IEKP-KA/2009-32

Neural Network based B^0 Flavor Tagging at the Belle Experiment

Michael Prim

Diplomarbeit

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

Referent: Prof. Dr. M. Feindt
Korreferent: Prof. Dr. T. Müller

Institut für Experimentelle Kernphysik

Januar 2010

B^0 Flavor-Tagging mit Neuronalen Netzwerken am Belle Experiment

Deutsche Zusammenfassung

Diese Diplomarbeit beschäftigt sich mit dem Flavor-Tagging neutraler B-Mesonen am Belle-Experiment. Neben der Entwicklung eines neuartigen und auf neuronalen Netzwerken basierenden Flavor-Taggers, umfasst sie auch Teile der zugehörigen Validierung.

Das Standardmodell der Teilchenphysik ist eine weithin anerkannte Theorie zur Beschreibung von drei der vier fundamentalen Wechselwirkungen. Das in den 1960er und 70er Jahren entwickelte Modell erlaubt die Beschreibung der starken, schwachen und elektromagnetischen Wechselwirkung und umfasst die zugehörigen Austauschteilchen, sowie die bekannten Formen der Materie: Quarks und Leptonen. Ein wesentlicher Bestandteil des Standardmodells ist die Cabibbo-Kobayashi-Maskawa (CKM) Matrix. Sie beschreibt Übergänge zwischen verschiedenen Flavor-Zuständen von Quarks. Darüber hinaus ermöglicht die Kobayashi- Maskawa-Theorie eine Erklärung der CP-Verletzung im Rahmen des Standardmodells. 2001 konnten das Belle- sowie das BaBar-Experiment CP-Verletzung im Zerfall von B-Mesonen beobachten und damit die Kobayashi-Maskawa-Theorie bestätigen. Kobayashi und Maskawa bekamen deshalb für ihre Theorie 2008 den Nobelpreis für Physik verliehen.

Von dieser besonderen Messung abgesehen, wird das Standardmodell permanent Tests seiner Gültigkeit unterzogen. Eine wichtige Form stellt hierbei die Messung zeitabhängiger CP-Verletzung dar. Diese kann als Asymmetrie in den zeitabhängigen Zerfallsraten von B^0 und $\overline{B}^0$ in CP-Eigenzustände f_{CP} beobachtet werden:

$$\mathcal{A}_{f_{\mathrm{CP}}}(\Delta t) = \frac{\Gamma\left(\overline{B}^0(\Delta t) \to f_{\mathrm{CP}}\right) - \Gamma\left(B^0(\Delta t) \to f_{\mathrm{CP}}\right)}{\Gamma\left(\overline{B}^0(\Delta t) \to f_{\mathrm{CP}}\right) + \Gamma\left(B^0(\Delta t) \to f_{\mathrm{CP}}\right)},$$

wobei Δt die Zeitdifferenz zwischen dem B^0 und $\overline{B}^0$ Zerfall ist. Es ist offensichtlich, dass für diese Messung die Flavor-Information, also ob es sich um B^0 oder $\overline{B}^0$ gehandelt hat, notwendig ist. Diese Information ist nicht durch den CP-Endzustand gegeben. In einer e^-e^+-Kollision wird ein quantenmechanisch verschränktes $B^0\overline{B}^0$-Paar erzeugt. Somit ist zum Zeitpunkt des Zerfalls des ersten B-Mesons der Flavor des zweiten B-Mesons festgelegt. Daher wird ein B-Meson im zu untersuchenden Zerfallskanal f_{CP} explizit rekonstruiert, wohingegen durch Analyse der Zerfallsprodukte des zweiten B-Mesons die Flavor-Information über das erste B-Meson extrahiert wird. Diesen Prozess bezeichnet man als Flavor-Tagging, der eine essentielle Komponente in der Messung zeitabhängiger CP-Verletzung darstellt.

Um den nötigen experimentellen Rahmen für eine derartige Messung zu schaffen, wurden der KEKB-Beschleuniger sowie der Belle-Detektor gebaut. Beim KEKB-Beschleuniger handelt es sich um einen asymmetrischen e^-e^+-Beschleuniger, welcher bei einer Schwerpunktsenergie von $\sqrt{s} = 10.58\,\mathrm{GeV}$ arbeitet. Diese Schwerpunktsenergie

entspricht der Masse der $\Upsilon(4S)$-Resonanz, welche fast ausschließlich in $B\overline{B}$-Mesonen zerfällt. Die B-Mesonen werden hierbei nahezu in Ruhe produziert. Durch die asymmetrische Strahlenergie ist das ganze System jedoch in eine Richtung geboostet, wodurch eine Messung von Δt erst möglich wird.

Der Belle-Detektor entspricht dem üblichen Schema eines Teilchendetektors. Im Innersten befindet sich ein Silizium-Vertexdetektor zur Messung der Zerfallsorte der produzierten B-Mesonen. Anschließend folgt eine Driftkammer zur Spurrekonstruktion der geladenen Zerfallsprodukte. Außerhalb der Driftkammer befinden sich Cherenkov-Zähler sowie Flugzeit-Zähler und ermöglichen weitere Rückschlüsse auf die Art der Zerfallsprodukte. Ein abschließendes elektromagnetisches Kalorimeter misst die Energie von Elektronen und Photonen. Alle bisher genannten Komponenten befinden sich innerhalb eines supraleitenden Magneten, wodurch eine Impulsmessung der geladenen Zerfallsprodukte in der Driftkammer möglich wird. Myon-Kammern bilden die äußerste Hülle des Detektors und detektieren Myonen, welche alle anderen Detektorkomponenten passieren.

Die oben erwähnte Analyse der Zerfallsprodukte zur Ermittlung des B-Meson-Flavors stellt eine große Herausforderung dar. Es gibt viele Zerfallsmöglichkeiten für B-Mesonen, aber nicht alle sind geeignet, um daraus Informationen über den Flavor zu gewinnen. Darüber hinaus sind die Informationen, welche sich aus den verschiedenen Endprodukten eines B-Meson-Zerfalls gewinnen lassen, stark korreliert. Um diese Korrelationen zu berücksichtigen, wurde ein auf neuronalen Netzwerken basierender Ansatz für den Flavor-Tagger gewählt. Da sich letztlich die Flavor-Information, in Abhängigkeit der Zerfallsprodukte, nur mit einer gewissen Wahrscheinlichkeit korrekt vorhersagen lässt, ist eine gute Kalibrierung des Flavor-Taggers notwendig. Nur so ist eine Interpretation des Flavor-Tagger-Ausgabewertes als Wahrscheinlichkeit für B^0 oder $\overline{B}^0$ möglich.

Der Flavor-Tagger gliedert sich in drei Ebenen. Zunächst wird auf einer Spuren-Ebene versucht, anhand der einzelnen Spuren den B-Flavor zu bestimmen. Hierzu werden die Spuren der Zerfallsprodukte in verschiedene Kategorien (Langsame Pionen, Lambda-Baryonen, Kaonen mit und ohne zusätzlichem K_S^0 im Ereignis, Elektronen und Myonen) eingeteilt. In jeder Kategorie versucht ein speziell auf diese Aufgabe trainiertes neuronales Netzwerk den Flavor zu ermitteln. In der nächsten, der Event-Ebene, werden Informationen von mehreren Spuren-Ebenen-Netzwerken kombiniert. Zum Beispiel werden die Informationen aus dem Elektronen- und Myonen-Netzwerk zu einem gemeinsamen Leptonen-Netzwerk zusammengefasst. In der letzten Ebene werden alle Informationen im Ereignis in einem einzigen Netzwerk kombiniert. Der Ausgabewert dieses Netzwerkes entspricht bei korrekter Kalibrierung der Wahrscheinlichkeit dafür, dass es sich beim untersuchten B-Meson um ein B^0 oder $\overline{B}^0$ gehandelt hat.

Mit diesem neuen Ansatz konnte eine relative Verbesserung der Flavor-Tagger-Leistung von 2.7% gegenüber bestehenden Algorithmen erreicht werden. Dieser Wert wurde auf simulierten Daten ermittelt und vermittelt einen guten Eindruck von der Größenordnung der Verbesserung. Da jedoch keine Simulation die Wirklichkeit korrekt beschreiben kann, können systematische Abweichungen zwischen simulierten und echten Daten auftreten. Da der Flavor-Tagger später auf echte Daten angewendet wird, sollte

seine Validierung auch mit Hilfe echter Daten erfolgen. Dies ist anhand einer kombinierten Parameterschätzung der B^0-Mischungsfrequenz und der Rate falsch getaggter B-Mesonen möglich. Obwohl es weitere Möglichkeiten zur Optimierung des Flavor-Taggers gibt, sind diese vorerst, bis zum Abschluss der Validierung und Verifizierung der Verbesserung auf echten Daten, aufgeschoben. Der Flavor-Tagger ist, in der im Rahmen dieser Arbeit beschriebenen Form, seit Dezember 2009 Teil der Belle-Analyse-Software.

Zur Validierung auf echten Daten wird zunächst der Zerfall eines B-Mesons in einen Flavor-Eigenzustand explizit rekonstruiert, d.h. die Flavor-Information ist anhand der Zerfallsprodukte gegeben. Die Anwendung des Flavor-Taggers auf das andere B-Meson ermöglicht somit eine Bestimmung der Rate von falsch getaggten B-Mesonen und eine Überprüfung, ob die vorhergesagte Wahrscheinlichkeit des Flavor-Taggers korrekt ist.

Durch eine Überarbeitung von großen Teilen der Spuren-Rekonstruktions-Algorithmen haben sich systematische Unterschiede in den aktuellen Daten gegenüber früheren Daten ergeben. Analysen, welche auf alten Daten zur Validierung von Flavor-Tagger-Algorithmen erarbeitet wurden, sind daher nur noch bedingt anwendbar. Es wurde daher beschlossen, die explizite Rekonstruktion des Flavor-Eigenzustands $B^0 \to D^*(2010)^- \ell^+ \nu$ von Grund auf neu zu entwickeln.

Bei der Neuentwicklung fanden gegenüber der alten Analyse ebenfalls neuronale Netzwerke Anwendung. Die Selektion des besten B^0 Kandidaten in einem Ereignis wird durch ein neuronales Netzwerk durchgeführt. Dadurch werden harte Schnitte zur Selektion und Untergrund-Reduktion vermieden. Auf diese Weise wurde eine Steigerung der Effizienz um etwa 120% auf $\epsilon = (8.85 \pm 0.04)\%$ erreicht. Die Reinheit der Selektion wurde zu $p = (63.76 \pm 0.09)\%$ ermittelt und ist etwa 15% schlechter als in alten Analysen. Da jedoch das Produkt aus Effizienz und Reinheit ausschlaggebend ist, wird der Verlust an Reinheit durch den Zugewinn an Effizienz mehr als kompensiert.

Die Entwicklung der Selektion ist damit abgeschlossen und einer Anwendung auf echte Daten steht nichts mehr im Wege. Die abschließende Phase der Validierung kann somit im Anschluss an diese Diplomarbeit beginnen.

Contents

Daß ich erkenne, was die Welt
Im Innersten zusammenhält

Goethe, Faust I

1 Introduction

The world is excited about the Large Hadron Collider (LHC) [1] at Geneva. The world's largest, most energetic and coldest particle collider. Together with ATLAS [2] and CMS [3], two of the largest detectors for high energetic particle physics, the LHC project is one of superlatives even before being started.

But there are other experiments of superlatives in the world of high energy particle physics. One is located at Tsukuba, Japan, northwest of Tokyo. It's the e^-e^+ collider KEKB, which operates at the $\Upsilon(4S)$ resonance and has reached a world record luminosity of $2.11 \cdot 10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ in June 2009.

This allows the Belle collaboration to study physics using one of the world's largest data samples. About 1 billion B-meson pairs have been recorded. With such statistics, high precision measurements can be performed to probe the Standard Model of particle physics. One of the most important results is the measurement of large CP violation in B-meson decays [4]. This was one of two measurements that confirmed the Kobayashi-Maskawa theory of CP violation in the Standard Model [5], which earned the Nobel prize in physics 2008.

This thesis covers the topic of neutral B flavor tagging at the Belle experiment. Flavor tagging is a crucial part in a time dependent CP violation measurement. Therefore, a brief introduction to the Standard Model of particle physics will be given in chapter 2. An review of time dependent CP violation measurement will be given and the decay of neutral B-mesons will be explained in detail as its understanding is essential for flavor tagging. Also the process of neutral B-meson mixing and its use for validation of flavor tagging algorithms will be explained. The experimental setup formed by the KEKB accelerator and the Belle detector will be described in chapter 3. Chapter 4 will give an introduction to multivariate methods of data analysis with neural networks, using the program package NeuroBayes® . A neural network based neutral B flavor tagging algorithm will then be presented in chapter 5. Its structure will be explained and the results of a Monte Carlo (MC) comparison study with existing algorithms will be shown. MC simulations are, however, only as good as the models that are used for simulation. In chapter 6 the procedure on how to validate such a tagging algorithm only on real data with the measurement of B^0 mixing in the flavor specific decay channel $B^0 \to D^*(2010)^-\ell^+\nu$ will be covered. The reconstruction and selection of a $B^0 \to D^*(2010)^-\ell^+\nu$ enriched sample for validation will then be described. To conclude this thesis, chapter 7 will summarize the results and an outlook will be given.

2 Theoretical Overview

Currently, a theory known as the Standard Model (SM) of particle physics is commonly accepted. It has been probed by various experiments over the last decades and, up to now, no inconsistencies have been found. However, the theory can only explain three out of four fundamental interactions known today. But as the fourth interaction, gravity, can be neglected on the energy scale of particle physics, this is not a serious problem for the confidence of the Standard Model.

2.1 The Standard Model

The Standard Model [6] describes the properties of particles and the three fundamental interactions known as strong, weak and electromagnetic interaction. From a mathematical point of view the Standard Model is a combination of three local symmetry groups $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$, where the indices indicate the *color* charge (C) of the strong interaction, the *left* chirality (L) of the weak interaction and the *hypercharge* (Y) of the electroweak interaction.

This description was introduced in the 1960's, beginning with the electroweak theory [7–9], and early 1970's, describing the strong interaction [10–12]. One of the main principles of the SM is, that the forces are mediated by spin 1 particles, called bosons, between spin 1/2 particles called fermions. The exchanged particles are related to gauge fields of the respective symmetry group:

- $SU(3)$ is related to the gauge fields G_μ^α ($\alpha = 1 \dots 8$), which represent 8 gluons g, physical particles that mediate the strong force.

- $SU(2) \otimes U(1)$ is related to the gauge fields W_μ^α ($\alpha = 1 \dots 3$) and B_μ respectively. The two charged $W^\pm$ bosons are represented by a combination of W_μ^1 and W_μ^2. Due to electroweak unification, the mixing between W_μ^3 and B_μ represents the neutral Z^0 boson as well as the photon γ. $W^\pm$ and Z^0 are mediators of the weak force and the photon mediates the electromagnetic force.

All particles of the Standard Model are illustrated in Figure 2.1. Matter only consists of fermions, which are divided into quarks and leptons. All fermions carry weak isospin charge and can interact weakly but only quarks carry color charge and can interact strongly. Only the particles with electric charge take part in electromagnetic interactions. Sometimes quarks are also categorized into up-type (u, c, t) and down-type (d, s, b) with electric charges $+2/3\,\mathrm{e}_0$ or $-1/3\,\mathrm{e}_0$ respectively. Quarks and leptons

are classified into three generations or families. Stable matter only consists of first generation particles.

All particles have corresponding anti-particles with opposite charges, e.g. electron e^- and positron e^+ or up quark u and anti-up quark $\overline{u}$. The term flavor is used to describe which kind of particle is meant.

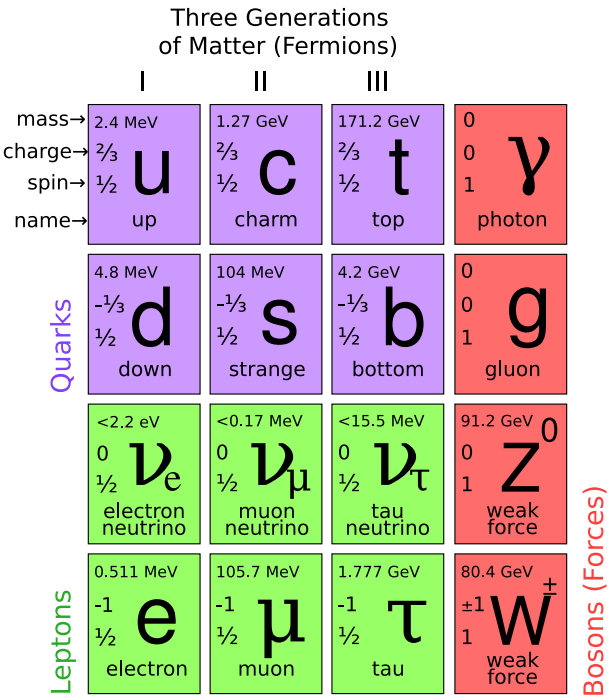

Figure 2.1: Particles of the Standard Model.

In total there are 12 particles, 12 anti-particles, 8 gluons, 3 weak interaction bosons and 1 photon known in the Standard Model. One missing and not yet discovered particle called the Higgs boson, together with a process called electroweak symmetry breaking, is responsible for giving mass to all particles. Direct search for this important missing part is done at the LHC project.

Due to the fact that gluons carry color charge themselves, they can self-interact. Therefore, quarks never occur in isolation but only in bound states, called hadrons. This phenomenon is called quark- or color-confinement. The process from single quark to several hadrons is called fragmentation. Protons and neutrons are such hadrons and not fundamental particles, since they are bound states of three quarks. The proton is composed of uud quarks whereas the neutron is composed of udd quarks. Bound states of 3 quarks are called baryons. Bound states formed by one quark and one anti-quark are called mesons. Mesons composed of a bottom quark b and another quark are called B-mesons. The notation B_q^{charge} is used, where $charge$ is the charge of the entire meson and q the flavor of the second quark. A B_c^+ meson thus consists of a $\overline{b}$ quark and a c quark. If the q subscript is omitted, first generation quarks are inferred and, therefore, $charge$ defines the composition sufficiently, i.e. B^0 consists of $\overline{b}d$ and B^- of $b\overline{u}$.

2.2 CKM Matrix

The only flavor changing process allowed in the Standard Model is the exchange of charged $W^{\pm}$ bosons of the weak interaction. In the quark sector, this requires a transition between up-type and down-type quarks due to charge and weak isospin conservation. The transition is realized by the coupling of the $W^{\pm}$ bosons to the weak eigenstates q', which do not coincide with the mass eigenstates q. The particular coupling strength to the mass eigenstates is given by the 3×3 Cabibbo-Kobayashi-Maskawa (CKM) matrix V_{CKM} [5]. Its elements are fundamental parameters of the Standard Model. The transformation between weak and mass eigenstates is given by

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}. \tag{2.1}$$

The CKM matrix is a unitary matrix and can be parameterized by three rotation angles θ_{ij} and a complex phase δ. The latter was introduced by Kobayashi and Maskawa, together with the 3rd generation of quarks, to explain CP violating processes in the Standard Model.

The hierarchy $s_{13} \ll s_{23} \ll s_{12} \ll 1$, where $s_{ij} = \sin\theta_{ij}$, is known experimentally [6]. This hierarchy motivates the parameterization

$$s_{12} = \lambda = \frac{|V_{us}|}{\sqrt{|V_{ud}|^2 + |V_{us}|^2}}, \qquad s_{23} = A\lambda^2 = \lambda|\frac{V_{cb}}{V_{us}}|, \tag{2.2}$$

$$s_{13}e^{i\delta} = V_{ub}^* = A\lambda^3(\rho + i\eta) = \frac{A\lambda^3(\bar{\rho} + i\bar{\eta})\sqrt{1 - A^2\lambda^4}}{\sqrt{1-\lambda^2}\left[1 - A^2\lambda^4(\bar{\rho} + i\bar{\eta})\right]}, \tag{2.3}$$

where $\bar{\rho} = (1 - \lambda^2/2)\,\rho$ and $\bar{\eta} = (1 - \lambda^2/2)\,\eta$. This parameterization also ensures that

$$\bar{\rho} + i\bar{\eta} = -\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*} \tag{2.4}$$

is phase convention independent and it also allows to write the CKM matrix in terms of λ, A, $\bar{\rho}$ and $\bar{\eta}$. The matrix can be expanded in powers of λ and is unitary to all orders of λ. This parameterization is called the Wolfenstein parameterization [13] and commonly used:

$$V_{\mathrm{CKM}} = \begin{pmatrix} 1 - \lambda^2/2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1 - \lambda^2/2 & A\lambda^2 \\ A\lambda^3(1 - \rho - i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}\left(\lambda^4\right). \tag{2.5}$$

As $\lambda = s_{12} \approx 0.2$ it can be easily seen that the diagonal elements V_{ii} are close to 1, $|V_{us}| \simeq |V_{cd}| \approx 0.2$ and the remaining off diagonal elements are of $\mathcal{O}\left(10^{-3}\right)$.

Due to its unitarity, the CKM matrix allows a graphical interpretation. The unitarity requires that scalar products of two out of three columns or rows, respectively, vanish:

$$\sum_{i=1}^{3} = V_{ij}V_{ik}^* = \delta_{jk} \qquad \text{and} \qquad \sum_{j=1}^{3} = V_{ij}V_{kj}^* = \delta_{ik} \qquad \text{with } k = 1, 2, 3. \tag{2.6}$$

Each condition from equation 2.6 can be represented as triangle in a complex plane. The most commonly used triangle, also known as The Unitarity Triangle, arises from

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0. \tag{2.7}$$

It is illustrated in figure 2.2, where each side of equation 2.7 was divided by $V_{cd}V_{cb}^*$. This is the experimentally best known value and, therefore, this triangle allows the best constraint of the CKM matrix elements. In the Wolfenstein parameterization the triangles vertices are exactly at $(0,0)$, $(1,0)$ and $(\bar{\rho},\bar{\eta})$. The angles of the triangle are given by

$$\beta = \phi_1 = \arg\left(-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}\right), \tag{2.8}$$

$$\alpha = \phi_2 = \arg\left(-\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*}\right), \tag{2.9}$$

$$\gamma = \phi_3 = \arg\left(-\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*}\right). \tag{2.10}$$

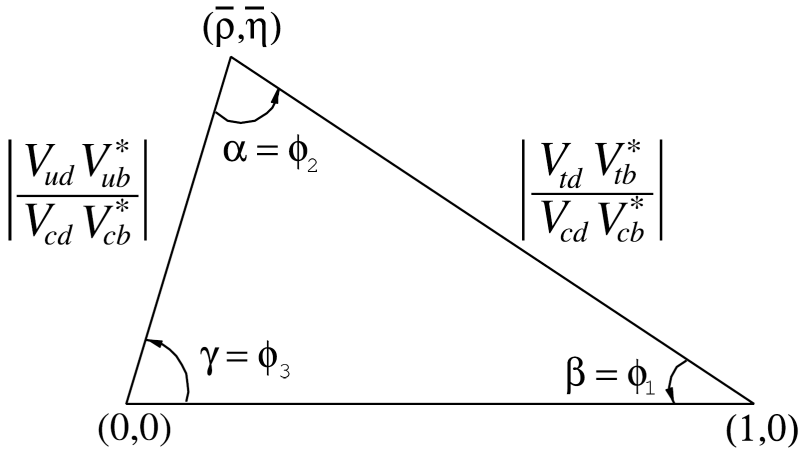

Figure 2.2: Graphical illustration of the unitarity triangle.

Over the years, various independent measurements of the triangle's angles and sides have been performed to determine the CKM matrix elements as precisely as possible. The CKMfitter group [14] as well as the UTfit group [15] are combining all these measurements in a global fit to constrain $\bar{\rho}$ and $\bar{\eta}$. As Standard Model constraints such as three generations of quarks are included, a significant deviation from unitarity would indicate new physics. However, no such deviation has been observed up to now. Recent results of the CKMfitter group from summer 2009 are shown in figure 2.3. Constraints from different measurements are shown in different colors. The red hashed region around the triangle's top corresponds to a 68% confidence level.

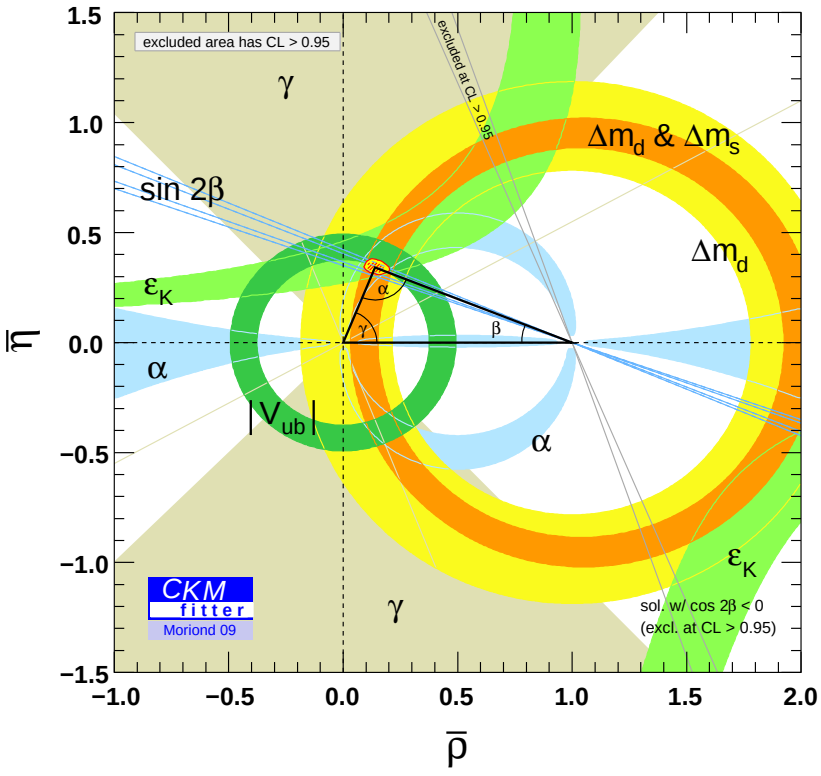

Figure 2.3: Global CKM fit in the $\bar{\rho} - \bar{\eta}$ plane [14].

2.3 Time Dependent CP Violation Measurement

One of the important types of measurements to constrain the unitarity triangle, by measuring $\sin 2\beta$, is the time dependent CP violation measurement. In neutral B meson decays, the interference between a decay with and without mixing leads to CP violation. This time dependent CP violation can be observerd as an asymmetry in the time dependent B^0 and $\overline{B}^0$ decay rates to final CP eigenstates f_{CP}:

$$\mathcal{A}_{f_{\mathrm{CP}}}(\Delta t) = \frac{\Gamma\left(\overline{B}^0(\Delta t) \to f_{\mathrm{CP}}\right) - \Gamma\left(B^0(\Delta t) \to f_{\mathrm{CP}}\right)}{\Gamma\left(\overline{B}^0(\Delta t) \to f_{\mathrm{CP}}\right) + \Gamma\left(B^0(\Delta t) \to f_{\mathrm{CP}}\right)}. \tag{2.11}$$

It can be seen in equation 2.11 that the flavor information, whether the B was B^0 or $\overline{B}^0$, is needed. As it can not be obtained from the decay products to the final CP eigenstate one uses flavor tagging algorithms to determine the B's flavor. Such algorithms are crucial for time dependent CP violation measurements.

Figure 2.4 illustrates the basic principles of a time dependent CP violation measurement. The $\Upsilon(4S)$ resonance decays to B-meson pairs. The mesons form an quantum mechanical entangled state. If the flavor of one B is known, the other B's flavor is automatically given. Therefore, if the first meson decays, the flavor of the second meson

has to be opposite at this moment. With time proceeding the second B could of course oscillate and change its flavor as the system is no longer entangled after the decay of the first B.

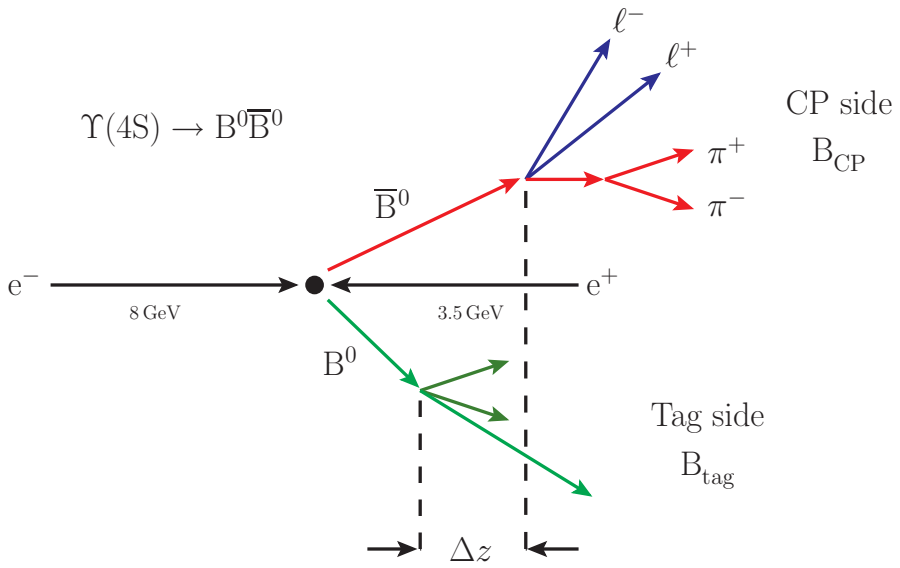

Figure 2.4: Schematic drawing of $\Upsilon(4S)$ decay and time dependent CP violation measurement.

First the signal or CP side B is reconstructed and its decay vertex position is determined. In figure 2.4 the decay $\overline{B}^0 \rightarrow J/\psi K_S^0$ is shown. Because it is a final CP eigenstate it is not clear whether the original B was B^0 or $\overline{B}^0$. In the next step the vertex position of the tag side B is determined from the remaining tracks. Finally the flavor of the tag side B has to be determined with a flavor tagging algorithm. The thesis topic of neutral B flavor tagging means therefore to distinguish whether the tag side B was B^0 or $\overline{B}^0$ at the time of its decay.

Knowing the B vertex positions one can obtain the decay time difference Δt from the difference between the decay points of the two B-mesons along the z-axis:

$$\Delta t = \frac{\Delta z}{\beta\gamma \cdot c}, \tag{2.12}$$

where $\beta\gamma$ is the boost of the system (see Chapter 3.2). Together with the information about the tag side B-meson's flavor one can then fit the asymmetry and measure the CP violation.

