TECHNISCHE UNIVERSITÄT WIEN

DIPLOMARBEIT

Towards a Precision Measurement of Angular Observables in B decays at Belle II

ausgeführt an der
Fakultät für Physik der Technischen Universität Wien

in Zusammenarbeit mit dem
Institut für Hochenergiephysik der Österreichischen Akademie der Wissenschaften

Zur Erlangung des akademischen Grades
Diplom-Ingenieur

im Rahmen des Studiums
Technische Physik

eingereicht von
Matthias MAURER, BSc.
Matrikelnummer: 01426473

Betreuung: Privatdoz. Dipl.-Ing. Dr.techn. Christoph SCHWANDA
Mitbetreuung: Dr. Gianluca INGUGLIA

Ort, Datum

Unterschrift des Betreuers

Abstract

The flavour changing neutral current $b \to s\ell^+\ell^-$ is considered to be one of the most promising particle transitions to search for new physics at the current point of time. It cannot occur directly in the Standard Model, but only through higher order processes, and is therefore highly suppressed. This suppression of the Standard Model amplitude makes the influences on angular distribution and branching ratio predicted by various new physics models easier to measure. Previous experiments, most notably Belle I in 2016 and LHCb with its most recent update in 2020, have reported hints of lepton flavour universality violation at a level of $2.6\,\sigma$ and $2.5\,\sigma$ respectively for the angular observable $P_5^{'}$ in the decay $B \to K^*\ell^+\ell^-$. The thesis at hand is part of Belle II's undertaking of improving these results and thereby either verifying or excluding these tensions.

As the full angular analysis necessary for a $P_5^{'}$ measurement is out of scope for achieving within a single Master thesis, the present work's aim is to lay the necessary foundation towards doing so in future projects. Since the overarching analysis is not complete, no real data has been used in order to avoid the introduction of biases, thereby relying entirely on simulated Monte Carlo data. Nevertheless, the analysis is entirely compatible with real data as well. After the selection and reconstruction of signal candidates from raw data, an extensive analysis of higher order correlations was conducted to avoid biases in later parts of the angular analysis. Background was suppressed using gradient boosted decision trees, and the branching ratios were calculated using signal yields extracted from $1D$ fits on the beam constrained mass. Last but not least, the stability of the fitter against statistical fluctuations was tested using 10 000 poisson distributed toys.

Unfortunately, major delays in data taking have made it very difficult to improve the precision achieved by previous experiments within the foreseeable future, with only $\sim 350\,\mathrm{fb}^{-1}$ being available for now in comparison to the $711\,\mathrm{fb}^{-1}$ available to Belle I. While the predicted signal yields and purity for the B^+ decays are competitive with the Belle I results, this is nowhere near the case for the B^0. As a result, it seems unlikely that an improvement in $P_5^{'}$ precision can be achieved with the current amount of data available. An interesting option to circumvent the current statistical limits might be to use both Belle I and Belle II data sets simultaneously for a combined measurement.

Contents

1. Introduction

The Standard Model (SM) of elementary particle physics describes the world at the smallest known scale and is, together with the theory of gravity, the foundation of the modern understanding of nature. Today's formulation of the SM was finalized in the 1970s, and merges the theories of the the electroweak force (the unification of weak and electromagnetic force) and the strong force, into one common framework. It continues to be the best available model for elementary particle physics to date, supported by the great successes of its experimental predictions. Nevertheless, it is well agreed in science that the Standard Model cannot be the end of the story. It does not capture gravity at all, and cannot be made compatible with General Relativity, which is the corresponding accepted theory at the largest scales. Neither theory is able to explain dark matter, nor can they explain the big mismatch of matter and anti-matter in the universe. Furthermore, the SM predicts neutrinos to be mass-less, in contrast to experimental evidence. Many alternative theories exist and attempt to fix these shortcomings, but no convincing proof for any of these theories was found as of yet. In addition to the shortcomings just mentioned, which science has been aware of for decades, some more recent high-precision experiments reported new tensions with the Standard Model, such as in the muon g-factor or in flavour changing neutral currents. Confirmation that one of these tensions is indeed new physics rather than the result of statistical fluctuation or errors in the respective analysis would be a huge step towards finding new physics. Flavour changing neutral currents, the type of decay this thesis is concerned with, are particularly well suited to the search for new physics. In the SM they do not occur directly, but only through higher order processes, leading to a strong suppression of the decay. Exactly this strong suppression of the SM amplitude makes contributions of hypothetical new physics, be it due to new particles or any other effect, easier to detect and measure in high-precision experiments, should they exist in the first place.

The thesis at hand is dedicated to laying the groundwork for an angular measurement of the so-called P_5' anomaly in the $B \to s\ell^+\ell^-$ transition in the $B \to K^*\ell^+\ell^-$ channel, where hints of possible Lepton Flavour Universality (LFU) violation between electrons and muons were reported in multiple previous analysis. For example Wehle et al. at Belle I in 2016 or LHCb in 2020 measured discrepancies to the Standard Model in said P_5' observable with a statistical significance of $2.6\,\sigma$ and $2.5\,\sigma$

respectively (see chapter 2.4 for more information) [1][2]. While these tensions might be promising, their statistical significance is not good enough to be sure that they are not the result of simple fluctuations. The current analysis pursued at the Belle II experiment, which this thesis is part of, takes most of its inspiration from Wehle et al.'s work at the predecessor experiment Belle I. Using more advanced technology with respect to Belle I, both on hardware and software side, the goal is to achieve a higher accuracy than past results and thereby either verify or exclude the P_5' anomaly, despite having less data currently available at this point of time than Belle I had for their study. While the full angular analysis necessary to measure P_5' is out of scope for a master thesis, it shall pave the way towards doing so. The current stage of the analysis presented in this thesis deals with Monte Carlo (MC) simulation data only, not only because not sufficient data has been produced by Belle II as of yet, but also because unblinding real data before the analysis frame has been fully developed and finished could lead to unwanted information and therefore biases leaking in. Of course, the analysis is developed in such a way that it can be applied on real data as well.

The thesis itself will start with the theoretical background of the physics involved, beginning with a brief overview of the Standard Model, followed by a detailed description of $b \to s\ell^+\ell^-$ and observables relevant for the angular analysis. In the next part, the Belle II experiment will be described, with details on all subdetectors. The section after that gives an introduction to machine learning, presenting selected topics leading up to gradient boosted decision trees for binary classification, which is what is used in this thesis. The last chapter before the analysis itself deals with dependencies in data sets and both linear and non-linear correlations. The practical part starts off with the MC data sets used, followed by a description of how the decays chains are reconstructed and individual particle candidates are chosen. Then, all variables of relevance are shown, including an analysis of non-linear correlations, before the preselection cuts applied are motivated. The next section deals with background suppression using machine learning to filter the very rare decays of interest out of the large background, doing so in such a way that unwanted higher-order correlations that could interfere with the angular analysis are kept at bay. After background suppression, the shape of signal and background components are modeled using a $1D$ fit on the beam constrained mass, thereby also preparing the higher dimensional fits that will be part of the angular analysis. Signal yields are

then extracted from the $1D$ models and used to further further calculate the BRs. Last but not least, toy data samples are generated to test the stability of the fitter against statistical fluctuations.

Unless noted otherwise, the following conventions will be used in this work:

- Natural units ($\hbar = c = 1$).
- The charge conjugated case is always implied for all particles and decays.

2. Theoretical background

2.1. The Standard Model of particle physics

The Standard Model is a relativistic Quantum Field Theory (QFT), where all elementary particles are described as excitations of the corresponding quantum field. Not only quarks and leptons, characterized as Spin-1/2 fermions, are viewed as particles in this framework, but also the interactions between particles are mediated by particles themselves. Interaction particles, also called gauge bosons, are described as Spin-1 vector bosons. The massless photon is the interaction particle of the electromagnetic force, the eight massless gluons the interaction particles of the strong force and the massive $W^{\pm}$ and Z^0 the interaction particles of the weak force. The elementary fermions of the SM consist of six quarks and six leptons, which are both split up into three generations of pairs. These generations differ by mass only and share identical physical properties otherwise. Within the pairs, the partners differ by mass well, and exactly one unit of charge. The final particle of the Standard Model is the Higgs boson, a spin-0 scalar. It is the particle mediating interactions with the Higgs field, leading to a mechanism that gives other particles mass. In fig. 2.1, an overview of all Standard Model particles can be found.

Mathematically, the Standard Model can be described as a quantum gauge theory with three gauge groups

$$\mathrm{SU(3)_C} \times \mathrm{SU(2)_L} \times \mathrm{U(1)_Y}. \tag{2.1}$$

$\mathrm{SU(3)_C}$ is the gauge symmetry group corresponding to quantum chromodynamics (QCD), the theory of the strong force, where the subscript C indicates its color charge. As the name strong force already suggests, it is by many orders of magnitude the strongest of the three fundamental interactions known in nature (four when including gravity, which is technically not a force, but the curvature of space-time). The only particles carrying color charge are quarks and the interaction's mediator particles, the gluons. Since the gluons carry color charge too, they interact with each other as well, rather than just mediating between quarks. As a result, the field of the strong force does not diminish when separating colored particles like an electric field between electrically charged particles would, and instead elongates to a narrow flux tube and stays constant. At some point of separation, enough energy has been

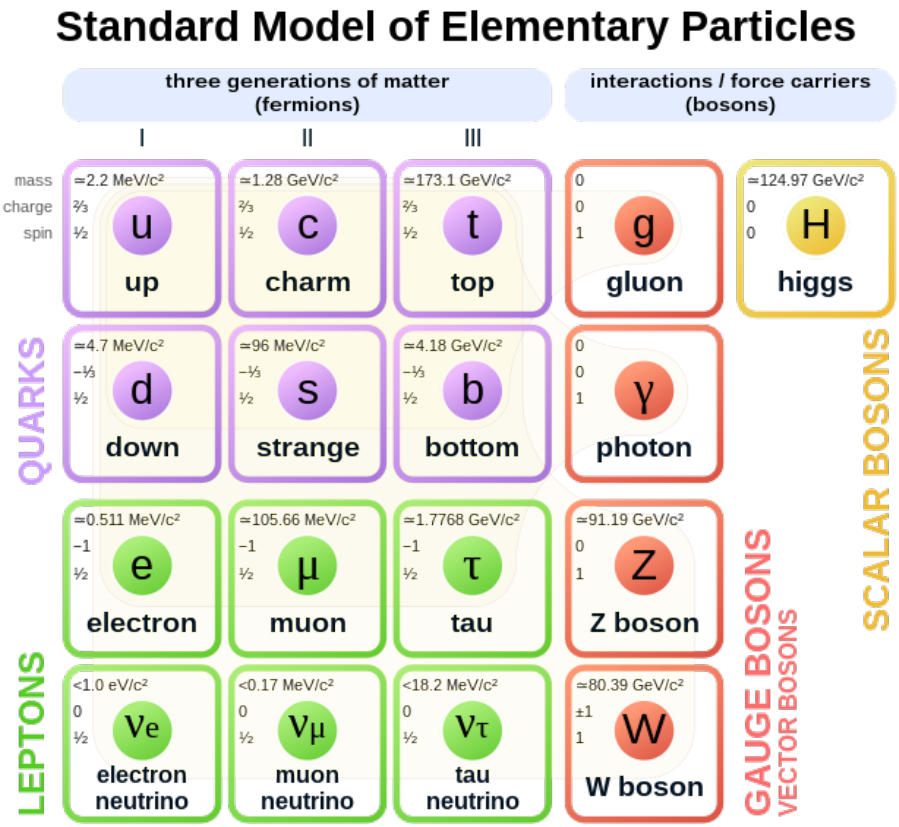

Figure 2.1: The 17 elementary particles in the SM.

put in for a new quark-antiquark pair to spontaneously appear, thereby forming two new composite colorless particles. Composite quarks are called hadrons, with the most important ones being mesons, consisting of quark-antiquark combinations (integer spin), and baryons/anti-baryons consisting of three quarks/anti-quarks (half numbered spin). Observations of tetraquarks (two quarks and two anti-quarks) and pentaquarks (three quarks and two anti-quarks) have confirmed as well in recent years.

$SU(2)_L \times U(1)_Y$ is the unified electroweak interaction described by the Glashow-Weinberg-Salam model. This interaction consists of $SU(2)_L$, the gauge symmetry group of the left-handed isospin, where L indicates the group's left chirality, and $U(1)_Y$, the gauge symmetry group of the weak hypercharge Y. The Glashow-Weinberg-Salam model predicts four massless gauge bosons, namely the B^0 boson

(not to be confused with the non-elementary B mesons), and three W bosons W^0, W^1 and W^2. Those four bosons are not observed in nature though, since the symmetry in the corresponding mass matrix breaks spontaneously at temperatures below the order of 246 GeV ($\sim 10^{15}$ K) and off-diagonal terms appear. Diagonalizing this matrix leads to the four particles observed in nature, namely three massive gauge bosons W^+, W^- and Z^0, and the massless photon, which are all mixed states of the four original bosons. This spontaneous symmetry breaking is called the Higgs mechanism. Without this mechanism, the Standard Model would break, because its intrinsic gauge symmetry demands all gauge bosons to be massless. Note that the Higgs mechanism just described is not identical to the Yukawa interaction of fermions with the Higgs field. It is the coupling strength of this Yukawa interaction that determines the fermions masses. All leptons interact with the weak force part of the electroweak interaction. Only electrons, muons and taus carry electric charge though, and are therefore only those leptons interact with the electromagnetic force, in contrast to the neutral neutrinos.

The symmetry broken model then contains 25 free parameters that need to be chosen to match experimental observations and cannot be derived from within the theory. They correspond to the masses, or rather Yukawa couplings, of the 12 fermions, one coupling constant each for the three gauge interactions, the vacuum expectation value of the Higgs field, the mass of the Higgs boson, and, last but not least, four mixing angles each of the Pontecorvo–Maki–Nakagawa–Sakata (PMNS) and Cabibbo-Kobayashi-Maskawa (CKM) matrices describing the mismatch between mass eigenstates and weak interaction eigenstates of neutrinos and quarks respectively. The arbitrariness of choosing that many parameters to match observations rather than deriving them from a higher theoretical principle, plus various patterns emerging within the groups of parameters, make it seem likely that there are new symmetry principles to be found [3].

