^- and π^+ vertex information; and the χ^2 fit probability of D^0 (vertex-fit), K^+ (track-fit), π^+ (track-fit) and π^0 (mass-vertex-fit).

In total 745000 charged B^+B^- Monte Carlo simulated events were processed, and split into a training (355000 events) and an independent validation sample (390000 events). The signal fraction in both samples is 6.6%.

3.2.2.1. Comparison

The benchmark introduced in Section 3.2.2 was used to compare the performance (in terms of fitting runtime, inference runtime, classification quality and `WeightFile` size) of different backends.

Two sets of classifiers were compared among each other on an Intel(R) Core(TM) i7-4770 CPU (@ 3.40GHz) with a main memory of 32 GigaByte. To ensure a sound comparison, all methods were restricted to one core and typical hyper-parameters were chosen.

- Stochastic gradient-boosted decision tree implementations (`FastBDT`, `TMVA`-BDT, `SKLearn`-BDT, `XGBoost`) with the same hyper-parameters (one hundred trees with a depth of three).
- Artificial neural network implementations (`FANN`, `TMVA`-NN, `SKLearn`-NN, `Tensorflow`, `NeuroBayes`) with the same hyper-parameters (one hidden layer with 29 neurons and maximally one hundred iterations through the dataset during the fitting). The input features were normalized (shifted to a mean of zero, and scaled to a standard deviation of one).

The presented results depend on the chosen preprocessing steps, hyper-parameters and the specific task. In consequence, this comparison is only a rough guideline of the strengths and weaknesses of the backends. The results are summarized in Table 3.1.

The BDT methods are faster than the Neural Networks during the fitting-phase, at the same time they outperform the Neural Networks in the quality of the separation (measured by the area under the receiver operating characteristic AUC ROC [59]). `FastBDT` is the fastest method and obtains the best AUC ROC score on the benchmark. The convergence of `FANN` for this benchmark was unstable i.e. the obtained ROC values fluctuated strongly.

The runtime required to load the validation data from disk dominates the inference-phase (as can be seen from the `Trivial` method, which only loads the validation data and does not perform computations). To ensure a significant result the validation dataset statistics was increased by factor 10. Contrary to the naive expectation that the Neural Networks are faster during the inference because they rely only on linear algebra operations, which can be performed very efficiently on modern hardware; `FastBDT` is the fastest method. `SKLearn`-BDT is nearly as fast as `FANN` and `SKLearn`-NN, faster than `TMVA`-NN and `Tensorflow`. `NeuroBayes` is the slowest contestant.

Several effects have to be considered to explain the unexpected result: preprocessing steps can be very expensive (e.g. `NeuroBayes` does several non-linear transformations as preprocessing); careful optimizations for caching and pipelining have a large impact on modern hardware (as can be seen by the results of `FastBDT`). In consequence, even the same algorithm can have very different runtimes depending on the implementations (as can be seen by the different BDT implementations, which all implement stochastic gradient-boosting).

3.2.3. Conditions Database Integration

The `Belle II Conditions Database` contains experiment and run-dependent information required for the processing of Monte Carlo simulated events and detector-data. Traditionally machine parameters like: beam energies; luminosity; detector alignment and status; and calibration constants are stored in the condition database. Due to the expected extensive usage of MVA methods in Belle II, and their dependence on the run-dependent background conditions, the `WeightFiles` of MVA methods used in the online and offline reconstruction (likely also mDST analysis) can be fitted per experiment (or run) and stored in the `Belle II Database`.

Furthermore, the database provides a revision system for `WeightFiles`, an infrastructure to automatically distribute the required `WeightFiles` to the computing sites, and ensures reproducibility.

The `mva` package transparently loads and stores `WeightFiles` in the `Belle II Conditions Database`. Optionally, the `WeightFiles` can be loaded from and stored on disk in the `ROOT` and `XML` format. Depending on the backend, used method and its hyper-parameters, the `WeightFiles` usually require between 10 kilobyte and 1 megabyte disk-space. Table 3.1 states the size of the `WeightFiles` for the different backends in the `ROOT` format.

3.2.4. Automatic Evaluation

Evaluating the fitted models is important: to spot possible issues in the input features; to check for convergence of the algorithm; to avoid biases due to over-fitting

Table 3.1.: Comparison of the different backends using the benchmark $D^0 \to K^-\pi^+\pi^0$. The first column shows the fitting time on 28 features and 355000 events measured in seconds. The second column shows the inference time on 28 features and 3900000 events measurement in seconds. The third column shows the area under the receiver operating characteristics as a measure of the classification quality (more is better). Different quantities can be used to construct the receiver operating characteristic, here the signal efficiency and purity are used. The last column shows the size of the `WeightFile` in kilobytes. All measurements were performed 10 times, the minimal achieved time in fitting and inference, and the average AUC ROC is stated in the table. The `Trivial` method does nothing during the fitting-phase (time equals overhead of the `mva` package), and returns the signal-fraction during the inference-phase (time equals overhead due to data loading).

Method	Fitting time in s	Inference time in s	AUC ROC	`WeightFile` size in KB
`Trivial`	0.2	4.9	0.066	2
Stochastic Gradient Boosted Decision Tree				
`FastBDT`	3.7	6.9	0.435	58
`SKLearn`-BDT	32.1	7.8	0.429	69
`XGBoost`	18.0	11.4	0.415	34
`TMVA`-BDT	19.8	16.5	0.297	101
Artificial Neural Network				
`SKLearn`-NN	27.6	7.2	0.401	32
`Tensorflow`	201.9	9.4	0.399	30
`NeuroBayes`	112.3	75.4	0.377	182
`FANN`	50.6	7.1	0.316 ± 0.061	21
`TMVA`-NN	510.6	16.8	0.156	53

and under-fitting; and compare different models to one another. The automatic evaluation of the `mva` package creates a summary in form of a `PDF` document, which can later be examined by the user. It is also possible to observe the quality of an MVA method in an automated fitting environment and to sent emails to quality control if the quality drops below a predefined threshold.

The `PDF` summary of the `mva` package contains:

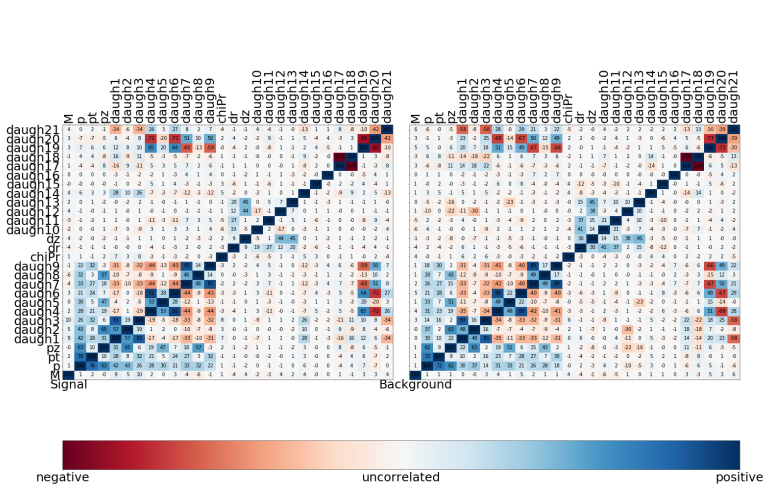
- the configuration of the used methods;
- correlations of all features for signal and background (see Figure 3.4a);
- importance of all features for all methods (see Figure 3.4b and Section B.3);
- distributions of all features and spectators (see Figure 3.4c);
- receiver operating characteristics on the training and an independent test datasets for all methods (see Figure 3.4d);
- distributions on the training and an independent test dataset of the response for all methods (see Figure 3.4e);
- diagonal plots, which shows purity of the binned response for all methods.

Additional plots (like `tSNE` plots, and control regions plots for different response-cuts) are provided on demand.

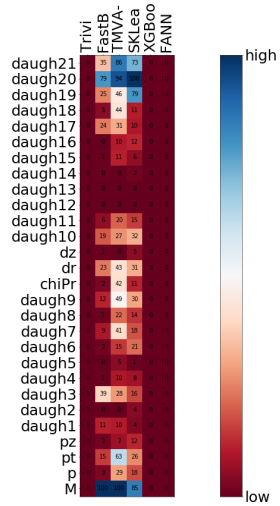
3.3. Applications

This section summarizes the usages of machine learning and the `mva` package in the Belle II collaboration. Section 3.3.1 describes applications during mDST analysis; these were developed during this thesis and many are used in the analysis $B \to \tau\nu_\tau$. Furthermore, the `mva` package is used extensively during online and offline reconstruction: in the identification of K_L^0 particles from clusters in the KLM; the energy and shape determination of clusters in the ECL; and the track finding and fitting algorithms in the CDC and VXD.

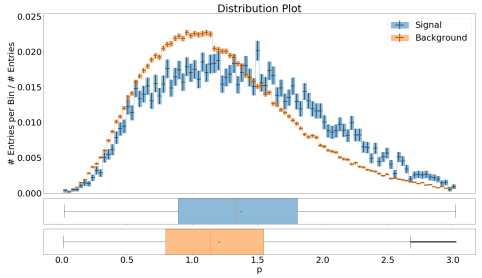
This thesis focused in particular on the development of basic building blocks for complex algorithms. Often, some features have to be reserved for subsequent steps in the analysis like the extraction of the signal yield using a fit. Therefor, discriminators with an enforced uniform signal selection efficiency were investigated in Section 3.3.2. Section 3.3.3 introduces data-driven techniques, which are a key to reduce the dependency on Monte Carlo simulation. Finally, Section 3.3.4 outlines algorithms used to optimize the hyper-parameters of an MVA method.



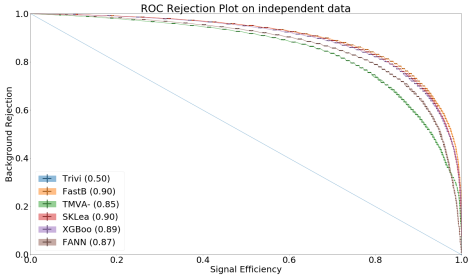
(a) Correlation among features for signal and background.



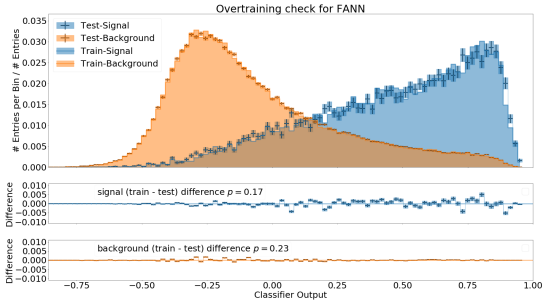
(b) Feature importances for different methods.



(c) Distribution of signal and background for one feature.



(d) Receiver operating characteristic for different methods.



(e) Distribution of the response of one method for signal and background. Useful to check for over-fitting. A Kolmogorov-Smirnov test is performed to check for statistical compatibility of the distributions on the training and an independent test dataset.

Figure 3.4.: Exemplary evaluation plots for the benchmark problem presented in Section 3.2.2.

3.3.1. Analysis Algorithms

All multivariate algorithms used in the `analysis` package of `BASF2` are based on the `mva` package. In this section I briefly describe the three most important algorithms, which were implemented during this thesis (Section 3.3.1.1) or during supervised master theses (Section 3.3.1.2 and Section 3.3.1.3).

3.3.1.1. Full Event Interpretation

The Full Event Interpretation (`FEI`) is used for hadronic and semileptonic tagging. It automatically reconstructs hadronic and semileptonic B meson decay-chains and infers a signal probability for each reconstructed candidate. This algorithm is used in the measurement of branching fraction of rare decays with missing kinematic information like $B \to \tau\nu_\tau$, $B \to \ell\gamma\nu$ and $B \to \nu\nu$.

This signal probability is calculated using a hierarchical network of multivariate classifiers, which is visualized in Figure 4.3. A boosted decision tree based on `FastBDT` (see Section 3.2.1.1) is trained for each final state particle and each decay-channel of the intermediate particles. For each particle candidate a signal probability is calculated using the corresponding BDT. The signal probabilities of the daughter particles are used as features in the training of the BDT of a composite intermediate particle. In this manner the complete information about the entire decay-chain is encoded in the signal probability of the final B candidates.

The training, application and evaluation of the more than 100 BDTs employed by the `FEI` is fully automatized with the tools provided by the `mva` package.

The FEI is described in greater detail in Chapter 4.

3.3.1.2. Flavour Tagging

`BASF2` includes two flavour tagging algorithms, which distinguish the decay products of a B^0 meson from a $\overline{B}^0$ meson. Flavour-tagging is used in time dependent CP violation analyses like the measurement of $\mathcal{A}_{CP}(B^0 \to J/\psi K^0_S)$.

The category-based flavour tagger trains either boosted decision trees (see Section 3.2.1.1) or neural networks (see Section 3.2.1.3) to identify certain flavour-specific signatures (so-called categories) like primary lepton decays or $b \to c \to s$ decay-chains [60]. Each classifier receives hand-crafted features, which are known to be correlated to the flavour-specific signature. The output of the category classifiers are combined by another BDT or NN to the final flavour tagging information (1 for B^0 and -1 for $\overline{B}^0$) and the estimated uncertainty on the tag (0 for a random choice and 1 for absolute certainty).

The deep learning-based (see Section 3.1.4.3) flavour tagger receives only low-level features as its input [61]. The four momenta, POCA (point of closes approach of the track to the beam pipe) and PID (particle identification) information of all tracks associated to the decay products. The quantities are fed into a deep neural network ordered by the total momentum of the associated track and separated by charge. It is assumed, that the deep neural network learns the relevant flavour-specific signatures and correlated high-level features automatically.

The training and application of both taggers is fully automatized with the tools provided by the `mva` package.

The performance of the flavour tagger based on deep learning is out-of-the-box equal to the traditionally used category-based flavour tagger. Both taggers can be evaluated and calibrated on data using the `b2bii` package.

3.3.1.3. Continuum Suppression

Most analyses at the Belle II experiment investigate the decay of B mesons from an $e^-e^+ \to \Upsilon(4S) \to B\overline{B}$ decay. An important background to these analyses are continuum events $e^-e^+ \to q\bar{q}, \tau^-\tau^+, \mu^-\mu^+, e^-e^+$.

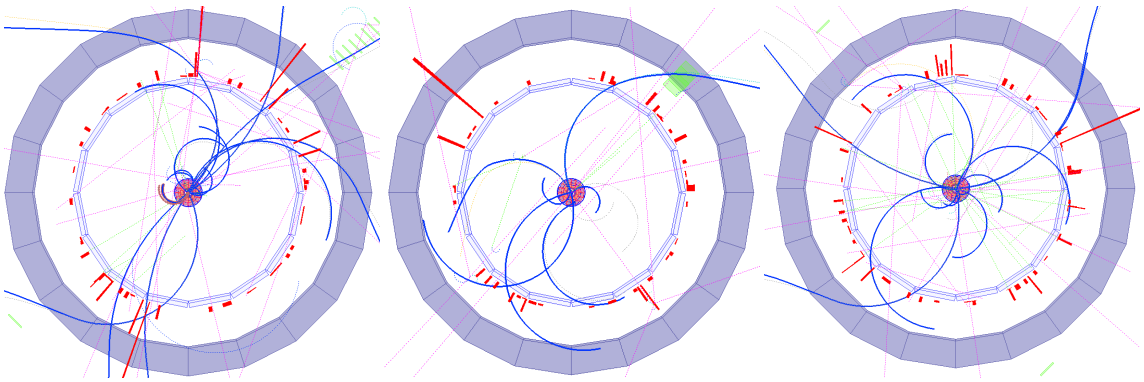
Historically, the difference in the event topology was used to suppress continuum events. Due to the low momentum of the B mesons in the center of mass (CMS) frame, the topology of a $B\overline{B}$ event is spherical in the CMS frame. The continuum events show a twofold cone-shaped back-to-back structure in the CMS frame. Hand-crafted high-level features were designed to systematically describe the event topology like the CLEO cones [62] and the KSFW (Kakuno Super Fox Wolfram) moments (see [63] and [1, Chapter 9]). Figure 3.5 shows example event topologies of a continuum event in Figure 3.5a, a possible signal-decay $B^- \to \tau^-\overline{\nu}_\tau$ event in Figure 3.5b and a generic charged $\Upsilon(4S)$ event in Figure 3.5c. As can be seen from the examples, the signal event topology is more similar to the continuum than the $\Upsilon(4S)$ event, due to the one-prong decay on the signal-side. Hence, it is important to consider the specific signal-side in the continuum suppression algorithm.

`BASF2` includes two continuum suppression algorithms.

The traditional continuum suppression algorithm uses a boosted decision tree[6] (see Section 3.2.1.1), which was fitted to suppress the continuum background, using the hand-crafted high-level features (see Section 3.1.4.2).

The deep-learning-based (see Section 3.1.4.3) continuum suppression algorithm receives in addition low-level features as its input [64]. The four-momenta and POCA of the charged particles are transformed into the CMS frame and rotated with respect

[6] FastBDT: One hundred trees with a depth of three.



(a) $e^-e^+ \to c\bar{c}$ (b) $B^- \to \tau^- \overline{\nu}_\tau$ (c) $e^-e^+ \to \Upsilon(4S) \to B^+B^-$

Figure 3.5.: Event displays showing the event topologies of different event types. The difference of the spherical topology of $\Upsilon(4S)$ and the twofold cone-shaped back-to-back structure of continuum events is exploited by hand-crafted continuum suppression features to separate them from one another.

to the thrust axis. They are fed to a deep neural network in spherical coordinates ordered by the magnitude of the momentum of the particles and separated by the charge and their assumed affiliation to the signal or tag side.

Both continuum suppression algorithms are trained using a dataset containing signal events (instead of generic $\Upsilon(4S)$ event) and continuum events. The addition of low-level features leads to a significant improvement in the separation (see Figure 3.6). No significant difference between a BDT and a simple deep neural network were observed. However, more advanced network architectures lead to further improvements [64].

3.3.2. Uniformity-constraint

Physics analyses often employ MVA classifiers to improve the signal-to-noise ratio using multiple *features*, which can separate between signal and background. On the other hand, many analyses extract physical observables by fitting a signal and background model to one or more *fit-variables*. A typical example is a two-dimensional fit to the Dalitz plane of the benchmark decay $D^0 \to K^-\pi^+\pi^0$ to extract the fraction, mass and width of intermediate resonances in the decay. In this example, the *fit-variables* are the invariant masses of the $K^-\pi^+$ system $M_{K^-\pi^+}$ and the $K^-\pi^0$ system $M_{K^-\pi^0}$; whereas the remaining quantities described in Section 3.2.2 are used as *features* in the classifier. Figure 3.7 shows the signal and background distribution of the benchmark decay in the *fit-variables*.

It is advantageous to require a uniform selection efficiency of the classifier for signal and background in the *fit-variables*, separately.

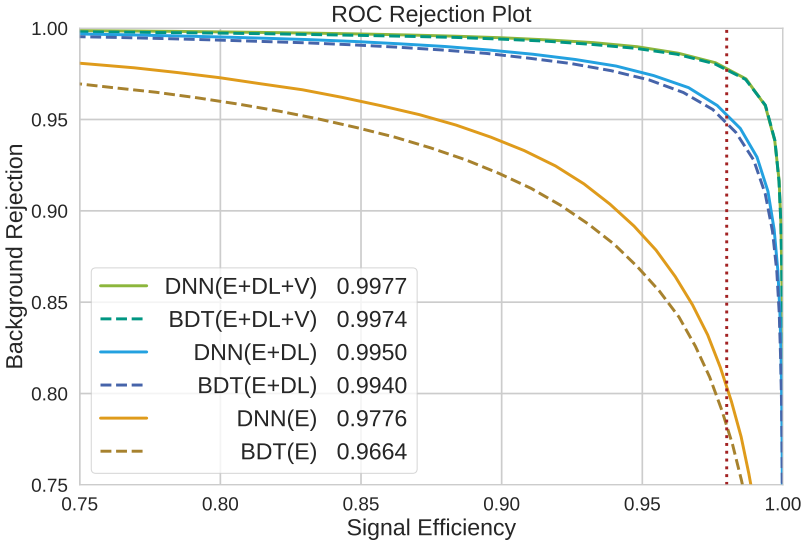


Figure 3.6.: Comparison between different feature sets: traditional hand-crafted features E (30), low-level kinematic features DL (440) and low-level vertex features (60). The addition of low-level features increases the quality of the separation. Taken from [64].

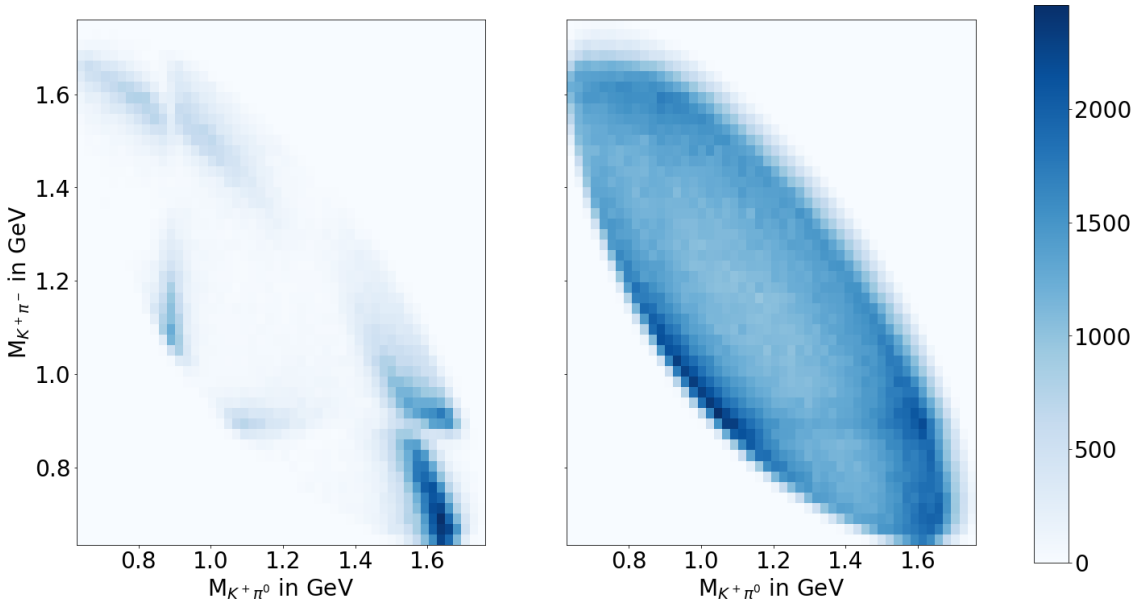


Figure 3.7.: Dalitz plane of the benchmark decay $D^0 \to K^-\pi^+\pi^0$. (Left) correctly reconstructed D^0; showing the rich substructure due to the intermediate resonances in the decay. (Right) wrongly reconstructed D^0.

- A non-uniform selection efficiency of background events creates an artificial peaking background if a cut on the classifier output is performed. The shape of the background model depends on the cut on the classifier output. This can lead to large systematic uncertainties if the background model cannot be checked independently in a control region.
- A non-uniform selection efficiency of signal events changes the shape of the signal model depending on the cut on the classifier output. This increases the reliance on Monte Carlo simulation, because usually the signal model cannot be checked independently in a control region.
- Contrastingly, a non-uniform selection efficiency of signal and background combined is desired, because otherwise the signal-to-noise ratio would not increase by cutting on the classifier output.

In the following I describe three algorithms, which construct classifiers with a uniform selection efficiency in the *fit-variables*. They are compared to two baseline models with non-uniform selection efficiency (Section 3.3.2.1) on the benchmark (Section 3.2.2).

The performance is compared in terms of the classification quality, measured as usual by the area under the receiver operating characteristic curve[7] (AUC ROC), and the uniformity of the selection efficiency. As described in [65], the uniformity of the selection efficiency can be measured by different approaches using: the standard deviation of the efficiency in bins (SDE); the Theil index (Theil); and the Cramér-von-Mises similarity (CvM). For all measures, a lower value indicates a more uniform selection efficiency.

All investigated algorithms were included in the `mva` package during this thesis.

3.3.2.1. Baseline

Two baseline models trained on the benchmark are used: a boosted decision tree implemented in `hep_ml` (BDT) and a neural network implemented in `tensorflow` (NN). The uniformity of the selection efficiency of the BDT and the NN is visualized in Figure 3.8 and Figure 3.9, respectively. As can be seen from the figures, both algorithms have a non-uniform selection efficiency for both signal and background.

3.3.2.2. Feature Drop

The simplest method to enforce a uniform selection efficiency is to remove all *features* which are dependent on the *fit-variables*. The dependency can be quantified

[7] Different quantities can be used to construct the receiver operating characteristic, here the signal efficiency and background rejection are used.

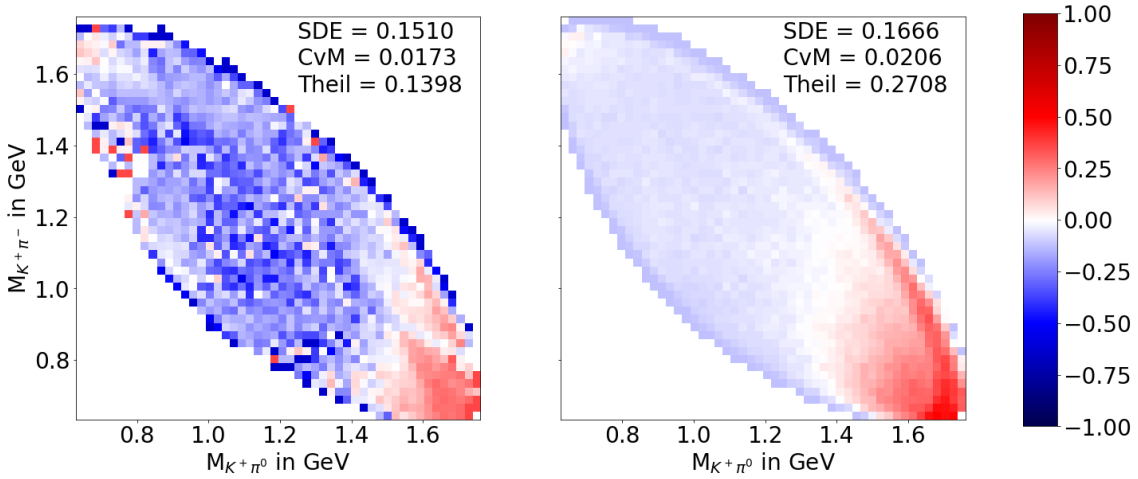


Figure 3.8.: **Baseline BDT**: Deviation from the average selection efficiency for the cut on the BDT response, which maximizes the signal-to-noise ratio. The left (right) plot shows signal (background) D^0 candidates.

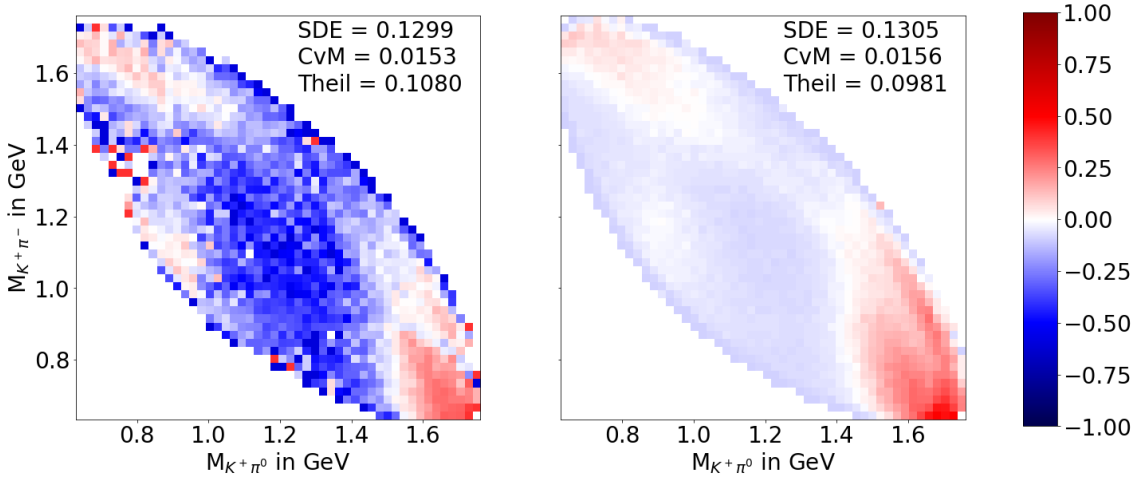


Figure 3.9.: **Baseline NN**: Deviation from the average selection efficiency for the cut on the neural network response, which maximizes the signal-to-noise ratio. The left (right) plot shows signal (background) D^0 candidates.

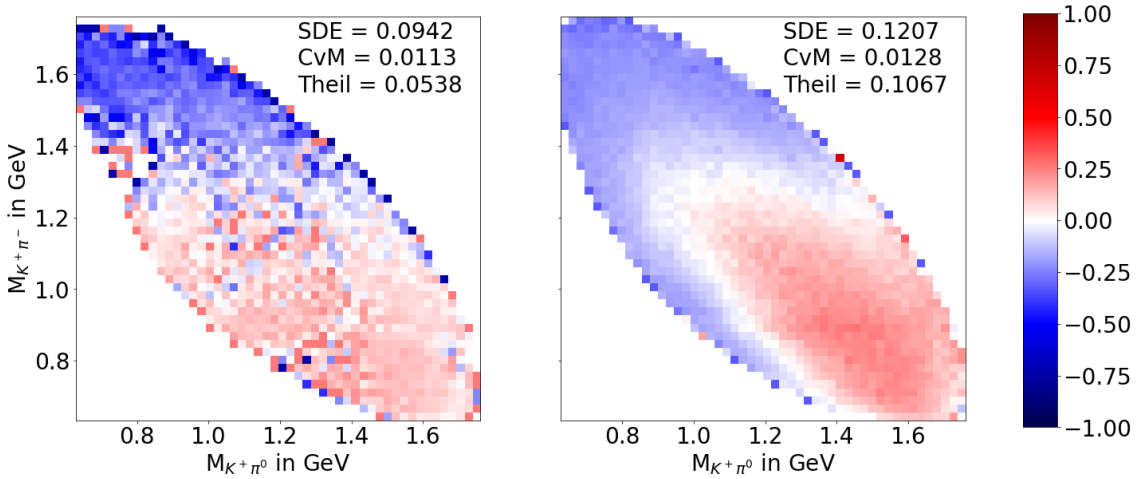


Figure 3.10.: **BDT with FeatureDrop**: Deviation from the average selection efficiency for the cut on the BDT response, which maximizes the signal-to-noise ratio. The left (right) plot shows signal (background) D^0 candidates. The BDT was trained on a subset of *features* without kinematic information of the daughter particles.

by: **linear correlation coefficients** (linear and ignores multivariate correlations); **statistical hypothesis tests** for two features being independent [66] (non-linear and ignores multivariate correlations); or the feature importance of the *features* in a **multivariate regression** to predict the *fit-variable* (non-linear and considers multivariate correlations).

Using handcrafted features designed to be uncorrelated to the *fit-variables*, e.g., the beam-constrained mass and ΔE as a parametrization of the four-momentum of a B meson, can further improve this approach.

A major drawback of this method is that the *features* with high separation power are often also dependent on the *fit-variables*. Therefore, one potentially looses a lot in terms of classification quality as is evident from Figure 3.14 and Figure 3.15. Furthermore, as can be seen from Figure 3.10 and Figure 3.11 even aggressively removing dependent *features* does not guarantee a uniform selection efficiency. Empirically, there is always some dependency between the *features* and the *fit-variables*, so enforcing an exact uniform selection efficiency would require removing all *features* from the training.

To sum up, more powerful techniques are required to enforce a uniform selection efficiency.

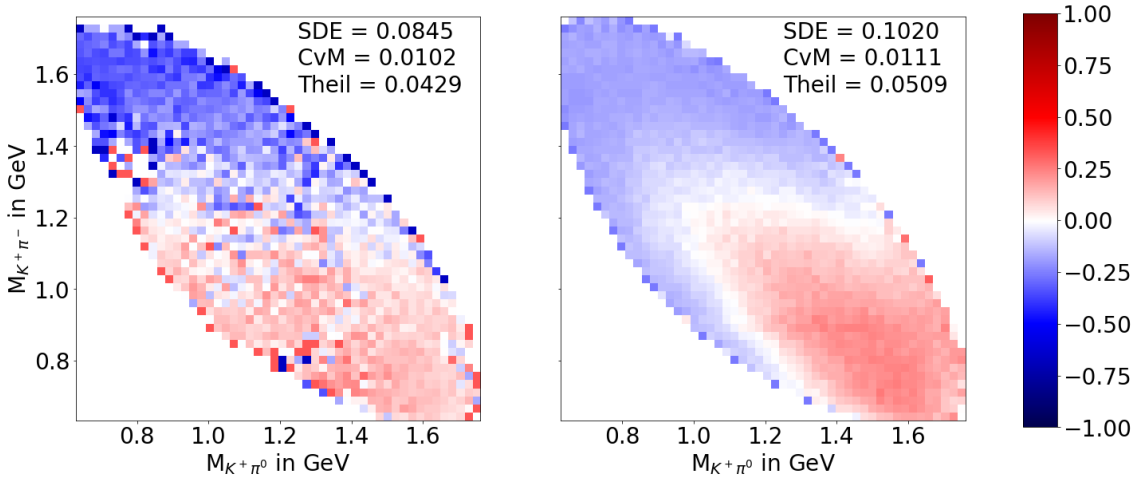


Figure 3.11.: **NN with FeatureDrop**: Deviation from the average selection efficiency for the cut on the NN response, which maximizes the signal-to-noise ratio. The left (right) plot shows signal (background) D^0 candidates. The NN was trained on a subset of *features* without kinematic information of the daughter particles.

3.3.2.3. Boosting to uniformity

As described in [65], boosting (re-weighting of events depending on the output of a classifier) can be used to enforce a uniform selection efficiency. This technique can be naturally incorporated in a boosting algorithm, like it is used for stochastic gradient-boosted decision trees. The loss-function is extended by an additional term $\mathcal{L}_{\text{flat}}$, which penalizes a non-uniform selection efficiency

$$\mathcal{L}_{\text{flat}} = \alpha \, \text{CvM}_S + \beta \, \text{CvM}_B, \tag{3.4}$$

where CvM is the Cramér-von-Mises similarity introduced in Section 3.3.2, and α and β are new hyper-parameters controlling the strictness of the uniformity constraint. The boosting algorithm will increase (decrease) the weight of events in bins[8] with lower (higher) selection efficiency compared to the inclusive distribution.

As can be seen from Figure 3.12 this algorithm yields a nearly uniform selection efficiency on the benchmark while maintaining a high classification quality (see Figure 3.14).

This algorithm was implemented into `FastBDT` (see Section 3.2.1.1) during this thesis, and is also (independently) available in the `hep_ml` package.

[8] A joint region in the *fit-variables*.

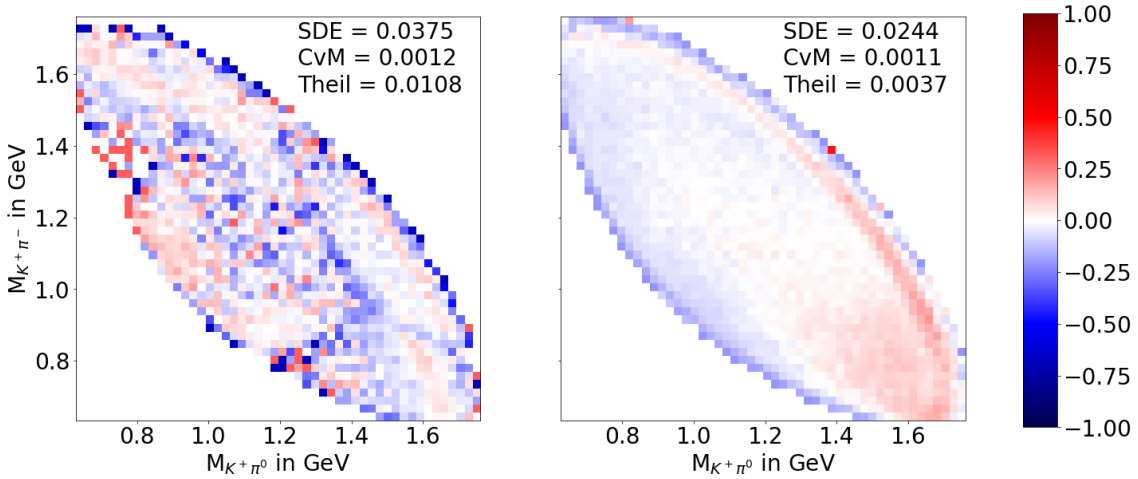


Figure 3.12.: **BDT with uniformity boosting**: Deviation from the average selection efficiency for the cut on the BDT response, which maximizes the signal-to-noise ratio. The left (right) plot shows signal (background) D^0 candidates.

3.3.2.4. Adversarial neural networks

Adversarial neural networks are commonly used in generative neural networks [41]. [42] applied the technique to create a neural network classifier $f(\vec{x}, \vec{w})$, whose output s is independent of the values $\vec{z}$ of some nuisance parameters Z

$$p(f(\vec{x}, \vec{w}) = s|\vec{z}) = p(f(\vec{x}, \vec{w}) = s|\vec{z}^\star) \tag{3.5}$$

for all $\vec{z}, \vec{z}^\star \in Z$ and all possible outputs s of the classifier. [42] proofs that the obtained classifier is both optimal[9] and fulfills Equation 3.5, if such a solution exists. Equation 3.5 implies that $f(X), Z$ are independent random variables, therefore a cut on the output of f has a uniform selection efficiency in Z by construction.

The technique was adapted during this thesis to create a uniform discriminator by identifying the nuisance parameters with the *fit-variables*. As explained in Section 3.3.2, a uniform selection efficiency is desired for signal and background separately. This can be achieved by additionally conditioning on the event class c

$$p(f(\vec{x}, \vec{w}) = s|\vec{z}, c) = p(f(\vec{x}, \vec{w}) = s|\vec{z}^\star, c) \tag{3.6}$$

for all $\vec{z}, \vec{z}^\star \in Z$, $c \in \{S, B\}$ and all possible outputs s of the classifier. In particular the random variables $f(X), Z$ are now conditionally independent instead of independent.

[9] In the sense defined in [42].

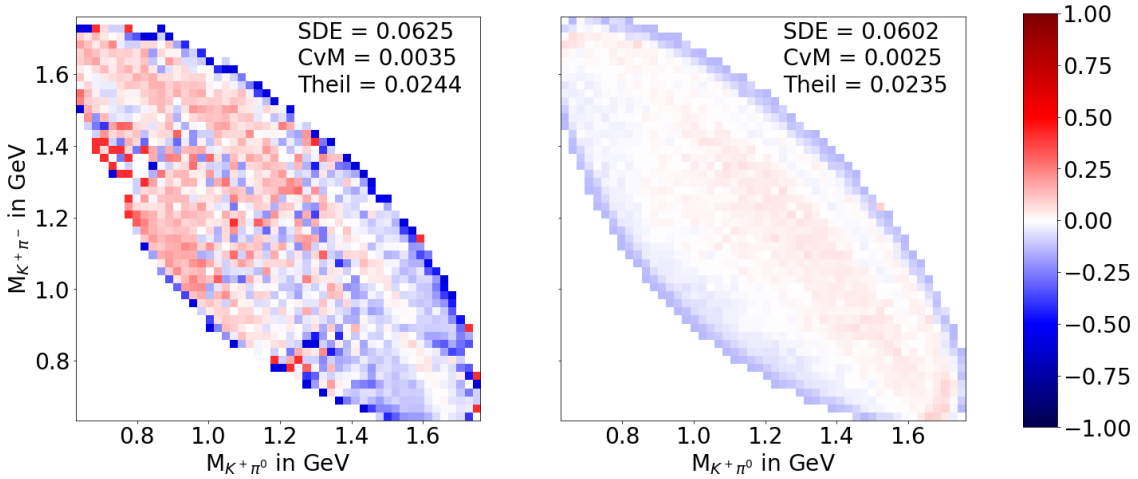


Figure 3.13.: **NN with adversary**: Deviation from the average selection efficiency for the cut on the NN response, which maximizes the signal-to-noise ratio. The left (right) plot shows signal (background) D^0 candidates.

For each pair (z_i, c_i) of *fit-variable* z_i and class c_i, an adversarial neural network is trained in parallel to the neural network f responsible for the classification. Each adversarial neural network learns to predict the probability density function $p(z_i|s, c)$ of its *fit-variable* and class; using the output s of f as the sole input. This can be accomplished by minimizing the loss-function $\mathcal{L}_i = -\log(p(z_i|s, c_i))$. The sum of all adversarial loss-functions $\mathcal{L}_i$ is subtracted from the loss-function $\mathcal{L}_f$ of the classification neural network

$$\mathcal{L} = \mathcal{L}_f - \alpha \sum_i \mathcal{L}_i, \tag{3.7}$$

where α is a new hyper-parameter controlling the strictness of the uniformity constraint. Hence, the adversarial neural networks learn to extract any information on the *fit-variables* contained in the output of f. At the same time the classification neural network is penalized if the adversarial neural networks succeed in doing so. All networks are trained simultaneously until convergence.

As can be seen from Figure 3.13 this algorithm yields a nearly uniform selection efficiency on the benchmark while maintaining a high classification quality (see Figure 3.15).

The usage of adversarial neural networks as presented in this section is key to a successful deployment of deep learning algorithms in HEP, because it can prevent the network from learning specific *fit-variables*. Examples include: The decay length difference Δz in CP measurements, and the beam-constrained mass M_{bc} in hadronic tagging algorithms.

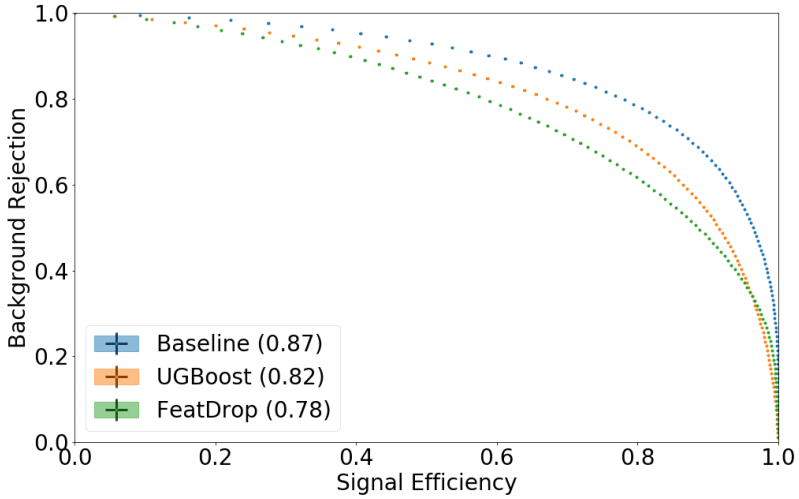


Figure 3.14.: Receiver operating characteristic (ROC) of: an ordinary BDT (baseline), a BDT using uniform boosting (UGBoost), an ordinary BDT trained on a subset of *features* without kinematic information of the daughter particles (FeatDrop). The number in parenthesis states the area under the corresponding ROC curve.

3.3.2.5. Conclusion

It is possible to ensure a uniform selection efficiency for signal and background separately: For BDTs the boosting to uniformity technique, and for NNs the adversary technique can be successfully employed. Both yield similar results in terms of classification quality and uniformity. The traditional Feature Drop approach is inferior.

The classification quality of the different algorithms, is compared in Figure 3.14 for the BDT and in Figure 3.15 for the NN. The uniformity constraint diminishes the classification quality, because information dependent on the *fit-variables* cannot be fully exploited by the classifiers.

Further algorithms exist to ensure a uniform selection efficiency. [67] used hand-crafted features to design decorrelated taggers for jet classification, and investigated the possibility to automatize the process with principal component analysis. However, the algorithm (as presented in [67]) still requires human input and is only shown to work with four *features* and one *fit-variable*. [68] introduced the first boosting to uniformity technique for boosted decision trees (distinct from the one presented in this thesis), and previously [46] presented already an approach based on event re-weighting by building a tree of neural networks. Both approaches suffer from

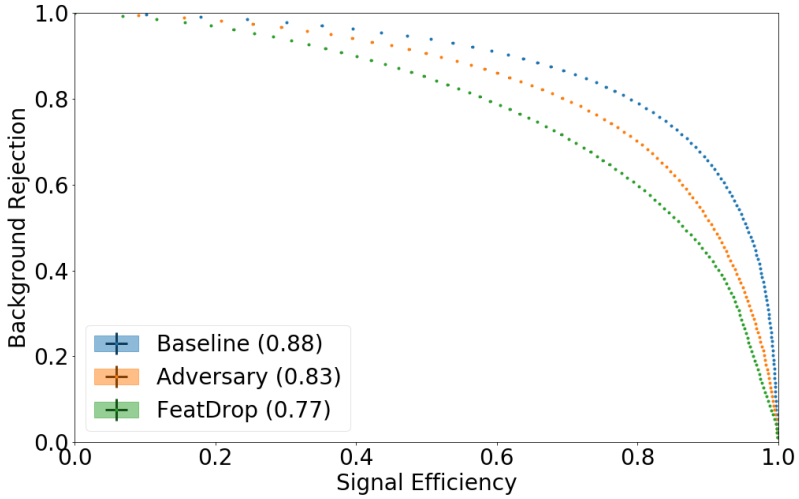


Figure 3.15.: Receiver operating characteristic (ROC) of: an ordinary NN (baseline), a NN with adversary (Adversary), an ordinary NN trained on a subset of *features* without kinematic information of the daughter particles (FeatDrop). The number in parenthesis states the area under the corresponding ROC curve.

a significant runtime performance drawback, because many classifiers have to be trained.

Finally, uniformity-constrained classifiers can be used to implicitly prevent using information available only on MC. Here, the *fit-variable* is a binary variable containing 0 for MC and 1 for detector data. For instance, the adversarial neural network is trained to distinguish MC and detector data, and the neural network responsible for the classification is punished if this is possible. The additional conditioning on the event class c is usually not desired (and usually not possible on data), because signal MC is easily distinguishable from detector data, if the signal is rare. The next section describes further data-driven techniques.

3.3.3. Data-driven

Many analyses in HEP employ Monte Carlo (MC) simulation to optimize cuts; determine the shape of signal and background components; and train multivariate methods. The simulated *MC events* are usually carefully calibrated and correction factors for the particle identification performance and branching fractions are used to ensure a sound modeling of the *detector data*. In addition, control regions and channels can be used to cross-check the simulation.

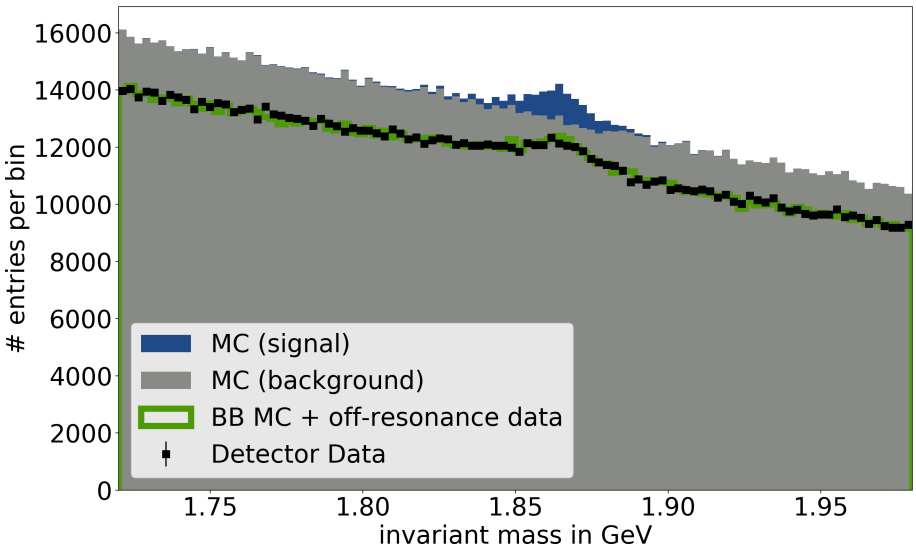


Figure 3.16.: The invariant mass of candidates reconstructed from the benchmark decay $D^0 \to K^-\pi^+\pi^0$. The filled histograms show the expected distribution from *Monte Carlo* simulation for signal (blue) and background (gray). The black points show the distribution obtained from *detector data*. The overall normalization differs from the *MC expectation*. There are two main background sources: $\Upsilon(4S)$ events, which is usually well understood; and continuum events, which have large uncertainties in the description of the hadronization and the branching fractions of the subsequent decay-chains. The green line shows *Monte Carlo* expectation, where the continuum component was replaced by off-resonance detector data. This distribution is in agreement with the *detector data*, indicating that the main difference is due to the continuum description.

As can be seen from Figure 3.16, already the well-known benchmark decay (Section 3.2.2) exhibits large differences between *detector data* and the *Monte Carlo* expectation. The benchmark decay was reconstructed on *Monte Carlo* and *detector data* using the `b2bii` package. The main differences arise from the poor description of the continuum component, that is non-resonant processes like $e^+e^- \to c\bar{c}$ (see Section 2.3.3.3).

Data-driven techniques can be used to reduce the reliance of analyses on *Monte Carlo* simulation. Hereby, the desired information is determined directly from data by exploiting physics knowledge.

In the remainder of this section I introduce several data-driven techniques which aim to reduce the dependency on *Monte Carlo* simulation during the fitting-phase of multivariate methods. The first three techniques are originally described in [46]. All of them were investigated and implemented in the `mva` package during this thesis and a supervised bachelor thesis [69].

3.3.3.1. Baseline

As a baseline for the following studies, a boosted decision tree (BDT) was trained to distinguish signal and background candidates on the benchmark decay. Figure 3.17 shows the distribution of the invariant mass of the benchmark decay after a cut on the response of the BDT. The cut was chosen to optimize the signal-to-noise ratio on *Monte Carlo*. The observed signal-to-noise ratio on data is 31.78. It was extracted by fitting a background-only model to the observed distribution on *detector data* excluding the signal peak between 1.8 GeV and 1.9 GeV.

The training of the same boosted decision tree is repeated in the following sections using different data-driven techniques. Throughout the text the applied cuts on the response of the BDTs are always chosen to maximize the signal-to-noise ratio on *MC events* and the cited signal-to-noise ratio is always extracted on *detector data* as described above.

To ensure a sound comparison, only features independent of the invariant mass were included in the training, otherwise a possible peaking background contribution could distort the evaluation on *detector data*. This corresponds to the Feature Drop ansatz explained in the last section.

3.3.3.2. Event re-weighting

The main idea of **event re-weighting** is to encode the difference of *MC events* and *detector data* in the response of a multivariate classifier. A classifier is trained to distinguish *MC events* and *detector data*. Its response is the probability $p = \frac{\text{detector}}{\text{detector+MC}}$ of selecting a *detector event* from the whole sample. An *MC event* with a high (low) probability is likely (unlikely) to appear in *detector data*, hence we want to increase (decrease) the importance of this event. The importance of each *MC event* is defined by a weight $w = \frac{\text{detector}}{\text{MC}} = \frac{p}{1-p}$ (see [46]).

The larger the differences between *MC events* and *detector data*, the better the classification quality of the classifier. In the absence of mis-modeling, the classifier would not be able to distinguish *MC events* and *detector data*.

In contrast to simple scaling factors derived from a binned distribution, this method takes multivariate dependencies, between all features used in the classifier, into account.

Subsequently, the user can train another multivariate classifier to distinguish signal from background on the re-weighted MC dataset. In particular deep learning methods, which typically employ numerous low-level features can profit from this technique (see Section 3.1.4.3). Furthermore, the derived weights w can be used for all tasks performed on MC like determining cuts and creating plots.

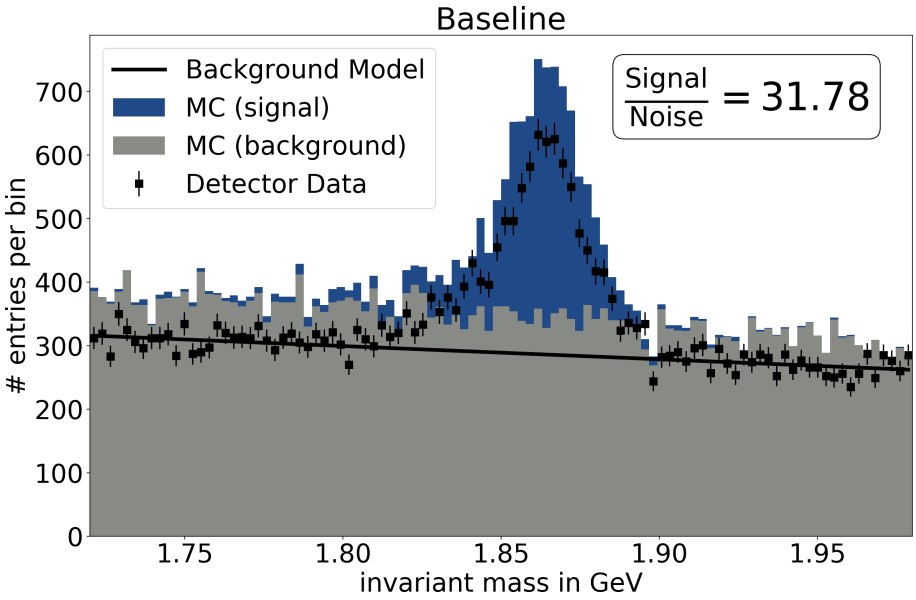


Figure 3.17.: The invariant mass distribution of the benchmark decay $D^0 \to K^-\pi^+\pi^0$, after a cut on the response of a boosted decision tree was applied. The BDT was trained on *MC events*. The filled histograms show the expected distribution from *Monte Carlo* simulation for signal (blue) and background (gray). The black points show the distribution obtained from *detector data*. The black line shows the background model fitted to the *detector data*

This methods cannot compensate differences in phase-space regions where no *MC events* exist. Also, the method is not aware of signal and background, and could re-weight signal (background) events from MC to match background (signal) events in data.

Alternatively, one can use a uniformity-constraint classifier to mitigate differences between *MC events* and *detector data* as described in Section 3.3.2.5. This approach changes the underlying classifier itself, hence it cannot be automatized for arbitrary classifiers.

The **event re-weighting** technique was implemented in the `mva` package, and is automatically applied as soon as the user provides the necessary *detector data*. Figure 3.18 shows the invariant mass distribution of the benchmark decay, after a cut on a BDT trained on re-weighted *MC events*. The achieved signal-to-noise ratio does not differ significantly from the baseline (see Figure 3.17). The re-weighted expectation from MC is in agreement with the measured spectrum on detector data.

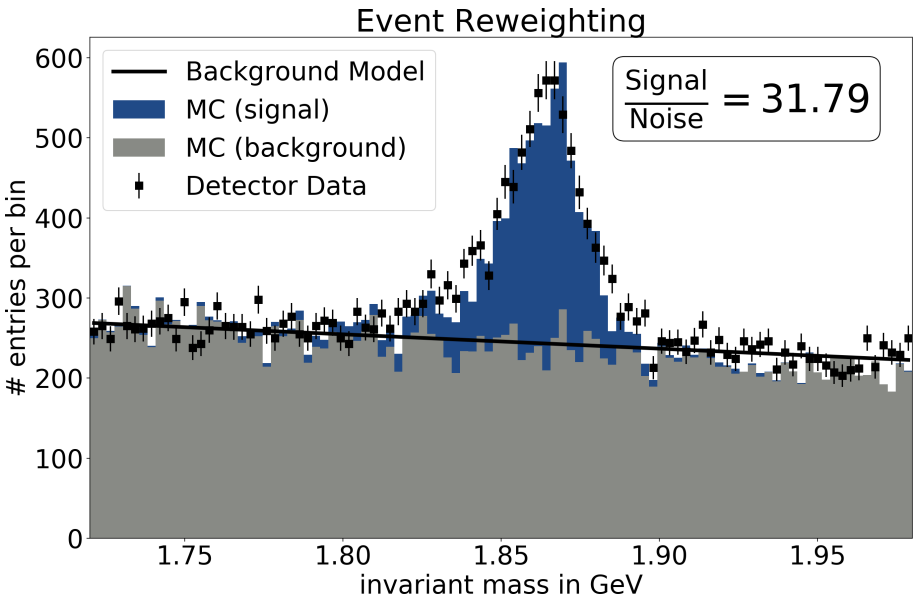


Figure 3.18.: The invariant mass distribution of the benchmark decay $D^0 \to K^-\pi^+\pi^0$, after a cut on the response of a boosted decision tree was applied. The BDT was trained on re-weighted *MC events*. The filled histograms show the expected distribution from *Monte Carlo* simulation for signal (blue) and background (gray). The black points show the distribution obtained from *detector data*. The black line shows the background model fitted to the *detector data*.

3.3.3.3. Sideband-Subtraction

Training a multivariate classifier directly on data is challenging, because usually there are no pure signal and background samples available. However, in some situations signal-enriched and signal-free phase-space regions can be identified. These regions can be used to train a multivariate classifier on data, by statistically subtracting the background in the signal-enriched region using events from a signal-free region with a negative weight.

Sideband-Subtraction requires the definition of three phase-space regions: a **signal-enriched region**, which contains signal and background events; a **background-region**, which contains only background events; and a **negative signal-region**; which contains only background events indistinguishable[10] from the background events in the signal-enriched region. In addition, the method requires the expected number of signal events in the signal-enriched region e.g. from *Monte Carlo*.

[10] Meaning that the events cannot be distinguished by the features used in the training of the classifier.

Since the ground truth is not known for individual *detector events*, all events (including the background events) in the signal-enriched region are used as signal in the training. To counteract the background events contained in the signal region, the background events from the negative-signal region are used as signal as well, but with a negative weight. Finally, the events from the background region are used as background.

Figure 3.19 shows the invariant mass distribution of the benchmark decay, after a cut on a BDT trained on *detector data* using Sideband-Subtraction. The exact choice of the regions, as long as they fulfill the assumptions and contain enough statistics, is not important. The different regions were chosen as follows: the signal-enriched region $1.82\,\text{GeV} < M < 1.9\,\text{GeV}$, the background region $1.77\,\text{GeV} < M < 1.78\,\text{GeV}$ and $1.94\,\text{GeV} < M < 1.95\,\text{GeV}$, and the negative signal-region $M < 1.8\,\text{GeV}$ and $1.92\,\text{GeV} < M$. The achieved signal-to-noise ratio is worse than the baseline (see Figure 3.17).

The method can be used to increase the signal-to-noise ratio in situations were *Monte Carlo events* cannot be used. The expected number of signal events in the signal-enriched region has to be known. If the classifier is able to distinguish background events in the signal and negative signal regions using its features, the classification quality deteriorates, because the negative signal events are partly rejected and cannot subtract all background events in the signal-enriched region. In consequence, only a subset of features can be used during the Sideband-Subtraction training.

3.3.3.4. sPlot

sPlot is a statistical technique, to reconstruct the distribution of different *sources* in so-called *control variables*, using the known distributions of these sources in a *discriminating variable* [45]. In the context of machine learning it can be used to perform a multivariate training on data.

The *sources* are chosen to be signal and background decays, the *control variables* are the features we want to use in the training of the multivariate classifier, and the *discriminating variable* is a variable for which the signal and background distribution is known e.g. the invariant mass of a particle.

The underlying idea is similar to Sideband-Subtraction. The known distributions in the *discriminating variable* define implicitly signal-enriched and background-enriched regions. In contrast to Sideband-Subtraction, sPlot does not require a signal-free region.

In the sPlot based multivariate training every event is used twice, once as signal and once as background, but with different weights. The weights for signal w_s and background w_b are derived from the probability density distributions of the

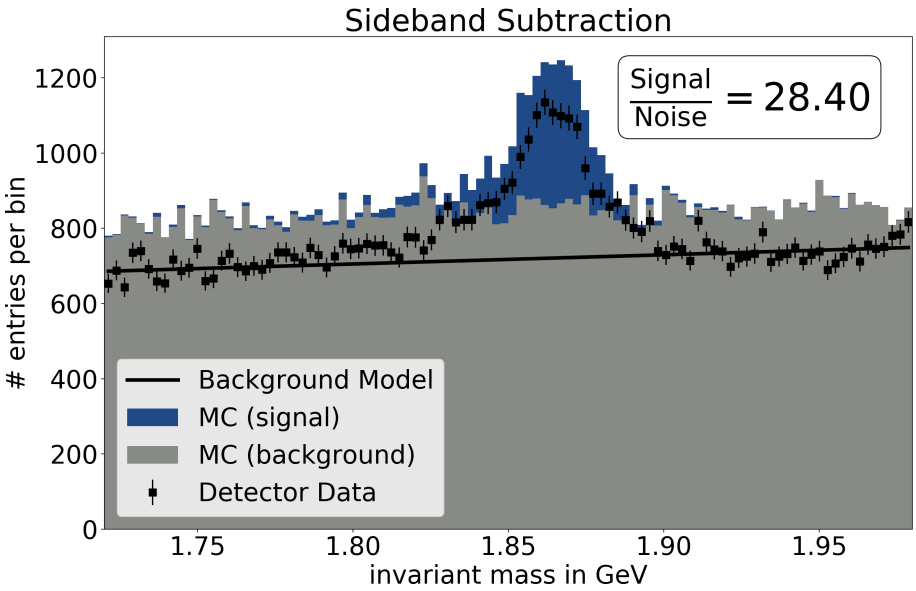


Figure 3.19.: The invariant mass distribution of the benchmark decay $D^0 \to K^-\pi^+\pi^0$, after a cut on the response of a boosted decision tree was applied. The BDT was trained using sideband subtraction. In particular, no *MC events* were used during this training. The filled histograms show the expected distribution from *Monte Carlo* simulation for signal (blue) and background (gray). The black points show the distribution obtained from *detector data*. The black line shows the background model fitted to the *detector data*.

discriminating variable for signal f_s and background f_b. These are fitted to the *detector data* in order to extract the number of signal N_s and background events N_b:

$$\begin{pmatrix} w_s \\ w_b \end{pmatrix} = \frac{V}{N_s \cdot f_s + N_b \cdot f_b} \cdot \begin{pmatrix} f_s \\ f_b \end{pmatrix},$$

where V is the covariance matrix of the extracted yields. The approach can be generalized to more than one background source.

sPlot requires the probability density distribution of the *discriminating variable* for signal and all background sources. In addition, the events of each source must be conditional independent of the *discriminating variable*, meaning the classifier should not be able to estimate the value of the *discriminating variable* using its features. In consequence, only a subset of features can be used during the sPlot training. A violation of the above assumptions leads to a deterioration of the achieved classification quality. This is the same effect already encountered in Section 3.3.3.3.