The whole procedure of a time dependent CP violation measurement is, of course, much more complicated than described here and various other effects (e.g. detector resolution, inefficiencies, etc...) have to be taken into account. However this section was only trying to give an overview and explain the placement and importance of flavor tagging within this widely-used type of measurement.

2.4 B-Meson Decay and Flavor Tagging

For the task of flavor tagging, the determination whether a meson was B^0 or $\overline{B}^0$ at the time of decay, it is essential to understand the decay modes of B mesons, which are given by the decay of the b quark. In this section the decay of B^0, composed of $\overline{b}d$, will be assumed. The $\overline{B}^0$ decay is similar and can be derived by charge conjugation, neglecting small differences due to direct CP violation.

In figure 2.5 a possible decay chain of the $\overline{b}$ quark is shown. In this figure the spectator quark d is not drawn, however, it takes part in all decay processes. On the basis of flavor, charge and momentum of the final state particles it is possible to determine the B meson flavor. Sometimes the tag side B is also called associated B and written as B_{asc}.

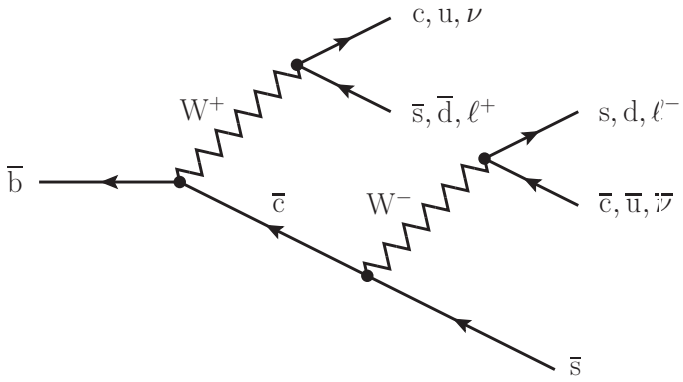

Figure 2.5: Possible decay chain of a $\overline{b}$ quark.

There are several flavor specific decay modes of the $\overline{b}$ quark that can be used to determine its flavor:

1. The charge of leptons from $\overline{b} \to X\ell^+\nu$ decay can be used to determine the flavor. On average it is the lepton with the highest momentum.

2. The charge of leptons from $\overline{b} \to \overline{c} \to \overline{s}\ell^-\overline{\nu}$ decays is opposite to those directly coming from $\overline{b}$. Those leptons arise from the $\overline{c}$ decay and on average they have intermediate momentum.

3. The cascade process $\overline{b} \to \overline{c} \to \overline{s}$ is the dominant decay for $\overline{b}$ as the CKM elements V_{cb} and V_{cs} are big compared to other possible transitions. Therefore K^+ mesons, composed of $\overline{s}u$, are very likely to be found in the final state.

4. It is also possible, that in the cascade decay $\overline{b} \to \overline{c} \to \overline{s}$ a $\overline{\Lambda}$ baryon, composed of $\overline{u}\overline{d}\overline{s}$, is formed during fragmentation. The branching fraction is very low compared to item 3 but due to the unique V shape of the decay and the occurrence of a proton in the final state it can be used for tagging the B.

5. It is also possible to tag the flavor with high momentum π^+, composed of $u\overline{d}$, that comes from $\overline{b}$ decay. Such pions can be found as final state particles of $B^0 \to D\pi^+X$ or $B^0 \to D^*\pi^+X$ decays.

6. In $\overline{b} \to \overline{c}$ it is possible that the $\overline{c}$ quark and a d quark form an excited D^{*-} meson. This excitation decays immediately via strong interaction $D^{*-} \to \overline{D}^0\pi^-$. The decay has only very limited phase space and therefore the pion momentum is very slow. The term slow pion tag is used and sometimes π_s is written.

There is no perfect flavor tagging algorithm which can always determine the flavor of the B meson from its final state particles. For instance, there can be misidentified particles or inefficiencies in particle detection, B mesons can decay to non flavor specific final states or non dominant physical processes could indicate a flavor opposite to the true one. In Figure 2.6 such a non dominant double Cabibbo suppressed (DCS) decay is shown. The charge of the on average high momentum π^- would indicate a $\overline{B}^0$ meson, whereas the true flavor was a B^0. On the other hand the K^+ would indicate the true flavor although it doesn't arise from the most probable cascade decay $\overline{b} \to \overline{c} \to \overline{s}$.

This simple example shows that flavor tagging is not a simple task. Correlations between different decay channels have to be taken into account. Based on a statistical approach a flavor tagging algorithm can only determine which flavor is more likely.

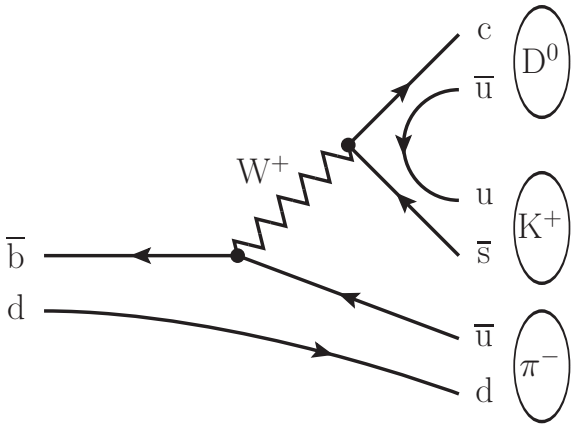

Figure 2.6: Double Cabibbo suppressed B^0 decay channel.

2.5 B-Meson Mixing and Wrong Tag Fraction

Charged weak currents in the Standard Model allow a transition between different quark flavors. This gives neutral B mesons the ability to oscillate into their own anti-particles. This process is a second order weak interaction induced by $W^\pm$ exchange. This loop process is illustrated in figure 2.7 in so-called box-diagrams. Each coupling of a $W^\pm$ boson is proportional to the corresponding element of the CKM matrix. In the Wolfenstein parameterization (see equation 2.5) up, charm and top quark transitions

are of order λ^3. But due to the huge mass difference of those quarks the process is dominated by virtual top quark transitions inside the box. Only small contributions arise from charm and up quarks.

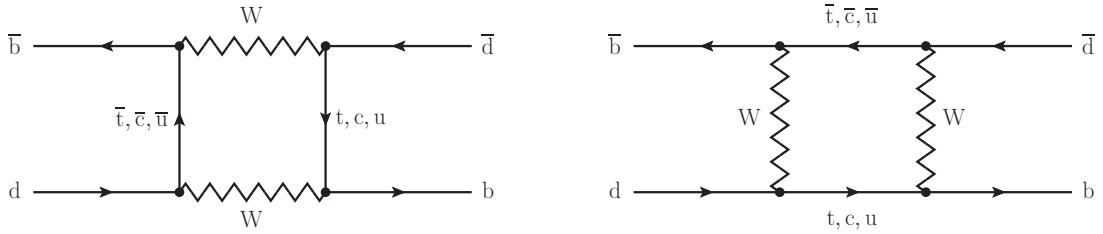

Figure 2.7: The two lowest order Feynman box-diagrams for B^0 mixing.

Experimentally such mixing was observed [16] when in a single event the decay products only allowed the conclusion that there must have been B^0B^0 or $\overline{B}^0\overline{B}^0$ at the time of decay, although at the time of production $B^0\overline{B}^0$ was guaranteed. Within the Standard Model only oscillation of the B mesons could explain this.

If the signal side B does not decay into a CP eigenstate but via a flavor specific decay channel, such as $B^0 \to D^*(2010)^-\ell^+\nu$, the flavor of the signal or reconstructed B is fixed by its final state particles. The other B's flavor is determined by a flavor tagging method. Events can be classified either as SF, if both mesons have the **s**ame **f**lavor, or as OF if both have **o**pposite **f**lavor. The probability to observe either SF or OF is given by

$$\mathcal{P}_{SF} = \frac{1}{2c\tau_{B^0}} \exp\left(-\frac{\Delta t}{c\tau_{B^0}}\right)\left(1 + \cos\left(\Delta m_d \Delta t\right)\right), \tag{2.13}$$

$$\mathcal{P}_{OF} = \frac{1}{2c\tau_{B^0}} \exp\left(-\frac{\Delta t}{c\tau_{B^0}}\right)\left(1 - \cos\left(\Delta m_d \Delta t\right)\right), \tag{2.14}$$

where τ_{B^0} is the B^0 mean life time and Δm_d the mass difference between the eigenvalues of the mass eigenstates $|B^H\rangle$ and $|B^L\rangle$. Δt is the proper decay time difference and, as described in section 2.3, it can be obtained from the decay length difference Δz. The mass eigenstates in the base of flavor eigenstate are

$$|B^L\rangle = p|B^0\rangle + q|\overline{B}^0\rangle \quad \text{and} \quad |B^H\rangle = p|B^0\rangle - q|\overline{B}^0\rangle \quad \text{with} \quad |p|^2 + |q|^2 = 1. \tag{2.15}$$

$\mathcal{P}_{SF}$ and $\mathcal{P}_{OF}$ both include the usual exponential decay of instable particles, such as B mesons. The second part of the equations arises from the Schroedinger equation and the time dependent evolution of the entangled $B^0\overline{B}^0$ state.

As explained in the last section, there is no perfect flavor tagging algorithm and therefore flavor misidentification has to be taken into account. This misidentification rate is called the wrong tag fraction w. So the experimental observed $\mathcal{P}^{rec}_{SF}$ and $\mathcal{P}^{rec}_{OF}$ become

$$\mathcal{P}^{rec}_{SF} = (1-w)\mathcal{P}_{OF} + w\mathcal{P}_{SF} \;\propto\; 1 + (1-2w)\cos\left(\Delta m_d \Delta t\right), \tag{2.16}$$

$$\mathcal{P}^{rec}_{OF} = w\mathcal{P}_{OF} + (1-w)\mathcal{P}_{SF} \;\propto\; 1 - (1-2w)\cos\left(\Delta m_d \Delta t\right). \tag{2.17}$$

The mixing induced asymmetry between SF and OF final state can be written as

$$\mathcal{A}_{mix}^{raw} = \frac{\mathcal{P}_{OF} - \mathcal{P}_{SF}}{\mathcal{P}_{OF} + \mathcal{P}_{SF}} = (1 - 2w)\cos\left(\Delta m_d \Delta t\right). \tag{2.18}$$

By fitting this distribution in data, one can not only obtain the mass difference Δm_d but also the wrong tag fraction w of the tagging algorithm. The amplitude of the asymmetry is affected by the wrong tag fraction. Figure 2.8 illustrates this for different values of w. This fit allows one to do a validation of a given flavor tagging algorithm on data only.

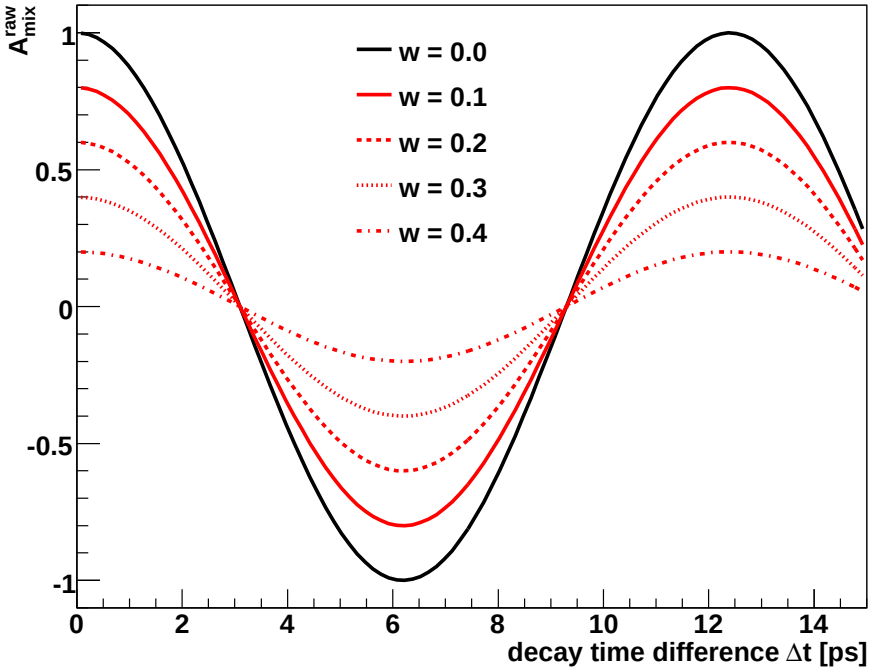

Figure 2.8: Illustration of the influence of wrong tag fraction w on the asymmetry.

2.6 Effective Efficiency and Dilution

Not only the amplitude of the mixing induced asymmetry $\mathcal{A}_{mix}^{raw}$ (equation 2.18) is affected by the wrong tag fraction. Also the amplitude of the asymmetry in the time dependent B^0 and $\overline{B}^0$ decay rates to final CP eigenstates $\mathcal{A}_{f_{\mathrm{CP}}}$ (equation 2.11) is diluted. The observed asymmetry is given by

$$\mathcal{A}_{f_{\mathrm{CP}}}^{obs} = (1 - 2w)\mathcal{A}_{f_{\mathrm{CP}}}, \tag{2.19}$$

where the term $(1-2w)$ is often called dilution factor or dilution. The statistical error on the observed asymmetry is given by

$$\sigma_{\mathcal{A}_{f_{\mathrm{CP}}}^{obs}} \propto \frac{1}{\sqrt{\epsilon_{tag} N_{rec}}}, \tag{2.20}$$

where ϵ_{tag} is the efficiency of the tagging algorithm and N_{rec} the number of reconstructed events on the signal side of the time dependent CP violation measurement. From equation 2.19 and 2.20 it follows, by using Gaussian error propagation, that the statistical error on the real asymmetry is given by

$$\sigma_{\mathcal{A}_{f_{\mathrm{CP}}}} \propto \frac{1}{\sqrt{\epsilon_{tag}(1-2w)^2 N_{rec}}}, \tag{2.21}$$

where $\epsilon_{eff} \equiv \epsilon_{tag}(1-2w)^2$ is defined as effective efficiency. Therefore it is obvious, that the effective efficiency of the tagging algorithm has to be maximized to increase the significance of the results of the time dependent CP violation measurement. The tagging algorithm should be able to tag the flavor of the tag side B in every event as precisely as possible.

3 Experimental Setup

High energy particle physics requires a complicated and expensive experimental setup. Usually a set of accelerators is needed to accelerate stable charged particles with electric fields to a desired energy. With bending and focussing systems that use magnetic fields, these particles are brought to collision at a specific location called the interaction region. For the purpose of data analysis, a detector is built around this location to detect and record the properties of the products of the interaction.

3.1 Basic Principles of a Collider

3.1.1 Energy

Everyone has heard of Albert Einstein's famous equation [17],

$$E = mc^2, \tag{3.1}$$

which states that energy E and mass m are equal and can be transformed into each other. The squared speed of light, c^2, is just a multiplicative factor. This relation can be exploited to produce new particles and the energy threshold is therefore given by the particle's mass.

To reach this threshold, charged particles are accelerated almost to the speed of light, using electric fields. When their kinetic energy is high enough they are brought to head-on collision. In the center of mass frame of this collision the available energy $\sqrt{s}$ can be calculated as the sum of the two particles' four momenta $\mathbf{p}_1$ and $\mathbf{p}_2$:

$$s = \left(\mathbf{p}_1 + \mathbf{p}_2\right)^2, \tag{3.2}$$

$$s = m_1^2c^4 + m_2^2c^4 + 2\left(E_1E_2 - \vec{p}_1\vec{p}_2c^2\right), \tag{3.3}$$

$$\sqrt{s} = \sqrt{4E_1E_2}. \tag{3.4}$$

where the approximation $m_ic^2 \ll E_i$ was made and $p_ic = \sqrt{E_i^2 - m_i^2c^4}$ was used. In general all kinds of charged particles could be used. However, in high energy particle physics it is common to use only stable particles to avoid decays within the acceleration process. In this context stable charged particles are electrons, positrons, protons and anti-protons. Accelerators can therefore be divided into three groups:

Electron-positron colliders such as LEP [18] or KEKB. Their advantage is that the colliding particles do not have a substructure. The initial state and the available energy in the center of mass system are well known. When built as circular

colliders they are practically limited in maximal energy. This limit is given by synchrotron radiation which is emitted by any charged particle bend in magnetic fields. The energy loss per turn can be expressed as

$$\Delta E = \frac{1}{3}\left(\frac{e^2\beta^3\gamma^4}{\rho}\right), \tag{3.5}$$

where e is electric charge, β the velocity of the charged particle and the bending radius ρ. With $\gamma = {}^{E}/_{mc^2}$ it follows that

$$\Delta E \sim \frac{1}{m^4} \tag{3.6}$$

and, therefore, lighter particles like electrons and positrons have a much higher energy loss then heavy particles like protons. The LEP collider with its beam energy of $\sim 100\,\mathrm{GeV}$ has had a loss of $\sim 2.9\,\mathrm{GeV}$ per turn. Therefore the next high energy electron-positron collider is supposed to be a linear collider, which does not have this disadvantage.

Hadron colliders usually collide protons with protons (LHC) or protons with anti-protons (Tevatron)[19]. Their loss due to synchrotron radiation is much smaller and therefore they are used to achieve the highest possible energies. Due to the substructure of the colliding particles, only constituents of each hadron interact with each other. These constituents carry only a fraction of the entire hadron energy. Therefore the initial state is not well known and, due to the fragmentation in the strong interaction, the multiplicity in the events is much higher. In figure 3.1 one can easily see this difference in e.g. a typical event recorded by the CDF [20] detector (hadronic $p\overline{p}$ interaction) and one recorded by the Belle detector (leptonic e^-e^+ interaction).

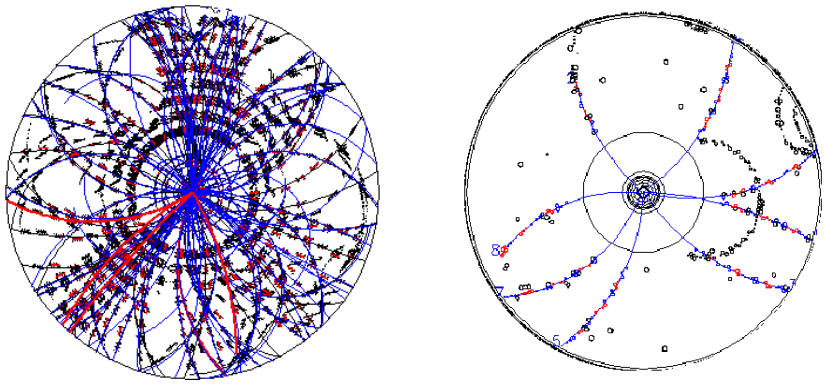

Figure 3.1: Typical CDF event (left) compared to typical Belle event (right).

Hadron-electron colliders are rare and apart from fix target experiments, the HERA [21] experiment was the only hadron-electron collider built and was located at the Deutsches Elektron Synchrotron (DESY) in Hamburg, Germany. It has made important contributions to the measurement of the proton substructure which were only possible due to its unique design.

3.1.2 Luminosity

As explained in the last section, energy determines whether a certain production mechanism is possible at all. Yet there is no information about how often a certain process occurs. The probability of a physics process at a given energy can be calculated and is known as a cross section σ. After accumulating data for some time, this process can be found N times recorded in data. The interaction rate $^{dN}/_{dt}$ of a certain process is given by

$$\frac{dN}{dt} = \mathcal{L} \cdot \sigma, \tag{3.7}$$

where $\mathcal{L}$ is the luminosity of the accelerator. When integrated over time one can directly obtain the number of events of a given process in the accumulated data of an experiment:

$$N = \int \mathcal{L} \cdot \sigma \; dt \; . \tag{3.8}$$

In theory we collide two single particles whereas in practice we have two particle beams. Each beam can consist of up to thousands of bunches, which themselves hold up to millions of particles. Such a topology is produced by the use of high frequency electric fields for the acceleration. During a collision one bunch from the first beam is crossing with one from the second beam. The luminosity of a collider depends only on the properties of the beam and is given by

$$\mathcal{L} = \frac{N_1 N_2 f}{4\pi\sigma_x\sigma_y}, \tag{3.9}$$

where $N_{1,2}$ are the numbers of particles in each bunch, $\sigma_{x,y}$ are the spatial dimensions of the bunches and f is the bunch crossing or collision rate.

3.2 KEKB Accelerator

The KEKB accelerator is an asymmetric electron-positron collider [22, 23] at Tsukuba, Japan, northwest of Tokyo. It was designed as a B-Factory whose main goal is to achieve a maximum production rate for B-meson pairs, i.e. $B^0\overline{B}^0$ and B^+B^- pairs. Figure 3.2 shows a schematic drawing of the KEKB accelerator complex.

Electrons and positrons are accelerated in a linear accelerator (Linac). Positrons are then filled into the low energy ring (LER) with energy $E_+ = 3.5\,\text{GeV}$ and a positron current of about 1600 mA. Electrons are filled into the high energy ring (HER) with energy $E_- = 8.0\,\text{GeV}$ and a electron current of about 1200 mA. The circumference of both rings is 3016 m. There is one interaction region (IR) at Tsukuba hall, where the Belle detector is located. In total 1584 bunches are filled into each beam. The bunch crossing rate is 509 MHz. Bunches are crossing at a finite angle of 22 mrad.

To keep the effective crossing area $4\pi\sigma_x\sigma_y$ as small as possible, bunches need to collide head-on. Using crab cavities, bunches get rotated in the interaction region shortly before the collision. This way they collide head-on in spite of the finite crossing angle.

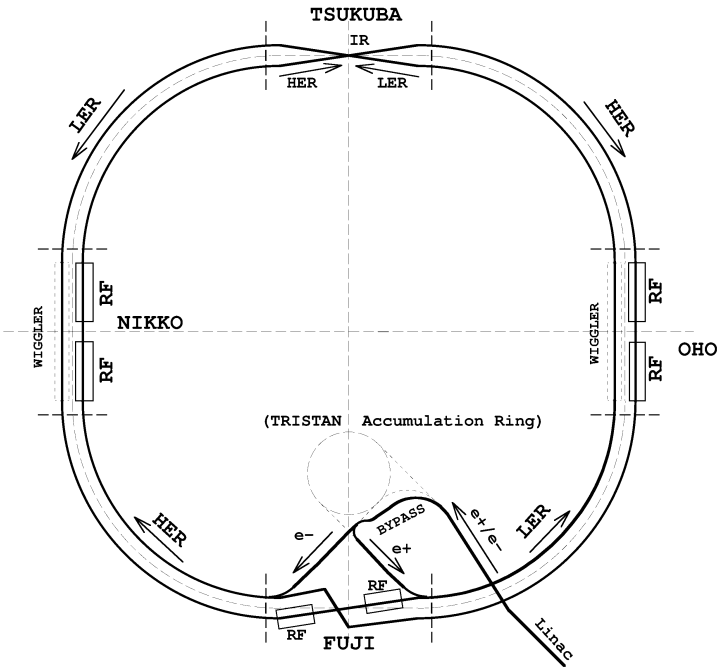

Figure 3.2: Schematic layout of KEKB accelerator complex.

The energy of the beams was chosen so that the resulting center of mass energy $\sqrt{s} = 10.58\,\text{GeV}$ corresponds to the mass of the $\Upsilon(4S)$ resonance:

$$\sqrt{s} = \sqrt{4E_+E_-} \approx 10.58\,\text{GeV} = m_{\Upsilon(4S)}. \tag{3.10}$$

This resonance, a bound state of $b\bar{b}$ quarks, is just above the threshold of B-meson pair production and decays in $\approx 96\%$ of cases [6] into B-meson pairs, which are almost at rest in the center of mass (CMS) frame of the $\Upsilon(4S)$ resonance:

$$\vec{p}_{B,\text{CMS}} \approx 0. \tag{3.11}$$

Due to the asymmetric beam energy, the B-meson pairs are boosted into the direction of the HER. The Lorentz-boost parameter $\beta\gamma$ of the system is given by

$$\beta\gamma = \frac{E_- - E_+}{\sqrt{s}} = 0.425 \tag{3.12}$$

and results in non-zero momenta for the B-meson pairs in the laboratory frame:

$$\vec{p}_{B,\text{lab}} \neq 0. \tag{3.13}$$

Therefore B-mesons can travel a measurable finite distance before decaying, thus allowing the Belle collaboration to measure their decay time.

It is notable that with the given configuration [24] the KEKB accelerator achieved a world record in luminosity of $2.11 \cdot 10^{34}\,\text{cm}^{-2}\text{s}^{-1}$ in June 2009. This is more than twice its design luminosity and allows the Belle collaboration to study very rare decay channels. Since Belle started to take data the integrated luminosity has reached nearly $1\,\text{ab}^{-1}$.

3.3 The Belle Detector

The Belle detector[25] is a particle detector, designed and constructed to perform high precision time dependent CP violation measurements and studies of rare B-meson decays. It is built around the interaction region of KEKB.

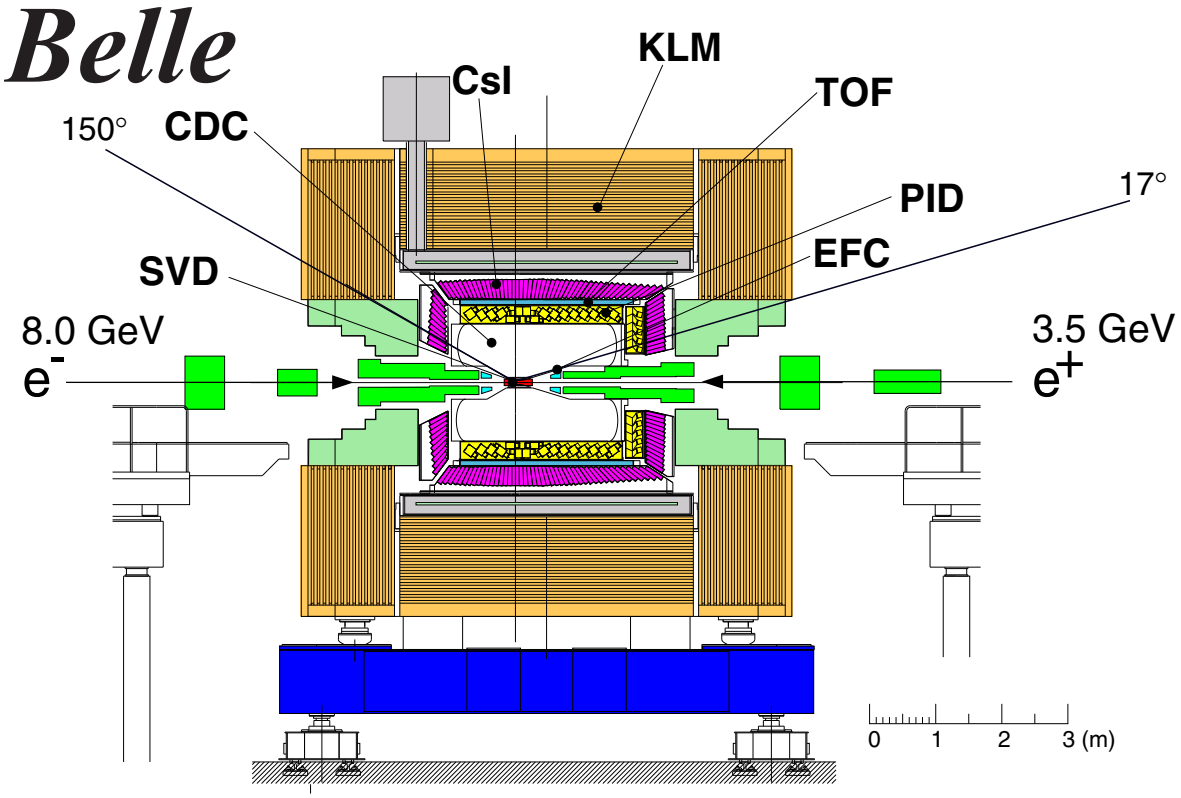

Figure 3.3: Side view of the Belle detector.

Figure 3.3 shows the layout of the Belle detector. The detector is constructed around the KEKB beam pipe. It has an iron structure, which is used as a yoke for a superconducting solenoid, which provides a magnetic field of 1.5 T. A silicon vertex detector (SVD) around the beam pipe is used to measure vertices of decaying particles. Charged particles are bent within the magnetic field and their momenta is measured from the curvature of their reconstructed tracks in the central drift chamber (CDC). Measurements of $^{dE}/_{dx}$ from the CDC, together with a photon yield from the aerogel threshold Cherenkov counter (ACC) and a time of flight measurement from a time of flight counter (TOF), are used for particle identification (PID). Both ACC and TOF are situated outside the CDC. Figure 3.4 shows the arrangement of the inner systems used for particle identification in more detail. Electromagnetic showers are detected in an electromagnetic calorimeter made of CsI(Tl) crystals. An additional extreme forward calorimeter (EFC) is situated close to the interaction point. Outside the solenoid, but built in the iron yoke, are resistive plate counters for K_L^0 and $\mu^\pm$ detection (KLM).

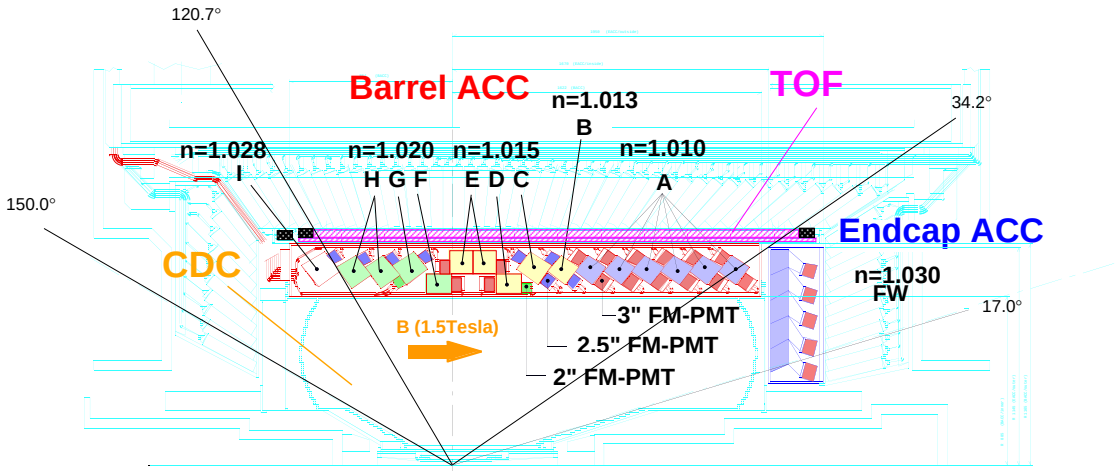

Figure 3.4: Schematic drawing of the inner region of the Belle detector used for PID.