2.2. Flavour changing neutral currents

As was already mentioned, the transition of $b \to s\ell^+\ell^-$ is mediated by flavour changing neutral currents. Those are forbidden at tree level in the standard model, since the mediators of those currents, Z^0 and γ, are not able to change quark flavours (e.g. $b \to s$ or $s \to d$) themselves, making the inclusion of higher order diagrams

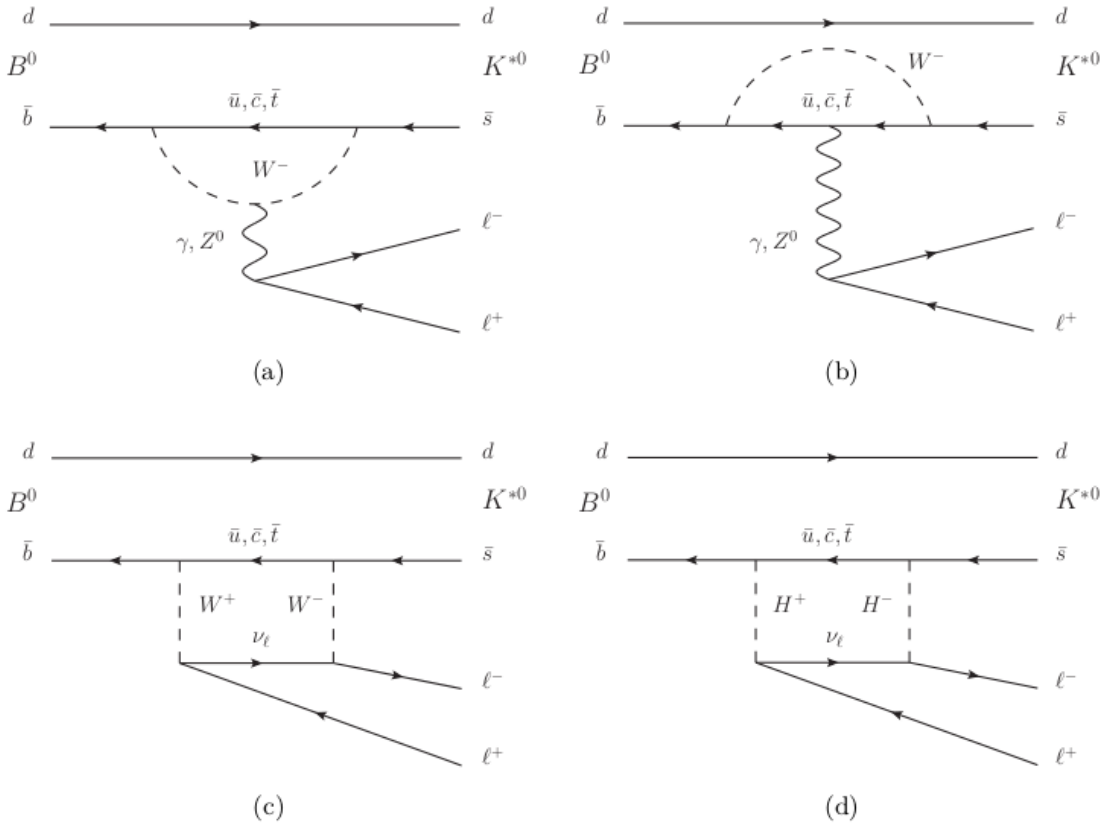

Figure 2.2: Feynman diagrams of the decay $B \to K^*\ell^+\ell^-$. (a) to (c) show the three lowest order feynman diagrams, with (a) and (b) being so-called penguin diagrams and (c) being a gluonic box diagram. (d) is an example of a hypothetical new physics diagram, where a non-SM supersymmetric charged Higgs boson is exchanged, taken from Ref. [1].

necessary. As a result, the decay is strongly suppressed, leading to decay rates of the order $\sim 10^{-6}$ [4]. The total decay rate of the decay $B \to K^*\ell^+\ell^-$ discussed in this thesis is dominated by the three lowest-order Feynman diagrams given in subfigures 2.2 (a) to (c). a) and b) are electroweak penguin diagrams, where a photon or a Z^0 is exchanged, while c) shows a gluonic box diagram. (d) is an example for a hypothetical new physics diagram that might also contribute to the decay rate via exchange of a non-SM supersymmetric charged Higgs boson. To connect the decay rate, comparatively easily accessible by experiment, to theory, one can use Fermi's golden rule in first order perturbation theory, which is given by

$$\Gamma_{i\to f} = \frac{2\pi |M_{i\to f}|^2 \rho}{\hbar}, \tag{2.2}$$

with ρ being the density of states and M the matrix element

$$M = \langle f | H_{eff} | i \rangle . \tag{2.3}$$

$\langle f|$ denotes the energy eigenstates of the particles on the left side of the Feynman diagram, $|i\rangle$ the energy eigenstates of the particles on the right side, while H_{eff} is the effective Hamiltonian. An effective Hamiltonian can used in this context because the flavour changing neutral processes in this thesis takes place at energies much lower than m_W, the mass of the mediator $W^{\pm}$ boson. An operator product expansion can then be used to simplify and integrate out long distance processes. The resulting Hamiltonian for the 10 lowest-level processes is then given by

$$H_{eff} = -\frac{4G_F}{\sqrt{2}} V_{tb} V_{ts}^* \sum_{n=1}^{10} \mathcal{C}_i(\mu) \mathcal{O}_i(\mu), \tag{2.4}$$

where $V_{qq'}$ are the matrix elements of the CKM matrix, $\mathcal{C}_i(\mu)$ coupling constants called Wilson coefficients and $\mathcal{O}_i(\mu)$ local non-perturbative operators. While the $\mathcal{O}_i(\mu)$ represent the long-distance physics, the Wilson coefficients contain the short-distance physics and are calculated theoretically via a perturbative expansion at the scale $\mu = m_{\mathrm{W}}$. Wilson coefficients can also be used to introduce new physics contributions.

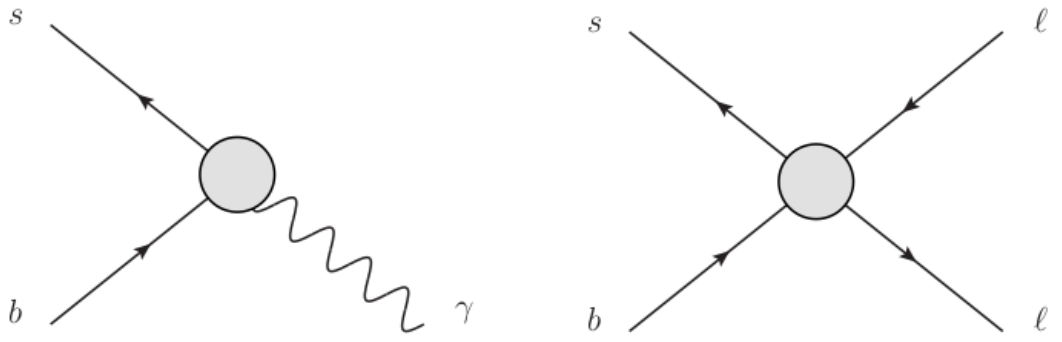

Figure 2.3: Feynman diagrams corresponding to the processes of the operators $\mathcal{O}_7$ (left) and $\mathcal{O}_9$ and $\mathcal{O}_{10}$ (right), taken from Ref. [1].

Out of the 10 operators $\mathcal{O}_i(\mu)$, only the electromagnetic operator $\mathcal{O}_7(\mu)$ (left side of fig. 2.3) and the vector component $\mathcal{O}_9(\mu)$ and axial vector component $\mathcal{O}_{10}(\mu)$ (right side of fig. 2.3) of the electroweak penguin operator contribute on one-loop level order. Those operators are

$$\mathcal{O}_7 = \frac{e}{16\pi^2}\bar{s}_\alpha[\gamma^\mu, \gamma^\nu](m_s L + m_b R)b_\alpha F^{\mu\nu}, \tag{2.5}$$

$$\mathcal{O}_9 = \frac{e^2}{16\pi^2}\bar{s}_\alpha\gamma^\mu L b_\alpha \bar{\ell}\gamma_\mu \ell, \tag{2.6}$$

$$\mathcal{O}_{10} = \frac{e^2}{16\pi^2}\bar{s}_\alpha\gamma^\mu L b_\alpha \bar{\ell}\gamma_\mu\gamma_5 \ell, \tag{2.7}$$

where ℓ, s and b refer to lepton, strange and bottom field, $F_{\mu\nu}$ to the EM field tensor, e to the electromagnetic constant and α to the color index. L and R are the left-handed and right-handed projection operators. Other operators for current-current, QCD penguin and chromomagnetic processes can be neglected [1][5].

2.3. Decay topology and differential decay rate

Hypothetical new physics could not only contribute to the total decay rate, but also to angular distributions via short distance interactions, which are contained in the above-mentioned Wilson coefficients. Since such hypothetical new short distance interactions are suspected in $B \to K^*\ell^+\ell^-$, this section is now going to introduce the decay's topology and the differential decay rate. It is important to note that the K^* is too short-lived to be measured directly and needs to be reconstructed from its decay products K and π. Therefore, four independent kinematic parameters are necessary to fully describe the decay that is actually being measured in this thesis, namely $B \to (K^* \to K\pi)\ell^+\ell^-$. The parametrization chosen in this thesis is the invariant mass of the dileptic system $q^2 = M_{\ell\ell}$, and the three angles θ_ℓ, θ_K and ϕ, all defined in the $B\bar{B}$ rest frame (see fig. 2.4). θ_ℓ is the angle between the flight direction of the dileptic system and the flight direction of the $\ell^+(\ell^-)$, θ_K is the angle between the K^* flight direction and the K flight direction and ϕ the angle between the decay planes of the K^* and the dileptic system.

While the theory discussed so far would be enough for an overview of the physics actually used in this thesis, one needs to go further and derive certain angular observables to describe the motivation of the analysis as a whole and certain decisions

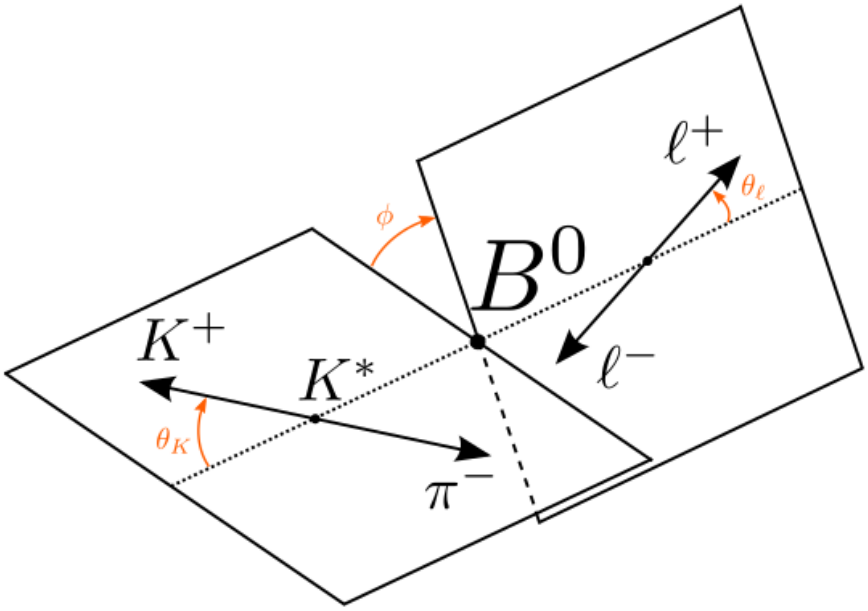

Figure 2.4: The three angles ϕ, θ_ℓ and θ_K in the $B\bar{B}$ rest frame, taken from Ref. [1].

taken. The derivation of the angular observables follows Ref. [6] for the most part, but also uses Ref. [1].

The first step is the definition of the differential decay rate. It can be calculated by squaring the matrix element of the four-body decay, summing over all spins of the final state particles and constraining the kinematics. The full derivation can be found in Ref. [5] and results in the differential decay rate

$$\begin{aligned}\frac{1}{\mathrm{d}\Gamma/\mathrm{d}q^2}\frac{\mathrm{d}^4\Gamma}{\mathrm{d}\cos\theta_K\mathrm{d}\cos\theta_\ell\ \mathrm{d}\phi\ \mathrm{d}q^2} &= \frac{9}{32\pi}\sum_{i=1}^{9} I_i f_i(\cos\theta_\ell,\cos\theta_K,\phi)\\ &= \frac{9}{32\pi}(I_1^{(s)}\sin^2\theta_K + I_1^{(c)}\cos^2\theta_K\\ &+ (I_2^{(s)}\sin^2\theta_K + I_2^{(c)}\cos^2\theta_K)\cos 2\theta_\ell\\ &+ I_3\sin^2\theta_K\sin^2\theta_\ell\cos 2\phi\\ &+ I_4\sin 2\theta_K\sin 2\theta_\ell\cos\phi\\ &+ I_5\sin 2\theta_K\sin\theta_\ell\cos\phi\\ &+ I_6\sin^2\theta_K\cos\theta_\ell\\ &+ I_7\sin 2\theta_K\sin\theta_\ell\sin\phi\\ &+ I_8\sin 2\theta_K\sin 2\theta_\ell\sin\phi\\ &+ I_9\sin^2\theta_K\sin^2\theta_\ell\sin 2\phi),\end{aligned} \tag{2.8}$$

where the eleven angular coefficients $I_i^{(a)}$ are bilinear combinations of six complex decay amplitudes $\mathcal{A}_{i=0,\|,\perp}^{j=L,R}$. The suffixes s and c on some of the I_i are a convention indicating the fraction of the I_i corresponding to the harmonics $\sin^2\theta_K$ and $\sin^2\theta_K$ respectively. The I_i are functions of q^2 only and correspond to the transversity states of the K^{*0} and the chiralities of the dimuon system.