Figure 3.20 shows the invariant mass distribution of the benchmark decay, after a cut on a BDT trained on *detector data* using sPlot. The probability density

functions for the invariant mass for signal and background is extracted from Monte Carlo simulation. The achieved signal-to-noise ratio is worse than the baseline (see Figure 3.17). The more detailed description of the method can be found in [69] and [46].

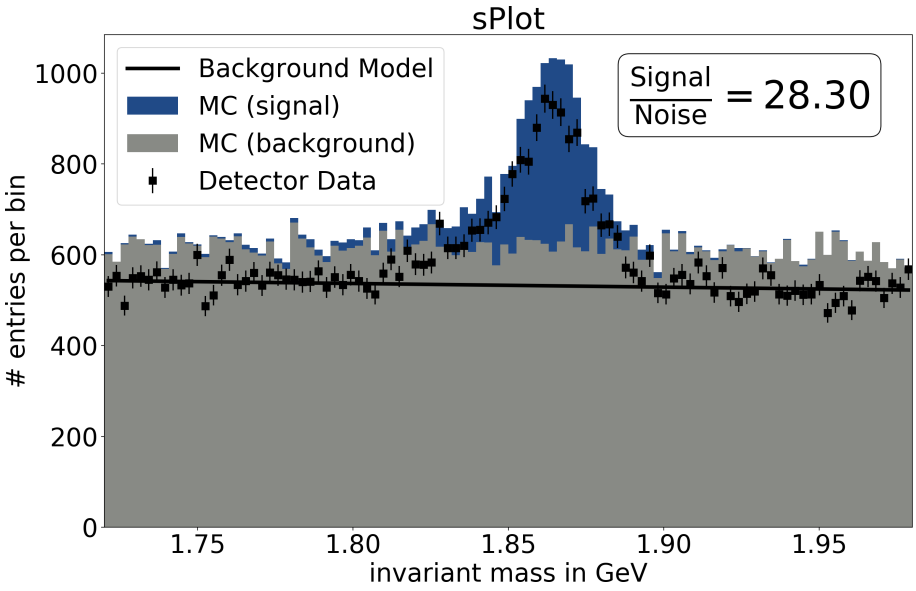


Figure 3.20.: The invariant mass distribution of the benchmark decay $D^0 \to K^-\pi^+\pi^0$, after a cut on the response of a boosted decision tree was applied. The BDT was trained using sPlot. In particular, no *MC events* were used during this training. The filled histograms show the expected distribution from *Monte Carlo* simulation for signal (blue) and background (gray). The black points show the distribution obtained from *detector data*. The black line shows the background model fitted to the *detector data*.

3.3.3.5. decorrelated sPlot

As stated in the last section, the classification quality of an sPlot training deteriorates if the classifier can predict the *discriminating variable* for background-events using its features. Using the uniformity-constraint is not straight-forward in this case (see Section 3.3.3.6).

During this thesis an extension of sPlot was developed to mitigate this problem, by automatically reducing the conditional dependency between the features and the *discriminating variable* for background-events.

The basic idea is to perform an event re-weighting (similar to Section 3.3.3.2) to eliminate the undesired conditional dependency. Therefor, a classifier is trained to

distinguish events with a low L and high H value of the *discriminating variable*. The main challenge is to leave the dependency between the features and the target (signal or background) untouched.

One possible approach would be to assign all events left of the signal peak to L and all events right of the signal peak to H. In case of a symmetrical signal peak distribution in the *discriminating variable*, there would be the same amount of signal in L and H, in consequence the classifier would learn to distinguish L and H but not signal and background, hence leaving the dependency between the features and the target untouched.

A more advanced approach uses all events twice (similar to sPlot), once as L and once as H, weighted using the cumulative distribution function for signal and normalized to the probability distribution function of background:

$$w_L = \frac{\int_{-\infty}^{x} f_s(x')\mathrm{d}x'}{f_b(x)}$$
$$w_H = \frac{\int_{x}^{\infty} f_s(x')\mathrm{d}x'}{f_b(x)},$$

where x is the *discriminating variable* and f_s (f_b) is the probability density function for signal (background). This approach is independent of assumptions on the signal distribution shape, but still distributes the amount of signal per background evenly between L and H.

In the next step the output p of the previously trained L vs. H classifier, is used to assign a decorrelation weight to the event

$$w_D = \frac{1}{2}\left(\frac{\int_{-\infty}^{x} f_s(x')\mathrm{d}x'}{p} + \frac{\int_{x}^{\infty} f_s(x')\mathrm{d}x'}{1-p}\right),$$

in order to remove the conditional dependency between features and the *discriminating variable* for background-events. The cumulative distribution function for signal is again used to leave the signal-events on average untouched. If the classifier could not decide between L and H, that is the probability is close to $p = 0.5$, the final weight is close to $w_D = 1$. In contrast, a stark prediction like a value close to $p = 1$, will lead to a small (large) weight if the prediction is correct (incorrect).

Finally, an sPlot training as described in Section 3.3.3.4 is performed using the decorrelation weights w_D multiplied with the sPlot weights.

Figure 3.21 shows the invariant mass distribution of the benchmark decay, after a cut on a BDT trained on *detector data* using decorrelated sPlot. As expected the obtained signal-to-noise ration is better than the one obtained by pure sPlot. Furthermore, the achieved signal-to-noise ratio is not significantly different from the baseline (see Figure 3.17).

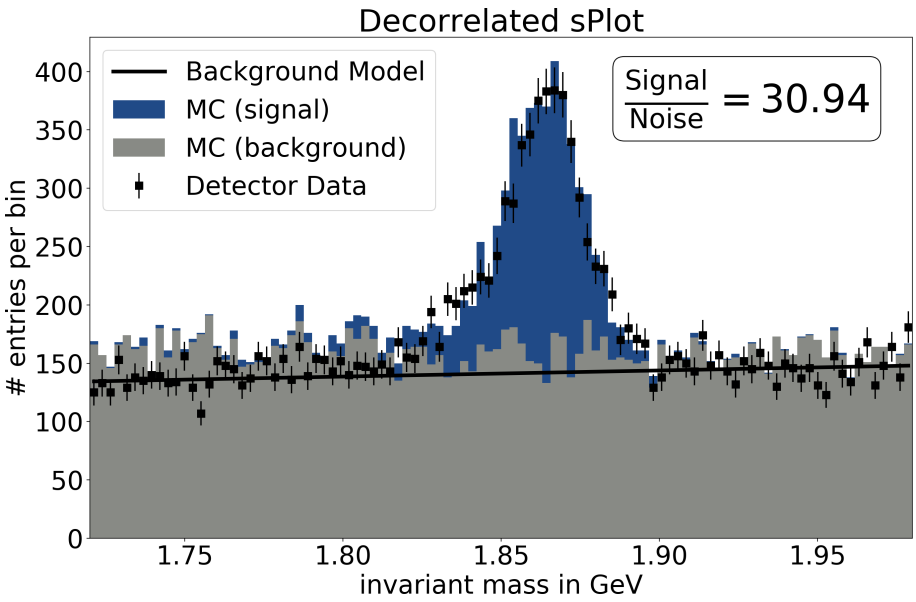


Figure 3.21.: The invariant mass distribution of the benchmark decay $D^0 \to K^-\pi^+\pi^0$, after a cut on the response of a boosted decision tree was applied. The BDT was trained using decorrelated sPlot. In particular, no *MC events* were used during this training. The filled histograms show the expected distribution from *Monte Carlo* simulation for signal (blue) and background (gray). The black points show the distribution obtained from *detector data*. The black line shows the background model fitted to the *detector data*.

3.3.3.6. Conclusion

All presented methods yield reasonable results on *detector data*. They differ in the amount of information they require from *Monte Carlo* simulation: event reweighting requires simulated *MC events* to which corrections can be applied; Sideband-Subtraction requires an estimate of the number of signal events in the signal-enriched region; (decorrelated) sPlot requires the distribution of the signal and background component. In addition, Sideband-Subtraction and sPlot require that the background composition does not change in the considered phase-space region, and that the features used during the training are independent of the *discriminating variable*. The last requirement can be mitigated by the decorrelated sPlot algorithm developed during this thesis.

In consequence, the preferred method depends on the quality of the *Monte Carlo* simulation and the correctness of the necessary assumptions for the investigated tasks.

In the benchmark example presented in this section, the classifier based on pure *Monte Carlo*, still performs slightly better or equally well compared to the data-

driven methods. This is expected since the benchmark decay is a well-known and well-simulated decay-channel, except for the continuum component, where the overall normalization is off. The event re-weighting method is particularly useful, because it yields the same results as the pure *Monte Carlo* training, if *MC events* and *detector data* are indistinguishable. It is key to a successful employment of deep learning methods, where low-level features are used, which are usually less well simulated compared to high-level features.

It is not clear yet, if the data-driven techniques presented in this section can be combined with the uniformity-constraint introduced in the last section. While most data-driven technique require a *discriminating-variable* to define signal and background events, the uniformity-constraint is enforced using this definition on the *fit-variable*, which could be the same as the *discriminating-variable* (like it is the case for sPlot). This was not further investigated during this thesis.

3.3.4. Hyper-parameter optimization

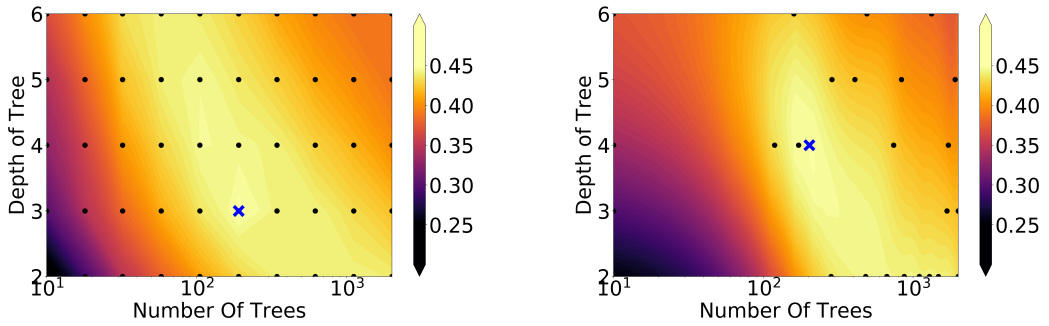
As described in Section 3.1.3, the complexity of the statistical model generated by an MVA algorithm can be controlled by so-called hyper-parameters. While the MVA algorithm optimizes the parameters of the statistical model to minimize the error on a given training dataset, the hyper-parameter optimization tunes the hyper-parameters of the MVA algorithm to minimize the error on an independent validation dataset.

In total, three independent datasets are required in this situation: the training dataset used by the MVA algorithm; the validation dataset used by the hyper-parameter optimization; and the test dataset used to evaluate the performance of the selected model. Techniques like cross-validation [20, Chapter 1.3] and Bayesian model comparison [20, Chapter 3.4] can be used to avoid introducing a dedicated validation dataset if the amount of available labeled data is limited.

Optimizing hyper-parameters by hand is cumbersome, error-prone and not suitable for automation. During this thesis different algorithms, used to automatize the search for the optimal hyper-parameter values, were investigated. The `mva` package includes examples for all of them. Figure 3.22 shows the application of the hyper-parameter optimization algorithms, described in the following, to the hyper-parameters of `FastBDT` on the benchmark.

3.3.4.1. Grid and Random Search

The most common approach is a grid search. The loss on the validation dataset is evaluated on an n-dimensional grid in the hyper-parameter space. The computational effort of this approach scales exponentially in the number of hyper-parameters. The



(a) Grid-search evaluated on 50 grid points. (b) Bayesian optimization after 26 evaluations.

Figure 3.22.: Hyper-parameter optimization of `FastBDT` on the benchmark (see Section 3.2.2). Two hyper-parameters were optimized: the `number of trees` and the `depth of the trees`. The contour map indicates the ROC AUC score predicted by the measured hyper-parameter settings. The blue cross indicates the best hyper-parameter configuration with a ROC AUC score of 0.45 (both algorithms found distinct but equivalent maximums)

accuracy of the method depends on the density of the grid. Quantitative[11] and qualitative[12] hyper-parameters can be optimized by this approach.

Figure 3.22a shows an application of the grid search algorithm on the benchmark. In total, 50 distinct configurations of `FastBDT` were tested, the best achieved ROC AUC score was 0.450, which is significantly better than the default configuration achieving 0.435 (see Table 3.1).

Recent studies suggest that random-search is superior to grid-search at least in high-dimensional scenarios [31]. This is due to the fact that often only a few out of many hyper-parameters really matter. In such a scenario, the computational effort of random-search is independent of the number of nuisance[13] hyper-parameters

3.3.4.2. Bayesian Optimization

A probabilistic $f(\vec{h})$ model of the error on the validation dataset given the hyper-parameters $\vec{h}$ is constructed. Using the Bayesian Optimization framework [33] this

[11] Quantitative parameters have an intrinsic ordering, examples are the `number of trees` in an `BDT`, or the `weight-decay` constant in the loss function of a `NN`.

[12] Qualitative parameters do not have an intrinsic ordering, examples are the separation gain measure in a `BDT` and the optimization algorithm of a `NN`.

[13] A hyper-parameter with minor or no influence on the score.

model can be used to determine the next hyper-parameter space-point to evaluate using all available information from previous evaluations. This approach can take into account prior knowledge about the hyper-parameters and the cost (runtime) of the evaluation for different hyper-parameters. Quantitative and qualitative hyper-parameters can be optimized by this approach.

Figure 3.22b shows an application of the Bayesian optimization algorithm on the benchmark. After 26 evaluations, the model found a configuration with an ROC AUC score of 0.450, which is significantly better than the default configuration achieving 0.435 (see Table 3.1). In comparison to grid-search, Bayesian optimization found an equivalent best hyper-parameter configuration in fewer evaluations.

3.4. Conclusion

In the early years of HEP cut-based analysis (see Section 3.1.4.1) dominated the field, in the final years of the last millennium HEP started to adopt multivariate analysis methods and machine learning (see Section 3.1.4.2), today the fields move on to deep learning (see Section 3.1.4.3).

The `mva` package enables Belle II physicists to keep up with the rapid developments in the field and to easily employ modern machine learning algorithms in their work. Most of the multivariate methods used in the reconstruction and analysis algorithms in `BASF2` are built on the `mva` package and use the default classification method `FastBDT`, both developed during this thesis.

In particular, tagging algorithms, which exploit the unique environment provided by B factories, such as the `Flavour Tagger` and `Full Event Interpretation`, rely heavily on the `mva` package.

The provided algorithms were tested, where appropriate, on data recorded by the Belle experiment by taking advantage of the `b2bii` package.

Chapter 4

Full Event Interpretation

The `Full Event Interpretation (FEI)` is a tagging algorithm based on machine learning. It exploits the unique experimental setup of B factory experiments such as the Belle and Belle II experiment. Both experiments operate on the $\Upsilon(4S)$ resonance, which decays at least 96% of the time into exactly two B mesons. Conceptually, the event is divided into two sides: The signal-side containing the tracks and clusters compatible with the assumed signal B_{sig} decay the physicist is interested in, e.g. $B^+ \to \tau^+\nu_\tau$; and the tag-side containing the remaining tracks and clusters compatible with an arbitrary B_{tag} meson decay. Figure 4.1 depicts this situation. Ideally, a **full** reconstruction of the entire **event** has to take all detected tracks and clusters into account to attain a correct **interpretation** of the measured data.

The `FEI` automatically reconstructs B_{tag} candidates[1] and calculates a signal probability to separate correctly reconstructed B_{tag} candidates from background. Using constraints derived from the unique experimental setup, the reconstructed B_{tag} meson can be used to recover information about the remaining B_{sig} meson in the event.

This enables the measurement of a wide range of decays with a minimum amount of detectable information, like $B^+ \to \tau^+\nu_\tau$, $B^+ \to \ell^+\nu\gamma$ and $B \to K\nu\bar{\nu}$, or no detectable information at all, like in the case of $B^0 \to \nu\bar{\nu}$, in the final state.

In the following I describe tagging at B factories in general (Section 4.1); the `FEI` algorithm itself (Section 4.2); the validation of the performance of the `FEI` (Section 4.3) using converted Belle detector-data and simulated Belle II Monte Carlo; and finally possible future extensions and use-cases (Section 4.4).

[1] A candidate consists of an assumed decay-chain, which happened in the detector and was detected by a fixed set of tracks and clusters.

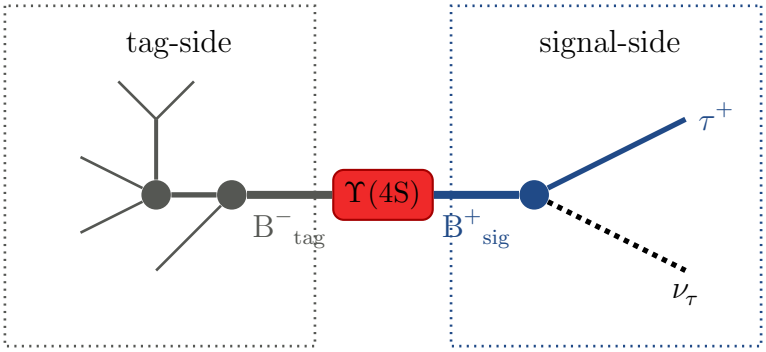


Figure 4.1.: (Left) a common tag-side decay $B^- \to D^0[\to K^0_S[\to \pi^-\pi^+]\pi^-\pi^+]\pi^-$ and (right) the signal-side decay $B^+ \to \tau^+\nu_\tau$ investigated in Chapter 5.

4.1. Tag-side reconstruction

The initial four-momentum of the produced $\Upsilon(4S)$ resonance is precisely known. Therefore, the reconstruction of the tag-side B meson allows to recover information about the signal-side, which would be otherwise inaccessible. The recovered information includes:

the consistency of quantities like the charge or the flavour;

the four-momentum of the B_{tag} and consequently the B_{sig} meson;

the decay-vertex and the decay time difference Δt between both mesons;

the event-type like $\Upsilon(4S) \to B^0\overline{B}^0$ or $\Upsilon(4S) \to B^+B^-$;

and the assignment of tracks and clusters to either the B_{tag} or the B_{sig} meson.

Historically, there were two distinct approaches to tagging at B factories: inclusive and exclusive.

They differ in their **tagging efficiency** (that is the fraction of $\Upsilon(4S)$ events which can be tagged), their **tag-side efficiency** (that is the fraction of $\Upsilon(4S)$ events with a correct tag) and in the quality of the recovered information, which determines the **purity** (that is the fraction of the tagged $\Upsilon(4S)$ events with a correct tag-side) of the tagged events. These three properties are the key performance indicators used in this thesis. They are closely related to important properties of a specific analysis: The tagging efficiency is important to judge the disk-space required for skimming, that is the number of events which have to be considered for the analysis; the tag-side efficiency influences the effective statistics of the analysis, and the purity is related to the signal-to-noise ratio of the analysis.

Exclusive tagging provides the assignment of tracks and clusters to either the tag-side and or the signal-side, whereas the inclusive tag requires this assignment as input. The advantages and disadvantages of the different approaches are visualized in Figure 4.2.

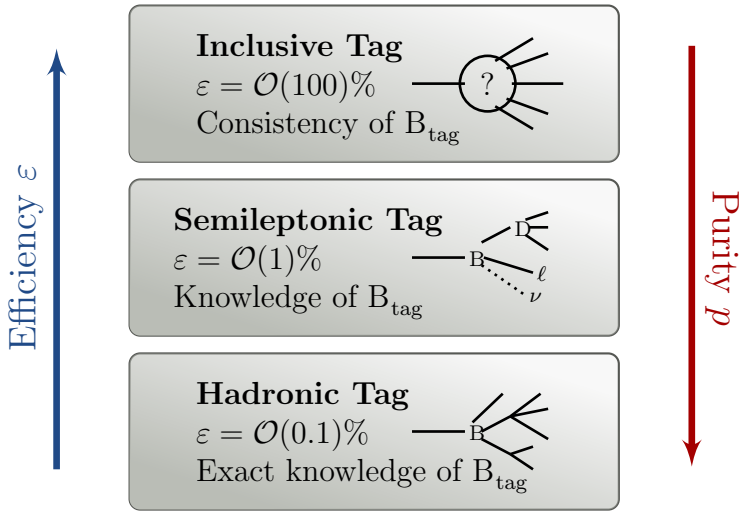


Figure 4.2.: Overview of the three tagging algorithms, which were used in the past to infer information about the B_{sig} meson using the other B_{tag} meson in the event.

4.1.1. Inclusive Tagging

Inclusive refers to the reconstruction of a particle (here the B_{tag}) without assuming an explicit decay-channel. Consequently, inclusive tagging combines the four-momenta of all tracks and clusters, which were not used during the reconstruction of the B_{sig} – the so-called rest of event. The decay-chain of the B_{tag} is not explicitly reconstructed, hence the assignment is not provided by the algorithm and cannot be used to discard wrong candidates. Only the overall consistency between the B_{sig} and B_{tag} meson can be checked. This approach has a high tagging efficiency of $\mathcal{O}(100)\%$, since it can always provide a valid B_{tag}, but suffers from a high background, and consequently the tagged sample is very impure.

Inclusive tagging is used in time-dependent CP violation analyses, to determine the decay vertex of the B_{tag} meson. It can also be used to improve the momentum resolution on the signal-side in analyses like $B^+ \to \mu^+\nu$, where the signal-side alone is suffice to provide a pure sample. Flavour tagging (see Section 3.3.1.2) and continuum suppression (see Section 3.3.1.3) can be seen as a special form of inclusive tagging. On the other hand, inclusive tagging cannot be used for an inclusive signal-side.

The `FEI` is not an inclusive tagging algorithm. However, it does try to always provide a valid B_{tag} like it is accomplished by the inclusive tagging. Therefore, it could be used instead of inclusive tagging as described in Section 4.4.1.

4.1.2. Exclusive Tagging

Exclusive refers to the reconstruction of a particle (here the B_{tag}) assuming an explicit decay-channel. Consequently, exclusive tagging reconstructs the B_{tag} independently of the B_{sig} using either hadronic or semileptonic B meson decay-channels. The decay-chain of the B_{tag} is explicitly reconstructed and therefore the assignment of tracks and clusters to the tag-side and signal-side is known.

If the signal-side is exclusively reconstructed as well, the entire decay-chain of the $\Upsilon(4S)$ is known. Consequently, all tracks and clusters measured by the detector should be accounted for. In particular, the requirement of no additional tracks, besides the ones used for the reconstruction of the $\Upsilon(4S)$, is an extremely powerful and efficient way to remove nearly all reducible[2] background. This requirement is called the **completeness-constraint** throughout the text.

The tagged sample is pure, but it suffers from a low tagging efficiency $\mathcal{O}(1)\%$, since only a tiny fraction of the B decays can be explicitly reconstructed, due to the large amount of possible decay-channels and their high multiplicity, as well as the imperfect reconstruction efficiency of tracks and clusters. For instance, the probability p to find all tracks of a hypothetical decay chain with a multiplicity of 10, assuming a reconstruction efficiency of 0.95 per track and a geometrical detector acceptance of 91.1% is only $p \approx 24\%$.

The quality of the recovered information and systematic uncertainties depend on the decay-channel of the B_{tag}, therefore one distinguishes further between hadronic and semileptonic exclusive tagging.

Additionally, exclusive tagging can also be used to perform an inclusive reconstruction of the signal-side, without assuming an exclusive signal decay-channel.

The `FEI` is an exclusive tagging algorithm, which supports hadronic and semileptonic tagging.

4.1.2.1. Hadronic Tagging

Hadronic tagging solely uses hadronic B decay-channels for the reconstruction. Hence, the four-momentum of the reconstructed B_{tag} is well-known and the tagged sample is very pure. However, hadronic tagging suffers from a low tagging efficiency. It is only possible for a tiny fraction of the recorded events, because the branching fraction of explicit fully hadronic decay-chains is very small. A typical hadronic B decay has a branching fraction of $\mathcal{O}(10^{-3})$.

[2] Reducible background has distinct final state products from the signal.

Figure 4.5 and Figure 4.10 show the beam-constrained mass distribution of hadronically reconstructed B_{tag} mesons for Belle and Belle II, respectively. The correctly reconstructed mesons peak at a beam-constrained mass of $M_{bc} = 5.279\,\text{GeV}$. The beam-constrained mass can be calculated from the initial beam-conditions [1, Chapter 7.1]:

$$M_{bc} = \sqrt{E_{beam}^2 - p_B^2}, \tag{4.1}$$

where all quantities are measured in the center-of-mass system and E_{beam} is the beam energy. M_{bc} is per definition independent of the mass hypotheses of the final state particles used during the reconstruction of the B meson candidate.

Another useful quantity is the deviation from the beam-energy [1, Chapter 7.1]

$$\Delta E = E_B - E_{beam}, \tag{4.2}$$

which is sensitive to mis-identification of the final state particles during the reconstruction of the B meson candidate.

The above beam-dependent quantities are less correlated with each other in comparison to the invariant mass M and energy E of the reconstructed B meson candidate.

The hadronic tag of the `FEI` can be accessed by the user using `BASF2` via the provided `B0:generic` and `B+:generic` `ParticleLists`. The different decay-channels can be distinguished by the associated `decayModeID`.

4.1.2.2. Semileptonic Tagging

Semileptonic tagging uses semileptonic B decay-channels. Due to the higher branching fraction of semileptonic B decays this approach usually has a higher tagging efficiency compared to hadronic tagging. A typical semileptonic B decay has a branching fraction of $\mathcal{O}(10^{-2})$. On the other hand, the semileptonic reconstruction suffers from missing kinematic information due to the neutrino in the final state of the decay. Hence, the sample is not as pure as in the hadronic case.

Figure 4.6 and Figure 4.11 show the angle $\cos\Theta_{BD\ell}$ between the true B meson and the measured $D\ell$ system of semileptonically reconstructed B_{tag} mesons for Belle and Belle II, respectively. The correctly reconstructed mesons are missing exactly one massless neutrino, and therefore peak between -1 and 1. The angle $\cos\Theta_{BD\ell}$ can be calculated from the initial beam-conditions [1, Chapter 7.2]:

$$\cos\Theta_{BD\ell} = \frac{2E_{beam}E_{D\ell} - m_B^2 - m_{D\ell}^2}{2p_B p_{D\ell}}, \tag{4.3}$$

where all quantities are measured in the center-of-mass system and E_{beam} is the beam energy, m_{B} is the nominal mass of the B meson and p_{B} is the total momentum of a B meson from a $\Upsilon(4S)$ decay.

The semileptonic tag of the `FEI` can be accessed by the user using `BASF2` via the provided `B0:semileptonic` and `B+:semileptonic ParticleLists`. The different decay-channels can be distinguished by the associated `decayModeID`.

4.2. Implementation

A proof of concepts (PoC) of the `Full Event Interpretation` was developed in [3] based on the original `Full Reconstruction (FR)` algorithm used by Belle [36]. During this thesis, the `FEI` was further developed from a PoC to a standard tool used in production. It was significantly extended by: further increasing the tagging and tag-side efficiency by getting more inclusive; reducing the runtime and memory consumption; adapting the algorithm for the usage on converted Belle data using `b2bii`; and migrating the algorithm to the `mva` package.

In the following I describe the complete `FEI` algorithm in an abstract manner (Section 4.2.1), the training of the employed multivariate classifiers (Section 4.2.2), and the performance optimizations (Section 4.2.4) which are key to a successful employment of the `FEI` in the future.

4.2.1. Algorithm

The FEI follows a hierarchical approach with six stages, visualized in Figure 4.3. All steps in the algorithm are configurable, therefore: the used decay-channels, employed cuts, and the input features and hyper-parameters of the multivariate classifiers depend on the configuration. A detailed description of the current default configuration of the `FEI` including a list of all decay-channels can be found in Section C.1.

4.2.1.1. Combination of Candidates

Charged final state particle candidates are created from `Tracks` assuming different particle hypotheses, whereas neutral final state particle candidates are created from `ECL clusters`, `KLM clusters`, or `V0 objects`. Each candidate can be correct (signal) or wrong (background). For instance, a `Track` used to create a π^+ candidate can originate from a pion traversing the detector (signal), from a kaon traversing

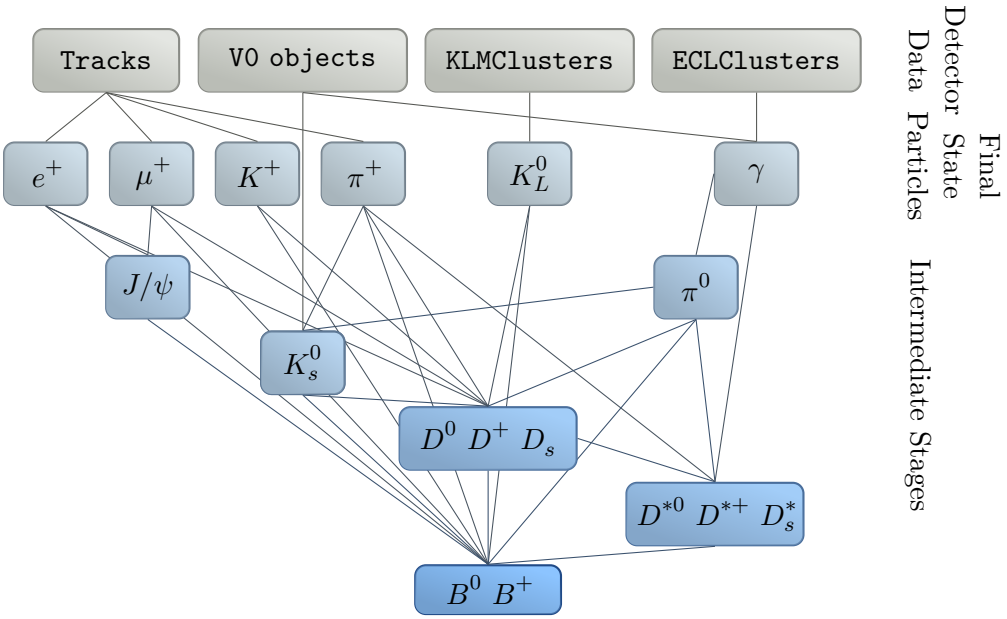


Figure 4.3.: Schematic overview of the `FEI`. The B mesons are reconstructed hierarchically; using the detector output, final state particle candidates are formed, and combined to intermediate particles until the final B candidates are formed. The probability of each candidate to be correct is estimated by an MVA algorithm. The probabilities are fed into the subsequent MVA methods.

the detector (background) or even consists of a random combination of hits from beam-background (also background).

All candidates available at the current stage are combined to intermediate particle candidates in the subsequent stages, until candidates for the desired B mesons are created. Each intermediate particle has multiple possible decay-channels, which can be used to create valid candidates. For instance, a B^- candidate can be reconstructed by combining a D^0 and a π^- candidate, or by combining a D^0, a π^- and a π^0 candidate. The used D^0 candidate could be reconstructed from a K^- and a π^+, or from a K^0_S and a π^0.

The `FEI` reconstructs more than $\mathcal{O}(100)$ explicit decay-channels, leading to more than $\mathcal{O}(10000)$ distinct decay-chains. All considered decay-channels are listed in Section C.1.

4.2.1.2. Multivariate Classification

The `FEI` employs multivariate classifiers (see Chapter 3) to estimate the probability of each candidate to be correct. Hence, each candidate created by the `FEI` (regardless at which stage) has an associated `SignalProbability` σ, which can be used to discriminate correctly reconstructed candidates from background.

In order to use all available information, a network of multivariate classifiers is built, following the hierarchical structure of the reconstruction. For each final state particle and for each decay-channel of an intermediate particle, a multivariate classifier is trained which estimates the probability that the candidate is correct.

For instance, the classifier built for the decay of $B^- \to D^0\pi^-$ would use the `SignalProbability`s of the used D^0 and π^- candidates, to estimate the `SignalProbability` of the B^- candidate created by combining the aforementioned D^0 and π^- candidates.

The input features of the classifier are among others the kinematics and vertex fitting information of the candidate and its daughters, as well as the `SignalProbability` of the daughters. The full list of input features and the chosen hyper-parameters for the multivariate classifiers can be found in Section C.1.

As can be seen in Figure 4.3 the available information flows from the data provided by the detector, through the intermediate candidates into the final B meson candidates, yielding a single number, which can be used to distinguish correctly reconstructed from incorrectly reconstructed B_{tag} mesons.

4.2.1.3. Combinatorics

It is not possible to consider all possible B meson candidates created by all possible combinations. The amount of possible combinations scales with the factorial in the number of tracks and clusters. This problem is known as **combinatorics** in HEP. Furthermore, it is not worthwhile to consider all possible B meson candidates, because all of them (except for two in the best-case scenario) are wrong.

The `FEI` uses two sets of so-called **cuts**. A cut is a criterion a candidate has to fulfill to be further considered. For instance one could demand that the invariant mass of the B meson candidate is near the nominal mass 5.28 GeV of a B meson particle, or that a μ^+ candidate has a large μ likelihood calculated from the measurements in the PID sub-detectors.

Directly after the creation of the candidate (either from a detector object, or by combining other candidates), but before the application of the multivariate method, the `FEI` uses loose and fast **pre-cuts** to remove wrongly reconstructed candidates (background), without loosing signal. The main purpose of these cuts is to save computing time and to reduce the memory consumption. These **pre-cuts** are applied separately for each decay-channel.

At first, a very loose fixed cut is applied on a quantity, which is fast to calculate, e.g. the energy for photons, the invariant mass for D mesons, the released energy in the decay for D^* mesons, or the beam-constrained mass for hadronic B mesons.

Secondly, the remaining candidates are ranked according to a quantity, which is fast to calculate (usually the same quantity as above is used here). Only the n best-candidates in each decay-channel are further considered, the others are discarded. This best-candidate selection ensures that each decay-channel and each event receives roughly the same amount of computing time.

Next, the computationally expensive parts of the reconstruction are performed on each candidate (see also Section 4.2.4): the Monte Carlo matching (in case of MC), the vertex fitting, and the multivariate classification .

After the multivariate classifiers have estimated the `SignalProbability` of each candidate, the candidates of different decay-channels can be compared to one another. Here the `FEI` uses tighter **post-cuts** to aggressively remove wrongly reconstructed candidates using all available information. The main purpose of these cuts is to restrict the number of candidates per particle to a manageable number.

At first, there is a loose fixed cut on the `SignalProbability`, to remove unreasonable candidates.

Secondly, the remaining candidates are ranked according to their `SignalProbability`. Only the m best-candidates of the particle (that is over all decay-channels) are further

considered, the others are discarded. This best-candidate selection ensures that the amount of candidates produced in the next stage is tractable by the computing system (see Section 4.2.4.1).

The exact definitions of the **pre-cuts** and **post-cuts** like the used values for n and m can be found in Section C.1.

4.2.2. Training

The multivariate classifiers used by the `FEI` are trained on Monte Carlo (MC) simulated events as explained in Chapter 3. Each multivariate classifier requires at least around $\mathcal{O}(10^3)$ signal and background candidates to be successfully trained, that is to prevent under-fitting and over-fitting. The branching fractions of the employed decay-channels for B and D mesons are of the order of $\mathcal{O}(10^{-3})$ and the reconstruction efficiency of some B decay-channels is as low as $\mathcal{O}(10^{-2})$. So, it is expected to require $\mathcal{O}(10^8)$ MC events containing B meson decays for a successful training of the `FEI`. This number was empirically confirmed.

During the training of the `FEI`, the provided MC is reconstructed following the hierarchical approach. At each stage the necessary training data for all classifiers used by the current stage is written out, and the reconstruction is suspended before the **post-cuts** would have been applied. The classifiers are fitted using the training data. Afterwards, the reconstruction is resumed, the **post-cuts** are now applied using the immediately previously fitted classifiers. The technical aspects of the training are discussed in Section 4.2.4.4.

There are three distinct types of MC events, which could be used for the training of the `FEI`: **double-generic events**: $e^+e^- \to \Upsilon(4S) \to B\overline{B}$, where both B mesons decay generically[3]; **continuum events**, that is non-resonant interactions $e^+e^- \to q\bar{q}, \ell\bar{\ell}, \gamma\gamma$; and **signal events**: $e^+e^- \to \Upsilon(4S) \to B\overline{B}$, where one B decays generically, and the other decays in an analysis-specific signal-channel like $B^+ \to \tau^+\nu_\tau$.

Depending on the training procedure and mixture of MC types in the training, the multivariate classifiers of the `FEI` are optimized for different objectives. The main objectives of the `FEI` are a **high tag-side efficiency on signal events**, that is to output correctly reconstructed B_{tag} mesons for signal events and assign them a high `SignalProbability`; and a **high purity** by rejecting background candidates, or assigning them a low `SignalProbability`. Moreover, if the signal-side can already provide a good B_{sig} candidate, it is also desired to obtain a **high tagging efficiency** by always providing at least one reasonable candidate. This can improve the final signal selection efficiency, because the event is not discarded if for instance

[3]The mesons decays into all possible final states with the correct branching fractions.

mis-identified or missing tracks prevent an entirely correct reconstruction of the tag-side.

Continuum events are not used during the training of the `FEI`. Firstly, continuum events do not contain usable signal candidates. B mesons are not present in continuum events, and other particles like D mesons have different properties (e.g. a higher momentum) than typically found in $\Upsilon(4S)$ events. Secondly, they can be suppressed efficiently by the dedicated `ContinuumSuppression` (Section 3.3.1.3) algorithm of `BASF2`.

In the following I describe the main training procedures and discuss their respective advantages and disadvantages.

4.2.2.1. Generic FEI

The generic `FEI` is trained on double-generic MC events. The training is done independently of any specific signal-side and can be performed centrally, once per Monte Carlo campaign.

The classifiers are trained to identify correctly reconstructed B_{tag} mesons on double-generic events. The training of the classifiers could be sub-optimal for signal events, due to the different multiplicity distribution of double-generic and signal events. One can mitigate this problem by training the `FEI` on double-generic MC events with a maximum number of $12 + N$ tracks, where 12 is the maximum number of tracks the `FEI` can combine to a **valid**[4] B_{tag} and N is the number of tracks required for the reconstruction of the B_{sig} (see Section 4.3.3.1).

After the user performed his analysis-specific signal-side selection and applied the completeness-constraint, most of the wrongly reconstructed B_{tag} candidates can be rejected. The generic `FEI` cannot take advantage of this fact during the training of the classifiers.

4.2.2.2. Specific FEI

The specific `FEI` is trained evenly on double-generic and signal MC events. First the analysis-specific signal-side selection is performed, afterwards the `FEI` is trained on the rest-of-event of the B_{sig} candidates.

The classifiers are specifically trained to identify correctly reconstructed B_{tag} mesons for signal events, which is the main objective of the `FEI`. Furthermore, B classifiers are only trained on background candidates from the rest-of-event of double-generic events, which fulfill the completeness-constraint. In consequence, the classifiers

[4] Empirically, 7 is the maximum number of tracks the `FEI` can combine to a **correct** candidate.

can focus on reducing the non-trivial background. Candidates from incorrectly reconstructed B_{sig} from signal events are not used at all, because the candidates produced by these events would be vastly over-represented in the training data, since the signal events are usually very rare on data.

Since only B candidates which fulfill the completeness-constraint are used, the procedure suffers from notoriously low statistics in signal and background. In consequence, large amounts of double-generic and signal MC are required to properly train the `FEI` using this procedure.

The training depends on the analysis-specific signal-side selection and has to be performed by the user for his analysis.

4.2.2.3. Mono FEI

The mono `FEI` is trained on signal MC events for the decay $B^0 \to \nu\bar{\nu}$. There is exactly one detectable B in each event. In effect, the `FEI` is trained on the rest-of-event of an arbitrary correctly reconstructed B_{sig}.

The classifiers are specifically trained to identify the most likely tag-side decay-chain, under the assumption of an correctly reconstructed signal-side. Such a training could be used instead of the traditional inclusive tagging algorithm. It can provide improved tag-side vertexing by exploiting the explicitly known tag-side decay-chain, e.g. using the decay-tree fitter (see [70]). This type of training was not further investigated during this thesis.

The training can be done independently of any signal-side and can be performed centrally, once per Monte Carlo campaign.

4.2.3. Application

The `FEI` can either be applied to the **entire event**, or only to a subset like the **rest-of-event** of a B_{sig} candidate.

After the application the `FEI` provides tag-side B candidates for B^+ and B^0 in hadronic and semileptonic decay-channels. Each candidate has an associated `SignalProbability` that can be used to reject wrongly reconstructed candidates.

An analysis with an inclusive signal-side such as $B^+ \to X_u \ell^+ \nu$ applies the generic `FEI` to the entire event without using the completeness-constraint. The rest-of-event of the B_{tag} candidate can be used to reconstruct the signal-side inclusively by assigning all remaining tracks and clusters to the signal-side.

An analysis with an exclusive signal-side, e.g. $B^+ \to \tau^+\nu_\tau$, can either apply the generic `FEI` to the entire event or the specific `FEI` to the rest-of-event of the B_{sig}. In the end, the completeness-constraint is always applied. The choice of the generic or specific `FEI` depends on the feasibility of a dedicated training of the specific `FEI` using large amounts of signal and double-generic MC events. Furthermore, the systematic uncertainties are different (see Section 4.3.4).

Finally, it is also possible to apply the generic `FEI` to the rest-of-event, although it was trained on the entire event. This allows a rough estimation of improvements which can be expected of the specific `FEI`.

4.2.4. Performance

Fitting the `FEI` on $\mathcal{O}(100)$ million events and applying it to $\mathcal{O}(1)$ billion events, is a CPU-intensive task. An optimized runtime and a small memory-footprint is key for the practicalness in production and saves computing resources i.e. money. As can be seen from Table 4.1 most of the time in the `FEI` application is spent in: vertex fitting, particle combination and classifier inference. All three tasks have been carefully optimized during this thesis. These optimizations are discussed in the following sections.

Furthermore, the time-consuming training procedure has been implemented using a distributable map-reduce approach (Section 4.2.4.4).

All runtime measurements in this section were performed on the KEKCC cluster. The measured values depend on the hardware (e.g., CPU, RAM, and caches), the load of the system (e.g., the amount of rivaling processes and the current IO traffic), the software (the exact `BASF2` version) and the data (e.g., the Belle experiment and event type). Therefore, only the relative differences between the stated numbers are meaningful.

4.2.4.1. Combination of Candidates

As explained in Section 4.2.1.3, the number of candidates which have to be processed is growing like the factorial of the multiplicity of the channel. In previous approaches the runtime and the maximum memory consumption was dominated by a few high multiplicity events and tight cuts had to be applied to high multiplicity channels.

In contrast, the `FEI` limits the combinatorics problem by performing best-candidate selections during the reconstruction of the decay-chain instead of fixed cuts.

In consequence, each event and each decay-channel is allowed to process the same number of candidates in vertex fitting and classifier inference i.e. consuming similar amounts of CPU time. Moreover, the maximum memory consumption is limited due to the fixed number of best-candidates per event, which is a key requirement by the computing infrastructure.

Table 4.1.: Relative time in percent spent by the `FEI` applied to $\Upsilon(4S)$ Belle II MC for various tasks. The fitting of the `FEI` is dominated by reading (writing) the cached `DataStore` from (to) disk space in between the stages, which is limited by the IO throughput of the system and cannot be optimized in the `FEI`. The application is still dominated by the vertex fitting, which originally required about 78% of the CPU time.

Task	Training	Application
read/write `DataStore`	30	0
vertex fitting	26	38
particle combination	19	27
classifier inference	11	15
training data & monitoring	6	0
best candidate selection	3	6
other	5	14

Table 4.2.: Application time of the `FEI` in ms for various types of MC7 on the KEKCC cluster. The uncertainty of the absolute values is about ± 3 ms.

MC Type	Time per event in ms
$\Upsilon(4S) \to B^+B^-$	128
$\Upsilon(4S) \to B^0\overline{B}^0$	119
$e^+e^- \to c\bar{c}$	76
$e^+e^- \to s\bar{s}$	48
$e^+e^- \to d\bar{d}$	45
$e^+e^- \to u\bar{u}$	43
$e^+e^- \to \tau^-\tau^+$	14
$e^+e^- \to \mu^-\mu^+$	12
$B^+ \to \tau^+\nu_\tau$	43

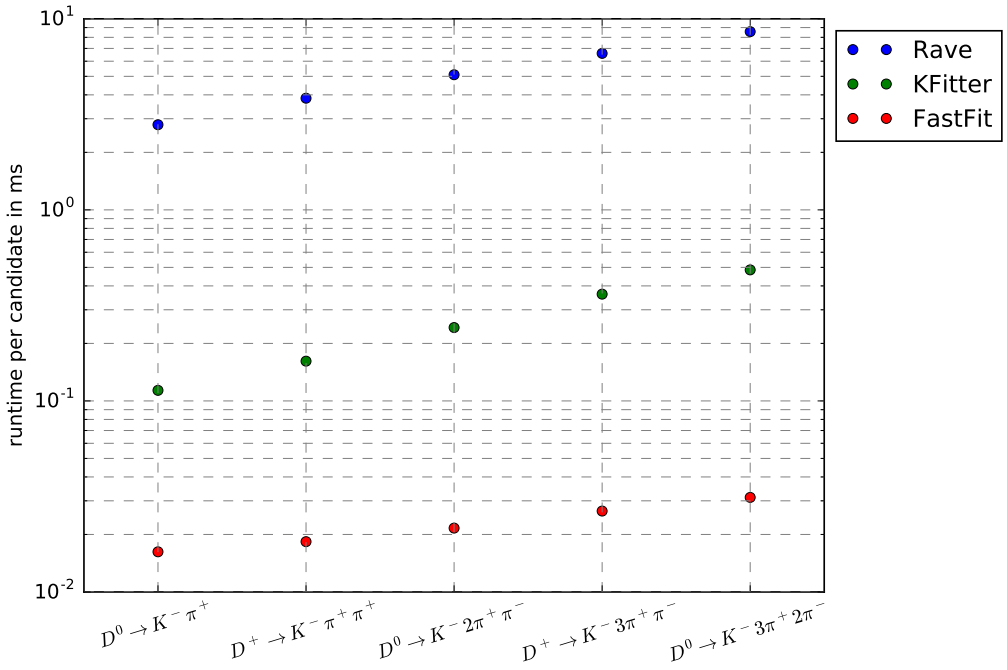


Figure 4.4.: Runtime of an unconstrained vertex fit for different D meson decay-channels in the `BASF2` framework. `FastFit` is one order of magnitude faster than `KFitter` and two orders of magnitude faster than `RAVE`.

4.2.4.2. Vertex fitting

The default vertex fitting implementation of Belle II was ported from the legacy Belle code and is named `KFitter` (Kinematic Fitter). It is based on a Kalman Filter and supports unconstrained, IP profile constrained, and mass-constrained fits. In addition, the `RAVE` (Reconstruction in an Abstract, Versatile Environment) toolkit [71] was integrated in `BASF2`. It has its roots in the CMS (Compact Muon Solenoid) vertex reconstruction software and supports a broad collection of vertex finding and fitting algorithms based on a Kalman Filter approach generalized to include adaptive track assignment and multiple vertices.

On the other hand, the `FEI` uses only a simple unconstrained vertex fit during the reconstruction, and feeds the calculated information into its multivariate classifiers. The user can refit the whole decay-chain of the final B candidates again, including mass and/or ip profile constraints if desired. Therefore, the `FEI` restricts itself to a

fast unconstrained vertex fitting. Still, the major part of the CPU time of the `FEI` is spent during vertex fitting.

During this thesis a dedicated fitter (called `FastFit`) based on a Kalman Filter [72] was implemented for the `FEI`, which outperforms the default `KFitter` implementation by one order of magnitude as can be seen in Figure 4.4. A speed-up of the vertex-fitting part of $s_{\text{FastFit}} \approx 5$ of the `FEI` was measured on generic B^+B^- Belle II events. With `KFitter` the application phase of the `FEI` spent $p \approx 78\%$ of the CPU time during the vertex fitting. According to Amdahl's law [73], a theoretical speed-up of the entire `FEI` of approximately

$$S_{\text{FEI}} = \frac{1}{(1-p) + \frac{p}{s_{\text{FastFit}}}} = 2.66,$$

is expected. Empirically, an overall speed-up of the `FEI` of 2.74 was observed.

The quality (deviation to the true vertex position) of the vertex fit of all three tested implementations is very similar, which is expected since all implementations calculate the same well-defined least-square solution. Deviations occur due to different stopping criterion of the algorithms. Consistently, the overall maximum tag-side efficiency of the FEI does not depend on the employed vertex fitter during fitting or application.

The `FastFit` code is licensed under GPLv3 and available on github [74].

4.2.4.3. Multivariate Classification

As described in Section 3.2.1.1, `FastBDT` is the default multivariate classification algorithm of `BASF2`. It was originally designed for the `FEI` to speed up the fitting and inference-phase. Compared to other popular BDT implementations such as those provided by `TMVA`, `SKLearn` and `XGBoost` it originally gained more than one order of magnitude in execution time, both in fitting and inference. Using `FastBDT`, most of the time is spent during the extraction of the necessary features used for the inference, therefore no further significant speedups can be achieved by employing a different method.

4.2.4.4. Distributed Training

The training of the `FEI` is fully automatized and distributed. The distributed training was already introduced in [3]. It follows a map-reduce approach.

The supplied Monte Carlo files are partitioned and processed by independent computing nodes. At first the total amount of supplied Monte Carlo events and their particle content (e.g. the total number of B^+ and B^0 mesons) is calculated. Subsequently, the six stages of the `FEI` hierarchy are reconstructed.

At each stage the necessary training data for all classifiers used by the current stage is written out. The reconstruction is suspended and the current content of the `DataStore` of all events on all nodes is cached on disk.

The training data is merged centrally and the associated classifiers are fitted. The generated `WeightFiles` are uploaded into the local `Belle II Conditions Database`.

Afterwards, reconstruction is resumed by reading in the cached `DataStore` files from disk. The classifiers are loaded from the corresponding `WeightFiles`, which are downloaded from the `Belle II Conditions Database`.

4.2.4.5. Comparison with Full Reconstruction

The FEI was compared to the hadronic `Full Reconstruction` (`FR`) algorithm used by Belle.

Out-of-the-box the `FR` requires on average 36 ms (39 ms) to reconstruct a simulated $B^0\overline{B}^0$ (B^+B^-) Belle event. The `FR` was implemented using dedicated `C++` code. In contrast, the `FEI` requires with the identical configuration[5] on average 24 ms (26 ms) to reconstruct the same Belle events using the `b2bii` interface. The `FEI` is implemented in `Python` and only uses general purpose modules implemented in `C++`.

The default configuration of the `FEI` requires 98 ms (100 ms) to reconstruct a simulated $B^0\overline{B}^0$ (B^+B^-) Belle event. This time includes the reconstruction of hadronic and semileptonic B_{tag} candidates using many additional channels (see Section C.1) and yielding a higher tag-side efficiency.

The average runtime 119 ms (128 ms) to reconstruct a simulated $B^0\overline{B}^0$ (B^+B^-) Belle II event is significantly larger because of increased event size in terms of number of reconstructed tracks and number of reconstructed clusters.

4.2.5. Automatic Reporting

The `FEI` includes an automatic reporting system called `Full Event Interpretation Report (FEIR)`. The `FEIR` contains efficiencies and purities for all particles and decay-channels at different points during the reconstruction. It is generated from so-called monitoring histograms, which can optionally be written out during the training and the application. In addition, reports for each multivariate classifier can be created by the `mva` package (see Section 3.2.4).

This built-in monitoring capability upgrades the `FEI` from a black-box to a white-box algorithm, which the user can understand and inspect on all levels.

An excerpt of the `FEIR` can be found in Section C.2.

[5] Only decay-channels were reconstructed, which were used by the `FR` as well.

4.3. Validation

The FEI was extensively validated on converted Belle MC from the last Belle Monte Carlo campaign, converted Belle data recorded by the Belle detector, and Belle II MC from the 7$^{\text{th}}$ Belle II Monte Carlo campaign. Unless stated otherwise all studies were conducted using $\Upsilon(4S)$ and continuum events (including $c\bar{c}$, $s\bar{s}$, $d\bar{d}$, $u\bar{u}$ and $\tau^-\tau^+$) scaled to their expected fractions on recorded data.

The validation includes a detailed discussion on the signal-definition used for the B_{tag} candidates (Section 4.3.1); the performance of the FEI on converted Belle events (Section 4.3.2) and Belle II events (Section 4.3.3); the calibration of the FEI on converted Belle data using control channels (Section 4.3.4); and a comparison to the Full Reconstruction (FR) algorithm used by Belle (Section 4.2.4.5).

Throughout the text, only the results for charged B mesons are shown, the results for neutral B mesons are very similar and can be found in Section C.3.

4.3.1. Signal Definitions

This thesis used the Monte Carlo matching algorithm of BASF2. The algorithm searches for the common mother of all tracks and clusters used during the reconstruction. It can distinguish a large number of possible errors, which allows for a fine-grained signal-definition:

- particles which were missed during the reconstruction: final state radiation (fsr), photons (excluding fsr), intermediate resonances; neutrinos, K_L^0 or other massive particles;
- charged final state particles, which were misidentified or decayed in flight;
- particles which were wrongly added during the reconstruction.

The **default signal-definition** of Belle II allows for missing final state radiation, missing intermediate resonances and missing neutrinos. If not stated otherwise, this thesis uses this definition, to be comparable to previous exclusive tagging algorithms.

However, since the B mesons reconstructed by the FEI are used for exclusive tagging, an **extended signal-definition** can be useful. Here, the mis-identification of a charged final state particle and the addition of a beam-induced background particle is allowed as well. The completeness-constraint is uncompromised by this extended signal-definition. Also important quantities like the beam-constrained mass and deviation from the nominal beam-energy are often unchanged.

As can be seen from Figure 4.5 (Belle) and Figure 4.10 (Belle II), the hadronically tagged extended-signal candidates peak at the same position in the beam-constrained

mass as the default-signal candidates. The same holds true in $\cos\Theta_{\mathrm{BD}\ell}$ for the semileptonically tagged B mesons (see Figure 4.6 (Belle) and Figure 4.11 (Belle II)).

The multivariate classifiers of the `FEI` use the default signal-definition, with an additional constraint for final state particles, which have to come from the primary interaction to be considered signal (for instance electrons from converted gammas are not considered signal in the electron classifier). Therefore the candidates fulfilling only the extended signal-definition are suppressed if a tight cut on the `SignalProbability` is chosen (see Figure 4.7a and Figure 4.12a).

Even the extended signal-definition does not include all peaking background contributions from mis-reconstructed B mesons, as can be seen from Figure 4.8, where a Gaussian distribution for the peaking B candidates and an ARGUS function for combinatorial background [1, Chapter 7.1] was fitted to converted Belle data. This additional **peaking background** is caused by nearly correctly reconstructed B mesons. There is no peaking background component visible on recorded off-resonance data, which does not contain B mesons (see Figure 4.8). Consequently, the peaking background is caused by physics and not by a non-uniform selection efficiency of the classifiers used by the `FEI` (see Section 3.3.2). It is debatable if those peaking B candidates can be (for some analyses) considered signal as well.

An analysis, for instance the benchmark analysis $\mathrm{B}^+ \to \tau^+\nu_\tau$ (see Chapter 5), usually accepts all events which contain the searched signal-decay as signal. Hence, the signal-definition of the tag-side is not used. Still, many analyses could profit from the relaxed extended signal-definition if the classifiers used by the `FEI` are explicitly trained with this definition.

4.3.1.1. Double Counting

It is possible that the same candidate is reconstructed in more than one decay-channel by the `FEI`. For instance, the decay-chain $\mathrm{B}^+ \to \overline{\mathrm{D}}^{0*}[\to \overline{\mathrm{D}}^0\pi^0]\pi^+$ can be reconstructed via the decay-channels $\mathrm{B}^+ \to \overline{\mathrm{D}}^{0*}\pi^+$ and $\mathrm{B}^+ \to \overline{\mathrm{D}}^0\pi^0\pi^+$, which are both used by the `FEI`. Both candidates will be correct, because the missing intermediate $\overline{\mathrm{D}}^{0*}$ resonance is allowed by the default signal definition. In general, their associated `SignalProbability` will be different.

To prevent double counting[6] in such corner cases, the `FEI` provides an additional `extraInfo` associated to each candidate called `uniqueSignal`, which will flag only one of the candidates as signal. In the remainder of this work, the default signal definition uses this information, while the extended signal definition regards both candidates as true, since both peak in the signal region.

[6] Two candidate constructed from the same final state particles are counted as signal separately, which can lead theoretically to efficiencies larger than one.

Table 4.3.: **Belle:** Maximum tag-side efficiency of the `FEI` on converted Monte Carlo simulated Belle events.

	Charged $B^{\pm}$	Neutral B^0
Hadronic	0.76 %	0.46 %
Semileptonic	1.80 %	2.04 %

Table 4.4.: **Belle:** Tagging efficiency of the `FEI` on converted data recorded by the Belle detector.

	All		signal region	
	Charged $B^{\pm}$	Neutral B^0	Charged $B^{\pm}$	Neutral B^0
Hadronic	48.0 %	37.9 %	5.6 %	4.6 %
Semileptonic	90.0 %	87.2 %	25.2 %	22.2 %

This effect is only of minor importance, as can be seen from the FEIR reports in Section C.2, where the tag-side efficiencies with and without the `uniqueSignal` requirement are stated.

In the benchmark analysis (see Chapter 5), a best-candidate selection is performed on the tag-side, therefore double-counting is completely prevented.

4.3.2. Performance on Belle

The performance of the `FEI` was studied on converted Belle MC from the last Belle Monte Carlo campaign and converted Belle data recorded by the Belle detector. The sample used for the training of the `FEI` contained 200 million simulated $\Upsilon(4S)$ events divided evenly between the processes $\Upsilon(4S) \to B^+B^-$ and $\Upsilon(4S) \to B^0\overline{B}^0$. The sample used to estimate the performance of the `FEI` during the application phase contained 6 million simulated events divided evenly between the $\Upsilon(4S)$ processes $\Upsilon(4S) \to B^+B^-$, $\Upsilon(4S) \to B^0\overline{B}^0$, and 6 million simulated events divided evenly between the continuum processes $e^+e^- \to c\bar{c}, s\bar{s}, d\bar{d}$ and $u\bar{u}$. In addition 10 million events recorded at the $\Upsilon(4S)$ resonance, and 10 million off-resonance events recorded 60 MeV below the $\Upsilon(4S)$ resonance were used. Throughout the text, the simulated events are scaled to the integrated luminosity of the recorded data they are compared to.

The generic `FEI` trained on all Belle events with less or equal to 14 tracks per event was applied. Table 4.3 shows the obtained tag-side efficiency and Table 4.4 shows the obtained tagging efficiency.

At this point an analysis with an exclusive signal-side would employ the completeness-constraint. Instead, only the candidate with the highest `SignalProbability` in each event is considered in the following. This best-candidate selection is independent of a specific signal-side, but not optimal (meaning the stated results can be understood as lower bounds for the tag-side efficiency and purity of the `FEI`).

Figure 4.5 shows the beam-constrained mass of the best hadronically tagged B^+ meson candidate per event. The correctly reconstructed B mesons peak clearly at the nominal B meson mass, as expected. Figure 4.6 shows $\cos\Theta_{BD\ell}$ of the best semileptonically tagged B^+ meson candidate per event. The correctly reconstructed B mesons peak clearly in the physically allowed region between -1 and 1, as expected. The dashed lines indicates the signal region which is used to calculate the tag-side efficiency and purity receiver operating characteristic curves in Figure 4.7. The hadronic tag can reach a higher purity than the semileptonic tag, whereas the semileptonic tag provides a larger tag-side efficiency, as expected.

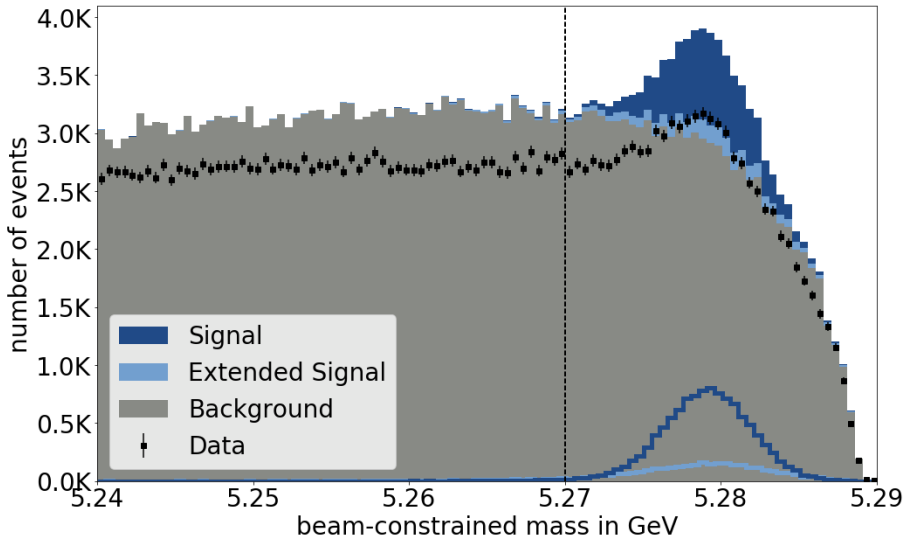
For the hadronic tag, the results were verified on data.

The observed differences in the background distribution between the Monte Carlo expectation and the recorded data can be explained by the poor simulation of the continuum component (see Section 2.3.3.3). The over-estimation of hadronic B candidates from continuum in the Monte Carlo simulation was cross-checked using off-resonance data (see the lower plot in Figure 4.8). The shape difference of semileptonic B candidates is also caused by the continuum description as can be seen from Figure 4.9.

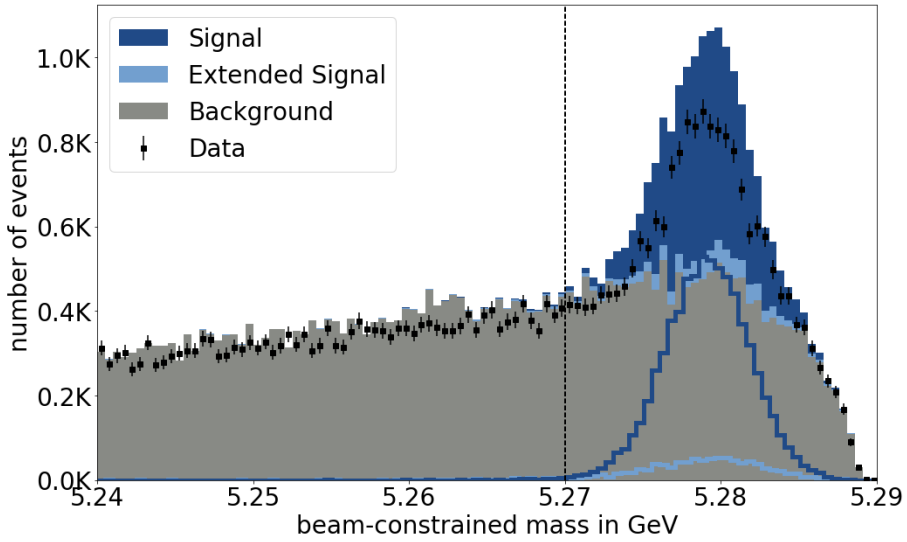
4.3.2.1. Hadronic tag-side efficiency estimation on data

The measurement of absolute branching fractions (in contrast to, e.g, branching fraction ratios) requires the knowledge of the selection efficiency on data. Therefore it is crucial to calibrate the tag-side efficiency of the `FEI` on data using control-channels (see Section 4.3.4). This section introduces an alternative method to estimate the tag-side efficiency, which does not require the reconstruction of a signal-side.