The above mentioned systems will be described below in more detail. The coordinate system used is defined in a way so that the positive z-axis points in the direction of the HER beam. This is also called the forward direction. The x-axis points out of the accelerator plane and the y-axis is perpendicular to the x- and z-axis and lies within the accelerator plane. The angle θ is measured with respect to the z-axis and ϕ is measured with respect to the x-axis.

Beam pipe Before particles reach the detector, they have to pass the beam pipe. Particles can multiple Coulomb scatter in the beam-pipe wall, which effects z-vertex resolution. The beam pipe is also exposed to beam-induced heating of a few hundred watts. Therefore a double wall beryllium cylinder design was chosen with each wall having a thickness of $d = 0.5\,\text{mm}$, a gap of 2.5 mm between both walls and an inner diameter of 40 mm. The gap is filled with helium-gas for cooling the beam pipe. Helium was chosen instead of water to minimize the material in the beam pipe.

EFC The extreme forward calorimeter was installed to increase the detector's polar angle coverage by the ECL in the extreme forward and backward direction. The EFC covers the polar angular range from $6.4° < \theta < 11.5°$ in the forward and $163.3° < \theta < 171.2°$ in the backward direction. Due to its location near the interaction point it has to be radiation-hard. A Bismuth Germanate ($Bi_4Ge_3O_{12}$) calorimeter was chosen to fulfil the requirements to radiation-hardness and simultaneously provide an excellent e/γ energy resolution of

$$\frac{\sigma_E}{E} = \frac{(0.3-1)\%}{\sqrt{E[\text{GeV}]}}.$$

SVD To measure time dependent CP violation in B-meson decays a z-vertex resolution of $\sim 100\,\mu$m is necessary. This resolution also allows one to use the vertex detector for D-meson and τ identification. In the beginning of data-taking a three layer silicon vertex detector (SVD1) was installed. It provided a polar angular coverage of $23° < \theta < 139°$ and the radii of the three layers were:

$$r_1 = 30.0\,\text{mm}, \quad r_2 = 45.5\,\text{mm}, \quad r_3 = 60.5\,\text{mm}.$$

Due to massive radiation damage, it was replaced with the SVD2[26] in 2003. The new SVD2 has a polar angular coverage of $17° < \theta < 150°$ and 4 layers with the radii:

$$r_1 = 20.0\,\text{mm}, \quad r_2 = 43.5\,\text{mm}, \quad r_3 = 70.0\,\text{mm}, \quad r_4 = 88.0\,\text{mm}.$$

With layers closer to the interaction point it has a better z-vertex resolution of $\sigma_{\Delta z} \sim 80\,\mu$m and covers the full nominal angular coverage of the Belle detector.

CDC Physics measurements without the central drift chamber would be impossible. It allows the efficient reconstruction of charged tracks and the measurement of their momenta. In addition $^{dE}/_{dx}$ of each track is measured in the CDC and is used in particle identification. The CDC has 50 layers of either axial or stereo wires that are arranged cylindrically around the beam axis. This configuration creates 8400 drift cells, each with a maximum drift distance of 8-10 mm and a radial thickness of 15.5-17 mm. It covers the polar angular range of $17° < \theta < 150°$. To minimize multiple Coulomb scattering a low-Z gas was chosen, namely a 50% helium and 50% ethane gas mixture. The large ethane component provides a good $^{dE}/_{dx}$ measurement. The CDC resolution parameters are:

$$\sigma_{r\phi} = 130\,\mu\text{m}, \quad \sigma_z = 200 - 1400\,\mu\text{m}, \quad \frac{\sigma_{p_t}}{p_t} = 0.3\%\sqrt{p_t^2[\text{GeV}] + 1}, \quad \frac{\sigma_{\frac{dE}{dx}}}{\frac{dE}{dx}} = 6\%.$$

ACC The aerogel Cherenkov counter is installed outside the CDC in the barrel and forward region. Only light particles like pions in the momentum range $1.2\,\text{GeV}/c < p < 3.5\,\text{GeV}/c$ emit Cherenkov radiation in the scintillators as they travel faster than the speed of light in the material the scintillator was made of, whereas kaons do not. Therefore the system is providing additional information for $K^\pm$ and $\pi^\pm$ separation. It consists of 960 counter modules in the ϕ direction and 228 modules in the forward direction. Fine mesh photomultipliers are used to read out the modules. They create a certain amount of photoelectrons $N_{pe} \geq 6$, which is high enough to not worry about photomultiplier efficiency.

TOF The time of flight detector is located outsite the ACC in the barrel region. Complementary to the ACC it provides information for particle identification in the momentum range of $p < 1.2\,\text{GeV}/c$. With a time resolution of $\sigma_t \sim 100\,\text{ps}$ it measures the time particles need to reach the TOF, which is 1.2 m away from the interaction point. Knowing flight length and travel time, one can caluclate

the particle's mass. The TOF consists of 64 scintillator modules, which are read out by fine mesh photomultipliers, and covers the polar angular range of $34° < \theta < 120°$.

ECL The electromagnetic calorimeter is used mainly for measuring the energy and position of electrons and photons with high efficiency and resolution. It has to cover the full energy range of 500 MeV photons at the end of a decay chain up to 4 GeV in two-body decay modes such as $B \to K^*\gamma$ or $B \to \pi^0\pi^0$. For π^0 detection a good angular resolution is needed to distinguish two nearby photons from each other. To satisfy this requirements a CsI(Tl) crystal calorimeter was chosen. It has a high segmentation, 8736 CsI(Tl) counters were installed in total, and is read out by silicon photodiodes. The polar angular range from $12.4° < \theta < 155.1°$ is covered. However, small gaps between barrel and end-cap regions are unavoidable due to construction reasons. This lowers the acceptance in the covered region by $\sim 3\%$. The ECL resolution parameters are:

$$\frac{\sigma_E}{E[\mathrm{GeV}]} = \frac{1.3\%}{\sqrt{E[\mathrm{GeV}]}}, \qquad \sigma_{pos} = \frac{0.5\,\mathrm{cm}}{\sqrt{E[\mathrm{GeV}]}}.$$

Magnet Up to now all systems were installed inside the superconducting solenoid, which provides a 1.5 T magnetic field and is made of NbTi/Cu. Charged tracks get bent due to the magnetic field and allow a measurement of their momenta by means of curvature inside the CDC. The magnet has a cylindrical volume of 3.4 m in diameter and 4.4 m in length. A liquid helium cryostat is installed around the solenoid to reach superconducting temperatures. The cool down time is about 6 days and it takes half an hour to charge the magnet to a nominal current of 4400 A. The iron structure of the Belle detector outside the magnet serves as a return path for the magnetic flux.

KLM The K_L^0 and $\mu^\pm$ detection system is installed inside the iron structure. The latter serves not only as a yoke for the magnet but also as absorber material for the KLM system. 14 iron plates, each 4.7 cm thick provide a total of 3.9 interaction lengths for particles travelling the plates perpendicular. Together with the ECL this sums up to 4.7 interaction lengths to convert K_L^0 particles. The direction of those showers is measured by 15 layers of glass-electrode-resistive plate counters, which detect charged particles. This also allows to identify weakly interacting muons very well. They get deflected by multiple scattering but still travel much further without being absorbed than charged hadrons such as $\pi^\pm$ or $K^\pm$. The KLM covers the polar angular range of $20° < \theta < 155°$ and has an angular resolution of 30 mrad in θ and ϕ direction.

The scientific goals of the Belle collaboration were defined in 1994. In 1995 the design report was written and after some design changes in 1997 Belle started to take data in 1999. In 2001 the Belle experiment measured time dependent CP violation in B-meson decays [4] and simultaneously the BaBar [27] experiment measured the same [28]. Those two results lead to the proof of the Kobayashi-Maskawa theory [5], which earned the Nobel prize in physics 2008.

4 Neural Networks

As often in science, the requirements on the techniques used are very demanding in high energy particle physics. For the purpose of data analysis, these techniques are of statistical nature. For classification problems the use of multivariate methods has many advantages compared to classical methods. However, the success of such complex methods rises and falls with their implementation. The artificial neural network package NeuroBayes® has shown its robustness and performance in various applications in the past.

4.1 Classification Problems

In data analysis one often has to deal with classification problems, e.g whether a particle was kaon or pion or in case of flavor tagging whether it was B^0 or $\overline{B}^0$. Generally speaking one wants to classify each event in a set of data to be either signal or background. Usually such a classification is made with a selection on a set of input variables, which carry information about the target variable.

To characterize the performance of a classification method one has to regard efficiency ϵ and purity p of the method:

$$\epsilon = \frac{N(\text{selected signal events})}{N(\text{total amount of signal events in data set})}, \tag{4.1}$$

$$p = \frac{N(\text{selected signal events})}{N(\text{total amount of selected events})}. \tag{4.2}$$

A classification method should ideally have maximum efficiency, i.e. selecting all signal events, and maximum purity, i.e. selecting only signal events. In figure 4.1 a simple example of three different selections is shown. The green dashed lines show a simple cut based selection in the 2-dimensional space of input variables. It is obvious that the selected region contains all signal events (red) but also lots of background events (blue). Reducing the amount of background by changing the cuts will automatically also reduce the efficiency. A more sophisticated method would be to first rotate the input variables and then perform a cut. This is shown by the orange dashed lines. The whole signal region was still selected but the amount of background has been reduced compared to the cut based selection. However, even such methods do not take into account the correlations between different input variables. Multivariate methods can do that and the result of such a selection is shown by the black dashed line. The pair of efficiency and purity can be optimized at the same time.

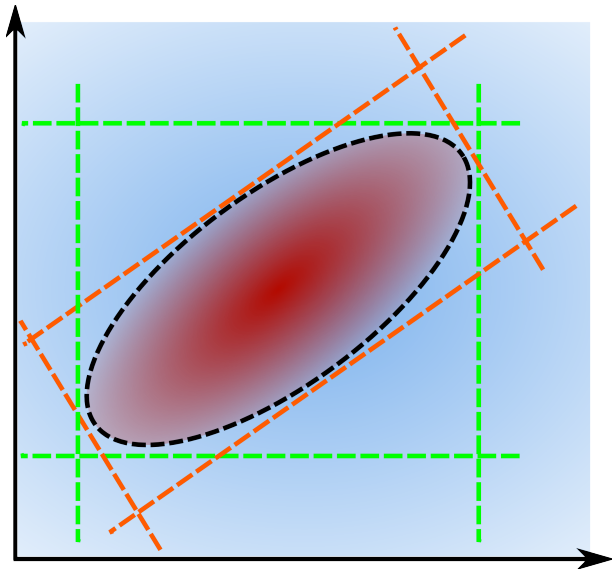

Figure 4.1: Simple example of different selections (dashed lines) for selecting signal (red) from background (blue) in 2-dimensional space.

4.2 Artificial Neural Networks

In general, the task of a multivariate method is to map the n-dimensional input variable space onto a single scalar which then contains all information and correlations of the input variables. This single variable can then be used as a selection variable instead of cutting on several input variables individually. Such mapping can be realized by an artificial neural network (ANN).

The basic element of an artificial neural network is the artificial neuron. Its design is inspired by the biological neuron in the brain. This cell can be divided into three parts: An input structure, called dendrites, a cell body and an output structure, called axon. The neurons in the brain communicate among each other by exchanging electrical impulses. They are fired by the axon of one cell and received by the dendrites of other neurons. If the power of the electrical signal exceeds a certain threshold, the neurons are triggered and fire output signals themselves along their axons. The junctions between axons and dendrites of different cells are called synapses. These are able to increase or decrease the electrical impulse, when received at a dendrite, using a biochemical process.

The layout of an artificial neuron is shown in figure 4.2. The input structure consists of an input vector $\vec{x}_i$. The weighted sum of the input vector is calculated, where the weights w_{ij} are in analogy to the synapses. A bias θ_j controls the signal threshold of the neuron. The output o_j of the neuron is given by the activation or transfer function φ. Often the symmetric sigmoid function

$$\varphi = S(x) = \frac{2}{1 - e^{-x}} - 1 \tag{4.3}$$

is used which is symmetric to the point of origin and maps the interval $]-\infty, +\infty[$ to

$[-1, +1]$. The output o_j of the artificial neuron j, with n inputs x_i, each weighted by w_{ij} is therefore given by

$$o_j = S\left(\sum_{i=1}^{n} w_{ij}x_i - \theta_j\right). \tag{4.4}$$

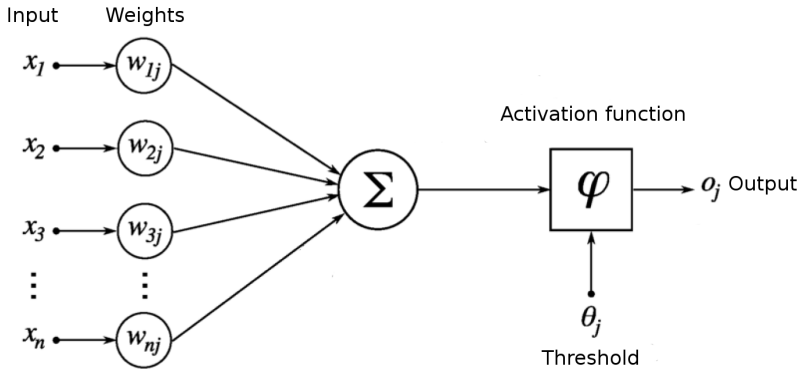

Figure 4.2: Layout of an artificial neuron.

4.3 Feed Forward Network

In the human brain, consisting of up to 30 billion neurons, each neuron has up to 10.000 synapses to other neurons. This structure is far too complicated for the purpose of data analysis. For classification problems the three layer feed forward network with a single output node is a sufficient choice. The information of the n input nodes x_i is transfered via m nodes y_i in a hidden layer to a single output node o. The term feed forward arises from the fact that information, in contrast to the brain, is only transfered in one direction. In the case of a three layer feed forward network the final output o can be calculated by

$$o = S\left(\sum_{j=1}^{m} w_j S\left(\sum_{i=1}^{n} w_{ij}x_i - \theta_j\right)\right), \tag{4.5}$$

where w_j is the weight of the connection of the hidden layer j to the output node, w_{ij} is the weight of the connection between input node i and hidden node j and θ_j is the signal threshold for the hidden node j. There is no signal threshold applied in the output node.

In general the number of input nodes is given by the number of variables, which carry information needed for the classification. The number of hidden nodes is arbitrary. However, with too many nodes the network can learn things by heart, called overtraining, and loose its ability to generalize. With too few nodes important information can not be learned. In various applications it was a good choice if the number of hidden nodes approximated the number of input nodes, thus giving the network the ability to learn important information but minimizing the risk of overtraining. In figure 4.3 the topology of a typical simple feed forward network is shown.

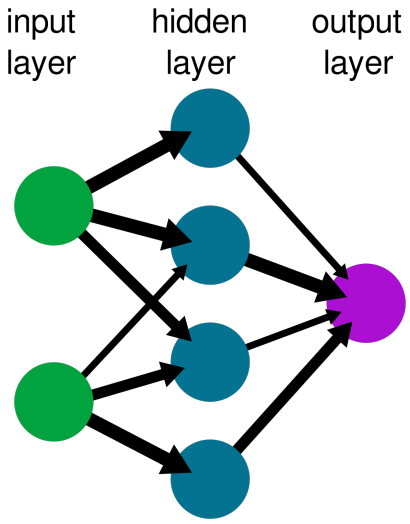

Figure 4.3: Topology of a simple feed forward network. The weight of each connection between nodes is indicated by the arrow's thickness.

4.4 Training a Network

To use a neural network for a classification problem it has to be trained first. For such training, one needs a set of data for which the truth is known. In general, this can be historical data, but in particle physics simulations are usually used. Using a data set for signal and background where the truth of each entry or event is known, the weights of each individual connection in the neural network are determined in a way that the network output for each event is as close to the known truth as possible. For a feed forward network with 5 input nodes, 5 hidden nodes and 1 output node, this means determination of $5 \times 5 + 5 \times 1 = 30$ weights.

To solve this problem, the weights are adjusted iteratively. The weights are usually adjusted using the whole training data set. For each event i the network output is computed with the current weights and compared to the truth t_i. This is usually expressed as a cost function. One possibility of a cost function is the sum over the squared difference between network output and truth or target value

$$\chi^2 = \sum_i^N \left(o_i\left(\vec{w}\right) - t_i\right)^2, \tag{4.6}$$

where N is the total number of training events and $\vec{w}$ is the vector of all weights. Another possible cost function is the entropy function

$$E = \sum_i^N \ln\left(\frac{1}{2}\left(1 + o_i\left(\vec{w}\right) \cdot t_i\right)\right). \tag{4.7}$$

In general, training a neural network is a non-trivial minimization problem in high dimensional space. To be more precise we minimize the cost function by adjusting each

weight. The gradient descent method can be used for this purpose and if the minimum has not been reached

$$\frac{\partial E}{\partial w_{ij}} \neq 0, \tag{4.8}$$

the change of weights Δw_{ij} is proportional to the respective gradient

$$\Delta w_{ij} = \eta \frac{\partial E}{\partial w_{ij}}, \tag{4.9}$$

where η is the proportionality constant. Starting with random weights, they are adjusted in each iteration of the training until the minimum of the cost function is reached. This kind of neural network training algorithm is called Backpropagation-Algorithm.

4.5 NeuroBayes®

One implementation of neural networks is called NeuroBayes® [29] and was originally developed at the University of Karlsruhe. It is now maintained and further developed by physicists in the company <phi-t>® – Physics Information Technologies [30], a spin-off, which transfered the algorithm beyond physics problems.

4.5.1 Bayes Theorem

As the name NeuroBayes® indicates, the package makes use of the Bayes' theorem [31]

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \tag{4.10}$$

which states that the conditional probability of A assuming a given B, $P(A|B)$, is connected to the conditional probability of B assuming a given A, $P(B|A)$. $P(A)$ is the a priori probability to measure A and $P(B)$ is the a priori probability to measure B, where $P(A)$ and $P(B)$ are independent of each other.

The aim of the neural network is to achieve an estimate of the conditional probability density function $f(t|\vec{x}_i)$ for a given input vector $\vec{x}_i$ and target t. Assuming that there is no information in the input vector, then $f(t|\vec{x}_i) = f(t)$. But if there is a correlation between input vector and target, $f(t|\vec{x}_i)$ should be a better estimator for event i than the inclusive distribution $f(t)$.

4.5.2 Preprocessing

In general a set of input data could directly be used to train a neural network. However, this can lead to various problems e.g. input values could have outliers or different input variables could have values of different orders of magnitude, leading to numerical problems, thus not creating the best conditions for finding the minimum of the cost function.

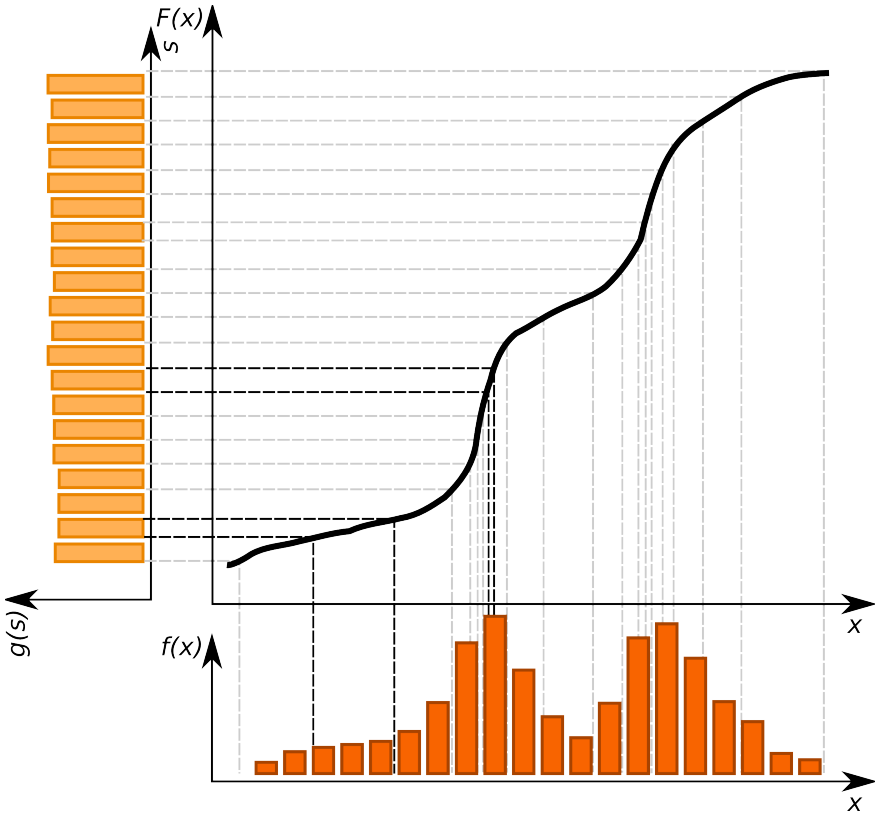

Figure 4.4: Transforming a distribution $f(x)$ via integral $F(x)$ to flat distribution $g(s)$.

NeuroBayes® provides a set of preprocessing options, which can be chosen individually for each input variable. In general the input data distribution is flattened first. This is illustrated in figure 4.4, where the integral $F(x)$ of the original distribution $f(x)$ is calculated and then used to create a distribution $g(s)$, which has equally filled bins. By calculating the purity in each bin, one can map this distribution to the interval $[0, +1]$. To minimize fluctuations, the distribution can be optionally fitted with a spline function. Finally, the distribution is transformed to a distribution with mean 0 and width 1. NeuroBayes® can also deal with special values, e.g. missing values, by setting them to a δ-function. Plots, which allow an evaluation of these steps, are created automatically and stored in an analysis file. Examples are shown in figure 4.5.

To decorrelate the transformed input variables, the covariance matrix of the transformed input variables is diagonalized by a series of rotations. This gives benefit to the minimization process.

NeuroBayes® also calculates the significance of each input variable. Its correlation to the target value is obtained during preprocessing and the correlations between all variables are calculated. A color coded plot of the correlation matrix, shown in figure 4.6, is added to the analysis file and numerical values of significance and correlation to the target are given in the output.

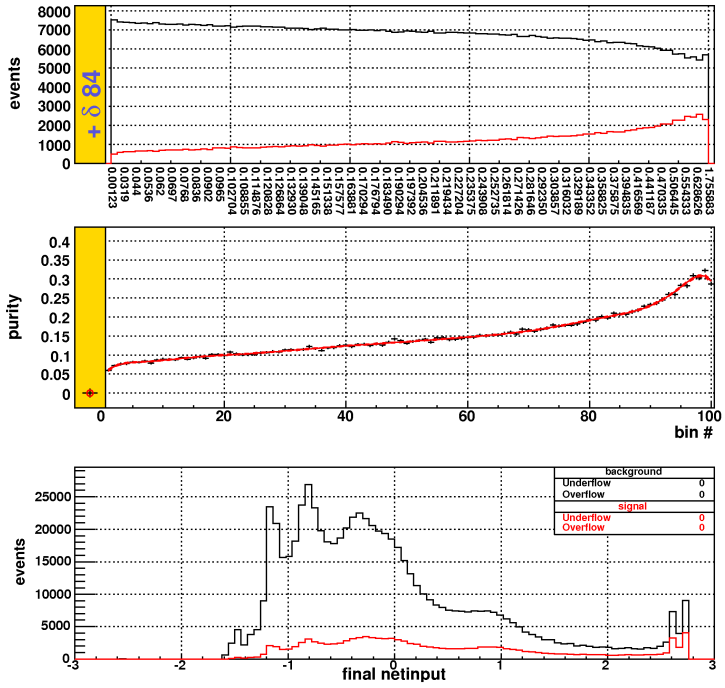

Figure 4.5: Preprocessing plots from analysis file. (top) Flattened distribution with same amount of signal (red) + background (black) in each bin. Values set to δ-function are shown independently (orange box). (middle) Purity per bin (black markers) with result of spline fit (red) and purity of δ-function values (orange box). (bottom) Transformed final distribution.

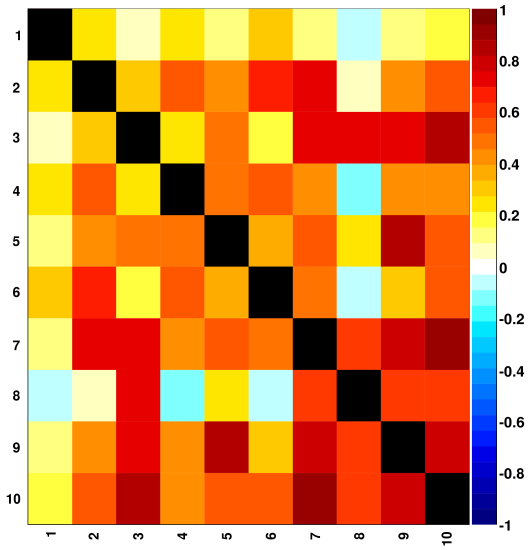

Figure 4.6: Color coded correlation matrix of input variables. The target is drawn as first variable and therefore column and row one show the correlation of each variable to the target. Color axis is going from 100% correlation (dark red) to 100% anti-correlation (dark blue).

4.5.3 Training Details

As described above, the learning process is complex and even with preprocessing, which gives good starting conditions, the search for the global or at least a good local minimum of the cost function is challenging. Among other techniques [29], two methods to improve the learning process shall be explained in more detail.

The random starting weights are chosen to be distributed as a Gaussian with mean 0 and width $1/\sqrt{n}$, where n is the number of incoming weights to a neuron. This way, if the input variables are distributed as a standard Gaussian, the output of the hidden layer follows a Gaussian distribution, which is then also valid for the output node. This gives optimal learning conditions from the beginning.

If a connection becomes insignificant ($< 0.001\sigma$) during learning it gets pruned away, therefore set to exactly zero. This changes the architecture of the network by decreasing the number of free parameters. At the same time the network's ability to generalize gets increased and overtraining is avoided.

4.5.4 Network Output

After training, the network output is evaluated and several plots for evaluating the network's performance and the quality of the training are added to the analysis file, which already contains the plots to evaluate the preprocessing.

Figure 4.7 shows a purity-efficiency plot for different cuts on the network output. The minimal purity at maximum efficiency is given by the a priori probability of the target in the training data set. In this example the signal to background ratio S/B of the training data set was about $4/6$. For the upper curve points the network output is bigger than the cut value, for the lower curve the output is smaller than the cut value.

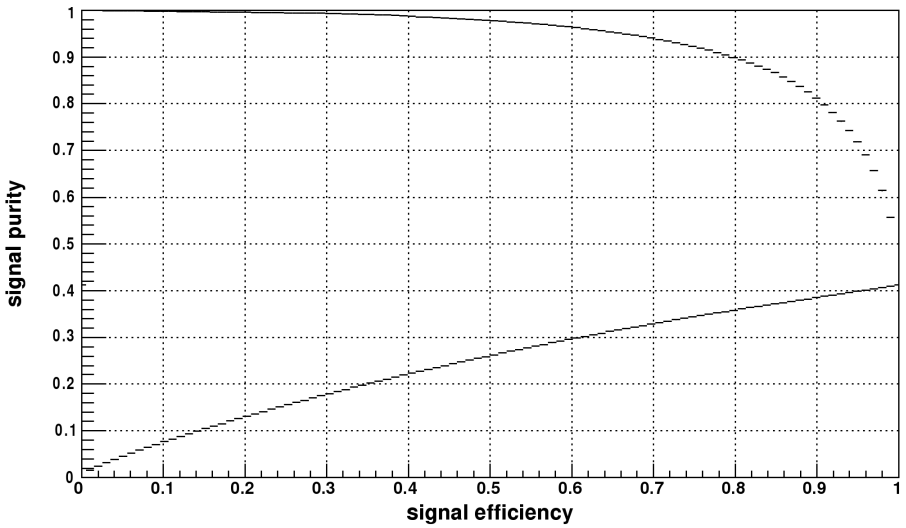

Figure 4.7: Purity-efficiency plot for different cuts on the network output.

Another plot to evaluate the network's power to separate is the Gini-plot, shown in figure 4.8. For each point of the blue curve, the network output is bigger than a certain cut value. The gray shaded region can not be reached and the diagonal would be a random decision. The upper left corner of the white triangle marks perfect separation and is given by the signal to background S/B ratio of the training data set, as in figure 4.7 it was about $4/6$.

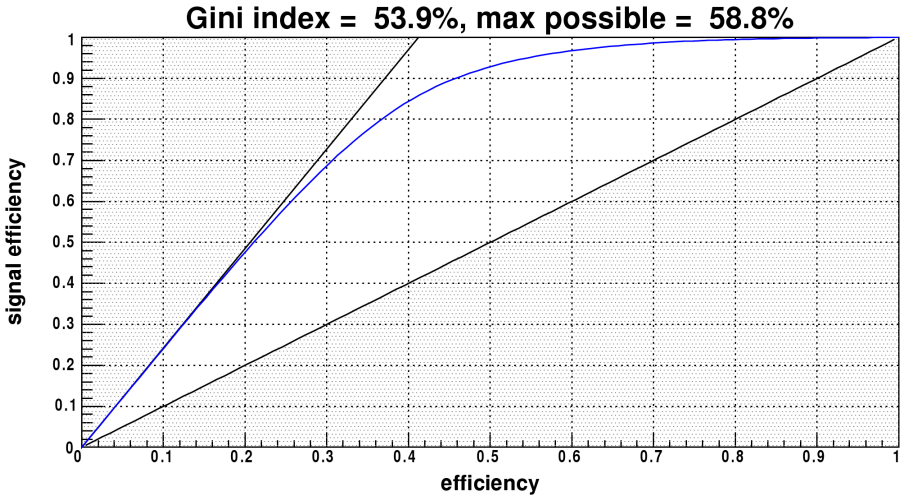

Figure 4.8: The Gini-plot gives an estimation of the network's separation power.

Not all problems allow good separation, therefore it is useful to look at the distribution of the network output itself. An example is shown in figure 4.9 and in case of good separation, background (black) should peek at -1 whereas signal (red) should peak at 1.