This formula by itself is not valid for the CP conjugated decay $\bar{B}\to\bar{K}^*\ell^+\ell^-$ though, which is why one defines a CP averaged differential decay rate for the combined measurement of B and $\bar{B}$

$$\frac{\mathrm{d}^4(\Gamma+\bar{\Gamma})}{\mathrm{d}\cos\theta_K\mathrm{d}\cos\theta_\ell\ \mathrm{d}\phi\ \mathrm{d}q^2} = \frac{9}{32\pi}\sum_{i=1}^{9}(I_i+\bar{I}_i)f_i(\cos\theta_\ell,\cos\theta_K,\phi), \tag{2.9}$$

where $\bar{I}_i^{(a)}$ is identical to $I_i^{(a)}$, but with all weak phases conjugated and the sign flipped for $i=5,6,8,9$. While in principle all information is already contained and accessible in the $I_i^{(a)}$ and $\bar{I}_i^{(a)}$, it is desirable to reduce the number of free parameters, since the expected signal yields are too little to accomplish a high-precision eleven dimensional fit. To do so, one assumes the massless limit $m_\ell\to 0$, which holds with very high precision for $q^2\gtrsim 1\,\mathrm{GeV}^2/c^4$, but still works reasonably well at lower q^2 values, resulting in a simplification of the $I_i^{(a)}$, $\bar{I}_i^{(a)}$ and f_i (see table 2.1). Using the new symmetries thereby arising, we define new CP symmetric observables

$$S_i = \frac{I_i + \bar{I}_i}{\frac{\mathrm{d}\Gamma}{\mathrm{d}q^2} + \frac{\mathrm{d}\bar{\Gamma}}{\mathrm{d}q^2}} \tag{2.10}$$

and *CP* asymmetric observables

$$A_i = \frac{I_i - \bar{I}_i}{\frac{\mathrm{d}\Gamma}{\mathrm{d}q^2}}. \tag{2.11}$$

Those variables are not fully independent and the following relationships can be found

$$S_{1s} = 3S_{2s} \tag{2.12}$$

$$S_{1c} = -S_{2c} \tag{2.13}$$

$$1 = \frac{3}{4}(2S_1^s + S_1^c) - \frac{1}{4}(2S_2^s + S_2^c), \tag{2.14}$$

reducing the number of independent observables from eleven to eight. The observable S_1^c, more commonly referred to as F_L, is the longitudinal polarization of the K^{*0} and can be written as

$$S_1^c = F_L = \frac{|\mathcal{A}_0^L|^2 + |\mathcal{A}_0^R|^2}{|\mathcal{A}_0^L|^2 + |\mathcal{A}_0^R|^2 + |\mathcal{A}_{||}^L|^2 + |\mathcal{A}_{||}^R|^2 + |\mathcal{A}_{\perp}^L|^2 + |\mathcal{A}_{\perp}^R|^2}. \tag{2.15}$$

Using these *CP* symmetric and asymmetric variables, various *optimized* observables can be built where most major form-factor uncertainties cancel. The ones relevant in this analysis are the so-called $P_i^{'}$, with

$$P_{i=4,5,7,8}^{'} = \frac{S_i}{\sqrt{F_L(1-F_L)}}. \tag{2.16}$$

To test for LFU violation, one can naturally define

$$Q_i = P_{i,e}^{'} - P_{i,\mu}^{'}, \tag{2.17}$$

where any deviation from zero would be a strong hint for new physics. This exact test is the final goal of the overarching analysis. Theoretical calculations for these observable are quite involved and available for certain bins of q^2 only. For ease of comparison, this analysis uses the same bins as the Belle I analysis, $q_1^2 = [0.1, 4], q_2^2 =$

i	I_i	f_i
$1s$	$\frac{3}{4}[\lvert\mathcal{A}_{\lvert\lvert}^L\rvert^2 + \lvert\mathcal{A}_\perp^L\rvert^2 + \lvert\mathcal{A}_{\lvert\lvert}^R\rvert^2 + \lvert\mathcal{A}_\perp^R\rvert^2\ ]$	$\sin^2\theta_K$
$1c$	$\lvert\mathcal{A}_0^L\rvert^2 + \lvert\mathcal{A}_0^R\rvert^2$	$\cos^2\theta_K$
$2s$	$\frac{1}{4}[\lvert\mathcal{A}_{\lvert\lvert}^L\rvert^2 + \lvert\mathcal{A}_\perp^L\rvert^2 + \lvert\mathcal{A}_{\lvert\lvert}^R\rvert^2 + \lvert\mathcal{A}_\perp^R\rvert^2\ ]$	$\sin^2\theta_K\cos 2\theta_\ell$
$2c$	$-\lvert\mathcal{A}_0^L\rvert^2 - \lvert\mathcal{A}_0^R\rvert^2$	$\cos^2\theta_K\cos 2\theta_\ell$
3	$\frac{1}{2}[-\lvert\mathcal{A}_{\lvert\lvert}^L\rvert^2 + \lvert\mathcal{A}_\perp^L\rvert^2 - \lvert\mathcal{A}_{\lvert\lvert}^R\rvert^2 + \lvert\mathcal{A}_\perp^R\rvert^2]$	$\sin^2\theta_K\sin^2\theta_\ell\cos 2\phi$
4	$\sqrt{\frac{1}{2}}[\mathcal{A}_0^L\mathcal{A}_{\lvert\lvert}^{L*} + \mathcal{A}_0^R\mathcal{A}_{\lvert\lvert}^{R*}]$	$\sin 2\theta_K\sin 2\theta_\ell\cos\phi$
5	$\sqrt{2}\mathrm{Re}(\mathcal{A}_0^L\mathcal{A}_\perp^{L*} - \mathcal{A}_0^R\mathcal{A}_\perp^{R*})$	$\sin 2\theta_K\sin\theta_\ell\cos\phi$
$6s$	$2\mathrm{Re}(\mathcal{A}_{\lvert\lvert}^L\mathcal{A}_\perp^{L*} - \mathcal{A}_{\lvert\lvert}^R\mathcal{A}_\perp^{R*})$	$\sin^2\theta_K\cos\theta_\ell$
7	$\sqrt{2}\mathrm{Im}(\mathcal{A}_0^L\mathcal{A}_{\lvert\lvert}^{L*} - \mathcal{A}_0^R\mathcal{A}_{\lvert\lvert}^{R*})$	$\sin 2\theta_K\sin\theta_\ell\sin\phi$
8	$\sqrt{\frac{1}{2}}\mathrm{Im}[\mathcal{A}_0^L\mathcal{A}_\perp^{L*} + \mathcal{A}_0^R\mathcal{A}_\perp^{R*}]$	$\sin 2\theta_K\sin 2\theta_\ell\sin\phi$
9	$\mathrm{Im}(\mathcal{A}_{\lvert\lvert}^L\mathcal{A}_\perp^{L*} + \mathcal{A}_{\lvert\lvert}^R\mathcal{A}_\perp^{R*})$	$\sin^2\theta_K\sin^2\theta_\ell\sin 2\phi$

Table 2.1: Angular observables I_i in the limit $m_\ell \to 0$ and the corresponding harmonics.

$[4.0, 8.0]$, $q_3^2 = [10.09, 12.9]$ and $q_4^2 = [14.8, 19.0]$ [1]. The bins do not cover the full q^2 range due to Dalitz decays, J/Ψ and $\Psi(2S)$ decays polluting the signal, as will be explained in more detail in chapter 6.7.

2.4. Previous angular analysis of $B \to K^*\ell^+\ell^-$

The first full angular analysis of $B^0 \to K^{*0}\mu^+\mu^-$ was presented by the LHCb experiment in 2013, using $1.0\,\mathrm{fb}^{-1}$ of data from p^+p^- collisions, where the observables A_{FB}, F_L, S_3, S_9, A_9, $A_T^{(2)}$, A_T^R where probed in multiple bins of q^2. Another LHCb analysis on the same data set tested S_4, S_5, S_7, S_8, P_4', P_5', P_6' and P_8' in six bins of q^2 and found agreement with the SM in 23 out of 24 measurements. In the interval $4.30 < q^2 < 8.68\,\mathrm{GeV}^2$, a SM deviation in P_5', was found at a level of $2.5\,\sigma$, which is known as the P_5' anomaly [7]. The measurement was repeated in 2020 using $4.70\,\mathrm{fb}^{-1}$ of LHCb p^+p^- collision data, showing $2.5\,\sigma$ deviation in the $4.0 < q^2 < 6.0$ bin and $2.9\,\sigma$ in $6.0 < q^2 < 8.0$ [2]. The latest $B^+ \to K^{*+}\mu^+\mu^-$ LHCb measurement on $9.0\,\mathrm{fb}^{-1}$ p^+p^- collision data showed similar results [8]. P_5' was also measured in 2016 by Belle II's predecessor Belle I, using $711\,\mathrm{fb}^{-1}$ of e^+e^- collision data, confirming the anomaly with a local significance of $2.6\,\sigma$. This analysis also serves as the main inspiration and threshold for the thesis at hand, since the Belle II experiment is much

more similar to Belle I than LHCb .

3. Experimental setup

The Belle II experiment is a detector at the e^+e^- collider SuperKEKB, located in Tsukuba, Japan. Both detector and collider are direct upgrades over their predecessors, the Belle detector (henceforth referred to as Belle I to avoid confusion) and the KEKB collider. KEKB and SuperKEKB are so-called B-factories, as they are specialized on the production of B mesons. The particle beams are set to a center of mass (COM) energy of $\sqrt{s} = 10.58\,\text{GeV}$, which corresponds to the rest mass of $\Upsilon(4S)$ (see table 3.1 for a list of relevant cross sections). The $\Upsilon(4S)$ is an excited $b\bar{b}$ meson that decays with a probability of 48.6% to a neutral $B^0\bar{B}^0$ pair, and with a probability of 51.4% to a charged B^+B^- pair, hence explaining the term B-factory. Belle I operated from 1999 to 2010, collecting around $1\,\text{ab}^{-1}$ of data, of which $711\,\text{fb}^{-1}$ were collisions at $\Upsilon(4S)$ resonance. Work on Belle II started in 2010, with beam commissioning of SuperKEKB starting in 2016. Its design luminosity is $8 \times 1035\,\text{cm}^{-2}\text{s}^{-2}$, topping KEKB's design luminosity by a factor of 40. The goal is to produce data corresponding to an integrated luminosity of $50\,\text{ab}^{-1}$ at $\Upsilon(4S)$ resonance by the end of its lifetime [9].

process	cross section [nb]
$e^+e^- \to \Upsilon(4S)$	1.11
$e^+e^- \to u\bar{u}(\gamma)$	1.61
$e^+e^- \to d\bar{d}(\gamma)$	0.40
$e^+e^- \to s\bar{s}(\gamma)$	0.38
$e^+e^- \to c\bar{c}(\gamma)$	1.30
$e^+e^- \to \tau^+\tau^-(\gamma)$	0.919

Table 3.1: Cross sections of e^+e^- interactions at Belle II. All other processes can be neglected, as they do not produce four charged tracks and are therefore filtered out during reconstruction.

3.1. SuperKEKB

SuperKEKB is an asymmetric particle accelerator colliding e^+ and e^- at energies of 7 GeV and 4 GeV respectively. Due to their difference in energy, the CMS frame is boosted by a Lorentz factor of 0.28 at the point of collision. Both e^+ and e^- are

produced in a 600 m linear accelerator before continuing into the high energy storage ring (HER) and the low energy storage ring (LER) respectively (see left side of fig. 3.1). The rings have a circumference of 3 km and a maximum capacity of about 1 500 particle bunches, with each bunch containing about 3×10^{10} particles within a length of about 0.5 cm.

Main reasons for the vastly improved luminosity with respect to KEKB are the roughly doubled beam currents in both storage rings (3.6 A for LER and 2.6 A for HER), a larger crossing angle of 4.8° and more strongly squeezed particle bunches, which can be achieved due to the application of superconducting magnets in the interaction region. The thereby resulting collision area has a width of about 0.3 mm and a height of about 100 nm [10][11].

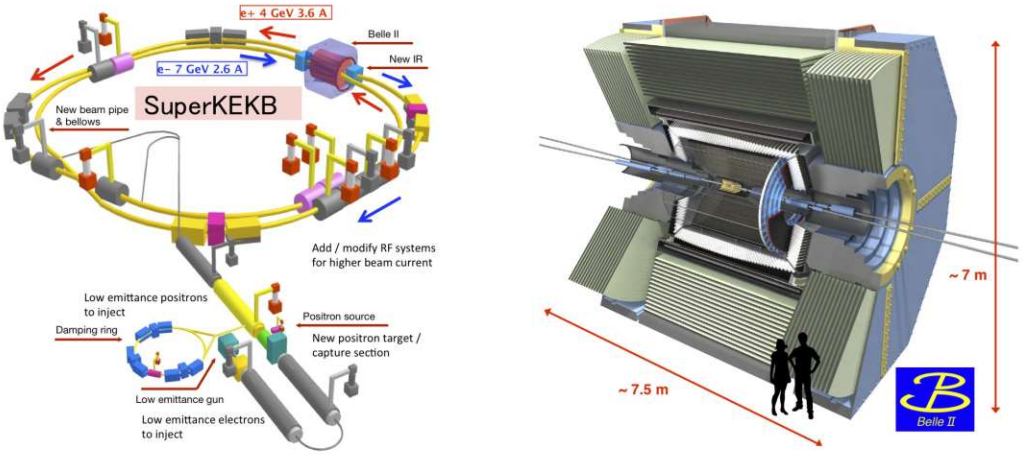

Figure 3.1: The image on the left side shows an overview of the e^+e^- accelerator SuperKEKB, with the detector Belle II being positioned in the blueish box on the top right of the image. The right side figure gives a closer look at Belle II. [12].

3.2. Belle II

The Belle II detector is positioned around the collision area of the e^+ and e^- bunches. The old shell of Belle I, a superconducting solenoid magnet with an iron yoke, is reused, but all other components are either fully new or at least received an upgrade. Major upgrades were necessary to accommodate for the increase in luminosity by a factor of 40 with respect to Belle I, which results in an increase of event rates

by the same amount the detector needs to handle. Belle II consists of several subdetectors (see fig. 3.2), each dedicated to specialized tasks such as charged track reconstruction, vertex detection or particle identification. The next chapters shall give an introduction to the individual components of Belle II. For more detailed information on the experiment, refer to the Belle II physics book [9] and the Belle II technical design report [13].

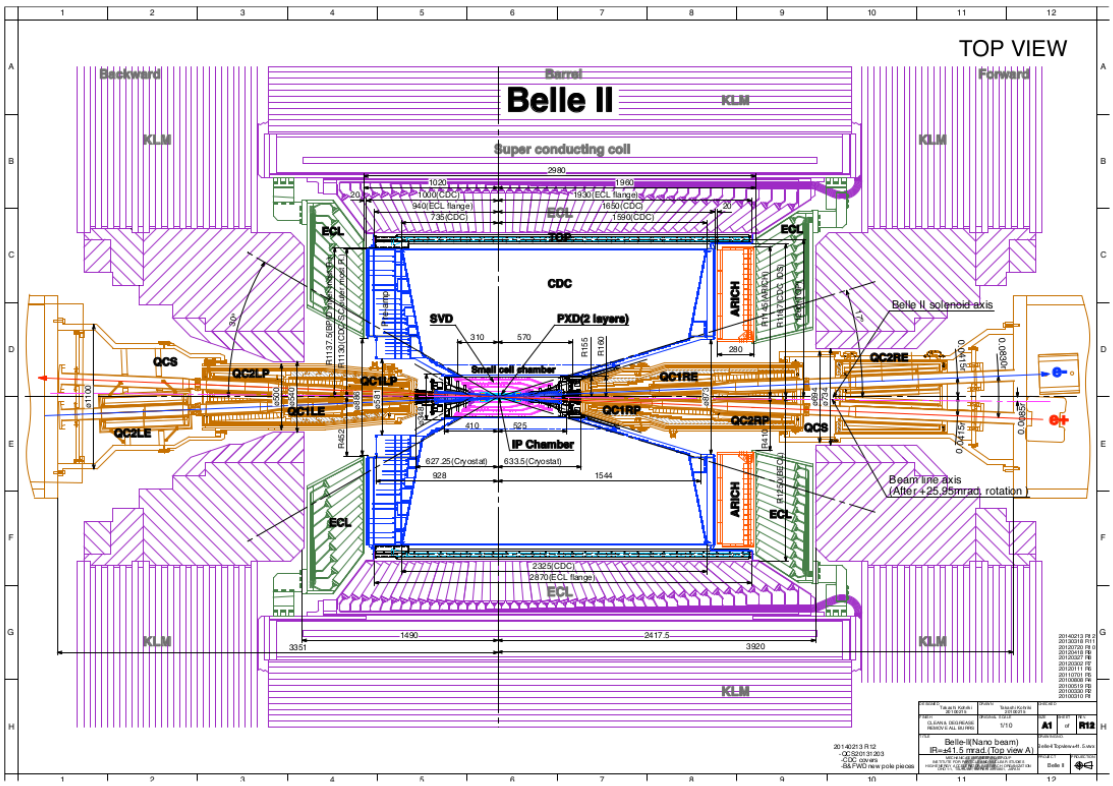

Figure 3.2: An overview showing the various components of the Belle II detector in cross section [9].