The hadronic tag-side efficiency can be estimated on recorded data by fitting the beam-constrained mass spectrum of the B meson candidates. The distribution of the signal candidates are modeled with a Gaussian distribution and the background candidates are modeled with an `ARGUS` function [1, Chapter 7.1]. The location and scale of the distributions were fixed to the Monte Carlo expectation. Hence, only the normalization and the shape parameter c of the `ARGUS` function are adjustable by the fit. The two components are fitted using an extended unbinned maximum likelihood (EUML) fit. Afterwards the tag-side efficiency and purity in a window of $5.24\,\text{GeV} < M_{bc} < 5.29\,\text{GeV}$ can be calculated using the fitted yields of the signal and background component.

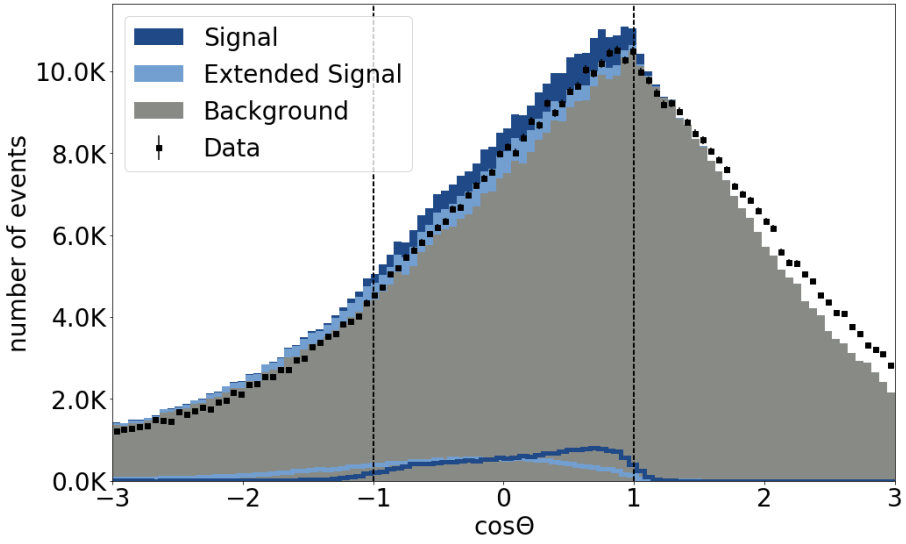


(a) A loose cut on the `SignalProbability` was performed.

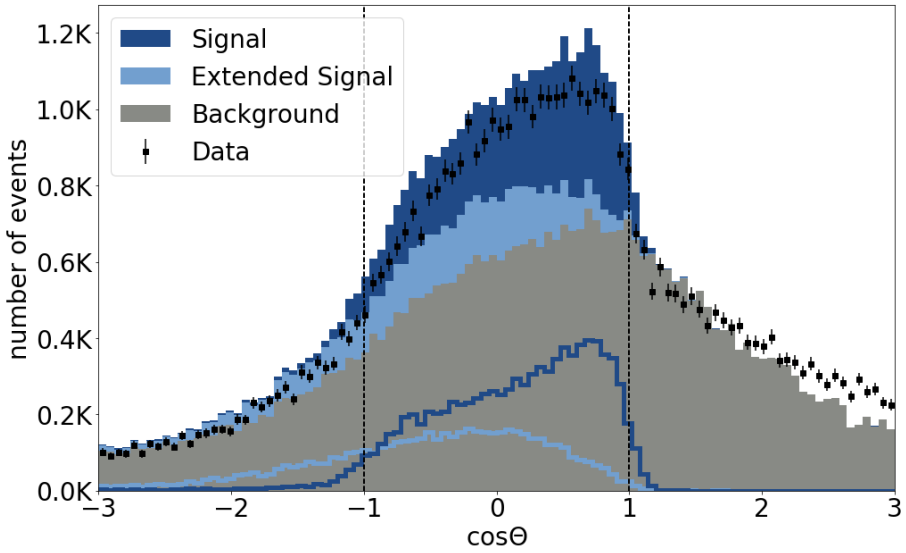


(b) A tight cut on the `SignalProbability` was performed.

Figure 4.5.: **Belle:** The beam-constrained mass of the best hadronically tagged B^+ meson candidate per event. The spectrum of the different signal-definitions is shown.

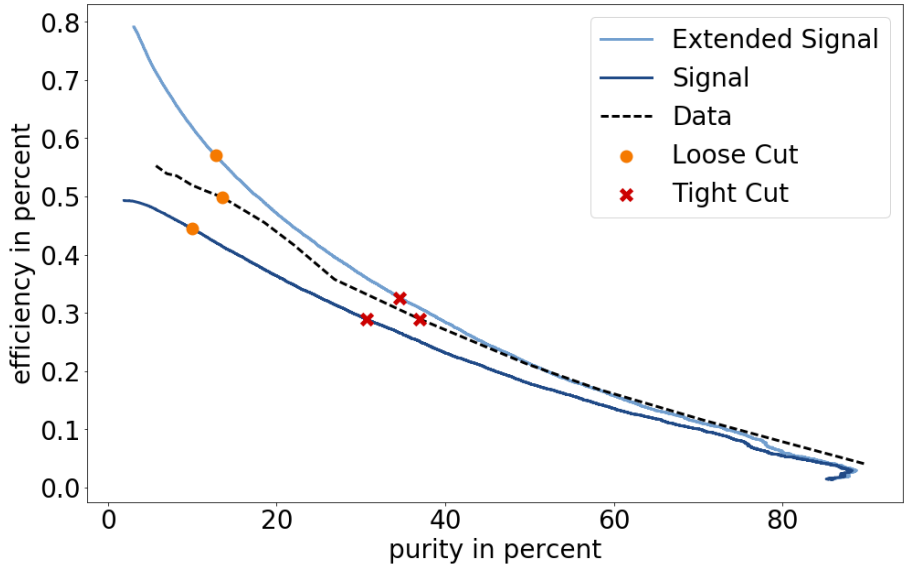


(a) A loose cut on the `SignalProbability` was performed.

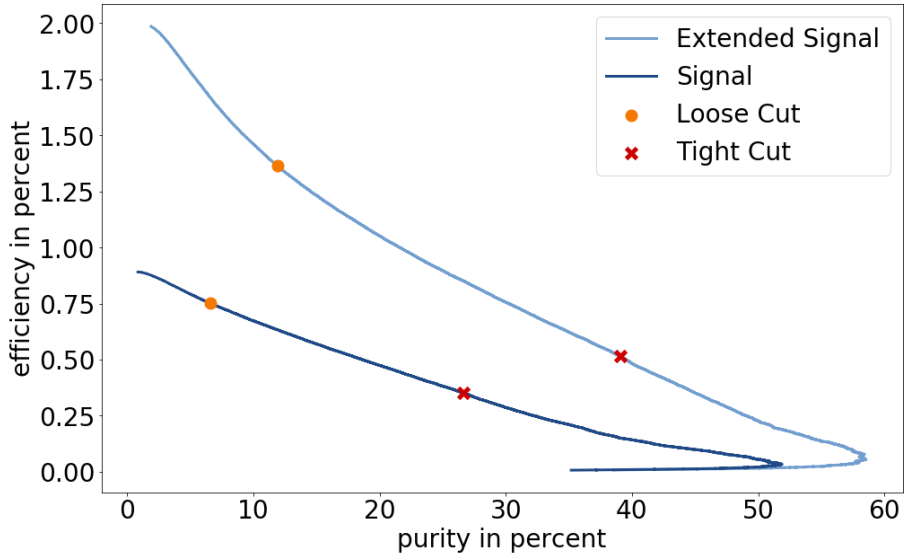


(b) A tight cut on the `SignalProbability` was performed.

Figure 4.6.: **Belle:** $\cos\Theta_{\mathrm{BD}\ell}$ of the best semileptonically tagged B^+ meson candidate per event. The spectrum of the different signal-definitions is shown.



(a) Hadronic Tag



(b) Semileptonic Tag

Figure 4.7.: **Belle:** Receiver operating characteristic (ROC) curve of tagged B^+ for the signal and extended signal-definition. The loose (`SignalProbability` > 0.01) and tight (`SignalProbability` > 0.1) cut are shown in orange and red, respectively.

This method overestimates the tag-side efficiency if the beam-constrained mass spectrum contains a background component that peaks at the same position as the signal component as can be seen in Figure 4.8. Although the signal peak is clearly smaller on recorded data than on simulated events, the extracted signal fraction is the same, because the peaking background was assigned to the signal peak by the fit (see also the discussion of the signal-definition in Section 4.3.1).

The resulting efficiencies and purities for different cuts on the `SignalProbability` of the candidates is shown in Figure 4.7a. It is compatible with the expectation from Monte Carlo expectation, but suffers from the over-estimation problem discussed above.

A complementary and superior method using control-channels is discussed in Section 4.3.4. This method is robust against peaking background, because it does not rely on any signal-definition of the tag-side candidates.

The semileptonic tag-side efficiency can be estimated on recorded data by fitting the $\cos\Theta_{\mathrm{BD}\ell}$ spectrum of the B meson candidates. This was not further investigated during this thesis, but is described in more detail in [75].

4.3.2.2. Continuum Background

The performance of the `FEI` was studied on continuum data recorded by the Belle detector 60 MeV below the $\Upsilon(4S)$ resonance.

At this energy, the production of B mesons is kinematically forbidden, hence the name "continuum data". Some kinematic properties like the beam-constrained mass are very sensitive to the reduced energy, therefore these properties are scaled to the correct energy scale. For instance, the beam-constrained mass is scaled by

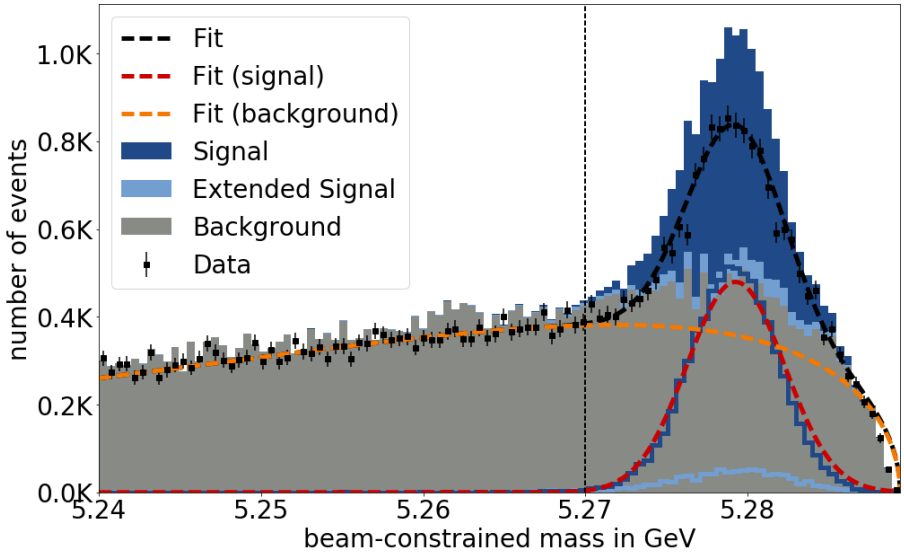
$$S_{\mathrm{energy}} = \frac{10.58\,\mathrm{GeV}}{10.58\,\mathrm{GeV} - 0.06\,\mathrm{GeV}}.$$

to account for the shift of the kinematic end-point. In the case of $\cos\Theta_{\mathrm{BD}\ell}$, the on-resonance beam-energy is used for the calculation (see Equation 4.3).

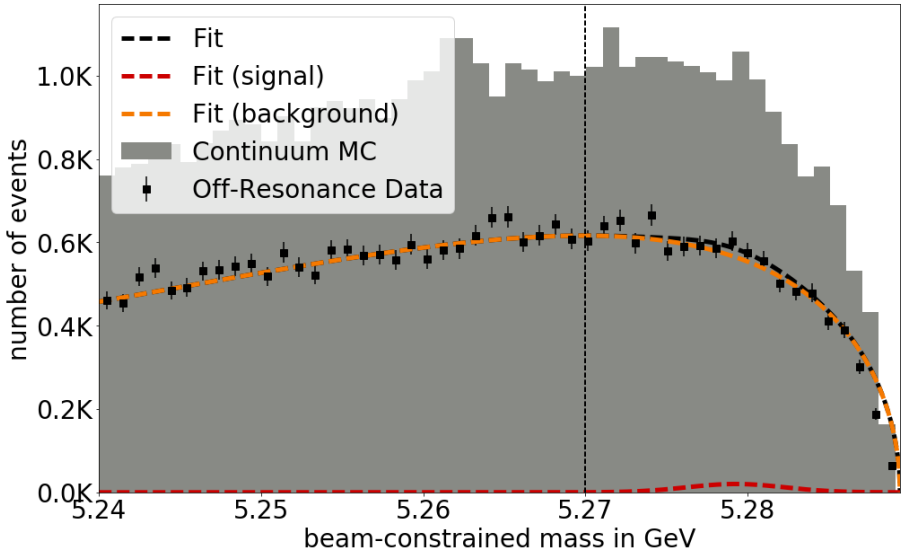
As can be seen from Figure 4.8 the continuum data does not contain any signal or peaking background, but the combinatorial background can be described well with an ARGUS function. The Monte Carlo simulation overestimates the amount of combinatorial background from continuum. The same holds true for the semileptonic B mesons (see Figure 4.9), in addition the distribution is shifted. This is a known problem (already encountered in Section 3.3.3 and described in Section 2.3.3.3). The continuum background on data is about 20% lower than expected from Monte Carlo.

The tagging efficiency on continuum data (see Table 4.5) differs, as expected because the $\mathrm{B\overline{B}}$ component is missing.

In total, the `FEI` performance does not seem to be influenced by the lower beam-energy of off-resonance data, which is a desired property because this allows the use of off-resonance data as a signal-free control region for continuum background.



(a) On-resonance: The signal peak on data is smaller. Hence the tag-side efficiency is lower on data compared to the Monte Carlo expectations. On the other hand, the fitted signal fraction is over-estimated due to the peaking background. Both effects cancel each other.



(b) Off-resonance: As expected the fitted signal fraction is nearly zero. Hence nearly no peaking background is observed on continuum data. The off-resonance data was scaled to account for the shift in the kinematic end-point of the beam-constrained mass distribution.

Figure 4.8.: **Belle:** The beam-constrained mass of the best hadronically tagged B^+ meson candidate per event. The Monte Carlo expectation is shown as filled histograms. The data is shown as black points with a Poisson uncertainty. The result of the EUML on 10 million recorded Belle events is shown as dashed lines.

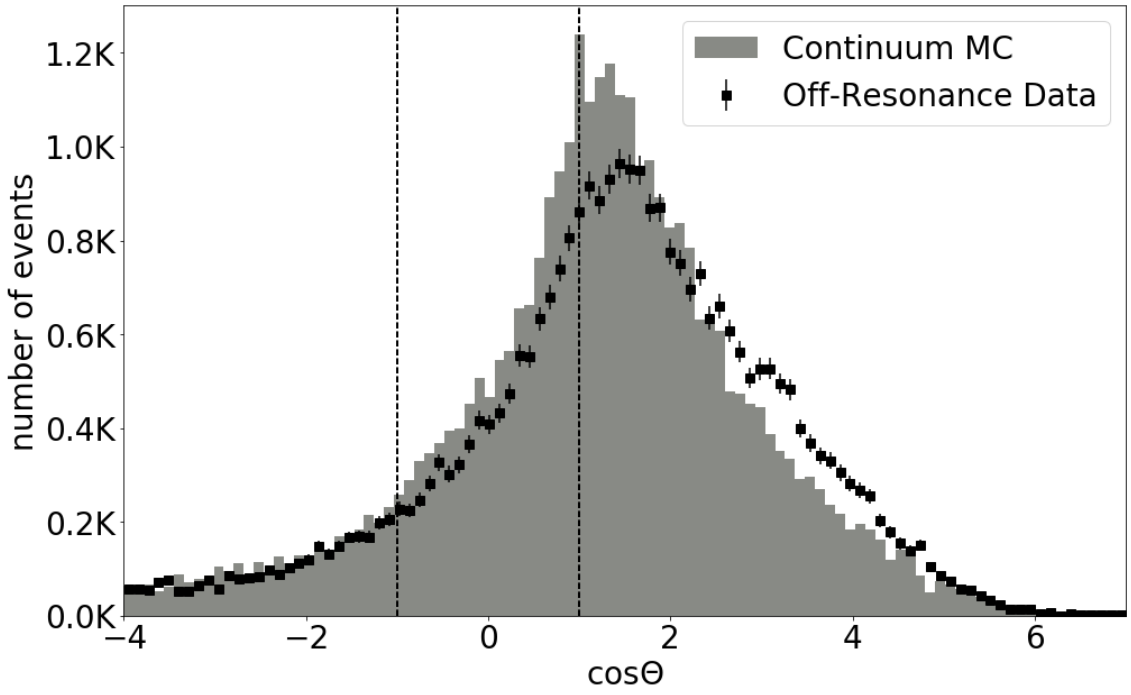


Figure 4.9.: **Belle:** $\cos\Theta_{\mathrm{BD}\ell}$ of the best semileptonically tagged B^+ meson candidate per event. The Monte Carlo expectation of the continuum component is shown as a filled histogram. The off-resonance data is shown as black points with a Poisson uncertainty. The off-resonance data-points were calculated using the on-resonance beam-energy.

Table 4.5.: **Belle:** Tagging efficiency of the `FEI` on converted continuum data recorded by the Belle detector.

	All		signal region	
	Charged $\mathrm{B}^{\pm}$	Neutral B^0	Charged $\mathrm{B}^{\pm}$	Neutral B^0
Hadronic	40.0 %	30.6 %	6.9 %	5.6 %
Semileptonic	86.7 %	83.6 %	24.8 %	21.5 %

Table 4.6.: **Belle II:** Maximum tag-side efficiency of the `FEI` on independent Monte Carlo simulated Belle II events.

	Charged $B^{\pm}$	Neutral B^0
Hadronic	0.66 %	0.38 %
Semileptonic	1.45 %	1.94 %

Table 4.7.: **Belle II:** Tagging efficiency of the `FEI` on independent Monte Carlo simulated Belle II events.

	All		signal region	
	Charged $B^{\pm}$	Neutral B^0	Charged $B^{\pm}$	Neutral B^0
Hadronic	35.5 %	28.0 %	4.7 %	3.7 %
Semileptonic	71.2 %	69.6 %	18.8 %	12.4 %

4.3.3. Performance on Belle II

The performance of the `FEI` was studied on Belle II MC from 7$^{\text{th}}$ Belle II Monte Carlo campaign. The sample used for the training of the `FEI` contained 180 million simulated $\Upsilon(4S)$ events divided evenly between the processes $\Upsilon(4S) \to B^+B^-$ and $\Upsilon(4S) \to B^0\overline{B}^0$. The sample used to estimate the performance of the `FEI` during the application phase contained 7 million simulated events divided evenly between the processes $\Upsilon(4S) \to B^+B^-$, $\Upsilon(4S) \to B^0\overline{B}^0$, and $e^+e^- \to c\overline{c}, s\overline{s}, d\overline{d}, u\overline{u}$ and $\tau^+\tau^-$, and were scaled to an integrated luminosity of $1\,\text{ab}^{-1}$.

The generic `FEI` was trained on all Belle II events with less than or equal to 14 tracks per event. Table 4.6 shows the obtained tag-side efficiency. Table 4.7 shows the obtained tagging efficiency.

At this point an analysis with an exclusive signal-side would employ the completeness-constraint. Instead, only the candidate with the highest `SignalProbability` in each event is considered in the following. This cut is independent of a specific signal-side, but not optimal (meaning the stated results can be understood as lower bounds for the tag-side efficiency and purity of the `FEI`).

Figure 4.10 shows the beam-constrained mass of the best hadronically tagged B^+ meson candidate per event. The correctly reconstructed B mesons peak clearly at the nominal B meson mass, as expected. Figure 4.11 shows $\cos\Theta_{BD\ell}$ of the best semileptonically tagged B^+ meson candidate per event. The correctly reconstructed B mesons peak clearly in the physically allowed region between -1 and 1, as expected. The dashed lines indicates the signal region which is used to calculate the tag-side efficiency and purity receiver operating characteristic curves in Figure 4.12. The

hadronic tag can reach a higher purity than the semileptonic tag, whereas the semileptonic tag provides a larger tag-side efficiency, as expected.

4.3.3.1. Influence of the event multiplicity

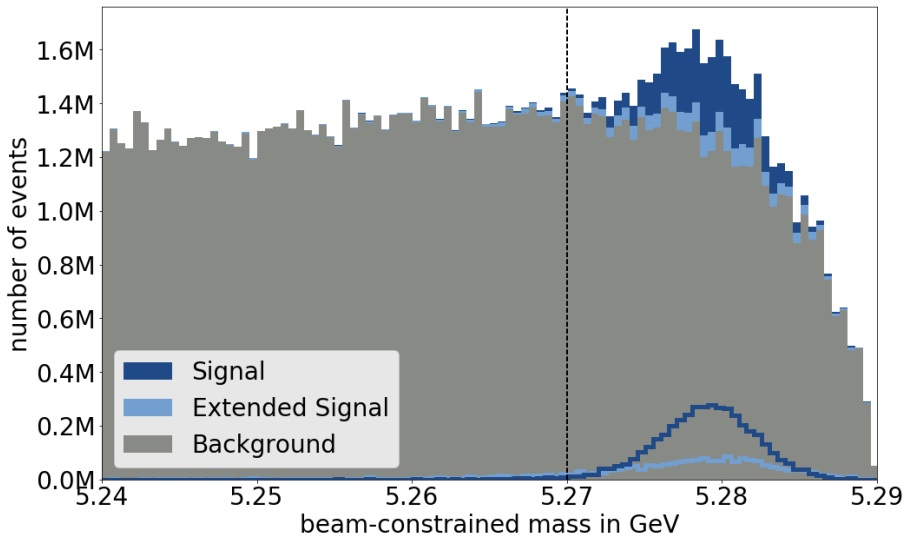
The number of reconstructed tracks (the **multiplicity**) is a fundamental property of an event. The larger the number of reconstructed tracks, the larger the combinatorial background and the more difficult it is to reconstruct the correct decay-chain. The influence of the event and candidate multiplicity on the performance of the `FEI` was investigated during this thesis.

Figure 4.13 shows the multiplicity of B_{tag} candidates. The majority of correctly reconstructed tag candidates consist of less than 8 tracks. Theoretically, the maximum number of tracks which can be combined to a valid charged (neutral) B_{tag} candidate by the `FEI` is 11 (12). However, in practice 9 (10) tracks per candidate is not exceeded.

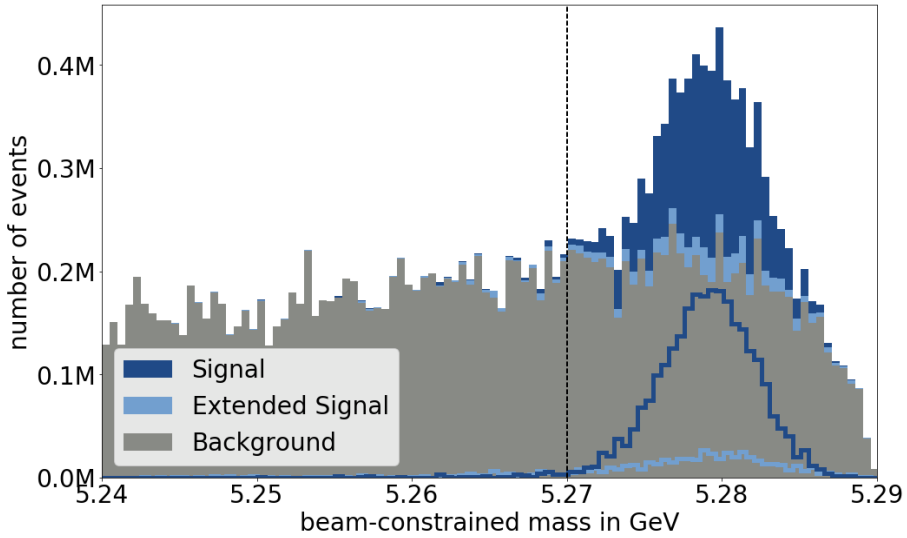
As can be seen from Figure 4.14 this effectively restricts the multiplicity of correctly tagged signal events. In this thesis the rare decay $B^+ \to \tau^+\nu_\tau$ is used as signal (see Chapter 5). Often the investigated rare B_{sig} decay-channel has a very low multiplicity N. Therefore it is reasonably to only consider events with less than $11 + N$ $(12 + N)$ reconstructed tracks in the training and/or in the application, because only those events will pass the completeness-constraint. Furthermore, it is expected that the `FEI` performs better on low-multiplicity events due to the reduced combinatorics, and that the required computing resources scales with the event multiplicity.

The influence of the maximum number of reconstructed tracks allowed in the training and in the application was studied. The study was conducted with version 3 of the `FEI`, whereas the rest of this thesis uses version 4. The results are shown in Figure 4.15 and should be independent of the used version. Only the best B_{tag} candidate per event was considered, and an additional cut on the beam-constrained mass was applied.

Surprisingly, the maximum number of reconstructed tracks allowed in the training barely influences the performance of the `FEI` if applied to all events (see Figure 4.15a). Hence, it is not important if the training is done on all events, or on low-multiplicity events. Only at a multiplicity as low as 6 a significant effect is visible. This can be explained by the hierarchical approach of the `FEI`. It allows the training of high-multiplicity B decay-channels using only low-multiplicity D channels. But during the application the high-multiplicity B decay-channels can also be combined with high-multiplicity D channels. In fact, a `FEI` trained on events with less than or equal to six tracks is still able to reconstruct tag candidates with a multiplicity of seven during the application quite well.

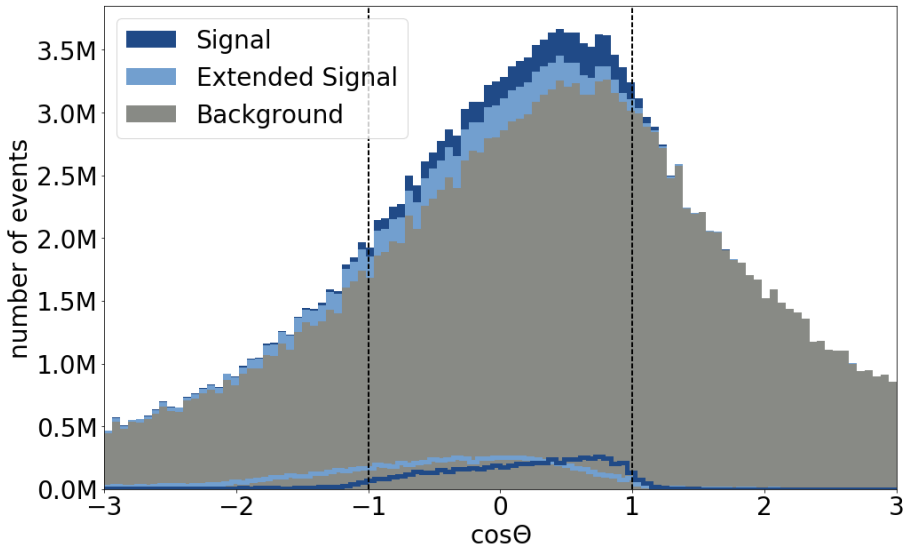


(a) A loose cut on the `SignalProbability` was performed.

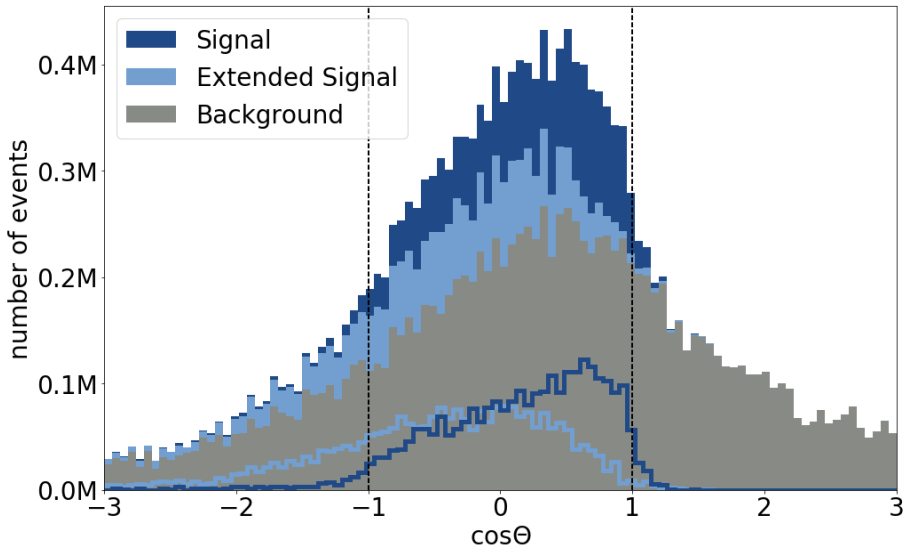


(b) A tight cut on the `SignalProbability` was performed.

Figure 4.10.: **Belle II:** The beam-constrained mass of the best hadronically tagged B^+ meson candidate per event. The spectrum of the different signal-definitions is shown.

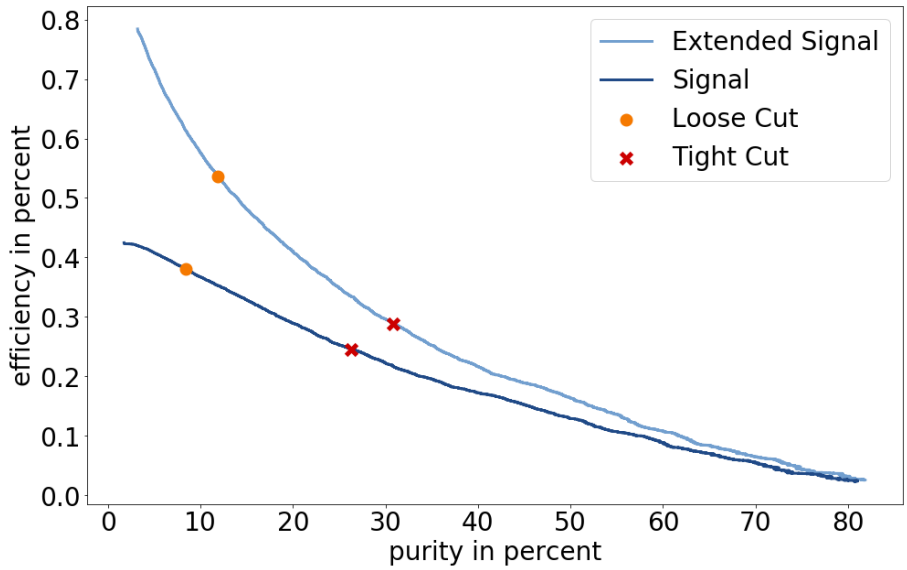


(a) A loose cut on the `SignalProbability` was performed.

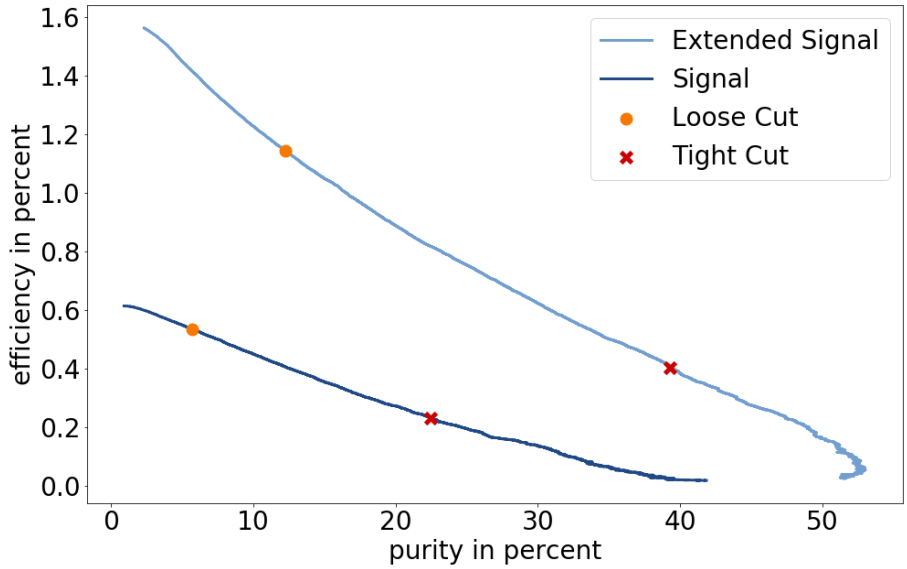


(b) A tight cut on the `SignalProbability` was performed.

Figure 4.11.: **Belle II:** $\cos\Theta_{\mathrm{BD}\ell}$ of the best semileptonically tagged B^+ meson candidate per event. The spectrum of the different signal-definitions is shown.

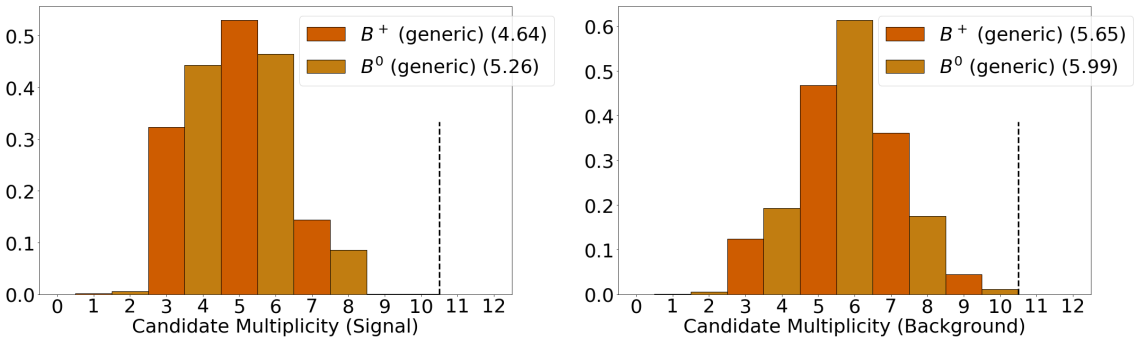


(a) Hadronic Tag



(b) Semileptonic Tag

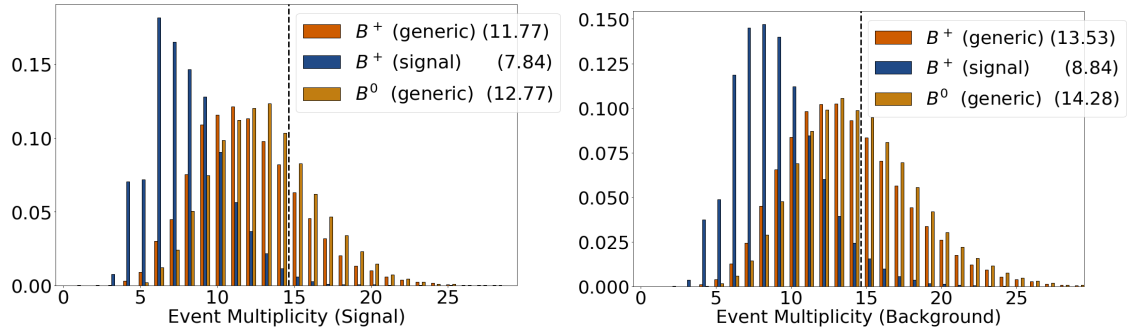
Figure 4.12.: **Belle II:** Receiver operating characteristic (ROC) curve of tagged B^+ for the signal and extended signal-definition. The loose (`SignalProbability` > 0.01) and tight (`SignalProbability` > 0.1) cut are shown in orange and red, respectively.



(a) The track multiplicity of correctly reconstructed B candidates.

(b) The track multiplicity of incorrectly reconstructed B candidates.

Figure 4.13.: Number of tracks (multiplicity) of the candidates. The dashed line indicates the default cut on the maximum numbers of tracks used during the application of the `FEI` in the benchmark analysis $B \to \tau\nu$ in this thesis. The number in parenthesis in the legend states the mean number of tracks for the corresponding component.



(a) The track multiplicity of the events with correctly reconstructed B candidates.

(b) The track multiplicity of the events with incorrectly reconstructed B candidates. An event is counted multiple times if it contains more than one background candidate.

Figure 4.14.: Number of tracks (multiplicity) of the events. The dashed line indicates the default cut on the maximum numbers of tracks used during the training of the `FEI` in this thesis. The rare decay $B^+ \to \tau^+\nu_\tau$ is used as signal. The number in parenthesis in the legend states the mean number of tracks for the corresponding component.

Moreover, the performance of the FEI on events with different multiplicities is barely influenced by the maximum number of reconstructed tracks allowed during the training as well (see Figure 4.15b). For instance, a FEI trained specifically on events with ≤ 6 tracks, does not perform better on events with 6 tracks than a FEI trained on all events. However, overall the FEI performs significantly better on low-multiplicity events, since it is much easier to reconstruct the correct B_{tag} meson. In particular, the tag-side efficiency for signal-sides with a low-multiplicity is usually significantly better (see Chapter 5).

The same behavior was found for hadronically tagged B^+, hadronically tagged B^0, semileptonically tagged B^+ and semileptonically tagged B^0 (see Section C.3.1).

The default training of the generic FEI considers only events with a maximum multiplicity of 14 during the training, and considers all events during the application. The number was chosen to save computing resources without any risk of decreasing the overall performance of the FEI on signal events (see Figure 4.13). The benchmark analysis in Chapter 5 uses the default training, but applies an additional cut on the multiplicity during the application to save computing time without discarding signal events.

4.3.3.2. Influence of beam-background

The exact beam-background conditions at Belle II are currently (at the time of writing) unknown. The current estimations range between a 20 to 30 fold increase compared to Belle. The exact numbers depend on the detector region. Therefore the influence of the beam-background on the performance of the FEI was studied.

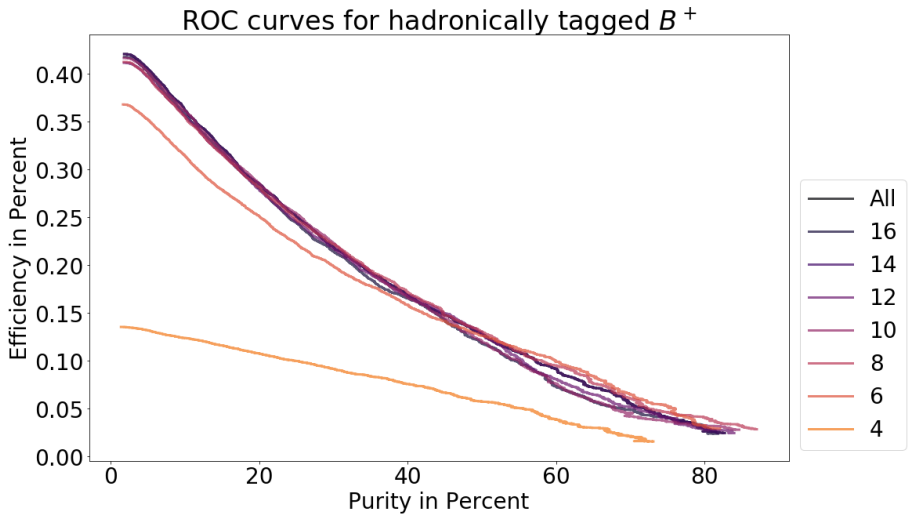
The generic FEI was **t**rained, with (T1) and without (T0) beam-background. Both were **a**pplied to independent generic $\Upsilon(4S)$ events, with (A1) and without (A0) beam-background. The results are shown in Figure 4.17. Only the best B_{tag} candidate per event was considered, and an additional cut on the beam-constrained mass was applied.

As expected, T0 performs better than T1 on A0, and T1 performs better than T0 on A1.

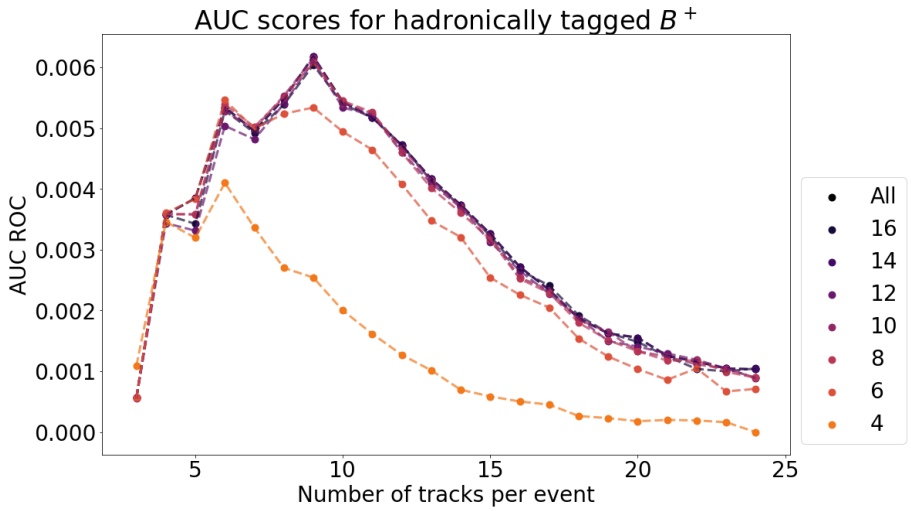
The training of the FEI is robust against beam-background. The loss in tag-side efficiency due to occurring beam-background in the training (that is the difference between T1 and T0) is of the order of 10%.

In contrast, the application of the FEI is sensitive to the beam-background. The loss in tag-side efficiency due to occurring beam-background in the application (that is the difference between A1 and A0) is of the order of 30%.

In particular the increased fake-rate (the fraction of reconstructed tracks which do not originate from a charged particle) caused by the increased beam-background



(a) Receiver operating characteristic (ROC) curves of the `FEI` applied to **all** events without a cut on the event multiplicity.



(b) The area under the ROC curve (AUC ROC) depending on the multiplicity of the event.

Figure 4.15.: Evaluation of the best hadronically tagged B^+ candidate per event, reconstructed by different configurations of the `FEI`. Each curve corresponds to a generic `FEI` with a cut on the maximum number of N tracks in the event during the **training**.

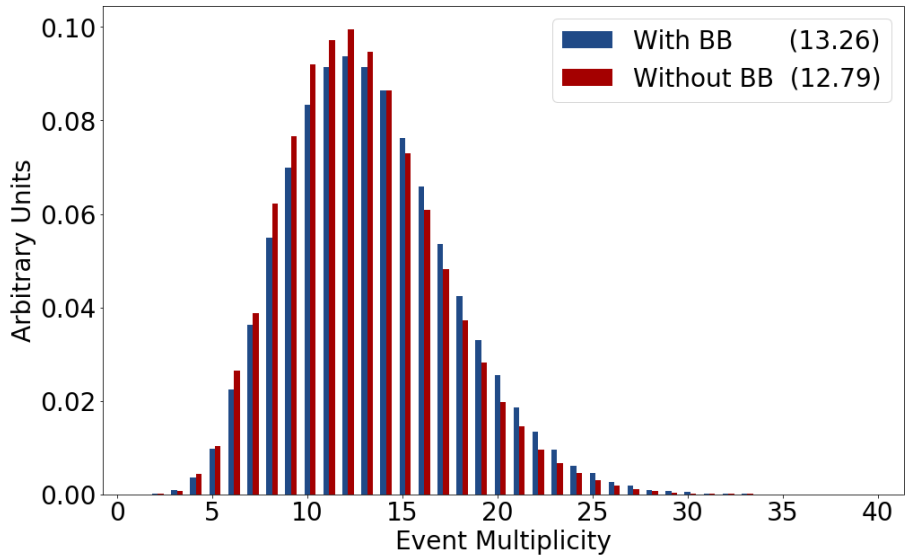


Figure 4.16.: The distribution of the event multiplicity for $\Upsilon(4S)$ events with and without beam-background (BB). On average, there is a fake track due to beam-background in every second event. The number in parenthesis in the legend states the mean number of tracks for the corresponding component.

is problematic (see Figure 4.16), because it leads to increased combinatorics and decreases the probability of identifying the correct B_{tag} candidate.

The same behavior was found for hadronically tagged B^+, hadronically tagged B^0, semileptonically tagged B^+ and semileptonically tagged B^0 (see Section C.3.2).

Finally, the completeness-constraint is very sensitive to an increased fake-rate. Also quantities like the extra energy in the electromagnetic calorimeter are sensitive to the amount of beam-background. However, it is expected that the fake-rate will be significantly reduced in the upcoming years, because the reconstruction software of Belle II is not yet optimized to reduce the fake-rate. This is further discussed in the benchmark analysis in Chapter 5.

4.3.4. Tag-side efficiency correction

The uncertainty on the tag-side efficiency of the `FEI` is (one of) the most important systematic uncertainties in the measurement of branching fractions of rare decays, e.g. $\mathcal{BR}(B^+ \to \tau^+\nu_\tau) = (1.25 \pm 0.28 \pm 0.27) \cdot 10^{-4}$ [15], with a systematic uncertainty due to the tag-side efficiency of $0.16 \cdot 10^{-4}$.

Therefore it is indispensable to calibrate and correct the tag-side efficiency on data using multiple control channels.

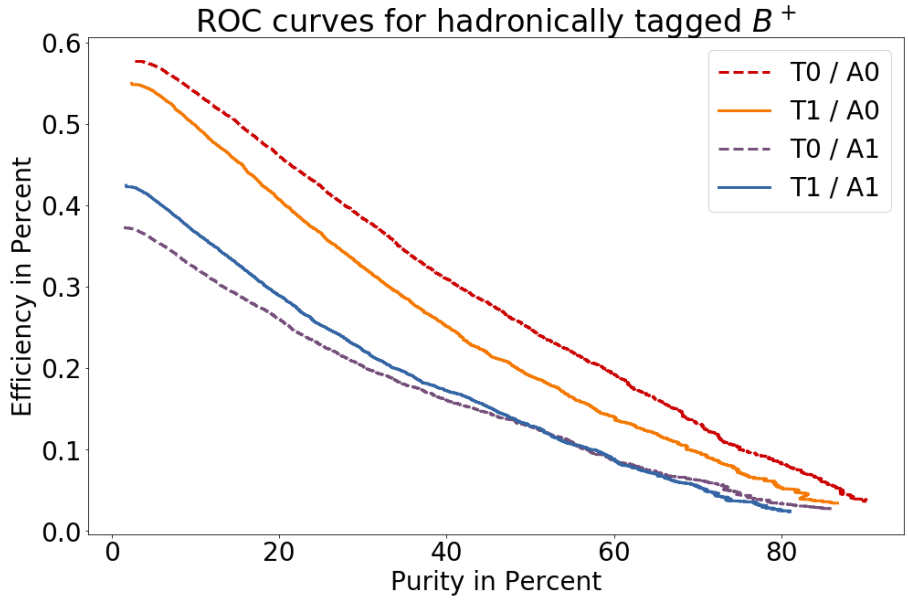


Figure 4.17.: Evaluation of the best hadronically tagged B^+ candidate per event, reconstructed by different configurations of the `FEI`. The solid (dashed) curves correspond to a generic `FEI` trained on simulated Belle II events with (without) beam-background. The blue and purple (red and orange) curves correspond to a generic `FEI` applied to simulated Belle II events with (without) beam-background.

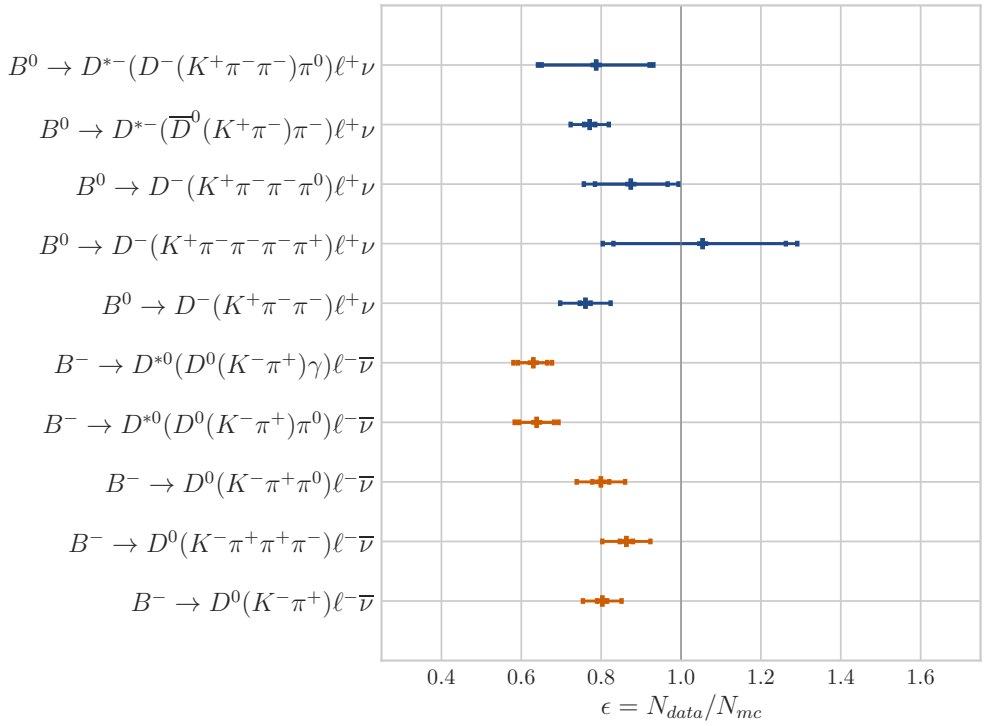


Figure 4.18.: The overall efficiency correction calculated by measuring the known branching fractions of 10 control channels on converted Belle data [76]. The calibration was performed on the latest `FEI` version 4.

The `FEI` was calibrated on converted Belle data within the scope of a supervised master's thesis [76]. I briefly summarize the results here.

The `FEI` was applied to the full $\Upsilon(4S)$ dataset recorded by Belle. The well-known branching fractions of ten control channels were measured using the `FEI`. The difference between the Monte Carlo expectation and the data yields tag-side channel dependent calibration factors for the tag-side efficiency, as well as their uncertainties.

Figure 4.18 summarizes the results for the ten control channels. The overall calibration factors for the latest `FEI` version 4 averaged over all control channels are

$$\varepsilon_{\text{charged}} = 0.74^{+0.014}_{-0.013} \pm 0.050$$
$$\varepsilon_{\text{mixed}} = 0.86^{+0.045}_{-0.050} \pm 0.054.$$

The measured values are compatible with the observed lower tag-side efficiency on recorded data in comparison with the Monte Carlo expectation in Figure 4.8.

The uncertainty of the tag-side efficiency has only a minor effect on searches of rare decays, which are statistically limited, e.g., in the search for $B \to K\nu\bar{\nu}$ (see [77]) or

$B \to \ell\nu\gamma$ (see [78]). In addition, the uncertainty on the tag-side efficiency can be removed by performing a relative measurement as the efficiency will cancel out due to the normalization, e.g., $R = \frac{\mathcal{BR}(B \to D\tau\nu_\tau)}{\mathcal{BR}(B \to D\ell\nu)}$ (see [79]).

4.3.5. Comparison with Full Reconstruction

The performance of the `FEI` was compared to the `FR` on converted Belle MC from the last Belle Monte Carlo campaign. All numbers shown here refer to the Monte Carlo expectation. As was shown in Section 4.3.4, the performance on data for the `FEI` is about 20% lower.

The tag-side efficiencies of the `FR` are stated in Table 4.8. The tag-side efficiencies of the `FEI` using an identical configuration, meaning the same decay-channels, are stated in Table 4.9.

Using the identical configuration, the `FEI` outperforms the `FR`, and increases the hadronic tag-side efficiency by 89% (83%) for charged (neutral) B mesons. This increase in tag-side efficiency is due to the best-candidate selection and improved classifiers.

Using the default configuration, the hadronic tag-side efficiency is increased by 171% (156%) for charged (neutral) B mesons. This further increase in tag-side efficiency is due to the additional new decay-channels.

The official `FR` did not provide a semileptonic tag. However, variants of the `FR` were used for semileptonic tagging (see [75] and [15]). The semileptonic tag-side efficiency of the default configuration of the `FEI` is increased by 177% for charged B mesons.

The stated results for the hadronic tag were verified on data (see Section 4.3.2.1 and Section 4.3.4). The stated results for the semileptonic tag are based on MC, but are consistent with the increases seen for the hadronic tag.

The performance of the `FEI` on Belle II MC events is worse than on Belle events, but still significantly better than the `FR` on Belle events. The hadronic tag-side efficiency is increased by 135% (111%) for charged (neutral) B mesons. As shown in Section 4.3.3.2 the `FEI` performance is sensitive to the fake-rate due to the increased beam-background. It is expected that the fake-rate will be drastically reduced as soon as the Belle II reconstruction software is finished and optimized.

In consequence, the stated increases in tag-side efficiency for Belle II are a lower bound of the expected improvements.

Altogether, the improvements of the `FEI` with respect to the `FR` are very consistent and explainable for all combinations of tag (hadronic or semileptonic), charge (charged or neutral), experiment (Belle or Belle II) and data type (MC events or recorded data).

Table 4.8.: Tag-side efficiency of the FR [75].

	Charged $B^{\pm}$	Neutral B^0
Hadronic	0.28 %	0.18 %
Semileptonic	0.65 %	–

Table 4.9.: Tag-side efficiency of the FEI using the identical configuration as the FR.

	Charged $B^{\pm}$	Neutral B^0
Hadronic	0.53 %	0.33 %

4.4. Outlook

The design of the FEI foresees the possibilities to extend the algorithms in the future. In this section, I summarize ideas and techniques that were not further investigated during this thesis. However, they have a large potential to further increase the key performance indicators (tagging efficiency, tag-side efficiency and purity) of the FEI.

4.4.1. Inclusive tagging using the FEI

The FEI can be used for of inclusive tagging. Preliminary studies on the signal channel $B^+ \to \mu^+\nu_\mu$ have shown that this is possible in principle. The FEI was applied to the rest-of-event of a reconstructed B_{sig}, and the intermediate pre- and post-cuts were relaxed as far as possible. A valid candidate was then created in $\mathcal{O}(10\%)$ of the cases.

The resulting four-momentum and vertex position is as good as that provided by the traditional inclusive tag. Now the suspected explicit decay-chain can be used to further improve the four-momentum and vertex position with the help of a decay tree fitter [70].

BASF2 does not include a usable decay tree fitter at the time of writing. Hence, further studies have been postponed.

4.4.2. $D \to X\ell\nu$

Semileptonic tagging provides a higher efficiency compared to its hadronic counterpart, but due to the missing kinematic information the tagged sample is not as pure. Until now, only semileptonic B decays were used for this.

However, semileptonic D decays can be used as well and behave similarly. 13% of the D^0 and 33% of the D^+ mesons decay semileptonically. The B meson is reconstructed in a hadronic decay-channel $B \to D(\dots)$, and the D meson is reconstructed in a semileptonic decay-channel $D \to (\dots)\ell\nu$. Using the known invariant mass of the D and the possibly occurring D^* meson, the missing kinematic information due to the neutrino can be partially recovered.

The current default `FEI` configuration (see Section C.1) already includes in addition semileptonic D decays.

Although, this new class of decay-channels yields impure tagged samples, it provides the next step towards even more inclusiveness, and can be beneficial in situations where the `FEI` would otherwise not output a valid candidate at all.

4.4.3. $X \to Y K_L^0$

A generic B decay contains in 29% of the cases a K_L^0. A K_L^0 can be detected by the Belle and the Belle II detector, but both provide only a rough energy and flight direction estimation. The situation can be compared to semileptonic decay-channels, where kinematic information is missing as well. Until now, decay-channels containing K_L^0 particles were not used.

Some Belle analyses employed a K_L^0 veto. In consequence, tagged events with a detected K_L^0 were rejected. This situation can be improved, especially in view of Belle II, which will provide a better K_L^0 identification.

The current default `FEI` configuration (see Section C.1) already includes in addition decay-channels with a K_L^0.

Although, this new class of decay-channels yields impure tagged samples, it provides the next step towards even more inclusiveness, and can be beneficial in situations where the `FEI` would otherwise not output a valid candidate at all.

4.4.4. $\pi^0 \to \gamma(\gamma)$

Many B and D decay-channels include π^0 particles, which decay immediately into two photons. A generic B decay contains in 89% of the cases a π^0.

Assuming a spherical distribution and a geometric detector acceptance of 91.1% (by taking the boost into consideration as shown in Section D.3.1) 14% of the π^0 will have one undetectable photon due to the acceptance. This number was confirmed using MC simulation.

The FEI could use π^0 candidates reconstructed from just one photon, and check if an assumed second photon could be outside of the detector acceptance. Using the known invariant mass of the π^0 and the detector acceptance, the momentum of the missing photon could be estimated, leading eventually to the recovery of the incomplete π^0 particles.

As in the case of semileptonic decay-channels and decay-channels with K^0_L, the FEI could profit even from incomplete π^0 particles in situations where the FEI would otherwise not output a valid candidate at all. Furthermore, many measurements of rare decays, which use exclusive tagging, extract the signal-yield from the extra energy in the electromagnetic calorimeter E_{ECL} (see Chapter 5). If the FEI assigns all tracks correctly to a tag-side B_{tag} meson, but misses an incomplete π^0, one additional photon from the tag-side will end up in E_{ECL}. Therefore, reconstructing incomplete π^0 has the potential to improve the resolution of the signal peak in E_{ECL}.

4.4.5. FEI on $\Upsilon(5S)$

The FEI can directly be applied to the $\Upsilon(5S)$ resonance. This resonance decays into a pair of $B\overline{B}$(5.5%), $B\overline{B}^*$(13.7%), $B^*\overline{B}^*$(38.1%), $B_s^*\overline{B}_s^*$(20.1%) mesons. The user has to add the B^* and B_s^* particles and their associated decay-channels to the FEI configuration. The powerful completeness-constraint can still be applied in this situation.

4.4.6. Deep FEI

Newly developed algorithms based on deep-learning can outperform established algorithms, as was shown in the case of flavour tagging (Section 3.3.1.2) and continuum suppression (Section 3.3.1.3). Both algorithms can be seen as a form of inclusive tagging, hence, it is obvious to study the applicability of deep-learning to exclusive tagging as well.

As described in Section 3.1.4.3, applications with a learnable distributed representation, are in particular suitable for deep-learning. The most successful applications of deep-learning use domain-specific knowledge to restrict the learnable representation, e.g. by using convolutional neural networks in image recognition (see [40, Chapter 9]), recurrent neural network in speech recognition (see [40, Chapter 10]), and relational neural networks for relational reasoning (see [80]).

In the case of exclusive tagging a reasonable representation of the input data (tracks and clusters) is already known from physics: it is the decay-chain of the $\Upsilon(4S)$. This representation has all desired properties defined in Section 3.1.4.3. The FEI is designed to reconstruct this decay-chain in an hierarchical approach: starting

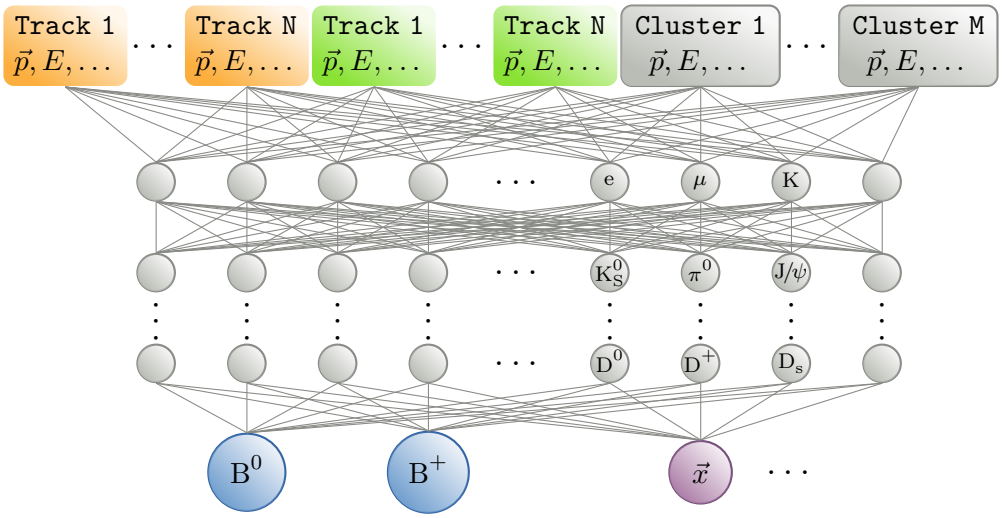


Figure 4.19.: Proposed network architecture for an exclusive tagging algorithm based on deep learning.

from the low-level information (tracks and cluster), creating intermediate particle candidates and finally providing the most likely B meson candidates.

In that sense the `FEI` is a deep-learning based algorithm, which enforces the decay-chain representation. It comes as no surprise that the hierarchical approach can be visualized by a network (see Figure 4.3), where the nodes are given by the particle and the weights are multivariate classifiers. In contrast to an ordinary deep neural network, the `FEI` can take advantage of complex existing algorithms in each step, e.g. vertex-fitting. Furthermore, the explicit decay-chain is automatically known, whereas it is not straight-forward to design a deep neural network that outputs the explicit decay-chain it reconstructed.

Nevertheless it is interesting to study the architecture of a possible exclusive tagging algorithm based on a deep neural network. The following paragraphs require basic knowledge of (deep) neural networks. A thorough introduction can be found in [40]. The proposed network architecture is shown in Figure 4.19.

4.4.6.1. Input Layer

The input layer takes the low-level information of each track and cluster. The main challenge is the variable input size. In the following I describe a possible input layer, based on the deep flavour tagger (see Section 3.3.1.2) and continuum suppression (see Section 3.3.1.3) algorithms.

A fixed-size input layer is used, which takes the N positive tracks (orange), N negative tracks (green) and M clusters (gray) with the highest energy (see Figure 4.19). For

each track the four-momentum, the POCA and the PID information is passed to the network. For each cluster the four-momentum, the cluster shape and the cluster timing is passed to the input layer.