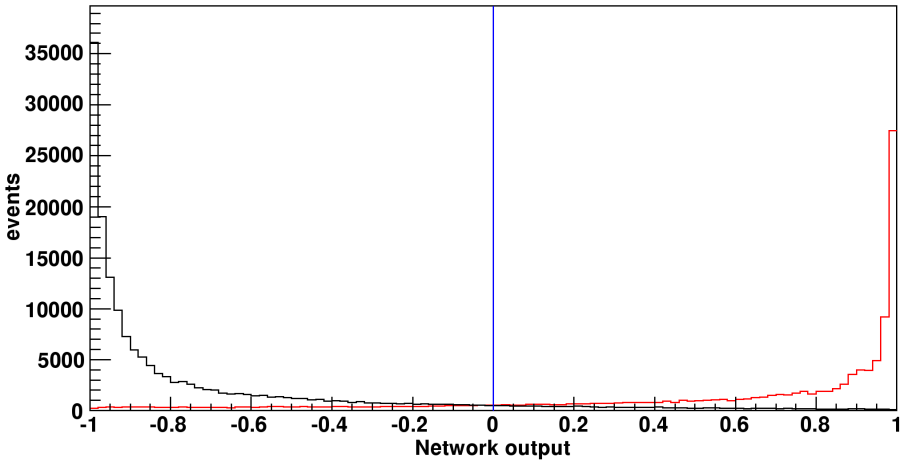

Figure 4.9: Distribution of the network output.

Assuming a well trained network, the purity as function of the network output (figure 4.10) should be on the diagonal, thus allowing one to interpret the output as probability for being signal when rescaled to the interval $[0, +1]$.

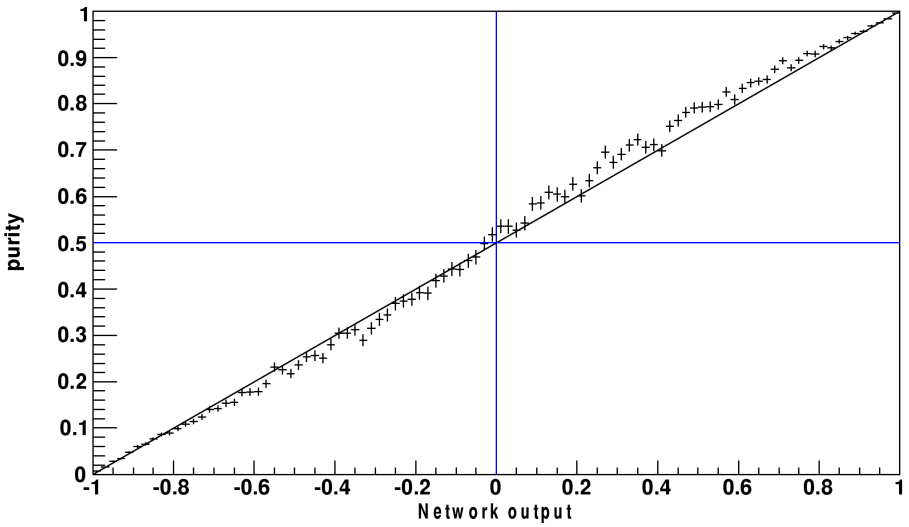

Figure 4.10: Purity as function of the network output in NeuroBayes® default output interval $[-1, +1]$.

5 Flavor Tagger

Due to their importance in time dependent CP violation measurements, the flavor tagging algorithms at the Belle experiment have been improved since the beginning of data taking. The neural network based flavor tagging algorithm is a novel and sophisticated method to determine the B-meson's flavor.

5.1 Multi-Dimensional Likelihood Flavor Tagger

The current default flavor tagging algorithm, used in many measurements, is a multi-dimensional likelihood (MDLH) method [32]. On a Monte Carlo data sample, tracks get assigned to certain categories and the probability density functions of measured quantities get binned. Afterwards the likelihood for having B^0 or $\overline{B}^0$ in each bin is calculated. Hence, one can create lookup tables, which allow one to determine the flavor in data.

The multi-dimensional likelihood method is a good approach to take correlations between different quantities into account. However, the smaller the binning gets, the less entries are in each bin and the statistical uncertainty increases. The output distribution of the multi-dimensional likelihood method also shows a peaking structure, due to the binning inside the algorithm, and not a continuous shape (see figure 5.6).

A replacement of the multi-dimensional likelihood tagger with a neural network (NN) based tagger can solve these problems by taking all correlations into account and by treating input variables as continuous variables and therefore removing the internal binning.

5.2 Neural Network based Flavor Tagger

The neural network based flavor tagger consists of 10 neural networks in total. The physical reason (see section 2.4) for choosing certain variables in the MDLH method still holds for the neural network method and therefore input variables are kept but combined with neural networks. Due to readability, detailed lists of input variables, results from each network training, including plots of correlations, network output distribution and purity as function of network output, have been transfered to appendix A.

In general, events are first evaluated by track level networks. Each track is assigned to a certain category and the network output is computed. The categories are slow pion, lambda, kaons with and without additional K_S^0 in the event, electrons and muons. In the next step, on event level, the slow pion tracks are combined in a slow pion network.

The lambda and kaon tracks are combined in a strangeness network. The electron and muon tracks are combined in a lepton network. Finally, all information is combined in a combined event level network. The output of this network is used directly for flavor tagging as the probability for having B^0 or $\overline{B}^0$. The layout of the flavor tagger is shown in figure 5.1.

The flavor tagging algorithm was integrated in the existing `hamlet` software library for flavor tagging at Belle. It doesn't replace existing tagging methods, but adds a new tagging method. A code listing of how to use the neural network based tagging method is given in appendix B.

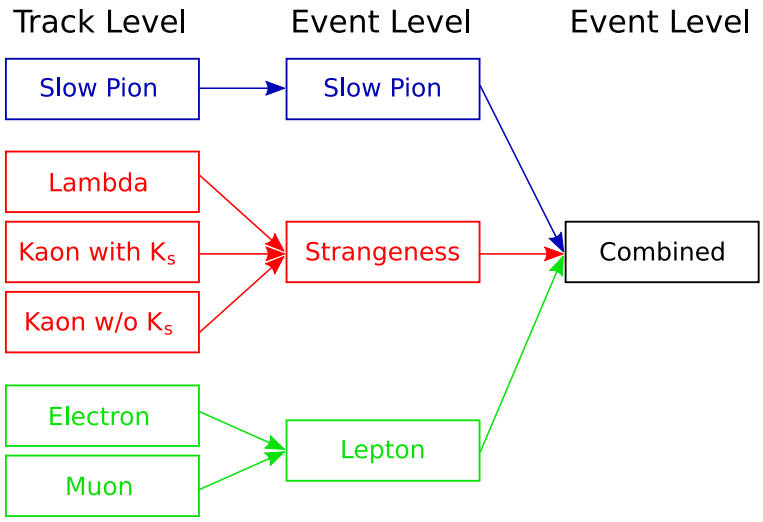

Figure 5.1: Layout of the neural network based flavor tagger.

5.2.1 Training Sample

The neural network based tagger was trained on about 1.5 million simulated $\Upsilon(4S) \rightarrow B^0\overline{B}^0$ events using version `b20090127_0910` of Belle Analysis Software Framework (BASF). This software version introduces a new tracking algorithm, which improves the track finding efficiency of the Belle software during event reconstruction. There are plans to reprocess all Belle data, which was taken with the SVD2 detector, with this software version by the end of 2009. There are no plans yet to reprocess the data taken with the SVD1 detector. Therefore it is reasonable to use events which already make use of the new tracking for training the flavor tagger's networks as the main part of all data taken by the Belle experiment will be reconstructed with this kind of tracking. The events used for training are given in table 5.1. Only MC data from experiments 61 and 63 are used as no other was available at the time the tagger was developed.

5.2.2 Track Level Networks

For the flavor tagging on track level, charged tracks, which don't belong to the CP side final state f_{CP}, are used. In addition they have to satisfy the impact parameters'

Experiment	Run-Start	Run-End	Event-Type	Stream	Events
61	1	150	evtgen-mixed	6	877596
63	1	75	evtgen-mixed	6	545244

Table 5.1: Simulated events used for training the neural network based flavor tagger.

requirements $|dr| < 2\,\text{cm}$ and $|dz| < 10\,\text{cm}$. K_S^0 and Λ candidates are selected from the corresponding candidate lists, which are obtained during event reconstruction by a secondary vertex reconstruction algorithm. Daughter tracks of K_S^0 and Λ candidates are removed from the charged track list.

Assignment to different track categories is decided by cuts on momentum and particle identification (PID) information. The particle identification likelihoods are created, using combined information of ACC, TOF, CDC, ECL and KLM (see section 3.3). A track is only assigned to the first category it fits in. The categories are checked in the same order as listed below. As written above, the criteria for these categories are the same as for the MDLH tagger and motivated by the physics of B-meson decays (see section 2.4).

Slow Pion

Slow pion tracks are selected by requiring a momentum of $p_{\text{cms}} < 0.25\,\text{GeV}/c$ and a ratio of kaon and pion likelihood $\mathcal{L}_K/(\mathcal{L}_K + \mathcal{L}_\pi) < 0.9$. The main background for this category is from low momentum pions from other than D^* decays. Due to its small momentum, the slow pion flight direction follows the D^* direction, therefore α_{thr}, the angle between the slow pion flight direction and the thrust axis calculated from the tag side particles, is used in the network. The thrust axis $\vec{t}$ of m particles is defined as the axis that maximizes

$$\vec{t} = \max_{\vec{n}} \frac{\sum_i^m \vec{n} \cdot \vec{p}_i}{\sum \vec{p}_i}, \tag{5.1}$$

where $\vec{p}_i$ is the momentum of particle i. A part of the low momentum electron background from photon conversion is rejected through secondary vertex reconstruction algorithms, applied during event reconstruction.

Electron

Electron tracks are selected by requiring a momentum of $p_{\text{lab}} > 0.4\,\text{GeV}/c$ and a ratio of electron and kaon likelihood $\mathcal{L}_e/(\mathcal{L}_e + \mathcal{L}_K) > 0.8$. The momentum in CMS frame p_{cms} of the electron tends to be higher for electrons from B decays than from charm decays. The hadronic recoil

$$M_{\text{recoil}} = |(\sum_i^n p_{\text{cms}}^i) - p_{\text{cms}}|, \tag{5.2}$$

where n is the number of all tracks and momenta are four vectors, should peak around the D-Meson mass. The missing momentum in the CMS frame

$$p_{\text{miss}} = |\sum_{i}^{n} \vec{p}^{\,i}_{\text{cms}}|, \tag{5.3}$$

where n is the number of all tracks, should be higher for primary electrons from B decays than for those from D decays.

Muon

Muon tracks are selected by requiring a momentum of $p_{\text{lab}} > 0.8\,\text{GeV}/c$ and a ratio of muon and kaon likelihood $\mathcal{L}_\mu/(\mathcal{L}_\mu + \mathcal{L}_K) > 0.95$. The same considerations for semileptonic decays with electrons are also true for semileptonic decays with muons. The tighter p_{lab} cut compared to electrons, is only due to the hard separation of low momentum muons, which don't reach the KLM, from pions.

Kaon with and without additional K_s^0 in the event

Kaon tracks are selected by requiring a ratio of proton and kaon likelihood $\mathcal{L}_p/(\mathcal{L}_p + \mathcal{L}_K) < 0.7$. However, this track category is split into two networks for events with and without an additional K_S^0 candidate. The reason is that kaons in events without K_S^0 tend to originate from the cascade decay $b \to c \to s$ whereas kaons in events with K_S^0 tend to originate from other (e.g. $s\bar{s}$ popping) processes.

Lambda

Lambda candidates are reconstructed from two oppositly charged tracks, one identified as proton. The candidates used in flavor tagging have to satisfy $1.1108\,\text{GeV}/c^2 < M_{p\pi} < 1.1208\,\text{GeV}/c^2$. The angle difference between Λ momentum and the vector from the interaction point to the Λ vertex is required to be $\theta_{\text{defl}} < 30°$ and the difference of Λ daughters at the Λ vertex to be $|\Delta z| < 4\,\text{cm}$. The distance of the secondary vertex from the interaction point in the $r - \phi$ plane is required to be above 0.5 cm.

5.2.3 Event Level Networks

On event level, the track level network outputs are combined. In each track category, the candidates are first ordered by their network output. The construction of likelihoods, which combine the network output of all candidates on track level as input for the corresponding event level network, is described in detail in appendix A.12. In general those likelihoods are constructed as

$$\mathcal{L} = \frac{\mathcal{L}_{B^0}}{\mathcal{L}_{B^0} + \mathcal{L}_{\overline{B}^0}}, \tag{5.4}$$

where $\mathcal{L}_{B^0}$ and $\mathcal{L}_{\overline{B}0}$ are given by

$$\mathcal{L}_{B^0} = \prod_{i=1}^{n} \mathcal{L}_{B^0,i} \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0} = \prod_{i=1}^{n} \mathcal{L}_{\overline{B}0,i}. \tag{5.5}$$

The likelihoods $\mathcal{L}_{B^0,i}$ and $\mathcal{L}_{\overline{B}0,i}$ for each track i are defined as

$$\mathcal{L}_{B^0,i} = 1 + \mathrm{NN}(i) \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,i} = 1 - \mathrm{NN}(i), \tag{5.6}$$

where $NN(i)$ is the network output of the corresponding track level network.

In contrast to the MDLH method, all input variables used on track level are reused and added to the event level network again. In general, the two best candidates, those with the highest network output, are used in each category. This is the reason for the block structure, which can be seen in all the event level network correlation plots shown in the appendix (see figure A.19 for instance).

The event level networks can also handle events that only have a single or even no candidate in a certain category. Therefore no combination of previous outputs can be made and a lot or even all of the track level input variables are missing. However, NeuroBayes® can deal well with missing values.

If one compares the plots of network output and purity as a function of the network output for the event level networks with those on track level, one can see that the event level networks are already able to separate B^0 from $\overline{B}^0$ and that those networks are better calibrated than those on track level. For instance, the network output for the kaon without K_S^0 network, compared to the strangeness event level network is shown in figure 5.2. The purity as function of network output for those two networks is shown in figure 5.3.

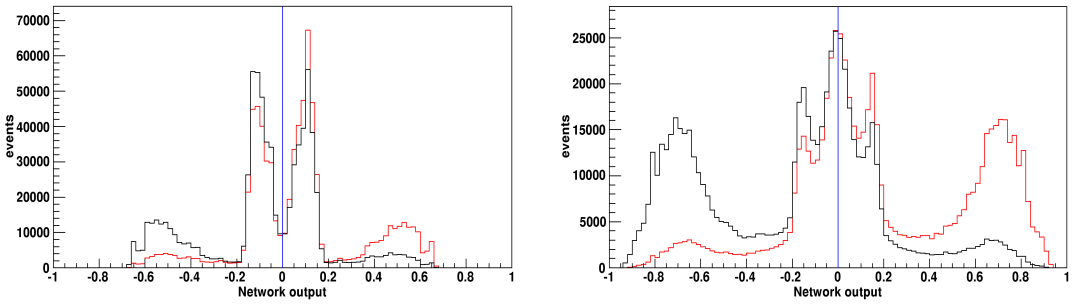

Figure 5.2: Network output of kaon without K_S^0 network (left) and strangeness event level network (right).

Slow Pion

The slow pion event level network combines the information of the slow pion candidates. The slow pion likelihood $\mathcal{L}_{pion}$ combines the network outputs of all slow pion candidates. In addition, the two best slow pion candidates are added with all their input variables and track level network output.

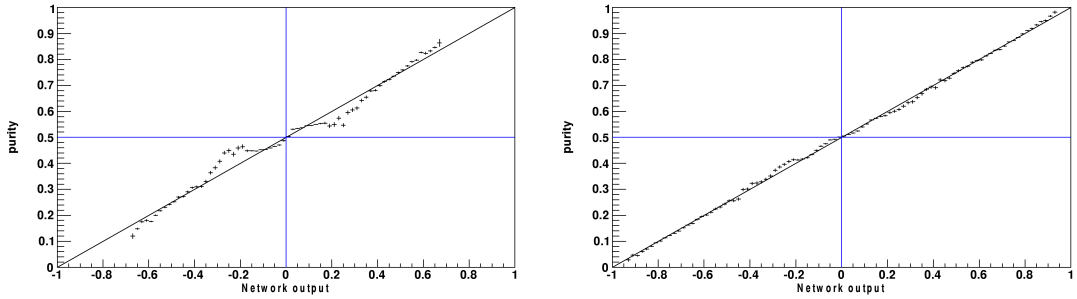

Figure 5.3: Purity as function of network output of kaon without K_S^0 network (left) and strangeness event level network (right).

Strangeness

The strangeness event level network combines the information of kaon and lambda candidates. The kaon likelihood $\mathcal{L}_{kaon}$ combines all kaon network outputs and the $\mathcal{L}_\Lambda$ likelihood combines all lambda network outputs. The strange likelihood $\mathcal{L}_{strange}$ combines $\mathcal{L}_{kaon}$ and $\mathcal{L}_\Lambda$. In addition, the two best kaons candidates and the best lambda candidate are added with all their input variables and track level network output.

Lepton

The lepton event level network combines the information of the electron and muon candidates. The lepton likelihood $\mathcal{L}_{lepton}$ combines all electron and muon network outputs. In addition, the two best electron and the two best muon candidates are added with all their input variables and track level network output.

5.2.4 Combined Event Level Network

The combined event level network combines the network outputs on track level as well as those on event level. In addition, the event level network outputs are combined by an event likelihood $\mathcal{L}_{event}$. As for the three event level networks, all input variables from track level, apart from Λ network inputs, are reused as input variables in the combined event level network. Thus resulting in a total amount of 71 input variables that show a clear block structure in the correlation plot.

The combined event level network is the only network, whose output is important for the user, as it should provide the probability for having B^0 or $\overline{B}^0$ and is used in the time dependent CP violation measurement. In figure 5.4 the network output and the purity as a function of the network output is shown. It can be seen, that over the whole network output range the purity is on the diagonal, i.e. the network is well calibrated and therefore allows one to interpret the network output as probability for having B^0 or $\overline{B}^0$.

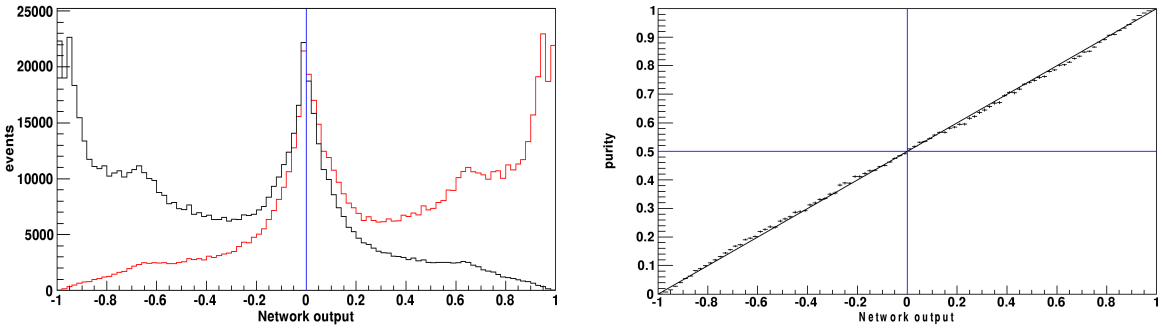

Figure 5.4: Network output (left) and purity as function of network output (right) of the combined event level network.

5.3 Interpretation of Tagger Output

As any new method has to compete with the existing methods, a validation of the neural network based flavor tagger and comparison with the multi-dimensional likelihood flavor tagger is necessary. One way of validation is the use of simulated $\Upsilon(4S) \rightarrow B^0\overline{B}^0$ events. One of the two B mesons can be selected as the signal side B and the tracks of the other B are used for flavor tagging. In that way no systematics due to wrongly reconstructed signal side B mesons or non-$B^0\overline{B}^0$ events are introduced. This method is mainly limited by the size of the sample used for validation. The use of about 4 million simulated events for SVD2 new tracking and about 3 million for SVD1 old tracking, guarantees a small statistical uncertainty on the obtained results.

As mentioned previously, the main part of data taken by Belle will be SVD2 data using the new tracking with a smaller part of the data using SVD1 data and old tracking. Therefore this validation will focus on these two types of data as SVD2 data with old tracking will soon no longer be used.

As described in section 2.6, the effective efficiency ϵ_{eff} is the main quantity to characterize a flavor tagger's performance. As the wrong tag fraction w depends on the tagger output itself, ϵ_{eff} is measured in bins of the tagger output $q \cdot r$, where q is the sign of the tagger output and returns the flavor, whereas r is the numeric value. In case of a well calibrated tagger, a higher value in r should have a smaller wrong tag fraction w, thus $r = 1 - 2w$.

The flavor tagger returns values in the interval $[-1, 1]$ and is divided into 13 bins. The bins are illustrated in figure 5.5. The central bin $(-0.1, 0.1)$ is defined as having a wrong tag fraction of $w = 0.5$ and therefore this bin does not have any flavor information, thus $q = 0$. This bin is also not used in Belle measurements. Positive output $q > 0$ is defined to be B^0 and negative output $q < 0$ is defined to be $\overline{B}^0$.

It was studied, whether the use of more than 13 bins can improve the effective efficiency as binning always implies to average over all events in one bin. However, no significant improvement was observed.

In first order, the wrong tag fractions w for B^0 and $\overline{B}^0$ are the same but due to detector effects in the detection of matter and anti-matter, it is more precise to introduce

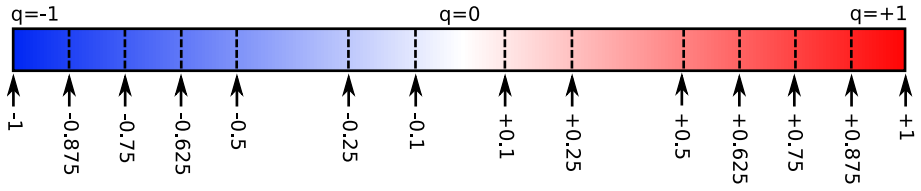

Figure 5.5: Binning of the tagger output $q \cdot r$.

flavor specific wrong tag fractions w_{B^0} and $w_{\bar{B}^0}$ per bin. Neglecting the central bin, one obtains 12 independent parameters, one for each remaining bin. One can introduce the average wrong tag fraction between two corresponding bins, e.g. $[-1, -0.875]$ and $[0.875, 1]$,

$$w_{ave} = \frac{w_{B^0} + w_{\bar{B}^0}}{2}, \tag{5.7}$$

which is equivalent to what was called w before. The difference between both flavor specific wrong tag fractions in corresponding bins is given by

$$\Delta w = w_{B^0} - w_{\bar{B}^0}. \tag{5.8}$$

One can group those 12 flavor specific wrong tag fractions to six first order w^l_{ave} and six second order Δw^l parameters, one of each order per bin l, where $l = 1 \ldots 6$.

The effect of the flavor specific wrong tag fraction is small, but measureable, and the equations 2.16 and 2.17 become

$$\mathcal{P}^{rec}_{SF} = (1 - w_{B^0})\mathcal{P}_{OF} + w_{\bar{B}^0}\mathcal{P}_{SF} \;\propto\; 1 - \Delta w + (1 - 2w_{ave}) \cos(\Delta m_d \Delta t), \tag{5.9}$$

$$\mathcal{P}^{rec}_{OF} = w_{B^0}\mathcal{P}_{OF} + (1 - w_{\bar{B}^0})\mathcal{P}_{SF} \;\propto\; 1 + \Delta w - (1 - 2w_{ave}) \cos(\Delta m_d \Delta t). \tag{5.10}$$

Therefore the values for w^l_{ave} and Δw^l obtained on Monte Carlo simulated data will be given in the next section as they are needed as input for time dependent CP violation measurements on simulated data. However, these values should not be used directly for real data. As described in section 2.5, the wrong tag fraction values can be obtained from real data by a mixing fit, thus being independent of any systematic effects introduced by Monte Carlo simulation.

When comparing two tagging methods, one can neglect those second order effects and simply use the values of w^l_{ave}. The effective tagging efficiency is then given by

$$\epsilon_{eff} = \epsilon_{tag} \sum_{l=1}^{6} \epsilon_l (1 - 2w^l_{ave})^2, \tag{5.11}$$

where ϵ_{tag} is the tagging efficiency and ϵ_l the event fraction per bin l. The tagging algorithms have been made so robust that even in case of e.g. unexpected NaN (not a number) values they return output. During validation on Monte Carlo simulated data, no event without output returned from the tagger was observed and therefore $\epsilon_{tag} \simeq 1$ can be neglected.

5.4 Validation on Monte Carlo

5.4.1 Validation on SVD2 New Tracking

A general achievement of the neural network based tagger is the smoother output distribution, compared to the MDLH output distribution. In figure 5.6 the true flavor times tagger output $q \cdot r$ is shown for the neural network based tagger as well as the MDLH tagger. The neural network based tagger shows a smooth distribution and no peaking structure.

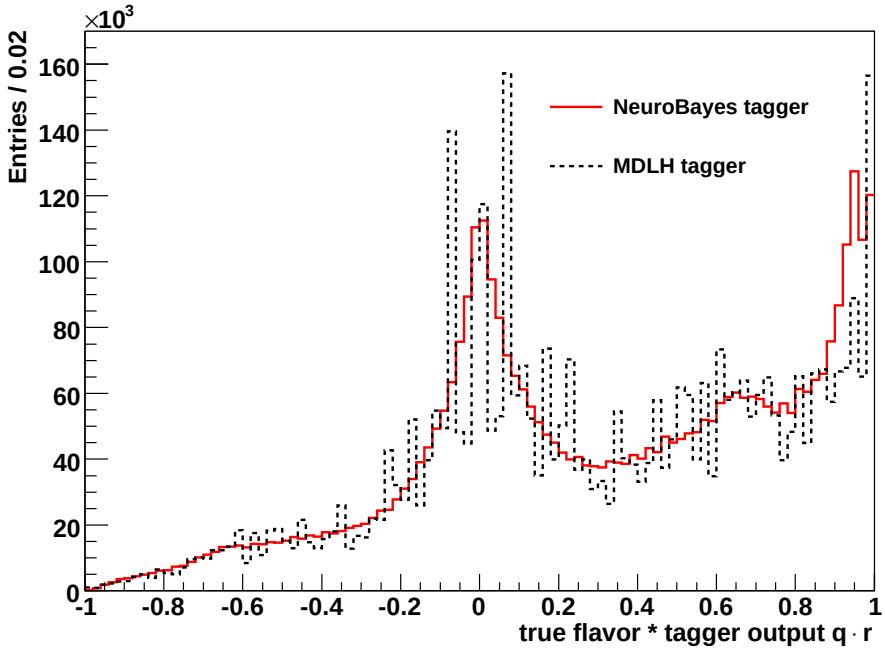

Figure 5.6: Distribution of true flavor times tagger output $q \cdot r$ on SVD2 new tracking.

On 4 million $B^0\overline{B}^0$ the flavor specific wrong tag fraction was obtained for B^0 and $\overline{B}^0$ in each bin of $q \cdot r$. From those values, the average wrong tag fraction w^l_{ave} was calculated as well as the difference between flavor specific wrong tag fraction Δw^l. The event fraction ϵ_l was also extracted. Figure 5.7 shows the flavor specific wrong tag fraction as function of the tagger output $|q \cdot r|$ for the neural network based tagger as well as the MDLH tagger. For an ideally calibrated tagger, the wrong tag fraction values would be on the blue diagonal.

The extracted values result in an effective efficiency for the neural network based tagger of

$$\epsilon^{NN}_{eff} = 32.51\% \pm 0.04\% \quad \text{and} \quad \epsilon^{MDLH}_{eff} = 31.64\% \pm 0.04\% \tag{5.12}$$

for the MDLH tagger. The errors given are statistical errors on Monte Carlo only and details on how they are calculated are given in appendix C. The relative improvement of the neural network based tagger over the MDLH tagger is 2.7%.

The detailed values that have been extracted for the neural network based tagger are shown in table 5.2.

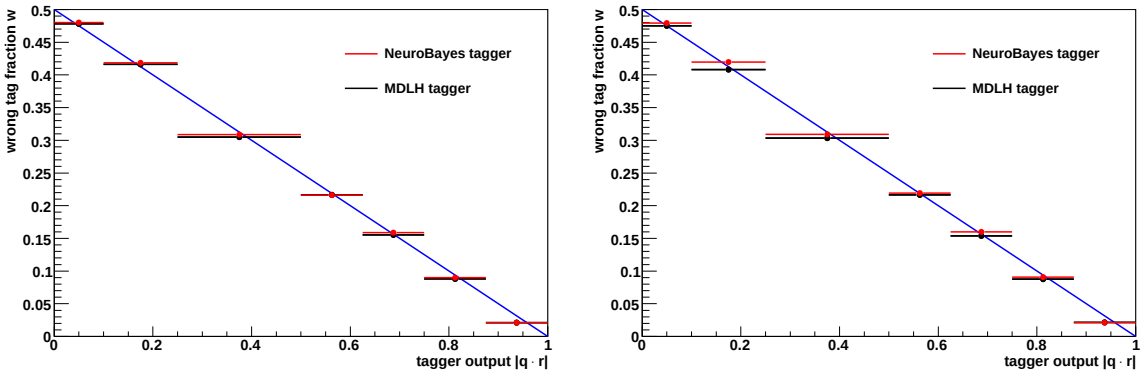

Figure 5.7: Flavor specific wrong tag fraction for B^0 flavor (left) and $\overline{B}^0$ flavor (right) of the neural network based tagger (red) and MDLH tagger (black) as function of tagger output $|q \cdot r|$. Ideal calibration is indicated by the diagonal line (blue). Results obtained on Monte Carlo for SVD2 data period with new tracking.

l	r **interval**	ϵ_l	w^l_{ave}	Δw^l	ϵ^l_{eff}
0	[0.000,0.100)	0.2010 ± 0.0002	0.5	0	0
1	[0.100,0.250)	0.1532 ± 0.0002	0.4190 ± 0.0006	-0.0011 ± 0.0012	0.0040 ± 0.0001
2	[0.250,0.500)	0.1803 ± 0.0002	0.3089 ± 0.0005	-0.0002 ± 0.0011	0.0263 ± 0.0002
3	[0.500,0.625)	0.0990 ± 0.0002	0.2179 ± 0.0006	-0.0030 ± 0.0013	0.0315 ± 0.0002
4	[0.625,0.750)	0.1067 ± 0.0002	0.1593 ± 0.0006	-0.0009 ± 0.0011	0.0495 ± 0.0002
5	[0.750,0.875)	0.1000 ± 0.0002	0.0903 ± 0.0004	-0.0008 ± 0.0009	0.0671 ± 0.0002
6	[0.875,1.000]	0.1598 ± 0.0002	0.0211 ± 0.0002	0.0002 ± 0.0004	0.1466 ± 0.0002

Table 5.2: Wrong tag fractions of neural network based flavor tagger. Results obtained on Monte Carlo for SVD2 data period with new tracking.