3.3. Vertex Detector

The innermost subdetector, positioned right around the beryllium beam pipe of radius $r = 10\,\text{mm}$, is the vertex detector (VXD) (see fig. 3.3), and consists of six concentric layers. Its purpose is to reconstruct decay vertices at a high resolution, aided by extrapolated particle tracks measured in the central drift chamber (CDC,

see below). The two innermost layers ($r = 14\,\text{mm}$ and $r = 22\,\text{mm}$) of the VXD are equipped with roughly 8×10^6 DEPFET (Depleted p-Channel Field Effect Transistor) sensors and are called silicon pixel detector (PXD). The remaining four layers at $r = 39, 80, 104$ and $135\,\text{mm}$ respectively are referred to as silicon vertex detector (SVD) and utilize double-sided silicon strip sensors [14].

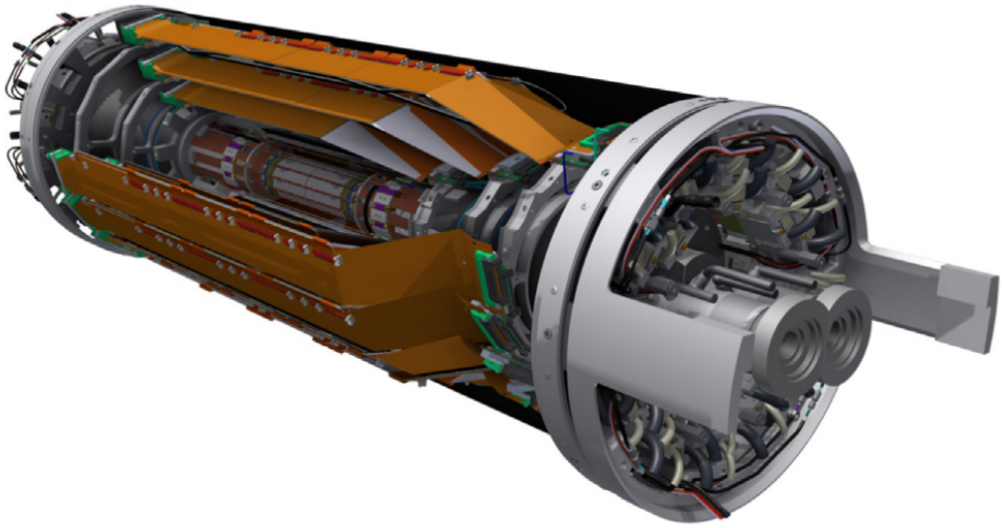

Figure 3.3: The Belle II VXD subdetector, image taken from [15].

3.4. Central Drift Chamber

The second subdetector is the CDC, a large-volume drift chamber between $r = 160\,\text{mm}$ and $r = 1130\,\text{mm}$ consisting of small drift cells. Its main purpose is the reconstruction of tracks, momenta $\vec{p}$ and energy losses $\frac{\mathrm{d}E}{\mathrm{d}x}$ of charged particles. Charged particles passing through ionize the gas in the chamber, which is a 50/50 mixture of He and C_2H_6. The electromagnetic field of the drifting ions is then picked up by sense wires of $30\,\mu\text{m}$ diameter. The in total 14 336 wires are arranged in 56 layers, which in turn are arranged to superlayers consisting of either six or eight adjacent layers. The superlayers' orientation alternates between aligned with the solenoidal magnetic field and skewed to it, see top and bottom of fig. 3.4 respectively. This variation in orientation allows a three-dimensional reconstruction of the particle's track. Further, the measurement of energy losses can be used to assign a particle

identity to the individual tracks. This feature is primarily used for particles that do not have enough energy to reach the Particle Identification Detector (PID), the subject of the next section.

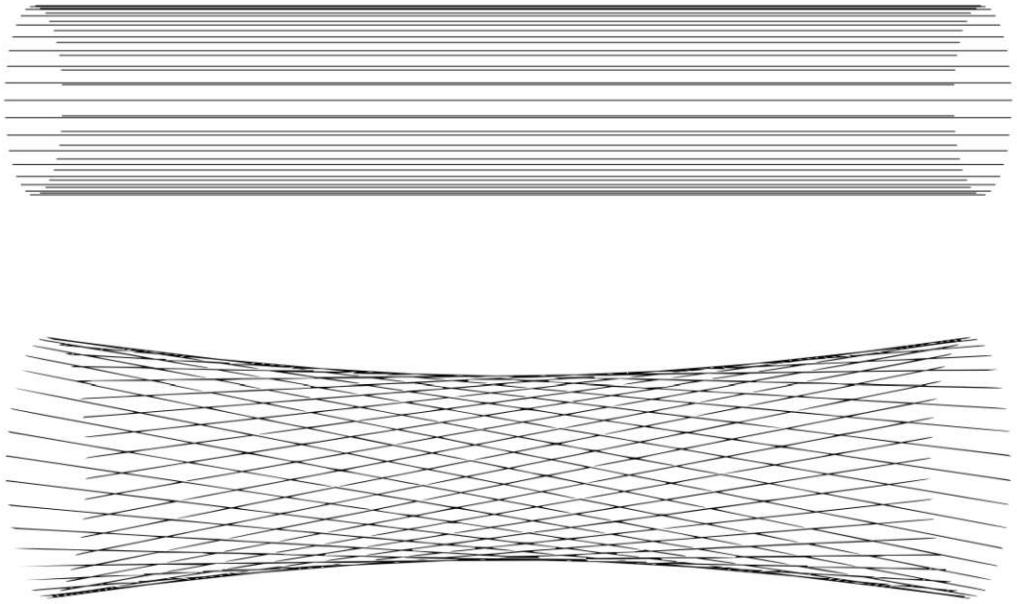

Figure 3.4: The two wire orientations of layers in the CDC. The upper plot shows wires aligned with the solenoidal magnetic field, the bottom plot wires skewed to it. Image taken from [15].

3.5. Particle Identification

3.5.1. Time-Of-Propagation detector

Following the CDC is the PID, which itself consists of two subdetectors. The first one, the Time-Of-Propagation counter (TOP), is built of sixteen modules that are located in the barrel region around the outer wall of the CDC. Each module is assembled from two quartz bars, a mirror and photomultiplier tubes. Charged particles passing through the quartz bars produce Cherenkov photons, which are internally reflected until they are detected by the photomultiplier tubes. The thereby measured position and detection time of photons can be used to calculate particle hypothesis probabilities via a comparison with the expected distributions for e, μ, π, K, p and d. The left side of fig. 3.5 schematically shows the difference in path of a Cherenkov

Photon in one of the modules, depending on whether the passing particle is a K or a π.

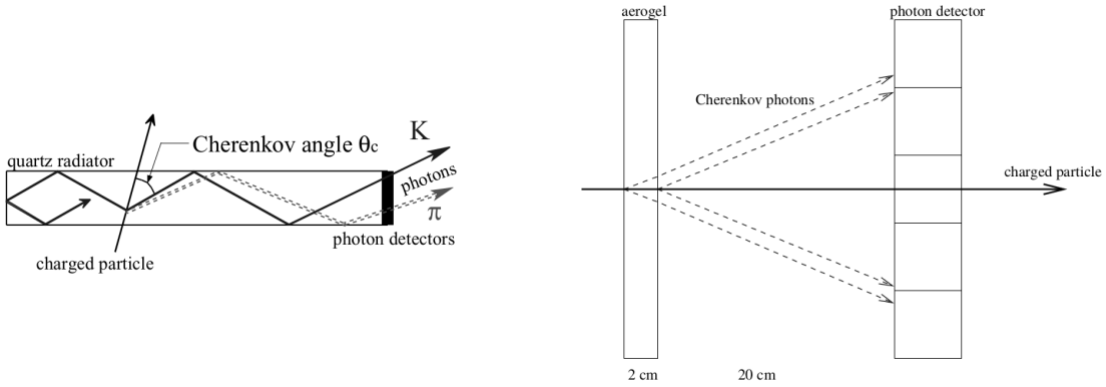

Figure 3.5: The left figure shows a schematic view of one of the quartz bars used in TOP. The Cherenkov photons are internally reflected until they reach the photomultiplier tube. Their path depends on the type of the original particle, here indicated for a K and a π. The right side shows a scheme of ARICH, which measures Cherenkov photons of charged particles passing through an aerogel. Images taken from [13].

3.5.2. Aerogel Ring Imaging Cherenkov detector

The second part of the PID is the Aerogel Ring Imaging Cherenkov detector (ARICH), which is located in the forward endcap region. Its purpose is the differentiation of K and π mesons in the momentum region between 0.4 GeV/c and 4 GeV/c. It is built of aerogel radiators consisting of two 2 cm thick layers with refractive indices of $n_1 = 1.045$ and $n_2 = 1.055$ respectively, and a plane of photon detectors. The Cherenkov photons of a single charged particle passing through the aerogel produce a ring on the photon detector, whose radius can be used to calculate the incidence angle, which in turn allows the differentiation of kaons and pions. The choice of refractive indices results in an increased focus of the Cherenkov photons compared to a single layer. A schematic side view of ARICH is shown on the right side of figure 3.5.

3.6. Electromagnetic Calorimeter

Positioned around the PID is the Electromagnetic Calorimeter (ECL). It is responsible for the detection of photons and measuring their energy and position, as well as separating electrons from hadrons, most importantly pions. The detector is built of a total of 8 736 thallium-doped caesium iodide CsI(Tl) crystals, and covers about 90% of the solid angle in the COM system. Two photo diodes are glued to the rear end of each crystals. Crystals, preamplifiers and support structures are recycled from Belle I, but there are considerations of upgrading them in the future.

3.7. K_L^0 and Muon detector

The last and outermost subdetector is the K_L^0 and Muon detector (KLM). It consists of alternating 4.7 cm thick iron plates and active detector elements surrounding the superconducting solenoid shell of the detector in both barrel, forward endcap and backward endcap region. The iron plates serve both as material for hadronic showers of K_L^0 (3.9 interaction lengths) and also as the magnetic flux return of the solenoid. The detector elements are built of glass-electrode resistive plate chambers (RPC). The high-voltage glass electrodes are separated by gas, which gets ionized by charged particles passing through it, leading to a current that can be measured. The KLM is used in combination with the ECL to detect the K_L^0 hadron showers. Muons are identified by combining extrapolated CDC particle tracks with KLM measurements. Only muons with an energy between ~ 0.7 GeV and ~ 1.5 GeV can be measured, as they are not able to reach the KLM if their energy is too low, or tend to pass through undetected when their energy is too high.

4. Machine learning

The practical part of this thesis makes heavy use of applied machine learning (ML) and statistics, which is why the next two sections shall give an introduction to some key concepts of relevance. The current section will deal with selected topics in ML, while the section after discusses dependencies in data sets, before finally heading on to the analysis itself.

ML, a relatively modern field, is concerned with algorithms that use input data to improve its performance on some kind of task. A generic data set consists of n multidimensional samples. Each dimension of the individual samples is referred to as feature or variable of the data set. This could be any kind of information about the sample, such as its mass, height or color.

There are two basic categories of learning problems that ML deals with. The first category is called supervised learning, where the goal is to predict the value of certain features on new data the algorithm has not seen before. To achieve this, a dedicated training data set, where the true value of these features are known, needs to be fed to the algorithm beforehand, such that it can 'learn' how to make the predictions. When the features the algorithm is supposed to predict are discrete, one speaks of classification. An example for such a discrete feature (or class) could be the color of an object, such as green, blue and red. If there are only two classes to choose from, it is called a binary classification problem. A typical example in high energy physics (HEP), and also the use case in this thesis, is the discrimination between a decay of interest, henceforth referred to as signal, and all other non-signal decays, commonly called background. When the predicted features are continuous rather than discrete, it is a regression problem. This could for example be the prediction of human height as a function of gender, age and weight.

The second overarching category of learning problems is called unsupervised learning. Here, the goal is not to predict the value of certain features, as it is in the case of supervised learning, but rather to find patterns in the data set. Examples for use cases are clustering, where the algorithm attempts to find subgroups within the data that are similar to each other, the compression of data, or density estimation, where distributions are modeled.

The type of problem this thesis deals with, as already mentioned, is of the binary classification type. In HEP, the most popular classification methods nowadays are

so-called deep learning methods and decision tree based methods. As this thesis makes use of the algorithm *XGBoost* (eXtreme Gradient Boosting) [16], which is based on an ensemble of decision trees, the rest of this section shall deal with the fundamentals of decision trees and subsequent topics of relevance for *XGBoost* itself. But before heading on to decision trees, an introduction to the concept of cuts, the foundation of every decision tree, needs to be given first.

4.1. Cut based methods

Using cuts on real variables is the most rudimentary and basic form of classification possible. Applying a cut refers to the introduction of a restriction on a specific input variable x_j^a of the data set $Q = \{x_j^i, y_j\}_{j=(1,n)}^{i=(1,l)}$, where the $x_j^i \in \mathbb{R}_n^l$ represent the n samples with l training features, and $y_j \in \mathbb{R}_n$ the label feature of the n samples that the algorithm shall learn to predict, which could in principle itself be a vector of dimension m. In the case of binary classification, such a condition takes the form $x_b^a > c \forall\, b \in (1, n), c \in \mathbb{R}$. Depending on whether a sample b fulfills the condition, $y_{b,\text{ predicted}}$ is then assigned to either the one class or the other. The overall goal of such cuts is the improvement of the overlap between $y_{j,\text{ predicted}}$ and $y_{j,\text{ true}}$. In the case of this thesis, the individual data samples are candidates for a $B \to K^* \ell^+ \ell^-$ decay, following an e^+e^- collision event. If the condition is fulfilled, the candidate is labeled as a signal decay, otherwise it is rejected as background. Whilst very simple, it is also this intrinsic simplicity that makes cuts very useful even on their own. For example, in this analysis cuts are applied by 'hand' before the machine learning section in order to reduce the vast data size to a manageable amount, by cutting such that only the physical regions of interest remain. Furthermore, they are also used to clean up the data set from various problematic candidate decays, e.g. in physical regions where the detector does not perform very well (more information in chapter 6.6). The effect of individual or groups of cuts can also be quantified, for example with a metric such as the figure of merit

$$\text{FOM} = \frac{N_{\text{signal}}}{\sqrt{N_{\text{signal}} + N_{\text{background}}}}, \tag{4.1}$$

which also plays an important role in general in this analysis. The subscript 'signal' refers to true signal that is correctly labeled, and the subscript 'background' refers

to background that is mislabeled as signal. Naturally, information regarding the true identity of an event is available for simulation data only, and unknown for real data. Whilst the effect of the same sequence of cuts on the same data must always be the same independent of the order of application due to their inherent linearity, the order becomes nevertheless important when evaluating the effect of individual cuts. For example, a situation could arise where both cut 1 on variable A and cut 2 on variable B appear to significantly improve the FOM, but when swapping their order it turns out that cut 1 has no effect after all, since the same result could also be achieved with cut 2 alone. As every cut applied in this analysis needs to be justified and also accounted for in the systematic uncertainties later on, one needs to take this into account when developing the preselection process.