In case of missing tracks and clusters the associated input neurons are set to zero, which is the natural choice for the four-momentum and PID information. In other words, we assume that our algorithm is "infrared-safe"[7]: additional inputs which suggest a particle with zero mass and momentum should not influence the result.

The inputs are ordered by their total momentum and are separated by charge. Preliminary studies for the continuum-suppression algorithm showed, that spherical coordinates and the center-of-mass frame should be used where appropriate.

There are several alternatives to deal with the variable input size, which I mention briefly.

One could use a recurrent neural network (RNN) and feed the network one track after another. RNNs exploit the sequential correlations in the provided sequence. However, there is no obvious order of tracks which has a sequential correlation. In fact the most high energetic tracks are most likely not from the same side.

A convolutional neural network (CNN) can be used, where the tracks and clusters are mapped onto spatial images, e.g., using their flight direction. CNNs exploit the spatial correlations in the provided images. Again, there is no obvious mapping of tracks, which has a high spatial correlation. In fact correlated[8] tracks are unlikely to have a similar flight direction.

Finally, multiple networks could be trained for each possible input size. This approach is likely to suffer from a high computational burden and low-statistics.

4.4.6.2. Hidden Layers

So far, the deep-learning based algorithms implemented in `BASF2` used ordinary fully connected feed-forward layers. Finding the optimal architecture of a neural network usually involves trial-and-error. Therefore, I only state some building blocks here.

Firstly, the gradient of the loss-function, used during the training phase to adapt the weights of the deep neural networks, is known to be unstable for the foremost layers. This is known as the vanishing or exploding gradient problem [81]. Several solutions for this problem exist e.g. greedy layer-wise pre-training.

I propose the usage of intermediate targets following the hierarchical structure of the `FEI` (symbolized by the inscribed particle name in the gray hidden neurons in

[7] The term is used as an analogy and not in the strict sense of the word.

[8] Track correlation can be defined using their degree of kinship, that is two tracks with a common mother have the highest correlation.

Figure 4.19). Hence, the neurons in the first hidden layers should learn to identify final state particles. In subsequent layers the presence and four-momentum of intermediate resonances can be learned. In other words, some neurons in each layer are enforced to learn physically interpretable quantities by adding additional terms to the loss-function.

Therefore the loss-function naturally has a direct influence on the foremost layers, and the network can learn the known reasonable representation (the decay-chain of the $\Upsilon(4S)$) more easily. Furthermore, this scheme allows an implicit greedy layer-wise training, by changing the importance of the terms in the loss functions during the training. At the start the terms associated with neurons near the input layer are more important, while at the end only the final target in the output layer is used.

Secondly, relational neural networks (RelNN) can be used (see [80]). RelNNs learn pairwise operations, and exploit permutation invariance. For instance one could imagine: the vertex of two tracks, the invariant mass of two tracks, or the angle between two clusters. The calculation of these quantities is possible and reasonable for each given pair of tracks (or any other intermediate representation of the network). The advantage compared to a fully-connected layer is the weight-sharing. While a fully-connected layer has to learn $\mathcal{O}(N^2)$ different functions to compare N objects using one example per event during the training phase, the relational neural network learns one function using $\mathcal{O}(N^2)$ examples per event during the training phase.

4.4.6.3. Output Layer

There are two possible architectures for the output layer.

If the algorithm only operates on the rest-of-event of an already reconstructed signal-side (the equivalent to the specific `FEI`), a single output neuron with a sigmoid activation function and a cross-entropy loss function is sufficient. The network learns to output 1 if the given rest-of-event forms a valid B_{tag} candidate, and 0 otherwise. This is shown in blue in Figure 4.19.

Additional outputs can estimate the four-momentum, the vertex, the charge, the flavour, the number of missing neutrinos or any other quantities, under the assumption that the formed candidate is valid. This is shown in purple in Figure 4.19.

If the algorithm operates on the entire event it has to assign each track and cluster either to the signal or to the tag-side. This is much more challenging. I propose to use $2N + M$ outputs, which output the assignment of the $2N$ track and M clusters to the tag-side 1 or rest-of-event (that is the signal-side) 0. However, each output is created by a small independent neural network, which gets the learned representation of the main network, and the original inputs associated with the track or cluster as

input. In other words, the small networks predict the assignment of each track or cluster using the learned representation.

This architecture has the additional advantage, that the main network could be the encoder part of an auto-encoder network, which can be (pre, post)-trained on data.

4.4.6.4. Final Remarks

The proposed architecture is based on the experiences of studies conducted for the flavour tagger and continuum suppression algorithm. It is expected that only parts of it would ultimately prove useful in a future deep `FEI`.

4.5. Conclusion

The `Full Event Interpretation` enables Belle II physicists to measure a wide range of interesting decays with a minimum of detectable information. It exploits the unique experimental setup of B factories.

During this thesis the `FEI` was further developed from a "Proof of Concept" to a standard tool used in production. The algorithm was extensively studied on Belle and Belle II MC, and the obtained results were verified on data recorded by the Belle detector using the `b2bii` package.

The `FEI` more than doubles the tag-side efficiency for all combinations of tag (hadronic/semileptonic), charge (charged/neutral) and experiment (Belle/Belle II), compared to its (already very successful) predecessor. The tag-side efficiency for hadronically tagged B mesons was calibrated on Belle data, and is ready for the productive usage in analyses on converted Belle data.

Besides, many software tools originally developed for the `FEI` like `FastBDT` and `FastFit` proofed useful in many other contexts. The `FEI` was the first algorithm to take full advantage of the `b2bii` (see Chapter 2) and `mva` (see Chapter 3) package.

Finally, there are many possible extensions and unexplored applications of the `FEI`, which provide an exciting and fruitful area for further research.

Chapter 5

$B \to \tau\nu$

The leptonic decays of charged B mesons belong to the golden decay-channels at B factories. The expected branching fractions in the Standard Model of particle physics (SM) are small, hence leptonic decays are so-called **rare-decays**. The theoretical uncertainties on the predicted branching fraction are small and the possible influence of physics beyond the Standard Model (BSM) can be large.

The experimental signature is (mostly) a single track produced by a lepton or its decay product. The accompanying neutrinos are not detected. The leptonic decays are currently only accessible via the unique experimental setup at B factories.

The tauonic decay $B \to \tau\nu$, is of particular interest. Its branching fraction is the largest among the leptonic decays, due to the high mass of the τ meson elevating the helicity suppression. On the other hand the τ is unstable in the detector, hence it can only be measured indirectly using its decay products. Multiple neutrinos in the final state render this decay particularly experimentally challenging. It is the most prominent use-case for exclusive tagging.

During this thesis the decay $B \to \tau\nu$ was investigated as a benchmark analysis for the developed software tools and algorithms presented in the preceding chapters. The previous Belle analyses ([16] and [15]) were repeated on the full $\Upsilon(4S)$ dataset recorded by Belle using the `b2bii` package. Furthermore, a sensitivity study on simulated Belle II Monte Carlo events was conducted.

The main goal was to validate the developed tools and establish a prototype analysis, which can be used as a starting point for other analyses based on `b2bii` and/or the `FEI`.

In the following chapter, I describe the theoretical background (Section 5.1), the experimental measurement at Belle and Belle II (Section 5.2), the validation of the analysis strategy using off-resonance data and sidebands (Section 5.3), and the obtained results (Section 5.4).

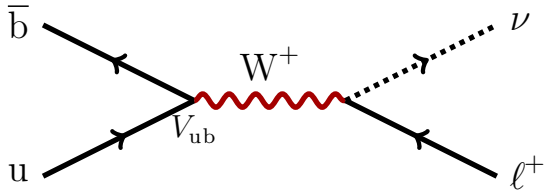


Figure 5.1.: Feynman Diagram of the leading Standard Model process mediated by a charged W boson for the purely leptonic decay of a charged B meson.

5.1. Theory

The Standard Model of particle physics (SM) describes the elementary particles and their fundamental interactions, except for gravity.

The SM is a relativistic quantum field theory, that is the particles are described by quantum fields, whose equations of motion are invariant under the Poincaré space-time symmetry. The fundamental interactions (the electromagnetic, strong and weak interaction) are described by local gauge symmetries. The field equations of the SM are encoded in a Lagrangian density $\mathcal{L}$, from which the field equations can be derived using the Lagrange formalism.

The SM (with massless left-handed neutrinos) has 19 free parameters. All of them can and have been experimentally determined. A thorough introduction to the SM can be found in [82].

5.1.1. Leptonic $B^{\pm}$ meson decays in the Standard Model

The branching fraction of the leptonic decay of charged $B^{\pm}$ mesons can be calculated in the SM using perturbation theory. The leading order Feynman Diagram is shown in Figure 5.1. It results in the following expression for the branching fraction [1, Chapter 7.10.2]

$$\mathcal{B}\left(B^+ \to \ell^+\nu\right)_{SM} = \frac{G_F^2 M_B M_\ell^2}{8\pi}\left(1 - \frac{M_\ell^2}{M_B^2}\right)^2 {f_B}^2 |V_{ub}|^2 \tau_B, \tag{5.1}$$

where all occurring constants can be theoretically calculated and/or independently experimentally measured. The numerical values and their relative uncertainties are summarized in Table 5.1. The measurement of the branching fraction of the leptonic decay of charged B mesons can reduce the existing uncertainty, and provides

Table 5.1.: Numerical values and relative uncertainties of all quantities required to calculate the branching fraction of $B^+ \to \ell^+\nu$ in the Standard Model. All values are taken from the PDG review 2016 [58].

	PDG Value	Relative uncertainty
G_F	$11.7\,\text{TeV}^{-2}$	$5 \cdot 10^{-7}$
m_B	5.28 GeV	$3 \cdot 10^{-5}$
m_τ	1.78 GeV	$7 \cdot 10^{-5}$
τ_B	1.64 ps	$2 \cdot 10^{-3}$
f_B	187.1 MeV	$2 \cdot 10^{-2}$
$\lvert V_{ub}\rvert_{inc}$	$4.49 \cdot 10^{-3}$	$5 \cdot 10^{-2}$
$\lvert V_{ub}\rvert_{exc}$	$3.72 \cdot 10^{-3}$	$5 \cdot 10^{-2}$
$\lvert V_{ub}\rvert_{avg}$	$4.09 \cdot 10^{-3}$	$1 \cdot 10^{-1}$

Table 5.2.: Standard model prediction for $B^+ \to \ell^+\nu$ calculated by Equation 5.1 using the numerical values given in Table 5.1 and the averaged $|V_{ub}|$ value.

	SM Prediction	PDG 2016
$\mathcal{B}\left(B^+ \to e^+\nu_e\right)$	$(1.09 \pm 0.21) \cdot 10^{-11}$	$< 9.8 \cdot 10^{-7}$ CL=90%
$\mathcal{B}\left(B^+ \to \mu^+\nu_\mu\right)$	$(4.65 \pm 0.91) \cdot 10^{-7}$	$< 1.0 \cdot 10^{-6}$ CL=90%
$\mathcal{B}\left(B^+ \to \tau^+\nu_\tau\right)$	$(1.03 \pm 0.2) \cdot 10^{-4}$	$(1.06 \pm 0.20) \cdot 10^{-4}$

an independent verification for the consistency of the SM with the experimental observations.

The leading dependency of the branching fraction on the squared lepton mass M_ℓ^2 is caused by the so-called helicity suppression [58, p. 1109]. In consequence, the τ lepton is expected to have the largest branching fraction among the leptonic B decay channels, as can be seen in Table 5.2.

5.1.1.1. The V_{ub} puzzle

The CKM matrix element $|V_{ub}|$ corresponds to one of the 19 free parameters of the SM. It induces the dominant theoretical uncertainty in the prediction of the branching fraction of $B^+ \to \ell^+\nu$. $|V_{ub}|$ can be determined by several distinct measurements, which differ in their associated theoretical and experimental uncertainties.

The **full decay width of the inclusive decay** $B \to X_u\ell\nu$ can be calculated with high precision, that is $< 5\%$ theoretical uncertainty, since QCD form-factors are not required. The inclusive decay has a high branching fraction,

but the experimental measurement is difficult due to the high background contamination from the CKM-favored $B \to X_c \ell\nu$ decay.

Therefore it is easier to measure **the partial decay width of the inclusive decay** $B \to X_u \ell\nu$, by demanding a high momentum for the lepton. However, the theoretical prediction of a partial decay width in this phase-space region is much more difficult and requires the knowledge of a non-perturbative distribution function, the so called "shape-functions" [58, p.1317].

The **exclusive measurement of the branching fraction** of $B \to \pi\ell\nu$ is experimentally clean, but has a lower branching fraction and requires in addition form-factors on the theoretical side. The measurement can be performed with and without exclusive tagging. Both approaches yield compatible results [58, p. 1319].

There is a slight tension between the determination of $|V_{ub}|$ from inclusive and exclusive measurements, as can be seen in Table 5.1. [58] states that the average value of $|V_{ub}|$ should be treated with caution "*given the poor consistency between the two determinations*". Nevertheless, this thesis uses the average value (see Table 5.1), but notes the differences between the $|V_{ub}|$ from exclusive and inclusive measurements where appropriate.

Finally, the measurement presented in this thesis can provide an independent determination of $|V_{ub}|$. The theoretical prediction of the branching fraction of $B \to \ell\nu$ is theoretically easy, since there are no strongly interacting particles in the final state. Experimentally it is challenging since there is usually only one measurable track.

5.1.1.2. The f_B decay constant

The contribution of the strong interaction in the decay $B \to \ell\nu$ is represented by the decay constant f_B. It is related to the overlap of the bottom and up quark wave-functions in the meson.

The decay constant induces the second largest theoretical uncertainty besides the CKM matrix element $|V_{ub}|$. It can be calculated using lattice QCD simulations and QCD sum rules. Both approaches yield compatible results. This thesis uses the average value (see Table 5.1) from lattice QCD simulations determined by [58, p.1117].

5.1.2. Type II Two Higgs Doublet Model

New physics contributions could influence the measurement of the branching fraction of $B \to \ell\nu$. The most famous model considered in the literature in connection

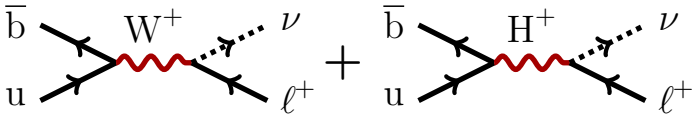


Figure 5.2.: Feynman Diagrams of the leading Standard Model process and a hypothetical new physics process mediated by a charged Higgs boson for the purely leptonic decay of a charged B meson.

with $B \to \ell\nu$ is the Type II Two Higgs Doublet Model [83]. In this model, the Standard Model Higgs Doublet is replaced with two Higgs Doublets, where one couples to down quarks and charged leptons, while the other couples to up quarks. The model introduces additional parameters $\tan\beta$, that is the ratio of the two vacuum expectation values of the two Higgs Doublets, and the masses $M_{H^\pm}$, M_{H^0} and M_{A^0} of the four additional Higgs particles.

The model provides another leading order Feynman Diagram shown in Figure 5.2, in which the $W^\pm$ boson is replaced by a hypothetical charged Higgs. The two diagrams (amplitudes) interfere destructively, hence the branching fraction expected in the SM is further suppressed by a multiplicative factor [1, Chapter 7.10.2]:

$$\mathcal{B}\left(B^+ \to \ell^+\nu\right)_{\text{2HDM}} = \mathcal{B}\left(B^+ \to \ell^+\nu\right)_{\text{SM}} \cdot \left(1 - \frac{M_B^2 \tan^2\beta}{M_{H^+}^2}\right)^2 . \tag{5.2}$$

The mass of the charged Higgs M_{H^+} is already tightly constrained by weak radiative B meson decays $M_{H^+} > 580$ GeV [84]. Nevertheless, a valid value of $\tan\beta \geq 35$ can still yield a modification of the branching fraction, as large as the current theoretical and experimental uncertainties.

Therefore, the measurement of $B \to \ell\nu$ can provide independent constraints on beyond the SM models like the Type II Two Higgs Doublet Model.

5.1.3. τ decay models

This thesis investigates $B \to \tau\nu_\tau$, that is the leptonic decay of the charged B meson involving the heaviest charged lepton. This decay is particularly challenging because the τ is not stable in the detector, consequently the measurement has to deal with one or more neutrinos and the kinematics of the τ decay.

The kinematics of the decay $\tau \to \ell\nu\bar{\nu}$ is well-known[1]. The same holds true for $\tau \to \pi\nu$, where only a single hadron has to be considered[2]. The decay into multiple hadrons (in particular $\tau \to \pi\pi\pi\nu$) is more complicated due to the occurring intermediate resonances and their interference. Belle explicitly simulated the decays $\tau \to \rho[\to \pi\pi\nu]\nu$ and $\tau \to a_1[\to \pi\pi\pi]\nu$ without taking further decay channels into the same final state into account[3]. In contrast, Belle II simulates the hadronic decay of $\tau \to \pi\pi\nu$ and $\tau \to \pi\pi\pi\nu$ inclusively based on [85] and takes interferences into account[4].

The implications of this different treatment in Belle and Belle II for this analysis are mostly technical details related to the Monte Carlo matching algorithm, and are not further discussed.

5.1.4. Previous measurements

The decay B $\to \tau\nu_\tau$ was already observed in previous measurements at B factories. Both, the Belle ([16] and [15]) and the BaBar ([86] and [87]) experiment conducted measurements using hadronic and semileptonic tagging. The previous measurements and the Standard Model predictions using V_{ub} from exclusive (SMExc) and inclusive (SMInc) measurements are shown in Figure 5.3. The Belle measurements are more precise due to the larger recorded dataset. The hadronic and semileptonic tags have similar statistical power. The current world average of the measured branching fraction stated by [58], is compatible with both predictions SMExc and SMInc.

5.2. Measurement

The measurement of the branching fraction of the decay B $\to \tau\nu_\tau$ involves five reconstruction steps, which are explained in detail in this chapter.

1. The events are preselected (see Section 5.2.1).
2. The signal side is reconstructed in five distinct decay channels of the τ lepton (see Section 5.2.2).
3. The tag side is reconstructed by the `FEI` (see Section 5.2.3).
4. Both sides are combined to a $\Upsilon(4S)$ candidates and the **completeness-constraint** is applied (see Section 5.2.4).

[1] The `EvtTaulnunu` model of `evtgen` is used.
[2] The `EvtTauScalarnu` model of `evtgen` is used.
[3] The `EvtTauVectornu` model of `evtgen` is used.
[4] The `EvtTauHadnu` model of `evtgen` is used.

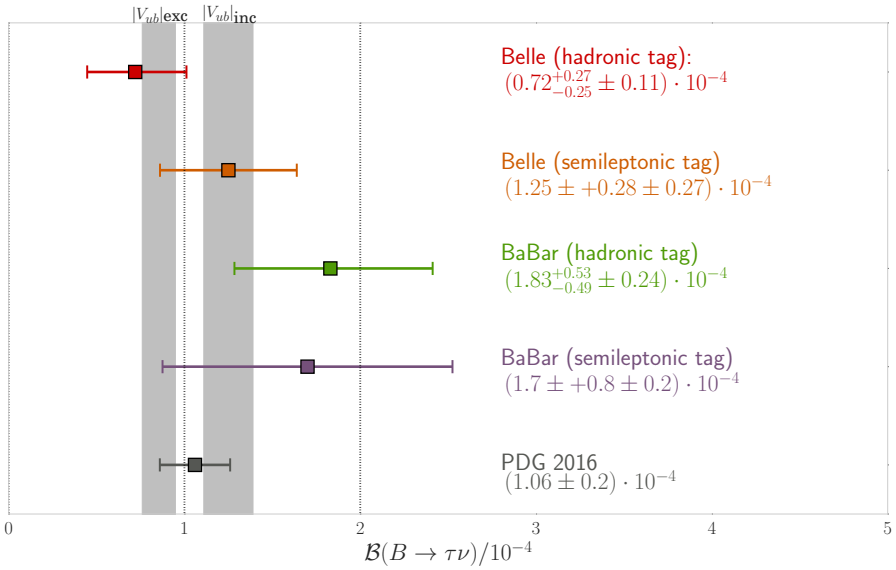


Figure 5.3.: Current experimental status of the measurement, including the Standard Model predictions (gray bands) using V_{ub} from exclusive and inclusive measurements.

5. The branching fraction is extracted with an extended unbinned maximum likelihood fit on the extra energy in the electromagnetic calorimeter of the selected candidates (see Section 5.2.5).

The measurement is performed on converted data recorded by the Belle detector and Monte Carlo simulated Belle II events. The hadronically and semileptonically tagged samples are treated as independent measurements in this thesis. The hadronic tag already includes the tag-side efficiency correction factors obtained by [76]. At the time of writing there was no tag-side efficiency correction available for the semileptonic tag. In consequence, the semileptonic results have to be taken with a grain of salt.

If not stated otherwise, all figures and tables for Belle were produced using one stream ($\equiv 711\,\mathrm{fb}^{-1}$) of Monte Carlo simulated events to model the background processes and a high-statistics sample of 100 million signal-events. For Belle II, high-statistics samples of 90 million events per component were used, and scaled to the equivalent of one stream.

The current version of the reconstruction software of Belle II is not fully optimized yet. During this benchmark analysis an extremely high fake-rate (up to 7 additional tracks in correctly reconstructed events), and a poor ECL resolution with large backgrounds caused by beam-background was observed. It is expected that these **reconstruction issues** will be solved in the future (see also Section 4.3.3.2). Nevertheless, the following figures and tables are given for both Belle and Belle II, for comparison. The

final fit is only performed on converted Belle data, because the current reconstruction issues prevent a meaningful result using the Belle II Monte Carlo simulation.

5.2.1. Skimming

The measurements consider only events with ten or less **good tracks**, that is tracks with impact parameters of less than 2 cm and 4 cm in transverse and beam axis direction, respectively.

This reduces the computational effort drastically without loosing any signal events. As was shown in Figure 4.13, the fraction of B_{tag} candidates with more than seven tracks, which are correctly tagged by the `FEI`, is negligible. The B_{sig} requires at most three tracks (for most channels only one). Therefore, the `FEI` cannot assign a correct tag to signal events with more than ten good tracks, and the events can be discarded without loosing correctly tagged signal events. Even incorrectly tagged signal events are only affected in the three-prong τ decay channel, as can be seen from Figure 4.13b.

Figure 5.4 shows this cut and the corresponding selection efficiencies for Belle and Belle II. The observed difference between Monte Carlo simulation and data was already investigated and discussed in Section 2.3.3.3. The increased number of tracks in Belle II events is caused by the reconstruction issues mentioned above.

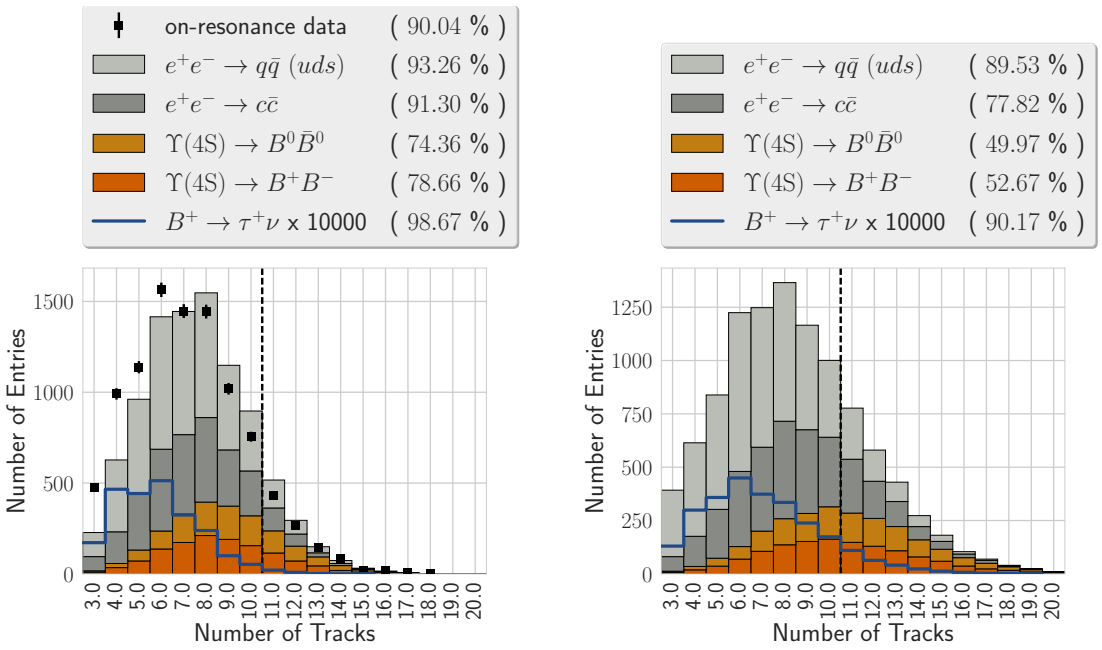
5.2.2. Signal side reconstruction

The B meson travels less than 1 mm inside the Belle II detector due to its short life-time of $1.64 \cdot 10^{-12}$ s. The τ lepton, produced by the investigated decay $B \to \tau\nu_\tau$, has an even shorter life-time of $2.9 \cdot 10^{-13}$ s. Therefore the τ lepton does not leave the beam-pipe and cannot be detected directly.

In this thesis the signal-side is reconstructed in five different τ decay-channels covering a total branching fraction of $\mathcal{BR}(\tau) = 80.8\%$ (see Figure 5.6).

Each decay-channel has distinct properties and consequently individual selection criteria, which are described below. As can be seen from Table 5.3, the signal-side selection criteria suppress the background by two orders of magnitude, while maintaining most of the signal. The measured selection efficiencies on converted data are compatible with the Monte Carlo simulation of background. The selection is in general better on Belle Monte Carlo than on Belle II, due to the reconstruction issues mentioned above.

Although the tag-side is reconstructed by the `FEI`, it proved useful to already discard candidates whose rest-of-event is not compatible with a single B meson. A fast,

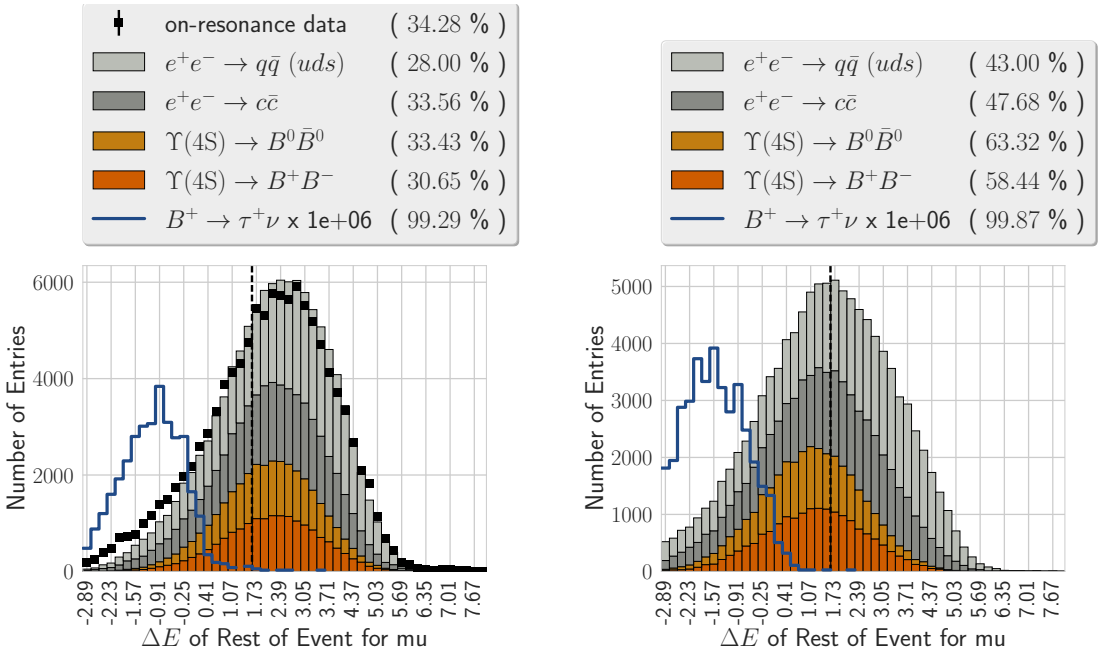


(a) Belle (b) Belle II

Figure 5.4.: Skim cut on the number of good tracks. Evaluated on 10000 events for each component, scaled according to their expected relative fractions.

Table 5.3.: The overall efficiencies of the signal-side selection: $\varepsilon_{\text{SignalEvents}}$, is the fraction of events containing the stated signal decay channel, in which a correct candidate was reconstructed. This includes the detector acceptance, trigger and reconstruction efficiencies and the skimming. The signal-side selection alone is characterized by: the fraction of correct candidates which survive the signal-side selection $\varepsilon_{\text{Signal}}$; the fraction of background candidates which survive the signal-side selection $\varepsilon_{\text{Background}}$; and (in the case of Belle) the fraction of candidates on converted data which survive the signal-side selection $\varepsilon_{\text{Data}}$. Evaluated on 10000 events for each component.

Channel	Belle				Belle II		
	$\varepsilon_{\text{SignalEvents}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{SignalEvents}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
μ	0.40	0.48	0.006	0.006	0.53	0.59	0.018
e	0.68	0.71	0.008	0.010	0.53	0.66	0.025
π	0.89	0.88	0.087	0.096	0.76	0.83	0.096
ρ	0.41	0.65	0.014	0.015	0.32	0.69	0.011
a_1	0.44	0.67	0.060	0.058	0.35	0.53	0.030



(a) Belle (b) Belle II

Figure 5.5.: ΔE distribution of the rest-of-event in GeV for the $\tau \to \mu\nu\bar{\nu}$ decay channel with a cut on $\Delta E < 1.5\,\text{GeV}$. Evaluated on 10000 events for each component, scaled according to their expected relative fractions.

efficient and easy to calculate selection variable is the deviation ΔE of the total energy in the center-of-mass system of the rest-of-event from the nominal beam-energy in the center-of-mass system. Figure 5.5 shows the efficiency of this selection for the $\tau \to \mu\nu\bar{\nu}$ decay channel, the distributions in the other channels look nearly identical. This selection saves computation time, once the tag-side is reconstructed by the `FEI`, a superior[5] selection on the same quantity can be performed.

5.2.2.1. $\tau \to \mu\nu\bar{\nu}$

The τ decay into a muon and two neutrinos has a branching fraction of 17.4%. It can be efficiently selected using the muon identification provided by the PID sub-detectors of Belle II (see Figure 5.7). We expect to find one track compatible with a muon hypothesis in the event besides the B_{tag} meson. The selection criteria and corresponding efficiencies are stated in Table 5.4.

[5]The knowledge of the explicit decay-chain of the B_{tag} increases the resolution of ΔE.

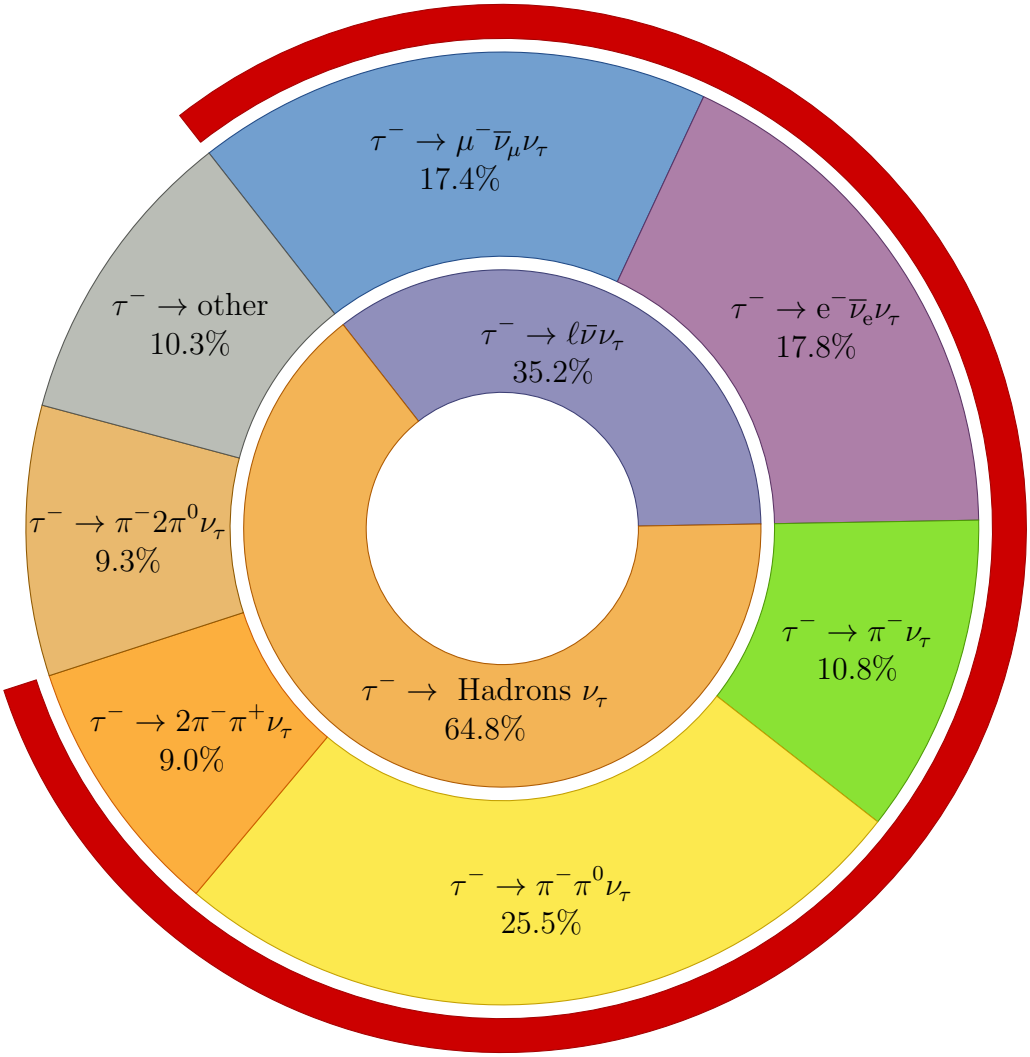
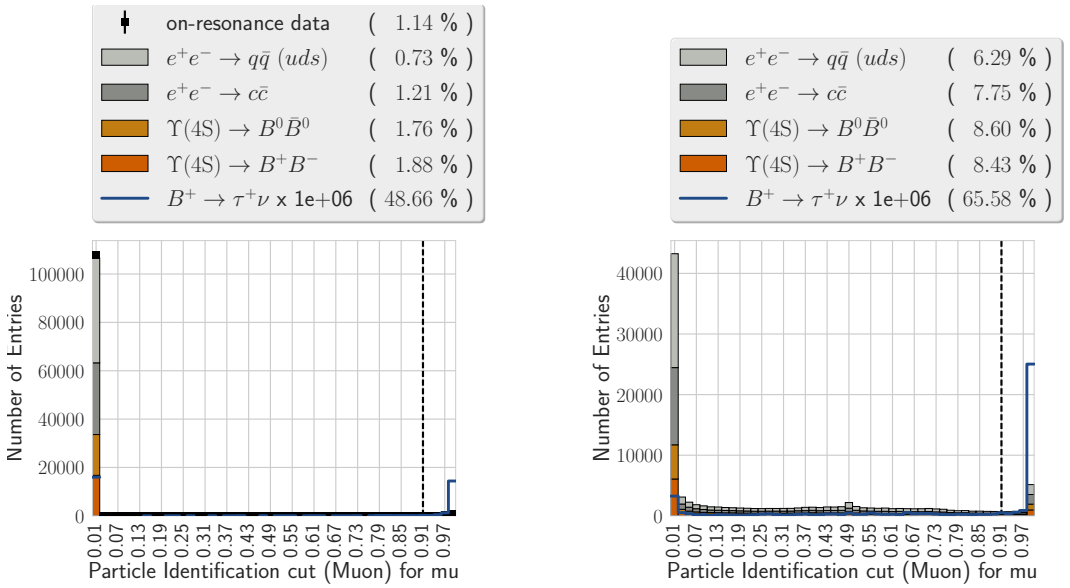


Figure 5.6.: Visualization of the τ branching fractions: used in this thesis (red arc); into exclusive decay-channels (outer circle); and into inclusive hadronic and leptonic decay-channels (inner circle).

Table 5.4.: The selection criteria for the $\tau \to \mu\nu\bar{\nu}$ decay-channel with the corresponding selection efficiencies for signal candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$ and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. Evaluated on 10000 events for each component.

Selection Criterion	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
# Tracks ≤ 10	0.99	0.82	0.84	0.92	0.65
$p^*_\tau > 0.1\,\text{GeV}$	1.00	0.97	0.97	0.99	0.91
$\Delta E \leq 1.5\,\text{GeV}$	0.99	0.31	0.34	1.00	0.50
dr < 2	0.98	0.75	0.74	0.99	0.91
$-4 <$ dz < 4	0.96	0.68	0.66	0.96	0.83
$\text{PID}_\mu > 0.9$	0.49	0.01	0.01	0.66	0.07



(a) Belle (b) Belle II

Figure 5.7.: PID_μ distribution for the $\tau \to \mu\nu\bar{\nu}$ decay channel with a cut on $\text{PID}_\mu > 0.9$. Evaluated on 10000 events for each component, scaled according to their expected relative fractions.

Table 5.5.: The selection criteria for the $\tau \to e\nu\bar{\nu}$ decay-channel with the corresponding selection efficiencies for signal candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$ and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. Evaluated on 10000 events for each component.

Selection Criterion	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
# Tracks ≤ 10	0.99	0.82	0.84	0.91	0.65
$p_\tau^* > 0.1\,\text{GeV}$	0.99	0.97	0.97	0.99	0.91
$\Delta E \leq 1.5\,\text{GeV}$	0.99	0.31	0.34	1.00	0.50
dr < 2	0.96	0.75	0.74	0.98	0.91
$-4 <$ dz < 4	0.94	0.68	0.66	0.94	0.83
$\text{PID}_\mu < 0.9$	1.00	0.99	0.99	0.90	0.93
$\text{PID}_\text{e} > 0.9$	0.74	0.04	0.04	0.86	0.16

5.2.2.2. $\tau \to e\nu\bar{\nu}$

The τ decay into an electron and two neutrinos has a branching fraction of 17.8%. It can be efficiently selected using the electron identification provided by the PID sub-detectors of Belle II (see Figure 5.8). We expect to find one track compatible with an electron hypothesis in the event besides the B_{tag} meson. The selection criteria and corresponding efficiencies are stated in Table 5.5.

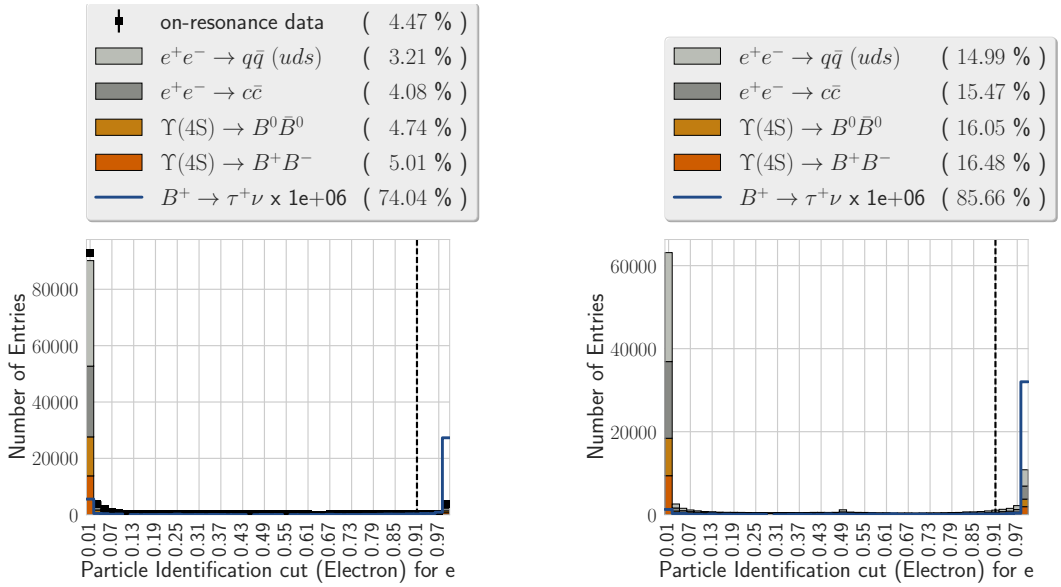
5.2.2.3. $\tau \to \pi\nu_\tau$

The τ decay into a pion and one neutrino has a branching fraction of 10.8%. It is a two-body decay, hence the momentum of the pion in the rest-frame of the τ is known

$$p_\pi^* = \frac{1}{2}\left(m_\tau - \frac{m_\pi^2}{m_\tau}\right) = 0.88\,\text{GeV} \tag{5.3}$$

Together with the momentum of the τ lepton in the rest-frame of the B meson and the momentum of the B meson in the center-of-mass system

$$p_\tau^* = 2.34\,\text{GeV} \tag{5.4}$$
$$p_\text{B}^* = 0.33\,\text{GeV} \tag{5.5}$$



(a) Belle (b) Belle II

Figure 5.8.: PID_e distribution for the $\tau \to \text{e}\nu\bar{\nu}$ decay channel with a cut on $\text{PID}_\text{e} > 0.9$. Evaluated on 10000 events for each component, scaled according to their expected relative fractions.

we can calculate the momentum of the pion in the center-of-mass system using the rapidities of the τ lepton and B^+ meson

$$\eta^*_\tau = \text{arsinh}\frac{p^*_\tau}{m_\tau} = 1.089 \tag{5.6}$$

$$\eta^*_\text{B} = \text{arsinh}\frac{p^*_\text{B}}{m_\text{B}} = 0.062 \tag{5.7}$$

and finally the minimal and maximal pion momentum in the center-of-mass system

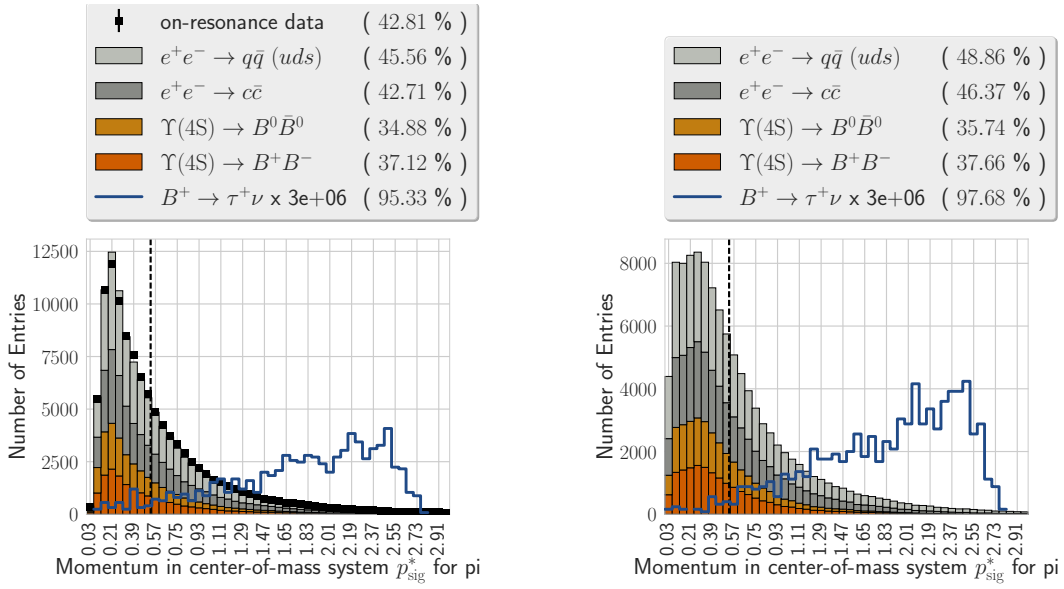
$$\eta = \pm\eta^*_\tau \pm \eta^*_\text{B} \tag{5.8}$$

$$p_\pi = p^*_\pi \cosh\eta - \sqrt{m^2_\pi + (p^*_\pi)^2}\sinh\eta \tag{5.9}$$

$$= 0.26\,\text{GeV} \quad \text{minimum} \tag{5.10}$$

$$= 2.80\,\text{GeV} \quad \text{maximum.} \tag{5.11}$$

The exact p_π distribution is skewed towards higher momenta due to the fixed helicity of the neutrino (see Figure 5.9). We expect to find one track with a compatible momentum in the event besides the B_tag meson. In addition, we expect a large cross-feed in this decay-channel from other signal decay-channels. For instance, a lepton which does not pass the PID criteria will automatically be picked up by the pion decay-channel. The selection criteria and corresponding efficiencies are stated in Table 5.6.



(a) Belle (b) Belle II

Figure 5.9.: p^*_{sig} distribution in GeV for the $\tau \to \pi\nu_\tau$ decay channel with a cut on $p^*_{\text{sig}} < 0.5\,\text{GeV}$. Evaluated on 10000 events for each component, scaled according to their expected relative fractions.

Table 5.6.: The selection criteria for the $\tau \to \pi\nu_\tau$ decay-channel with the corresponding selection efficiencies for signal candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$ and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. Evaluated on 10000 events for each component.

Selection Criterion	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
# Tracks ≤ 10	0.99	0.76	0.82	0.93	0.65
$p^*_\tau > 0.5\,\text{GeV}$	0.95	0.42	0.43	0.98	0.44
$\Delta E \leq 1.5\,\text{GeV}$	0.99	0.31	0.34	0.99	0.50
dr < 2	0.96	0.75	0.74	0.99	0.91
$-4 <$ dz < 4	0.93	0.68	0.66	0.97	0.83
$\text{PID}_\mu < 0.9$	0.99	0.99	0.99	0.96	0.93
$\text{PID}_\text{e} < 0.9$	1.00	0.96	0.96	0.99	0.84
$\text{PID}_\text{K} < 0.9$	0.97	0.88	0.87	0.96	0.84

Table 5.7.: The selection criteria for the $\tau \to \rho\nu_\tau$ decay-channel with the corresponding selection efficiencies for signal candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$ and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. Evaluated on 10000 events for each component.

Selection Criterion	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
# Tracks ≤ 10	0.99	0.81	0.82	0.92	0.60
$p_\tau^* > 0.6\,\text{GeV}$	0.96	0.45	0.46	0.95	0.51
$\Delta E \leq 1.5\,\text{GeV}$	1.00	0.35	0.37	1.00	0.58
dr < 2	0.93	0.74	0.72	0.99	0.91
$-4 <$ dz < 4	0.88	0.65	0.64	0.93	0.82
$\text{PID}_\mu < 0.9$	1.00	0.99	0.99	0.95	0.93
$\text{PID}_\text{e} < 0.9$	0.99	0.96	0.95	0.98	0.85
$\text{PID}_\text{K} < 0.9$	0.98	0.88	0.87	0.97	0.84
$0.5\,\text{GeV} < M_\rho < 1.2\,\text{GeV}$	0.88	0.56	0.56	0.95	0.62
$0.1\,\text{GeV} < M_{\pi^0} < 0.18\,\text{GeV}$	0.84	0.27	0.28	0.98	0.19

5.2.2.4. $\tau \to \rho\nu_\tau$

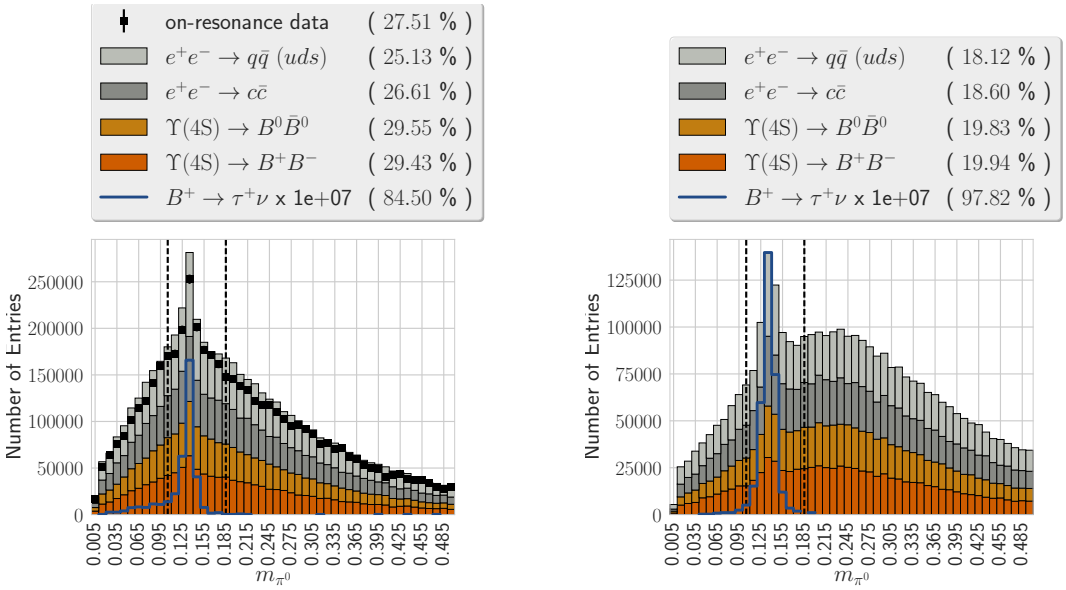
The τ decay into a ρ meson and one neutrinos has a branching fraction of 25.5%. It is the dominant decay channel of the τ with the largest branching fraction. The ρ resonance decays instantaneous into a charged and a neutral pion $\rho^+ \to \pi^+\pi^0[\to \gamma\gamma]$. The intermediate resonances ρ and π^0 provide efficient selection criteria based on their invariant masses

$$m_\rho = 0.775\,\text{GeV} \tag{5.12}$$
$$m_{\pi^0} = 0.135\,\text{GeV}. \tag{5.13}$$

We expect to find one track and two ECL clusters compatible with a π^0 (see Figure 5.10) and ρ (see Figure 5.11) in the event besides the B_{tag} meson. The selection criteria and corresponding efficiencies are stated in Table 5.7.

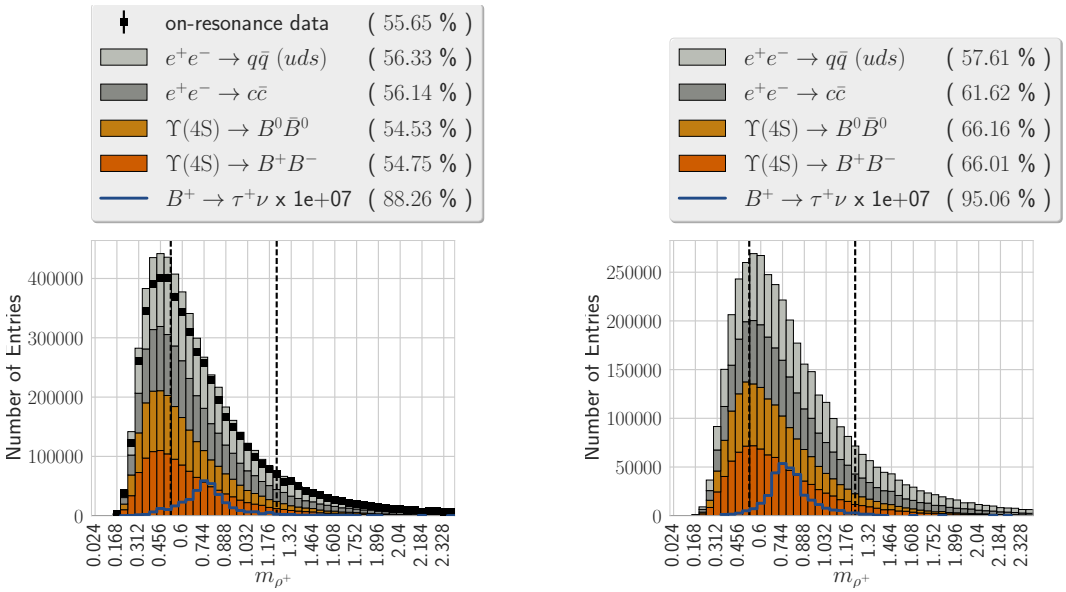
5.2.2.5. $\tau \to a_1\nu_\tau$

The τ decay into three pions $\tau \to 3\pi$ is dominated by the decay-chain $\tau^+ \to a_1^+[\to \rho^0\pi^+]\nu_\tau$ [88]. In this thesis, the τ decay into three charged pions (via $\rho^0 \to \pi^+\pi^-$) was investigated for the first time, it has a branching fraction of 9.5%. The other



(a) Belle (b) Belle II

Figure 5.10.: M_{π^0} distribution in GeV for the $\tau \to \rho\nu_\tau$ decay channel with a cut on $0.1\,\text{GeV} < M_{\pi^0} < 0.18\,\text{GeV}$. Evaluated on 10000 events for each component, scaled according to their expected relative fractions.



(a) Belle (b) Belle II

Figure 5.11.: M_ρ distribution in GeV for the $\tau \to \rho\nu_\tau$ decay channel with a cut on $0.5\,\text{GeV} < M_\rho < 1.2\,\text{GeV}$. Evaluated on 10000 events for each component, scaled according to their expected relative fractions.

Table 5.8.: The selection criteria for the $\tau \to a_1\nu_\tau$ decay-channel with the corresponding selection efficiencies for signal candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$ and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. The impact parameter selection criteria are the same for each of the three tracks. Evaluated on 10000 events for each component.

Selection Criterion	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
# Tracks ≤ 10	0.94	0.64	0.66	0.72	0.36
$p_\tau^* > 0.9\,\text{GeV}$	0.96	0.51	0.52	0.95	0.50
$\Delta E \leq 1.5\,\text{GeV}$	1.00	0.53	0.55	1.00	0.69
dr < 2	0.99	0.96	0.95	0.98	0.95
$-4 <$ dz < 4	0.97	0.90	0.88	0.95	0.86
$\chi^2_{\text{Vertex}} > 0.01$	0.79	0.53	0.50	0.80	0.41
$0.8\,\text{GeV} < M_{a_1} < 1.6\,\text{GeV}$	0.94	0.53	0.53	0.96	0.47

possible decay into a charged pion and two neutral pions (via $\rho^0 \to \pi^0\pi^0$) was not studied.

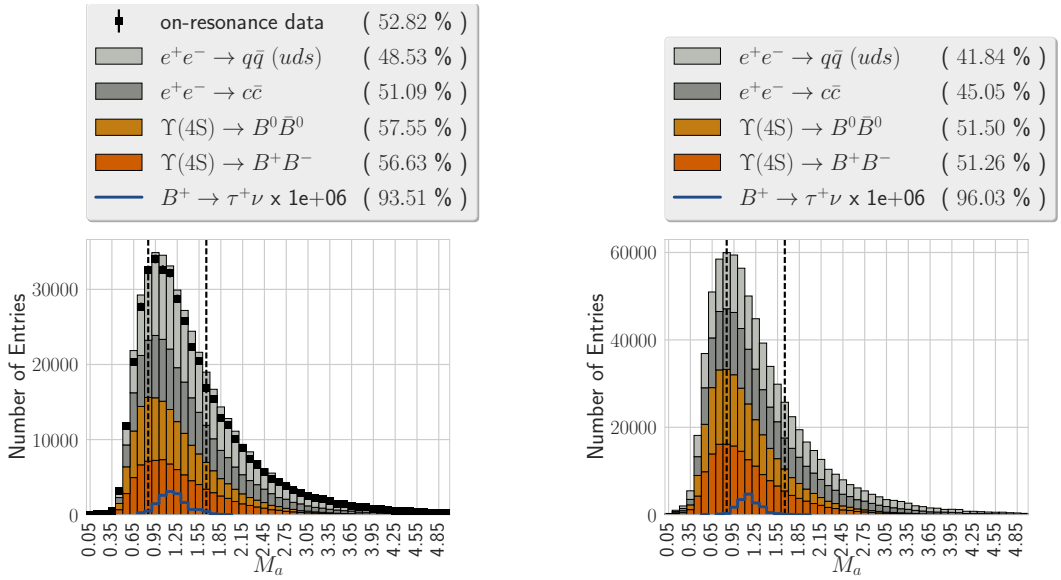
The resonances a_1 provides an efficient selection criteria based on its invariant mass (see Figure 5.12)

$$m_{a_1} = 1.230\,\text{GeV}. \tag{5.14}$$

We expect to find three tracks from a common vertex compatible with the expected intermediate resonances in the event besides the B_{tag} meson. This is the only signal decay-channel with a different multiplicity, that is number of tracks in the final state. Therefore it has distinct physics background contributions compared to all other studied decay-channels. Unfortunately, it also suffers from a large combinatorial background, which is the reason why it was not used in previous analyses. The selection criteria and corresponding efficiencies are stated in Table 5.8.

5.2.3. Tag side reconstruction

The default generic `FEI` (see Chapter 4) applied to the entire event was used to reconstruct the tag-side of the event. The `FEI` provides B meson tag-side candidates in **hadronic** and **semileptonic** decay channels. Both were used in this analysis. Because the different systematic uncertainties (in particular the missing calibration factors for the semileptonic tag), the tags are treated as independent measurements in this thesis, that is in principle the same event could appear in both tagged samples.



(a) Belle (b) Belle II

Figure 5.12.: $M_{\mathrm{a_1}}$ distribution in GeV for the $\tau \to \mathrm{a_1}\nu_\tau$ decay channel with a cut on $0.8\,\mathrm{GeV} < M_{\mathrm{a_1}} < 1.6\,\mathrm{GeV}$. Evaluated on 10000 events for each component, scaled according to their expected relative fractions.

The default `FEI` does not use information from the M_{bc} and $\cos\Theta_{\mathrm{BD}\ell}$ of the hadronically and semileptonically tagged B meson, respectively. In addition the `FEI` aims to always output a valid candidate, even if the associated `SignalProbability` σ is very low. However, this analysis requires a high-purity tag, because the signal-side selection alone does not provide enough separation power. Therefore the tag-side candidates have to fulfill the selection criteria stated in Table 5.9 to be considered in the analysis.

The same numerical value was chosen for the cut on the `SignalProbability` for Belle and Belle II. However, the cut is effectively tighter for Belle II, because the increased fake-rate reduces the prior-probability of each candidate to be correct.

5.2.4. Event reconstruction

The selected signal-side and tag-side candidates are combined to an $\Upsilon(4S)$ candidate.

5.2.4.1. Completeness-constraint(s)

The completeness-constraint is applied and all candidates with additional good tracks in the event (besides the ones, which were used for the reconstruction of the $\Upsilon(4S)$)

Table 5.9.: The selection criteria for the tagged B mesons with the corresponding selection efficiencies for correctly reconstructed candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$ and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. Evaluated on 10 million events for each component.

Selection Criterion	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
Hadronic Tag					
$M_{\text{bc}} \geq 5.27$	0.98	0.1205	0.1167	0.96	0.1374
$-0.15 < \Delta E < 0.1$	0.97	0.5928	0.5720	0.91	0.3475
$\sigma > 0.005$	0.94	0.1553	0.1469	0.81	0.0241
Total	0.91	0.0251	0.0242	0.76	0.0043
Semileptonic Tag					
$-1 < \cos\Theta_{\text{BD}\ell} < 1$	0.88	0.2926	0.2796	0.83	0.1993
$\sigma > 0.005$	0.90	0.1295	0.1262	0.72	0.0162
Total	0.80	0.0525	0.0496	0.63	0.0063

are discarded. The completeness-constraint is the most-effective selection criterion in this analysis. It has a very high signal efficiency, while it reduces the background by one order of magnitude in all signal-side (the five τ decay-channels) and tag-side (hadronic and semileptonic) combinations. The efficiencies are stated in Table 5.10. An example distribution for the $\tau \to \pi\nu_\tau$ is shown in Figure 5.13.

For Belle II, only good CDC tracks were considered for the completeness-constraint, due to the high-fake rate from SVD tracks. In consequence, the constraint discards less background compared to Belle.

The completeness-constraint is part of an entire class of constraints on additional physics objects in the event besides the reconstructed $\Upsilon(4S)$. Conveniently, the `FEI` provides candidates with an associated `SignalProbability` for many particle types. In particular, completeness-constraint like selections on the number of good[6] additional photons, K_S^0 and π^0 can be performed. The cut on additional photons is strongly correlated to the final fit variable E_{ECL}, hence only the cuts on the additional K_S^0 and π^0 are applied. The efficiencies are stated in Table 5.11 for π^0 and in Table 5.12 for K_S^0. An example distribution for the $\tau \to \pi\nu_\tau$ is shown in Figure 5.14 for π^0 and in Figure 5.15 for K_S^0.

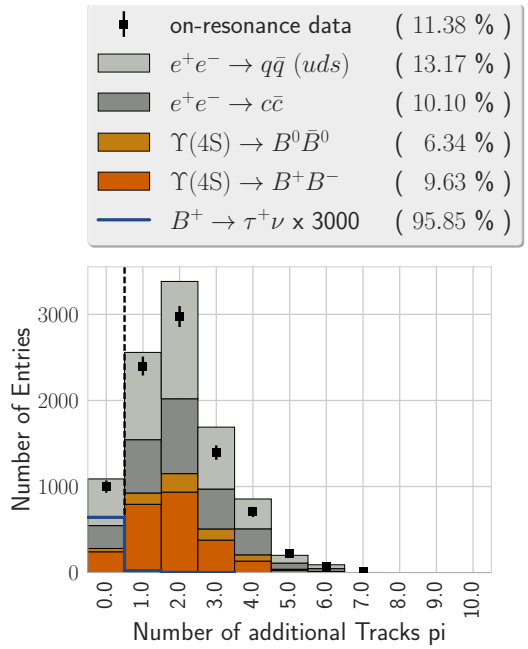
[6]Candidates with a `SignalProbability` larger than 0.1 are considered good.

Table 5.10.: Completeness-constraint on the number of additional good tracks. Stated are the selection efficiencies for: correctly reconstructed candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$, and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. Evaluated on 10 million events for each component.