5.4.2 Validation on SVD1 Old Tracking

The same validation process as in the last section was applied to 3 million MC events of the SVD1 data period, which uses old tracking. As shown in figure 5.8, the neural network based tagger output shows the same smooth behavior compared to the peaking MDLH tagger output.

The wrong tag fraction was obtained as in the last section and is shown in figure 5.9. The detailed values that have been extracted for the neural network based tagger are given in table 5.3.

The values lead to effective efficiencies of

$$\epsilon^{NN}_{eff} = 30.19\% \pm 0.04\% \quad \text{and} \quad \epsilon^{MDLH}_{eff} = 29.60\% \pm 0.04\%. \tag{5.13}$$

The errors given are statistical errors on Monte Carlo only and explained in appendix C. The relative improvement of the neural network based tagger over the MDLH tagger is 2.0%.

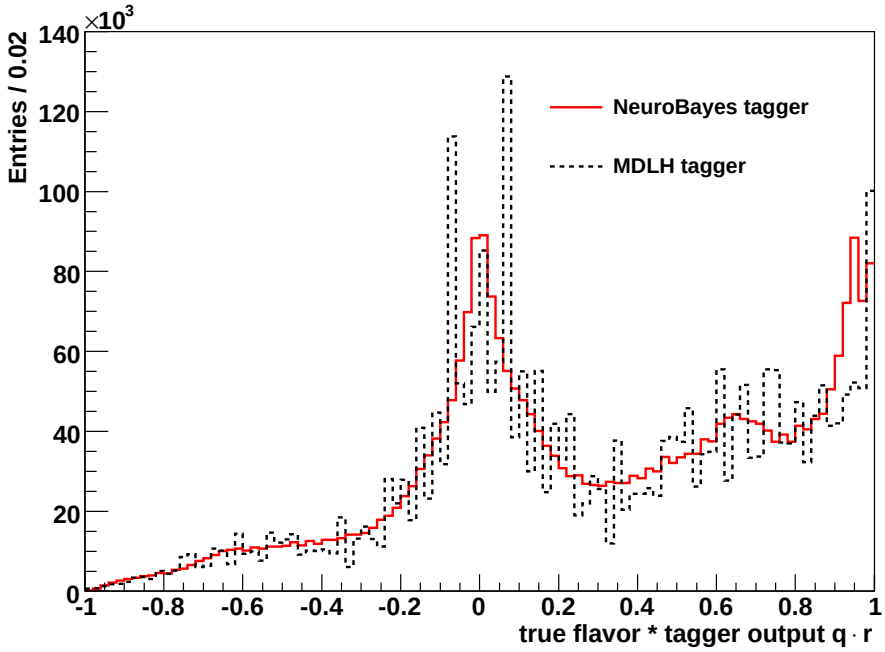

Figure 5.8: Distribution of true flavor times tagger output $q \cdot r$ on SVD1 old tracking.

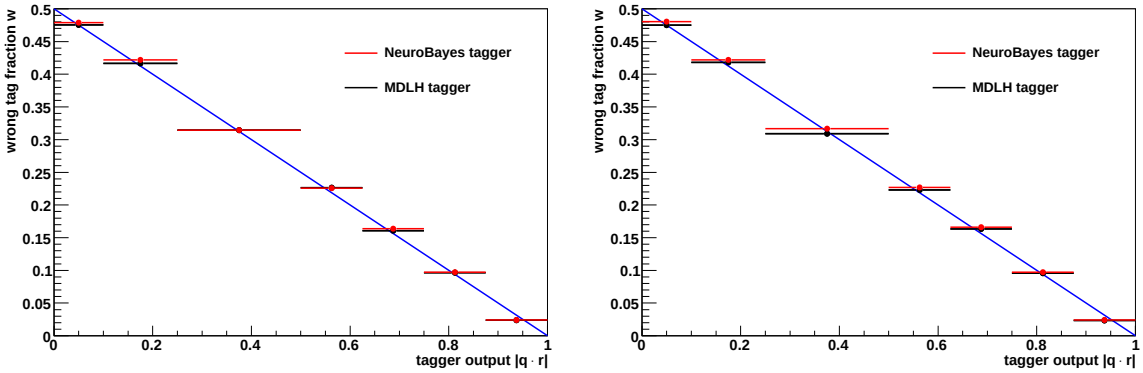

Figure 5.9: Flavor specific wrong tag fraction for B^0 flavor (left) and $\overline{B}^0$ flavor (right) of the neural network based tagger (red) and MDLH tagger (black) as function of tagger output $|q \cdot r|$. Ideal calibration is indicated by the diagonal line (blue). Results obtained on Monte Carlo for SVD1 data period with old tracking.

l	r **interval**	ϵ_l	w^l_{ave}	Δw^l	ϵ^l_{eff}
0	[0.000,0.100)	0.2139 ± 0.0003	0.5	0	0
1	[0.100,0.250)	0.1607 ± 0.0002	0.4219 ± 0.0007	-0.0002 ± 0.0014	0.0039 ± 0.0001
2	[0.250,0.500)	0.1758 ± 0.0002	0.3157 ± 0.0006	-0.0021 ± 0.0013	0.0239 ± 0.0002
3	[0.500,0.625)	0.0995 ± 0.0002	0.2264 ± 0.0007	-0.0013 ± 0.0015	0.0298 ± 0.0002
4	[0.625,0.750)	0.1064 ± 0.0002	0.1649 ± 0.0006	-0.0026 ± 0.0013	0.0478 ± 0.0002
5	[0.750,0.875)	0.0938 ± 0.0002	0.0972 ± 0.0005	0.0001 ± 0.0011	0.0608 ± 0.0002
6	[0.875,1.000]	0.1498 ± 0.0002	0.0242 ± 0.0002	0.0002 ± 0.0005	0.1357 ± 0.0003

Table 5.3: Wrong tag fractions of neural network based flavor tagger. Results obtained on Monte Carlo for SVD1 data period with old tracking.

5.4.3 Cross Validation Test

For testing purposes, another tagger was trained on SVD2 data with old tracking as well as one that was trained on SVD1 data with old tracking. Those taggers, as well as the official neural network based tagger trained on SVD2 new tracking, were applied to all other kinds of data, thus leading to nine different (x, y) combinations of tagger trained on x and applied to y, where x and y can be SVD2_{new}, SVD2_{old} and SVD1_{old}. However, no significant deviation has been found during these tests, that would indicate that a single tagger for all kinds of Belle data would introduce a bias depending on its training sample or to be not robust enough when applied to other than its training sample.

6 Validation on Data

As a Monte Carlo simulation can not simulate nature perfectly, its use may introduce a bias or systematic effects if results obtained on simulation are directly applied on data. When possible, it is usually a good idea to try to become independent of simulation. Therefore, a validation of flavor tagging algorithms on data is necessary.

6.1 Outline of Validation Procedure

As described in section 2.5, one can extract the wrong tag fraction of a flavor tagging algorithm from data by doing a $B^0\overline{B}^0$ mixing fit. The decay $B^0 \to D^*(2010)^-\ell^+\nu$ and its charged conjugate[1] are decays to a flavor eigenstate, i.e. the flavor of the B meson is given by the charge of the decay products. There are other decays to flavor eigenstates, such as $B^0 \to D^-\ell^+\nu$, $B^0 \to D^*(2010)^-\pi^+$ or $B^0 \to D^-\pi^+$, but this specific B^0 decay was chosen, as its branching ratio $BR = (5.16 \pm 0.11)\%$ [6] is large compared to other decays to flavor eigenstates and the $D^*(2010)^-$ provides a relatively clear signal, compared to modes without $D^*(2010)^-$.

Therefore, the first step for validation on data is obtaining a sample of $B^0 \to D^*(2010)^-\ell^+\nu$, thus knowing the flavor of the corresponding B^0. Then the flavor tagging algorithm is applied to the other side B and its result is stored together with vertex information of both B-mesons. These are, in general, the necessary inputs for the mixing fit, which is not covered in this thesis, to obtain the wrong tag fraction.

Througout this chapter, simulated data and off-resonance data of experiments 61 to 65 was used, thus entirely using the new tracking.

6.2 Simulated Signal Events

About 1 million simulated signal events have been produced, using the event generator EvtGen [33], which was designed to simulate the physics of B decays. A full detector simulation was done, using GEANT3 [34]. The EvtGen configuration file, which was used, is given in appendix D and for GEANT3 the default Belle configuration was applied. The signal Monte Carlo was produced using version `b20090127_0910` of the Belle Analysis Software Framework (BASF) and thus the new tracking, explained in section 5.2.1, was applied.

The initial process $\Upsilon(4S) \to B^0\overline{B}^0$ is simulated and mixing between both B-mesons is taken into account. One B decays via the signal decay mode $B^0 \to D^*(2010)^-\ell^+\nu$,

[1] Charged conjugate modes of any physics process are included, unless explicitly specified, throughout this chapter.

where ℓ^+ can be either e^+ or μ^+ and $D^*(2010)^- \to \overline{D}^0\pi^-$. The ratio between e^+ and μ^+ was chosen 1/1. For the $\overline{D}^0$ decay, the three modes $\overline{D}^0 \to K^+\pi^-$, $\overline{D}^0 \to K^+\pi^-\pi^0$ and $\overline{D}^0 \to K^+\pi^-\pi^+\pi^-$ have been simulated. The ratios between these modes have been chosen as measured in data [6]. The other B decays generically, using the default Belle decay table, which is wherever possible based on the results compiled by the Particle Data Group [6].

A BASF module has been written to count the exact number of events in each of the decay channels. The resulting number of simulated events per channel are given in table 6.1. From the measured branching ratios and the number of $B\overline{B}$ pairs in data one can estimate the amount of signal events in data (see section 6.5 for more details). The ratio between expected signal events and simulated signal events can be used to scale the amount of simulated signal events.

Channel	Simulated Signal Events	Expected Signal Events
$\overline{D}^0 \to K^+\pi^-$	174022	312156 ± 3456
$\overline{D}^0 \to K^+\pi^-\pi^0$	615600	1115416 ± 6299
$\overline{D}^0 \to K^+\pi^-\pi^+\pi^-$	359928	649991 ± 3887
Combined	1149550	2077563 ± 8168

Table 6.1: Number of simulated signal events and expected signal events in data per $\overline{D}^0$ decay channel and total amount.

6.2.1 Final State Radiation

In all decays with charged particles, the `PHOTOS` module [35] has been applied to simulate the final state radiation (FSR). Technically, the module adds additional γ daughters to the mother particle of the charged track that emits FSR. Therefore, FSR can come from B^0, $D^*(2010)^-$ or $\overline{D}^0$. However, due to a low energy $E_\gamma \geq 10\,\text{MeV}$ cut-off during simulation, the $D^*(2010)^-$ has no FSR as its charged daughter, the slow pion, has too low momentum.

For B^0 and D^0, the energy distribution of FSR photons is shown in figure 6.1. The distribution is shown for events where the lepton is either electron or muon. As expected there is no significant difference between the amount of FSR from D^0. The amount of FSR from B^0 differs, as the cross section for FSR has a dependence on the ratio between mass of the mother and mass of the daughter. As the electron mass is much smaller than the muon mass, more FSR is expected from electrons than from muons.

FSR photons are not explicitly reconstructed and candidates with FSR are matched as truly reconstructed signal if only FSR photons are missing. Events with FSR are signal and not counting them as signal would introduce a large amount of background, about 9.2% of all events have FSR, that behaves like signal. The missing mass due to the neutrino can be slightly increased due to FSR. However, this effect is rather small and no individual treatment of e^+ and μ^+ decay modes was necessary.

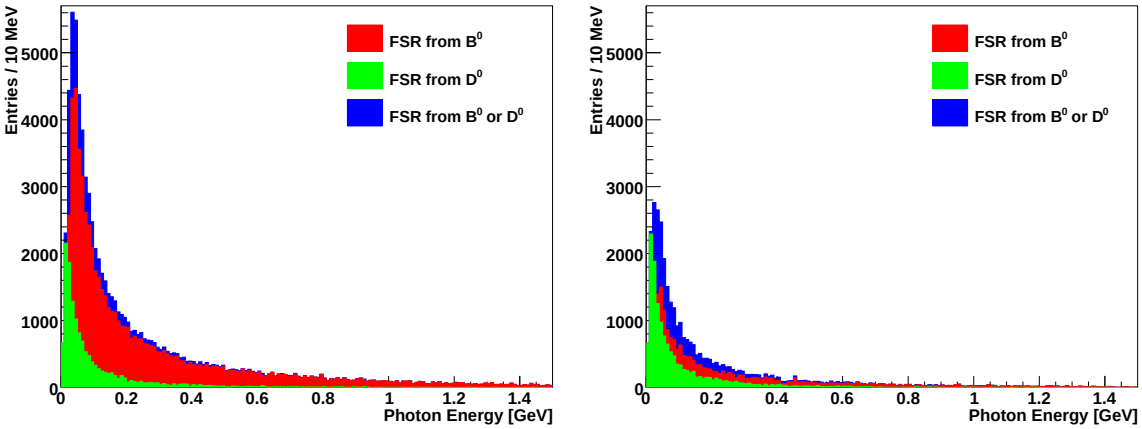

Figure 6.1: Energy of FSR photons in $B^0 \to D^*(2010)^- e^+ \nu$ (left) and $B^0 \to D^*(2010)^- \mu^+ \nu$ (right). FSR from D^0 is similar in both channels, whereas FSR from B^0 differs as expected.

6.2.2 Resonant Substructure in $\overline{D}^0 \to K^+\pi^-\pi^0$ Decay

The module `D_Daltiz` has been applied to the decay $\overline{D}^0 \to K^+\pi^-\pi^0$. The resonant substructure of this decay was measured [36] and the results can be used in the `D_Dalitz` module to create a better simulation than a simple three body phase space decay. The substructure arises from the fact that amplitudes $\mathcal{A}_i$ of different resonant decay modes i interfere. The three most important resonant decay modes are:

- $\overline{D}^0 \to K^+\rho^-$ with $\rho^- \to \pi^-\pi^0$
- $\overline{D}^0 \to K^*(892)^0\pi^0$ with $K^*(892)^0 \to K^+\pi^-$
- $\overline{D}^0 \to K^*(892)^+\pi^-$ with $K^*(892)^+ \to K^+\pi^0$

The squared amplitude $|\mathcal{A}|^2$ of the decay $\overline{D}^0 \to K^+\pi^-\pi^0$ is given by

$$|\mathcal{A}|^2 = |\sum_i \mathcal{A}_i|^2 \neq \sum_i |\mathcal{A}_i|^2. \tag{6.1}$$

The substructure has to be measured and interferences can result in very sharp resonances. For other 3-body decays it might be possible, that the $\neq$ in equation 6.1 can be replaced with a $\simeq$ but there is no general rule. For illustration purposes, the resonant substructure of $\overline{D}^0 \to K^+\pi^-\pi^0$ generated by EvtGen is shown in figure 6.2. The distribution shows a structure that clearly differs from an uniformly distributed phase space.

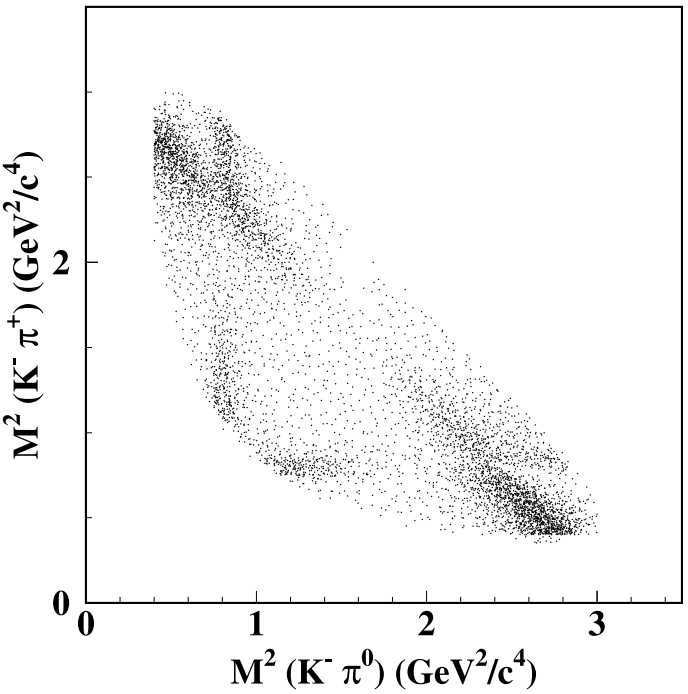

Figure 6.2: Dalitz plot for $D^0 \to K^- \pi^+ \pi^0$ in EvtGen.

6.3 Reconstruction of $B^0 \to D^*(2010)^- \ell^+ \nu$

6.3.1 Charged Track Selection

The tracks from the `Mdst_charged` table are used and they are required to have at least one hit in the $r - \phi$ plane of the SVD and to pass the impact parameter cuts $|dr| < 2\,\text{cm}$ and $|dz| < 4\,\text{cm}$.

Tracks that pass the impact parameter requirements are assigned to different categories. A track can be assigned to more than one category. If a track is assigned it is called a candidate of the corresponding category. However, it is taken into account that e.g. a kaon and a pion candidate, which are based on the same track, can never be combined directly or indirectly. Tracks are assigned to the electron category if the electron likelihood is $\mathcal{L}_e > 0.5$ and they are assigned to the muon category if the muon likelihood is $\mathcal{L}_\mu > 0.8$. The event is skipped if neither electron nor muon candidates are found. The ratio between kaon and pion likelihood is required $\mathcal{L}_K/(\mathcal{L}_K + \mathcal{L}_\pi) < 0.9$ for tracks in the pion category and $\mathcal{L}_K/(\mathcal{L}_K + \mathcal{L}_\pi) > 0.1$ for the kaon category. A subsample of the pion category is the slow pion category, which additionally requires the candidates to have $p_{cms} < 0.25\,\text{GeV}/c$.

6.3.2 π^0 Selection

π^0 candidates are taken from the `Mdst_pi0` table. The invariant mass of the candidates has to be within $0.011\,\text{MeV}/c^2$ of the nominal π^0 mass and the momentum is required

$p > 0.2\,\mathrm{GeV}/c$. Each child photon of the π^0 is required to have energy $E > 80\,\mathrm{MeV}$.

6.3.3 $\overline{\mathrm{D}}^0$ Reconstruction

$\overline{D}^0$ candidates are reconstructed by combining $K^+\pi^-$, $K^+\pi^-\pi^0$ and $K^+\pi^-\pi^+\pi^-$. For each $\overline{D}^0$ candidate a kinematic fit of all tracks, including the virtual track of the neutral π^0 in the $K^+\pi^-\pi^0$ mode, to a common vertex is done.

6.3.4 $\mathrm{D}^*(2010)^-$ Reconstruction

$D^*(2010)^-$ candidates are reconstructed by combining $\overline{D}^0$ candidates with negative charged slow pion π_s^- candidates. Each slow pion candidate is fitted to its production vertex, which is the B^0 vertex and obtained before (see below), to improve the invariant mass difference $\Delta m = m_{\bar{D}^0\pi_s^-} - m_{\bar{D}^0}$ resolution.

6.3.5 B^0 Reconstruction

Before a B^0 candidate is reconstructed by combining a $D^*(2010)^-$ candidate with an electron or muon candidate, the lepton is fitted with the $\overline{D}^0$ and the IP Tube constraint [37] to a common vertex. This vertex is assumed to be the B^0 vertex and the slow pion candidate is refitted to this vertex. Afterwards, the slow pion candidate is used to reconstruct the $D^*(2010)^-$ candidate.

6.3.6 Reconstruction Efficiency

The reconstruction efficiency ϵ_{rec} is given by the fraction of true reconstructed candidates over the amount of generated signal events. No selection criteria other than the charged track and π^0 selection explained above are applied. The efficiency is determined per channel and the results are shown in table 6.2. The combined reconstruction efficiency is the weighted sum of all three channels. The errors are Monte Carlo statistical errors only and their calculation is explained in appendix C.

Channel	**True Candidates**	**Generated Events**	$\boldsymbol{\epsilon_{rec}}$ [%]
$\overline{D}^0 \to K^+\pi^-$	48178	174022	27.69 ± 0.11
$\overline{D}^0 \to K^+\pi^-\pi^0$	50696	615600	8.24 ± 0.04
$\overline{D}^0 \to K^+\pi^-\pi^+\pi^-$	55814	359928	15.51 ± 0.06
Combined	154688	1149550	13.43 ± 0.03

Table 6.2: Reconstruction efficiency per channel.

6.4 B^0 Selection

For selecting the best B^0 from the reconstructed candidates a neural network was trained. The signal Monte Carlo sample was split to two equal and independent samples. Preselection cut optimization and neural network training have been done on one sample. All selection efficiency evaluation has been done on the other sample. For generic Monte Carlo background, two independent streams have been used for training and evaluation.

6.4.1 Derived Variables

For the best B^0 selection some derived variables are calculated and explained below.

Dalitz Weight

In the $K^+\pi^-\pi^0$ mode the squared amplitude $|\mathcal{A}|^2$ of the $\overline{D}^0$ candidate is calculated from $m_{\bar{D}^0}$, $m^2_{K^+\pi^-}$ and $m^2_{K^+\pi^0}$. To depict this calculation, one can say that the position in the Dalitz plot (see figure 6.2) is evaluated to obtain a weight for each $\overline{D}^0$ candidate, using the resonant substructure of the $\overline{D}^0$ decay. The $|\mathcal{A}|^2$ will be called Dalitz weight throughout the rest of this thesis.

Missing Mass

A difficult task is to select proper $D^*(2010)^-\ell^+$ combinations from the $D^*(2010)^-\ell^+\nu$ decay, as the neutrino does not show up in the detector. However, one can use the massless characteristics of the neutrino. In general, the mass of the neutrino m_ν is given by

$$m_\nu^2 = (E_B - E_{D^*l})^2 - |p_B|^2 - |p_{D^*l}|^2 + 2|p_B||p_{D^*l}|\cos\theta_{B,D^*l}, \tag{6.2}$$

where E_B and E_{D^*l} are the energy and $|p_B|$ and $|p_{D^*l}|$ are the CMS momentum of the B^0 and $D^*(2010)^-\ell^+$ system, respectively. $\cos\theta_{B,D^*l}$ is the cosine of the angle between the CMS momentum direction of the B^0 and $D^*(2010)^-\ell^+$ system. From the approximately massless characteristics $m_\nu \simeq 0$ of the neutrino, equation 6.2 becomes

$$0 = MM^2 + C\cos\theta_{B,D^*l}, \tag{6.3}$$

where $MM^2 = m_B^2 + m_{D^*l}^2 - 2E_BE_{D^*l}$ is the missing mass squared and $C = 2|p_B||p_{D^*l}|$. As the codomain of the cosine is $[-1, 1]$ one can require that

$$|\cos\theta_{B,D^*l}| = |-\frac{MM^2}{C}| \leq 1.0. \tag{6.4}$$

The quantity $\cos\theta_{B,D^*-l}$ is also calculated, where the lepton momentum in the CMS frame was flipped before the $\cos\theta_{B,D^*l}$ calculation. The reason for this is, that a real $D^*(2010)^-$ and ℓ^+ pair tends to be more back-to-back like, whereas uncorrelated combinations of $D^*(2010)^-$ and ℓ^+ do not show this behaviour. Therefore, $|\cos\theta_{B,D^*-l}| > 1.0$ is required.

6.4.2 Preselection Requirements

Before applying the best B^0 selection network, some simple precuts are applied to cut away obvious background events. To suppress continuum background, the 2$^{\text{nd}}$ Fox-Wolfram-Moment [38] of the event is required to be smaller than 0.6. In addition, to suppress background from $c\bar{c}$ continuum events, the $D^*(2010)^-$ momentum in the CMS frame is required to be $0.15\,\text{GeV}/c < p^{\text{cms}}_{D^*(2010)^-} < 2.5\,\text{GeV}/c$. The B^0 vertex fit is required to be successful and $\chi^2/_{dgf} < 20$ is required. The mass of D^0 candidates is required to be $1.82\,\text{GeV}/c^2 < m_{D^0} < 1.9\,\text{GeV}/c^2$ and the mass difference of $D^*(2010)^-$ and D^0 candidates is required to be $\Delta m < 0.17\,\text{GeV}/c^2$. To cut away outliers $|\cos\theta_{B,D^*l}| < 20$ is required. The latter cut is to reject obvious background candidates in non-physical regions which passed by chance the reconstruction process. Additionally, in the channel $\overline{D}^0 \to K^+\pi^-\pi^0$ only the D^0 candidates with the 10 highest Dalitz weights are kept. The efficiency ϵ_{pre} and Monte Carlo statistical error of this preselection is shown in table 6.3.

Channel	$\boldsymbol{\epsilon_{pre}}$ [%]
$\overline{D}^0 \to K^+\pi^-$	92.90 ± 0.16
$\overline{D}^0 \to K^+\pi^-\pi^0$	81.78 ± 0.24
$\overline{D}^0 \to K^+\pi^-\pi^+\pi^-$	90.69 ± 0.17
Combined	88.45 ± 0.11

Table 6.3: Preselection efficiency per channel.

6.4.3 Best B^0 Neural Network

For the best B^0 selection a neural network was trained which tries to use all information available for the B^0 candidate. It uses information of the B^0 candidate itself, such as invariant mass or momentum. In addition information of the children such as invariant mass, momentum and angles between children is used. Also information about the grandchildren and great-grandchildren of the B^0 is used. Apart of invariant mass, momemtum and angular distributions, the PID information of final state particles is used. A detailed list of input variables and other network details are shown in appendix E.

The mass difference Δm between $D^*(2010)^-$ and D^0 candidates was not used in this network. If used, it would not be possible to estimate the fake D^* background from the sideband. The requirement on $|\cos\theta_{B,D^*l}|$, explained in equation 6.4, was also not yet applied as this variable is needed to extract the signal yield and estimate other background fractions.

6.4.4 Best B^0 Selection

In each event, the B^0 candidate with the largest network output NN_{out} is determined and $NN_{out} \geq -0.5$ is required. In addition the $\overline{D}^0$ mass is required to be $1.83\,\mathrm{GeV}/c^2 < m_{D^0} < 1.886\,\mathrm{GeV}/c^2$ and the mass difference of $D^*(2010)^-$ and D^0 candidates is required to be $0.1435\,\mathrm{GeV}/c^2 < \Delta m < 0.1473\,\mathrm{GeV}/c^2$. For illustration the mass difference distribution is shown in figure 6.3, without the cut applied.

The mathematically correct $|\cos\theta_{B,D^*l}| \leq 1.0$ cut from equation 6.4 was changed to $|\cos\theta_{B,D^*l}| \leq 1.075$, as due to detector resolution effects and FSR, the measured missing mass can be larger than physically allowed. At the same time the $|\cos\theta_{B,D^*-l}|$ with flipped lepton momentum is required to be $|\cos\theta_{B,D^*-l}| > 1.075$. The distribution of $\cos\theta_{B,D^*l}$, without the $\cos\theta_{B,D^*l}$ cut applied, is shown in figure 6.4.

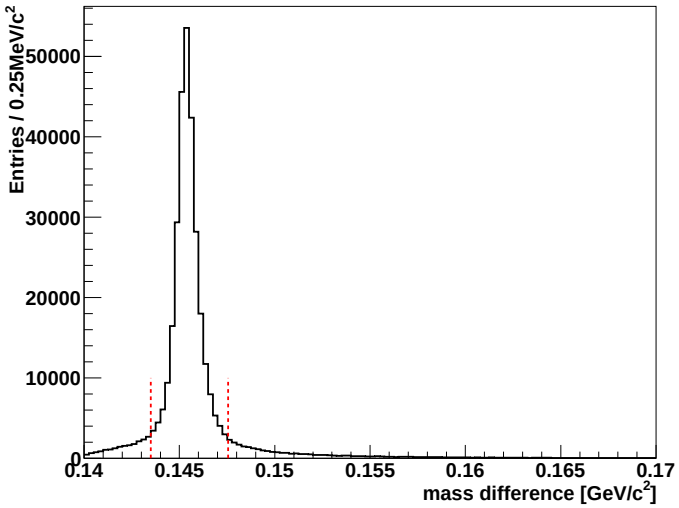

Figure 6.3: $D^*(2010)^- - D^0$ mass difference distribution of B^0 candidates. Cuts are indicated by red dashed lines. The signal MC was scaled to the number of expected events in data.

The cuts on NN_{out}, m_{D^0}, Δm, $|\cos\theta_{B,D^*l}|$ and $|\cos\theta_{B,D^*-l}|$ have been determined in a multi-dimensional minimization to optimize the figure of merit (FOM)

$$\mathrm{FOM} = \frac{S}{\sqrt{S+B}}, \tag{6.5}$$

where S and B are the number of signal and background events in the signal region.

The purity in the signal region on simulated data is $63.76 \pm 0.09\%$ and the efficiency of the best B^0 selection ϵ_{best} is shown in table 6.4. The efficiency is given as efficiency from the preselection to the final sample. The error of the efficiency is Monte Carlo statistical error only. The purity includes other uncertainties (see section 6.5 for details).

To check the continuum background model, the entire reconstruction and selection was applied to off-resonance data. As shown in figure 6.4, the continuum background Monte Carlo is in agreement with the off-resonance data.

Channel	ϵ_{best} [%]
$\overline{D}^0 \to K^+\pi^-$	79.46 ± 0.27
$\overline{D}^0 \to K^+\pi^-\pi^0$	70.61 ± 0.32
$\overline{D}^0 \to K^+\pi^-\pi^+\pi^-$	72.03 ± 0.28
Combined	74.03 ± 0.17

Table 6.4: Best B^0 selection efficiency per channel.