4.2. Decision Trees

Tree-based machine learning algorithms by definition always utilize one or more decision trees. A decision tree can be visualized as a flowchart-like structure, with a single node at its very top. This node, numbered as node 0, represents the full data set $Q_0 = \{x_j^i, y_j\}_{j=(1,n_0)}^{i=(1,l)}$ with n_0 samples. At each internal node k, the set $Q_k = \{x_{j,k}^i, y_{j,k}\}$ with its n_k samples is split into two subsets (or branches)

$$Q_k^{\text{left}}(\theta) = \{x_{j,k}^i, y_{j,k}|\theta\} \tag{4.2}$$

and

$$Q_k^{\text{right}}(\theta) = Q_k \backslash Q_k^{\text{left}}(\theta), \tag{4.3}$$

depending on the outcome of the condition $\theta = x_{b,k}^a > c_k \, \forall \, b \in (1, n_k)$. The external nodes (or leaves), where the tree terminates, are the outputs or prediction vectors $y_{j,\,\text{pred}}$. In its most basic form, with only a single root node with two branches and two corresponding leaves, the structure simplifies to a simple basic cut. Such a tree is also referred to as a tree of depth one. Trees can in principle reach an arbitrary depth, recursively splitting up the data set into smaller and smaller derived subsets at every step. One can not only associate an output value or class to an individual leaf, but also a probability of the association being correct, depending on the chosen algorithm. In order to build a decision tree, it needs to be assessed at every node k whether a candidate split improves the quality of the tree using the impurity

$$G(Q_k, \theta) = \frac{n_k^{\text{left}}}{n_k} H(Q_k^{\text{left}}(\theta)) + \frac{n_k^{\text{right}}}{n_k} H(Q_k^{\text{right}}(\theta)), \tag{4.4}$$

with $H(Q)$ being an arbitrary loss function. Common loss functions for classification are for example the Gini impurity

$$H(Q)_{\text{Gini}} = \sum_{r=0}^{R-1} p_k(1 - p_r) \tag{4.5}$$

and the logloss (also called entropy)

$$H(Q)_{\text{logloss}} = -\sum_{r=0}^{R-1} p_r \log p_r, \tag{4.6}$$

where R is the number of different classes, and p_r is the probability of all samples in the set Q belonging to the class r. To find the best candidate split θ^*, the impurity is minimized such that

$$\theta^* = \arg\min_{\theta}(G(Q_k, \theta)), \tag{4.7}$$

and if the thereby achieved impurity $G(Q_k, \theta^*)$ does not satisfy a certain threshold, the node k terminates and no new branches are built. The same procedure is recursively applied on the resulting subsets $Q_k^{\text{right}}(\theta^*) \equiv Q_{k+1}$ and $Q_k^{\text{left}}(\theta^*) \equiv Q_{k+2}$, until the the predefined maximum depth m_{max} or minimum number of samples n_{min} for every remaining node is reached and all nodes are terminated.

4.3. Random Forests

While it comes at no surprise that a decision tree is a direct improvement over simple cuts, a single decision tree on its own would need to grow very deep in order to achieve a high accuracy in its predictions for a generic use case. Unfortunately, too deep and complex trees usually tend to fit the the data it was trained on too well, as it starts to include stochastic fluctuations in the model as well. While this might lead to very impressive results on the training set, the model performs worse on data it has not seen before than it would otherwise, a phenomenon that occurs very often in ML and is referred to as *overfitting*.

This problem can be tackled by using an ensemble of shallow trees instead of a single deep one. Thereby, many weak learners (low correlation of the output with true classes) are combined into a single strong one (high correlation of the output with true classes). Another method to reduce variance is the introduction of various elements of randomness to the trees. In the case of an algorithm called *Random Forests*, a new subsample of the original data set for new tree is created by repeatedly selecting random samples with replacement (a method called bagging), and by also choosing a random subset of the feature space at each node. After a predefined number of trees has been built, the output for a single data sample is either the average over all tree outputs in the case of regression, or the majority label in the case of classification. More information on the topic of overfitting will be presented in chapter 4.5). While Random Forests' approach can improve the overfitting problem a single decision tree has, it comes at the cost of interpretability (due to the generally large number of trees), and can also lead to a minor increase in bias. Additionally, it is outperformed by a newer class of algorithms that apply the concept of gradient boosting.

4.4. Gradient boosting

The subsequent introduction to gradient boosting directly follows the approach by [17]. Just like Random Forests, gradient boosting algorithms combine many weak learners h, usually decision trees, into a single strong one, the summed model F. In contrast to Random Forests though, where the individual weak learners were built independent of each other, each new learner h_{m+1} in a gradient boosting algorithm is built in such a way that it compensates for the shortcomings of the combined existing model $F_m \equiv y_{i,\text{ prediction } m}$, such that

$$F_m(Q) + h_{m+1}(Q) \stackrel{!}{=} y_{i,\text{ true}}. \tag{4.8}$$

This directly yields the residual $y_{i,\text{ true}} - F_m(Q)$, which is the function the new learner h_{m+1} attempts to fit. The way the gradient boosting learners are built expands on the idea of Gradient Descent, where gradients are used to find the local minimum of a differentiable multi-variable function $G(a_i)$, with a_i being an arbitrary vector. Since the direction of fastest descent is $a_i - \nabla_i G(a_j)$, one can iteratively search for a local minimum using

$$a_{m+1,i} = a_{m,i} - \Delta \nabla_i G(a_{m,j}), \tag{4.9}$$

where $m \in \mathbb{N}_0$ indicates the m^{th} iteration step and $\Delta \in \mathbb{R}_+$ the step size. For our case, the vector $a_{m,i}$ translates to the summed model $F_m(Q) \equiv y_{i,\text{prediction } m}$ at iteration step m, and G to a differentiable loss function $H(Q, F_{m-1}(Q))$, resulting in the iterative formula

$$F_m(Q) = F_{m-1}(Q) - \Delta \nabla_{F_{m-1}} H(Q, F_{m-1}(Q)). \tag{4.10}$$

The iterative procedure can be further optimized by adjusting the step size at each iteration step $\Delta \to \Delta_m$ such that the loss function is minimized, leading to the minimization problem

$$\Delta_m = \arg\min_{\Delta} \sum_{i=1}^{n} H(y_i, F_m(Q)) \tag{4.11}$$

$$= \arg\min_{\Delta} \sum_{i=1}^{n} H(y_i, F_{m-1}(Q) - \Delta \nabla_{F_{m-1}} H(y_i, F_{m-1}(Q)) \tag{4.12}$$

In pseudocode, a generic gradient boosting algorithm for regression could look like the following:

Algorithm 1 Generic example for regression in gradient boosting

Input: training data $Q = \{x_i^j, y_i\}_{i=(1,n)}^{j=(1,l)}$
Input: maximum number of iterations M
Input: initial values $F_0(Q)$, e.g. a constant c_i
1: **for** $m = 1, M$ **do**
2: **for** $i = 1, n$ **do**
3: $g_{i,m} = -\dfrac{dH(Q, F_{m-1}(Q))}{dF_{m-1}(Q)}$
4: fit learner h_m to data set $\{x_i^j, g_{i,m}\}_{i=(1,n)}^{j=(1,l)}$
5: $\Delta_m = \arg\min_{\Delta} \sum_{i=1}^{n} H(Q, F_{m-1}(Q) - \Delta \nabla_{F_{m-1}} H(Q, F_{m-1}(Q)))$
6: $F_m(Q) = F_{m-1}(Q) + \Delta_m h_m$
7: **end for**
8: **end for**
Output: $F_M(Q)$

For a classification algorithm, this needs to be expanded. Every class r is assigned a separate model $F_r(Q)$, $r = 1, 2, 3, ...R$. To achieve a minimal summed loss of all classes, all R models are adapted in parallel, following the procedure described above. The output of the models is then used to calculate identification probabilities

$$P_r(Q) = \frac{e^{F_r(Q)}}{\sum_{j=1}^{R} e^{F_j(Q)}} \tag{4.13}$$

at every data point. The class of the model with the highest probability is predicted.

Example of a pseudocode for a generic classification gradient boost algorithm.:

Algorithm 2 Generic example for classification in gradient boosting

Input: training data $Q = \{x_i^j, y_i\}_{i=(1,n)}^{j=(1,l)}$
Input: maximum number of iterations M
Input: initial values $F_{r,0}(Q)$, e.g. a constant c_i
for $m = 1, M$ **do**
  for $i = 1, n$ **do**
    for $r = 1, R$ **do**
      $g_{i,r,m} = -\frac{dH(Q, F_{r,m-1}(Q))}{dF_{r,m-1}(Q)}$
      fit learner $h_{r,m}$ to data set $\{x_i^j, g_{i,r,m}\}_{i=(1,l)}^{j=(1,n)}$
      $\Delta_m = \arg\min \sum_{i=1}^{n} H(Q, F_{r,m-1}(Q) - \Delta \nabla_{F_{r,m-1}} H(Q, F_{r,m-1}(Q)))$
      $F_{r,m}(Q) \overset{\Delta}{=} F_{r,m-1}(Q) + \Delta_m h_{r,m}$
    end for
  end for
end for
Output: $F_{r,M}(Q)$

XGBoost, the ML-library used in this thesis, is based on gradient boosting as it was presented in this chapter. Unlike the gradient descent in regular gradient boosting though, which includes first derivatives (gradients) of the loss function only, *XGBoost* applies the Newton-Raphson method and therefore employs second derivates (hessians) of the loss function as well. Furthermore, *XGBoost* also makes heavy use of various forms of regularization, which shall be explained in the next chapter.

4.5. Regularization

A common problem in gradient boosting and machine learning in general is the phenomenon of overfitting, which occurs when the model starts to fit statistical fluctuations in the training set. While the accuracy on the training data might still increase while the overfitting advances, the model loses its generalizability and performs worse and worse on independent test data. Many methods of regularization are available in order to avoid this problem, and can also be used in conjunction, depending on the specific needs. The following gives a short overview for the parameters that were especially relevant while working for this thesis.

The possibly most important parameter is the learning rate λ, defined in the range $0 < \lambda \leq 1$, which can be introduced to the iterative formula from equ. 4.10 in the following way:

$$F_m(Q) = F_{m-1}(Q) - \lambda \Delta \nabla_{F_{m-1}} H(Q, F_{m-1}(Q)). \tag{4.14}$$

When $\lambda \equiv 1$, this of course reduces again to the original equ. 4.10. The lower λ is chosen, the more conservative the algorithm becomes. Generally speaking, the overall accuracy of the algorithm tends to increase with smaller λ, with computation costs increasing due to the higher demand of iterations.

Random subsampling of data and/or features can be used to bring introduce more randomness to the algorithm on the hand, thereby reducing overfitting, and to reduce computational cost, which needs to be balanced with losses in performance. In the case of decision tree ensembles one can define these parameters on the level of trees, level of tree and also for each node. γ defines the minimum loss reduction required to make a new a split, and therefore also influences how conservative the algorithm is. Further particularly useful is the concept of early stopping, where the model is evaluated on a test set after every new learner, and the iteration is stopped when the model does not improve for a predefined number of iterations in a row. This completely eliminates any danger of overfitting, but might come at significant computational costs since the model needs to be applied to the test set after every iteration [18].

5. Dependencies in data sets

5.1. Linear correlations

A very common occurrence in data is the interdependence or correlation of two features. In ML, this is in most cases not just good, but a sheer necessity, since such correlations between label feature and training features are what enables the algorithms to 'learn' in the first place. There are other cases though, where correlations can be a problem. Lets consider the following HEP example that is also highly relevant in this thesis: We want to find a model for a particles beam-constrained mass M_{bc} by fitting two different curves on the distribution, one for the signal decays and one for background decays. The signal decays have a sharp gaussian-like peak, while the background is more or less flat. In order to model the background distribution as accurately as possible, we want to fit the model on a M_{bc} range much larger than the actual signal range. When the two distributions are very different from each other, this works perfectly well. Now, if we were to feed the variable we want to fit into our machine learning algorithm, we would run into a problem. In a nutshell, the algorithm would first cut away everything outside of the signal range. Further, it would turn the remaining background distribution into a gaussian-like peak as well, as it learns that decays at the center of the peak are more likely to be signal, resulting in the two sub-distributions becoming fairly similar. While the performance of the algorithm might appear to have increased significantly when just looking at post-algorithm numbers such as the FOM, signal to background ratio etc., the following fit on M_{bc} becomes highly inaccurate, thereby impeding with the actual goal of a high-precision fit and worsening the overall performance.

Simply getting rid of M_{bc} for training might not be enough to solve this problem though. If there are variables in the data set significantly correlated with M_{bc}, the algorithm might 'learn' the distribution of M_{bc} by itself, resulting in the same problem we started with. Therefore, one needs to conduct a correlation study and eliminate all variables that would otherwise interfere.

Often, such correlations can be clearly seen when plotting features against each other, but this is not always the case. Furthermore, the more features a data set contains, the more time consuming it becomes to inspect all plots by hand, resulting in the need of quantitative (and automated) alternatives.