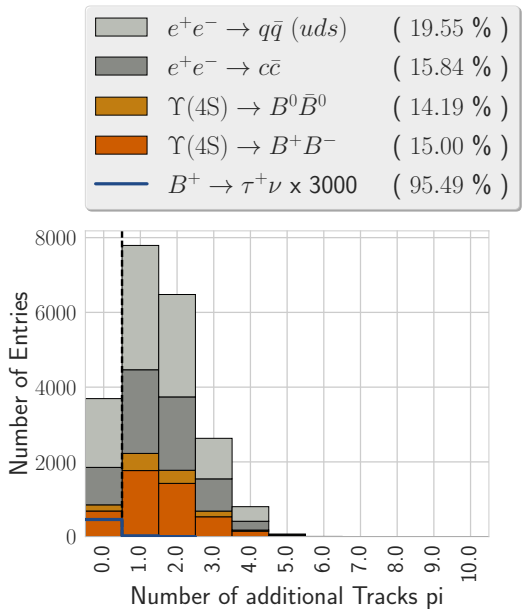
Decay-Channel	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
Hadronic Tag					
e^-	0.96	0.11	0.11	0.95	0.18
μ^-	0.96	0.12	0.13	0.96	0.17
π	0.96	0.11	0.11	0.95	0.17
ρ	0.95	0.12	0.14	0.95	0.18
a_1	0.96	0.17	0.17	0.94	0.28
Semileptonic Tag					
e^-	0.97	0.05	0.05	0.96	0.08
μ^-	0.97	0.05	0.06	0.96	0.09
π	0.97	0.03	0.03	0.96	0.08
ρ	0.95	0.03	0.03	0.95	0.08
a_1	0.95	0.05	0.05	0.94	0.14

Table 5.11.: Completeness-constraint on the number of additional π^0. Stated are the selection efficiencies for: correctly reconstructed candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$, and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. Evaluated on 10 million events for each component.

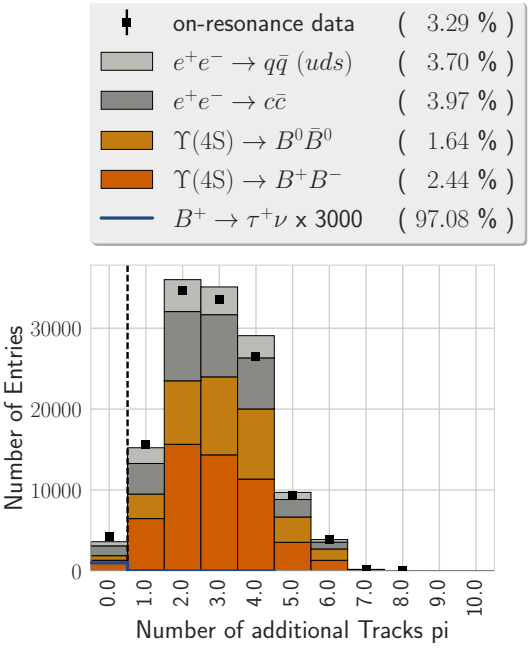
Decay-Channel	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
Hadronic Tag					
e^-	0.83	0.42	0.38	0.96	0.45
μ^-	0.85	0.39	0.42	0.96	0.45
π	0.82	0.36	0.34	0.96	0.41
ρ	0.86	0.33	0.32	0.96	0.40
a_1	0.79	0.33	0.34	0.95	0.39
Semileptonic Tag					
e^-	0.84	0.27	0.27	0.97	0.37
μ^-	0.87	0.28	0.27	0.98	0.38
π	0.84	0.21	0.21	0.96	0.34
ρ	0.88	0.19	0.18	0.97	0.31
a_1	0.79	0.18	0.18	0.96	0.30



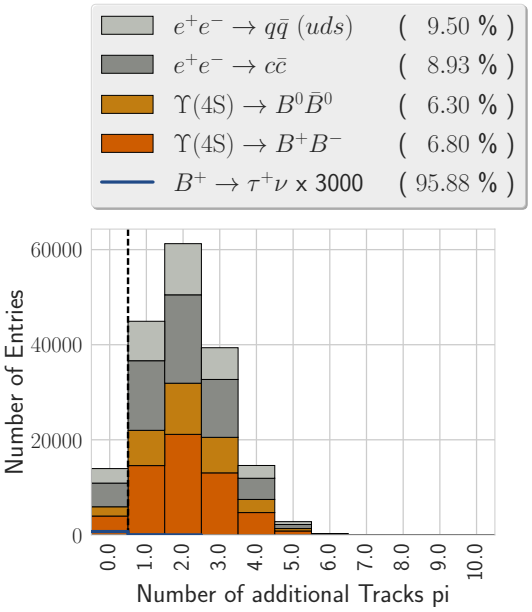
(a) Belle / Hadronic Tag



(b) Belle II / Hadronic Tag

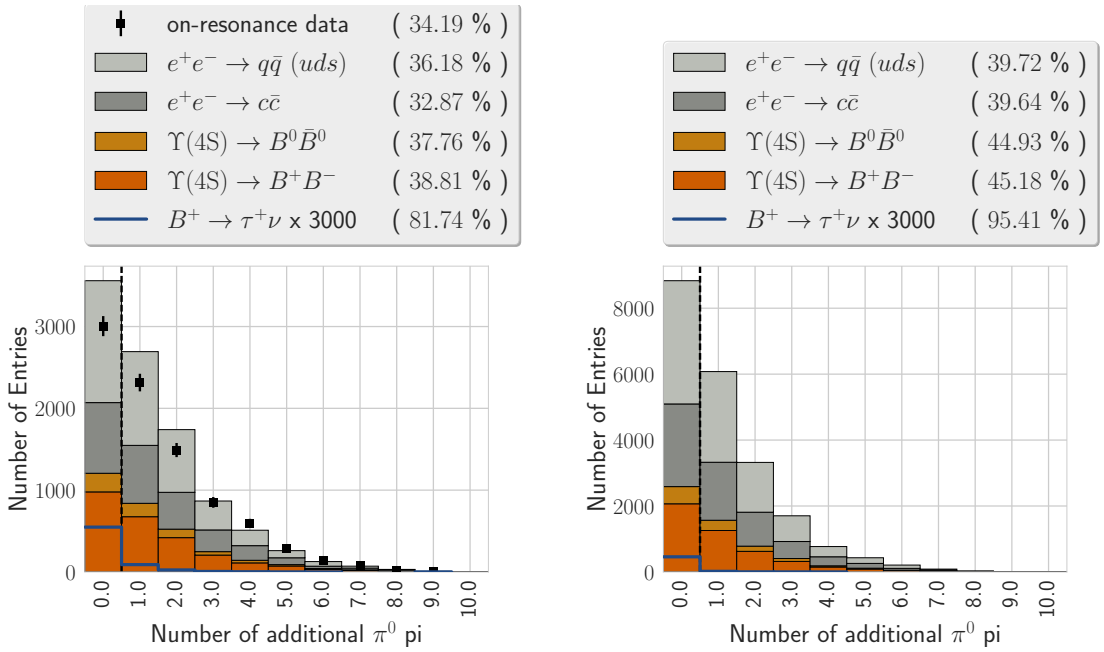


(c) Belle / Semileptonic Tag



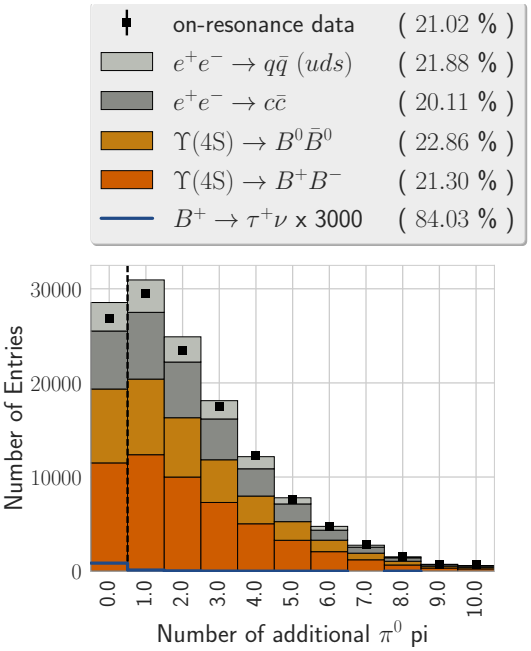
(d) Belle II / Semileptonic Tag

Figure 5.13.: Distribution of the number of additional good tracks in the event not used for the $\Upsilon(4S)$ reconstruction for the $\tau \to \pi\nu_\tau$ decay channel. Evaluated on 10 million events for each component, scaled according to their expected relative fractions.

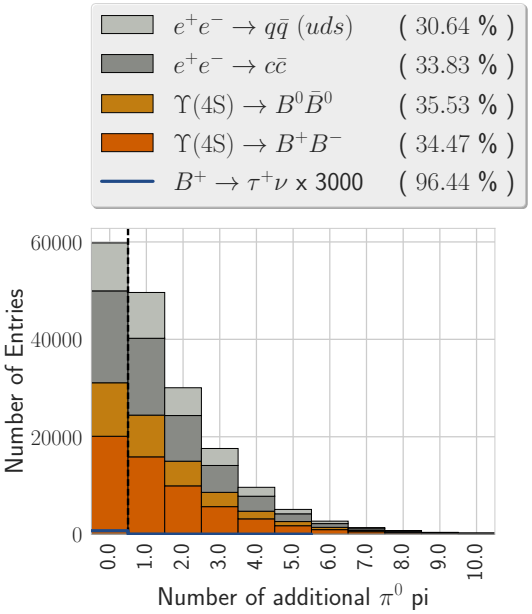


(a) Belle / Hadronic Tag

(b) Belle II / Hadronic Tag

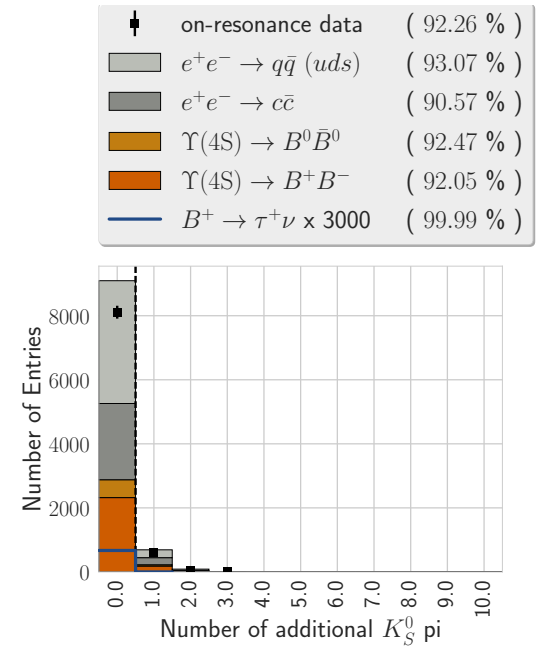


(c) Belle / Semileptonic Tag

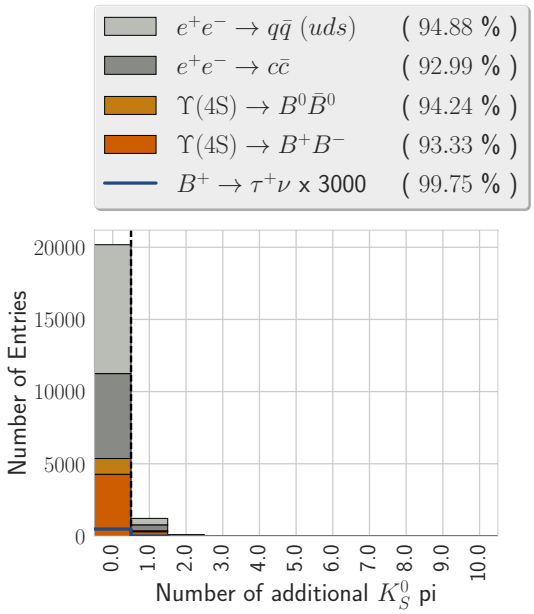


(d) Belle II / Semileptonic Tag

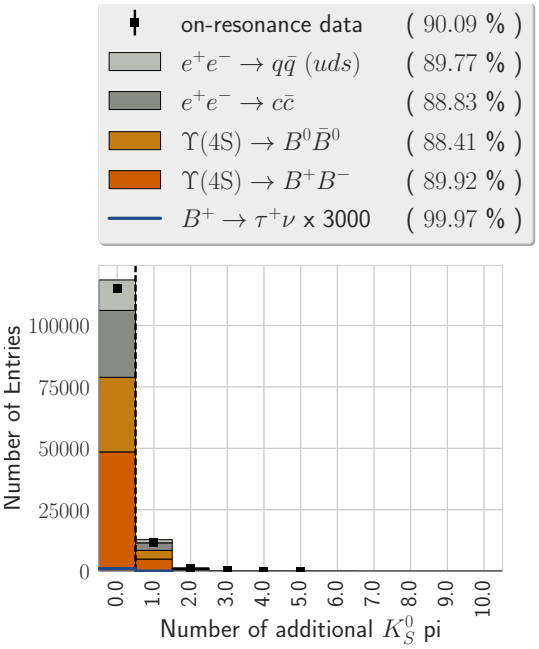
Figure 5.14.: Distribution of the number of additional good π^0 in the event not used for the $\Upsilon(4S)$ reconstruction for the $\tau \to \pi\nu_\tau$ decay channel. Evaluated on 10 million events for each component, scaled according to their expected relative fractions.



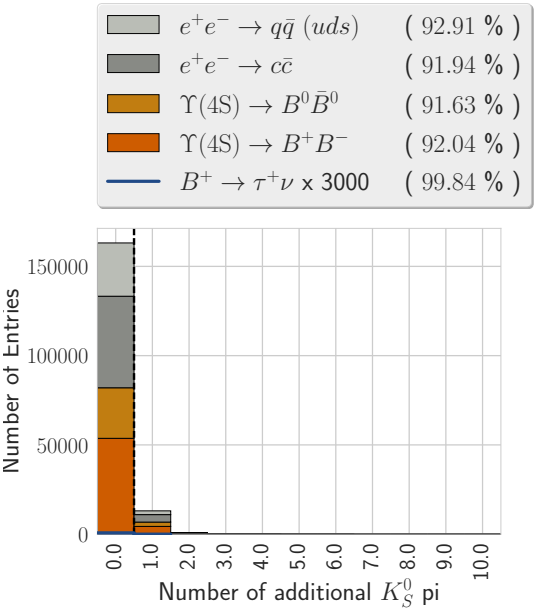
(a) Belle / Hadronic Tag



(b) Belle II / Hadronic Tag



(c) Belle / Semileptonic Tag



(d) Belle II / Semileptonic Tag

Figure 5.15.: Distribution of the number of additional good K^0_S in the event not used for the $\Upsilon(4S)$ reconstruction for the $\tau \to \pi\nu_\tau$ decay channel. Evaluated on 10 million events for each component, scaled according to their expected relative fractions.

Table 5.12.: Completeness-constraint on the number of additional K_S^0. Stated are the selection efficiencies for: correctly reconstructed candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$, and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. Evaluated on 10 million events for each component.

Decay-Channel	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
Hadronic Tag					
e^-	1.00	0.93	0.91	1.00	0.94
μ^-	1.00	0.93	0.91	1.00	0.93
π	1.00	0.92	0.92	1.00	0.94
ρ	1.00	0.93	0.93	1.00	0.94
a_1	1.00	0.94	0.33	1.00	0.96
Semileptonic Tag					
e^-	1.00	0.90	0.90	1.00	0.93
μ^-	1.00	0.90	0.92	1.00	0.92
π	1.00	0.89	0.90	1.00	0.92
ρ	1.00	0.90	0.90	1.00	0.92
a_1	1.00	0.92	0.92	1.00	0.94

5.2.4.2. Best-Candidate Selection

The chosen signal-side selection criteria and the completeness-constraint ensure that all signal-side candidates are (nearly[7]) mutual exclusive. Therefore a best-candidate selection solely based on the `SignalProbability` of the tag-side is performed. Hence, the same event can only contribute to one of the five investigated decay-modes. The efficiencies of the best-candidate selection are stated in Table 5.13.

5.2.4.3. Continuum Suppression

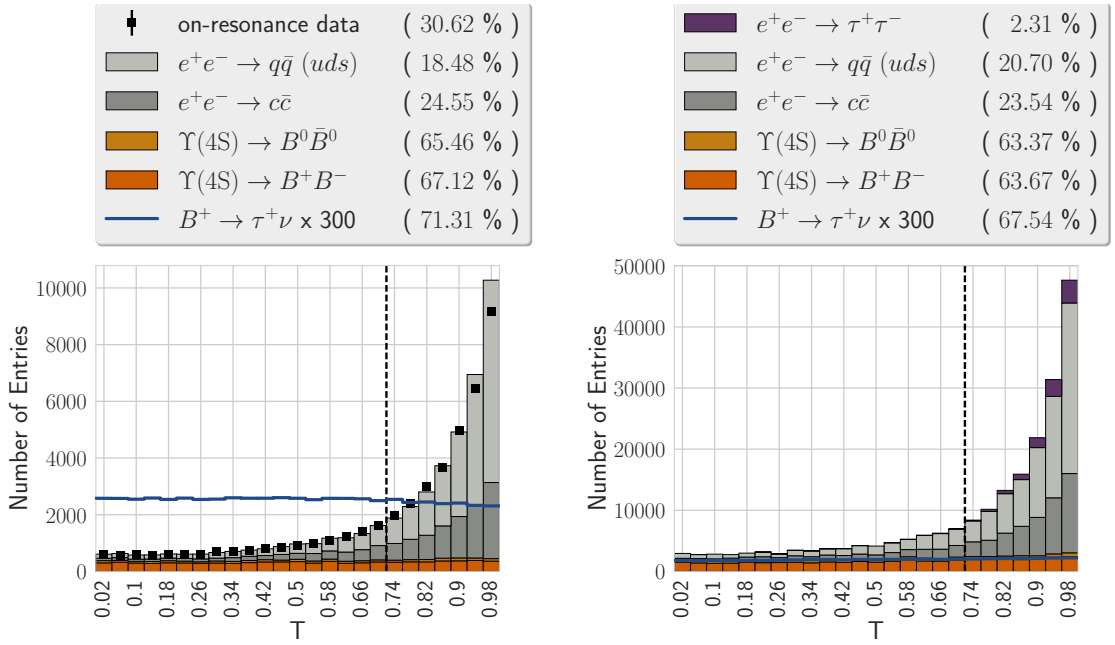
In previous and similar Belle analyses a continuum suppression algorithm based a multivariate method using event shape features was used to reduce the background from continuum processes (see Section 3.3.1.3).

In this benchmark analysis, I take a more conservative approach to prevent any systematic effects caused by an additional multivariate method besides the `FEI`.

[7] In rare cases a valid π and ρ candidate can be reconstructed in the same event, where the π^0 from the ρ is not considered a good π^0 from the `FEI`. In these cases the ρ candidate is discarded in favor of the π candidate.

Table 5.13.: Best-candidate selection efficiencies for correctly reconstructed candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$ and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. The efficiencies are calculated after all completeness-constraints were applied. The statistical power of the background and data sample was very limited, hence the stated efficiencies have large uncertainties ± 0.1. Evaluated on 10 million events for each component.

Decay-Channel	Belle			Belle II	
	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
Hadronic Tag					
e^-	0.86	1.0	1.0	0.95	1.0
μ^-	0.86	1.0	0.9	0.95	1.0
π	0.86	0.9	0.9	0.95	1.0
ρ	0.81	0.8	0.7	0.90	0.9
a_1	0.87	1.0	1.0	0.95	1.0
Semileptonic Tag					
e^-	0.97	1.0	1.0	0.97	1.0
μ^-	0.97	1.0	1.0	0.97	1.0
π	0.96	0.9	0.9	0.97	1.0
ρ	0.91	0.8	0.8	0.92	0.8
a_1	0.97	1.0	1.0	0.96	0.9



(a) **Belle:** The deviations between Monte Carlo simulation and data is discarded by a cut at 0.7.

(b) **Belle II:** The $e^-e^+ \to \tau^-\tau^+$ component, which is officially simulated in the Belle II Monte Carlo campaigns, is discarded by a cut at 0.7.

Figure 5.16.: Distribution of the angle between the thrust axes of the signal-side and tag-side for the hadronically tagged $\tau \to \pi\nu_\tau$ decay channel.

The continuum events are suppressed by a cut on a single well-understood variable

$$T = |\cos(\sphericalangle(T_{\text{tag}}, T_{\text{sig}}))|, \tag{5.15}$$

where T_{tag} and T_{sig} denote the thrust axis [1, Chapter 9.3] of the tag-side and signal-side, respectively. Correctly reconstructed $\Upsilon(4S)$ events have a flat distribution in this variable, because the two B meson decay spherically and independent of one another. Hence, the signal selection efficiency is uniform, and no additional systematic uncertainty is introduced. On the other hand, continuum events peak at large values of T, because of their jet-like decay topology. This theoretical assumption is verified (also for incorrectly reconstructed $\Upsilon(4S)$ decays) using Monte Carlo simulation by Figure 5.16.

5.2.4.4. Final Selection

A final selection optimized for each tag-side and signal-side combination separately is performed using: the `SignalProbability` σ of the tag-side, and the continuum suppression variable T.

In contrast to the previous selections, this final selection is performed during the **n-tuple analysis**. A multi-dimensional minimization algorithm optimizing the cuts by minimizing the expected negative $\frac{S}{S+B}$ was investigated. However, this approach was not used in the final analysis, because the optimum is located in a problematic phase-space region, with peaking background contributions from continuum processes not present in the simulated data (see Section 5.3). Instead, the cuts were chosen in order to avoid those problematic phase-space regions.

A loose cut on the `SignalProbability` $\sigma > 0.01$ was chosen for the hadronically tagged channels, whereas a tighter cut $\sigma > 0.05$ was chosen for the semileptonically tagged channels. This reflects the difference in the purity of the samples provided by the respective tags. The hadronic τ decay modes suffer from large continuum backgrounds, a tight cut $T < 0.7$ was chosen, to suppress those continuum events. In particular this selection discards the phase space region with the largest data/Monte Carlo disagreement. In contrast, the leptonic τ decay modes are expected to have only a small continuum contribution, nevertheless a loose cut $T < 0.85$ was applied to ensure that peaking contributions from unknown continuum processes are discarded (see Section 5.3.2.1).

The final selection efficiencies and the expected significance of a counting experiment of all events with an E_{ECL} below 100 MeV are stated in Table 5.14.

Previous analyses discarded in addition all events with additional non-good tracks. Although this selection is very effective and suppresses additional background it was not applied in this analyses for the following reasons. The non-good tracks are not well understood, they depend strongly on the beam-conditions and the reconstruction software (see Section 2.3.3.2). Furthermore the remaining off-resonance data sample after this selection does not allow a sound estimation of possible peaking background components. Previous analyses did not consider this potentially large systematic uncertainty (see [16]).

For Belle II, the strict requirement of no additional tracks $N = 0$ is very inefficient, due to the high-fake rate. Nevertheless, the cut is required to reduce the background to a manageable level. Compared to Belle, large parts of the continuum background remains, even in the usually clean hadronically tagged leptonic τ decay channels. It is expected that this will significantly improve as soon as the Belle II reconstruction is optimized. The current estimation does not provide much scientific value. In consequence, Belle II is not further considered in the following, except for the final sensitivity study in Section 5.4.2.

5.2.4.5. Self-Cross-Feed

All signal Monte Carlo events, which survive the final selection are considered as signal, regardless if they are reconstructed correctly or not. For instance, a signal event with

Table 5.14.: Final cuts on T and σ. Stated are the selection efficiencies for; correctly reconstructed candidates $\varepsilon_{\text{Signal}}$, background candidates $\varepsilon_{\text{Background}}$, and (in the case of Belle) for candidates on data $\varepsilon_{\text{Data}}$. $\frac{S}{\sqrt{B}}$ refers to the expected significance of a counting experiment of all events (number of signal events S and number of background events B) below 100 MeV.

Decay-Channel	Belle				Belle II		
	$\frac{S}{\sqrt{B}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$	$\varepsilon_{\text{Data}}$	$\frac{S}{\sqrt{B}}$	$\varepsilon_{\text{Signal}}$	$\varepsilon_{\text{Background}}$
Hadronic Tag							
e^-	2.40	0.73	0.39	0.39	0.5	0.80	0.45
μ^-	1.77	0.73	0.40	0.40	0.7	0.81	0.48
π	2.39	0.59	0.13	0.13	0.6	0.64	0.17
ρ	1.16	0.56	0.11	0.11	0.4	0.62	0.16
a_1	0.37	0.57	0.11	0.11	0.2	0.65	0.16
Semileptonic Tag							
e^-	2.12	0.35	0.10	0.09	0.6	0.38	0.13
μ^-	1.49	0.36	0.10	0.10	0.7	0.40	0.14
π	2.23	0.28	0.04	0.04	0.6	0.30	0.06
ρ	1.43	0.24	0.03	0.03	0.5	0.28	0.05
a_1	0.40	0.24	0.04	0.04	0.2	0.29	0.05

Table 5.15.: **Belle:** Expected number of events from the Monte Carlo simulation and the obtained number on on-resonance and off-resonance data. The uncertainty of the stated numbers is below 2%.

	Hadronic					Semileptonic				
	e	μ	π	ρ	a_1	e	μ	π	ρ	a_1
signal	80	52	124	79	21	79	48	115	89	23
background	4428	4056	6154	12244	6915	3961	3382	4223	8032	5135
total	4508	4109	6278	12323	6937	4041	3430	4339	8122	5158
on-resonance	4079	3545	6019	11572	5542	4131	3421	5070	9109	5050
continuum	264	180	2625	5481	2685	61	43	1089	1604	720
off-resonance	222	175	2306	5105	1702	135	103	1574	2059	707

Table 5.16.: **Belle:** The self-cross-feed of the reconstructed signal events. The rows contain the decay-channels in which the events where reconstructed. The upper five rows contain the hadronically tagged, and the lower five rows the semileptonically tagged reconstructed decay-channels. The columns indicate the Monte Carlo truth. Each cell contains the fraction of the events in the reconstructed decay-channel (row), which originate from the Monte Carlo decay-channel (column). The label *other* refers to the remaining τ decay-channels, which are not considered in this analysis. The uncertainty due to limited statistics is approximately ± 0.01.

	Hadronic						Semileptonic					
	e	μ	π	ρ	a_1	other	e	μ	π	ρ	a_1	other
e	0.91	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00
μ	0.00	0.93	0.01	0.02	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
π	0.04	0.19	0.20	0.27	0.04	0.25	0.00	0.00	0.00	0.00	0.00	0.00
ρ	0.01	0.02	0.03	0.68	0.18	0.08	0.00	0.00	0.00	0.00	0.00	0.00
a_1	0.00	0.00	0.00	0.01	0.82	0.17	0.00	0.00	0.00	0.00	0.00	0.00
e	0.05	0.00	0.00	0.00	0.00	0.01	0.86	0.01	0.00	0.00	0.00	0.08
μ	0.00	0.06	0.00	0.00	0.00	0.00	0.01	0.87	0.01	0.02	0.00	0.03
π	0.00	0.01	0.01	0.01	0.00	0.01	0.04	0.17	0.20	0.24	0.04	0.24
ρ	0.00	0.01	0.00	0.04	0.01	0.01	0.01	0.02	0.03	0.64	0.16	0.07
a_1	0.01	0.01	0.00	0.00	0.05	0.01	0.00	0.00	0.00	0.01	0.76	0.15

a decay chain B $\to \tau[\to e\nu\bar{\nu}]\nu$ might be reconstructed as B $\to \tau[\to \pi\nu_\tau]\overline{\nu}_\tau$, where the electron was mis-identified as a pion. Hence, the shape of the signal component contains two contributions: correctly reconstructed signal events (that is the signal event is reconstructed in the correct decay channel) and incorrectly reconstructed signal events (that is the signal event is reconstructed in a different decay channel). The last contribution is referred to as **self-cross-feed**.

Table 5.16 states the fraction of the events reconstructed in a decay-channel, which originates from a Monte Carlo decay-channel. There is nearly no self-cross-feed between the hadronic and semileptonic tag. This means, signal events, whose tag-side decayed semileptonically are not reconstructed by the hadronic tag. On the other hand, signal events, whose tag-side decayed hadronically are rarely reconstructed by the semileptonic tag.

The largest self-cross-feed can be found in the $\tau \to \pi\nu_\tau$ decay-channel. Leptons which fail to pass the PID criteria are collected by this channel, as well as $\tau \to \rho\nu_\tau$ decays, where the π^0 is missed. Furthermore, around 25% of the events reconstructed in the pion channel originate from τ decay channels which are not explicitly considered in this analysis. In consequence, the systematic uncertainty of the π channel is potentially large, and is influenced by the τ branching fractions, PID selection efficiencies, and the π^0 reconstruction efficiency.

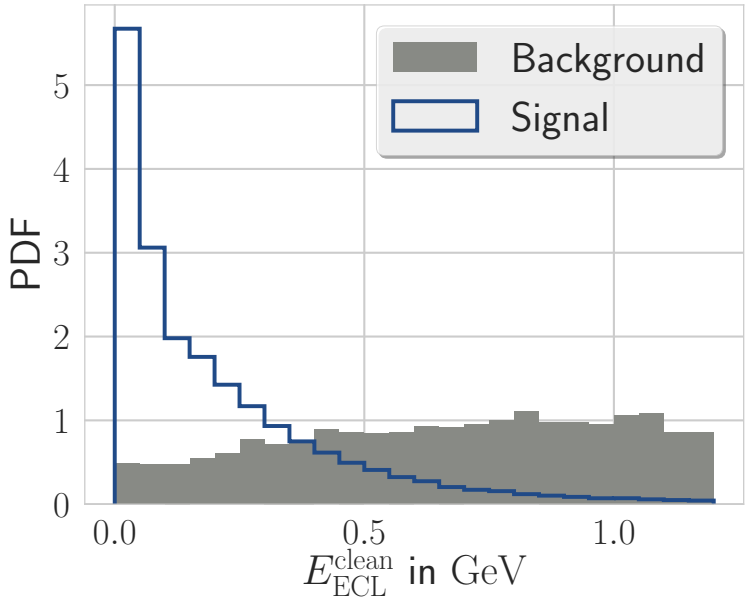


Figure 5.17.: Normed distribution of E_{ECL} for the $\tau \to \mathrm{e}\nu\bar{\nu}$ decay channel with hadronic tag.

5.2.4.6. Extra energy in the electromagnetic calorimeter

The sum of all depositions in the electromagnetic calorimeter, which (according to the reconstruction software) do not originate from the reconstructed $\Upsilon(4\mathrm{S})$ candidate, is called the raw extra energy in the electromagnetic calorimeter $E_{\mathrm{ECL}}^{\mathrm{raw}}$.

In order to reduce the dependency on the varying beam-background conditions between the experiments and runs (see Section 2.3.3.2), analyses using the extra energy usually consider only ECL clusters for the calculation of E_{ECL}, which fulfill additional requirements.

In this thesis, only clusters are considered without an associated track and with an energy greater than 50, 100 or 150 MeV if they are located in the barrel, forward end-cap or backward end-cap, respectively. The resulting quantity is the cleaned extra energy in the electromagnetic calorimeter $E_{\mathrm{ECL}}^{\mathrm{clean}}$.

The quantity is shown in Figure 5.17 for the $\tau \to \mathrm{e}\nu\bar{\nu}$ decay channel with hadronic tag. The signal component peaks at zero, while the background is slowly rising towards higher values. In consequence, $E_{\mathrm{ECL}}^{\mathrm{clean}}$ can be used as a fit variable, to determine the branching fraction of B $\to \tau\nu_\tau$.

Previous analysis considered more advanced cluster cleanings, e.g. [89] vetoed clusters, which were located close to a reconstructed track if certain criteria based on the ECL

cluster shape were fulfilled. The E_{ECL} variable was extensively studied in [90]. It is used in many analyses and its distribution for the signal component was validated on data using double-tagged events [91]. The distribution for background components can be validated on off-resonance data and sidebands (see Section 5.3).

For Belle II, a cluster cleaning based on the ECL cluster shapes and cluster timing will be important to mitigate the increased beam-background. However, at the time of writing the necessary reconstruction algorithms are not available in the official Belle II Monte Carlo campaign.

For the sake of clarity, I drop the index clean in the following, hence $E_{\text{ECL}} \equiv E_{\text{ECL}}^{\text{clean}}$.

5.2.5. Branching Fraction Extraction

The absolute branching fraction of $B \to \tau\nu_\tau$ is determined by extracting the number of events containing this decay $N_{\text{sig},d}$ in each decay-channel d with an extended unbinned maximum likelihood fit on the extra energy in the electromagnetic calorimeter E_{ECL} (see Section 5.2.4.6)

$$\mathcal{BR}(B \to \tau\nu_\tau) = \frac{N_{\text{sig},d}}{N_{B\overline{B}} \cdot \varepsilon_d} \tag{5.16}$$

where $N_{B\overline{B}} = (7.72 \pm 0.1) \cdot 10^8$ is the number of produced $B\overline{B}$ events[8] determined by the Belle experiment, and ε_d is the overall reconstruction efficiency in the decay-channel (see Table 5.18).

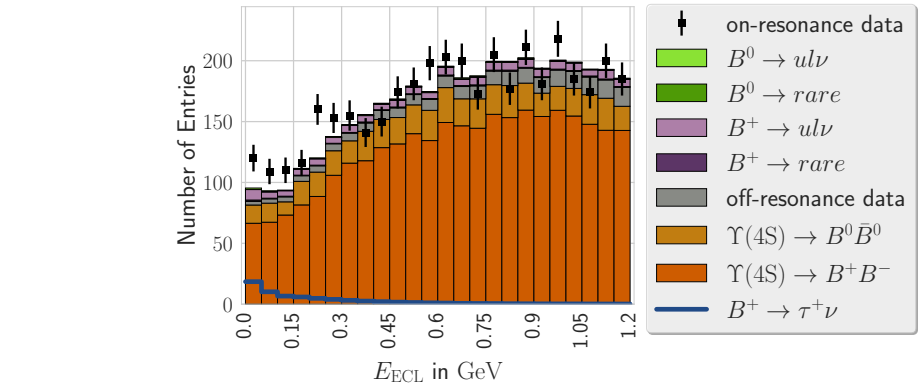
Figure 5.18 and Figure 5.19 show the distribution of E_{ECL} for all components used in the fit. The continuum component was estimated with a linear regression using the off-resonance sample.

The overall PDF is defined as

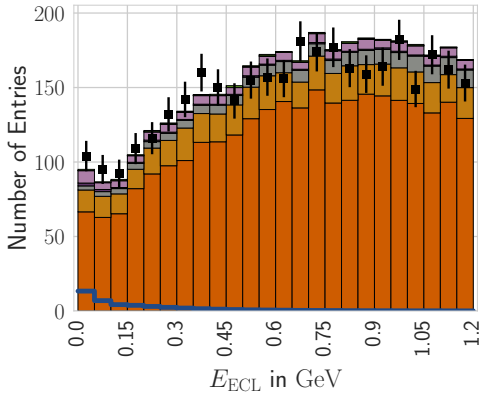
$$P(E_{\text{ECL}}) = \sum_{d=e,\mu,\pi,\rho,a_1} \left(\varepsilon_d N_{B\overline{B}} \mathcal{BR}(B \to \tau\nu_\tau) P_{\text{sig},d}(E_{\text{ECL}}) + N_{\text{bkg},d} P_{\text{bkg},d}(E_{\text{ECL}})\right), \tag{5.17}$$

where d denotes the reconstructed decay-channel, and $P_{\text{sig},d}$ and $P_{\text{bkg},d}$ the PDFs obtained from Monte Carlo simulation and off-resonance data for signal and background, respectively. There are six free parameters in the fit: the branching fraction $\mathcal{BR}(B \to \tau\nu_\tau)$ and the five background normalizations $N_{\text{bkg},d}$. The relative fractions of the different background components was fixed. The fit is performed independently for the hadronic and semileptonic tag.

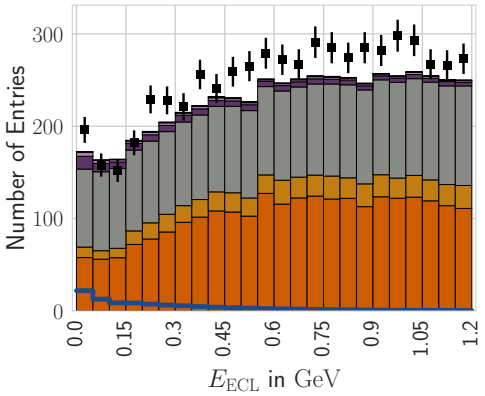
[8] The Belle analyses assumed equal branching fractions for charged and neutral B mesons. Hence the approximation $N_{B\overline{B}} \approx 2N_{B^+B^-}$ is used here. The correction is small compared to the uncertainties of the measurements and therefore neglected.



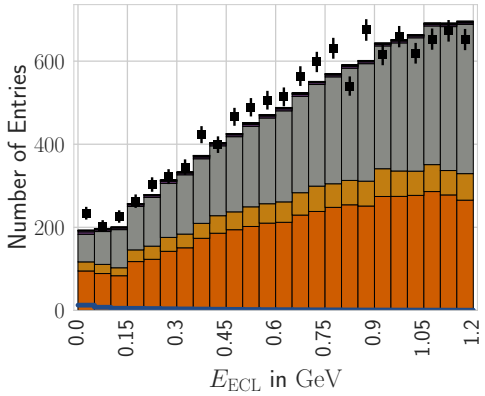
(a) $\tau \to \mathrm{e}^-\nu_\mathrm{e}\nu_\tau$



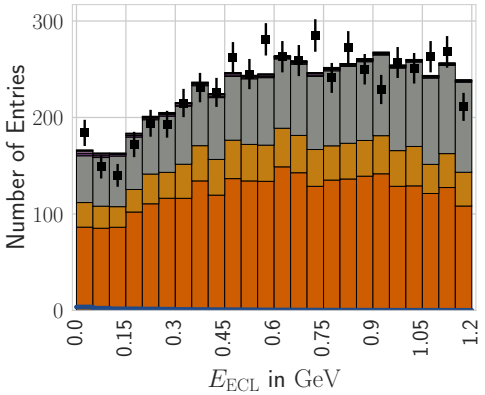
(b) $\tau \to \mu^-\nu_\mu\nu_\tau$



(c) $\tau \to \pi^+\nu_\tau$

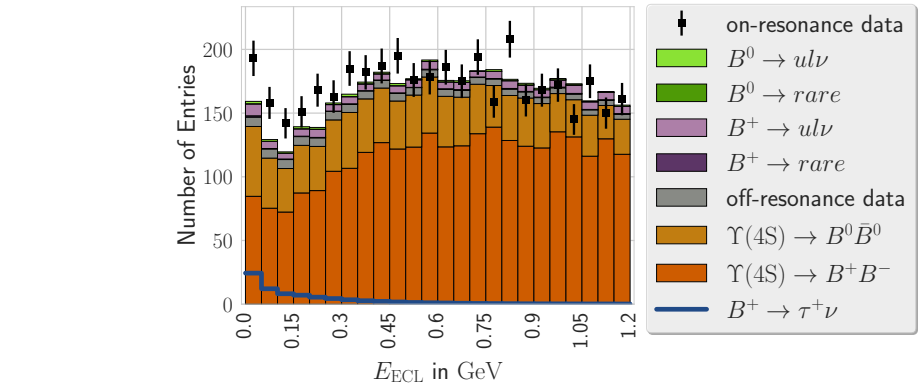


(d) $\tau \to \rho\nu_\tau$

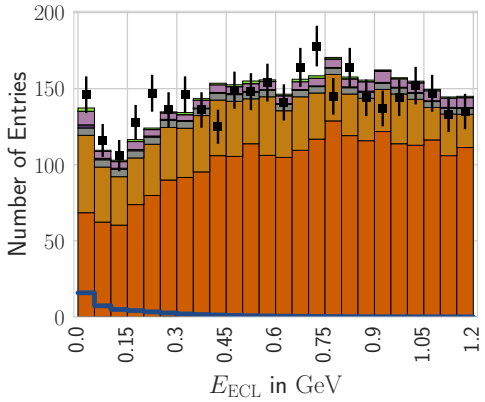


(e) $\tau \to \mathrm{a}_1\nu_\tau$

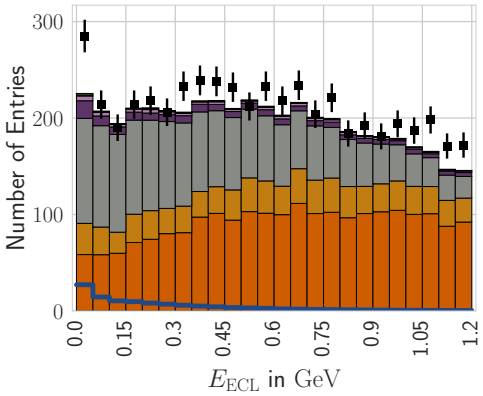
Figure 5.18.: Distribution of E_{ECL} for the hadronically tagged candidates. The stacked histograms show the expectation from Monte Carlo simulation and off-resonance data. The continuum component was estimated with a linear regression using the off-resonance sample.



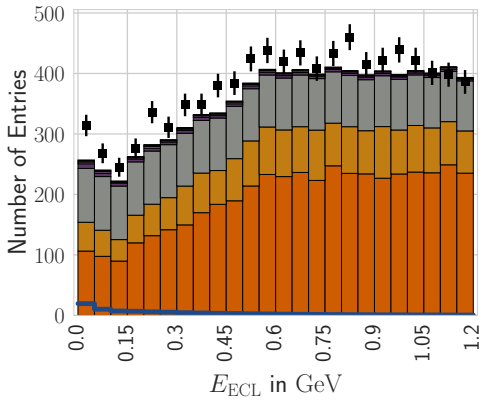
(a) $\tau \to \mathrm{e}^-\nu_\mathrm{e}\nu_\tau$



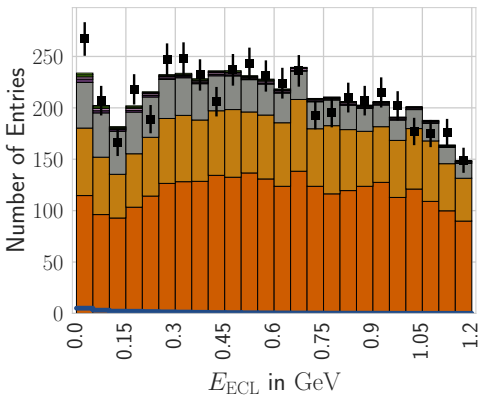
(b) $\tau \to \mu^-\nu_\mu\nu_\tau$



(c) $\tau \to \pi^+\nu_\tau$



(d) $\tau \to \rho\nu_\tau$



(e) $\tau \to \mathrm{a}_1\nu_\tau$

Figure 5.19.: Distribution of E_{ECL} for the semileptonically tagged candidates. The stacked histograms show the expectation from Monte Carlo simulation and off-resonance data. The continuum component was estimated with a linear regression using the off-resonance sample.

5.2.5.1. Component description

The extraction of the branching fraction from E_{ECL} relies on the precise knowledge of the shapes of all background processes. The following background processes were considered during the fit. Each component is described by a histogram PDF using 24 equidistant bins in the region between 0 GeV and 1.2 GeV.

$\Upsilon(4S) \to B^+B^-$ is background from charged $B^{\pm}$ decays. Also called generic charged $\Upsilon(4S)$ background. It is the main background component in all decay-channels. In total six (out of ten) streams of Monte Carlo were used to determine the shape of this component.

$\Upsilon(4S) \to B^0\overline{B}^0$ is background from neutral B^0 decays. Also called generic neutral (or mixed) $\Upsilon(4S)$ background. This type of background is suppressed by the tag-side requirements. In total six (out of ten) streams of Monte Carlo were used to determine the shape of this component.

$e^+e^- \to c\bar{c}$ is background from charmed quarks. Also called charm continuum background. Is is the main continuum background in semileptonically tagged decay-channels. This component is estimated from off-resonance data in the final fit. For the alternative shape description based on the Monte Carlo simulation in total six (out of six) streams of Monte Carlo were used.

$e^+e^- \to q\bar{q}$ (uds) is background from light quarks. Also called uds continuum background. It is the main continuum background in hadronically tagged decay-channels. This component is estimated from off-resonance data in the final fit. For the alternative shape description based on the Monte Carlo simulation in total six (out of six) streams of Monte Carlo were used.

$B^+ \to$ rare is background from rare $B^{\pm}$ decays such as $B^+ \to \ell^+\nu\gamma$. Although rare decays have an extremely low branching fraction, their final state can be nearly indistinguishable from $B^+ \to \tau^+\nu_\tau$. In fact, $B^+ \to \tau^+\nu_\tau$ is contained in this component as well, but was vetoed using Monte Carlo information. In total fifty (out of fifty) streams of Monte Carlo were used to determine the shape of this component.

$B^0 \to$ rare is background from rare B^0 decays such as $B^0 \to K\nu\nu$. This type of background is suppressed by the tag-side requirements compared to the charged rare decays. In total fifty (out of fifty) streams of Monte Carlo were used to determine the shape of this component.

$B^+ \to \bar{u}\ell^+\nu$ is background from CKM-suppressed charged $b \to u$ transitions. If the decay products from the u are lost during the reconstruction, this component can fake a signal component. In total twenty (out of twenty) streams of Monte Carlo were used to determine the shape of this component.

$B^0 \to \bar{u}\ell^+\nu$ is background from CKM-suppressed neutral b → u transitions. This type of background is suppressed by the tag-side requirements compared to the charged b → u transitions. In total twenty (out of twenty) streams of Monte Carlo were used to determine the shape of this component.

$B^+ \to \tau^+\nu_\tau$ In total ≈ 1200 streams of Monte Carlo were used to determine the shape of this component.

off-resonance data is continuum background recorded 60 MeV below the $\Upsilon(4S)$ resonance. It can be used to estimate the complete continuum background without using Monte Carlo simulation. This includes the above mentioned continuum background from quarks, QED processes and other (even unknown) continuum processes. In total ≈ 0.08 stream of off-resonance data were available and used to determine the shape of the continuum component.

The usually used Belle Monte Carlo campaign did not include QED processes. However, there are Monte Carlo simulated QED events available for Belle, those could not be used during this thesis, due to technical issues. Previous Belle analysis did not use this type of Monte Carlo. Nevertheless, this thesis studied the QED processes using the corresponding Belle II Monte Carlo simulated event for $e^-e^+ \to \gamma\gamma$, $e^-e^+ \to e^-e^+$, $e^-e^+ \to \mu^-\mu^+$, and $e^-e^+ \to \tau^-\tau^+$. The $\tau^-\tau^+$ component yields a small number of additional continuum events, which are suppressed by the continuum suppression selection (see Figure 5.16). No candidates from other QED processes were found to pass the selection criteria of this analysis. However, differences between the Monte Carlo simulation of the continuum component and the off-resonance data are observed (see Section 5.3). In consequence, the off-resonance data was used to estimate the continuum background shape.

5.2.5.2. Statistical uncertainty

The statistical uncertainty of the fit is estimated from the likelihood profile with respect to the branching fraction $\mathcal{BR}(B \to \tau\nu_\tau)$. The significance of the fit is determined by Wilk's theorem (see [89, Chapter 7.7]) using the likelihood ratio of the maximum likelihood obtained from the fit $L_{\max}$ and the likelihood under the null-hypothesis L_0 assuming a branching fraction of zero

$$\sigma = \sqrt{2 \ln \frac{L_{\max}}{L_0}}. \tag{5.18}$$

5.3. Validation

Before the final fit on on-resonance data was performed, the distribution of the extra energy in the calorimeter in the different tag-side and signal-side combinations was

validated. The background processes were studied using Monte Carlos simulation (Section 5.3.1) and data-driven methods (Section 5.3.2). The fit procedure was validated (Section 5.3.4). And finally possible systematic uncertainties of the analyses were investigated (Section 5.3.5).

The following sections summarize the findings of this validation. Tag-side and signal-side combinations which are not explicitly mentioned are in agreement with the expectation.

5.3.1. Monte Carlo based study

Monte Carlo simulated events were used to investigate the known background processes. In total six streams were investigated. An overview of the Monte Carlo matching error flags obtained from this sample on the tag-side and signal-side can be found in Section D.2.1.

5.3.1.1. Leptonic τ decay channels

The leptonic τ decay channels were investigated using Monte Carlo simulated events.

The main background originates from semileptonic $B^{\pm}$ decays accompanied by a correctly reconstructed B_{tag} decay: $\Upsilon(4S) \to B_{tag}B^{-}[\to D^{0}\ell^{-}\bar{\nu}]$. The additional D^0 meson decays into neutral final state particles like $D^0 \to K^0_L K^0_L(K^0_L)$ or $D^0 \to K^0_L \pi^0$.

The contributing decay channels from neutral $B^0\overline{B}^0$ pairs are very similar $\Upsilon(4S) \to B_{tag}B^0[\to D^{-}\ell^{+}\nu]$. The D^- meson decays further into neutral final state particles and an additional charged track, which is assigned to the tag-side.

The $B^+ \to u\ell^+\nu$ component is dominated by the decay $B^+ \to \pi^0\ell^+\nu$, where the π^0 is missed during the reconstruction.

The contribution from rare decays like $B^+ \to \ell^+\nu\gamma$ is negligible small assuming either the branching fractions predicted by the Standard Model or the current experimental limits.

The background from continuum processes described by the Monte Carlo simulation is very small as well. Here the lepton is usually produced via a semileptonic D decay, whereas the tag-side is reconstructed from the remaining charged tracks.

5.3.1.2. Hadronic τ decay channels

The hadronic τ decay channels were investigated using Monte Carlo simulated events.

The background from charged B^+B^- pairs is diverse. The main contribution originates from decay-channels with a high branching fraction and accompanied by a correctly reconstructed tag-side. For instance, $\Upsilon(4S) \to B_{tag}B^-[\to D^0\pi^-]$, where the D meson decays further into neutral final state particles like $D^0 \to K^0_L K^0_L(K^0_L)$ or $D^0 \to K^0_L\pi^0$. Another common example is a semileptonic decay-chain $\Upsilon(4S) \to B_{tag}B^-[\to D^0[\to K^-\ell^+\nu]\ell^-\bar{\nu}]$, where the K meson is mis-identified as a pion.

The contributions from neutral $B^0\overline{B}^0$ pairs are even more diverse, since both B mesons are wrongly reconstructed. There is no noticeable pattern, except for the obvious influence of the branching fractions.

The $B^+ \to u\ell^+\nu$ component is dominated by the decay $B^+ \to \pi^0\ell^+\nu$, where the π^0 is missed during the reconstruction and the lepton is mis-identified as a pion. However, due to the additional mis-identification this component is negligible, in contrast to the leptonic decay-channels.

On the other hand, the contribution from rare decays like $B \to K^0_L\pi$ is small.

The background from continuum processes described by the Monte Carlo simulation is as large as the background from $\Upsilon(4S)$ decays. The semileptonically tagged continuum background events are dominated by $e^-e^+ \to c\bar{c}$, because the tag-side requires a lepton, which is rarely produced by light quark pairs: $u\bar{u}$, $d\bar{d}$ and $s\bar{s}$. Hence, light quark pairs are suppressed by the mis-identification rate of leptons.

On the other hand, the hadronically tagged continuum background events have a large contribution from light quarks pairs, because no reconstructed lepton is required. This contribution is as large as the one from $e^-e^+ \to c\bar{c}$. This can be explained by the fact, that the D meson on the tag-side is usually mis-reconstructed, i.e. not matched to a D meson Monte Carlo particle. Probably, true D mesons are suppressed due to their different kinematics compared to the D mesons typically found in $\Upsilon(4S)$ decays. In consequence $c\bar{c}$ pairs and light quark pairs contribute in equal parts to the continuum background in hadronically tagged hadronic τ decay channels.

5.3.2. Off-resonance based study

Off-resonance data was used to validate the shape of the continuum component and to identify unknown continuum background processes. The off-resonance was scaled to account for the difference in luminosity recorded on-resonance and off-resonance, and the shifted center-of-mass energy.

5.3.2.1. Peaking continuum background in semileptonically tagged $\tau \to e\nu\bar{\nu}$

A large peaking background component in the semileptonically tagged $\tau \to e\nu\bar{\nu}$ decay channel was observed on off-resonance data by [89]. This peaking background is not described by the Monte Carlo simulation. It was suspected to be caused by $e^-e^+ \to \gamma[\to D\overline{D}]\gamma[\to e^-e^+]$, which is not part of the standard Belle Monte Carlo. A selection on the minimal invariant mass $M_{\text{sig}} > 0.2\,\text{GeV}$ of the signal-side electron with all other tracks in the event, and on the minimal invariant mass $M_{\text{tag}} > 0.2\,\text{GeV}$ of the tag-side lepton with all other tracks in the event was introduced by [89] to suppress this background. This selection is called invariant-mass-selection in the following.

This analysis observed the same peaking background on on-resonance and off-resonance data. However, the chosen cuts on the `SignalProbability` $\sigma > 0.05$ of the B_{tag} and on the continuum suppression variable $T < 0.85$ (see Section 5.2.4.3), discard already all events, which would be discarded by the invariant-mass-selection. This selection is called final-semileptonic-electron-channel-selection in the following.

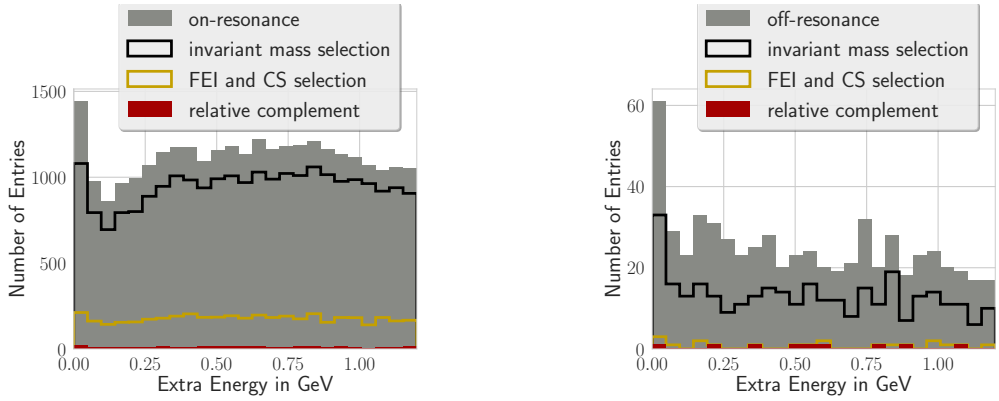
Figure 5.20 shows the effect of the two selection sets and their relative complement, that is the events which survive the final-semileptonic-electron-channel-selection, but not the invariant-mass-selection. Since the relative complement is negligible small, an additional invariant-mass-selection besides the final-semileptonic-electron-channel-selection does not have any effect. Hence, the invariant-mass-selection was not applied in this thesis.

5.3.2.2. Peaking continuum background in hadronic τ decay channels

A large peaking background component in the hadronic τ decay channels is observed on on-resonance and off-resonance data before the final selection on the `SignalProbability` σ and T are applied. This peaking background is not described by the Monte Carlo simulation.

Off-resonance data was used to verify that the final selection $\sigma > 0.01(0.05)$ for hadronic (semileptonic) B_{tag} and $T < 0.7$ discards this peaking background component. From this one can conclude, that the observed peak is caused by an unknown continuum component, and that the component is heavily suppressed by the final selection. The on-resonance and off-resonance distributions of E_{ECL} can be found in Section D.2.2.

Note that this peaking background is different from the one observed in the semileptonically tagged $\tau \to e\nu\bar{\nu}$ decay channel. The invariant-mass-selection introduced in Section 5.3.2.1 (where the leptons are replaced by the corresponding hadrons) does not suppress this peaking component.



(a) on-resonance (b) off-resonance

Figure 5.20.: Distribution of E_{ECL} for the $\tau \to \mathrm{e}\nu\bar{\nu}$ decay channel. A cut on $M_{\mathrm{sig}} > 0.2\,\mathrm{GeV}$ and $M_{\mathrm{tag}} > 0.2\,\mathrm{GeV}$ is shown in **black** and discards parts of the suspected peaking background. The selection on $\sigma > 0.05$ and $T < 0.85$, used for the semileptonically tagged $\tau \to \mathrm{e}\nu\bar{\nu}$ decay channel is shown in **yellow** and is discards large amounts of the background including its peaking contribution. Finally the relative complement of the invariant mass selection with respect to the σ and T selection is shown in **red**.

The previous Belle analysis which used the hadronic tag did not report any observation of this peaking background [16], whereas the previous Belle analysis which used the semileptonic tag did observe severe differences between Monte Carlo simulation and data as well, and obtained the shape for the continuum component solely from off-resonance data. This led to larger systematic uncertainties in the semileptonically tagged measurement compared to the hadronically tagged measurement, as can be seen in Table 5.17. This thesis estimates the continuum shape from off-resonance as well.

5.3.3. K^0_S Sideband based study

The K^0_S sideband was used to validate the shape of the overall background and to identify unknown $\Upsilon(4S)$ background processes. The sideband is defined by the presence of two additional tracks forming a good K^0_S candidate reconstructed by the `FEI`, which is not used for the reconstruction of the $\Upsilon(4S)$ candidate. Subtracting the processes identified using the off-resonance data, this sideband can be used to investigate the shape of $\Upsilon(4S)$ components, in particular the contributions due to K^0_L mesons.

By the approximate symmetry under the exchange of $K^0_S \to K^0_L$, it is expected to

detect peaking background due to processes containing K_L^0 in the signal region, where additional K_S^0 candidates are vetoed.

Background processes with a K_L^0 can be mitigated by the introduction of a K_L^0 veto, as investigated by [92]. However, K_L^0 are the least understood physics objects at the Belle experiment, hence this leads to a large systematic uncertainty (see Section 5.3.5).

There was no significant deviation from the Monte Carlo expectation observed in the hadronically tagged channels.

A possible peaking background component in the semileptonically tagged leptonic τ decay channels was observed after the final selection. The distributions can be found in Section D.2.3.

In contrast to the hadronic-tag, the semileptonic-tag used in this thesis was not calibrated. This can cause an incorrect relative fraction between correctly and incorrectly tagged B mesons, and in consequence could explain the observed deviation. The calibration factors for the semileptonic tag-side efficiency are not available, yet. Therefore, this observation was not further investigated during this thesis.

5.3.4. Fitting Procedure

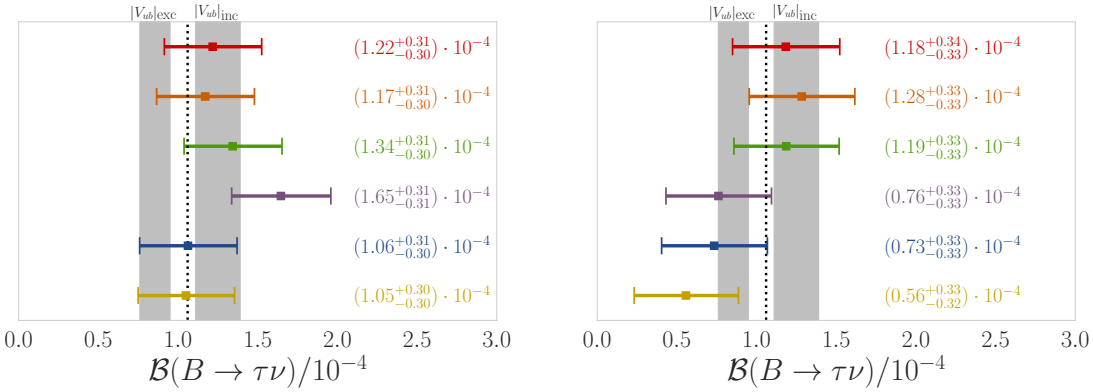
The fitting procedure was validated using six streams of Monte Carlo simulated events. Different options for the smoothing of the PDFs and the constraints used in the fit were investigated. For the final fit it was decided to use the same fitting procedure as in the previous Belle analysis [89]. This ensured a valid comparison.

However, the result of the fit depends heavily on the fit procedure. For instance, using the mis-modeled continuum Monte Carlo simulation, using b-spline based smoothing or adding additional free parameters for the normalization of the continuum component does change the final result and potentially increases the statistical uncertainty. To the best knowledge of the author, previous analyses did not take these differences, caused by the fitting procedure, into account. They used either Monte Carlo or off-resonance data, and fixed all relative branching fractions of the background components.

5.3.4.1. Monte Carlo Measurements

The final fit was performed independently on six streams of Monte Carlo simulated events. The tag-side calibration was not used, because it is only applicable to recorded data. In addition, the continuum shape was taken from Monte Carlo for these tests.

As can be seen from Figure 5.21 the results on Monte Carlo simulation reproduce the simulated branching fraction of $B \to \tau\nu_\tau$ of $1.06 \cdot 10^{-4}$ within the statistical uncertainties.



(a) hadronic tag (b) semileptonic tag

Figure 5.21.: Results of the final fit performed on the six streams of Monte Carlo simulated events. The gray bands show the predicted branching fractions by the SM, and the dotted black line the current world average, which was used in the Monte Carlo simulation.

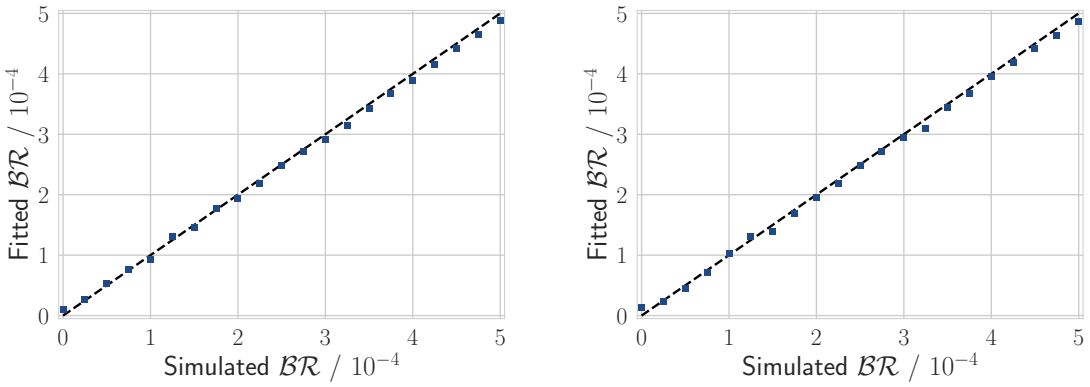
5.3.4.2. Continuum shape description

Several alternative procedures were investigated to estimate the continuum shape from off-resonance data and Monte Carlo simulation. For the final fit the shape is extracted using a linear regression (i.e. a first-order polynomial was fitted) of the off-resonance data. To estimate the systematic uncertainty, the fit was repeated 100 times by re-sampling the off-resonance data using the bootstrap method (see [34, Chapter 4.3.5]). These fits were performed using first-order, second-order and third-order polynomials. The standard deviation of the fit-results is quoted as the systematic uncertainty due to the continuum description.

The procedure yields a relative uncertainty of 36.3% and 21.7% for the hadronically and semileptonically tagged measurements, respectively. Compared to previous results, this increases the systematic uncertainty of the overall measurements. Previous results did either use the Monte Carlo description [16], or compared the first-order polynomial only with a single second-order polynomial, fitted on the original off-resonance data [15].