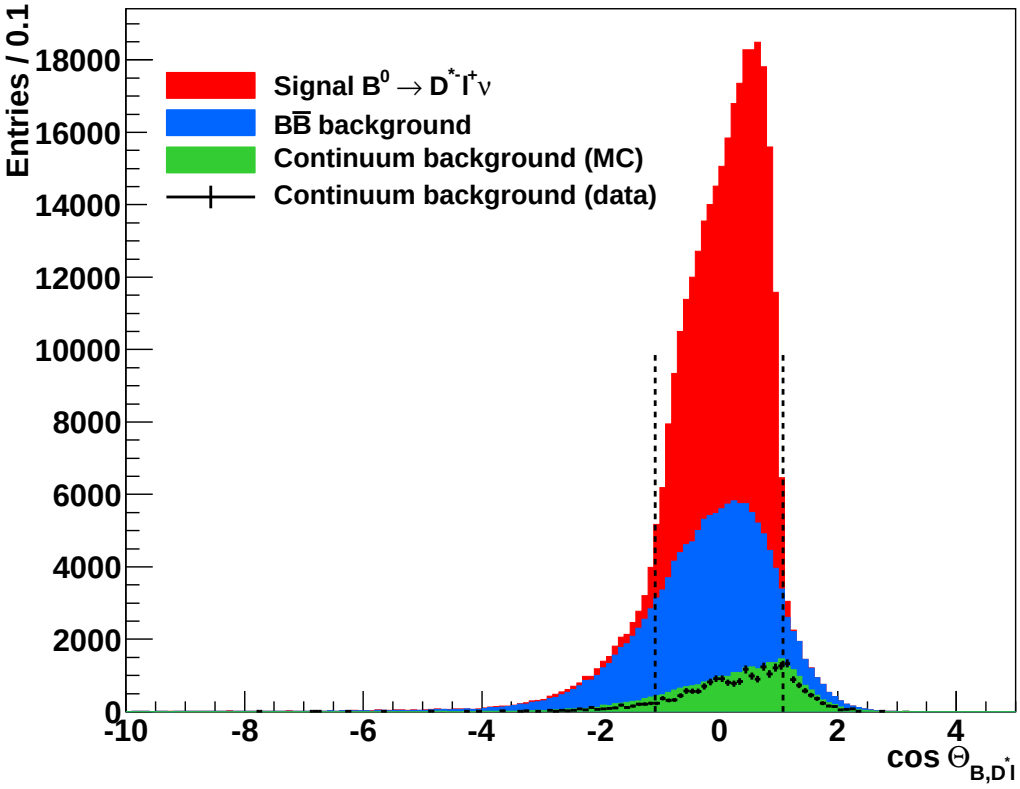

Figure 6.4: $\cos\theta_{B,D^*l}$ distribution for B^0 candidates. The signal MC (red) was scaled to the number of events expected in data. Background from $B^0\overline{B}^0$ and B^+B^- events (blue) is taken from generic MC. Background from continuum events was estimated from generic MC (green) as well as from off-resonance data (black). The cuts which define the signal region are indicated by black dashed lines.

6.5 Expected Signal Yield and Purity in Data

The expected amount of signal events in data can be estimated by using the measured branching ratios [6], the total number of $B\overline{B}$ pairs in data and the total efficiency for each decay mode. The measured branching ratios for each decay step in the decay chain are:

$$BR(B^0 \to D^*(2010)^- l^+ \nu_l) = (5.16 \pm 0.11)\% \Rightarrow (10.32 \pm 0.22)\% \quad (e \text{ and } \mu), \tag{6.6}$$
$$BR(D^*(2010)^- \to D^0 \pi^+) = (67.7 \pm 0.5)\%, \tag{6.7}$$
$$BR(D^0 \to K^- \pi^+) = (3.89 \pm 0.05)\%, \tag{6.8}$$
$$BR(D^0 \to K^- \pi^+ \pi^0) = (13.9 \pm 0.5)\%, \tag{6.9}$$
$$BR(D^0 \to K^- \pi^+ \pi^+ \pi^-) = (8.10 \pm 0.20)\%. \tag{6.10}$$

The collected data by Belle for experiments 61 to 65 is estimated to have $N_{B\bar{B}} = 114.856 \pm 1.720$ million $B\overline{B}$ pairs. Assuming an equal decay of $\Upsilon(4S)$ to $B^0\overline{B}^0$ and B^+B^-, the number of B^0-Mesons is $N_{B^0} = N_{B\bar{B}}$. The total efficiency ϵ_{tot} per decay channel is given by

$$\epsilon_{tot} = \epsilon_{rec} \cdot \epsilon_{pre} \cdot \epsilon_{best}, \tag{6.11}$$

where ϵ_{rec} is the reconstruction efficiency, ϵ_{pre} is the preselection efficiency and ϵ_{best} is the best B^0 selection efficiency.

The expected signal yield in data is then given by

$$N_{exp} = N_{B^0} \cdot BR \cdot \epsilon_{tot} - \frac{N_{B^0}}{2} \cdot BR^2 \cdot \epsilon_{tot}^2, \tag{6.12}$$

where BR is the branching ratio for the decay chain and ϵ_{tot} the total efficiency. The second term takes into account that we select only one candidate per event. The expected signal yield in data and the total efficiency are summarized in table 6.5 for each channel. The table shows also the purity of the selection per channel. The errors of purity and expected signal yield are statistical errors including the uncertainties of the branching ratio, the number of $B\overline{B}$ pairs and the efficiency. The error of the total efficiency is Monte Carlo statistical error only.

Channel	**BR** [%]	$\boldsymbol{N_{exp}}$	$\boldsymbol{\epsilon_{tot}}$ [%]	$\boldsymbol{p}$ [%]
$\overline{D}^0 \to K^+\pi^-$	0.2718 ± 0.0030	63788 ± 1239	20.57 ± 0.14	69.57 ± 0.15
$\overline{D}^0 \to K^+\pi^-\pi^0$	0.9711 ± 0.0055	53061 ± 932	4.79 ± 0.04	62.92 ± 0.17
$\overline{D}^0 \to K^+\pi^-\pi^+\pi^-$	0.5659 ± 0.0034	65837 ± 1129	10.14 ± 0.07	59.53 ± 0.15
Combined	1.8088 ± 0.0071	182685 ± 1917	8.85 ± 0.04	63.76 ± 0.09

Table 6.5: Branching ratio, expected signal yield, total efficiency and purity for each channel.

6.6 Tag Side B-Meson

The vertex of the accompanying B-meson is obtained by using `TagV` [37], the Belle default class for obtaining the tag side vertex. The B flavor is determined by using the neural network based flavor tagger algorithm, described in chapter 5, as well as the multi-dimensional likelihood tagger [32]. Vertex and flavor information of the tag side B are stored together with vertex and flavor information of the signal B and serve as input to the mixing and wrong tag fraction fit.

6.7 Comparison with Previous Analysis

Previous reconstruction and selection code, which was used for a $B^0\overline{B}^0$-mixing measurement [39] as well as flavor tagger validation [32], is estimated to have a total efficiency of $\epsilon_{tot} \approx 2.3\%$ and a purity of $p \approx 80\%$. These results have been obtained with the old SVD1 as well as old tracking code. Those values are similar for SVD2 data with old tracking and change to $\epsilon_{tot} \approx 4.2\%$ and $p \approx 74\%$ on SVD2 data with new tracking as Belle internal studies have shown.

To estimate the individual effect of the new tracking code and the improved selection due to neural networks, the presented analysis was applied to SVD2 data with old tracking code and $\epsilon_{tot} \approx 5.3\%$ and $p \approx 62\%$ was obtained. With the results on SVD2 new tracking $\epsilon_{tot} \approx 8.9\%$ and $p \approx 64\%$ one can estimate the improvement in efficiency due to the new tracking to be about $70 - 80\%$. The improvement due to new analysis techniques using neural networks, is estimated to be about $110 - 130\%$.

For a better comparison between the methods, one can regard ϵ_{tot} times p

$$\epsilon_{tot} \cdot p = \frac{S}{S_0} \cdot \frac{S}{S+B} = \frac{\mathrm{FOM}^2}{S_0}, \tag{6.13}$$

where S_0 is the initial number of signal events in a given sample and FOM the figure of merit. The results, assuming an amount of data which corresponds to $S_0 = 10000$ $B^0 \to D^*(2010)^- \ell^+ \nu$ events, are shown in table 6.6.

Scenario	**FOM**
old analysis on old tracking	13.6
old analysis on new tracking	17.6
new analysis on old tracking	18.1
new analysis on new tracking	23.9

Table 6.6: Comparison of the figure of merit (FOM) for both analysis.

7 Conclusion and Outlook

The goal of this thesis was the improvement of existing flavor tagging algorithms, used in time dependent CP violation measurements at Belle. In addition, the validation process of such tagging algorithms on real data was started.

In the first part, a new flavor tagging algorithm was presented. It uses artificial neural networks to replace the likelihood based method used in existing algorithms. In addition, it uses more information in later steps of the algorithm. A validation on simulated data indicates a relative improvement of 2.7% on SVD2 data with new tracking and 2.0% on SVD1 data with old tracking. The new tagging algorithm also returns continuous, non-peaking output. By December 2009, the algorithm was included in the Belle software library and made available for the collaboration.

For fast development and testing of the general ability of artificial neural networks to improve the flavor tagging algorithms, some concepts, such as the track level categories, have been adopted from the existing tagging algorithm. An idea for further improvement is to remove these cut based categories and use artificial neural networks, thus using the probability interpretation of their output to decide which category or categories a track most likely belongs to.

In the theory chapter, the general principle of validating a flavor tagging algorithm on real data, by a $B^0\overline{B}^0$-mixing and wrong tag fit, was described. Due to the changed systematics, introduced by the new tracking code, results from the old analysis can not be applied directly. A careful study is necessary and therefore it was decided to do so by developing a new analysis, based only on the new tracking and using advanced analysis techniques, such as artificial neural networks.

The first step, selecting a $B^0 \to D^*(2010)^-\ell^+\nu$ enriched sample, was presented in the second part of this thesis. The efficiency $\epsilon = (8.85 \pm 0.04)\%$ of this selection could be improved by $\approx 120\%$, compared to previous ones, due to neural networks. Another $\approx 75\%$ improvement is due to the new tracking, which shows clearly that systematics have changed. The purity $p = (63.76 \pm 0.09)\%$ of the new selection is slightly worse than in previous ones, however, this can be taken into account in the $B^0\overline{B}^0$-mixing and wrong tag fit.

The selection was studied and developed using simulated data. The continuum background was cross-checked by using real off-resonance data. It is necessary to further investigate the $B\overline{B}$ background as this is the main background to the $B^0\overline{B}^0$-mixing fit and has finite lifetime. In general, however, the selection is ready to be applied to real data and all necessary input variables for the $B^0\overline{B}^0$-mixing and wrong tag fit are available.

A Flavor Tagger Network Details

This appendix will present detailed information about the training results of the networks in the flavor tagger. First the definition of variable names will be given. Then for each network the plots for correlation of input variables, network output and purity as function of network output are presented. A table with the variables used for each training and their significance and contribution to the final results will be given.

A.1 Definitions of Variables and Abbreviations

q	Electric charge of track
p_{cms}	Momentum of track in CMS frame
p_{lab}	Momentum of track in laboratory frame
p_{miss}	Missing momentum in CMS frame
M_{recoil}	Hadronic recoil mass
PID	Particle identification (PID) likelihood ratios (see section 5.2.2)
θ_{lab}	Polar angle of track in laboratory frame
α_{thr}	Cosine of angle between track and thrust axis on tag side in CMS frame
$class$	Variable to assign event to certain class (see caption text for details)
$mass$	Difference of invariant mass of Λ to nominal Λ mass
Δz	z difference of the Λ daughters at vertex
θ_{defl}	Angle difference between Λ momentum and vector from IP to Λ vertex
V_{perp}	Distance in $r - \varphi$ plane between Λ vertex and IP
NN	Neural network output
$\mathcal{L}$	Likelihood combination of network outputs (see section A.12)

Table A.1: Definitions of the variable names used in the tables in the appendix.

rank	Variables ranked by importance for network output
prep.	Preprocessing flag
add. sig.	Added significance of this variable
only this	Significance of this variable alone
sig. loss	Significance loss when removing this variable
correl.	Global correlation to other variables
node	Input node number (for comparison with correlation plot)

Table A.2: Abbreviations of column names used in the tables in the appendix.

A.2 Electron Track Level Network

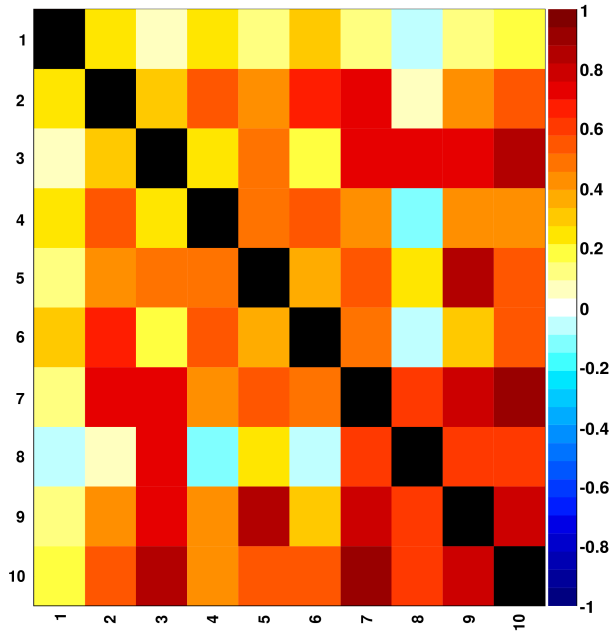

Figure A.1: Correlation matrix of input variables of the electron track level network.

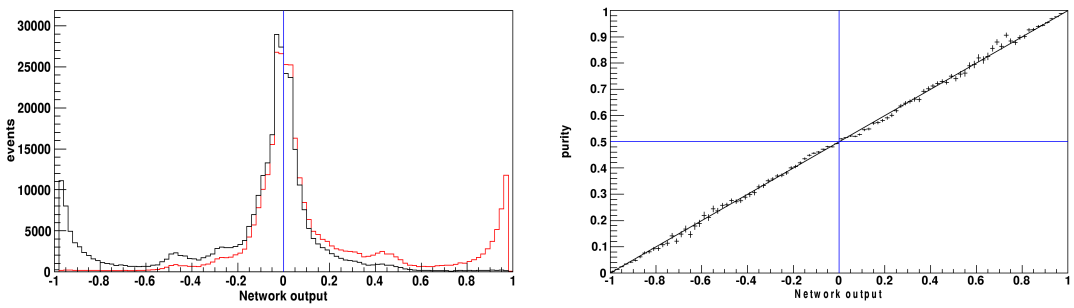

Figure A.2: Network output (left) and purity as function of network output (right) of the electron track level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$q \cdot p_{\text{cms}}$	94	224.83	224.83	125.90	79.0%	6
2	$q \cdot M_{\text{recoil}}$	94	47.68	161.62	45.44	72.4%	4
3	$q \cdot \text{PID}$	93	12.20	102.12	15.65	95.6%	7
4	$q \cdot p_{\text{miss}}$	94	15.91	98.56	11.46	92.4%	5
5	$q \cdot M_{\text{recoil}}$	93	16.10	21.14	16.38	90.0%	8
6	$q \cdot \text{PID}$	94	5.76	146.25	4.58	87.1%	2
7	$q \cdot p_{\text{miss}}$	93	3.99	82.86	3.63	96.8%	9
8	$q \cdot p_{\text{cms}}$	93	0.64	119.27	0.35	95.5%	10
9	$q \cdot \theta_{\text{lab}}$	93	0.26	46.95	0.26	90.8%	3

Table A.3: Results from electron track level training. The total significance of the training is $231.37\,\sigma$.

A.3 Muon Track Level Network

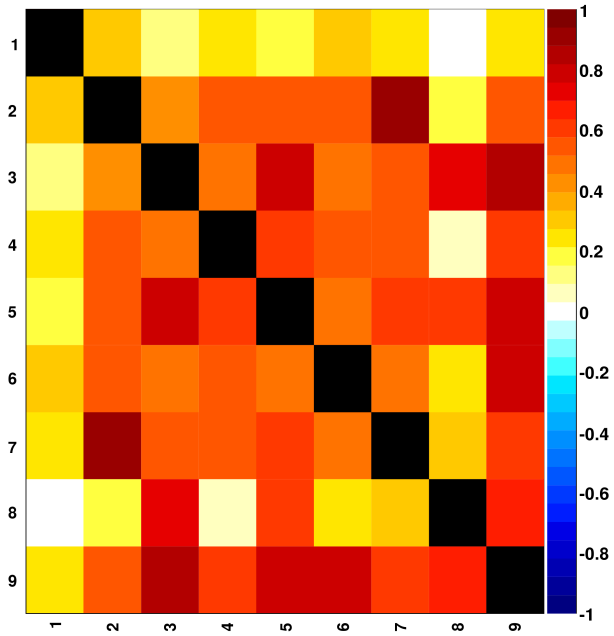

Figure A.3: Correlation matrix of input variables of the muon track level network.

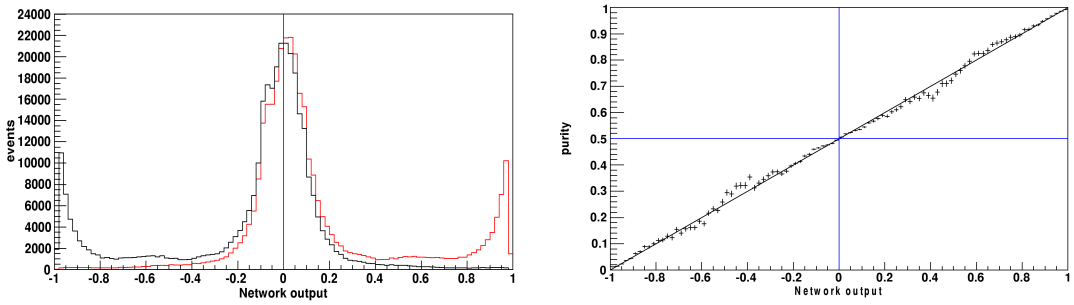

Figure A.4: Network output (left) and purity as function of network output (right) of the muon track level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$q \cdot p_{\text{cms}}$	94	202.88	202.88	65.36	89.4%	6
2	$q \cdot \text{PID}$	94	102.65	193.75	44.38	90.4%	2
3	$q \cdot M_{\text{recoil}}$	93	42.76	14.01	12.32	86.3%	8
4	$q \cdot M_{\text{recoil}}$	94	23.73	163.24	24.63	79.4%	4
5	$q \cdot \theta_{\text{lab}}$	93	8.85	97.08	5.48	89.1%	3
6	$q \cdot \text{PID}$	93	6.44	170.58	6.71	90.5%	7
7	$q \cdot p_{\text{miss}}$	93	6.02	113.66	4.89	86.7%	5
8	$q \cdot p_{\text{cms}}$	93	3.62	158.78	3.62	96.4%	9

Table A.4: Results from muon track level training. The total significance of the training is $232.93\,\sigma$.

A.4 Lepton Event Level Network

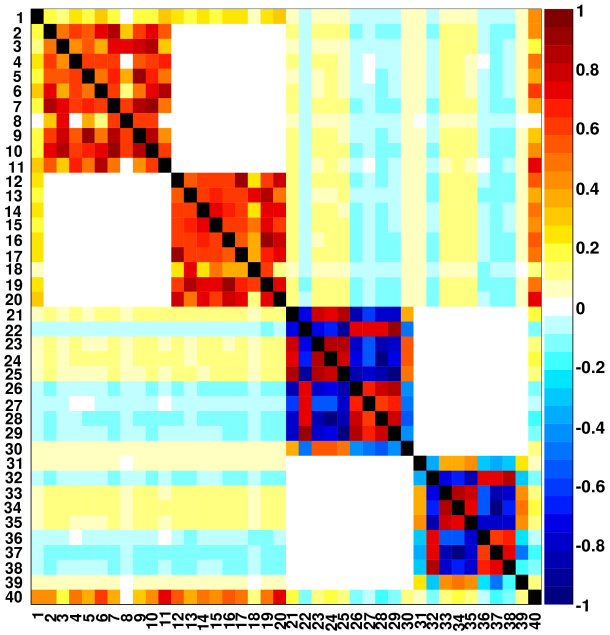

Figure A.5: Correlation matrix of input variables of the lepton event level network.

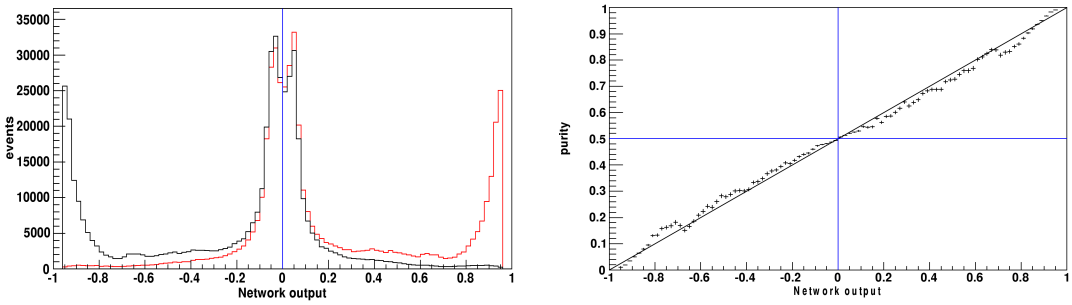

Figure A.6: Network output (left) and purity as function of network output (right) of the lepton event level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$\mathcal{L}_{lepton}$	94	366.43	366.43	29.67	99.4%	40
2	$q \cdot p_{\text{miss}}(\mu, 1)$	94	22.49	168.95	17.16	83.3%	15
3	$q \cdot p_{\text{miss}}(e, 1)$	94	25.43	145.95	6.08	95.8%	5
4	$q \cdot p_{\text{miss}}(\mu, 2)$	93	10.81	52.30	1.04	99.7%	37
5	$q \cdot p_{\text{miss}}(e, 2)$	94	10.87	55.16	1.04	99.1%	24
6	$q \cdot p_{\text{cms}}(e, 1)$	93	5.78	163.01	5.64	96.5%	10
7	$q \cdot \theta_{\text{lab}}(e, 1)$	93	7.88	84.15	5.87	90.2%	3
8	$q \cdot M_{\text{recoil}}(\mu, 1)$	93	5.61	43.08	8.22	87.5%	18
9	$q \cdot p_{\text{cms}}(\mu, 1)$	93	8.89	196.02	5.12	98.3%	19
10	$q \cdot M_{\text{recoil}}(e, 1)$	94	6.72	184.51	4.34	83.0%	4
11	$q \cdot M_{\text{recoil}}(e, 2)$	93	5.02	26.26	4.26	89.9%	27
12	$q \cdot \text{PID}(e, 2)$	94	4.94	52.44	3.37	89.8%	21
13	$\text{NN}(e, 1)$	94	3.18	257.84	7.51	98.8%	11
14	$\text{NN}(\mu, 1)$	94	5.12	257.82	6.56	98.8%	20
15	$q \cdot M_{\text{recoil}}(\mu, 2)$	93	4.87	25.51	7.41	88.5%	36
16	$q \cdot p_{\text{cms}}(\mu, 2)$	93	4.90	32.45	7.22	91.5%	38
17	$q \cdot M_{\text{recoil}}(\mu, 2)$	94	3.93	47.69	5.71	91.0%	33
18	$q \cdot p_{\text{cms}}(\mu, 2)$	94	4.46	42.19	3.64	92.7%	35
19	$q \cdot M_{\text{recoil}}(e, 2)$	94	4.21	50.17	3.19	91.4%	23
20	$q \cdot \text{PID}(e, 2)$	93	3.46	45.80	3.32	91.3%	26
21	$q \cdot \text{PID}(e, 1)$	94	3.39	162.50	4.31	93.8%	2
22	$\text{NN}(\mu, 2)$	94	3.34	40.25	3.31	66.7%	39
23	$q \cdot \text{PID}(e, 1)$	93	2.77	140.43	2.83	97.6%	7
24	$q \cdot \theta_{\text{lab}}(\mu, 1)$	93	1.96	135.40	2.43	89.0%	13
25	$q \cdot \text{PID}(\mu, 1)$	93	2.28	192.45	0.60	94.7%	17
26	$q \cdot \text{PID}(\mu, 2)$	94	2.15	27.29	2.40	43.3%	31
27	$q \cdot \theta_{\text{lab}}(\mu, 2)$	93	2.00	37.13	2.00	89.8%	32
28	$\text{NN}(e, 2)$	94	1.86	38.69	1.80	72.9%	30
29	$q \cdot p_{\text{miss}}(e, 1)$	93	0.99	132.67	0.80	98.0%	9
30	$q \cdot p_{\text{cms}}(e, 2)$	94	0.95	46.68	1.39	99.0%	25
31	$q \cdot p_{\text{cms}}(\mu, 1)$	94	0.94	216.68	0.78	96.9%	16
32	$q \cdot p_{\text{cms}}(e, 2)$	93	0.89	45.06	0.87	97.9%	29
33	$q \cdot p_{\text{miss}}(e, 2)$	93	0.77	54.31	0.75	99.4%	28
34	$q \cdot \text{PID}(\mu, 1)$	94	0.54	205.48	0.55	95.1%	12
35	$q \cdot p_{\text{cms}}(e, 1)$	94	0.50	228.78	0.46	93.3%	6
36	$q \cdot \theta_{\text{lab}}(e, 2)$	93	0.46	40.98	0.46	92.1%	22
37	$q \cdot M_{\text{recoil}}(\mu, 1)$	94	0.45	194.96	0.45	85.3%	14
38	$q \cdot M_{\text{recoil}}(e, 1)$	93	0.25	3.09	0.24	91.2%	8
39	$q \cdot p_{\text{miss}}(\mu, 2)$	94	0.10	52.83	0.10	99.5%	34

Table A.5: Results from lepton event level training. Variable name (i, j) gives affiliation to tracks in the event, where i indicates whether the track was electron or muon and j whether it was the track with best (1) or second best (2) track level network output. The total significance of the training is $369.01\,\sigma$.

A.5 Lambda Track Level Network

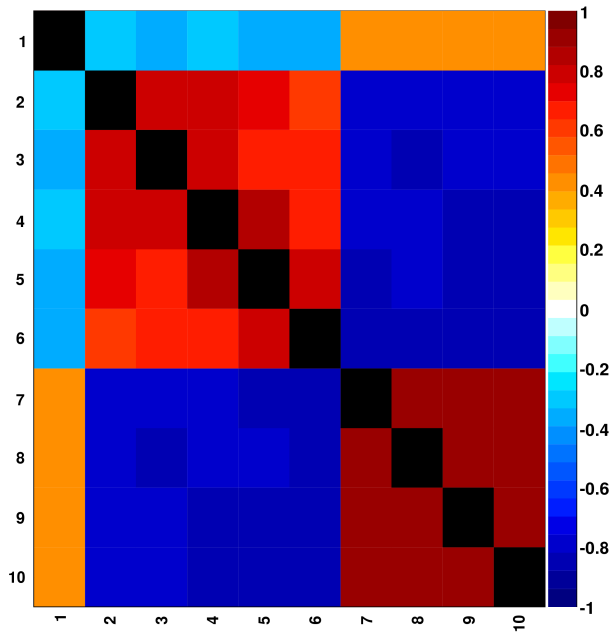

Figure A.7: Correlation matrix of input variables of the lambda track level network.

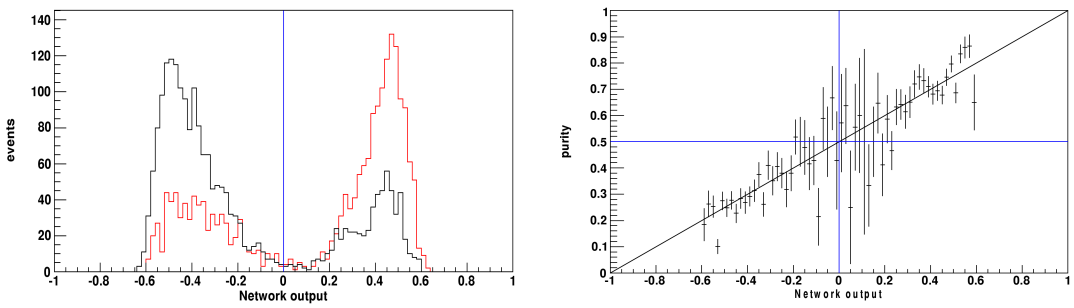

Figure A.8: Network output (left) and purity as function of network output (right) of the lambda track level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$q \cdot \theta_{\text{defl}}$	94	26.75	26.75	4.67	94.9%	8
2	$q \cdot V_{\text{perp}}$	94	5.53	26.52	2.72	95.8%	10
3	$q \cdot mass$	93	3.76	20.52	3.43	88.3%	2
4	$q \cdot mass$	94	3.17	26.63	2.69	94.9%	7
5	$q \cdot \Delta z$	94	1.29	25.91	1.96	94.6%	9
6	$q \cdot \Delta z$	93	1.46	21.53	2.40	91.2%	4
7	$q \cdot class_{\Lambda}$	93	1.72	24.42	1.85	91.3%	6
8	$q \cdot V_{\text{perp}}$	93	1.48	23.56	1.63	91.3%	5
9	$q \cdot \theta_{\text{defl}}$	93	0.83	23.00	0.83	88.5%	3

Table A.6: Results from Λ track level training. $class_{\Lambda}$ specifies quality of Λ candidate and whether K_S^0 was found in same event. The total significance of the training is $27.93\,\sigma$.

A.6 Kaon without K_S^0 Track Level Network

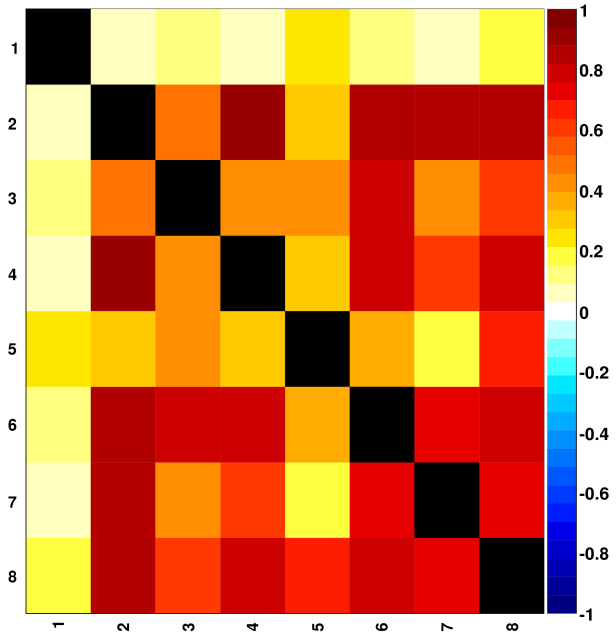

Figure A.9: Correlation matrix of input variables of the kaon without K_S^0 track level network.