The most common metric used to quantify correlations is Pearson's product-moment coefficient, also known as (linear) correlation coefficient. It can be derived from the covariance

$$\mathrm{cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \tag{5.1}$$

of two random variables X and Y, with E being the expectation operator and

$$\mu_X = \mathrm{E}[X] \tag{5.2}$$

and

$$\mu_Y = \mathrm{E}[Y] \tag{5.3}$$

the averages. Since $\mathrm{cov}(X,Y)$ depends on the magnitude of the random variables, it is hard to interpret it without further modification. By normalizing $\mathrm{cov}(X,Y)$ using the variances

$$\mathrm{var}(X) \equiv \mathrm{cov}(X,X) = \mathrm{E}[(X-\mathrm{E}[X])^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \tag{5.4}$$

and

$$\mathrm{var}(Y) \equiv \mathrm{cov}(Y,Y) = \mathrm{E}[(Y-\mathrm{E}[Y])^2] = \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2, \tag{5.5}$$

one can define the correlation coefficient

$$\rho_{X,Y} \equiv \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var(X)var(Y)}}} \tag{5.6}$$

$$= \frac{\mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]}{\sqrt{\mathrm{E}[X^2]-(\mathrm{E}[X])^2}\sqrt{\mathrm{E}[Y^2]-(\mathrm{E}[Y])^2}}, \tag{5.7}$$

whose magnitude can now be directly interpreted as a measure of correlation. It satisfies the relation $0 \leq \rho_{X,Y} \leq 1$, with 1 implying maximum linear correlation and 0 no linear correlation.

When X and Y are independent, it directly follows that $\rho_{X,Y} = 0$. The opposite does not hold though, since the Pearson operator does not take into account non-linear relationships. This can nicely be seen in fig. 5.1, where all distributions in the bottom row have a correlation coefficient of exactly 0, even though one can clearly observe strong non-linear dependencies. Therefore, with the considerations at the start of the chapter and the goal of a high-precision measurement in mind, a better measure of dependencies is needed.

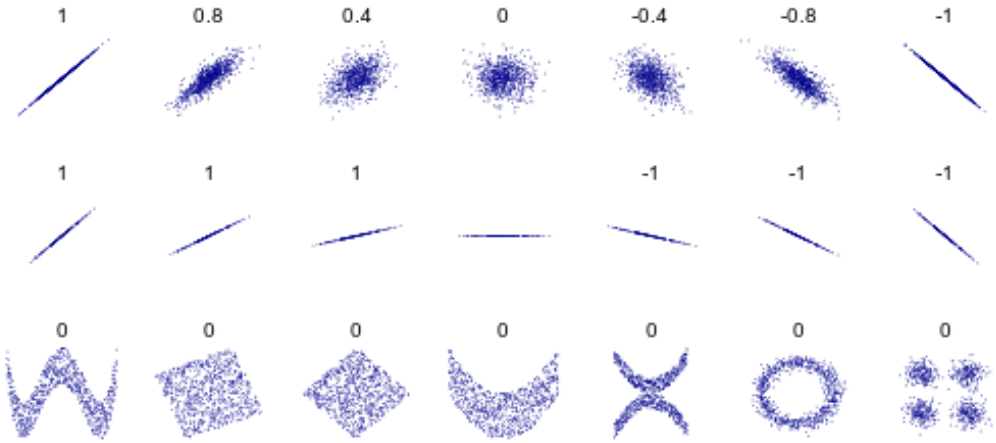

Figure 5.1: Various distributions and the corresponding values for $\rho_{X,Y}$. Note the bottom row, where $\rho_{X,Y} = 0$ for all distributions, despite the strong dependencies that can be clearly observed. Image taken from Ref. [19].

5.2. Distance Correlation

A solution is provided by the relatively recent (2007) concept of distance correlation $\mathrm{dcor}(X,Y)$ [20]. Just as $\mathrm{corr}(X,Y)$, $\mathrm{dcor}(X,Y)$ satisfies $0 \leq \mathrm{dcor}(X,Y) \leq 1$, but in contrast to Pearson's correlation, dcor is 0 if and only if the two random variables X and Y are completely independent of each other.

In order to define $\mathrm{dcor}(X,Y)$, one needs to start with the distance covariance $\mathrm{dcov}(X,Y)$. Let the primed variables X' and X'' be independently and identically distributed (iid) copies of X, and Y' and Y'' iid copies of Y, then dcov is defined as the square root of

$$\mathrm{dcov}^2(X,Y) := \mathrm{E}[|X-X'||Y-Y'|] + \mathrm{E}[|X-X'|]\mathrm{E}[|Y-Y'|] - 2\mathrm{E}[|X-X''||Y-Y''|]. \tag{5.8}$$

With that, one can define dcor analogously to $\rho_{X,Y}$, such that

$$\mathrm{dcor}(X,Y) = \frac{\mathrm{dcov}(X,Y)}{\sqrt{\mathrm{dcov}(X,X)\mathrm{dcov}(Y,Y)}} \tag{5.9}$$

$$\equiv \frac{\mathrm{dcov}(X,Y)}{\sqrt{\mathrm{dvar}(X)\mathrm{dvar}(Y)}}, \tag{5.10}$$

thereby gaining the metric used for the correlation analysis in this thesis. Examples of various distributions and their distance correlation value can be found in fig. 5.2, completely analogous to the linear correlations shown last chapter in fig. 5.1. For more information, consult Ref. [20] and [21].

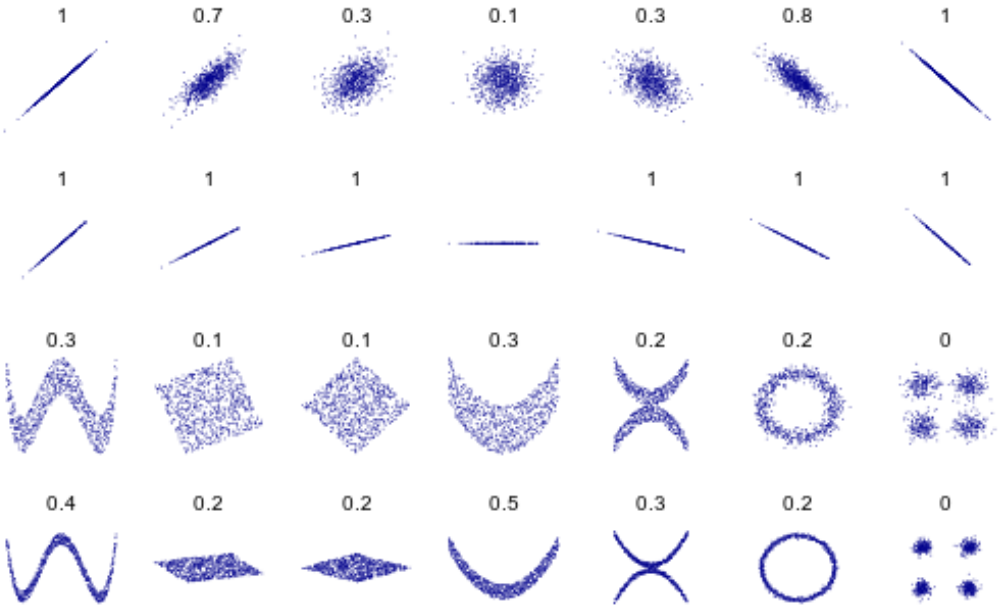

Figure 5.2: Various distributions and the corresponding values for $\mathrm{dcor}(X, Y)$, taken from Ref. [22].

6. Analysis of $B \to K^* \ell^+ \ell^-$

6.1. Overview

In this chapter and the ones following, the current state and preliminary results of the still ongoing $B \to K^* \ell^+ \ell^-$ analysis will be presented. The results shown are from simulated data only, as the full angular analysis needs to be finished first before using real data in order to avoid the introduction of biases. While the thesis at hand does not include the angular analysis itself and ends with a measurement of the branching ratios of $B \to K^* \ell^+ \ell^-$ and a corresponding statistical validation on toy data sets, all steps have been performed with the overarching goal of the angular analysis and its demands in mind.

6.2. Monte Carlo data samples

The simulated data samples used in this study were generated with the Monte Carlo (MC) method. Not only particle-particle and particle-detector interactions are simulated, but also the response of the detector to those interactions. Thereby, the simulation aims to mimic real data as well as possible, including misidentified particles, misreconstructed events in the later stages, and other biases. Of course, as the intention of this study is to apply it to real data at its very end, it needs to be developed using information that will be available on real data as well only. Nevertheless, having access to the true values of every detail in a generated event comes in very handy, as it allows to track the performance during the study by calculating signal efficiencies and purities at various steps, or as starting values for the fitting procedure.

Background data samples are centrally provided by the Belle II collaboration for all $e^+e^- \to X$ processes of relevance. For this analysis, only processes producing four tracks in the detector need to be considered, namely $e^+e^- \to \Upsilon(4S), u\bar{u}, d\bar{d}, s\bar{s}, c\bar{c}, \tau\bar{\tau}$ (see table 3.1 in chapter 3.2 for the corresponding cross sections). Since handling the full data sets is not only unfeasible due their sheer size, but also unnecessary, trimmed down versions, so-called skims, are available, which are reduced to the physical regions of interest for each individual analysis (see chapter 6.4 for the corresponding list of cuts in the skim). The Belle II internal name of the skim used in this analysis

is *BtoXll*.

The background data samples used were generated in Belle II data production campaign MC13a. While not being the most up-to-date version anymore at the time of writing, MC13a has more MC data available than newer versions, namely $5\,\mathrm{ab}^{-1}$ for the $\Upsilon(4S)$ subdecays B^+B^- and $B^0\bar{B}^0$ and $4\,\mathrm{ab}^{-1}$ for $u\bar{u}, d\bar{d}, s\bar{s}, c\bar{c}$ and $\tau\bar{\tau}$ versus $2\,\mathrm{ab}^{-1}$ each for the newest one, MC14a. The motivation for this decision is the very low BR ($\sim 10^{-7}$) and therefore scarcity of the decays of interest, making more statistics highly desirable. With more simulated data available, there is more information available for the BDT to learn from, thereby improving its predictions. For all samples, $1\,\mathrm{ab}^{-1}$ is reserved for the final validation and fitting procedure, and the rest used for training and testing.

Three signal data samples each are generated for the six decays of interest

$$B^0 \to (K^{*0} \to K^+\pi^-)e^+e^- \tag{6.1}$$
$$B^0 \to (K^{*0} \to K^+\pi^-)\mu^+\mu^- \tag{6.2}$$
$$B^+ \to (K^{*+} \to K^+\pi^0)e^+e^- \tag{6.3}$$
$$B^+ \to (K^{*+} \to K^+\pi^0)\mu^+\mu^- \tag{6.4}$$
$$B^+ \to (K^{*+} \to K_s^0\pi^+)e^+e^- \tag{6.5}$$
$$B^+ \to (K^{*+} \to K_s^0\pi^+)\mu^+\mu^-. \tag{6.6}$$

The first file contains 15 million events for the decays with e^+e^- in the final state, and 10 million events for the ones with $\mu^+\mu^-$ and is used for training and testing. The other two files per decay contain $1\,\mathrm{ab}^{-1}$ and 200 000 events respectively and are used in the final fit procedure and the following simulated measurement of the branching ratios.

6.3. Reconstruction

Since only certain particles can be measured by the detector directly, the full decay chains need to be reconstructed to properly identify signal candidates and further calculate all variables that are needed for this analysis. All particles of the signal candidate's decay chain up to the $\Upsilon(4S)$ are reconstructed, including the second B meson and its decays, with the latter also being called *Rest Of Event* (*ROE*). The

reconstruction procedure uses the software package *basf2* (belle analysis framework 2), release $06-00-03$.

A full decay chain, including the *ROE*, is referred to as an event. Notably, also all combinatorial permutations of particles assigned to an event are allowed. This means that a single event might have multiple event candidates, for example once correctly reconstructed and once with the correct K misidentified as π and the other way round, or a lepton from the *ROE* side of the event might be swapped with a signal-side lepton, or any other possible combination. Therefore, the reconstructed data contains more event candidates than actual events. Excess candidates will be eliminated only after the BDT selection, see chapter 6.8. Events are reconstructed by placing cuts upon numerous features of the data set, and only events satisfying all conditions are allowed. In the following, all cuts applied and variables used will be given and explained where necessary. Most of these cuts are motivated by the physical constraints of the decay, thereby cutting regions only where no signal is expected in the first place. Further cuts are necessary to accommodate for limitations of the detector. Starting point for the set of cuts was another established Belle 2 analysis with similar decay topology (see Ref. [23]), with some modifications taken due to the different demands of the analyses.

6.4. Skim

The first stage of cuts corresponds to the ones included in the skim. In some cases, the cuts are redundant, as stronger demands will be set at a later stage, but they shall be given for completeness' sake anyway.

The first cut uses so-called Fox Wolfram moments H_k [24]. They are defined as

$$H_k = \sum_{i,j}^{N} \frac{|\vec{p}_i||\vec{p}_j| P_k(\theta_{ij})}{E_{\text{vis}}^2}, \tag{6.7}$$

with N being the number of charged particles in the event, $|p_i|$ the momentum of charged particle i, P_k the k^{th} Legendre polynomial, $\theta_{i,j}$ the angle between particle i and particle j and E_{vis} is the total visible energy of the event.

The cut is then applied at event level on

$$R_2 = H_2/H_0 < 0.5, \tag{6.8}$$

and has the purpose of separating $B\bar{B}$ from continuum ($u\bar{u}, d\bar{d}, s\bar{s}, c\bar{c}$, background) events.

For the lepton candidates $|p_\ell| > 0.395$ GeV/c is demanded due to detector limitations. They also need to satisfy $\text{leptonID}_{\text{global}} \equiv \text{LID} > 0.1$, where lepton is a placeholder for either electron or muon, depending on the channel. LID is defined as

$$\text{LID} = \frac{L_\ell}{L_\ell + \tilde{L}_\ell}, \tag{6.9}$$

where L_ℓ denotes the likelihood of the particle's identity corresponding to ℓ and $\tilde{L}_\ell$ the likelihood of the particle not being the corresponding to ℓ. These likelihoods are calculated using information from the various detectors. Further, the energy of the dileptic system $E_{\ell\ell}$ in COM frame needs to be larger than 1.5 GeV, and E_{cluster} larger than 0.1 GeV. The last set of cuts is applied on all charged tracks ($e^\pm, \mu^\pm, K^\pm$ and $\pi^\pm$ candidates). The differentiation between charged tracks and neutral tracks is necessary since charged tracks are far easier to measure due to their magnetic field (which will also reflect on the results in later sections). Note that in all six channels, either three charged tracks are directly measured, namely two leptons and either a charged K or π due to $B^+ \to (K^{*+} \to \pi^0 K^+)\ell^+\ell^-$ or $B^+ \to (K^{*+} \to \pi^+ K^0_s)\ell^+\ell^-$, or four charged tracks due to $B^0 \to (K^{*0} \to \pi^+ K^-)\ell^+\ell^-$. For all charged tracks, the transverse momentum p_t is demanded to be larger than 0.1 GeV/c, $|dr|$ needs to be smaller than 0.5 cm, and $|dz|$ smaller than 2.0 cm. dr is defined as the distance of closest approach to the interaction point (IP) in the $r-\phi$ plane, and dz on the z axis. This restriction makes sure that it is reasonably likely that the particles corresponds to said IP. Finally, the sample only includes events with at least 3 charged tracks, a requirement of course also induced by the intrinsic demands of the analysis.