5.3.4.3. Closure Tests

The linearity of the fit (see [34, Chapter 10.5]) was tested for 21 different simulated branching fractions between 0 and $5 \cdot 10^{-4}$. For each simulated branching fraction 100 fits were performed using toy Monte Carlo data-samples. The toy Monte Carlo data-samples were created by randomly drawing events from the expected signal and



(a) hadronic tag (b) semileptonic tag

Figure 5.22.: Linearity test of the combined fit for the hadronically and semileptonically tagged sample. The fitted branching fraction $\mathcal{BR}$ is in agreement with the simulated branching fraction over a wide range of possible branching fractions between 0 and $5 \cdot 10^{-4}$.

background PDFs. On average the fitted branching fraction reproduces the simulated branching fraction, i.e. no bias is observed, as can be seen from Figure 5.22.

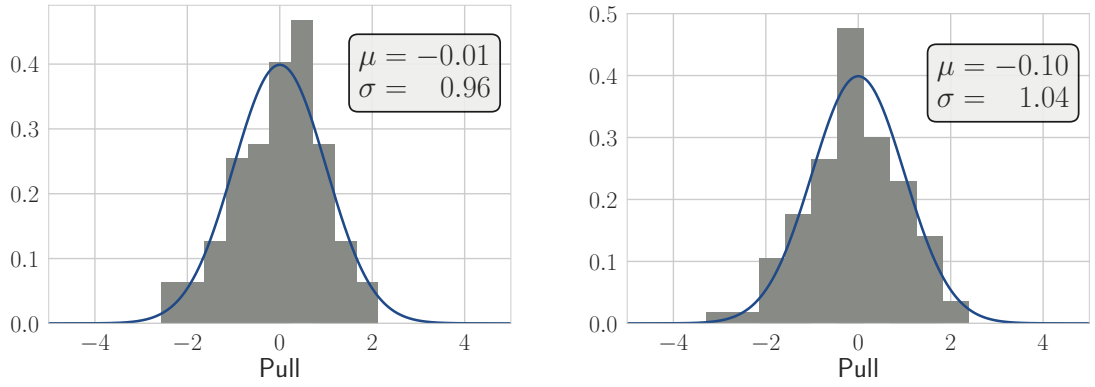
In addition, the pull distribution (see [34, Chapter 10.5]) of the fit was investigated by repeating the final fit using 100 toy Monte Carlo data-samples. The toy Monte Carlo data-samples were created by randomly drawing events from the fitted signal and background PDFs.

The obtained pull distributions for the hadronically and semileptonically tagged samples is compatible with a standard normal distribution, as can be seen from Figure 5.23. In other words: the distribution of the fit is approximately Gaussian, no bias is observed and the calculated statistical uncertainty covers the correct interval.

5.3.4.4. K^0_S Sideband

The fit was performed on the K^0_S sideband using the signal shape of the signal region. The E_{ECL} can be found in Figure D.5 and Figure D.6. As expected, there was no signal observed using the hadronically tagged sample.

A significant excess was observed on the semileptonically tagged sample with a local statistical significance of 2.76 σ. This excess was already discussed in Section 5.3.3. Further research is required as soon as the semileptonic tag-side calibration factors for the `FEI` are available.



(a) hadronic tag (b) semileptonic tag

Figure 5.23.: Pull distributions of the combined fit for the hadronically and semileptonically tagged sample. The distribution is in agreement with a standard normal distribution. μ and σ denote the mean and standard deviation of the pull distribution, respectively.

5.3.5. Systematic uncertainties

Correctly quantifying the uncertainty of a measured value is one of the most important part of a scientific measurement. The statistical uncertainty, caused by the finite amount of available data, is theoretically well understood and is either calculated analytically using the likelihood profile of the fit, or is extracted from a Monte Carlo generated distribution.

The systematic uncertainties, caused by deviations from assumptions made by the analyst, are notoriously ill-defined. Usually, in order to estimate the systematic uncertainties, those assumptions are relaxed or alternatives are investigated. Typical sources for systematic uncertainties are the selection efficiencies obtained on Monte Carlo, assumed branching fractions and PDF shapes used in the fit.

The dominating systematic uncertainties in the measurement of B $\to \tau\nu_\tau$ are the uncertainty on the tag-side efficiency correction and the PDF shapes used in the fit, in particular the continuum description. Table 5.17 summarizes all systematic uncertainties investigated by previous Belle measurements and the estimation for this analysis.

This thesis focuses on the dominating systematic uncertainties mentioned above. In a complete analysis it is best practice to quantify all known systematic uncertainties. The known systematic uncertainties and their scaling behavior with increasing luminosity are briefly summarized below.

The tag-side efficiency correction directly enters the calculation of the branching fraction in the denominator of Equation 5.16. The determination of the hadronic tag-side efficiency and its uncertainty was already discussed in Section 4.3.4. The semileptonic tag-side efficiency is not known yet, but is expected to be similar to the hadronic tag-side efficiency, because large parts of the tag-side reconstruction are identical. The associated systematic uncertainty is estimated by combining the statistical and systematic uncertainty of the tag-side efficiency correction factors. This uncertainty decreases with increasing luminosity, because the control channels used in the calibration can be measured with higher precision.

The PDF shapes used in the fit are usually determined from Monte Carlo simulation. The shapes can be sensitive to the statistical fluctuations (discussed here) and mis-modeling of the underlying Monte Carlo sample (see discussion on branching fraction and efficiencies below). The associated systematic uncertainty is estimated by the standard deviation of 100 fits performed using slightly varied PDF shapes, generated by re-sampling the Monte Carlo simulated events using the bootstrap method (see [34, Chapter 4.3.5]). This uncertainty decreases with increasing luminosity, because usually a larger luminosity is accompanied by a larger Monte Carlo sample. Alternatively, PDF shapes can be modeled by (theoretically motivated) analytical functions. Here the difference obtained by choosing different functions can be used to estimate the uncertainty associated with the specific choice used in the fit.

The continuum description determines the shape of the continuum component in the fit. The shape can be either determined by the Monte Carlo simulation (as described above) or taken from the off-resonance data. While the former is poorly modeled (as discussed in Section 2.3.3.3), the later suffers from low statistics. In particular the hadronic τ decays are affected by continuum background. In the case of off-resonance data, the systematic uncertainty is estimated by fitting different analytical functions (usually first and second order polynomials) to the off-resonance data. The detailed procedure used in this thesis is described in Section 5.3.4.2. The uncertainty decreases with increasing luminosity, because more off-resonance data is available. Furthermore, physical parameters determining the continuum spectra, like branching fractions and fragmentation constants, can be measured more precisely.

The branching fractions of the τ are only known with a finite precision. They enter the calculation of the branching fraction in the denominator of Equation 5.16. The associated systematic uncertainty is estimated by varying the branching fractions by their known uncertainty. The uncertainty decreases with increasing luminosity, because a B factory like Belle II is as well a τ factory, hence the branching fractions can be measured more precisely. However, the precision is eventually systematically limited.

The background branching fractions which are assumed during the simulation of the Monte Carlo sample, used to determine the PDF shapes, are only known with a finite precision. In particular the branching fractions of peaking background are important, because the induced uncertainty on the normalization of a peaking component cannot be determined by the fit. The associated systematic uncertainty is estimated by varying the weight of the corresponding events and repeating the fit procedure. The uncertainty decreases with increasing luminosity, because the contributing background processes can be measured more precisely. However, the precision of background branching fractions is eventually systematically limited, hence it is expected that at some points the scaling of this uncertainty decouples from the luminosity.

The track reconstruction efficiency has been studied by [93]. It has been found that a systematic uncertainty of 0.35% has to be assigned for each charged track on the signal-side. The uncertainty of the reconstruction efficiency for the tag-side is already accounted for in the tag-side efficiency correction.

The π^0 reconstruction efficiency has been studied by [94]. The deviation between Monte Carlo events and data has been found to be 4%. This uncertainty influences only the $\tau \to \rho\nu_\tau$ decay-channel, hence the associated systematic uncertainty is reduced by the fraction of signal events reconstructed in this channel (see [89]).

The particle identififcation selection efficiency has been studied by [95] and [96]. The deviation between Monte Carlo events and data influences the leptonic τ decays on the signal-side. The uncertainty of the PID selection efficiency for the tag-side is already accounted for in the tag-side efficiency correction.

The signal reconstruction efficiency directly enters the calculation of the branching fraction in the denominator of Equation 5.16. The statistical uncertainty is usually small, because large amounts of signal Monte Carlo events are used to determine this efficiency. The systematic uncertainty depends on the validity of the τ decay models used in the Monte Carlo generator (see Section 5.1.3).

The best-candidate selection can potentially have a different efficiency and modify the shape observed on data. The associated systematic uncertainty can be estimated by choosing a different best-candidate selection criterion, for instance a random best-candidate selection, or by not applying a best-candidate selection at all (as done by [89]).

The K_L^0 veto efficiency has been studied by [92]. It can influences the shape and potentially the selection efficiency for signal and background. Previous Belle analyses (e.g. [16]) used this veto to suppress background processes including K_L^0 particles. The associated systematic uncertainty is estimated by varying the weight of the corresponding events and repeating the fit procedure. Since

the KLM clusters used to reconstruct K_L^0 are the least understood physics object of the Belle experiment, the K_L^0 veto was not used during this thesis. In addition the K_L^0 could also influence the E_{ECL} distribution, however this was not studied by previous analyses.

The completeness-constraint efficiency can be studied using double-tagged samples. The fraction of events with additional charged tracks can be compared between Monte Carlo and data. [89] studied this uncertainty using a double-tagged sample $B^+ \to D^0\pi^+$ and obtained compatible results on Monte Carlo and data with a deviation of 1.9%, which can be used as an estimate for the systematic uncertainty.

The number of $B\overline{B}$ pairs directly enters the calculation of the branching fraction in the denominator of Equation 5.16. The Belle collaboration determined $N_{\Upsilon(4S)} = (7.72 \pm 0.1) \cdot 10^8$. Therefore the associated systematic uncertainty is 1.3%. Strictly speaking, this uncertainty is already contained in the tag-side efficiency uncertainty, because the final absolute branching fraction is measured with respect to the investigated control channels, hence the uncertainty already enters in the measurement of the control channels. Nevertheless, the previous Belle measurements added this uncertainty on the number of $B\overline{B}$ pairs to their list of systematic uncertainties.

5.4. Results

The results obtained from the Belle data and a sensitivity estimation for Belle II are summarized in this section.

5.4.1. Belle I

The rare decay channel $B \to \tau\nu_\tau$ was chosen as benchmark, because it was regarded as well-understood and established, however a previously neglected or under-estimated systematic uncertainty of the continuum background description was found to be sizeable.

In consequence, the overall uncertainty of the measurement could not profit from the increased statistics, due to the necessary relaxation of the selection criteria and the large systematic uncertainties associated with the peaking backgrounds in the hadronic τ decay modes. These backgrounds are unlikely to be produced by `b2bii` or the `FEI`, since they were neither encountered on the validation of tag-side only (see Section 4.3.2) nor in the control channels used for the tag-side calibration (see

Table 5.17.: Relative systematic uncertainties in percent. The previous measurements by Belle based on the hadronic tag [16] and the semileptonic tag [15] are shown in the first and second column respectively. The estimated leading systematic uncertainties of this analysis are shown in the third and fourth column. The tag-side efficiency correction for the semileptonic tag is not yet available, therefore the corresponding uncertainty was estimated with the uncertainty of the hadronic tag-side efficiency correction. The symbol - indicates that the corresponding uncertainty was not investigated.

Source	Relative uncertainty in %			
	[16]	[15]	Hadronic	Sem.-lep
Tag-side efficiency	7.1	12.6	7.0	≈ 7
PDF shapes Signal 4.2 Background 8.9		8.5	9.0	6.6
Continuum Description	-	14.1	36.3	21.7
τ branching fraction	0.6	0.2	0.4	0.4
Background branching fraction	3.8	3.1	-	-
tracking efficiency	0.3	0.4	0.4	0.4
π^0 efficiency	0.5	1.1	0.9	1.0
PID efficiency	1.0	0.5	-	-
signal reconstruction efficiency	0.4	0.6	0.4	0.3
best candidate selection efficiency	-	0.4	-	-
K^0_L veto efficiency	7.3			
completeness-constraint efficiency	-	1.9	-	-
number of $B\overline{B}$ pairs	1.3	1.4	1.3	1.3
Total	14.7	21.2	38.1	23.8

Section 4.3.4). Further research and studies of the encountered background processes are required.

The detailed results of the final fit, including the result of fitting the decay channels independently, are stated in Table 5.18 together with the results of the previous Belle analysis for comparison. The fitted distributions can be found in Figure 5.24 and Figure 5.25. The combined fits are compatible with the SM and the previous Belle results.

There are two conspicuous individual fits. Firstly, no significant signal was observed in the hadronically tagged $\tau \to \pi\nu_\tau$ channel, which is compatible with the previous individual fit. A more in-depth investigation hints that the linear regression of continuum off-resonance data, is not sufficient to describe the data in the signal region, and hence the signal contribution is suppressed although there is clearly a signal peak visible (see Figure 5.24). The large systematic uncertainty due to the continuum description of the hadronically tagged measurements reflects this issue in the final result.

Secondly, the semileptonic $\tau \to a_1\nu_\tau$ channel yields a large branching fraction with large uncertainties. This excess has a local statistical significance of 1.56 σ. This could be caused either by a statistical fluctuation or by the missing tag-side calibration factors for the semileptonic tag.

The achieved relative statistical uncertainty is very similar to the previous analyses. However, in comparison with previous analyses, the analysis presented in this thesis is very conservative.

The final selection was not optimized for maximal statistical significance, instead a phase-space region was chosen which allows an accurate estimation and validation of the background shapes on off-resonance data and sidebands. In particular, the completeness-constraint was not fully exploited, in contrast to previous analyses.

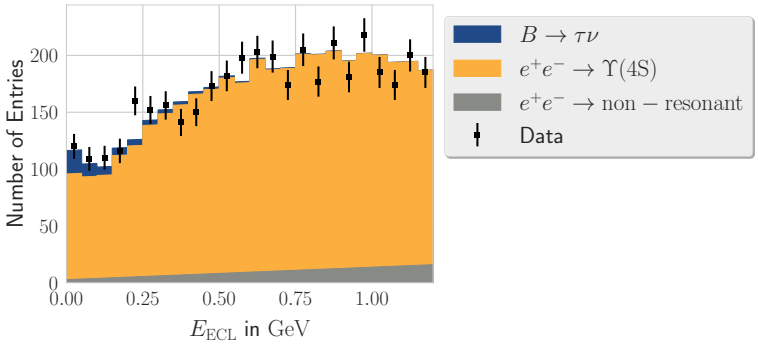
The shape of the continuum component was estimated from off-resonance data and does not rely on Monte Carlo simulation. The previous Belle analysis based on the semileptonic tag took the same approach. Whereas the previous Belle analysis based on the hadronic tag relied on Monte Carlo simulation and consequently reported a smaller systematic uncertainty (see Table 5.17).

This work advocates the use of machine learning. However, to ensure a sound validation of `b2bii` and the `FEI`, only simple cuts on well-understood quantities were used for the signal-side reconstruction and final selection. In particular, this work did not employ a multivariate continuum suppression, which could have introduced additional systematic uncertainties due to the encountered mis-modeling of the continuum processes.

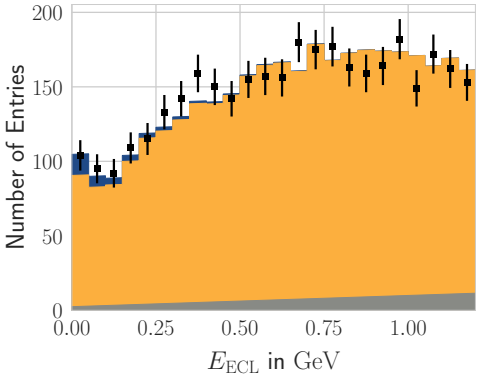
The K_L^0 veto described in [92] was not used, because KLM clusters used to reconstruct K_L^0 are the least understood physics object of the Belle experiment. However, a

Table 5.18.: Results obtained by previous Belle analyses compared to this analysis. The columns show the number of fitted B $\to \tau\nu_\tau$ events N_{sig}, the selection efficiency including the τ branching fractions ε, and the fitted branching fraction $\mathcal{BR}$. The stated selection efficiencies for the hadronic tag include the hadronic tag-side efficiency correction estimated from data. The tag-side efficiency correction for the semileptonic tag was not available at the time of writing, and is therefore **not** included. The combined hadronic tag has a statistical significance of 3.82 σ. The combined semileptonic tag has a statistical significance of 5.20 σ.

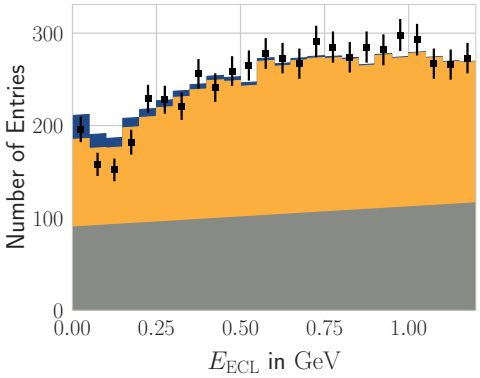
Decay-Channel	Previous			This analysis		
	N_{sig}	$\varepsilon(10^{-4})$	$\mathcal{BR}(10^{-4})$	N_{sig}	$\varepsilon(10^{-4})$	$\mathcal{BR}(10^{-4})$
Hadronic Tag [16]						
e^-	16^{+11}_{-9}	3.0	$0.68^{+0.49}_{-0.41}$	116^{+34}_{-33}	6.3	$2.38^{+0.71}_{-0.68}$
μ^-	26^{+15}_{-14}	3.1	$1.06^{+0.63}_{-0.58}$	71^{+32}_{-31}	4.1	$2.21^{+0.99}_{-0.96}$
π	8^{+10}_{-8}	1.8	$0.57^{+0.70}_{-0.59}$	-43^{+64}	9.8	$-0.58^{+0.84}$
ρ	14^{+19}_{-8}	3.4	$0.52^{+0.72}_{-0.62}$	225^{+64}_{-63}	6.2	$4.67^{+1.33}_{-1.31}$
a_1	–	–	–	-5^{+45}	1.6	$-0.39^{+3.54}$
Combined	62^{+23}_{-22}	11.2	$0.72^{+0.27}_{-0.25}$	327^{+89}_{-88}	28.1	$1.51^{+0.41}_{-0.40}$
Semileptonic Tag [89]						
e^-	47 ± 25	7.4	0.90 ± 0.47	152^{+40}_{-40}	8.4	$2.34^{+0.62}_{-0.61}$
μ^-	13 ± 21	5.5	0.34 ± 0.55	73^{+36}_{-35}	5.2	$1.81^{+0.90}_{-0.87}$
π	57 ± 21	4.7	1.82 ± 0.68	57^{+60}_{-57}	12.3	$0.60^{+0.63}_{-0.60}$
ρ	119 ± 33	7.3	2.16 ± 0.60	202^{+68}_{-67}	9.6	$2.73^{+0.92}_{-0.90}$
a_1	–	–	–	96^{+63}_{-62}	2.5	$4.97^{+3.24}_{-3.19}$
Combined	222 ± 50	25.0	1.25 ± 0.28	528^{+106}_{-103}	38.0	$1.80^{+0.36}_{-0.35}$



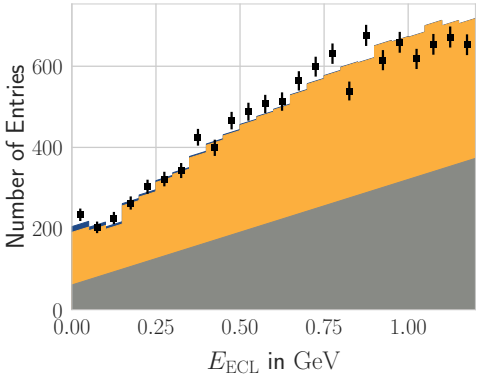
(a) $\tau \to e^- \nu_e \nu_\tau$



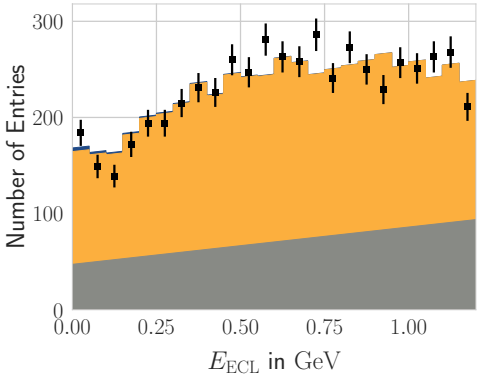
(b) $\tau \to \mu^- \nu_\mu \nu_\tau$



(c) $\tau \to \pi^+ \nu_\tau$

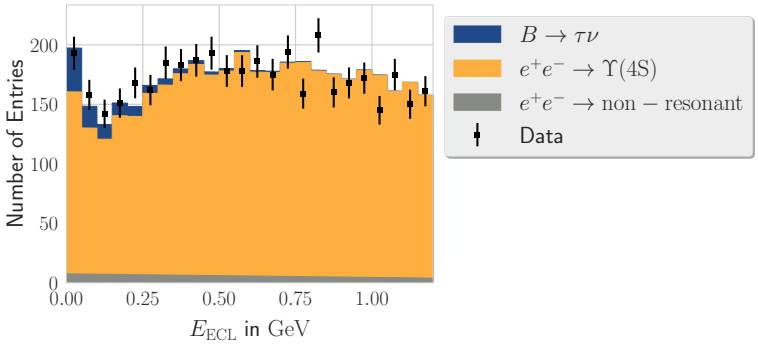


(d) $\tau \to \rho \nu_\tau$

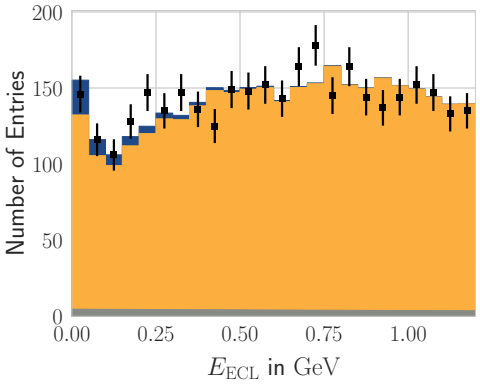


(e) $\tau \to a_1 \nu_\tau$

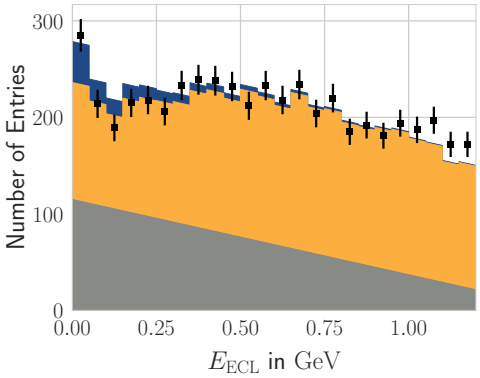
Figure 5.24.: Fitted distribution of E_{ECL} for the hadronically tagged candidates.



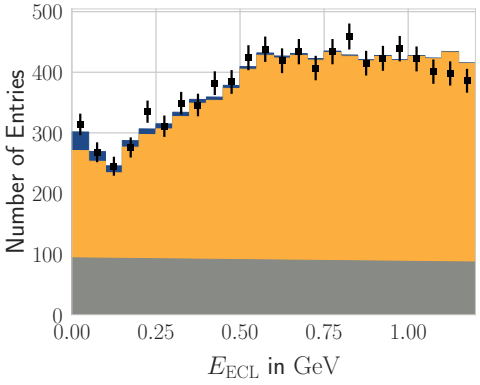
(a) $\tau \to \mathrm{e}^-\nu_\mathrm{e}\nu_\tau$



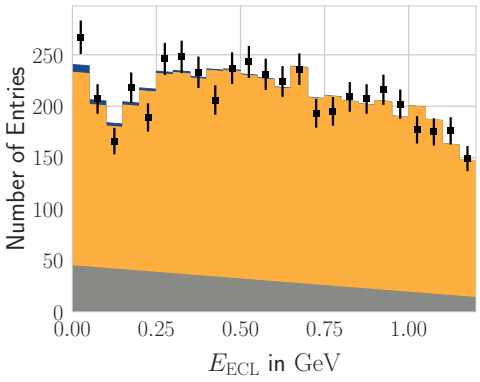
(b) $\tau \to \mu^-\nu_\mu\nu_\tau$



(c) $\tau \to \pi^+\nu_\tau$



(d) $\tau \to \rho\nu_\tau$



(e) $\tau \to \mathrm{a}_1\nu_\tau$

Figure 5.25.: Fitted distribution of E_{ECL} for the semileptonically tagged candidates.

possible effect on the extra-energy in the electromagnetic calorimeter was studied using the K_S^0 sideband.

The increase in the expected reconstruction efficiency for the **hadronic tag** is compatible with the expected improvements due to the increased tag-side efficiency. In particular the leptonic τ decay-modes agree well with the expectation. The quoted numbers include the tag-side efficiency correction obtained from control channels on data.

The result for the combined fit using the hadronic tag including statistical and systematic uncertainties is

$$\mathcal{BR}(B \to \tau\nu_\tau) = \left(1.51^{+0.41}_{-0.40} \pm 0.57\right) \cdot 10^{-4}. \tag{5.19}$$

The shift of the central values with respect to the previous Belle analysis can be explained with the additional uncorrelated systematic uncertainty discovered during this thesis. The previous Belle analysis was too optimistic in the estimation of the systematic uncertainty due to the continuum background. Furthermore, although the same dataset was used, the overlap between the investigated signal events is maximally 40%, due to the relaxed selection criteria and the improved overall signal selection efficiency.

The increase in the expected reconstruction efficiency for the **semileptonic tag** is not as large as expected. One reason is the tight cut on the `SignalProbability` of > 0.05, which was necessary due to the encountered peaking backgrounds. Moreover this efficiency does not yet include the tag-side efficiency correction.

The result for the combined fit using the semileptonic tag including statistical and systematic uncertainties is

$$\mathcal{BR}(B \to \tau\nu_\tau) = \left(1.80^{+0.36}_{-0.35} \pm 0.43\right) \cdot 10^{-4}. \tag{5.20}$$

The shift in the central value with respect to the previous Belle analysis can be explained by the additional systematic uncertainty due to the continuum background. Although the previous analysis used off-resonance data as well to estimate the continuum shape, this thesis relaxed the selection criteria and has a larger overall signal selection efficiency. In consequence, the continuum background description is different. Furthermore, the presented result does not contain the semileptonic tag-side calibration, which was not available at the time of writing. The tag-side calibration influences both: the signal selection efficiency, and the shape of the dominating charged $B\overline{B}$ background. Finally, although the same dataset was used, the overlap between the investigated events is maximally 65%, due to the relaxed selection criteria and the improved overall signal selection efficiency.

Taking into account the partly uncorrelated statistical uncertainty and the uncorrelated dominating component of the systematic uncertainties, the obtained central values for the branching fraction of $B \to \tau\nu_\tau$ are compatible with the previous Belle analyses, the previous BaBar analyses, and the SM, as can be seen from Figure 5.26.

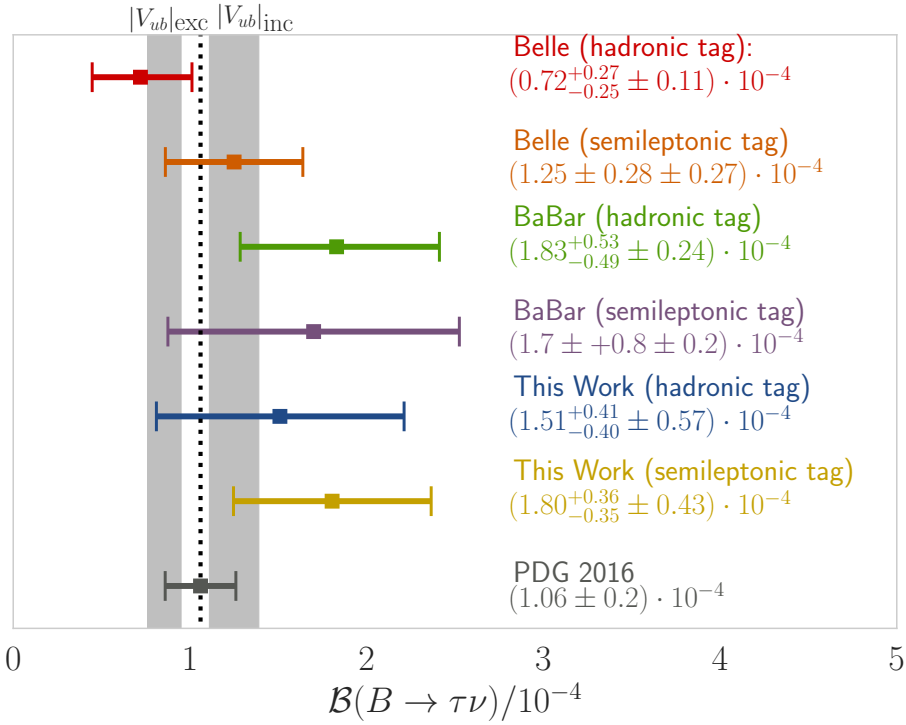


Figure 5.26.: Results obtained by the previous Belle analysis and this thesis. The gray bands show predicted branching fractions by the SM, and the dotted black line the current world average, which was used in the Monte Carlo simulation.

Table 5.19.: Statistical uncertainty: Scaled with $\frac{1}{\sqrt{\mathcal{L}}}$ for Belle II.

Scenario	Luminosity $\mathcal{L}$ in ab^{-1}	Relative Uncertainty in % Hadronic	Semileptonic	State
Belle	0.711	37.5	22.4	Measured
b2bii	0.711	27.1	20.0	Measured
Belle II	1	22.9	16.9	Scaled
Belle II	5	10.2	7.5	Scaled
Belle II	50	3.2	2.4	Scaled

5.4.2. Belle II

The rare decay $B \to \tau\nu_\tau$ was used to validate the entire software stack developed during this thesis. The initial goal of validating the software was accomplished. The `b2bii` package, the `mva` package and the `FEI` are ready to be used for physics analyses.

Assuming that the current reconstruction issues in `BASF2` will be fixed, one can estimate the statistical uncertainty on the branching fraction $B \to \tau\nu_\tau$ achievable by Belle II, based on the statistical uncertainty obtained using `b2bii`.

Table 5.19 states the expected relative statistical uncertainty with increasing integrated luminosity. As described in the previous section, the leading systematic uncertainties are expected to scale with the luminosity as well.

Based on the findings described in this thesis, several recommendations for the Belle II experiment can be deduced.

1. The Belle II reconstruction software has to be (and is currently) improved. The fake-rate of tracks has to be decreased significantly.
2. The Monte Carlo simulation does not describe the continuum background satisfactorily. A larger off-resonance data sample is paramount to describe the continuum background.
3. The data-driven techniques presented in Section 3.3.3 should be used to ensure to suppress background from continuum processes.
4. The extra energy in the electromagnetic calorimeter has to be better understood. Previous analysis primarily tested the shape of the signal component using double-tagged samples [91]. However, the shape of the background distributions (in particular from continuum processes) were not validated to the same degree of accuracy.
5. The tag-side calibration is key to the successful employment of the `FEI`, it should be provided by the collaboration for both: the hadronic and the semileptonic tag.

5.5. Conclusion

The benchmark analysis B $\to \tau\nu_\tau$ successfully used the entire stack of software developed during this thesis. The `b2bii` package was used to convert the Monte Carlo simulated Belle events and the data recorded by the Belle experiment. The hadronic and semileptonic tag was provided by the `FEI`, which in turn uses the `mva` package.

The expected improvements due to the larger tag-side efficiency were observed in all tag-side and signal-side combinations on Monte Carlo simulated Belle events. The obtained results on recorded data are consistent: with the Monte Carlo expectation, among each other, with the previous Belle analyses and with the Standard Model prediction. The Belle II reconstruction software is not yet mature enough to allow a detailed sensitivity study.

The limitations of the Belle Monte Carlo simulation and current Belle II reconstruction became apparent. The availability of recorded data, provided by the `b2bii` package, proved indispensable. In particular the continuum processes require further research, in order to reduce the systematic uncertainties in measurements which rely on the extra energy in the electromagnetic calorimeter. An increased off-resonance sample size would allow to constrain the shape of the continuum component more precisely.

The analysis presented in this chapter successfully reproduced the previous results reported by Belle and BaBar. It can be used as a template for future analyses which use `b2bii` or the `FEI`.

Chapter 6

Conclusion

This PhD thesis covered four major topics.

The Belle to Belle II Conversion package (`b2bii`) enables Belle II physicists to analyze the dataset recorded by Belle using `BASF2`. Thus, the entire Belle II analysis software stack was validated on recorded data, years before recorded data from the Belle II experiment was available.

The multivariate analysis package (`mva`) enables Belle II physicists to keep up with the rapid developments in the field and to easily employ modern machine learning algorithms in their work. Most of the multivariate methods used in the reconstruction and analysis algorithms in `BASF2` are built on the `mva` package and use the default classification method `FastBDT`, both developed during this thesis.

The Full Event Interpretation algorithm (`FEI`) enables Belle II physicists to measure a wide range of interesting decays with a minimum amount of detectable information. The `FEI` more than doubles the tag-side efficiency compared to its (already very successful) predecessor.

The $B \to \tau\nu$ benchmark analysis validated the entire Belle II analysis software stack. Unresolved reconstruction issues in the Belle II framework and previously unknown background contributions were discovered before Belle II started recording data. It successfully reproduced the previous results reported by Belle. The analysis serves as a prototype for other current (using `b2bii`) and upcoming exclusively tagged analyses.

All software packages developed during this thesis were validated on data and are used in production by Belle II physicists all over the world.

Appendices

Appendix A

B2BII

A.1. Monitoring Histograms

The full list of extracted quantities used for validating the Belle to Belle II Conversion. The number in parenthesis states the number of quantities.

Beam Parameters (16):

- the experiment, run and event number (3);
- the energy in the HER (High Energy Ring), LER (Low Energy Ring) and CMS (Center Of Mass System) (3);
- the crossing angle between the HER and LER (1);
- and the position and uncertainty of the interaction point (9).

K^0_S, Λ, and converted γ (201):

- invariant mass, four-momentum and vertex position (8);
- four-momentum and POCAs (point of closest approach) of the daughters (14);
- PID information calculated by the K^0_S finder (4);
- invariant mass, four-momentum and vertex position after a mass-constraint vertex fit (8);
- uncertainties of the four-momenta and vertex position calculated by the mass-constrained vertex-fit (28);
- p-value of the mass-constrained vertex fit (1);

- and Monte Carlo truth of the total momentum and PDG value of the daughters (4);

K_L^0 (9):

- KLM cluster position (3);
- number of KLM layers with hits (1);
- innermost KLM layer with hits (1);
- and Monte Carlo truth of the four-momentum and PDG value (4);

Tracks (41):

- PID information including quality indicators (6);
- four-momentum and POCA (7);
- uncertainties of the four-momentum and POCA (28);

γ (48):

- four-momentum and ECL cluster position (7);
- uncertainties of the four-momentum and cluster position (28);
- spherical coordinates of the ECL cluster position (3);
- ECL cluster shape quantities (6);
- and Monte Carlo truth of the four-momentum (4);

π^0 (41):

- four-momentum and vertex-position (7);
- invariant mass calculated using four-momentum and the daughter four-momenta (2)
- uncertainties of the four-momentum and vertex position (28);
- and Monte Carlo truth of the four-momentum (4);

MC Information (12):

- four-momentum of the MC particle (4);
- PDG code of the MC particle (1);
- vertex position or POCA of the MC particle (3);
- PDG code of the mother of charged and neutral pions (3);
- and number of daughters (1)

Appendix B

MVA

B.1. Loss Functions

In HEP the quantity we care about, e.g. the combined statistical and systematical uncertainties, usually cannot be optimized efficiently. Instead a surrogate loss-function $\mathcal{L}$ is used, to optimize the statistical model by minimizing the empirical risk (defined in Equation 3.2). This loss-function is not uniquely defined and the choice is not straightforward. A detailed discussion can be found in [23].

In this section I present a framework to derive suitable loss-functions based on the maximum likelihood principle.

The likelihood to observe the data $\mathcal{D} = \{y_1, y_2, \ldots, y_n\}$ given a parametrized distribution assumption $g(\vec{w})$ is

$$p(\mathcal{D}|\vec{w}) = \prod_i^n g(y_i, \vec{w}). \tag{B.1}$$

The maximum likelihood principle states that the optimal values of the parameters $\vec{w}$ can be estimated by maximizing the likelihood of observing the data.

Instead of maximizing the likelihood, one can equivalently minimize the negative logarithm of the likelihood

$$-\log\left(p(\mathcal{D}|\vec{w})\right) = -\sum_i^n \log\left(g(y_i, \vec{w})\right). \tag{B.2}$$

We can now identify the right-hand side of Equation B.2 with the empirical risk defined in Equation 3.2 to extract the maximum likelihood loss-function

$$\mathcal{L}_{\text{ML}} = -\log\left(g(y, \vec{w})\right). \tag{B.3}$$

In multivariate analysis, the parameters $\vec{w}$ are a function of the observed features $\vec{x}$. In consequence, although the assumed distribution g can be simple, the underlying statistical model can be extremely complex.

B.1.1. Regression

In the case of regression, a reasonable distribution assumption would be a Gaussian distribution

$$g = \mathcal{N}(y, \mu(\vec{x}), \sigma) \tag{B.4}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{B.5}$$

Using the maximum likelihood principle we can derive the ubiquitous square loss-function

$$\mathcal{L}_{\text{square}} = -\log\left(\mathcal{N}(y, \mu)\right) \tag{B.6}$$

$$\sim \quad (y-\mu)^2, \tag{B.7}$$

where terms and factors independent of y and $\vec{x}$ were omitted. So assuming a Gaussian distribution for the target y, we can estimate the mean value of y using our statistical model $\mu = \widehat{f}(\vec{x})$ by minimizing the square loss in Equation B.7.

B.1.2. Classification

In the case of classification the underlying distribution is a Bernoulli distribution

$$g = \mathcal{B}(y, p(\vec{x})) \tag{B.8}$$

$$= p^y \cdot (1-p)^{1-y} \tag{B.9}$$

Using the maximum likelihood principle we can derive the ubiquitous cross-entropy loss-function

$$\mathcal{L}_{\text{cross-entropy}} = -\log\left(\mathcal{B}(y, p)\right) \tag{B.10}$$

$$= \quad y \log p + (1-y)\log(1-p). \tag{B.11}$$

So, using the Bernoulli distribution for the target y, we can estimate the mean value of y using our statistical model $p = \widehat{f}(\vec{x})$ by minimizing the cross-entropy loss in Equation B.11.

It is important to notice, that while the maximum likelihood principle produces a suitable loss function to optimize the parameters $\vec{w}$ of the distribution g, it does not state the optimal estimator $\widehat{f}(\vec{x})$ to predict y. For instance, instead of the mean value, one could also use the median of the estimated distribution g. The correct choice depends on the cost of a correct or wrong decision, respectively.

B.1.3. Cross Entropy

There is an interesting connection to information theory here. By observing that the right-hand side of Equation B.2 is an approximation of an expectation value over the true distribution f of y for fixed $\vec{x}$

$$\underset{y\sim f}{\mathrm{E}} \left(-\log\left(g(y,\vec{w})\right)\right) \approx -\sum_{i}^{n} \log\left(g(y_i,\vec{w})\right) \tag{B.12}$$

we can identify the left-hand side with the cross-entropy

$$H(p,q) = -\underset{p}{\mathrm{E}} \log(q). \tag{B.13}$$

The cross-entropy measures the difference between two probability distributions p and q. It measures the number of bits required on average to identify an event drawn from p, if the employed coding scheme was optimized for q. The larger the difference between q and p the larger is their cross-entropy.

The presented maximum likelihood loss-function $\mathcal{L}_{\mathrm{ML}}$ does not only maximize the likelihood, it also minimizes the cross-entropy between the true distribution f and the estimated distribution g.

B.2. Python Interface

The `mva` package contains an API (application programming interface) defining a set of `Python` **hook-functions**, which can be used to build, fit, save, load and apply arbitrary `Python`-based MVA methods. By implementing the **hook-functions**, the user can employ all MVA frameworks which provide a `Python` interface.

For the most popular `Python`-based MVA frameworks, the `mva` package already predefines all the necessary **hook-functions**, however the user can still override them in a user-defined `Python` file.

In the following I describe the standard **hook-functions** used to fit and apply a `Python`-based MVA method. During the fitting-phase:

- the total number of events, features and spectators, and a user-defined configuration string is passed to **get_model** returning a state-object, which represents the statistical model of the method in memory and is passed to all subsequent calls;
- a validation dataset is passed to **begin_fit**, which can be used during the fitting to monitor the performance;

- the training dataset is streamed to **partial_fit**, which may be called several times if the underlying method is capable to perform out-of-core fitting;
- finally **end_fit** is called returning a serializable object, which is stored together with the user-defined `Python` file as a backend-agnostic `WeightFile` in the `Belle II Conditions Database`, and can be used later to load the fitted method during the inference-phase.

During the inference-phase:

- the user-defined `Python` file is loaded from the `WeightFile` into the `Python` interpreter and the serialized object is passed to **load** returning the state-object, which represents the statistical model of the method in memory;
- the state-object and a dataset is passed to **apply** returning the response of the statistical model, usually either the signal-probability (classification) or an estimated value (regression).

In addition, the user can override the **feature_importance** function, to include the feature importances in the automatic evaluation of the `mva` package.

B.3. Feature Importance on the Benchmark

The importance of features is a useful information for physicists, to gain knowledge about the (usual) black-box behavior of an MVA method. However, as can be seen from Figure 3.4b, the importance estimations of features strongly depends on the method. In fact the contribution of an individual feature to the classification is an ill-posed problem, due to the multivariate correlations.

To mitigate this problem, the `mva` package implements three different approaches to estimate the feature importance; these are listed below in order of increasing accuracy and computing time.

Internal The importance estimation of the method itself. Not all methods provide this (most neural network implementations do not), and some can give only a rough estimate (total information gain of a feature in a boosted decision tree).

Iterative The method is fitted $N + 1$ times (where N is the number of features). Each time one of the features is excluded from the fitting-phase. The difference in the area under the receiver operating characteristic curve (AUC ROC) is used to estimate the feature importance. This approach is independent of the used method but underestimates the importance of features whose information is highly correlated to another feature.

Recursive The method is fitted $N(N+1)/2$ times (where N is the number of features). Here, the iterative approach is applied recursively, by removing the most-important feature after each iteration. This will take the correlations of variables into account.

The three methods of the `mva` package using `FastBDT` as backend, are compared to the methods provided by `NeuroBayes` (used by Belle). The iterative approach of the `mva` package corresponds roughly to the loss (the loss in significance if the variable is removed) method of `NeuroBayes`. The recursive approach of the `mva` package corresponds roughly to the significance calculated by `NeuroBayes`.

As can be seen from Table B.1 different feature importance estimation methods do not provide a compatible estimate of the feature importance, except for the most important features. This is due to the many correlations between the variables. For instance: of the three pairwise invariant masses $\mathrm{M}_{\mathrm{K}^-\pi^+}$, $\mathrm{M}_{\mathrm{K}^-\pi^0}$, and $\mathrm{M}_{\pi^+\pi^0}$, only two are independent; and the PID variables $\mathrm{Kaon}-\mathrm{ID}_{\mathrm{K}^-}$ and $\mathrm{Pion}-\mathrm{ID}_{\mathrm{K}^-}$ contain exactly the same information.

Nevertheless, the feature importances can provide valuable insight into the information utilized by the MVA method, and facilitates the detection of errors in the software (for instance physical observables which include MC information by mistake).

Table B.1.: Feature importance ranks (most important feature is 1, least important 28) for the benchmark problem presented in Section 3.2.2 estimated by different approaches.

Feature	FastBDT			NeuroBayes	
	Internal	Iterative	Recursive	Loss	Significance
M	1	1	1	1	2
$M_{K^-\pi^0}$	2	4	2	2	1
pt_{K^-}	3	2	4	14	14
$M_{K^-\pi^+}$	4	7	7	5	9
dr	5	22	6	10	3
$\text{Kaon} - \text{ID}_{K^-}$	6	26	9	4	6
$M_{\pi^+\pi^0}$	7	12	23	3	4
pt	8	11	17	6	5
p_{π^0}	9	14	19	13	7
dr_{K^-}	10	16	13	24	24
pt_{π^0}	11	15	24	11	13
p_{K^-}	12	24	12	18	22
p	13	23	26	15	16
dr_{π^+}	14	9	15	9	11
χ^2_{prob}	15	17	27	7	10
$\text{Pion} - \text{ID}_{K^-}$	16	10	25	28	28
pz	17	25	10	17	17
dz	18	27	11	22	19
pt_{π^+}	19	18	5	27	27
dz_{K^-}	20	13	14	21	21
pz_{K^-}	21	28	18	26	26
$\chi^2_{\text{prob},\pi^+}$	22	8	20	12	12
pz_{π^0}	23	21	8	16	15
dz_{π^+}	24	6	28	25	25
p_{π^+}	25	3	21	8	8
χ^2_{prob,K^-}	26	5	3	23	23
pz_{π^+}	27	19	16	19	20
$\chi^2_{\text{prob},\pi^0}$	28	20	22	20	18

Appendix C

FEI

C.1. Default Configuration

A detailed description of the current default configuration of the `Full Event Interpretation` is given in this section. In the default configuration all employed multivariate classifiers have the same configuration: a boosted decision tree (`FastBDT`) with 400 trees, a depth of 3 per tree, a shrinkage of 0.1 and a sub-sampling rate (or bagging rate) of 0.5. The output of the boosted decision tree is named σ below. There is no cut applied on the output of the vertex fitting, however the information of the vertex fitting is used in the multivariate classifier.

All candidates reconstructed from one specific decay-channel are ranked according to a variable and only the n best candidates are kept for each decay-channel – this is known as **pre-cut**. A multivariate classifier is applied to all candidates. Afterwards, all candidates reconstructed from a specific particle are ranked according to the output of the associated multivariate classifier and only the m best candidates are kept for each particle – this is known as **post-cut**. The pre-cut and post-cut of all particles is summarized in Table C.1.

In the remainder of this chapter, the particles explicitly reconstructed by the `FEI` are discussed in detail. The `FEI` does not explicitly reconstruct intermediate resonances like ρ, ω or K^*, however these resonances are implicitly exploited by the multivariate classifiers applied to all candidates. The explicitly reconstructed decay-channels are listed below. The **highlighted** decay-channels were already used by the `Full Reconstruction (FR)` algorithm employed by Belle. The charge-conjugated particle and decay-channels are always implied throughout the text.

Table C.1.: pre-cut: ranking criteria and number of best candidates which is kept for each decay-channel before the vertex fitting and the multivariate classifier application. post-cut: ranking criteria and number of best candidates which is kept in total after the multivariate classifier was applied. P-ID is the particle identification information of particle P from the PID sub-detectors, E is the energy of the particle, σ is the signal probability of the particle outputted by the corresponding multivariate classifier, M is the invariant mass, M_{P} is the nominal mass of particle P, Q is the released energy in the decay, Q_{P} is the nominal released energy in the decay of the particle P, and $\prod_i \sigma_i$ is the product over the signal probabilities of all daughter particles.

Particle	pre-cut			post-cut					
e^+	10	highest	e-ID	5	highest	σ	and	0.01	$< \sigma$
μ^-	10	highest	μ-ID	5	highest	σ	and	0.01	$< \sigma$
π^+	20	highest	π-ID	10	highest	σ	and	0.01	$< \sigma$
K^+	20	highest	K-ID	10	highest	σ	and	0.01	$< \sigma$
γ	40	highest	E	20	highest	σ	and	0.01	$< \sigma$
π^0	20	lowest	$\lvert M - M_{\pi^0}\rvert$	10	highest	σ	and	0.01	$< \sigma$
$\mathrm{K}^0_{\mathrm{S}}$	20	lowest	$\lvert M - M_{\mathrm{K}^0_{\mathrm{S}}}\rvert$	10	highest	σ	and	0.01	$< \sigma$
$\mathrm{K}^0_{\mathrm{L}}$	20	lowest	$\lvert M - M_{\mathrm{K}^0_{\mathrm{L}}}\rvert$	10	highest	σ	and	0.01	$< \sigma$
D^0 (had)	20	lowest	$\lvert M - M_{\mathrm{D}^0}\rvert$	10	highest	σ	and	0.001	$< \sigma$
D^0 (sem)	20	highest	$\prod_i \sigma_i$	10	highest	σ	and	0.001	$< \sigma$
D^0 (klong)	20	highest	$\prod_i \sigma_i$	10	highest	σ	and	0.001	$< \sigma$
D^+ (had)	20	lowest	$\lvert M - M_{\mathrm{D}^+}\rvert$	10	highest	σ	and	0.001	$< \sigma$
D^+ (sem)	20	highest	$\prod_i \sigma_i$	10	highest	σ	and	0.001	$< \sigma$
D^+ (klong)	20	highest	$\prod_i \sigma_i$	10	highest	σ	and	0.001	$< \sigma$
D^{+*} (had)	20	lowest	$\lvert Q - Q_{\mathrm{D}^{+*}}\rvert$	10	highest	σ	and	0.001	$< \sigma$
D^{+*} (sem)	20	lowest	$\lvert Q - Q_{\mathrm{D}^{+*}}\rvert$	10	highest	σ	and	0.001	$< \sigma$
D^{+*} (klong)	20	lowest	$\lvert Q - Q_{\mathrm{D}^{+*}}\rvert$	10	highest	σ	and	0.001	$< \sigma$
$\mathrm{D}^+_{\mathrm{s}}$ (had)	20	lowest	$\lvert M - M_{\mathrm{D}^+_{\mathrm{s}}}\rvert$	10	highest	σ	and	0.001	$< \sigma$
$\mathrm{D}^+_{\mathrm{s}}$ (klong)	20	highest	$\prod_i \sigma_i$	10	highest	σ	and	0.001	$< \sigma$
$\mathrm{D}^{+*}_{\mathrm{s}}$ (had)	20	lowest	$\lvert Q - Q_{\mathrm{D}^{+*}_{\mathrm{s}}}\rvert$	10	highest	σ	and	0.001	$< \sigma$
$\mathrm{D}^{+*}_{\mathrm{s}}$ (klong)	20	lowest	$\lvert Q - Q_{\mathrm{D}^{+*}_{\mathrm{s}}}\rvert$	10	highest	σ	and	0.001	$< \sigma$
B^+ (had)	20	highest	$\prod_i \sigma_i$	20	highest	σ			
B^+ (sem)	20	highest	$\prod_i \sigma_i$	20	highest	σ			
B^+ (klong)	20	highest	$\prod_i \sigma_i$	20	highest	σ			
B^0 (had)	20	highest	$\prod_i \sigma_i$	20	highest	σ			
B^0 (sem)	20	highest	$\prod_i \sigma_i$	20	highest	σ			
B^0 (klong)	20	highest	$\prod_i \sigma_i$	20	highest	σ			

C.1.1. Final State Particles

Particles which can directly be reconstructed from `Tracks`, `ECL clusters`, `KLM clusters`, or `V0 objects` are considered final state particles by the `FEI`.

C.1.1.1. Charged

The charged final state particles e^+, μ^+, π^+ and K^+ are reconstructed from `Tracks` which are close to the IP, that is a maximum radial distance of 2cm and maximum distance along the beam-pipe of 4cm).

Only particles from the primary interaction are regarded as signal. The input features for the multivariate classifiers include: the likelihoods of the PID detectors; kinematics and tracking variables; the position in the best-candidate ranking of the pre-cut.

Protons are not used in the default configuration of the `FEI`.

C.1.1.2. Neutral

Photons γ are reconstructed from `ECL clusters` and `V0 objects`. Only photons with an energy of at least 140MeV (100MeV) in the forward, 130MeV (50MeV) in the barrel, and 200MeV (150MeV) in the backward region, are considered. The numbers in parenthesis are for converted Belle data.

Only particles from the primary interaction are regarded as signal. The input features for the multivariate classifiers include: kinematic variables and the position in the best-candidate ranking of the pre-cut. Additionally, cluster information like the numbers of hits and the timing is used for photons reconstructed from `ECL clusters`.

Λ mesons (reconstructed from `V0 objects`) and neutrons are not used in the default configuration of the `FEI`. K_S^0 and K_L^0 mesons are discussed below.

Neutrinos ν are not explicitly reconstructed by the `FEI`, because they exit the Belle and Belle II detector undetected. Therefore, the `FEI` treats decay-channels containing a neutrino separately from other decay-channels, due vastly different uncertainties on the kinematics.

C.1.2. Light Neutral Mesons

C.1.2.1. π^0

Neutral pions π^0 are reconstructed from two photons.

1. $\pi^0 \to \gamma\gamma$

For Belle II the `FEI` does this reconstruction itself, on converted Belle data the **default π^0 reconstruction** of the `B2BII` package is used. Only π^0 candidates with an invariant mass between 80MeV and 180MeV are considered.

The input features for the multivariate classifiers include: kinematic variables and the position in the best-candidate ranking of the pre-cut. In addition the signal probability σ of the two daughter photons is used for Belle II.

C.1.2.2. K^0_S

Neutral kaons K^0_S are reconstructed from two charged pions, two neutral pions or a `V0 object`.

1. $K^0_S \to \pi^+\pi^-$
2. $K^0_S \to \pi^0\pi^0$

For converted Belle data only the `V0 objects` are used. Only K^0_S candidates with an invariant mass between 400MeV and 600MeV are considered.

The input features for the multivariate classifiers include: kinematic and vertex position variables, and the position in the best-candidate ranking of the pre-cut. In addition the information provided from the Belle K^0_S finder is used for converted Belle data. For Belle II the signal probability σ, tracking variables and the total momentum in the rest frame of the K^0_S, of each daughter is used.

C.1.2.3. K^0_L

Neutral kaons K^0_L are reconstructed from `KLM clusters`. All candidates are considered. The usual signature of hits in the KLM without an associated track does only provide a very rough estimate of the direction and energy of the assumed K^0_L. Therefore, the `FEI` treats decay-channels containing K^0_L particles separately from other decay-channels, due to the vastly different uncertainties on the kinematics.

The input features for the multivariate classifiers include: the energy and the `KLM cluster` timing information.

C.1.3. Charmed Mesons

All charmed mesons have the same input features for the multivariate classifiers: lorentz invariant kinematic and vertex variables of the charmed meson; kinematic, tracking and vertex variables of all daughters evaluated in the rest frame of the charmed meson; the decay mode identifier and signal probability σ of all daughters; the invariant mass of all possible combinations of daughter particles; and the position of the candidate in the best candidate ranking of the pre-cut.

Since all kinematic and vertex informations are either Lorentz-invariant like the invariant mass or evaluated in the rest frame of the charmed meson, the performance of the multivariate classifier is independent of the B meson decay from which the charmed meson originated.

C.1.3.1. D^0 meson from hadronic decay-channels

The following hadronic channels are used to reconstruct D^0 mesons:

1. $\mathbf{D^0 \to K^- \pi^+}$
2. $\mathbf{D^0 \to K^- \pi^+ \pi^0}$
3. $D^0 \to K^- \pi^+ \pi^0 \pi^0$
4. $\mathbf{D^0 \to K^- \pi^+ \pi^+ \pi^-}$
5. $D^0 \to K^- \pi^+ \pi^+ \pi^- \pi^0$
6. $\mathbf{D^0 \to \pi^- \pi^+}$
7. $\mathbf{D^0 \to \pi^- \pi^+ \pi^0}$
8. $D^0 \to \pi^- \pi^+ \pi^0 \pi^0$
9. $D^0 \to \pi^- \pi^+ \pi^+ \pi^-$
10. $\mathbf{D^0 \to K^0_S \pi^0}$
11. $\mathbf{D^0 \to K^0_S \pi^+ \pi^-}$
12. $\mathbf{D^0 \to K^0_S \pi^+ \pi^- \pi^0}$
13. $\mathbf{D^0 \to K^- K^+}$
14. $D^0 \to K^- K^+ \pi^0$
15. $\mathbf{D^0 \to K^- K^+ K^0_S}$

Only hadronic D^0 candidates with an invariant mass between 1.7GeV and 1.95GeV are considered. If not stated otherwise D^0 always refers to a D^0, which decayed in one of the hadronic decay modes listed above, throughout the text.

C.1.3.2. D^0 meson from semileptonic decay-channels

Due to the different kinematic properties, D^0 mesons from semileptonic decay-channels are processed separately by the `FEI`. The following semileptonic channels are used to reconstruct D^0 mesons:

1. $D^0 \to K^- e^+ \nu$
2. $D^0 \to K^- \mu^+ \nu$
3. $D^0 \to K^- \pi^0 e^+ \nu$
4. $D^0 \to K^- \pi^0 \mu^+ \nu$
5. $D^0 \to K^0_S \pi^- e^+ \nu$
6. $D^0 \to K^0_S \pi^- \mu^+ \nu$

There are no further cuts applied to semileptonic D^0 candidates.

C.1.3.3. D^0 meson from decay-channels with K^0_L

Due to the different kinematic properties, D^0 mesons from decay-channels which include a K^0_L are processed separately by the `FEI`. The following channels are used to reconstruct D^0 mesons:

1. $D^0 \to K^0_L \pi^0$
2. $D^0 \to K^0_L \pi^+ \pi^-$
3. $D^0 \to K^0_L \pi^+ \pi^- \pi^0$
4. $D^0 \to K^- K^+ K^0_L$

There are no further cuts applied to D^0 candidates with K^0_L.

C.1.3.4. D^+ meson from hadronic decay-channels

The following hadronic channels are used to reconstruct D^+ mesons:

1. $\mathbf{D^+ \to K^- \pi^+ \pi^+}$
2. $\mathbf{D^+ \to K^- \pi^+ \pi^+ \pi^0}$
3. $\mathbf{D^+ \to K^- K^+ \pi^+}$
4. $\mathbf{D^0 \to K^- K^+ \pi^+ \pi^0}$
5. $D^+ \to \pi^+ \pi^0$
6. $D^+ \to \pi^+ \pi^+ \pi^-$
7. $D^+ \to \pi^+ \pi^+ \pi^- \pi^0$
8. $\mathbf{D^+ \to K^0_S \pi^+}$
9. $\mathbf{D^+ \to K^0_S \pi^+ \pi^0}$
10. $\mathbf{D^+ \to K^0_S \pi^+ \pi^+ \pi^-}$
11. $D^+ \to K^+ K^0_S K^0_S$

Only hadronic D^+ candidates with an invariant mass between 1.7GeV and 1.95GeV are considered. If not stated otherwise D^+ always refers to a D^+, which decayed in one of the hadronic decay modes listed above, throughout the text.

C.1.3.5. D^+ meson from semileptonic decay-channels

Due to the different kinematic properties, D^+ mesons from semileptonic decay-channels are processed separately by the `FEI`. The following semileptonic channels are used to reconstruct D^+ mesons:

1. $D^+ \to K^0_S e^+ \nu$
2. $D^+ \to K^0_S \mu^+ \nu$
3. $D^+ \to K^- \pi^+ e^+ \nu$
4. $D^+ \to K^- \pi^+ \mu^+ \nu$

There are no further cuts applied to semileptonic D^+ candidates.

C.1.3.6. D^+ meson from decay-channels with K^0_L

Due to the different kinematic properties, D^+ mesons from decay-channels which include a K^0_L are processed separately by the `FEI`. The following channels are used to reconstruct D^+ mesons:

1. $D^+ \to K^0_L \pi^+$
2. $D^+ \to K^0_L \pi^+ \pi^0$
3. $D^+ \to K^0_L \pi^+ \pi^+ \pi^-$
4. $D^+ \to K^+ K^0_L K^0_S$
5. $D^+ \to K^+ K^0_L K^0_L$

There are no further cuts applied to D^+ candidates with K^0_L.

C.1.3.7. J/ψ meson

The following hadronic channels are used to reconstruct J/ψ mesons:

1. $\mathbf{J/\psi \to e^+ e^-}$
2. $\mathbf{J/\psi \to \mu^+ \mu^-}$

Only hadronic J/ψ candidates with an invariant mass between 2.8GeV and 3.5GeV are considered.

C.1.3.8. D^{+*} meson from decay-channels

The following channels are used to reconstruct D^{+*} mesons:

1. $\mathbf{D^{+*} \to D^0 \pi^+}$
2. $\mathbf{D^{+*} \to D^+ \pi^0}$
3. $D^{+*} \to D^+ \gamma$

Only D^{+*} candidates with a released energy in the decay between 0GeV and 0.3GeV are considered. Separate particle lists are created for D^{+*} meson reconstructed from hadronic, semileptonic and klong D mesons. If not stated otherwise D^{+*} always refers to a D^{+*}, which decayed in one of the hadronic decay modes listed above, throughout the text.

C.1.3.9. $D^*(2010)^0$ meson from decay-channels

The following channels are used to reconstruct $D^*(2010)^0$ mesons:

1. $\mathbf{D^*(2010)^0 \to D^0 \pi^0}$
2. $\mathbf{D^*(2010)^0 \to D^0 \gamma}$

Only $D^*(2010)^0$ candidates with a released energy in the decay between 0GeV and 0.3GeV are considered. Separate particle lists are created for $D^*(2010)^0$ meson reconstructed from hadronic, semileptonic and klong D mesons. If not stated otherwise $D^*(2010)^0$ always refers to a $D^*(2010)^0$, which decayed in one of the hadronic decay modes listed above, throughout the text.

C.1.3.10. D_s^+ meson from hadronic decay-channels

The following hadronic channels are used to reconstruct D_s^+ mesons:

1. $\mathbf{D_s^+ \to K^+ K_S^0}$
2. $\mathbf{D_s^+ \to K^+ \pi^+ \pi^-}$
3. $\mathbf{D_s^+ \to K^+ K^- \pi^+}$
4. $\mathbf{D_s^+ \to K^+ K^- \pi^+ \pi^0}$
5. $\mathbf{D_s^+ \to K^+ K_S^0 \pi^+ \pi^-}$
6. $\mathbf{D_s^+ \to K^- K_S^0 \pi^+ \pi^+}$
7. $\mathbf{D_s^+ \to K^+ K^- \pi^+ \pi^+ \pi^-}$
8. $\mathbf{D_s^+ \to \pi^+ \pi^+ \pi^-}$
9. $D_s^+ \to K_S^0 \pi^+$
10. $D_s^+ \to K_S^0 \pi^+ \pi^0$

Only hadronic D_s^+ candidates with an invariant mass between 1.68GeV and 2.1GeV are considered. If not stated otherwise D_s^+ always refers to a D_s^+, which decayed in one of the hadronic decay modes listed above, throughout the text.