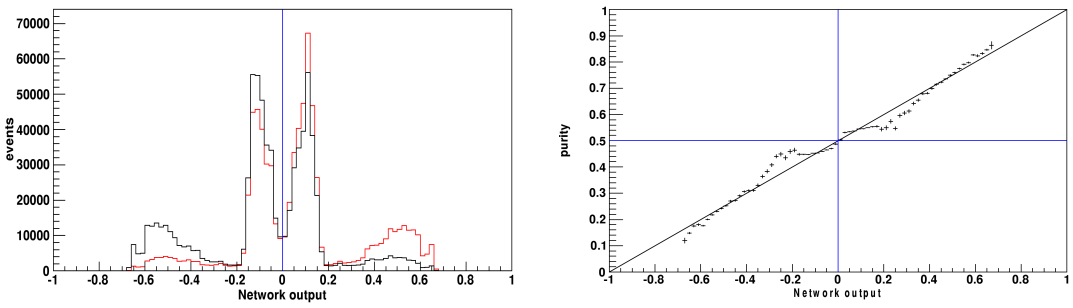

Figure A.10: Network output (left) and purity as function of network output (right) of the kaon without K_S^0 track level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$q \cdot \mathrm{PID}$	94	327.19	327.19	165.61	84.6%	5
2	$q \cdot p_{\mathrm{cms}}$	94	42.06	181.14	43.30	90.0%	3
3	$q \cdot p_{\mathrm{cms}}$	93	30.67	139.13	21.06	96.6%	6
4	$q \cdot \theta_{\mathrm{lab}}$	93	11.71	60.44	13.04	93.6%	7
5	$q \cdot \theta_{\mathrm{lab}}$	94	9.55	96.79	11.27	94.8%	4
6	$q \cdot class_K$	94	7.44	90.38	5.57	99.0%	2
7	$q \cdot \mathrm{PID}$	93	0.37	208.82	0.37	96.3%	8

Table A.7: Results from kaon without K_S^0 track level training. $class_K$ specifies whether K_S^0 was found in same event. The total significance of the training is $331.73\,\sigma$.

A.7 Kaon with K_S^0 Track Level Network

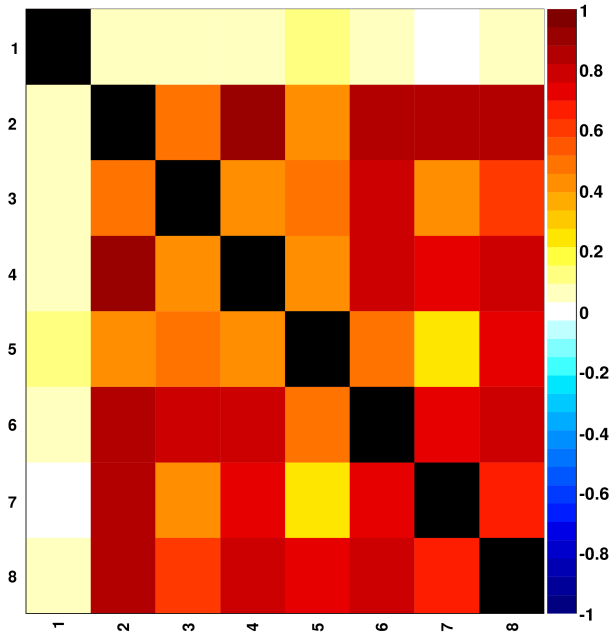

Figure A.11: Correlation matrix of input variables of the kaon with K_S^0 track level network.

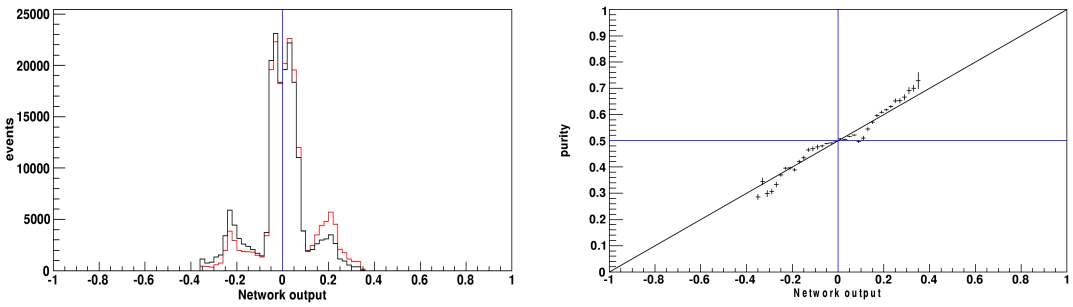

Figure A.12: Network output (left) and purity as function of network output (right) of the kaon with K_S^0 track level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$q \cdot \mathrm{PID}$	94	70.60	70.60	32.77	84.8%	5
2	$q \cdot p_{\mathrm{cms}}$	94	17.34	49.07	20.92	87.1%	3
3	$q \cdot p_{\mathrm{cms}}$	93	15.51	35.99	11.42	95.7%	6
4	$q \cdot \theta_{\mathrm{lab}}$	93	6.40	17.87	5.62	89.6%	7
5	$q \cdot \theta_{\mathrm{lab}}$	94	4.66	25.68	4.94	92.3%	4
6	$q \cdot class_K$	94	1.81	25.41	3.00	98.4%	2
7	$q \cdot \mathrm{PID}$	93	2.50	50.07	2.50	95.8%	8

Table A.8: Results from kaon with K_S^0 track level training. $class_K$ specifies whether K_S^0 was found in same event. The total significance of the training is $74.82\,\sigma$.

A.8 Strangeness Event Level Network

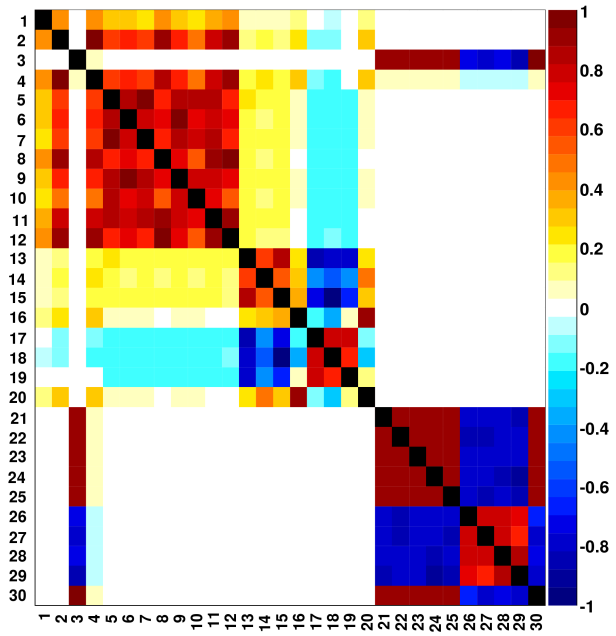

Figure A.13: Correlation matrix of input variables of the strangeness event level network.

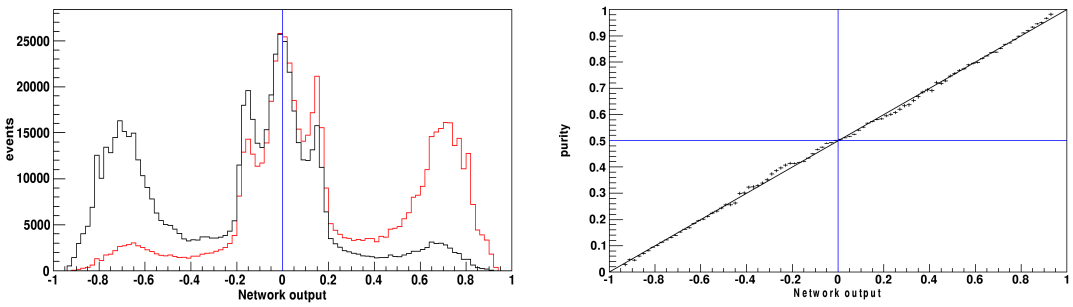

Figure A.14: Network output (left) and purity as function of network output (right) of the strangeness event level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$\mathcal{L}_{kaon}$	94	415.49	415.49	23.35	99.6%	2
2	PID$(K,1)$	94	37.05	382.68	11.48	98.1%	8
3	PID$(K,2)$	93	27.77	7.04	11.16	91.3%	19
4	$\mathcal{L}_{\Lambda}$	94	26.44	24.20	4.13	98.2%	3
5	PID$(K,2)$	94	13.27	100.39	18.05	91.8%	16
6	NN$(K,1)$	94	16.72	390.66	10.26	98.3%	12
7	NN$(K,2)$	94	10.85	107.00	5.82	94.8%	20
8	$q \cdot \theta_{\mathrm{lab}}(K,2)$	93	8.28	41.09	2.24	98.6%	18
9	PID$(K,1)$	93	4.56	352.35	7.86	97.6%	11
10	$q \cdot \theta_{\mathrm{lab}}(K,1)$	94	6.25	253.59	6.70	95.6%	7
11	$q \cdot p_{\mathrm{cms}}(K,1)$	94	6.89	288.46	8.38	98.9%	6
12	$q \cdot p_{\mathrm{cms}}(K,1)$	93	6.88	284.75	6.61	99.0%	9
13	$q \cdot p_{\mathrm{cms}}(K,2)$	94	4.36	59.79	4.29	75.6%	14
14	$q \cdot \theta_{\mathrm{lab}}(K,1)$	93	3.63	206.02	4.27	85.9%	10
15	$q \cdot class_{\Lambda}(K,1)$	94	3.02	260.79	3.17	96.1%	5
16	$\mathcal{L}_{strange}$	94	2.79	415.06	2.68	99.6%	4
17	$q \cdot class_{\Lambda}(\Lambda)$	94	2.05	22.53	1.34	96.8%	25
18	$q \cdot class_{\Lambda}(K,2)$	94	1.43	40.27	1.23	93.2%	13
19	$q \cdot \theta_{\mathrm{defl}}(\Lambda)$	94	0.66	22.82	0.75	98.1%	22
20	$q \cdot mass(\Lambda)$	93	0.82	17.98	0.47	93.2%	26
21	$q \cdot mass(\Lambda)$	94	0.58	22.57	0.61	95.7%	21
22	$q \cdot \theta_{\mathrm{defl}}(\Lambda)$	93	0.46	19.70	0.53	89.1%	27
23	$q \cdot \Delta z(\Lambda)$	94	0.47	22.62	0.50	95.0%	23
24	$q \cdot \Delta z(\Lambda)$	93	0.47	18.39	0.42	92.2%	28
25	$q \cdot V_{\mathrm{perp}}(\Lambda)$	94	0.40	22.81	0.39	97.3%	24
26	$q \cdot p_{\mathrm{cms}}(K,2)$	93	0.31	23.64	0.31	86.1%	17
27	$q \cdot V_{\mathrm{perp}}(\Lambda)$	93	0.10	20.08	0.10	92.0%	29
28	$q \cdot \theta_{\mathrm{lab}}(K,2)$	94	0.03	41.75	0.03	98.6%	15
29	NN(Λ)	94	0.03	23.61	0.03	98.9%	30

Table A.9: Results from strangeness event level training. Variable name (i,j) gives affiliation to tracks in the event, where i indicates whether the track was Λ or kaon and j whether it was the track with best (1) or second best (2) track level network output. In case of Λ, index j is omitted. The total significance of the training is $419.92\,\sigma$.

A.9 Slow Pion Track Level Network

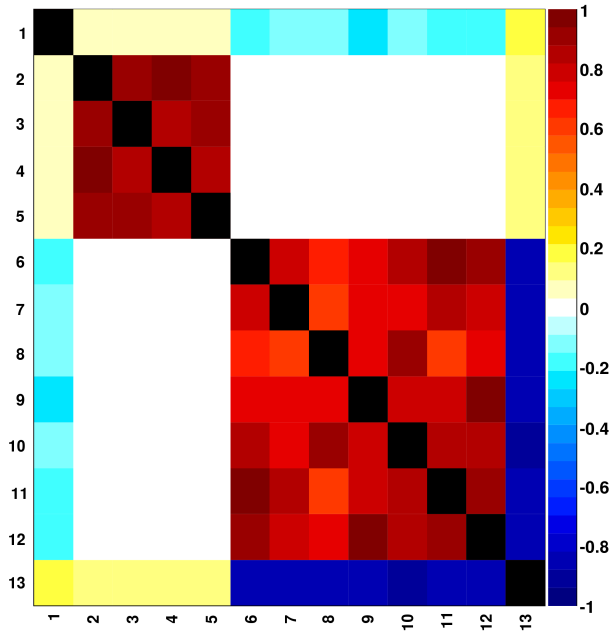

Figure A.15: Correlation matrix of input variables of the slow pion track level network.

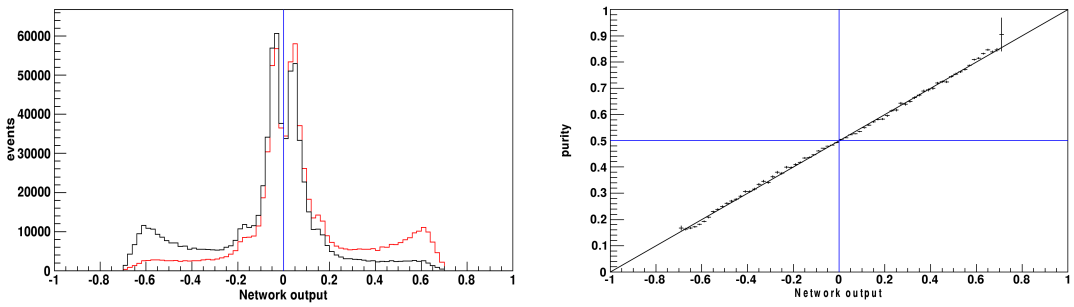

Figure A.16: Network output (left) and purity as function of network output (right) of the slow pion track level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} > 0.1)$	94	227.25	227.25	75.55	98.3%	9
2	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} > 0.1)$	94	73.27	119.77	25.72	98.6%	8
3	$q \cdot \mathrm{PID}/p_{\mathrm{lab}}$	94	34.01	194.04	20.91	99.0%	11
4	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} > 0.1)$	94	62.94	141.02	73.59	88.5%	7
5	$q \cdot \mathrm{PID} \cdot \alpha_{\mathrm{thr}}$	94	47.69	214.15	29.99	99.1%	12
6	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} < 0.1)$	94	41.46	40.73	12.32	94.3%	4
7	$q \cdot \mathrm{PID} \cdot \theta_{\mathrm{lab}}$	94	19.98	189.02	21.19	98.6%	13
8	$q \cdot \mathrm{PID}(\mathrm{PID} > 0.1)$	94	8.21	178.35	7.85	97.8%	6
9	$q \cdot \mathrm{PID}(\mathrm{PID} < 0.1)$	94	3.96	38.90	5.51	96.9%	2
10	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} < 0.1)$	94	3.20	36.42	2.24	93.2%	5
11	$q \cdot \mathrm{PID} \cdot p_{\mathrm{lab}}$	94	2.58	147.54	2.60	98.6%	10
12	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} < 0.1)$	94	1.60	36.30	1.60	93.1%	3

Table A.10: Results from slow pion track level training. For some variables, tracks have been splitted to two variables according to their PID value. The total significance of the training is $258.12\,\sigma$.

A.10 Slow Pion Event Level Network

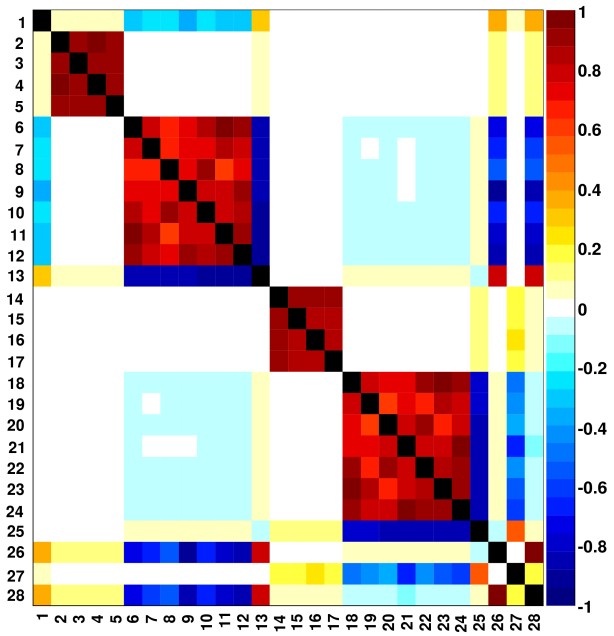

Figure A.17: Correlation matrix of input variables of the slow pion event level network.

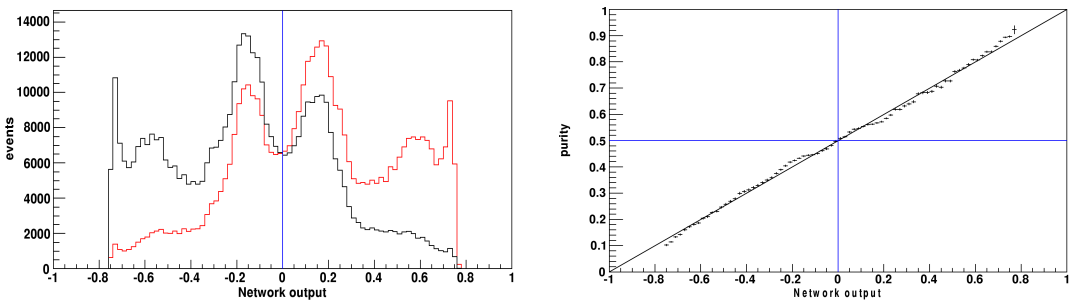

Figure A.18: Network output (left) and purity as function of network output (right) of the slow pion event level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$\mathcal{L}_{pion}$	94	293.49	293.49	42.07	98.5%	28
2	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} > 0.1)(1)$	94	17.18	265.06	8.00	98.8%	9
3	$q \cdot \mathrm{PID}/p_{\mathrm{lab}}(1)$	94	11.32	239.62	4.78	98.9%	11
4	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} > 0.1)(2)$	94	5.20	19.58	4.15	98.9%	21
5	$q \cdot \mathrm{PID} \cdot \alpha_{\mathrm{thr}}(2)$	94	6.12	17.90	2.53	99.3%	24
6	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} < 0.1)(1)$	94	4.50	39.94	3.10	92.9%	5
7	$q \cdot \mathrm{PID} \cdot \alpha_{\mathrm{thr}}(1)$	94	4.49	257.13	4.47	99.0%	12
8	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} > 0.1)(2)$	94	3.32	9.10	5.35	98.0%	20
9	$q \cdot \mathrm{PID} \cdot p_{\mathrm{lab}}(2)$	94	3.26	11.30	4.82	98.9%	22
10	$q \cdot \mathrm{PID}/p_{\mathrm{lab}}(2)$	94	3.99	16.05	2.31	98.4%	23
11	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} > 0.1)(1)$	94	3.19	194.70	2.44	93.2%	7
12	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} < 0.1)(2)$	94	2.62	12.36	1.48	92.4%	15
13	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} < 0.1)(1)$	94	2.13	38.83	2.49	94.4%	3
14	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} < 0.1)(1)$	94	2.43	42.31	1.59	95.7%	4
15	$q \cdot \mathrm{PID} \cdot \theta_{\mathrm{lab}}(2)$	94	1.88	16.71	1.52	94.5%	25
16	NN(1)	94	1.68	287.71	1.24	98.9%	26
17	$q \cdot \mathrm{PID} \cdot p_{\mathrm{lab}}(1)$	94	0.69	200.02	1.50	98.0%	10
18	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} > 0.1)(1)$	94	1.08	171.69	1.66	97.7%	8
19	$q \cdot \mathrm{PID} \cdot \theta_{\mathrm{lab}}(1)$	94	1.31	238.06	1.25	98.5%	13
20	$q \cdot \mathrm{PID}(\mathrm{PID} > 0.1)(2)$	94	1.26	14.22	1.31	97.5%	18
21	NN(2)	94	0.88	33.89	0.76	86.0%	27
22	$q \cdot \mathrm{PID}(\mathrm{PID} > 0.1)(1)$	94	0.48	223.80	0.49	97.9%	6
23	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} < 0.1)(2)$	94	0.34	11.93	0.54	92.0%	17
24	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} < 0.1)(2)$	94	0.48	12.96	0.26	93.7%	16
25	$q \cdot \mathrm{PID}(\mathrm{PID} < 0.1)(1)$	94	0.22	41.47	0.22	97.7%	2
26	$q \cdot \mathrm{PID}(\mathrm{PID} < 0.1)(2)$	94	0.18	12.56	0.18	96.8%	14
27	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} > 0.1)(2)$	94	0.09	12.74	0.09	92.5%	19

Table A.11: Results from slow pion event level training. For some variables, tracks have been split to two variables according to their PID value. Variable name (i) indicates whether the variable belongs to slow pion candidate with best (1) or second best (2) track level network output. The total significance of the training is $294.52\,\sigma$.

A.11 Combined Event Level Network

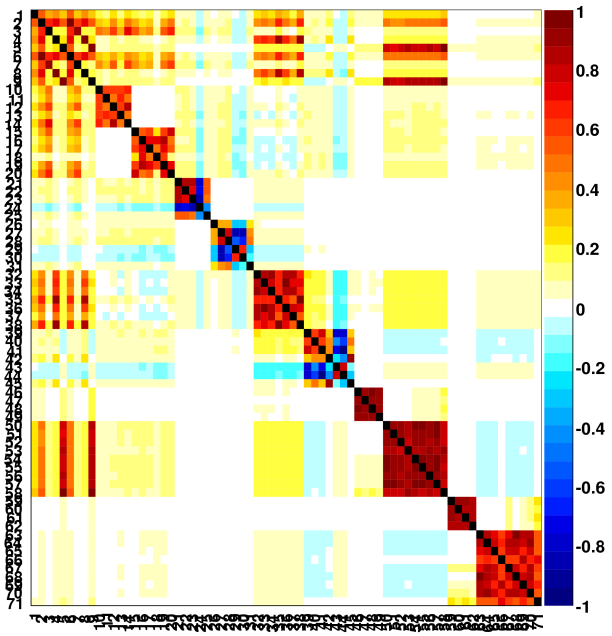

Figure A.19: Correlation matrix of input variables of the combined event level network.

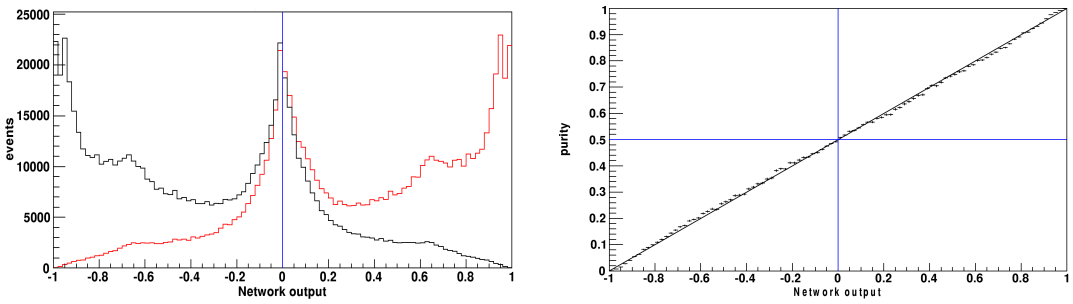

Figure A.20: Network output (left) and purity as function of network output (right) of the combined event level network.

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
1	$\mathcal{L}_{event}$	94	563.71	563.71	37.33	99.7%	2
2	NN($pion$)	93	22.32	293.60	19.73	99.0%	9
3	NN$(\pi, 1)$	94	29.82	287.76	15.15	97.3%	58
4	$q \cdot \mathrm{PID} \cdot \theta_{\mathrm{lab}}(\pi, 1)$	94	16.02	239.28	4.54	98.4%	57
5	NN($lepton$)	93	11.76	329.87	14.62	93.3%	7
6	NN($pion$)	94	8.36	299.67	9.18	99.0%	5
7	NN($strange$)	93	7.07	415.46	13.98	98.1%	8
8	$\mathcal{L}_{event}$	93	8.22	560.29	11.68	99.5%	6
9	NN($strange$)	94	8.98	428.55	8.27	99.0%	4
10	NN$(\mu, 1)$	94	4.61	257.87	5.93	97.7%	20
11	NN$(e, 1)$	94	9.41	257.50	8.09	97.9%	14
12	$q \cdot p_{\mathrm{cms}}(K, 1)$	93	8.02	283.35	3.87	98.8%	36
13	$q \cdot \theta_{\mathrm{lab}}(K, 1)$	94	6.38	252.59	3.33	94.6%	34
14	$q \cdot \mathrm{PID}(K, 1)$	94	8.10	383.36	3.52	97.8%	35
15	$q \cdot \mathrm{PID} \cdot \theta_{\mathrm{lab}}(\pi, 2)$	94	7.96	17.52	2.21	94.2%	70
16	$q \cdot p_{\mathrm{cms}}(\mu, 1)$	93	7.06	195.91	6.66	93.5%	19
17	$q \cdot \mathrm{PID}(e, 1)$	93	6.73	137.78	6.31	81.7%	13
18	NN$(\pi, 2)$	94	6.66	34.22	5.25	80.0%	71
19	$q \cdot \theta_{\mathrm{lab}}(K, 2)$	94	5.57	44.46	1.18	98.9%	41
20	$q \cdot p_{\mathrm{cms}}(K, 2)$	93	5.94	24.49	5.72	84.1%	43
21	$q \cdot p_{\mathrm{cms}}(\mu, 2)$	93	5.75	32.64	4.19	87.1%	30
22	$q \cdot p_{\mathrm{miss}}(e, 1)$	94	4.74	146.06	5.82	74.4%	11
23	$q \cdot M_{\mathrm{recoil}}(\mu, 1)$	93	4.63	42.48	3.69	82.4%	18
24	$q \cdot p_{\mathrm{miss}}(\mu, 1)$	94	4.05	169.60	5.83	81.1%	17
25	$q \cdot \mathrm{PID}(\mu, 1)$	94	4.01	205.79	3.60	82.0%	15
26	$q \cdot p_{\mathrm{cms}}(K, 2)$	94	3.93	61.70	2.15	74.6%	40
27	$q \cdot \mathrm{PID}(\mathrm{PID} > 0.1)(\pi, 2)$	94	3.30	14.70	1.64	97.9%	63
28	$q \cdot \mathrm{PID}/p_{\mathrm{lab}}(\pi, 1)$	94	1.98	245.07	2.42	98.9%	55
29	$q \cdot \mathrm{PID}(\mathrm{PID} > 0.1)(\pi, 1)$	94	3.21	238.70	3.63	98.9%	50
30	$q \cdot p_{\mathrm{miss}}(e, 2)$	94	2.74	54.69	2.27	85.1%	22
31	$q \cdot class_{\Lambda}(K, 1)$	94	2.57	260.98	3.42	95.4%	32
32	$q \cdot \mathrm{PID} \cdot \alpha_{\mathrm{thr}}(\pi, 1)$	94	1.99	260.49	1.96	98.3%	56
33	$q \cdot \mathrm{PID} \cdot p_{\mathrm{lab}}(\pi, 1)$	94	2.35	246.87	2.87	97.9%	54
34	$q \cdot class_{\Lambda}(K, 2)$	94	2.18	42.57	2.00	86.8%	39
35	$q \cdot \theta_{\mathrm{lab}}(\mu, 1)$	93	2.15	138.75	2.13	88.6%	16
36	NN($lepton$)	94	1.49	369.81	2.20	99.2%	3
37	NN$(e, 2)$	94	2.03	40.07	2.01	66.0%	25
38	$q \cdot \mathrm{PID}(K, 2)$	94	1.91	105.20	2.63	94.8%	42
39	NN$(K, 2)$	94	1.95	107.66	1.95	94.6%	45
40	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} > 0.1)(\pi, 1)$	94	1.94	233.26	1.60	93.4%	52
41	$q \cdot \theta_{\mathrm{lab}}(\mu, 2)$	93	1.90	43.76	2.64	94.5%	27
42	$q \cdot p_{\mathrm{miss}}(\mu, 2)$	94	1.85	53.27	2.13	83.2%	28
43	$q \cdot \theta_{\mathrm{lab}}(K, 1)$	93	1.62	206.13	1.61	85.5%	37
44	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} > 0.1)(\pi, 1)$	94	1.53	231.49	1.48	95.1%	51
45	$q \cdot M_{\mathrm{recoil}}(e, 1)$	94	1.45	183.43	1.52	75.1%	10
46	$q \cdot M_{\mathrm{recoil}}(\mu, 2)$	93	1.48	25.24	1.47	86.4%	29
47	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} < 0.1)(\pi, 1)$	94	1.03	39.01	1.13	94.2%	47
48	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} < 0.1)(\pi, 1)$	94	1.36	41.79	0.75	96.2%	48
49	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} < 0.1)(\pi, 2)$	94	1.22	11.72	1.06	87.8%	62
50	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} < 0.1)(\pi, 2)$	94	0.93	13.41	0.48	93.8%	61

rank	variable name	prep.	add. sig.	only this	sig. loss	correl.	node
51	$q \cdot \mathrm{PID} \cdot \alpha_{\mathrm{thr}}(\pi, 2)$	94	0.72	21.13	1.03	91.0%	69
52	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} > 0.1)(\pi, 2)$	94	0.74	24.40	0.65	87.4%	66
53	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} < 0.1)(\pi, 1)$	94	0.73	39.86	0.81	93.0%	49
54	$q \cdot p_{\mathrm{cms}}(K, 1)$	94	0.71	287.97	0.73	98.7%	33
55	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} < 0.1)(\pi, 2)$	94	0.57	12.58	0.70	91.5%	60
56	$q \cdot \mathrm{PID}(\mathrm{PID} < 0.1)(\pi, 2)$	94	0.51	12.51	0.51	95.1%	59
57	$q \cdot p_{\mathrm{cms}}(e, 2)$	94	0.44	45.80	0.43	89.7%	23
58	$q \cdot M_{\mathrm{recoil}}(e, 2)$	94	0.49	49.11	0.50	88.6%	21
59	$q \cdot \mathrm{PID} \cdot p_{\mathrm{lab}}(\pi, 2)$	94	0.48	19.09	0.59	91.8%	67
60	$q \cdot p_{\mathrm{lab}}(\mathrm{PID} > 0.1)(\pi, 2)$	94	0.47	16.78	0.40	86.5%	65
61	$q \cdot \theta_{\mathrm{lab}}(\mathrm{PID} > 0.1)(\pi, 2)$	94	0.46	16.78	0.44	92.4%	64
62	$q \cdot \mathrm{PID}(\mathrm{PID} < 0.1)(\pi, 1)$	94	0.42	41.35	0.41	97.8%	46
63	$q \cdot \mathrm{PID}/p_{\mathrm{lab}}(\pi, 2)$	94	0.28	16.64	0.28	97.9%	68
64	$q \cdot \theta_{\mathrm{lab}}(K, 2)$	93	0.24	43.96	0.24	99.0%	44
65	$q \cdot p_{\mathrm{cms}}(e, 1)$	94	0.23	228.01	0.23	90.9%	12
66	$\mathrm{NN}(K, 1)$	94	0.23	392.22	0.23	98.4%	38
67	$\mathrm{NN}(\mu, 2)$	94	0.18	41.55	0.17	63.4%	31
68	$q \cdot \alpha_{\mathrm{thr}}(\mathrm{PID} > 0.1)(\pi, 1)$	94	0.15	265.81	0.15	97.2%	53
69	$q \cdot \mathrm{PID}(e, 2)$	93	0.14	45.32	0.14	86.9%	24
70	$q \cdot \mathrm{PID}(\mu, 2)$	94	0.13	27.82	0.13	41.3%	26

Table A.12: Results from combined event level training. Variable name (i, j) gives affiliation to tracks in the event, where i indicates whether the track was electron, muon, kaon or pion and j whether it was the track with best (1) or second best (2) track level network output. The Λ tracks are not reused on this level. For some pion variables, tracks have been split to two variables according to their PID value, as in the track level training. The total significance of the training is $566.21\,\sigma$.