6.5. Correlation analysis

In order to be able to extract P_5' in the final angular fit, it is important that the variables θ_k, θ_l and ϕ (see chapter 2.3 and fig. 2.4) remain as unbiased as possible. The same is true for the beam-constrained mass M_{bc}, defined as

$$M_{bc} = \sqrt{E^2_{\text{beam}} - \vec{p}^2_B}, \tag{6.10}$$

with $\vec{p}_B$ being the 3-momentum of the B meson candidate, and is used to model

background and signal yields and calculate the BR. Therefore it is demanded that all variables used in training and preselection are below a predefined dcor (see chapter 5.2) threshold with M_{bc}, θ_k, θ_ℓ and ϕ on all combined background samples for every channel individually. For this analysis, only unavoidable cuts (skims, vetoes on irreducible J/Ψ, $\Psi(2S)$ and Dalitz decays on q^2, see below) are included. The procedure is deemed successful when the outputs of the BDTs satisfy the same condition, thereby proofing that the BDTs did not learn any undesired correlations. dcor < 0.1 was chosen as threshold, as it turned out to be a reasonable middle ground between performance and accuracy at the current stage of the analysis. A list of correlations for all variables considered can be found in tables A.1 to A.6 in the appendix.

6.6. Particle selection

This part will present how the individual particle candidates are selected. The only particles that are directly reconstructed by the detector are the charged particles $e^\pm$, $\mu^\pm, \pi^\pm, K^\pm$, and the neutral particles K_s^0, π^0 and γ. All other particles in the decay chain are traced back from those. The charged particles directly reconstructed in the detector all have a particleID (PID) associated to them, which is calculated by combining the likelihoods given by the various subdetectors for each individual particle hypothesis. Hypothetical electrons and muons are selected using a global $\text{PID}_{\text{global}}$ as introduced in chapter 6.4, while charged kaons and charged pions are selected using a binary $\text{PID}_{\text{binary}}$. While $\text{PID}_{\text{global}}$ uses the likelihood of a specific particle hypothesis and the likelihood of that hypothesis being wrong following equation 6.9, therefore taking into account all six possible particles, $\text{PID}_{\text{binary}}$ only considers the likelihoods of two specific particle hypothesis, which can be advantageous when only two types need to be considered.

For kaons, this binary PID is defined as

$$R(K/\pi) \equiv \text{kaonID}_{\text{binary}} = \frac{L_K}{L_K + L_\pi} \tag{6.11}$$

and for pions as

$$R(\pi/K) \equiv \text{pionID}_{\text{binary}} = \frac{L_\pi}{L_K + L_\pi}. \tag{6.12}$$

PIDs can only be used at specific thresholds, as provided by the responsible working group. For these thresholds, particle weights correcting for known discrepancies between MC and real data depending on momentum and energy are available. Furthermore, information of systematic uncertainties are available as well. Since the main limitation of this study is the low amount of statistics, higher efficiencies, defined as

$$\epsilon = \frac{N_{\text{signal, reconstructed}}}{N_{\text{signal, generated}}}, \tag{6.13}$$

are desirable, which is why it was attempted to leave the thresholds as low as possible. For both electrons and muons, global electronID > 0.5 and global muonID > 0.5 respectively was chosen, which are the smallest available thresholds. While this leads to an increase in fake rate of up to 40% with respect the next lowest available threshold value of 0.9, the gain in signal yield was considered worth it. A commonly used trick is to apply leptonID only on one of the two hypothetical leptons and assuming that the second particle must be a lepton as well due to all of the other constraints set. Tests using that trick did not improve the results though, and would complicate the evaluation of systematic errors, which is why this idea was given up for the time being. For the kaon and pion, binary kaonID > 0.6 and binary pionID > 0.6 was used respectively. This is only the second lowest threshold provided by Belle II, but using 0.1 did not result in significant improvements with respect to signal yield and would have negatively impacted the amount of background before the BDT, while about doubling the fake rate.

Furthermore, all four of those particles are required to have at least 20 registered hits in the CDC subdetector, which is the recommended value by the responsible Belle II working group. Also, muon momentum must be bigger than $0.8\,\text{GeV/c}$ such that it is physically possible for them to reach the KLM. For electron momentum, the recommended threshold is $0.4\,\text{GeV/c}$, and their cluster energy is required to be > 0.075 in forward endcap, > 0.05 in the barrell, and $> 0.1\,\text{GeV}$ in backward endcap.

Neutral particles do not have a PID associated to them. For K_s^0, the requirement is a successful vertex fit of its decay products and an invariant mass in the range of $[0.4867, 0.5076]\,\text{GeV/c}^2$, which is about $3\,\sigma$ of its nominal invariant mass. π^0 are reconstructed from two photons (decay ratio of ~ 0.98 [4]). The recommended requirements by the responsible working group are a minimum cluster energy in the

ECL of 0.08 for forward endcap, 0.03 in barrell and 0.06 GeV in backward endcap, and a cluster polar angle within $[0.2967, 2.6180]$ rad. Furthermore, the sum of weights of all crystals in an ECL cluster needs to be larger than 1.5. This value can be non-integer if the energy is split between nearby clusters. The absolute angle between the two photons must be smaller than 1.4 rad, and in the ϕ plane smaller than 1.5. Finally, the invariant mass of the reconstructed π^0 is demanded to be within $[0.1215, 0.1415]$ Gev/c^2.

For K^*, all candidates are required to have an invariant mass between 0.796 and 0.996 GeV/c^2 , which is roughly $4\,\sigma$ of its nominal invariant mass. K^* are reconstructed either as K^{*0} in $B^0 \to K^*\ell^+\ell^-$ decays, or as K^{*+} in $B^+ \to K^{*+}\ell^+\ell^-$ decays. Candidates for K^{*0} are reconstructed as $K^{*0} \to K^+\pi^-$, which occurs with a probability of 2/3, as can be calculated using Clebsch-Gordan coefficients. The second possible decay channel $K^{*0} \to \pi^0 K^0$ has a decay probability of 1/3 and is not taken into account due to the problems associated with detecting neutral particles and the corresponding low signal yield at high background levels.

Candidates for K^{*+} are reconstructed from $K^{*+} \to K^0\pi^+$, (probability of 2/3), and $K^{*+} \to K^+\pi^0$ (probability of 1/3). Those channels are not neglected, as there is only one neutral particle per channel, which makes the measurement easier in comparison to the case with two neutral particles. Unfortunately, the signal yield for the π^0 channel is still very low, as will be shown in chapter 6.9. Of the two K^0 *CP* eigenstates, only the short-lived K^0_s (decay modes with $> 5\%$ of decays $K^0_s \to \pi^+\pi^-$ and $K^0_s \to \pi^0\pi^0$) are reconstructed. The reconstruction of K^0_l is not possible as it is too long-lived to decay within the detector and leaves no track in the detector due to its neutral charge. Therefore, only 1/2 of neutral kaons are reconstructed, resulting in a total of 1/3 of events for $B^+ \to K^{*+}\ell^+\ell^-$ that are neglected. To select $K^{*+} \to K^+\pi^0$, two additional cuts are applied on the helicity angles in order to suppress combinatorial background. In the K^{*+} rest frame, the absolute value of the cosine of the angle between the hypothetical B^+ and the K^+ must be smaller than 0.8. In the π^0 rest frame, the absolute value of the cosine of the angle between the hypothetical K^{*+} and γ (first daughter of π^0) must also be smaller than 0.8. Last but not least, the B meson is reconstructed via a vertex fit of a K^* and two oppositely charged leptons of the same flavour.

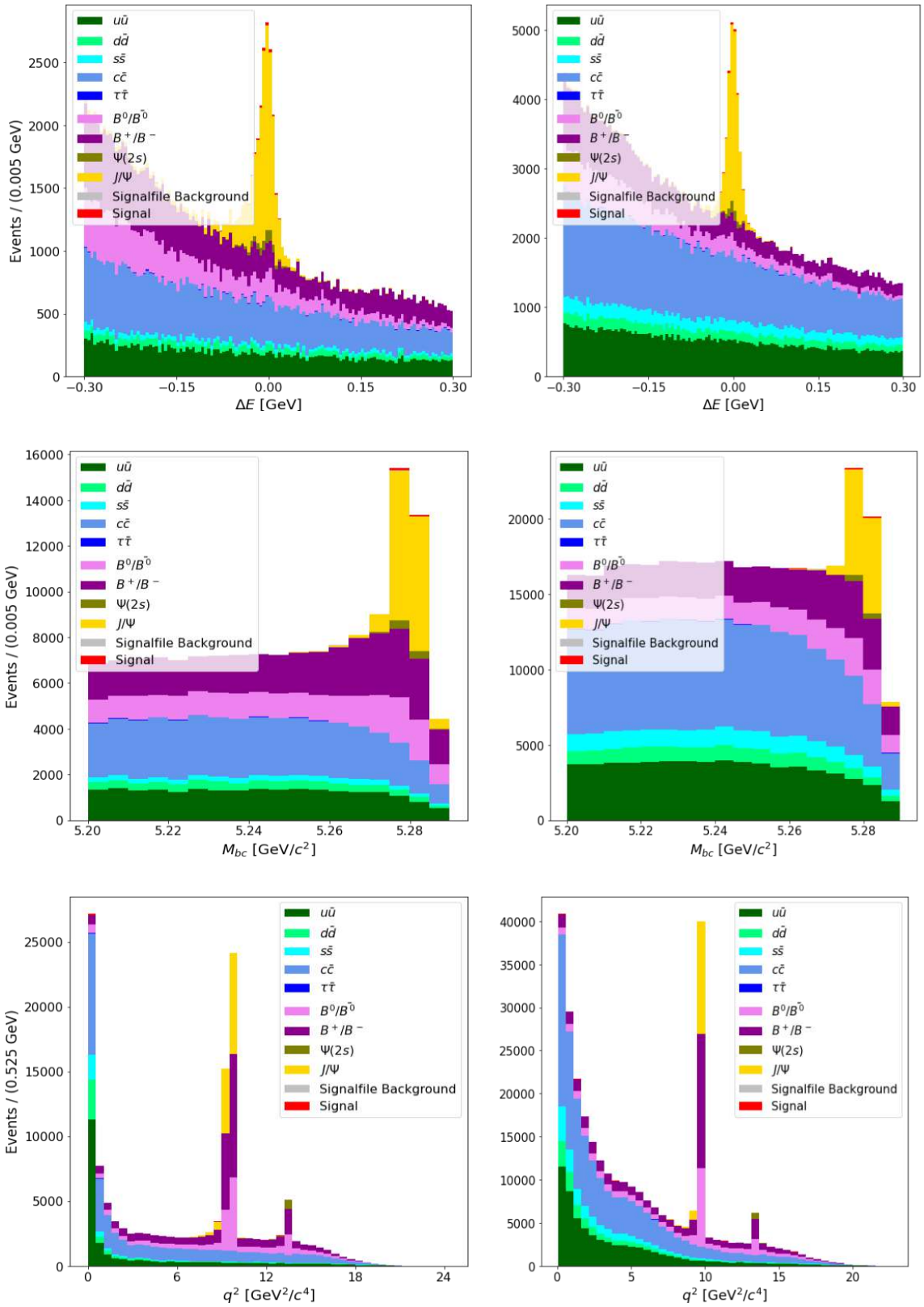

Figure 6.1: Plots of ΔE, M_{bc} and q^2 for the channels $B^0 \to (K^{*0} \to K^+\pi^-)e^+e^-$ (left side) and $B^0 \to (K^{*0} \to K^+\pi^-)\mu^+\mu^-$ (right side) directly after reconstruction with no further preselection cuts applied.

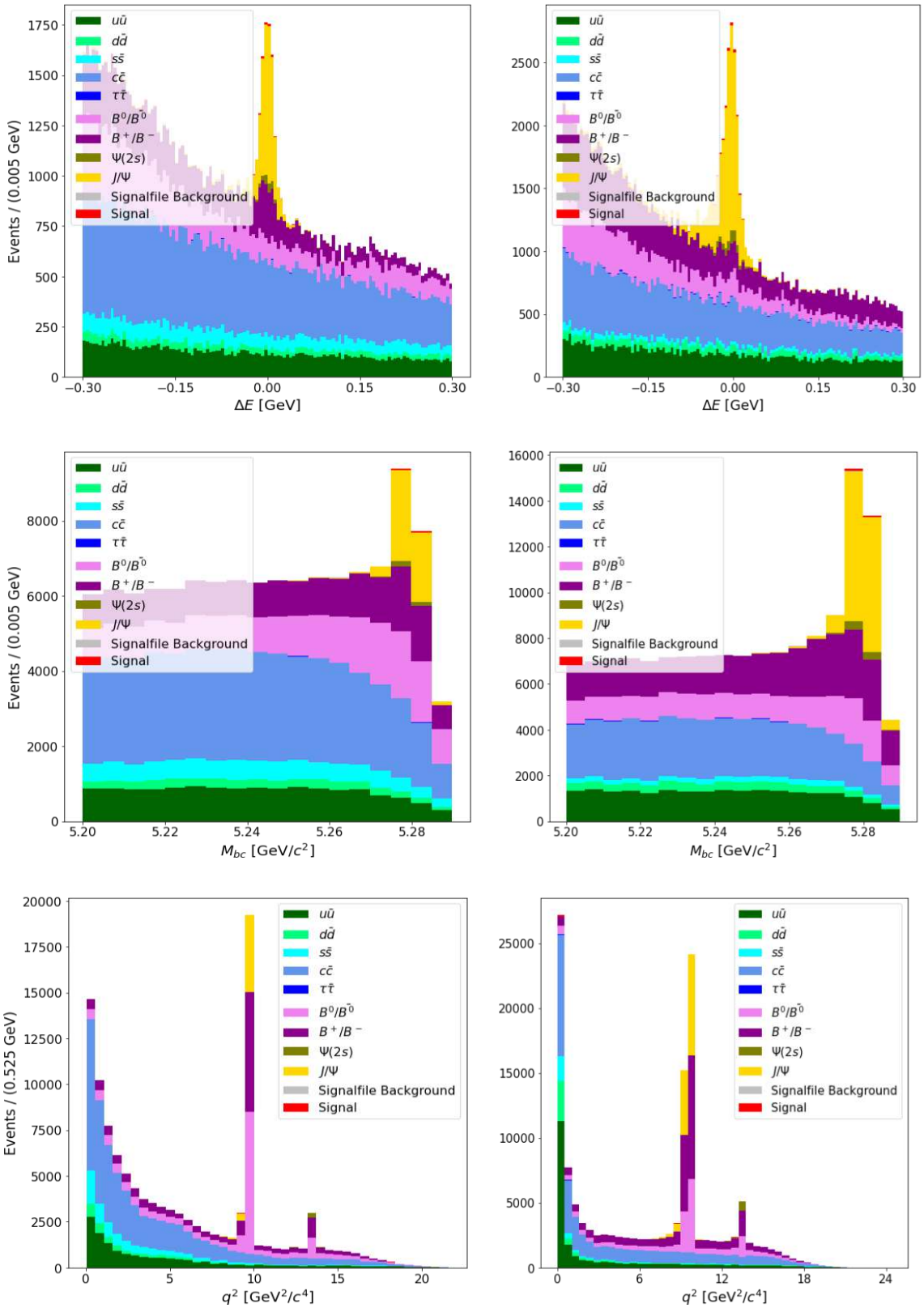

Figure 6.2: Plots of ΔE, M_{bc} and q^2 for the channels $B^+ \to (K^{*+} \to K_s^0\pi^+)e^+e^-$ (left side) and $B^+ \to (K^{*+} \to K_s^0\pi^+)\mu^+\mu^-$ (right side) directly after reconstruction with no further preselection cuts applied.