C.1.3.11. D_s^+ meson from decay-channels with K_L^0

Due to the different kinematic properties, D_s^+ mesons from decay-channels which include a K_L^0 are processed separately by the `FEI`. The following channels are used to reconstruct D_s^+ mesons:

1. $D_s^+ \to K^+ K_L^0$
2. $D_s^+ \to K^+ K_L^0 \pi^+ \pi^-$
3. $D_s^+ \to K^- K_L^0 \pi^+ \pi^+$
4. $D_s^+ \to K_L^0 \pi^+$
5. $D_s^+ \to K_L^0 \pi^+ \pi^0$

There are no further cuts applied to D_s^+ candidates with K_L^0.

C.1.3.12. D_s^{+*} meson from decay-channels

The following channels are used to reconstruct D_s^{+*} mesons:

1. $\mathbf{D_s^{+*} \to D_s^+ \gamma}$
2. $D_s^{+*} \to D_s^+ \pi^0$

Only D_s^{+*} candidates with a released energy in the decay between 0GeV and 0.3GeV are considered. Separate particle lists are created for D_s^{+*} meson reconstructed from hadronic D_s^+ and klong D_s^+ mesons. If not stated otherwise D_s^{+*} always refers to a D_s^{+*}, which decayed in one of the hadronic decay modes listed above, throughout the text.

C.1.4. B mesons

All B mesons have the same input features for the multivariate classifiers: kinematic and vertex variables excluding quantities correlated to the beam-constrained mass; kinematic, tracking and vertex variables of all daughters evaluated in the rest frame of the B meson; the decay mode identifier and signal probability σ of all daughters; and the position of the candidate in the best candidate ranking of the pre-cut.

C.1.4.1. B^+ meson from hadronic decay-channels

The following hadronic channels are used to reconstruct B^+ mesons:

1. $\mathbf{B^+ \to \overline{D}^0 \pi^+}$
2. $\mathbf{B^+ \to \overline{D}^0 \pi^+ \pi^0}$
3. $B^+ \to \overline{D}^0 \pi^+ \pi^0 \pi^0$
4. $\mathbf{B^+ \to \overline{D}^0 \pi^+ \pi^+ \pi^-}$
5. $B^+ \to \overline{D}^0 \pi^+ \pi^+ \pi^- \pi^0$
6. $B^+ \to \overline{D}^0 D^+$
7. $B^+ \to \overline{D}^0 D^+ K_S^0$
8. $B^+ \to \overline{D}^{0*} D^+ K_S^0$
9. $B^+ \to \overline{D}^0 D^{+*} K_S^0$
10. $B^+ \to \overline{D}^{0*} D^{+*} K_S^0$
11. $B^+ \to \overline{D}^0 D^0 K^+$
12. $B^+ \to \overline{D}^{0*} D^0 K^+$
13. $B^+ \to \overline{D}^0 D^*(2010)^0 K^+$
14. $B^+ \to \overline{D}^{0*} D^*(2010)^0 K^+$
15. $\mathbf{B^+ \to D_s^+ \overline{D}^0}$
16. $\mathbf{B^+ \to \overline{D}^{0*} \pi^+}$
17. $\mathbf{B^+ \to \overline{D}^{0*} \pi^+ \pi^0}$
18. $B^+ \to \overline{D}^{0*} \pi^+ \pi^0 \pi^0$
19. $\mathbf{B^+ \to \overline{D}^{0*} \pi^+ \pi^+ \pi^-}$
20. $\mathbf{B^+ \to \overline{D}^{0*} \pi^+ \pi^+ \pi^- \pi^0}$
21. $\mathbf{B^+ \to D_s^{+*} \overline{D}^0}$
22. $\mathbf{B^+ \to D_s^+ \overline{D}^{0*}}$
23. $\mathbf{B^+ \to \overline{D}^0 K^+}$
24. $\mathbf{B^+ \to D^- \pi^+ \pi^+}$
25. $B^+ \to D^- \pi^+ \pi^+ \pi^0$
26. $\mathbf{B^+ \to J/\psi K^+}$
27. $\mathbf{B^+ \to J/\psi K^+ \pi^+ \pi^-}$
28. $\mathbf{B^+ \to J/\psi K^+ \pi^0}$
29. $\mathbf{B^+ \to J/\psi K_S^0 \pi^+}$

Only hadronic B^+ candidates with an beam-constrained mass above 5.2GeV and an absolute deviation of the energy from the beam energy below 0.5GeV are considered.

C.1.4.2. B^+ meson from semileptonic decay-channels

Due to the different kinematic properties, B^+ mesons from semileptonic decay-channels are processed separately by the `FEI`. The following semileptonic channels are used to reconstruct B^+ mesons:

1. $B^+ \to \overline{D}^0 e^+ \nu$
2. $B^+ \to \overline{D}^0 \mu^+ \nu$
3. $B^+ \to \overline{D}^{0*} e^+ \nu$
4. $B^+ \to \overline{D}^{0*} \mu^+ \nu$
5. $B^+ \to D^- \pi^+ e^+ \nu$
6. $B^+ \to D^- \pi^+ \mu^+ \nu$
7. $B^+ \to D^{+*} \pi^+ \mu^+ \nu$
8. $B^+ \to D^{+*} \pi^+ e^+ \nu$

In addition all hadronic channels listed in Section C.1.4.1 are used, where one of the hadronic D mesons is replaced by a semileptonic D meson. There are no further cuts applied to semileptonic B^+ candidates.

C.1.4.3. B^+ meson from decay-channels with K_L^0

Due to the different kinematic properties, B^+ mesons from decay-channels which include a K_L^0 are processed separately by the `FEI`. All hadronic channels listed in Section C.1.4.1 are used, where either one of the hadronic D mesons is replaced by a klong D meson, or one of the K_S^0 is replaced by a K_L^0. There are no further cuts applied to B^+ candidates with K_L^0.

C.1.4.4. B^0 meson from hadronic decay-channels

The following hadronic channels are used to reconstruct B^0 mesons:

1. $\mathbf{B^0 \to D^- \pi^+}$
2. $\mathbf{B^0 \to D^- \pi^+ \pi^0}$
3. $B^0 \to D^- \pi^+ \pi^0 \pi^0$
4. $\mathbf{B^0 \to D^- \pi^+ \pi^+ \pi^-}$
5. $B^0 \to D^- \pi^+ \pi^+ \pi^- \pi^0$
6. $B^0 \to \overline{D}^0 \pi^+ \pi^-$
7. $B^0 \to D^- D^0 K^+$
8. $B^0 \to D^- D^*(2010)^0 K^+$
9. $B^0 \to D^{+*} D^0 K^+$
10. $B^0 \to D^{+*} D^*(2010)^0 K^+$
11. $B^0 \to D^- D^+ K_S^0$
12. $B^0 \to D^{+*} D^+ K_S^0$
13. $B^0 \to D^- D^{+*} K_S^0$
14. $B^0 \to D^{+*} D^{+*} K_S^0$
15. $\mathbf{B^0 \to D_s^+ D^-}$
16. $\mathbf{B^0 \to D^{+*} \pi^+}$
17. $\mathbf{B^0 \to D^{+*} \pi^+ \pi^0}$
18. $B^0 \to D^{+*} \pi^+ \pi^0 \pi^0$
19. $\mathbf{B^0 \to D^{+*} \pi^+ \pi^+ \pi^-}$
20. $\mathbf{B^0 \to D^{+*} \pi^+ \pi^+ \pi^- \pi^0}$
21. $\mathbf{B^0 \to D_s^{+*} D^-}$
22. $\mathbf{B^0 \to D_s^+ D^{+*}}$
23. $\mathbf{B^0 \to D_s^{+*} D^{+*}}$
24. $\mathbf{B^0 \to J/\psi K_S^0}$
25. $\mathbf{B^0 \to J/\psi K^+ \pi^-}$
26. $\mathbf{B^0 \to J/\psi K_S^0 \pi^+ \pi^-}$

Only hadronic B^0 candidates with an beam-constrained mass above 5.2GeV and an absolute deviation of the energy from the beam energy below 0.5GeV are considered.

The channel $B^0 \to \overline{D}^0 \pi^0$ was used by the `FR`, but is not yet used in the `FEI` due to unexpected technical restrictions in the `KFitter` algorithm.

C.1.4.5. B^0 meson from semileptonic decay-channels

Due to the different kinematic properties, B^0 mesons from semileptonic decay-channels are processed separately by the `FEI`. The following semileptonic channels are used to reconstruct B^0 mesons:

1. $B^0 \to D^- e^+ \nu$
2. $B^0 \to D^- \mu^+ \nu$
3. $B^0 \to D^{+*} \mu^+ \nu$
4. $B^0 \to D^{+*} e^+ \nu$
5. $B^0 \to \overline{D}^0 \pi^- e^+ \nu$
6. $B^0 \to \overline{D}^0 \pi^- \mu^+ \nu$
7. $B^0 \to \overline{D}^{0*} \pi^- e^+ \nu$
8. $B^0 \to \overline{D}^{0*} \pi^- \mu^+ \nu$

In addition all hadronic channels listed in Section C.1.4.4 are used, where one of the hadronic D mesons is replaced by a semileptonic D meson. There are no further cuts applied to semileptonic B^0 candidates.

C.1.4.6. B^0 meson from decay-channels with K^0_L

Due to the different kinematic properties, B^0 mesons from decay-channels which include a K^0_L are processed separately by the `FEI`. All hadronic channels listed in Section C.1.4.4 are used, where either one of the hadronic D mesons is replaced by a klong D meson, or one of the K^0_S is replaced by a K^0_L. There are no further cuts applied to B^0 candidates with K^0_L.

C.2. FEIR

C.2.1. Belle

Excerpt of the `FEIR` for a generic training using 200*million* Monte Carlo simulated $\Upsilon(4S)$ events.

Table C.2.: Per-particle efficiency before and after the applied pre- and post-cut.

Particle	pre-cut		post-cut		
	user	ranking	absolute	ranking	unique
π^+	0.788	0.788	0.788	0.779	0.773
K^+	0.755	0.755	0.750	0.750	0.746
μ^+	0.892	0.881	0.865	0.853	0.852
e^+	0.775	0.769	0.757	0.757	0.754
γ	0.640	0.640	0.640	0.640	0.632
π^0	0.571	0.433	0.427	0.398	0.397
K^0_S	0.459	0.457	0.444	0.443	0.415
J/ψ	0.086	0.086	0.086	0.086	0.086
D^0	0.140	0.134	0.128	0.125	0.123
D^+	0.104	0.100	0.093	0.092	0.091
D^+_s	0.071	0.069	0.052	0.052	0.051
D^{0*}	0.053	0.051	0.048	0.047	0.047
D^{+*}	0.050	0.049	0.049	0.049	0.047
D^{+*}_s	0.035	0.034	0.031	0.031	0.030
B^0	0.005	0.005	0.005	0.005	0.004

B^+	0.008	0.008	0.008	0.007	0.007
$D^0{}_{SL}$	0.056	0.055	0.052	0.051	0.051
$D^+{}_{SL}$	0.059	0.058	0.052	0.051	0.050
$D^{0*}{}_{SL}$	0.023	0.020	0.015	0.015	0.015
$D^{+*}{}_{SL}$	0.022	0.022	0.020	0.020	0.020
$B^0{}_{SL}$	0.022	0.021	0.021	0.021	0.020
$B^+{}_{SL}$	0.020	0.019	0.018	0.018	0.018

Table C.3.: Per-particle purity before and after the applied pre- and post-cut.

Particle	pre-cut		post-cut		
	user	ranking	absolute	ranking	unique
π^+	0.639	0.639	0.696	0.713	0.707
K^+	0.137	0.137	0.394	0.394	0.392
μ^+	0.046	0.047	.086	0.094	0.094
e^+	0.050	0.052	0.193	0.195	0.194
γ	0.587	0.587	0.652	0.652	0.645
π^0	0.044	0.102	0.133	0.187	0.186
K^0_S	0.067	0.068	0.210	0.210	0.197
J/ψ	0.109	0.109	0.207	0.207	0.207
D^0	0.001	0.002	0.018	0.027	0.026
D^+	0.000	0.001	0.015	0.016	0.015
D^+_s	0.000	0.000	0.014	0.014	0.014
D^{0*}	0.000	0.001	0.011	0.012	0.011
D^{+*}	0.001	0.001	0.051	0.051	0.050
D^{+*}_s	0.000	0.000	0.010	0.010	0.009
B^0	0.000	0.000	0.000	0.001	0.001
B^+	0.000	0.000	0.000	0.001	0.001
$D^0{}_{SL}$	0.001	0.001	0.012	0.014	0.014
$D^+{}_{SL}$	0.001	0.001	0.007	0.007	0.007
$D^{0*}{}_{SL}$	0.000	0.000	0.005	0.005	0.005
$D^{+*}{}_{SL}$	0.000	0.000	0.012	0.012	0.012
$B^0{}_{SL}$	0.000	0.000	0.000	0.001	0.001
$B^+{}_{SL}$	0.000	0.000	0.000	0.001	0.001

C.2.2. Belle II

Excerpt of the `FEIR` for training using 180*million* Monte Carlo simulated $\Upsilon(4S)$ events.

Table C.4.: Per-particle efficiency before and after the applied pre- and post-cut.

Particle	pre-cut		post-cut		
	user	ranking	absolute	ranking	unique
π^+	0.757	0.757	0.756	0.747	0.737
K^+	0.793	0.793	0.790	0.790	0.782
μ^+	0.854	0.844	0.818	0.813	0.809
e^+	0.728	0.726	0.716	0.715	0.708
γ	0.579	0.579	0.579	0.579	0.577
π^0	0.349	0.342	0.341	0.326	0.325
K_S^0	0.425	0.425	0.269	0.269	0.263
J/ψ	0.091	0.091	0.091	0.091	0.090
D^0	0.145	0.139	0.134	0.129	0.127
D^+	0.092	0.088	0.082	0.082	0.080
D_s^+	0.063	0.060	0.054	0.053	0.052
D^{0*}	0.042	0.038	0.034	0.034	0.033
D^{+*}	0.047	0.047	0.046	0.046	0.045
D_s^{+*}	0.038	0.036	0.031	0.031	0.031
B^0	0.004	0.004	0.004	0.004	0.004
B^+	0.007	0.007	0.007	0.007	0.007
${D^0}_{SL}$	0.064	0.063	0.060	0.059	0.058
${D^+}_{SL}$	0.103	0.102	0.096	0.093	0.090
${D^{0*}}_{SL}$	0.019	0.015	0.011	0.011	0.011
${D^{+*}}_{SL}$	0.023	0.023	0.021	0.021	0.021
${B^0}_{SL}$	0.023	0.022	0.022	0.021	0.020
${B^+}_{SL}$	0.017	0.016	0.016	0.015	0.015

Table C.5.: Per-particle purity before and after the applied pre- and post-cut.

Particle	pre-cut		post-cut		
	user	ranking	absolute	ranking	unique
π^+	0.606	0.606	0.705	0.734	0.723
K^+	0.127	0.127	0.500	0.500	0.495

μ^+	0.042	0.046	0.106	0.112	0.111
e^+	0.050	0.055	0.213	0.216	0.214
γ	0.570	0.570	0.644	0.644	0.642
π^0	0.144	0.164	0.181	0.228	0.228
K_S^0	0.025	0.025	0.146	0.146	0.143
J/ψ	0.182	0.182	0.233	0.233	0.232
D^0	0.001	0.003	0.018	0.028	0.027
D^+	0.000	0.001	0.014	0.016	0.015
D_s^+	0.000	0.000	0.011	0.011	0.011
D^{0*}	0.000	0.001	0.007	0.008	0.007
D^{+*}	0.001	0.001	0.053	0.053	0.051
D_s^{+*}	0.000	0.001	0.007	0.007	0.007
B^0	0.000	0.001	0.001	0.001	0.001
B^+	0.000	0.000	0.000	0.001	0.001
${D^0}_{SL}$	0.001	0.002	0.014	0.016	0.016
${D^+}_{SL}$	0.002	0.002	0.009	0.012	0.012
${D^{0*}}_{SL}$	0.000	0.000	0.005	0.005	0.005
${D^{+*}}_{SL}$	0.000	0.000	0.015	0.015	0.015
${B^0}_{SL}$	0.000	0.000	0.000	0.001	0.001
${B^+}_{SL}$	0.000	0.000	0.000	0.001	0.001

C.3. Validation

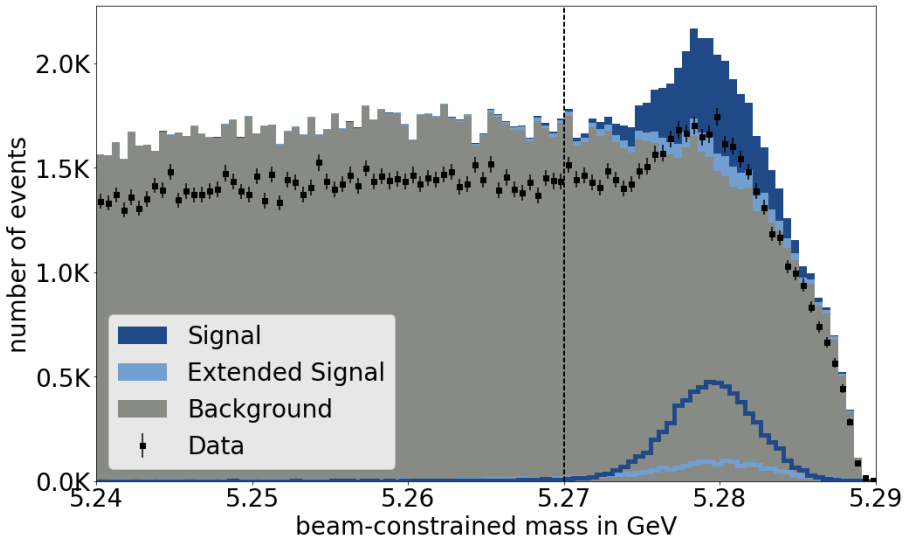
This section contains additional results of the validation of the `FEI` for neutral B^0 mesons.

C.3.1. Influence of the event multiplicity

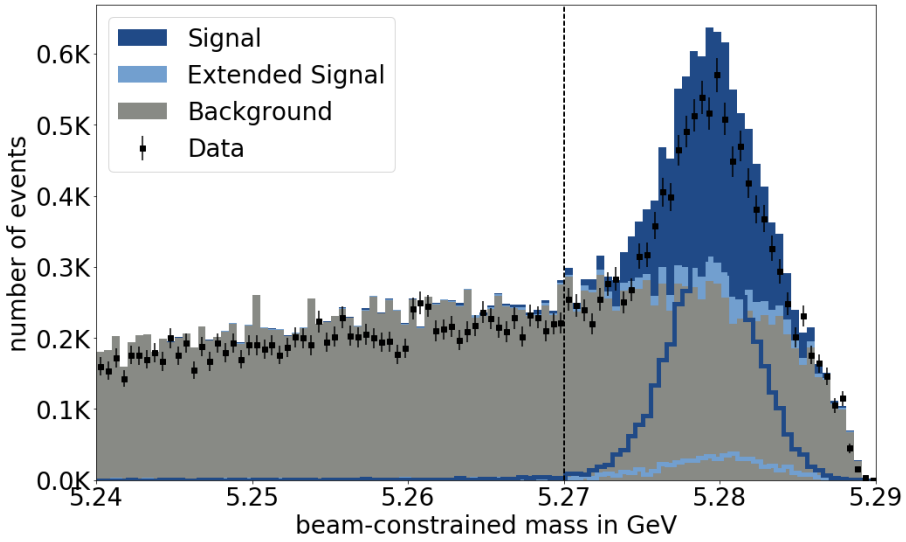
This section contains additional results of the investigation on the influence of the maximum number of reconstructed tracks per event allowed during the training and/or application of the `FEI`.

Only the best B_{tag} candidate per event was considered, and additional cuts on the beam-constrained mass and $\cos\Theta_{BD\ell}$ [1] were applied.

[1] The angle between the Dℓ system and the flight direction of the B meson.

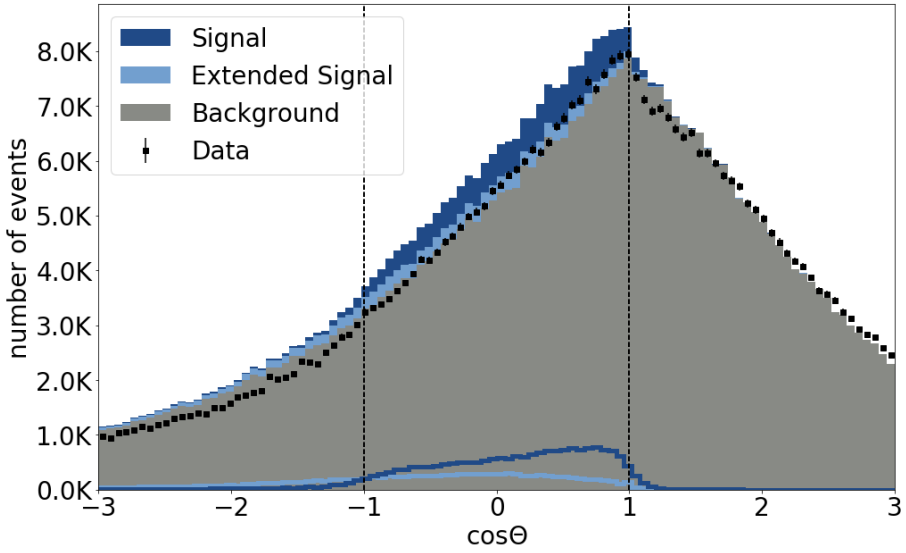


(a) A loose cut on the `SignalProbability` was performed.

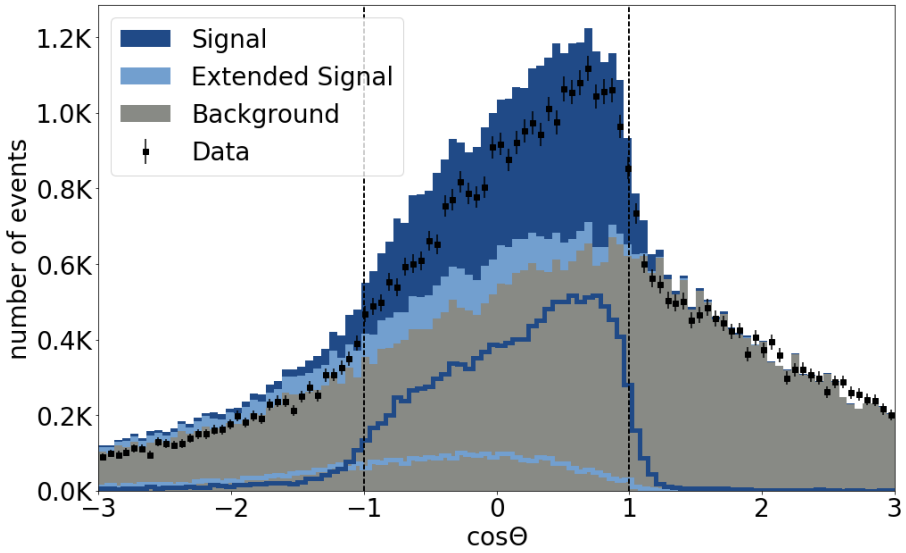


(b) A tight cut on the `SignalProbability` was performed.

Figure C.1.: **Belle** The beam-constrained mass of the best hadronically tagged B^0 meson candidate per event. The spectrum of the different signal-definitions is shown.

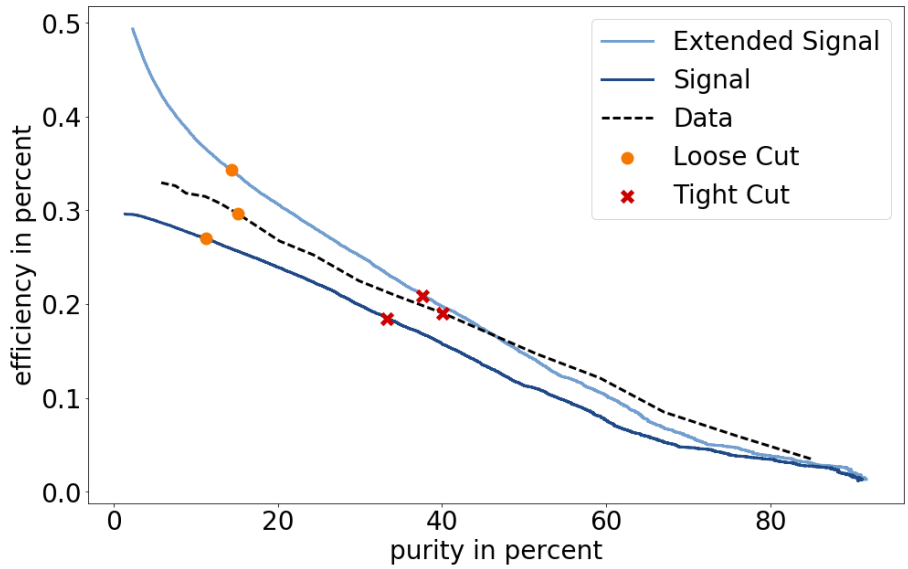


(a) A loose cut on the `SignalProbability` was performed.



(b) A tight cut on the `SignalProbability` was performed.

Figure C.2.: **Belle** $\cos\Theta_{\mathrm{BD}\ell}$ of the best semileptonically tagged B^0 meson candidate per event. The spectrum of the different signal-definitions is shown.



(a) Hadronic Tag

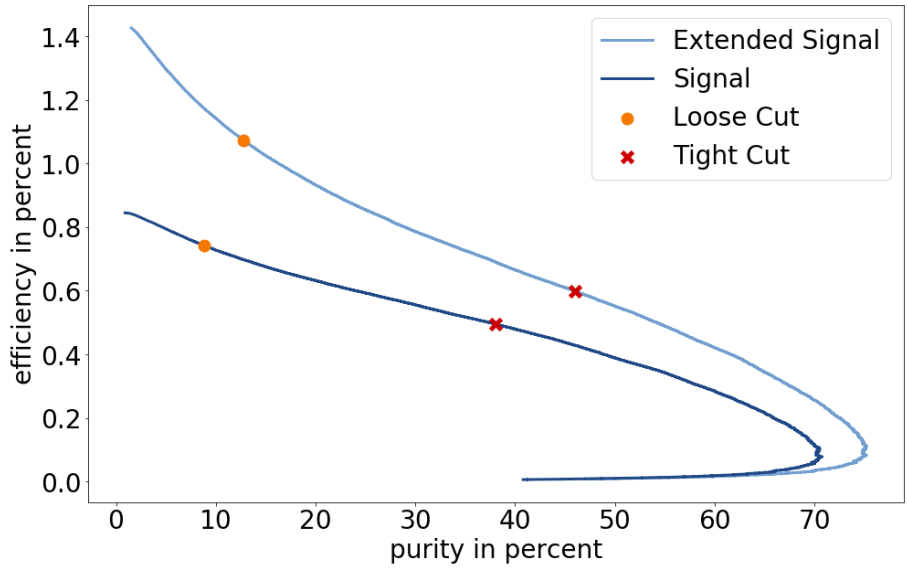
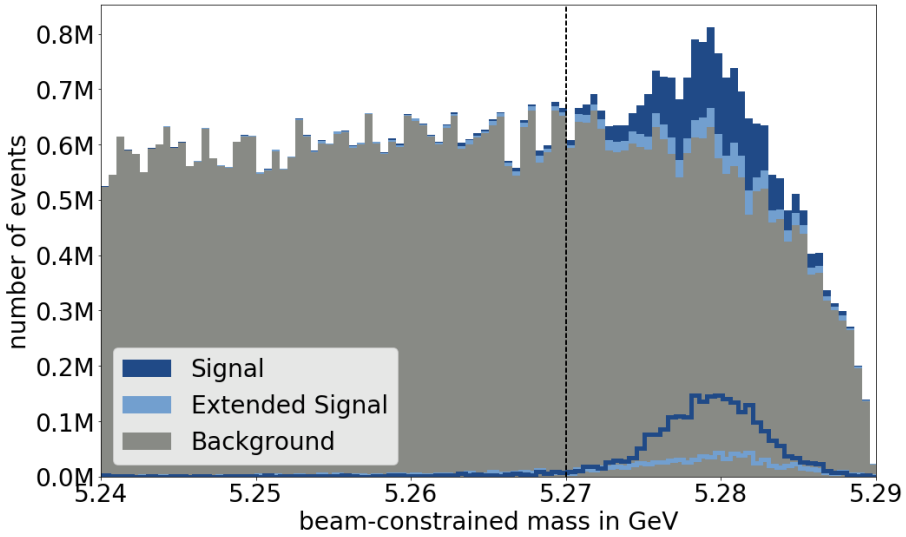
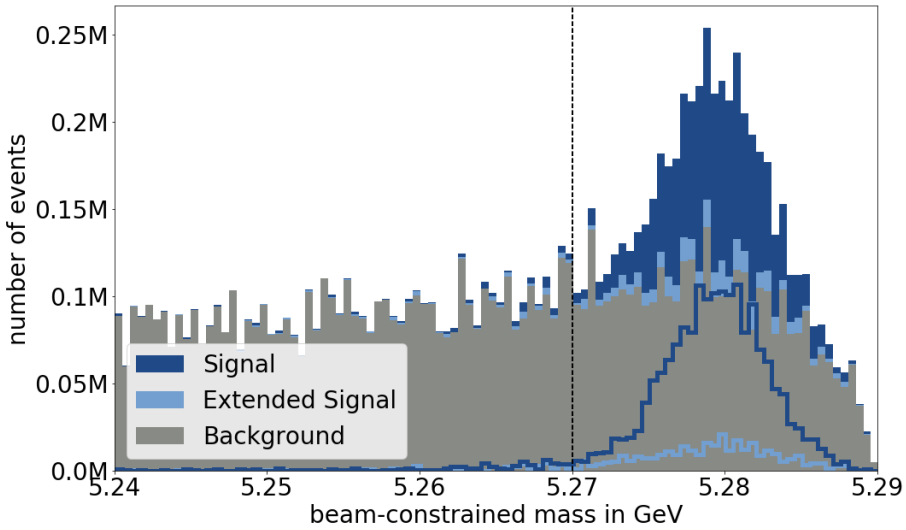


Figure C.3.: **Belle** Receiver operating characteristic (ROC) curve of tagged B^0 for the signal and extended signal-definition.

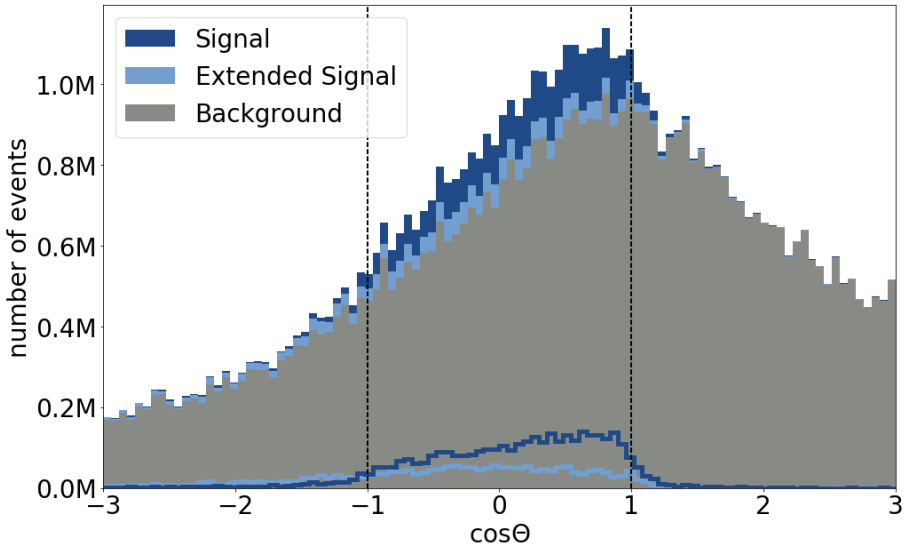


(a) A loose cut on the `SignalProbability` was performed.

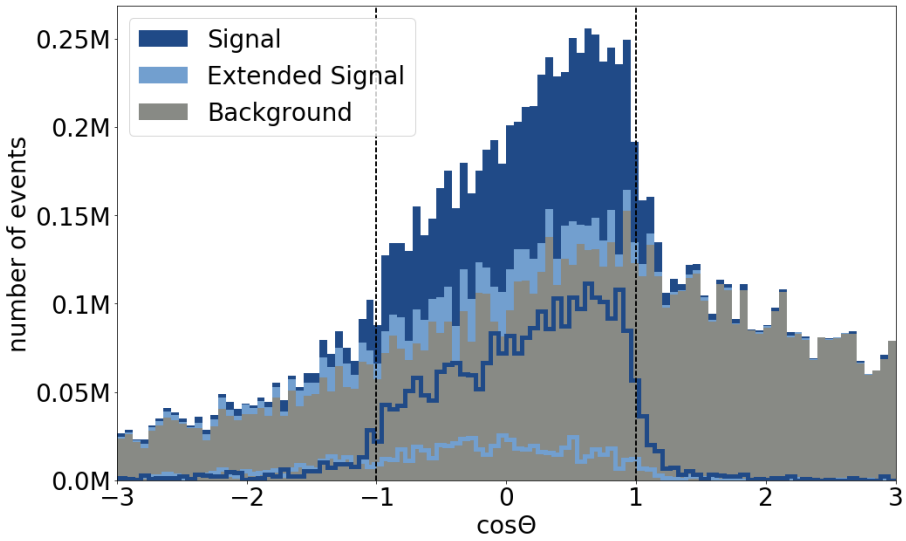


(b) A tight cut on the `SignalProbability` was performed.

Figure C.4.: **Belle II** The beam-constrained mass of the best hadronically tagged B^0 meson candidate per event. The spectrum of the different signal-definitions is shown.

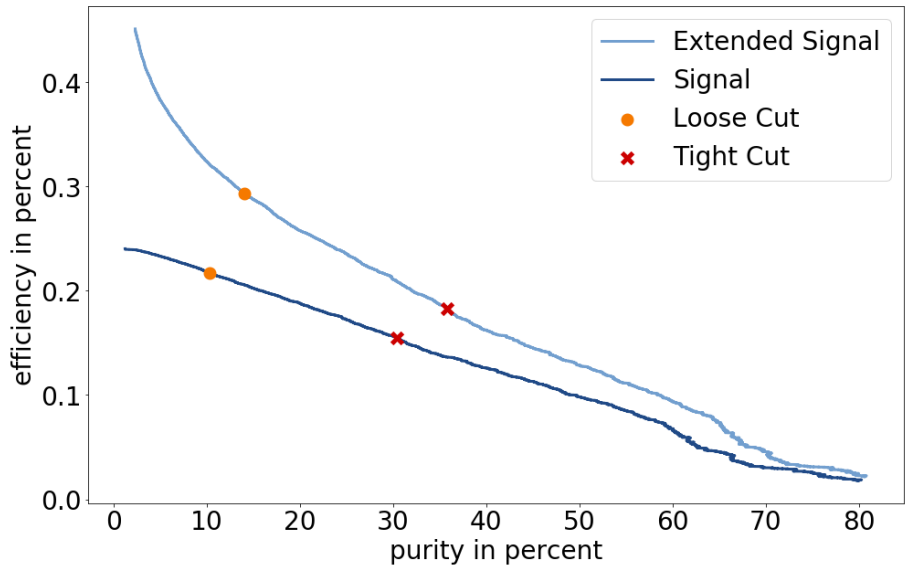


(a) A loose cut on the `SignalProbability` was performed.



(b) A tight cut on the `SignalProbability` was performed.

Figure C.5.: **Belle II** $\cos\Theta_{\mathrm{BD}\ell}$ of the best semileptonically tagged B^0 meson candidate per event. The spectrum of the different signal-definitions is shown..



(a) Hadronic Tag

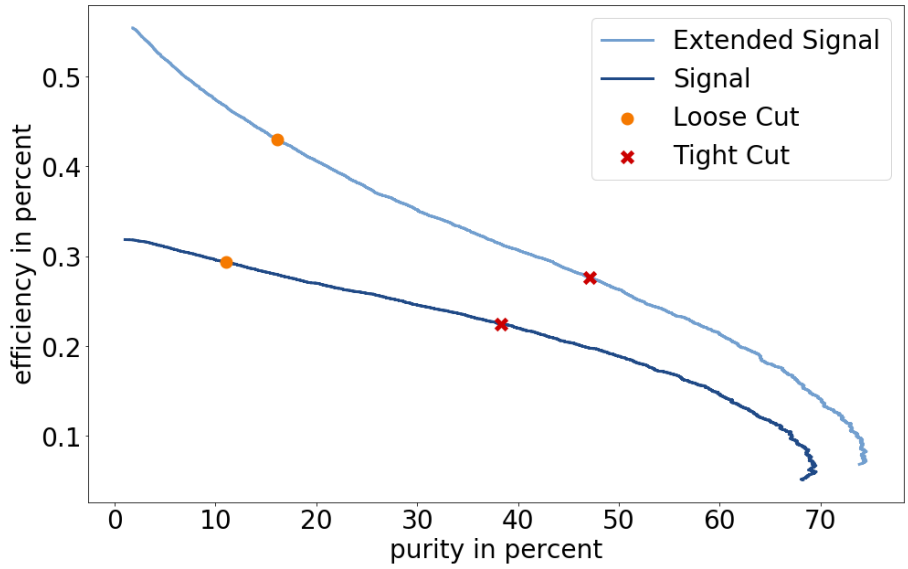
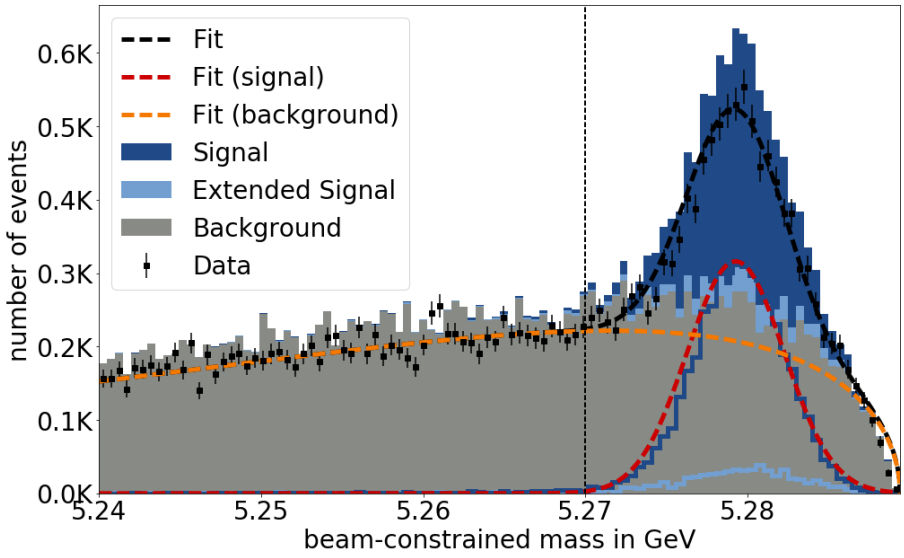
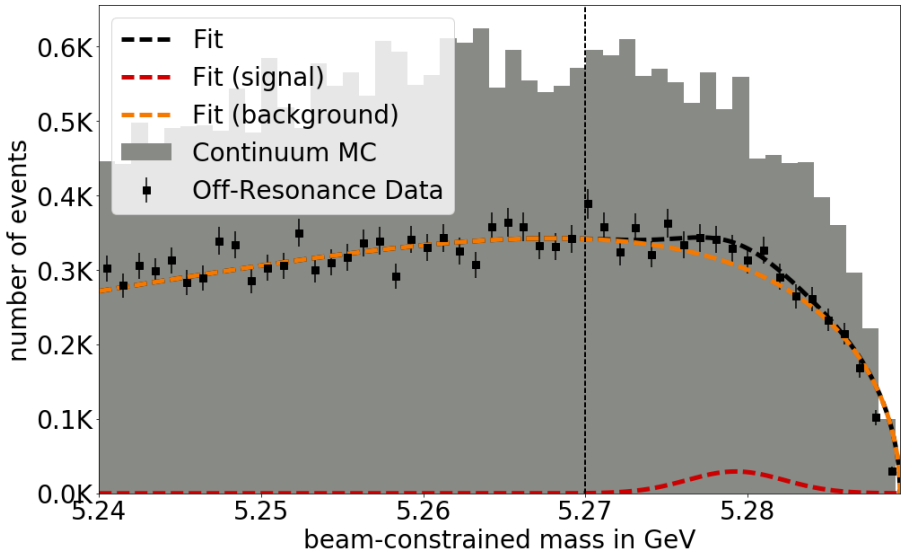


Figure C.6.: **Belle II** Receiver operating characteristic (ROC) curve of tagged B^0 for the signal and extended signal-definition.



(a) On-resonance: The signal peak on data is smaller. Hence the tag-side efficiency is lower on data compared to the Monte Carlo expectations. On the other hand, the fitted signal fraction is over-estimated due to the peaking background. Both effects cancel each other.



(b) Off-resonance: As expected the fitted signal fraction is nearly zero. Hence nearly no peaking background is observed on continuum data. The off-resonance data was scaled with a factor $\frac{10.58}{10.58-0.06}$.

Figure C.7.: **Belle** The beam-constrained mass of the best hadronically tagged B^0 meson candidate per event. The Monte Carlo expectation is shown as filled histograms. The data is shown as black points with a Poisson uncertainty. The result of the EUML on 10 million recorded Belle events is shown as dashed lines.

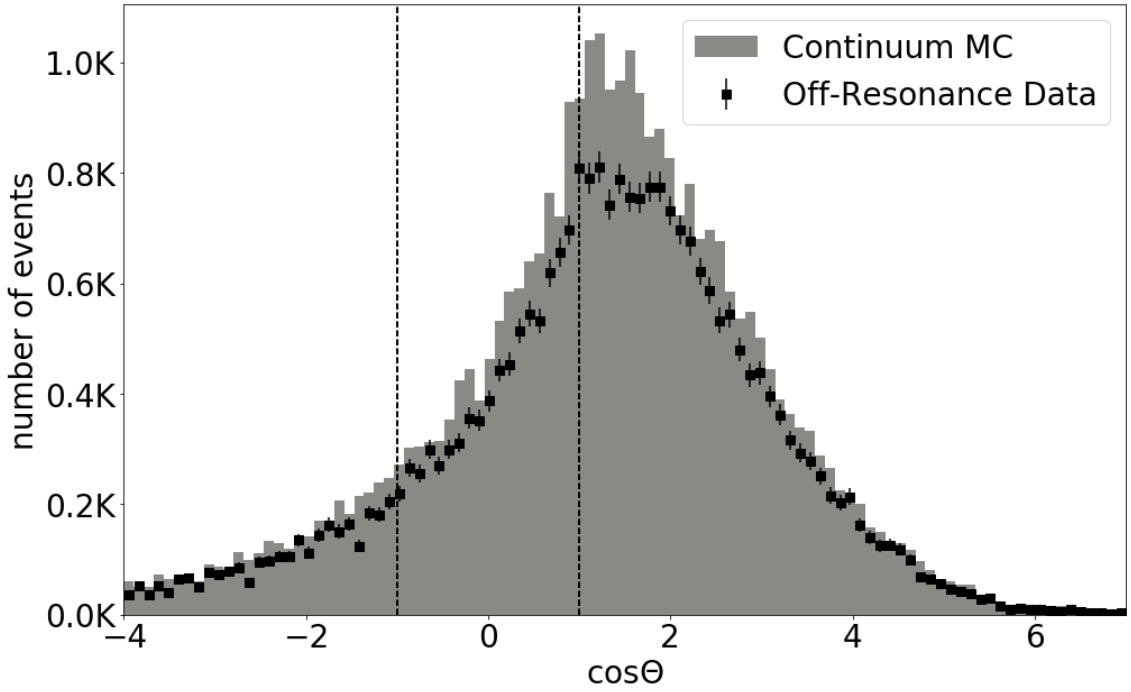


Figure C.8.: **Belle:** $\cos\Theta_{\mathrm{BD}\ell}$ of the best semileptonically tagged B^0 meson candidate per event. The Monte Carlo expectation of the continuum component is shown as a filled histogram. The off-resonance data is shown as black points with a Poisson uncertainty. The off-resonance data-points were calculated using the on-resonance beam-energy.

Hadronically tagged B^0 (see Figure C.9) and semileptonically tagged B^+ (see Figure C.10) show the same behavior as hadronically tagged B^+ (see Figure 4.15).

In contrast, the semileptonically tagged B^0 candidates show an influence of the maximum number of reconstructed tracks per event in the training (see Figure C.11). This anomalous behavior was traced back to invalid classifier outputs in certain decay-channels. The difference can therefore be explained by a technical issue.

C.3.2. Influence of beam-background

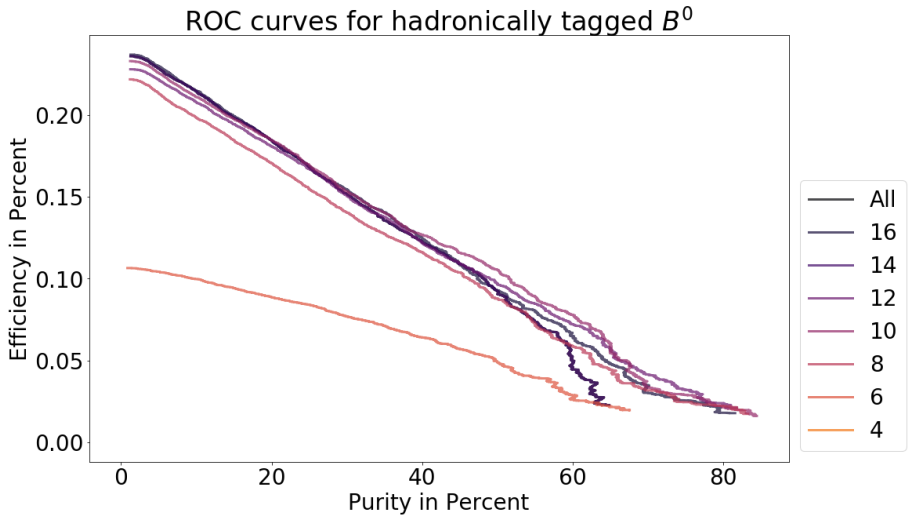
This section contains additional results of the investigation on the influence of beam background on the performance of the `FEI`.

Only the best B_{tag} candidate per event was considered, and additional cuts on the beam-constrained mass and $\cos\Theta_{BD\ell}$ [2] were applied.

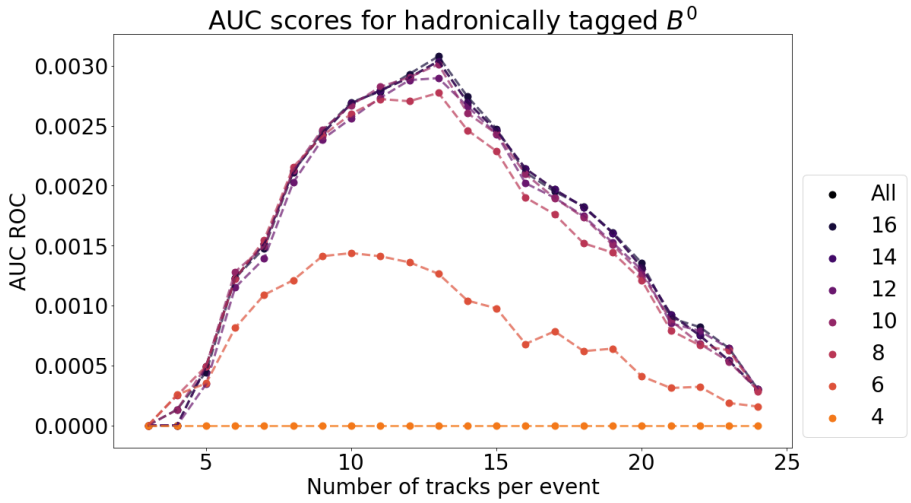
Hadronically tagged B^0 (see Figure C.12) and semileptonically tagged B^+ (see Figure C.13) show the same behavior as hadronically tagged B^+ (see Figure 4.17).

In contrast, the semileptonically tagged B^0 candidates show a different behavior (see Figure C.14). This anomalous behavior was traced back to invalid classifier outputs in certain decay-channels. The difference can therefore be explained by a technical issue.

[2] The angle between the $D\ell$ system and the flight direction of the B meson.

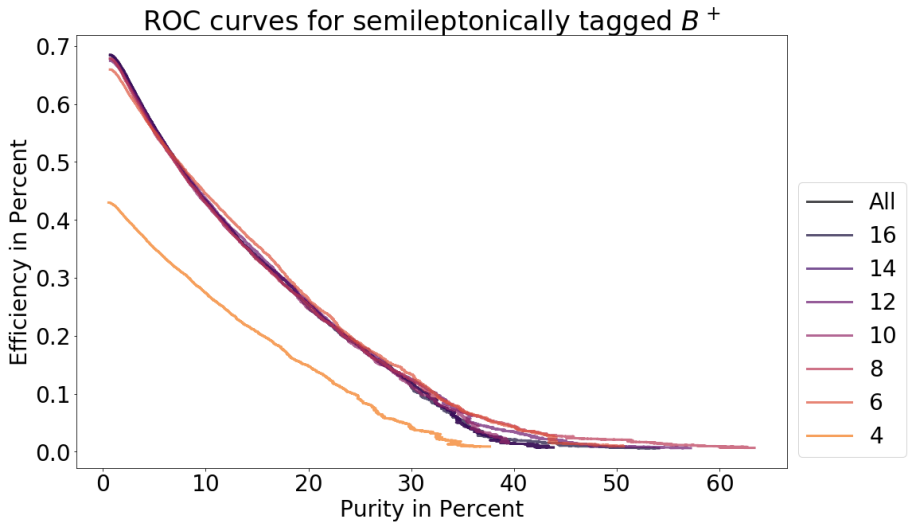


(a) Receiver operating characteristic (ROC) curves applied to **all** events without a cut on the event multiplicity.

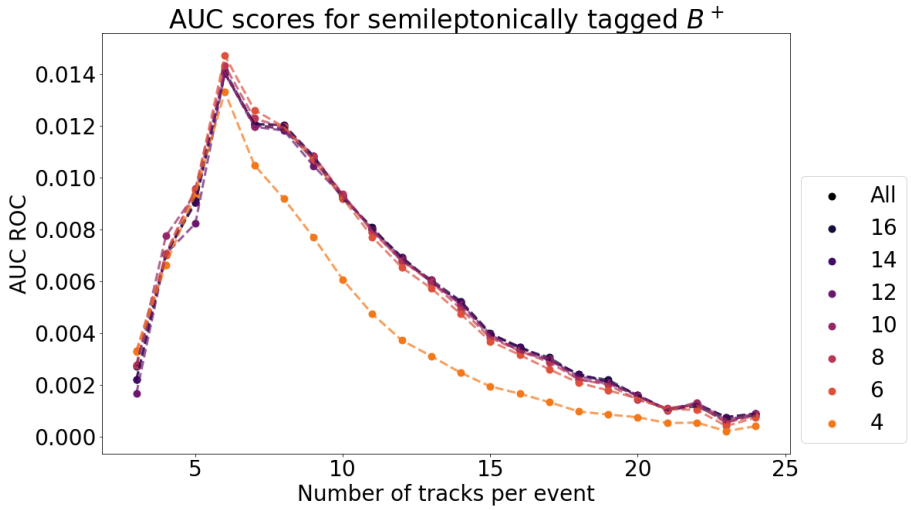


(b) The area under the ROC curve (AUC ROC) depending on the multiplicity of the event.

Figure C.9.: Evaluation of the best hadronically tagged B^0 candidate per event, reconstructed by different configurations of the `FEI`. Each curve corresponds to a generic `FEI` with a cut on the maximum number of N tracks in the event during the **training**.

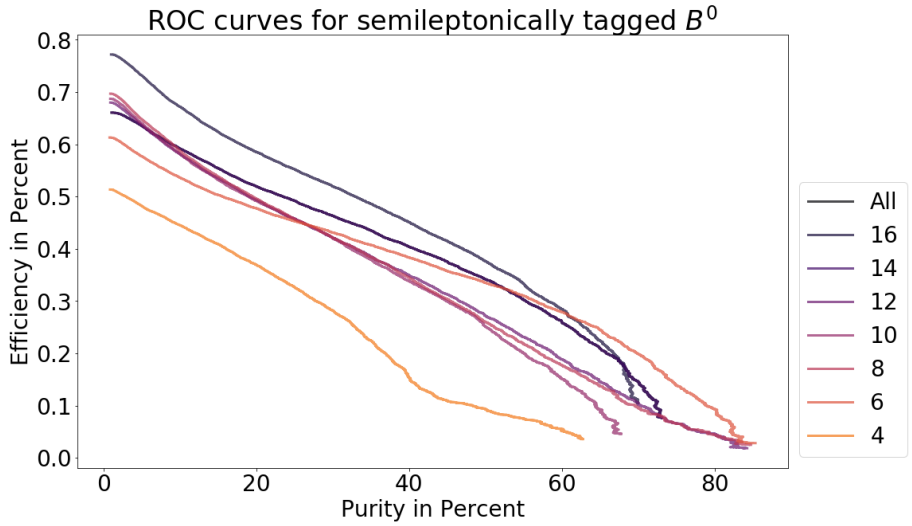


(a) Receiver operating characteristic (ROC) curves applied to **all** events without a cut on the event multiplicity.

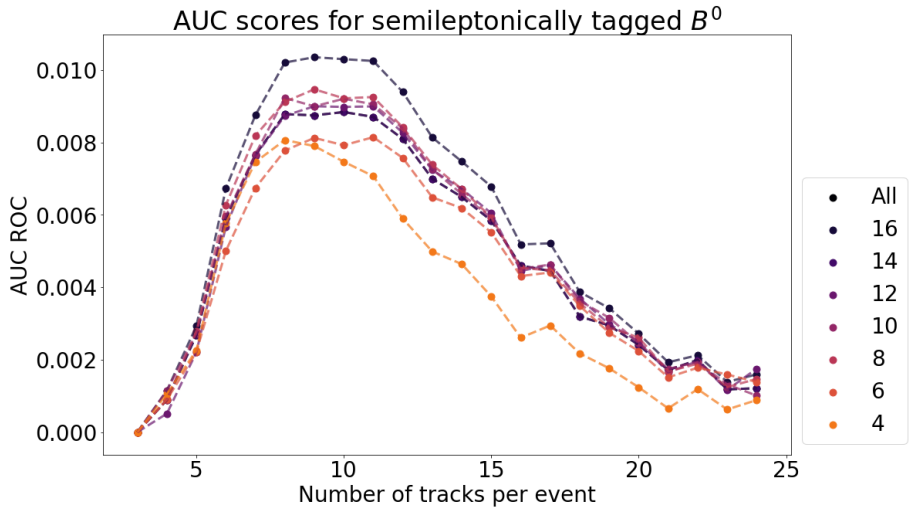


(b) The area under the ROC curve (AUC ROC) depending on the multiplicity of the event.

Figure C.10.: Evaluation of the best semileptonically tagged B^+ candidate per event, reconstructed by different configurations of the `FEI`. Each curve corresponds to a generic `FEI` with a cut on the maximum number of N tracks in the event during the **training**.



(a) Receiver operating characteristic (ROC) curves applied to **all** events without a cut on the event multiplicity.



(b) The area under the ROC curve (AUC ROC) depending on the multiplicity of the event.

Figure C.11.: Evaluation of the best semileptonically tagged B^+ candidate per event, reconstructed by different configurations of the `FEI`. Each curve corresponds to a generic `FEI` with a cut on the maximum number of N tracks in the event during the **training**.

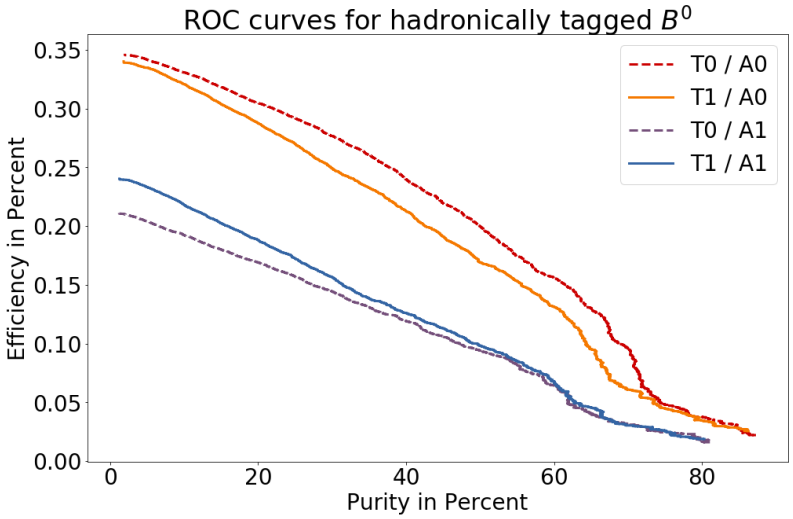


Figure C.12.: Evaluation of the best hadronically tagged B^0 candidate per event, reconstructed by different configurations of the `FEI`. The solid (dashed) curves correspond to a generic `FEI` trained on simulated Belle II events with (without) beam-background. The blue and purple (red and orange) curves correspond to a generic `FEI` applied to simulated Belle II events with (without) beam-background.

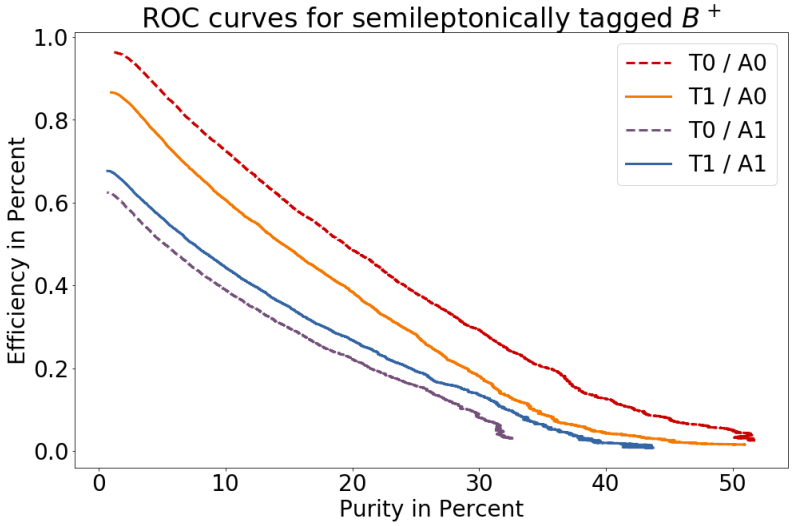


Figure C.13.: Evaluation of the best semileptonically tagged B^+ candidate per event, reconstructed by different configurations of the `FEI`. The solid (dashed) curves correspond to a generic `FEI` trained on simulated Belle II events with (without) beam-background. The blue and purple (red and orange) curves correspond to a generic `FEI` applied to simulated Belle II events with (without) beam-background.

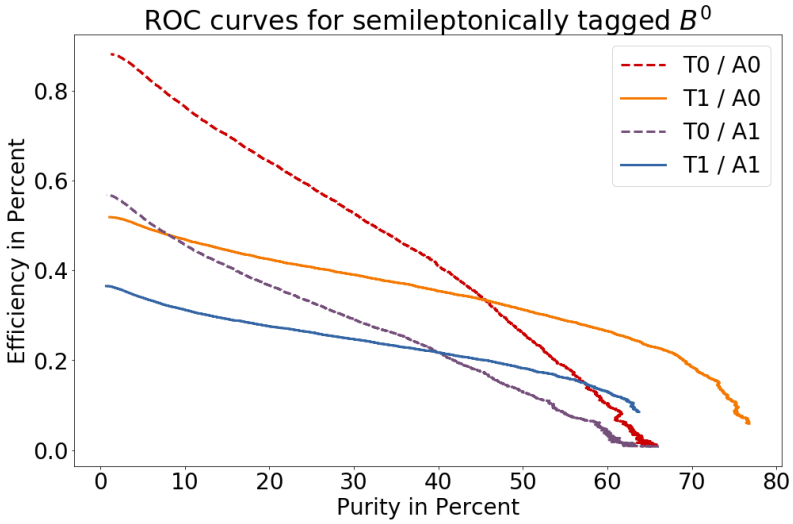


Figure C.14.: Evaluation of the best semileptonically tagged B^0 candidate per event, reconstructed by different configurations of the `FEI`. The solid (dashed) curves correspond to a generic `FEI` trained on simulated Belle II events with (without) beam-background. The blue and purple (red and orange) curves correspond to a generic `FEI` applied to simulated Belle II events with (without) beam-background.

Appendix D

$B \to \tau\nu$

D.1. The $B \to \tau\tau$ puzzle

The decay $B \to \tau\tau$ is extremely suppressed in the Standard Model [97]

$$\mathcal{B}(B^0 \to \tau^-\tau^+)_{SM} = (2.22 \pm 0.19) \cdot 10^{-8}. \tag{D.1}$$

Surprisingly, Belle measured the decay $B \to \tau\tau$ and found a significant 5 sigma excess over the SM expectation [97]

$$\mathcal{B}(B^0 \to \tau^-\tau^+)_{Belle} = (4.39^{+0.80}_{-0.83} \pm 0.45) \cdot 10^{-3}. \tag{D.2}$$

The measurement used the FR for exclusive hadronic tagging and reconstructed the τ particles in one-prong decays. It was not accepted by the collaboration and not published in a peer-reviewed paper.

The BaBar experiment searched for the same decay and set an upper limit compatible with the Belle measurement (see [98])

$$\mathcal{B}(B^0 \to \tau^-\tau^+)_{BaBar} < 4.1 \cdot 10^{-3} \quad @90\%C.L.. \tag{D.3}$$

The measurement used exclusive hadronic tagging and reconstructed the τ particles in one-prong decays. The exclusive hadronic tagging of BaBar relies on a different algorithm than used by Belle.

Finally, the LHCb experiment searched for the same decay and set an upper limit incompatible with the Belle measurement (see [99])

$$\mathcal{B}(B^0 \to \tau^-\tau^+)_{LHCb} < 2.1 \cdot 10^{-3} \quad @95\%C.L.. \tag{D.4}$$

Since LHCb is not located at a B factory, the measurement did not use exclusive hadronic tagging. Instead a high statistics sample was used and the τ was reconstructed in three-prong decays.

The second-most[1] likely[2] explanation for these results is an unaccounted SM background process, which generates a peaking background contribution in one-prong, but not in three-prong decays.

The analyses strategies used by Belle and BaBar (exclusive tagging and one-prong decays) are very similar to the analyses presented in Section 5.1.4. In fact, the Belle measurement would indicate that $B^0 \to \tau^-\tau^+$ is a sizeable peaking background component in the measurement of $B^+ \to \tau^+\nu_\tau$.