A.12 Likelihoods used in Flavor Tagger Training

Likelihood $\mathcal{L}_{lepton}$

The likelihood $\mathcal{L}_{lepton}$, which combines the individual network outputs $\mathrm{NN}(i,j)$ of n electrons and m muons used on track level, is constructed as

$$\mathcal{L}_{lepton} = \frac{\mathcal{L}_{B^0,\ell}}{\mathcal{L}_{B^0,\ell} + \mathcal{L}_{\overline{B}0,\ell}}, \tag{A.1}$$

where $\mathcal{L}_{B^0,\ell}$ and $\mathcal{L}_{\overline{B}0,\ell}$ are given by

$$\mathcal{L}_{B^0,\ell} = \left(\prod_{i=1}^{n} \mathcal{L}_{B^0,\ell,e,i}\right)\left(\prod_{i=1}^{m} \mathcal{L}_{B^0,\ell,\mu,i}\right) \quad \text{and} \quad \mathcal{L}_{\overline{B}0,\ell} = \left(\prod_{i=1}^{n} \mathcal{L}_{\overline{B}0,\ell,e,i}\right)\left(\prod_{i=1}^{m} \mathcal{L}_{\overline{B}0,\ell,\mu,i}\right) \tag{A.2}$$

The likelihoods $\mathcal{L}_{B^0,\ell,i,j}$ and $\mathcal{L}_{\overline{B}0,\ell,i,j}$ for each lepton track are defined as

$$\mathcal{L}_{B^0,\ell,i,j} = 1 + \mathrm{NN}(i,j) \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,\ell,i,j} = 1 - \mathrm{NN}(i,j). \tag{A.3}$$

Likelihood $\mathcal{L}_{kaon}$

The likelihood $\mathcal{L}_{kaon}$, which combines the individual network outputs $\mathrm{NN}(K,i)$ of n kaons used on track level, is constructed as

$$\mathcal{L}_{kaon} = \frac{\mathcal{L}_{B^0,K}}{\mathcal{L}_{B^0,K} + \mathcal{L}_{\overline{B}0,K}}, \tag{A.4}$$

where $\mathcal{L}_{B^0,K}$ and $\mathcal{L}_{\overline{B}0,K}$ are given by

$$\mathcal{L}_{B^0,K} = \prod_{i=1}^{n} \mathcal{L}_{B^0,K,i} \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,K} = \prod_{i=1}^{n} \mathcal{L}_{\overline{B}0,K,i}. \tag{A.5}$$

The likelihoods $\mathcal{L}_{B^0,K,i}$ and $\mathcal{L}_{\overline{B}0,K,i}$ for each kaon track i are defined as

$$\mathcal{L}_{B^0,K,i} = 1 + \mathrm{NN}(K,i) \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,K,i} = 1 - \mathrm{NN}(K,i). \tag{A.6}$$

Likelihood $\mathcal{L}_{\Lambda}$

The likelihood $\mathcal{L}_{\Lambda}$, which combines the n Λ track level network outputs $\mathrm{NN}(\Lambda,i)$ is constructed as

$$\mathcal{L}_{\Lambda} = \frac{\mathcal{L}_{B^0,\Lambda}}{\mathcal{L}_{B^0,\Lambda} + \mathcal{L}_{\overline{B}0,\Lambda}}, \tag{A.7}$$

where $\mathcal{L}_{B^0,\Lambda}$ and $\mathcal{L}_{\overline{B}0,\Lambda}$ are given by

$$\mathcal{L}_{B^0,\Lambda} = \prod_{i=1}^{n} \mathcal{L}_{B^0,\Lambda,i} \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,\Lambda} = \prod_{i=1}^{n} \mathcal{L}_{\overline{B}0,\Lambda,i}. \tag{A.8}$$

The likelihoods $\mathcal{L}_{B^0,\Lambda,i}$ and $\mathcal{L}_{\overline{B}0,\Lambda,i}$ for each Λ track i are defined as

$$\mathcal{L}_{B^0,\Lambda,i} = 1 + \mathrm{NN}(\Lambda,i) \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,\Lambda,i} = 1 - \mathrm{NN}(\Lambda,i). \tag{A.9}$$

Likelihood $\mathcal{L}_{strange}$

The combined strangeness likelihood $\mathcal{L}_{strange}$ is constructed as

$$\mathcal{L}_{strange} = \frac{\mathcal{L}_{B^0,s}}{\mathcal{L}_{B^0,s} + \mathcal{L}_{\overline{B}0,s}}, \tag{A.10}$$

where $\mathcal{L}_{B^0,s}$ and $\mathcal{L}_{\overline{B}0,s}$ are given by

$$\mathcal{L}_{B^0,s} = \mathcal{L}_{B^0,\Lambda} \cdot \mathcal{L}_{B^0,K} \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,s} = \mathcal{L}_{\overline{B}0,\Lambda} \cdot \mathcal{L}_{\overline{B}0,K}. \tag{A.11}$$

Likelihood $\mathcal{L}_{pion}$

The likelihood $\mathcal{L}_{pion}$, which combines the individual network outputs $\mathrm{NN}(i)$ of n pions used on track level, is constructed as

$$\mathcal{L}_{pion} = \frac{\mathcal{L}_{B^0,\pi}}{\mathcal{L}_{B^0,\pi} + \mathcal{L}_{\overline{B}0,\pi}}, \tag{A.12}$$

where $\mathcal{L}_{B^0,\pi}$ and $\mathcal{L}_{\overline{B}0,\pi}$ are given by

$$\mathcal{L}_{B^0,\pi} = \prod_{i=1}^{n} \mathcal{L}_{B^0,\pi,i} \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,\pi} = \prod_{i=1}^{n} \mathcal{L}_{\overline{B}0,\pi,i}. \tag{A.13}$$

The likelihoods $\mathcal{L}_{B^0,\pi,i}$ and $\mathcal{L}_{\overline{B}0,\pi,i}$ for each pion track i are defined as

$$\mathcal{L}_{B^0,\pi,i} = 1 + \mathrm{NN}(i) \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,\pi,i} = 1 - \mathrm{NN}(i). \tag{A.14}$$

Likelihood $\mathcal{L}_{event}$

The likelihood $\mathcal{L}_{event}$, which combines the event level network outputs $\mathrm{NN}(i)$, where $i = lepton, pion$ or $strange$, is constructed as

$$\mathcal{L}_{event} = \frac{\mathcal{L}_{B^0,event}}{\mathcal{L}_{B^0,event} + \mathcal{L}_{\overline{B}0,event}}, \tag{A.15}$$

where $\mathcal{L}_{B^0,event}$ and $\mathcal{L}_{\overline{B}0,event}$ are given by

$$\mathcal{L}_{B^0,event} = \prod_{i} \mathcal{L}_{B^0,event,i} \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,event} = \prod_{i} \mathcal{L}_{\overline{B}0,event,i}. \tag{A.16}$$

The likelihoods $\mathcal{L}_{B^0,event,i}$ and $\mathcal{L}_{\overline{B}0,event,i}$ for each event level network i are defined as

$$\mathcal{L}_{B^0,event,i} = 1 + \mathrm{NN}(i) \qquad \text{and} \qquad \mathcal{L}_{\overline{B}0,event,i} = 1 - \mathrm{NN}(i). \tag{A.17}$$

B Usage of Flavor Tagger

The code listing below shows how to use the neural network based flavor tagging method in a BASF module. It is integrated to the `hamlet` software library and can therefore be run in parallel to the multi-dimensional likelihood or any other tagging method.

```
1  // you need to include those header files
2  #include "hamlet/Hamlet.h"
3  #include "hamlet/Fbtag_NN1.h"
4  #include "hamlet/Fbtag_MultDimLikelihood0.h"

6  // init hamlet in your module init method
7  void your_module::init ( int* )
8  {
9          Hamlet::init();
10 }

12 // load each tagging method
13 void your_module::begin_run ( BelleEvent*, int* )
14 {
15         Hamlet::begin_run(Hamlet::MULT_DIM_LH); // MDLH method
16         Hamlet::begin_run(Hamlet::NN1); //Neural network
17 }

19 // run both tagging methods in parallel during e.g. event method
20 void your_module::event ( BelleEvent*, int* )
21 {
22         // NN tagger
23         Hamlet hamlet_NN;
24         hamlet_NN.setBcp(brec); // set B_CP side tracks
25         hamlet_NN.setTagMethod(Hamlet::NN1);
26         Fbtag_NN1 tagger_NN=hamlet_NN.fbtg_NN1();
27         double fq_nn = hamlet_NN.q(); // will return flavor*q

29         // MDLH tagger
30         Hamlet hamlet_MDL;
31         hamlet_MDL.setBcp(brec); // set B_CP side tracks
32         hamlet_MDL.setTagMethod(Hamlet::MULT_DIM_LH);
33         Fbtag_MultDimLikelihood0 tagger_MDL=hamlet_MDL.
              fbtg_mult_dim_likelihood();
34         double fq_mdl = hamlet_MDL.q(): // will return flavor*q
35 }
```

C Error Calculation

C.1 Gaussion Error Propagation

The error σ_f of a function $f(x, y)$ with x and y being variables, can be derived by using Gaussian error propagation formula

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y} COV_{xy}, \tag{C.1}$$

where σ_x and σ_y are the error of x and y, respectively. In case of independent variables x and y, the off-diagonal element of the covariance matrix is $COV_{xy} = 0$.

C.2 Errors for Validation on MC

Below the error propagation to get the statistical error of the effective efficiency in each $q \cdot r$ bin of the tagger output will be explained. The index l for the current bin, used in chapter 5, will be skipped as all calculations are done in a single bin. The wrong tag fraction per bin is defined as

$$w = \frac{W}{W + T}, \tag{C.2}$$

where W is the number of wrongly tagged events in the given bin and T is the number of correctly tagged events. The errors $\sigma_W = \sqrt{W}$ and $\sigma_T = \sqrt{T}$ are given by the common statistical error per bin. As W and T are independent, the error σ_w can be obtained by Gaussion error propagation formula

$$\sigma_w = \sqrt{\left(\frac{\partial}{\partial W} w\right)^2 (\sigma_W)^2 + \left(\frac{\partial}{\partial T} w\right)^2 (\sigma_T)^2}, \tag{C.3}$$

$$\sigma_w = \sqrt{\left(\frac{1}{W + T} - \frac{W}{(W + T)^2}\right)^2 (\sigma_W)^2 + \left(-\frac{W}{(W + T)^2}\right)^2 (\sigma_T)^2}, \tag{C.4}$$

for each flavor specific wrong tag fraction w_{B^0} and $w_{\bar{B}^0}$. The average wrong tag fraction w_{ave} and difference in wrong tag fraction Δw are defined as

$$w_{ave} = \frac{w_{B^0} + w_{\bar{B}^0}}{2}, \tag{C.5}$$

$$\Delta w = w_{B^0} - w_{\bar{B}^0} \tag{C.6}$$

and the errors $\sigma_{w_{ave}}$ and $\sigma_{\Delta w}$ are given by

$$\sigma_{w_{ave}} = \sqrt{\left(\frac{1}{2}\right)^2 (\sigma_{w_{B^0}})^2 + \left(\frac{1}{2}\right)^2 (\sigma_{w_{\bar{B}^0}})^2}, \tag{C.7}$$

$$\sigma_{\Delta w} = \sqrt{(\sigma_{w_{B^0}})^2 + (\sigma_{w_{\bar{B}^0}})^2}. \tag{C.8}$$

The event fraction ϵ per bin is given by

$$\epsilon = \frac{N}{N_{tot}}, \tag{C.9}$$

where N is the number of entries per bin and N_{tot} the total number of entries in all bins. If N_{tot} is written as sum of $N + N_{other}$, where N_{other} is the number of entries in all other bins, one can derive the error σ_ϵ the same way as for w. The error σ_ϵ is then given by

$$\sigma_\epsilon = \sqrt{\frac{N(N_{tot} - N)}{N_{tot}^3}}. \tag{C.10}$$

Finally the effective efficiency per bin is given by

$$\epsilon_{eff} = \epsilon(1 - 2w_{ave})^2 \tag{C.11}$$

and its error by

$$\sigma_{\epsilon_{eff}} = \sqrt{(1 - 2w_{ave})^2(\sigma_\epsilon)^2 + (-4\epsilon + 8\epsilon w_{ave})^2(\sigma_{w_{ave}})^2}. \tag{C.12}$$

C.3 Efficiency Errors

The efficiency for reconstruction efficiency, preselection efficiency and best B^0 selection efficiency is in general given by

$$\epsilon = \frac{N}{N_{tot}}. \tag{C.13}$$

where N is the number of events that pass a certain selection or reconstruction and N_{tot} is the total amount of all events this selection or reconstruction is applied to. As the denominator can always be written as the sum of events that pass and that don't pass this certain selection or reconstruction, one can derive the error as shown in equations C.4 and C.10. The error on the efficiency is then given by

$$\sigma_\epsilon = \sqrt{\frac{N(N_{tot} - N)}{N_{tot}^3}}. \tag{C.14}$$

D Signal Monte Carlo Configuration

Below, the config file for producing $B^0 \to D^*(2010)^- \ell^+ \nu$ signal Monte Carlo with the EvtGen generator is given.

```
# Aliases
Define dm 0.507e12

Alias MyB0 B0
Alias Myanti-B0 anti-B0
ChargeConj MyB0 Myanti-B0

Alias MyD*+ D*+
Alias MyD*- D*-

Alias MyD0 D0
Alias Myanti-D0 anti-D0

# Y(4S) -> B0 B0bar
Decay Upsilon(4S)
1.0 B0 anti-B0 MyB0 Myanti-B0      VSS_BMIX dm;
Enddecay

# B0 and B0bar -> D*lnu
Decay MyB0
0.5   MyD*-       e+   nu_e        PHOTOS HQET2 1.3 1.18 0.71;
0.5   MyD*-       mu+  nu_mu       PHOTOS HQET2 1.3 1.18 0.71;
Enddecay

Decay Myanti-B0
0.5   MyD*+       e-   anti-nu_e   PHOTOS HQET2 1.3 1.18 0.71;
0.5   MyD*+       mu-  anti-nu_mu  PHOTOS HQET2 1.3 1.18 0.71;
Enddecay

# D* -> D0
Decay MyD*-
1.0    Myanti-D0  pi-              PHOTOS VSS;
Enddecay

Decay MyD*+
1.0    MyD0  pi+                   PHOTOS VSS;
Enddecay

# D0 -> Kpi, Kpipi0, Kpipipi (PDG 2009 values scaled to sum of 1)
Decay MyD0
0.150   K-  pi+                    PHOTOS PHSP;
0.537   K-  pi+  pi0               PHOTOS D_DALITZ;
```

```
43 0.313   K-  pi+  pi+   pi-          PHOTOS PHSP;
44 Enddecay

46 Decay Myanti-D0
47 0.150   K+  pi-                      PHOTOS PHSP;
48 0.537   K+  pi-  pi0                 PHOTOS D_DALITZ;
49 0.313   K+  pi-  pi+   pi-          PHOTOS PHSP;
50 Enddecay

52 End
```

E Best B^0 Network Details

For the best B^0 network a new version of NeuroBayes was used. This version allows to do an internal boost. The events are weighted with the output of a first classification. Afterwards, the second classification, the boost, is applied. This boost learns only the deviation from the first classification and can focus on details, rather than learning the overall difference between signal and background.

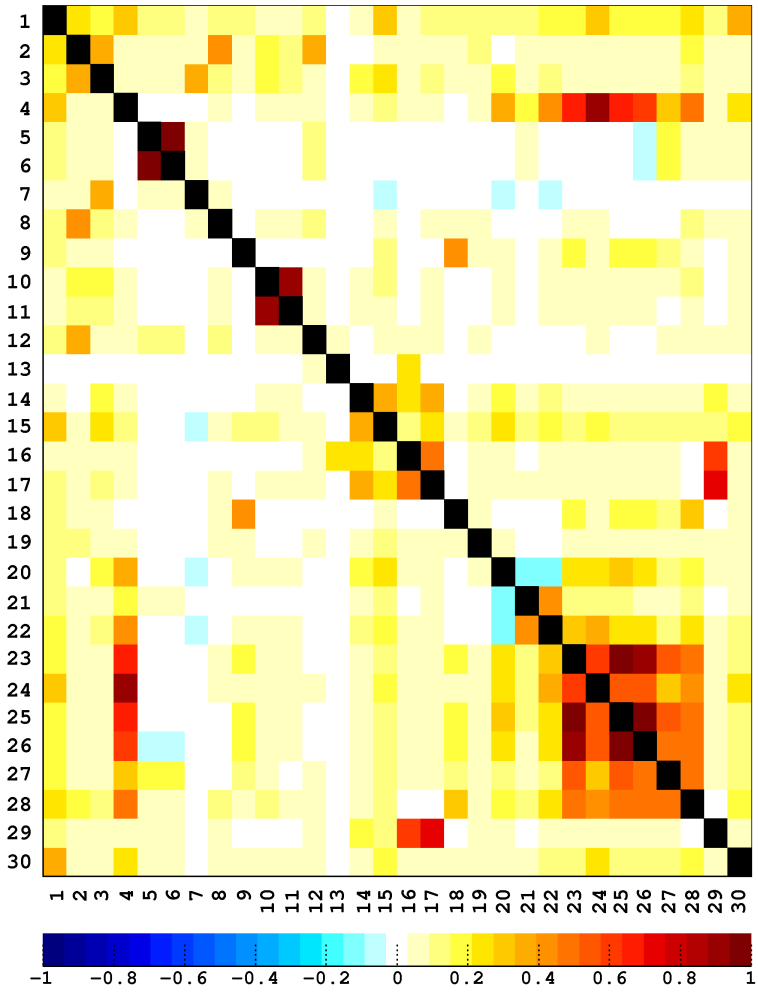

Figure E.1: Correlation matrix of input variables for the best B^0 network.

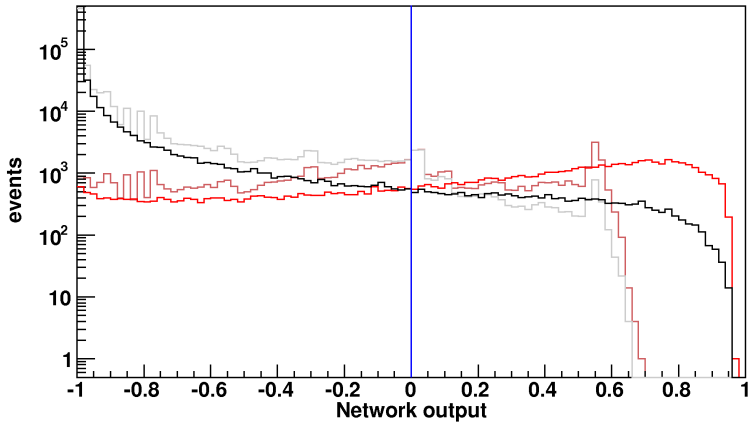

Figure E.2: Network output for the best B^0 network. Final output for background (black) and signal (red) as well as before internal boost (grey and brown).

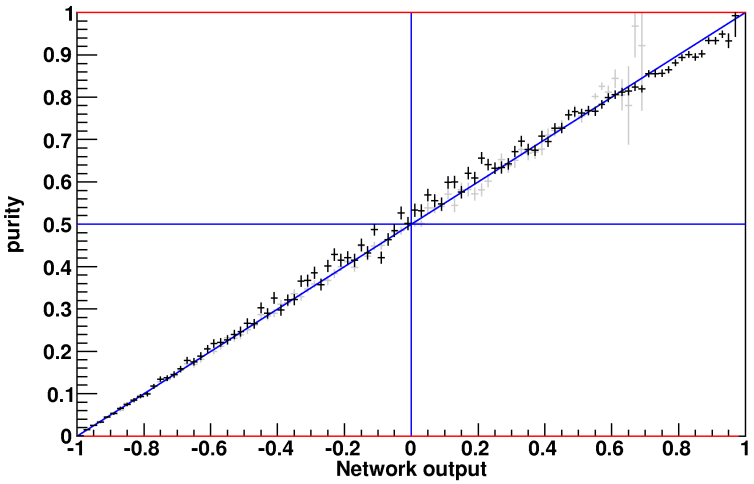

Figure E.3: Purity as function of network output for the best B^0 network (black) and before internal boost (grey).

Rank	Variable	Prepro	add. sig.	only this	sig. loss	correl.	index
1	m_{D^0}	34	316.76	316.76	202.49	32.3%	30
2	$\measuredangle(p_{\text{cms}}(D^0), p_{\text{cms}}(\pi_s))$	34	209.08	262.10	146.39	51.7%	15
3	$p_{\text{cms}}(3^{\text{rd}}\ D^0$ child)	34	171.57	267.99	78.96	93.2%	24
4	$p_{\text{cms}}(\ell)$	34	162.57	205.66	108.41	61.5%	2
5	$p_{\text{cms}}(B^0)$	34	81.76	79.98	75.85	48.5%	7
6	Dalitz Weight	34	69.47	114.75	49.87	47.3%	9
7	D^0 VF rank	34	62.54	231.78	45.74	66.9%	28
8	χ^2/dgf VF(B^0)	34	56.46	90.70	50.50	27.3%	5
9	χ^2/dgf VF(π_s)	34	33.62	81.04	24.82	82.5%	29
10	PID($1^{\text{st}}\ D^0$ child)	34	24.10	92.24	23.67	17.2%	19
11	PID($3^{\text{st}}\ D^0$ child)	34	19.67	166.16	29.11	95.6%	23
12	$p_{\text{cms}}(4^{\text{th}}\ D^0$ child)	34	17.15	167.66	21.77	97.6%	26
13	Decay Channel Flag	18	16.10	247.59	26.39	95.4%	4
14	$p_{\text{cms}}(2^{\text{nd}}\ D^0$ child)	34	21.67	146.77	22.98	71.6%	22
15	$p_{\text{cms}}(D^{*-})$	34	21.13	76.63	18.69	52.5%	14
16	PID(ℓ)	34	19.92	117.32	19.47	40.4%	8
17	Dalitz Weight rank	39	16.64	100.56	18.02	52.5%	18
18	χ^2/dgf VF(D^0)	34	16.85	155.23	16.42	64.6%	27
19	PID($2^{\text{nd}}\ D^0$ child)	34	14.26	83.55	15.66	45.3%	21
20	$p_{\text{cms}}(1^{\text{st}}\ D^0$ child)	34	15.14	94.81	18.49	67.5%	20
21	$\measuredangle(p_{\text{cms}}(D^{*-}), p_{\text{cms}}(\ell))$	34	11.02	153.08	9.76	62.1%	3
22	2^{nd} FWM	34	8.48	55.20	11.32	89.4%	10
23	Thrust	34	8.41	50.91	8.38	89.2%	11
24	PID($4^{\text{nd}}\ D^0$ child)	34	5.00	165.98	5.01	98.0%	25
25	$dr(\ell)$	34	4.60	88.37	4.64	38.1%	12
26	$dz(\pi_s)$	34	4.44	55.29	5.21	68.3%	16
27	$dr(\pi_s)$	34	4.54	86.46	4.59	78.2%	17
28	$dz(\ell)$	34	2.30	4.44	2.30	32.6%	13
29	Conf. level VF(B^0)	34	0.00	90.70	0.00	100.0%	6

Table E.1: Results from the best B^0 training. The total significance of the training is $473.12\,\sigma$.

List of Figures

List of Tables

Bibliography

[1] LHC, Large Hadron Collider. http://lhc.web.cern.ch/lhc.

[2] ATLAS, A Toroidal LHC ApparatuS. http://atlas.ch.

[3] CMS, Compact Muon Solenoid. http://cms.web.cern.ch/cms/index.html.

[4] K. Abe et al. Observation of large CP violation in the neutral B meson system. *Phys. Rev. Lett.*, 87(9):091802, 2001 arXiv:hep-ex/0107061v2.

[5] M. Kobayashi and T. Maskawa. CP-violation in the renormalizable theory of weak interaction. *Prog. Theor. Phys.*, 49:652–657, 1973.

[6] C. Amsler et al. (Particle Data Group), Physics Letters B667, 1 (2008) and 2009 partial update for the 2010 edition. http://pdg.lbl.gov.

[7] S.L. Glashow. Partial-symmetries of weak interactions. *Nucl. Phys.*, 22(4):579–588, 1961.

[8] S. Weinberg. A model of leptons. *Phys. Rev. Lett.*, 19(21):1264–1266, 1967.

[9] A. Salam. *Elementary particle theory*, 1968.

[10] D.J. Gross and F. Wilczek. Ultraviolet Behavior of Non-Abelian Gauge Theories. *Phys. Rev. Lett.*, 30(26):1343–1346, 1973.

[11] H.D. Politzer. Reliable Perturbative Results for Strong Interactions? *Phys. Rev. Lett.*, 30(26):1346–1349, 1973.

[12] D.J. Gross and F. Wilczek. Asymptotically Free Gauge Theories. I. *Phys. Rev. D*, 8(10):3633–3652, 1973.

[13] L. Wolfenstein. Parametrization of the Kobayashi-Maskawa Matrix. *Phys. Rev. Lett.*, 51(21):1945–1947, 1983.

[14] CKM fitter group. http://ckmfitter.in2p3.fr/.

[15] UT fit group. http://www.utfit.org/.

[16] H. Albrecht et al. Observation of B0 - anti-B0 Mixing. *Phys. Lett. B*, 192:245–266, 1987.

[17] A. Einstein. Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig? *Annalen der Physik*, 323(13):639–641, 1905.

[18] LEP, Large Electron-Positron Collider. http://public.web.cern.ch/public/en/Research/LEP-en.html.

[19] Tevatron. http://www-bdnew.fnal.gov/tevatron/.

[20] CDF, Collider Detector at Fermilab. http://www-cdf.fnal.gov/.

[21] HERA, Hadron Elektron Ring Anlage. http://adweb.desy.de/mpy/hera/.

[22] S. Kurokawa and E. Kikutani. Overview of the KEKB Accelerators. *Nucl. Inst. Meth. A*, 499(1):1–7, 2003.

[23] K. Akai et al. Commissioning of KEKB. *Nucl. Inst. Meth. A*, 499(1):192–227, 2003.

[24] KEK Report. KEK B Machine Parameters, June 2009. http://www-acc.kek.jp/kekb/Commissioning/Machineparam2009Jun17.pdf.

[25] A. Abashian et al. The Belle Detector. *Nucl. Inst. Meth. A*, 479(1):117–232, 2002.

[26] Y. Ushiroda. Belle silicon vertex detectors. *Nucl. Inst. Meth. A*, 511(1):6–10, 2003.

[27] BaBar. http://www.slac.stanford.edu/BF/.

[28] B. Aubert et al. Observation of CP violation in the B^0 meson system. *Phys. Rev. Lett.*, 87(9):091801, 2001 arXiv:hep-ex/0107013v1.

[29] M. Feindt. A Neural Bayesian Estimator for Conditional Probability Densities, 2004 arXiv:physics/0402093v1.

[30] <phi-t>® – Physics Information Technologies. http://www.phi-t.de.

[31] T. Bayes. An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

[32] H.Kakuno et al. Neutral B Flavor Tagging for the Measurement of Mixing-induced CP Violation at Belle. *Nucl. Inst. Meth. A*, 533(1):516–531, 2004 arXiv:hep-ex/0403022.

[33] David J. Lange. The EvtGen particle decay simulation package. *Nucl. Inst. Meth. A*, 462(1):152–155, 2001.

[34] R. Brun et al. GEANT3. *CERN Report*, 1987. CERN-DD/EE/84-1.

[35] P. Golonka and Z. Was. PHOTOS Monte Carlo: A Precision tool for QED corrections in Z and W decays. *Eur. Phys. J. C*, 45:97–107, 2006 arXiv:hep-ph/0506026.

[36] S. Kopp et al. Dalitz analysis of the decay $D^0 \to K^-\pi^+\pi^0$. *Phys. Rev. D*, 63:092001, 2001 arXiv:hep-ex/0011065.

[37] T. Higuchi. Vertex Reconstruction for ICHEP06. *Belle Note*, 924, 2009.

[38] G. Fox and S. Wolfram. Event Shapes in e+ e- Annihilation. *Nucl. Phys. B*, 149:413, 1979.

[39] K.Hara et al. Measurement of the $B^0 - \overline{B}^0$ Mixing Parameter Δm_d using Semileptonic B^0 Decays. *Phys. Rev. Lett.*, 89(25), 2002 arXiv:hep-ex/0207045.

Danksagung

Für die Möglichkeit meine Diplomarbeit am Institut für Experimentelle Kernphysik anzufertigen, die hervorragende Betreuung und sehr lehrreiche Diskussionen bedanke ich mich bei meinem Referent Prof. Dr. Michael Feindt.

Für die Übernahme des Korreferats dieser Diplomarbeit bedanke ich mich bei Prof. Dr. Thomas Müller.

Für die hervorragende Betreuung, die Möglichkeit jederzeit Fragen zu stellen und fundierte Antworten zu erhalten, sowie das Korrekturlesen dieser Arbeit bedanke ich mich bei Dr. Michal Kreps, Dr. Thomas Kuhr und Dr. Anže Zupanc.

Für zusätzliches Korrekturlesen bedanke ich mich bei Tobias Volkenandt und meiner Tante Siglinde sowie meinem Onkel Mike Krause.

Für interessante Diskussionen, stete Hilfsbereitschaft und eine tolle Arbeitsatmosphäre bedanke ich mich bei allen Mitgliedern der B-Physik Gruppe des Instituts für Experimentelle Kernphysik. Insbesondere bei meinen Zimmerkollegen Dr. Claudia Marino, Daniel Zander und Fabian Keller.

Für Verständnis und Unterstützung über die Dauer des gesamten Studiums bedanke ich mich bei allen meinen Freunden.

Für stete Unterstützung in allen erdenklichen Situationen bedanke ich mich von ganzem Herzen bei meinen Eltern Christine und Thomas Prim.

Da steh ich nun, ich armer Tor!
Und bin so klug als wie zuvor...

`Goethe, Faust I`