The measurement of $B^+ \to \tau^+\nu_\tau$ using the new (independently developed) `BASF2` software applied to the $\Upsilon(4S)$ dataset recorded by Belle, can provide additional insight into this situation. In particular, the encountered large systematic uncertainty due to the continuum description could be a possible explanation.

D.2. Validation

D.2.1. Monte Carlo Study

The background events, which survive the final selection were studied, to identify the most important background processes. The Belle II Monte Carlo matching error flags for the tag-side and signal-side are stated in Table D.1 and Table D.2, respectively.

D.2.2. On-resonance and Off-resonance study

The distributions of E_{ECL} on on-resonance data is shown in this section before and after the final selection. The hadronically tagged τ decay channels are shown in Figure D.1. The semileptonically tagged τ decay channels are shown in Figure D.2.

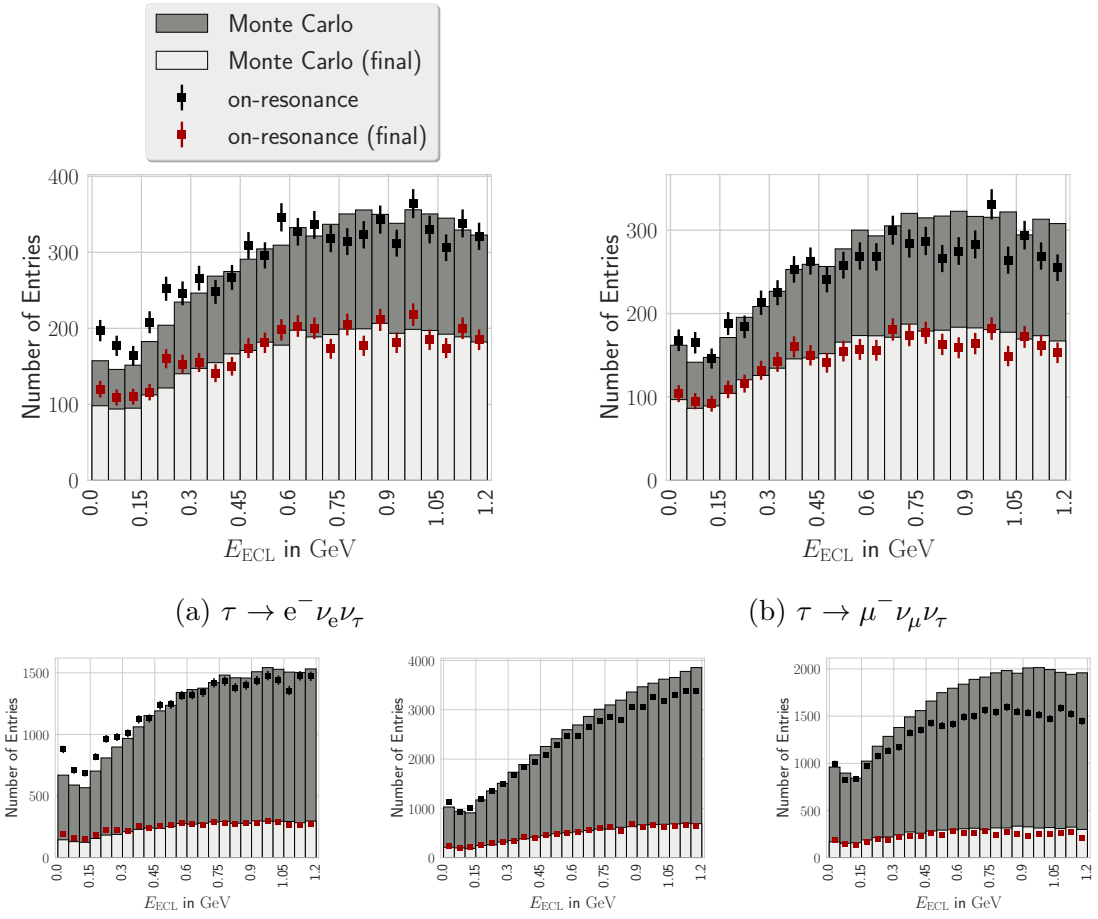
The distributions of E_{ECL} on off-resonance data is shown in this section before and after the final selection. The hadronically tagged τ decay channels are shown in Figure D.3. The semileptonically tagged τ decay channels are shown in Figure D.4.

Table D.1.: **Belle:** Fraction of background events after the final selection which contain a certain Monte Carlo error flag on the tag-side. The last stated digit is not significant.

	Hadronic					Semileptonic				
	e	μ	π	ρ	a_1	e	μ	π	ρ	a_1
Correct	0.34	0.34	0.21	0.18	0.19	0.40	0.42	0.31	0.29	0.29
MissFSR	0.11	0.09	0.37	0.42	0.36	0.10	0.08	0.29	0.25	0.19
MissResonance	0.79	0.80	0.79	0.85	0.80	0.79	0.79	0.82	0.84	0.80
DecayInFlight	0.01	0.01	0.02	0.02	0.02	0.00	0.00	0.01	0.01	0.01
MissNeutrino	0.39	0.39	0.21	0.23	0.23	0.92	0.92	0.88	0.89	0.88
MissGamma	0.44	0.44	0.54	0.66	0.57	0.40	0.40	0.48	0.59	0.45
MissMassiveParticle	0.42	0.43	0.56	0.65	0.60	0.41	0.40	0.52	0.53	0.53
MissKlong	0.24	0.25	0.29	0.34	0.30	0.23	0.22	0.26	0.28	0.24
MissID	0.25	0.24	0.38	0.40	0.37	0.17	0.15	0.18	0.18	0.17
AddedWrongParticle	0.46	0.47	0.59	0.67	0.62	0.43	0.41	0.54	0.54	0.54
InternalError	0.18	0.16	0.18	0.13	0.17	0.08	0.08	0.07	0.06	0.09
MissPHOTOS	0.11	0.09	0.37	0.42	0.36	0.10	0.08	0.29	0.25	0.19

Table D.2.: **Belle:** Fraction of background events after the final selection which contain a certain Monte Carlo error flag on the signal-side. The last stated digit is not significant.

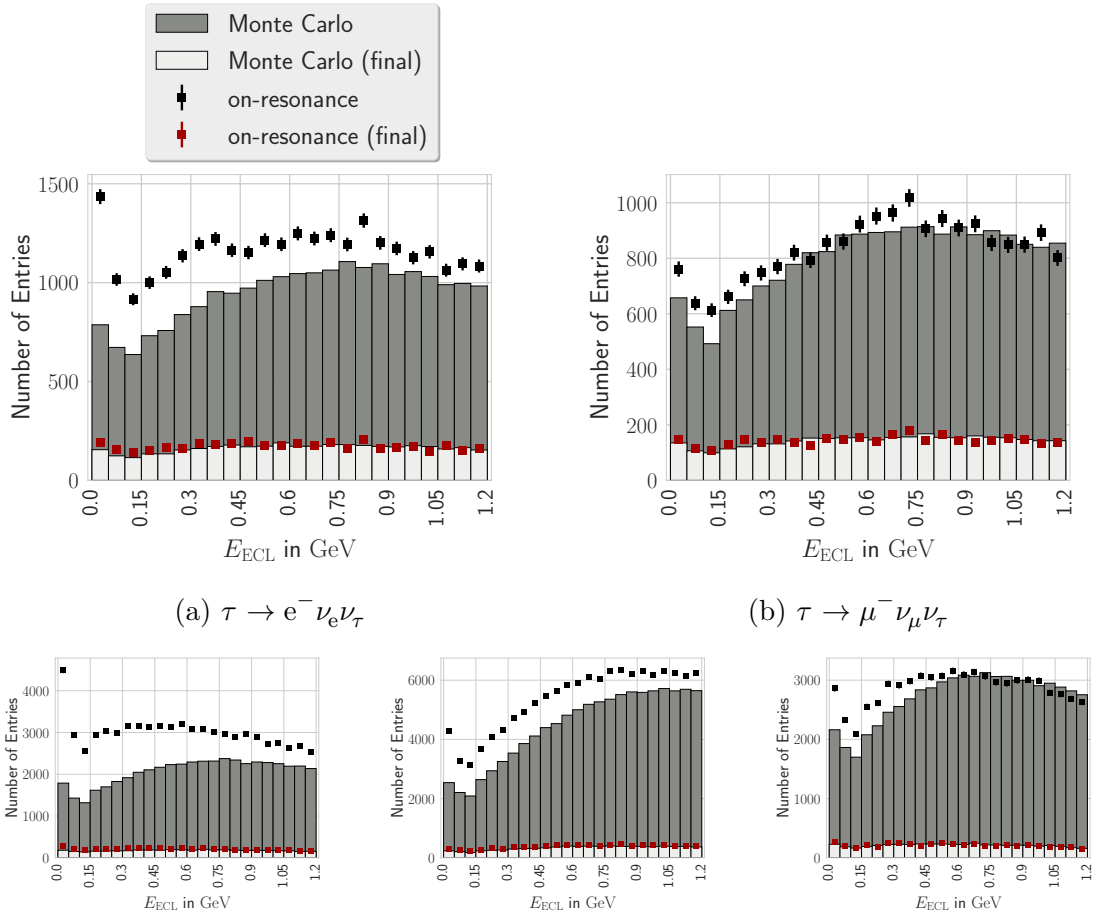
	Hadronic					Semileptonic				
	e	μ	π	ρ	a_1	e	μ	π	ρ	a_1
Correct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MissFSR	0.15	0.11	0.22	0.05	0.05	0.16	0.12	0.12	0.06	0.08
MissResonance	0.97	0.98	0.75	0.33	0.45	0.98	0.98	0.76	0.50	0.60
DecayInFlight	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MissNeutrino	0.96	0.96	0.35	0.19	0.32	0.99	0.98	0.61	0.42	0.55
MissGamma	0.95	0.96	0.68	0.31	0.41	0.90	0.91	0.65	0.44	0.54
MissMassiveParticle	0.99	0.98	0.79	0.33	0.45	0.99	0.99	0.82	0.50	0.60
MissKlong	0.61	0.61	0.42	0.18	0.27	0.64	0.64	0.45	0.28	0.37
MissID	0.03	0.05	0.28	0.17	0.28	0.01	0.03	0.37	0.26	0.36
AddedWrongParticle	0.96	0.97	0.81	0.34	0.45	0.97	0.97	0.84	0.52	0.61
InternalError	0.00	0.01	0.16	0.66	0.55	0.00	0.00	0.12	0.48	0.39
MissPHOTOS	0.15	0.11	0.22	0.05	0.05	0.16	0.12	0.12	0.06	0.08



(a) $\tau \to \mathrm{e}^{-}\nu_{\mathrm{e}}\nu_{\tau}$ (b) $\tau \to \mu^{-}\nu_{\mu}\nu_{\tau}$

(c) $\tau \to \pi^{+}\nu_{\tau}$ (d) $\tau \to \rho\nu_{\tau}$ (e) $\tau \to \mathrm{a}_1\nu_{\tau}$

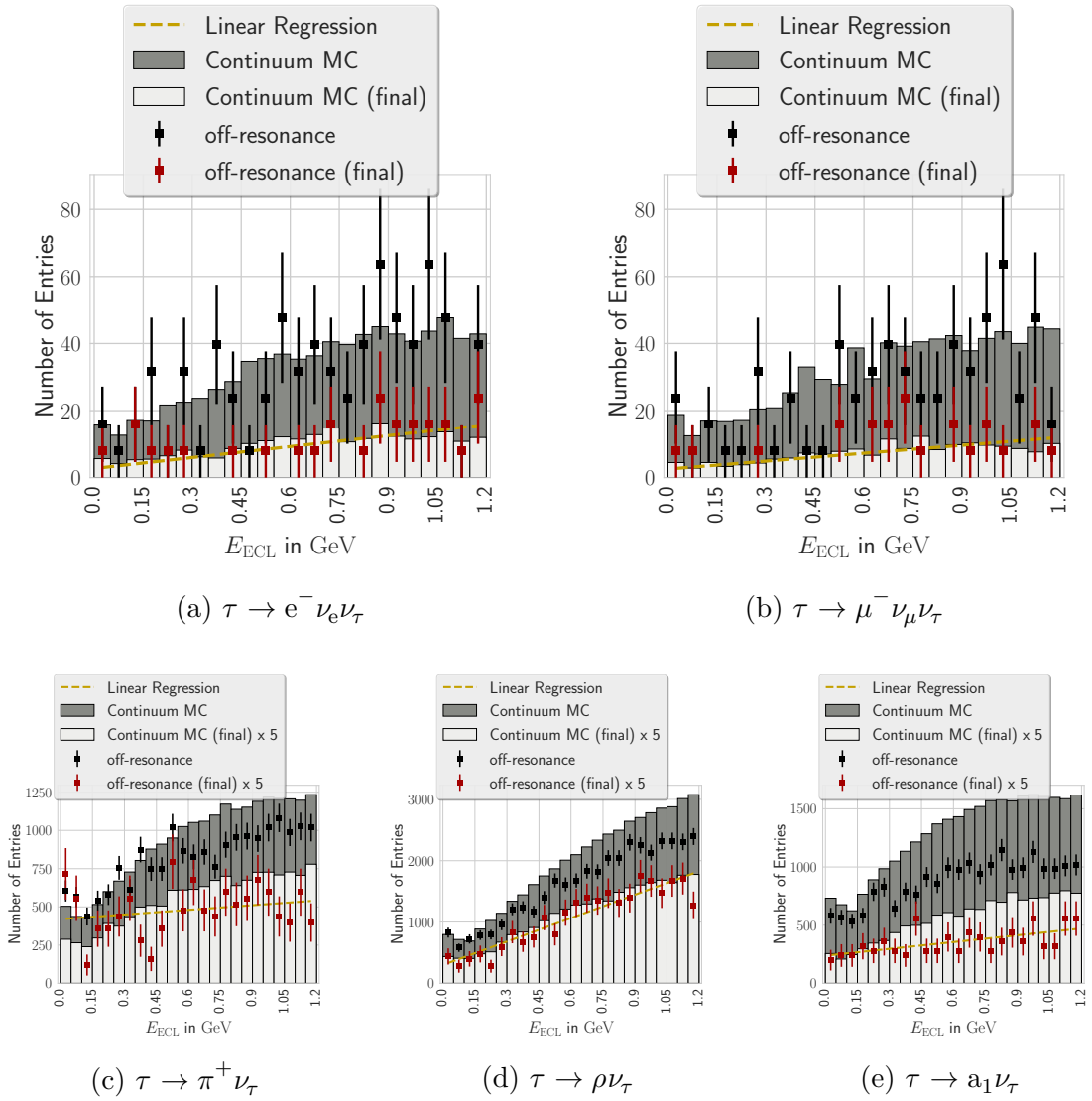
Figure D.1.: Distribution of E_{ECL} for the hadronically tagged candidates on on-resonance data. The final selection is shown in **blue** and discards large amounts of the background including its peaking contribution.



(a) $\tau \to \mathrm{e}^{-}\nu_{\mathrm{e}}\nu_{\tau}$ (b) $\tau \to \mu^{-}\nu_{\mu}\nu_{\tau}$

(c) $\tau \to \pi^{+}\nu_{\tau}$ (d) $\tau \to \rho\nu_{\tau}$ (e) $\tau \to \mathrm{a}_1\nu_{\tau}$

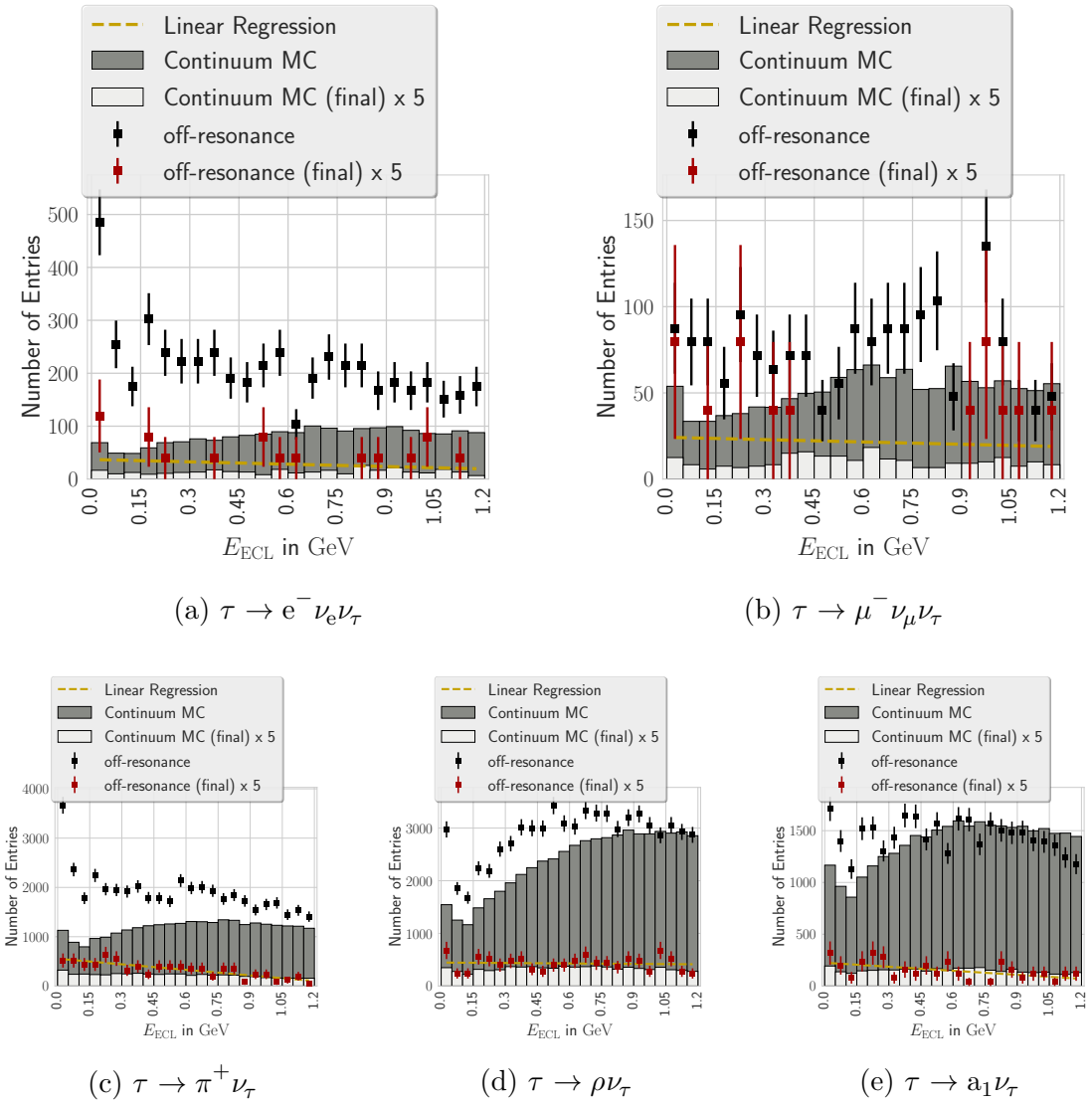
Figure D.2.: Distribution of E_{ECL} for the semileptonically tagged candidates on on-resonance data. The final selection is shown in **blue** and discards large amounts of the background including its peaking contribution.



(a) $\tau \to \mathrm{e}^{-}\nu_{\mathrm{e}}\nu_{\tau}$ (b) $\tau \to \mu^{-}\nu_{\mu}\nu_{\tau}$

(c) $\tau \to \pi^{+}\nu_{\tau}$ (d) $\tau \to \rho\nu_{\tau}$ (e) $\tau \to \mathrm{a}_1\nu_{\tau}$

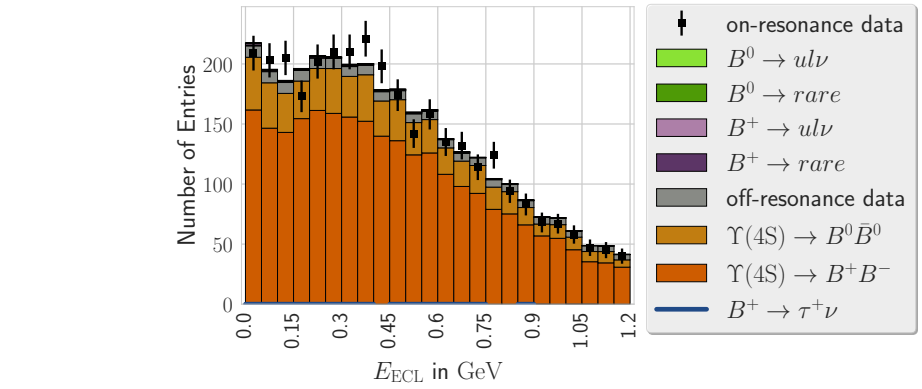
Figure D.3.: Distribution of E_{ECL} for the hadronically tagged candidates on off-resonance data. The final selection is shown in **red** and discards large amounts of the background including its peaking contribution. The shape of the continuum component in the fit is estimated by a linear regression shown in **yellow**.



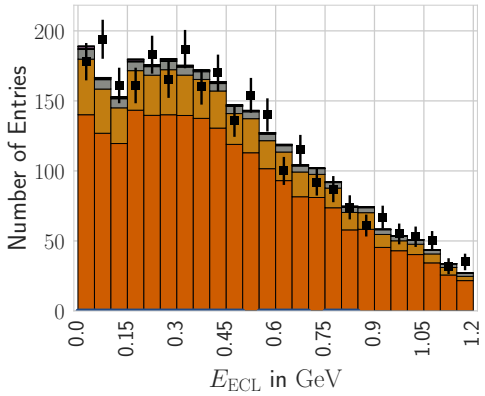
(a) $\tau \to \mathrm{e}^{-}\nu_{\mathrm{e}}\nu_{\tau}$ (b) $\tau \to \mu^{-}\nu_{\mu}\nu_{\tau}$

(c) $\tau \to \pi^{+}\nu_{\tau}$ (d) $\tau \to \rho\nu_{\tau}$ (e) $\tau \to \mathrm{a}_1\nu_{\tau}$

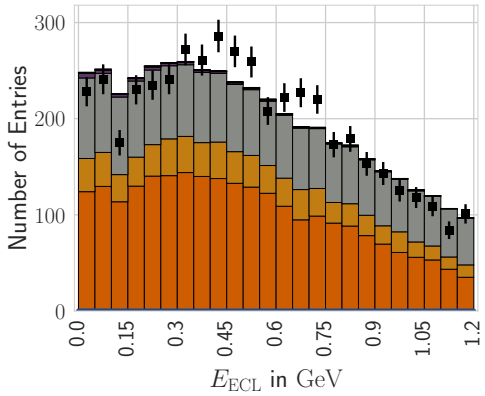
Figure D.4.: Distribution of E_{ECL} for the semileptonically tagged candidates on off-resonance data. The final selection is shown in **red** and discards large amounts of the background including its peaking contribution. The shape of the continuum component in the fit is estimated by a linear regression shown in **yellow**.



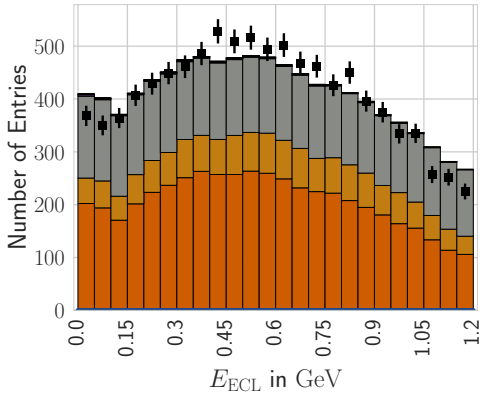
(a) $\tau \to \mathrm{e}^-\nu_\mathrm{e}\nu_\tau$



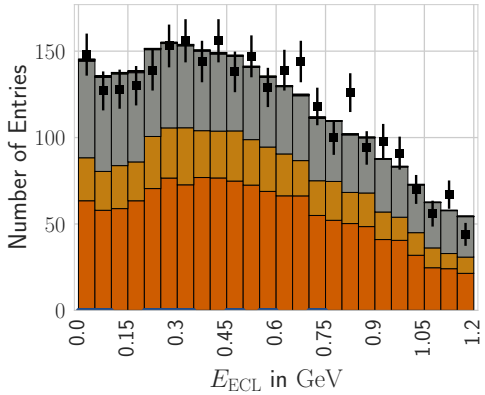
(b) $\tau \to \mu^-\nu_\mu\nu_\tau$



(c) $\tau \to \pi^+\nu_\tau$

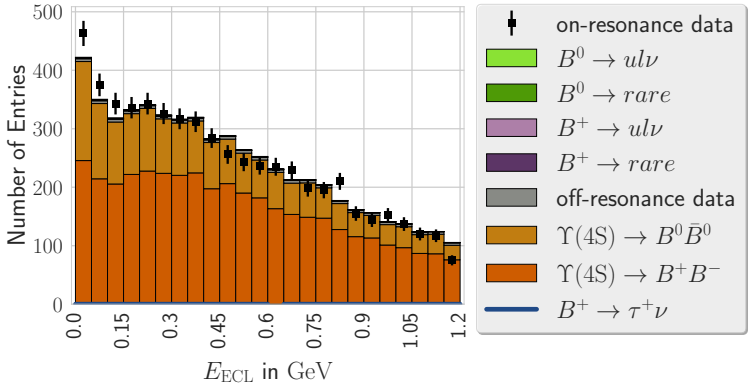


(d) $\tau \to \rho\nu_\tau$

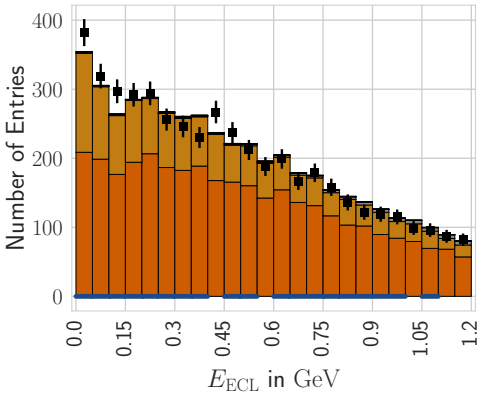


(e) $\tau \to \mathrm{a}_1\nu_\tau$

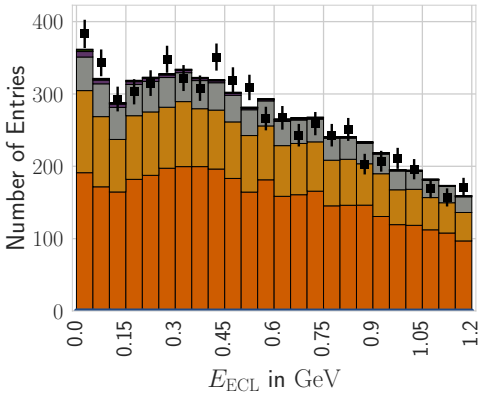
Figure D.5.: Distribution of E_{ECL} for the hadronically tagged candidates in the K^0_S sideband. The continuum component was estimated from off-resonance data using a linear regression. The overall normalization of the expectation was scaled to the observed data.



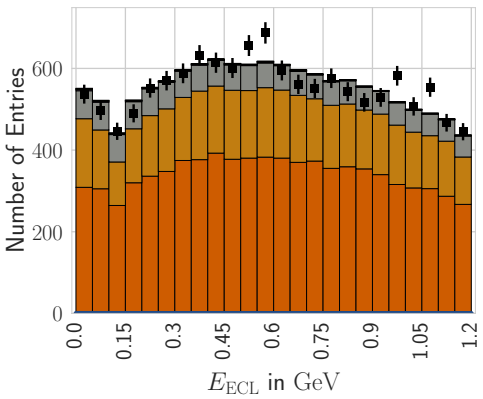
(a) $\tau \to \mathrm{e}^-\nu_\mathrm{e}\nu_\tau$



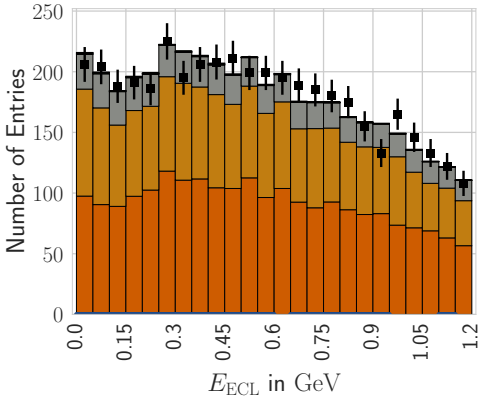
(b) $\tau \to \mu^-\nu_\mu\nu_\tau$



(c) $\tau \to \pi^+\nu_\tau$



(d) $\tau \to \rho\nu_\tau$



(e) $\tau \to \mathrm{a}_1\nu_\tau$

Figure D.6.: Distribution of E_{ECL} for the semileptonically tagged candidates in the K^0_S sideband. The continuum component was estimated from off-resonance data using a linear regression. The overall normalization of the expectation was scaled to the observed data.

D.2.3. Sideband Study

The distribution of E_{ECL} in the $\mathrm{K}^0_{\mathrm{S}}$ sideband is shown after the final selection. The hadronically tagged τ decay channels are shown in Figure D.5. The semileptonically tagged τ decay channels are shown in Figure D.6.

In addition, the shape of the momentum of the τ in the center-of-mass frame for all candidates in the signal region $E_{\mathrm{ECL}} < 100\,\mathrm{MeV}$ was investigated. These distributions are sensitive to unknown peaking contributions from decays and $\mathrm{b} \to \mathrm{u}\ell\nu$ processes. Noe unexpected deviations were observed within the statistical fluctuations. The hadronically tagged τ decay channels are shown in Figure D.7. The semileptonically tagged τ decay channels are shown in Figure D.6.

D.3. Results

D.3.1. Influence of reduced Boost

The asymmetry between the electron and positron beam of SuperKEKB was reduced with respect to KEKB, as stated in Table 2.2. The associated change of the Lorentz boost reduces the resolution of the decay time difference in time-dependent CP violation measurements, on the other hand it increases the detector acceptance. In spherical coordinates the Belle (and Belle II) detector covers $\Theta \in [17^\circ, 150^\circ]$ and $\phi \in [0^\circ, 360^\circ]$. The geometric detector acceptance in percent is

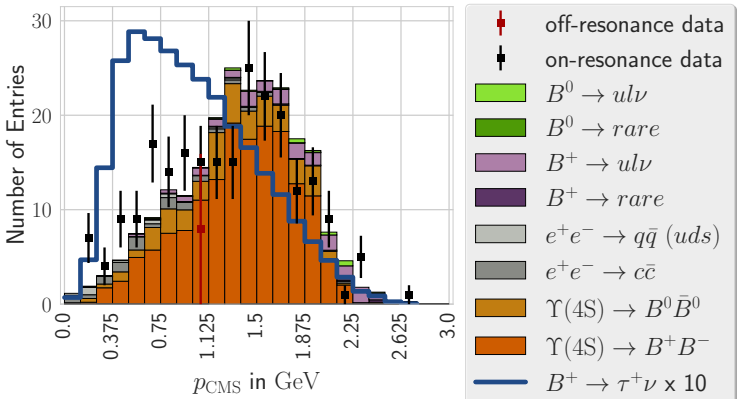
$$\Omega = \frac{1}{4\pi} \int_{17^\circ}^{150^\circ} \int_{0^\circ}^{360^\circ} \mathrm{d}\Omega = 91.1\%.$$

Assuming a uniform spherical distribution of the tracks and neglecting the mass of the final state particles compared to their momenta, the angle $\Theta^\star$ in the boosted frame of reference is

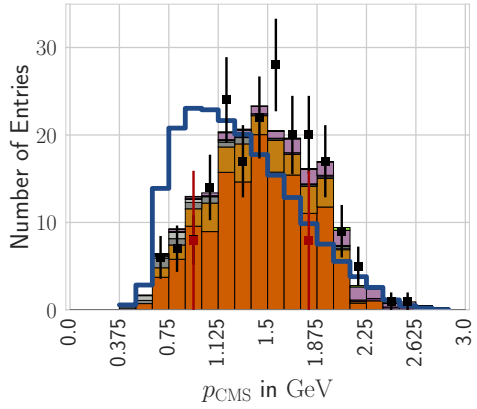
$$\tan\left(\Theta^\star\right) = \frac{1}{\gamma}\left(\frac{\sin\Theta}{\beta + \cos\Theta}\right)$$

where $\gamma = \frac{E_{\mathrm{H}} - E_{\mathrm{L}}}{E_{\mathrm{CMS}}}$ and $\beta = \frac{E_{\mathrm{H}} - E_{\mathrm{L}}}{E_{\mathrm{H}} + E_{\mathrm{L}}}$. In consequence the geometrical detector acceptance in the boosted frame of reference depends on the Lorentz boost of the collider. As can be seen from Figure D.9 this effect is of the order of sub-percent. The measurement of the branching fraction of B $\to \tau\nu$ requires the reconstruction of all tracks and clusters in the event, consequently the reduced boost does have a significant influence on the sensitivity [100].

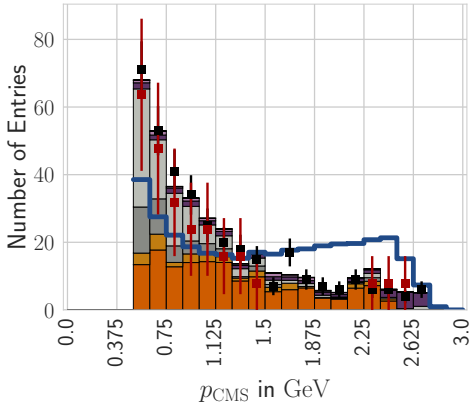
[1] The most likely explanation is a technical error in the Belle measurement.
[2] In a Bayesian sense.



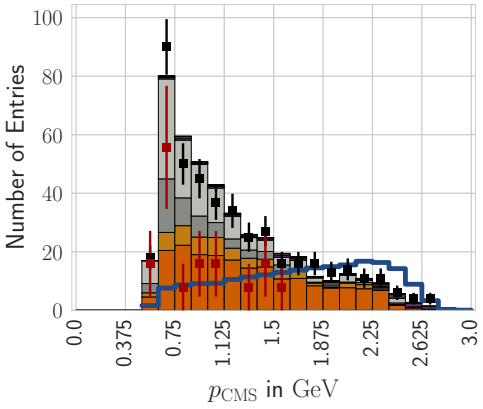
(a) $\tau \to \mathrm{e}^-\nu_\mathrm{e}\nu_\tau$



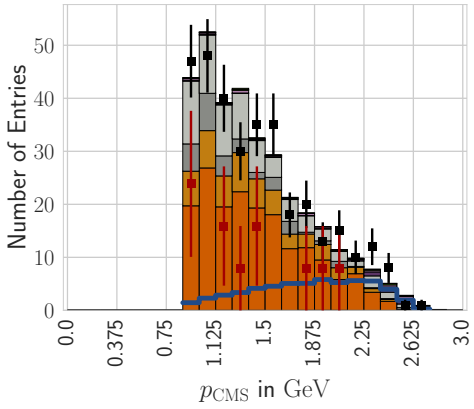
(b) $\tau \to \mu^-\nu_\mu\nu_\tau$



(c) $\tau \to \pi^+\nu_\tau$

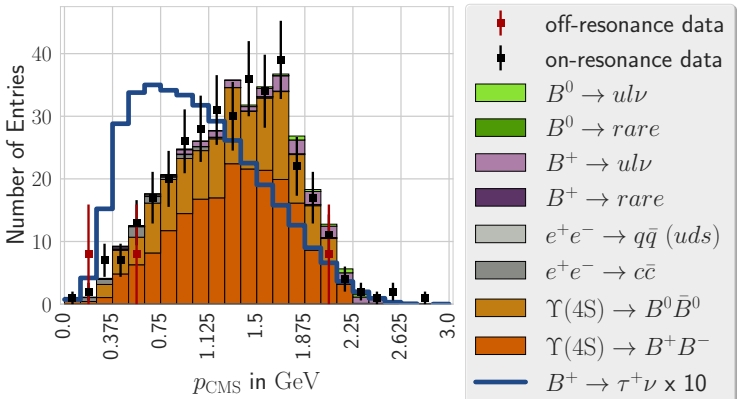


(d) $\tau \to \rho\nu_\tau$

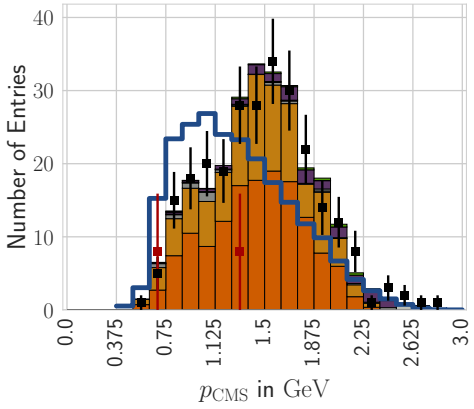


(e) $\tau \to \mathrm{a}_1\nu_\tau$

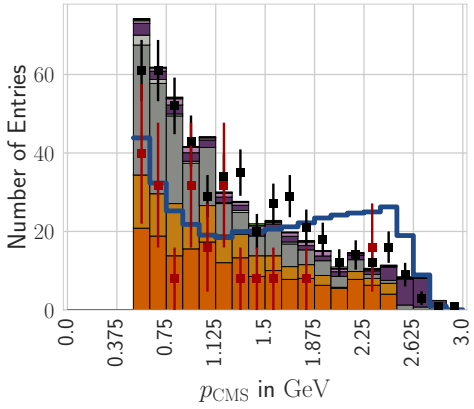
Figure D.7.: Distribution of p_{CMS} for the hadronically tagged candidates in the signal region $E_{\text{ECL}} < 100\,\text{MeV}$. The overall normalization of the expectation was scaled to the observed data.



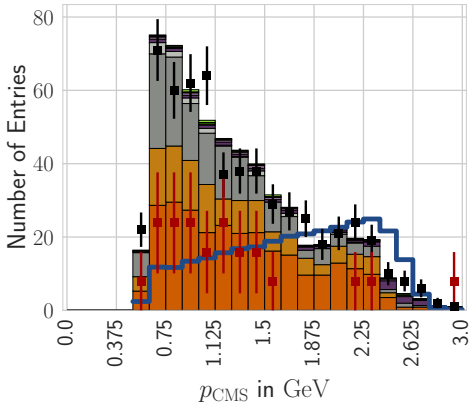
(a) $\tau \to \mathrm{e}^-\nu_\mathrm{e}\nu_\tau$



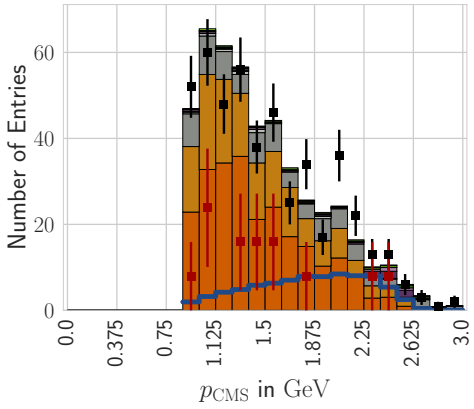
(b) $\tau \to \mu^-\nu_\mu\nu_\tau$



(c) $\tau \to \pi^+\nu_\tau$



(d) $\tau \to \rho\nu_\tau$



(e) $\tau \to \mathrm{a}_1\nu_\tau$

Figure D.8.: Distribution of p_{CMS} for the semileptonically tagged candidates in the signal region $E_{\mathrm{ECL}} < 100\,\mathrm{MeV}$. The overall normalization of the expectation was scaled to the observed data.

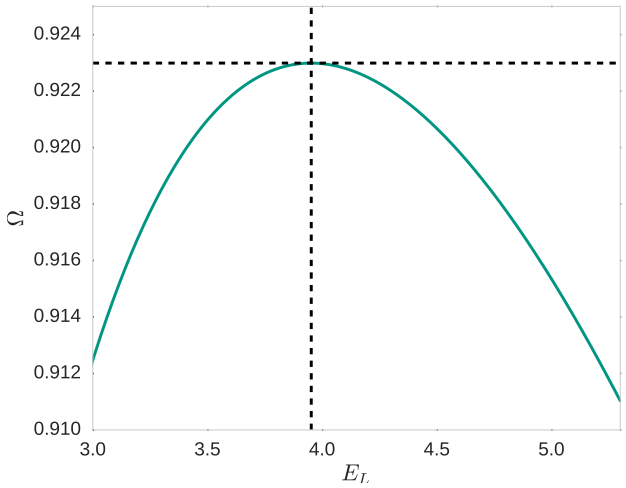


Figure D.9.: Geometrical detector acceptance in a boosted frame of reference depending on the energy E_{L} of the positron beam.

Bibliography

[1] A. J. Bevan et al. „The Physics of the B Factories“. In: *Eur. Phys. J.* C74 (2014), p. 3026. DOI: 10.1140/epjc/s10052-014-3026-9.

[2] A. Abashian, K. Gotow, N. Morgan, and L. Piilonen. „The Belle detector“. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 479.1 (2002), pp. 117–232. DOI: 10.1016/S0168-9002(01)02013-7.

[3] T. Keck. „The Full Event Interpretation for Belle II“. MA thesis. KIT, 2014. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48602.

[4] Z. Doležal and S. Uno. *Belle II Technical Design Report*. Tech. rep. KEK, 2010. arXiv: 1011.0352 [hep-ex].

[5] C. Schwanda. „SuperKEKB machine and Belle II detector status“. In: *Nuclear Physics B - Proceedings Supplements* 209.1 (2010), pp. 70–72. DOI: 10.1016/j.nuclphysbps.2010.12.012.

[6] C. Pulvermacher. „Analysis Software and Full Event Interpretation for the Belle II Experiment“. PhD thesis. KIT, 2015. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48741.

[7] 20.01.2017. URL: http://www-superkekb.kek.jp/documents/luminosityProjection_160802.pdf.

[8] D. Kim. „The software library of the Belle II experiment“. In: *Nuclear and Particle Physics Proceedings* 273 (2016), pp. 957–962. DOI: 10.1016/j.nuclphysbps.2015.09.149.

[9] R. Itoh. „BASF - BELLE AnalysiS Framework“. In: *Proceedings, 9th International Conference on Computing in High-Energy Physics (CHEP 1997)*. 1997. URL: http://www.ifh.de/CHEP97/paper/244.ps.

[10] K. Hara. *Calibration of low momentum K_L^0 efficiency for veto usage in missing E analyses*. Belle Note 1228. 2012.

[11] D. J. Lange. „The EvtGen particle decay simulation package“. In: *Nucl. Instrum. Meth.* A462 (2001). DOI: 10.1016/S0168-9002(01)00089-4.

[12] E. Barberio, B. van Eijk, and Z. Was. „Photos — a universal Monte Carlo for QED radiative corrections in decays“. In: *Computer Physics Communications* 66.1 (1991). DOI: https://doi.org/10.1016/0010-4655(91)90012-A.

[13] S. Agostinelli et al. „GEANT4: A Simulation toolkit“. In: *Nucl. Instrum. Meth.* A506 (2003). DOI: 10.1016/S0168-9002(03)01368-8.

[14] A. Ryd et al. *EvtGen – A Monte Carlo Generator for B - Physics*. User manual. URL: http://evtgen.warwick.ac.uk/static/docs/EvtGenGuide.pdf.

[15] B. Kronenbitter et al. „Measurement of the branching fraction of $B^+ \to \tau^+\nu_\tau$ decays with the semileptonic tagging method“. In: *Phys. Rev.* D92.5 (2015), p. 051102. DOI: 10.1103/PhysRevD.92.051102.

[16] K. Hara, Y. Horii, and T. Iijima. „Evidence for $B^- \to \tau^-\overline{\nu}_\tau$ with a Hadronic Tagging Method Using the Full Data Sample of Belle“. In: *Phys. Rev. Lett.* 110 (Mar. 2013). DOI: 10.1103/PhysRevLett.110.131801.

[17] F. Metzner. „Analysis of $B^+ \to \ell^+\nu_\ell\gamma$ decays with the Belle II Analysis Software Framework“. MA thesis. KIT, 2016. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48845.

[18] M. Gelb. *Search for the decay $B^+ \to l^+\nu\gamma$ with the Full Event Interpretation*. Belle Note 1474. 2017.

[19] F. Fichter. „Search for $B_S \to \phi\pi^0$ Decays at the Belle Experiment“. MA thesis. KIT, 2016. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48880.

[20] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer New York Inc., 2001. ISBN: 978-0-387-84858-7.

[22] V. Vapnik. „Principles of Risk Minimization for Learning Theory“. In: *NIPS*. 1991.

[23] L. Rosasco et al. „Are Loss Functions All the Same?“ In: *Neural Comput.* 16.5 (May 2004), pp. 1063–1076. DOI: 10.1162/089976604773135104.

[24] P. McCullagh and J. A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall, 1989. ISBN: 9780412317606.

[25] E. Parzen. „On Estimation of a Probability Density Function and Mode“. In: *The Annals of Mathematical Statistics* 33.3 (1962), pp. 1065–1076. DOI: 10.1214/aoms/1177704472.

[26] R. A. Rigby and D. M. Stasinopoulos. „Generalized additive models for location, scale and shape“. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54.3 (2005), pp. 507–554. DOI: 10.1111/j.1467-9876.2005.00510.x.

[27] R. W. Koenker and G. Bassett. „Regression Quantiles“. In: *Econometrica* 46.1 (1978), pp. 33–50.

[28] J. Neyman and E. S. Pearson. „On the Problem of the Most Efficient Tests of Statistical Hypotheses“. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (1933), pp. 289–337. DOI: 10.1098/rsta.1933.0009.

[29] R. A. Fisher. „The Use of Multiple Measurements in Taxonomic Problems“. In: *Annals of Eugenics* 7.7 (1936), pp. 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

[30] J. H. Friedman. „Stochastic gradient boosting“. In: *Computational Statistics and Data Analysis* 38.4 (2002), pp. 367–378. DOI: 10.1016/S0167-9473(01)00065-2.

[31] J. Bergstra and Y. Bengio. „Random Search for Hyper-parameter Optimization“. In: *J. Mach. Learn. Res.* 13 (Feb. 2012), pp. 281–305.

[32] D. Maclaurin, D. Duvenaud, and R. Adams. „Gradient-based Hyperparameter Optimization through Reversible Learning“. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 2015, pp. 2113–2122. URL: http://jmlr.org/proceedings/papers/v37/maclaurin15.pdf.

[33] J. Snoek, H. Larochelle, and R. P. Adams. „Practical Bayesian Optimization of Machine Learning Algorithms“. In: *Advances in Neural Information Processing Systems 25*. 2012, pp. 2951–2959. arXiv: 1206.2944 [stat.ML].

[34] O. Behnke, K. Kroeninger, T. Schoerner-Sadenius, and G. Schott. *Data analysis in high energy physics*. Wiley-VCH, 2013. ISBN: 9783527410583.

[35] K. Cranmer and I. Yavin. „RECAST: Extending the Impact of Existing Analyses“. In: *JHEP* 04 (2011), p. 038. DOI: 10.1007/JHEP04(2011)038.

[36] M. Feindt et al. „A Hierarchical NeuroBayes-based Algorithm for Full Reconstruction of B Mesons at B Factories“. In: *Nucl. Instrum. Meth.* A654 (2011), pp. 432–440. DOI: 10.1016/j.nima.2011.06.008.

[37] K. Hornik. „Approximation capabilities of multilayer feedforward networks“. In: *Neural Networks* 4.2 (1991), pp. 251–257. DOI: 10.1016/0893-6080(91)90009-T.

[38] H. W. Lin, M. Tegmark, and D. Rolnick. „Why Does Deep and Cheap Learning Work So Well?“ In: *Journal of Statistical Physics* (July 2017). DOI: 10.1007/s10955-017-1836-5.

[39] P. Baldi, P. Sadowski, and D. Whiteson. „Searching for Exotic Particles in High-Energy Physics with Deep Learning“. In: *Nature Commun.* 5 (2014), p. 4308. DOI: 10.1038/ncomms5308.

[40] Y. Lecun, Y. Bengio, and G. Hinton. „Deep learning“. In: *Nature* 521 (May 2015), pp. 436–444. DOI: 10.1038/nature14539.

[41] I. Goodfellow et al. „Generative Adversarial Nets“. In: *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[42] G. Louppe, M. Kagan, and K. Cranmer. „Learning to Pivot with Adversarial Networks“. In: (2016). arXiv: 1611.01046 [stat.ME].

[43] Y. Bengio, A. Courville, and P. Vincent. „Representation Learning: A Review and New Perspectives“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (Aug. 2013), pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

[44] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. „Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (Apr. 2017), pp. 652–663. DOI: 10.1109/TPAMI.2016.2587640.

[45] M. Pivk and F. R. Le Diberder. „SPlot: A Statistical tool to unfold data distributions“. In: *Nucl. Instrum. Meth.* A555 (2005), pp. 356–369. DOI: 10.1016/j.nima.2005.08.106.

[46] D. Martschei, M. Feindt, S. Honc, and J. Wagner-Kuhr. „Advanced event reweighting using multivariate analysis“. In: *Proceedings, 14th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2011)*. Vol. 368. 2012, p. 012028. DOI: 10.1088/1742-6596/368/1/012028.

[47] T. Keck. „FastBDT: A Speed-Optimized Multivariate Classification Algorithm for the Belle II Experiment“. In: *Computing and Software for Big Science* 1.1 (Sept. 2017). DOI: 10.1007/s41781-017-0002-8.

[48] 02.10.2017. URL: https://github.com/thomaskeck/FastBDT.

[49] J. Therhaag et al. „TMVA - Toolkit for multivariate data analysis“. In: *AIP Conference Proceedings* 1504.1 (2012), pp. 1013–1016. DOI: 10.1063/1.4771869.

[50] S. Nissen. *Implementation of a Fast Artificial Neural Network Library (fann)*. Tech. rep. http://fann.sf.net. Department of Computer Science University of Copenhagen (DIKU), Dec. 2003.

[51] M. Feindt and U. Kerzel. „The NeuroBayes neural network package“. In: *Nucl. Instrum. Meth.* A559 (2006), pp. 190–194. DOI: 10.1016/j.nima.2005.11.166.

[52] F. Pedregosa et al. „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[53] T. Chen and C. Guestrin. „XGBoost: A Scalable Tree Boosting System“. In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[54] M. Abadi et al. „TensorFlow: A system for large-scale machine learning“. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 2016, pp. 265–283. URL: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

[55] F. Chollet et al. *Keras*. 2015. URL: https://github.com/fchollet/keras.

[56] I. J. Goodfellow et al. „Pylearn2: a machine learning research library.“ In: (2013). arXiv: 1308.4214 [stat.ML].

[57] R. Al-Rfou et al. „Theano: A Python framework for fast computation of mathematical expressions“. In: (May 2016). arXiv: 1605.02688 [cs.SC].

[58] C. Patrignani et al. „Review of Particle Physics“. In: *Chin. Phys.* C40.10 (2016), p. 100001. DOI: 10.1088/1674-1137/40/10/100001.

[59] J. A. Hanley and B. J. McNeil. „The meaning and use of the area under a receiver operating characteristic (ROC) curve.“ In: *Radiology* 143.1 (1982), pp. 29–36. DOI: 10.1148/radiology.143.1.7063747.

[60] M. Gelb. „Neutral B Meson Flavor Tagging for Belle II“. MA thesis. KIT, 2015. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48719.

[61] J. Gemmler. „Study of B Meson Flavor Tagging with Deep Neural Networks at Belle and Belle II“. MA thesis. KIT, 2016. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48849.

[62] D. M. Asner, M. Athanas, D. W. Bliss, et al. „Search for exclusive charmless hadronic B decays“. In: *Phys. Rev. D* 53 (3 Feb. 1996), pp. 1039–1050. DOI: 10.1103/PhysRevD.53.1039.

[63] G. C. Fox and S. Wolfram. „Observables for the Analysis of Event Shapes in e^+e^- Annihilation and Other Processes“. In: *Phys. Rev. Lett.* 41 (23 Dec. 1978), pp. 1581–1585. DOI: 10.1103/PhysRevLett.41.1581.

[64] D. Weyland. „Continuum Suppression with Deep Learning techniques for the Belle II Experiment“. MA thesis. KIT, 2017. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48934.

[65] A. Rogozhnikov et al. „New approaches for boosting to uniformity“. In: *JINST* 10.03 (2015), T03002. DOI: 10.1088/1748-0221/10/03/T03002.

[66] M. Feindt and M. Prim. „An algorithm for quantifying dependence in multivariate data sets“. In: *Nucl. Instrum. Meth.* A698 (2013), pp. 84–89. DOI: 10.1016/j.nima.2012.09.043.

[67] J. Dolen et al. „Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure“. In: *JHEP* 05 (2016), p. 156. DOI: 10.1007/JHEP05(2016)156.

[68] J. Stevens and M. Williams. „uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers“. In: *JINST* 8 (2013), P12013. DOI: 10.1088/1748-0221/8/12/P12013.

[69] B. Lipp. „sPlot-based Training of Multivariate Classifiers in the Belle II Analysis Software Framework“. BA thesis. KIT, 2015. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48717.

[70] W. D. Hulsbergen. „Decay chain fitting with a Kalman filter“. In: *Nucl. Instrum. Meth.* A552 (2005), pp. 566–575. DOI: 10.1016/j.nima.2005.06.078.

[71] W. Waltenberger, W. Mitaroff, and F. Moser. „RAVE—a Detector-independent vertex reconstruction toolkit“. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 581.1–2 (2007), pp. 549–552. DOI: 10.1016/j.nima.2007.08.048.

[72] R. Fruhwirth. „Application of Kalman filtering to track and vertex fitting“. In: *Nucl. Instrum. Meth.* (1987). DOI: 10.1016/0168-9002(87)90887-4.

[73] G. M. Amdahl. „Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities“. In: *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*. 1967, pp. 483–485. DOI: 10.1145/1465482.1465560.

[74] 02.10.2017. URL: https://github.com/thomaskeck/FastFit.

[75] K. Kirchgessner. „Semileptonic Tag Side Reconstruction“. MA thesis. KIT, 2012. URL: http://ekp-invenio.physik.uni-karlsruhe.de/record/48181.

[76] J. Schwab. „Calibration of the Full Event Interpretation for the Belle and the Belle II experiment“. MA thesis. KIT, 2017. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48931.

[77] J. Grygier. „Search for $B \to h\nu\nu$ decays with semileptonic tagging at Belle“. In: (Feb. 2017). arXiv: 1702.03224 [hep-ex].

[78] A. Heller. „Search for $\mathrm{B}^+ \to \ell^+\nu\gamma$ decays with hadronic tagging using the full Belle data sample“. PhD thesis. KIT, 2015. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48743.

[79] M. Huschle. „Measurement of the branching ratio of $B \to D^{(*)}\tau\nu_\tau$ relative to $B \to D^{(*)}\ell\nu_\ell$ decays with hadronic tagging at Belle“. PhD thesis. KIT, 2015. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48622.

[80] A. Santoro et al. „A simple neural network module for relational reasoning“. In: (June 2017). arXiv: 1706.01427 [cs.CL].

[81] J. Schmidhuber. „Deep learning in neural networks: An overview“. In: *Neural Networks* (2015). DOI: https://doi.org/10.1016/j.neunet.2014.09.003.

[82] M. D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 2014. ISBN: 9781107034730.

[83] O. Eberhardt, U. Nierste, and M. Wiebusch. „Status of the two-Higgs-doublet model of type II“. In: *JHEP* (2013). DOI: 10.1007/JHEP07(2013)118.

[84] M. Misiak and M. Steinhauser. „Weak radiative decays of the B meson and bounds on $M_{H^\pm}$ in the Two-Higgs-Doublet Model“. In: *Eur. Phys. J.* (2017). DOI: 10.1140/epjc/s10052-017-4776-y.

[85] J. H. Kühn and A. Santamaria. „Tau decays to pions“. In: *Z. Phys.* (1990). DOI: 10.1007/BF01572024.

[86] J. P. Lees et al. „Evidence of $B^+ \to \tau^+\nu$ decays with hadronic B tags“. In: *Phys. Rev.* D88 (2013), p. 031102. DOI: 10.1103/PhysRevD.88.031102.

[87] B. Aubert et al. „Search for $B^+ \to \ell^+\nu_\ell$ recoiling against $B^- \to D^0\ell^-\overline{\nu}X$“. In: *Phys. Rev. D* 81 (5 Mar. 2010), p. 051101. DOI: 10.1103/PhysRevD.81.051101.

[88] Edward I. Shibata. „Hadronic structure in $\tau^- \to \pi^-\pi^-\pi^+\nu_\tau$ decays“. In: *Tau Lepton Physics: Proceedings*. Vol. C0209101. 2002, TU05. DOI: 10.1016/S0920-5632(03)80305-5.

[89] B. Kronenbitter. „Measurement of the branching fraction of $B^+ \to \tau^+\nu_\tau$ decays at the Belle experiment“. PhD thesis. KIT, 2014. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48604.

[90] J. Grygier. *Development and Evaluation of a newCalorimeter Based Variable for Missing Energy Analyses at the Belle Experiment*. 2013. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48043.

[91] R. Barlow. „Experimental status of $Btot\nu$ and $Btol\nu(\gamma)$“. In: *CKM unitarity triangle. Proceedings, 6th International Workshop, CKM 2010, Warwick, UK, September 6-10, 2010*. 2011. arXiv: 1102.1267 [hep-ex].

[92] K. Hara. *Calibration of low momentum K_L^0 efficiency for veto usage in missing E analyses*. Belle Note 1228 (internal).

[93] B. Bhuyan. *High P_T Tracking Efficiency Using Partially Reconstructed D^* Decays*. Belle Note 1165 (internal).

[94] S. W. Lin. *π^0 systematics Using Inclusive η*. Belle Note 645 (internal).

[95] S. Nishida. *Study of Kaon and Pion Identification Using Inclusive D^* Sample*. Belle Note 779 (internal).

[96] L. Hinz. *Lepton ID efficiency correction and systematic error*. Belle Note 954 (internal).

[97] M. Ziegler. „Search for the decay B $\to \tau^-\tau^+$ with the belle experiment“. PhD thesis. KIT, 2016. URL: http://dx.doi.org/10.5445/IR/1000063630.

[98] B. Aubert et al. „Search for the Rare Decay $B^0 \to \tau^+\tau^-$ at BABAR“. In: *Phys. Rev. Lett.* 96 (24 June 2006). DOI: 10.1103/PhysRevLett.96.241802.

[99] R. Aaij et al. „Search for the decays $B_s^0 \to \tau^+\tau^-$ and $B^0 \to \tau^+\tau^-$“. In: *Phys. Rev. Lett.* 118.25 (2017). DOI: 10.1103/PhysRevLett.118.251802.

[100] I. Adachi et al. „sBelle Design Study Report“. In: (2008). arXiv: 0810.4084 [hep-ex].

Danksagung

Ich danke allen die zum Erfolg dieser Arbeit beigetragen haben.

> If I have seen further it is by standing on the shoulders of giants.
> – Isaac Newton

Ich blicke auf vier wundervolle Jahre am Institut für experimentelle Teilchenphysik zurück. Daher danke ich meinen Kollegen und Freunden ganz herzlich für die schöne Zeit und das kreative Umfeld. Wegen euch habe ich mich auf jeden einzelnen Tag am Institut gefreut, und durch euch habe ich an jedem dieser Tage etwas Neues gelernt.

Die fachlichen Beiträge meiner Kollegen erwähne ich hier gesondert, und möchte ihnen auf diesem Weg meinen besonderen Dank zukommen lassen.

Die Idee zu einer Belle zu Belle II Datenkonvertierung erhielt ich von Bastian Kronenbitter. Als Ausgangspunkt für die `belle_legacy` Programmbibliothek diente mir eine bereinigte Version des `Belle Analysis Software Frameworks`, welche ich von Dr. Ryosuke Itoh erhalten hatte. Bei der Integration dieser Programmbibliothek in BASF2 und bei der Anbindung an die `Belle II Condition Database` unterstützte mich Dr. Martin Ritter. Das `b2bii` Paket entstand schließlich in Zusammenarbeit mit der Belle II `b2bii` task-force unter der Leitung von Dr. Anze Zupanc. Zu den ersten Nutzern, die für Feedback besonders wichtig waren, zählten Felix Metzner, Moritz Gelb und Markus Prim.

Mein Wissen über multivariate Methoden wurde stark von Prof. Dr. Michael Feindt geprägt. Die ursprünglichen Ideen zu den Daten-getriebenen Methoden und zur Weiterentwicklung der sPlot Technik stammen von ihm. Die erste Implementierung der sPlot Technik im Belle II Software Framework entstand während der Bachelorarbeit von Benjamin Lipp. Den Begriff Deep Learning hörte ich bewusst das erste Mal von Dr. Martin Heck. Wesentliche Beiträge zur Integration von Deep Learning Technologien in das `mva` Paket entstanden während der Masterarbeit und Doktorarbeit von Jochen Gemmler, und der Masterarbeit von Dennis Weyland. Eine

treibende Kraft bei der Integration des `mva` Pakets in der Rekonstruktionssoftware waren Nils Braun und Oliver Frost.

Die erste Version der `FEI` orientierte sich stark an der `Full Reconstruction`, welche von vielen Doktoranden am EKP zwischen 2008 – 2012 entwickelt wurde. Diese erste Version entstand in enger Zusammenarbeit mit Dr. Christian Pulvermacher. Viele inhaltliche Anregungen und die Idee zur spezifischen `FEI` erhielt ich von Dr. Martin Heck. Die Kalibrierung der FEI auf Belle Daten wurde von Judith Schwab in ihrer Masterarbeit durchgeführt. Die erste Anwendung der FEI auf Belle Daten wurde von Felix Metzner in seiner Masterarbeit durchgeführt.

Die Analyse $B^+ \to \tau^+ \nu_\tau$ baut auf den Doktorarbeiten von Dr. Bastian Kronenbitter, Dr. Johannes Grygier und Dr. Michael Ziegler auf.

Die inhaltliche und sprachliche Korrektur dieser Arbeit übernahmen: Nils Braun, Moritz Gelb, Jochen Gemmler, Dr. Pablo Goldenzweig, Dr. Thomas Hauth, Dr. Martin Heck, Felix Metzner, Dr. William Sutcliffe und Judith Schwab.

Gleichwertig zu diesen fachlichen Beiträgen, beruht der Erfolg dieser Arbeit auf der Unterstützung durch meine Betreuer.

Ich danke Dr. Martin Heck, der die stellvertretende Leitung der Arbeitsgruppe Feindt übernommen hat, und die Gruppe mit viel Engagement fortgeführt hat. Ich danke Prof. Dr. Günter Quast für die Übernahme des Korreferats und insbesondere auch für die motivierenden Gespräche. Ich danke meinem Doktorvater und Referenten Prof. Dr. Michael Feindt. Er hat mir während meiner Zeit als Doktorand viele wertvolle Erfahrungen ermöglicht, die ich in keiner anderen Arbeitsgruppe hätte machen können.