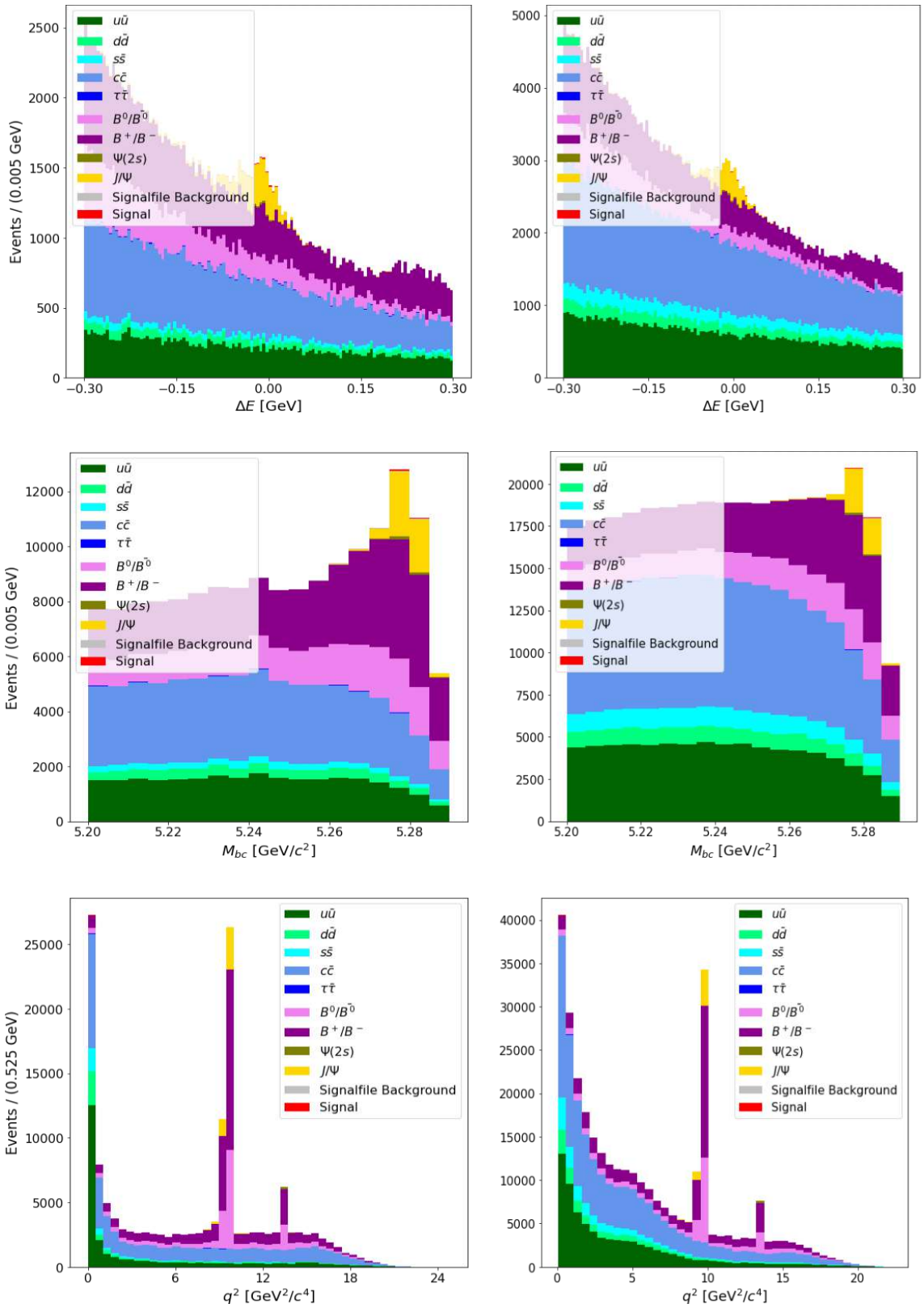

Figure 6.3: Plots of ΔE, M_{bc} and q^2 for the channels $B^+ \to (K^{*+} \to K^+\pi^0)e^+e^-$ (left side) and $B^+ \to (K^{*+} \to K^+\pi^0)\mu^+\mu^-$ (right side) directly after reconstruction with no further preselection cuts applied.

6.7. Event selection

Two very useful kinematic variables for discriminating background and signal are ΔE and M_{bc}. ΔE is defined as

$$\Delta E = E_B - E_{\text{beam}}, \tag{6.14}$$

where E_B is the energy of the reconstructed B meson and E_{beam} the beams energy, both given in the COM mass frame. Since collisions are produced at resonance of $\Upsilon(4S)$, which in turn decays into $B\bar{B}$, the B mesons are produced almost at rest, leading to signal events peaking around $\Delta E = 0$. As a result, many background B channels peak at $\Delta E \stackrel{!}{=} 0$, making them easy to veto. Another reason making ΔE useful is that misreconstructed and continuum background components are of a roughly exponential shape, in stark contrast to the gaussian-like peak with an asymmetric tail of signal decays. The plots in the top rows of figures 6.1 to 6.3 show the ΔE distributions of $1\,\text{ab}^{-1}$ of MC data for the six reconstructed decay channels, where one can observe the aforementioned shapes. The left side plots show the e^+e^- channels, and the right side plots the corresponding $\mu^+\mu^-$ channels. One can observe that the B-mesons (B^+ in violet, B^0 in pink) backgrounds exhibit both a gaussian peak similarly distributed to the signal decays in red, and also an exponential part due to misreconstructed decays, while all other mother particles produce exponential shapes only. The plots in the middle rows of the aforementioned figures show M_{bc}. Correctly reconstructed B decays, both signal and background, produce a gaussian-like peak around a value of $5.28\,\text{GeV/c}^2$, while the combinatorial background again exhibits a mostly flat shape.

Both M_{bc}'s and ΔE's signal and background shapes can be very well modeled due to the very distinct shapes of signal and most parts of the background, making them popular candidates to extract the signal yields from using a curve fitting procedure. A common strategy is to use both variables at once for a simultaneous $2D$-fit, with the advantage of a higher precision in comparison to a fit with just a single variable, but two major drawbacks. The first problem is that one would like to retain an as wide as possible range of the variables part of the fit in order to be able to gain as accurate estimates of the curve parameters as possible. This might lead to a drastic increase in data without any gain in signal, as it is the case in this study. Secondly, a variable used in the fitting procedure cannot be used for training, and

neither can any variables strongly correlated with it. Otherwise the BDT skews the background towards a signal-like shape, making it much harder if not impossible to find accurate parameters in the fitting procedure (see chapter 6.9. Ultimately, one trades systematic accuracy with statistical accuracy. Since we were not able to match Wehle et al.'s Belle I post-BDT results scaled to the same luminosity (see section 6.8) while sacrificing both M_{bc} and ΔE in training, which are both very strong variables, it was decided to use M_{bc} only for fitting, following the approach used in the Belle I analysis. For M_{bc}, the chosen upper boundary of $5.29\,\mathrm{GeV/c^2}$ is mostly cosmetic, because M_{bc} drops off sharply after the peak at 5.28 and no physical events are expected past $5.29\,\mathrm{GeV/c^2}$ anyway. The exact value of the lower bound is not critical either, as long as the range is large enough. Therefore it makes sense to just use the same use the same range as Belle I for ease of comparison, with a lower limit of $5.2\,\mathrm{GeV/c^2}$. The ΔE values follow Belle I as well, with

$$-0.1\,\mathrm{GeV} < \Delta E < 0.05\,\mathrm{GeV} \tag{6.15}$$

for electrons and

$$-0.05\,\mathrm{GeV} < \Delta E < 0.05\,\mathrm{GeV} \tag{6.16}$$

for muons. Just like before, the exact values are not critical either, as long as the bounds are not too tight, since the variable is fed into the BDT anyway. A summary of all cuts applied so far can be found in table 6.1.

The bottom row plots of figures 6.1 to 6.3 show the invariant mass of the dileptic system $M_{\ell\ell}$ squared, a variable also commonly called q^2. One large peak and two smaller ones can be observed. Part of the middle peak and the right side peak result from the decays $B \to (J/\psi(2S) \to \ell^+\ell^-)K^*$ and $B \to (\Psi(2S) \to \ell^+\ell^-)K^*$ respectively, which have the same final state particles as the signal decays. Both of these decays are also referred to as irreducible decays, as they have a very similar shape as the signal decay, and also very similar properties in general. Both peak around 0 GeV in ΔE and $5.28\,\mathrm{GeV}/c^2$ in M_{bc}, and the BDT has no way to distinguish them. As there is no other way to get rid of them, they are vetoed using the cuts

$$2.846^2\,\mathrm{GeV^2/c^4} < q^2 < 3.176^2\,\mathrm{GeV^2/c^4} \tag{6.17}$$

$$3.439^2\,\mathrm{GeV^2/c^4} < q^2 < 3.719^2\,\mathrm{GeV^2/c^4} \tag{6.18}$$

for electrons and

$$2.946^2\,\mathrm{GeV}^2/\mathrm{c}^4 < q^2 < 3.176^2\,\mathrm{GeV}^2/\mathrm{c}^4 \tag{6.19}$$

$$3.539^2\,\mathrm{GeV}^2/\mathrm{c}^4 < q^2 < 3.719^2\,\mathrm{GeV}^2/\mathrm{c}^4 \tag{6.20}$$

for muons.

The same is true for a fraction of left hand peak in the e^+e^- final state decays, which is solved by demanding $q^2 > 0.14^2\,\mathrm{GeV}^2/\mathrm{c}^4$. Two different decays are responsible for this background component, namely Dalitz decays $B \to K^*(\pi^0 \to e^+e^-\gamma)$ and photon conversion decays $B \to K^*(\gamma \to e^+e^-)$.

When vetoing these three peaks in q^2, the signal-like peaks in ΔE and M_{bc} disappear as well as a direct result, as can be nicely seen in figures 6.4 to 6.6. Again the top rows show Δ E, the middle rows M_{bc} and the bottom rows q^2, with the left hand sides showing the e^+e^- decays and the right hand sides the corresponding $\mu^+\mu^-$ decays. A summary of all cuts applied is given in table 6.1.

Table 6.1: List of all cuts applied to the data set before the BDT.

skim cuts		
all events:	R_2	$x < 0.5$
	p_ℓ	$x > 0.395\,\mathrm{GeV}$
	$E_{\ell\ell,\text{ in CMS frame}}$	$x > 1.5\,\mathrm{GeV}$
$e^\pm$ in final state:	$\mathrm{electronID}_{\mathrm{global}}$	$x > 0.1$
$\mu^\pm$ in final state:	$\mathrm{muonID}_{\mathrm{global}}$	$x > 0.5$
charged tracks:	$n_{\text{CDC Hits}}$	$x > 20$
	dr	$\lvert x\rvert < 0.5\,\mathrm{cm}$
	dz	$\lvert x\rvert < 2.0\,\mathrm{cm}$
preselections		
all events:	M_{bc}	$5.2\,\mathrm{GeV}/\mathrm{c}^2 < x < 5.29\,\mathrm{GeV}/\mathrm{c}^2$
	M_{K^*}	$0.796\,\mathrm{GeV}/\mathrm{c}^2 < x < 0.996\,\mathrm{GeV}/\mathrm{c}^2$
$e^\pm$ in final state:	$\mathrm{electronID}_{\mathrm{global}}$	$x > 0.5$
	p_ℓ	$x > 0.4\,\mathrm{GeV}$
	q^2	$x > 0.14^2\,\mathrm{GeV}^2/\mathrm{c}^4$
	ΔE	$-0.1\,\mathrm{GeV} < \Delta E < 0.05\,\mathrm{GeV}$

continued on next page

Table 6.1 – continued from previous page

	$E_{\ell,\text{ forward endcap}}$	$x > 0.075\,\text{GeV}$
	$E_{\ell,\text{ barrell}}$	$x > 0.05\,\text{GeV}$
	$E_{\ell,\text{ backward endcap}}$	$x > 0.1\,\text{GeV}$
	q^2	$2.846^2\,\text{GeV}^2/\text{c}^4 < x < 3.176^2\,\text{GeV}^2/\text{c}^4$
	q^2	$3.439^2\,\text{GeV}^2/\text{c}^4 < x < 3.719^2\,\text{GeV}^2/\text{c}^4$
$\mu^\pm$ in final state:	q^2	$2.946^2\,\text{GeV}^2/\text{c}^4 < x < 3.176^2\,\text{GeV}^2/\text{c}^4$
	q^2	$3.539^2\,\text{GeV}^2/\text{c}^4 < x < 3.719^2\,\text{GeV}^2/\text{c}^4$
	p_ℓ	$x > 0.8\,\text{GeV}$
$K^\pm$ in final state:	$\text{kaonID}_{\text{binary}}$	$x > 0.6$
	$n_{K,\text{CDC hits}}$	$x > 20$
$\pi^\pm$ in final state:	$\text{pionID}_{\text{binary}}$	$x > 0.6$
	$n_{\pi,\text{CDC hits}}$	$x > 20$
K_s^0 in final state:	M_K	$0.4876\,\text{GeV/c}^2 < x < 0.5076\,\text{GeV/c}^2$
π^0 in final state:	$n_{\gamma,\text{ECL hits}}$	$x > 1.5$
	$\theta_{\gamma,\gamma}$	$0.2967\,\text{rad} < x < 2.618\,\text{rad}$
	$\phi_{\gamma,\gamma}$	$x < 1.5\,\text{rad}$
	$\text{AbsoluteAngle}_{\gamma,\gamma}$	$x < 1.4\,\text{rad}$
	M_π	$0.1215\,\text{GeV/c}^2 x < 0.1415\,\text{GeV/c}^2$
	$E_{\gamma,\text{ forward endcap}}$	$x > 0.08\,\text{GeV}$
	$E_{\gamma,\text{ barrell}}$	$x > 0.03\,\text{GeV}$
	$E_{\gamma,\text{ backward endcap}}$	$x > 0.06\,\text{GeV}$
$K^{*+} \to K^+\pi^0$	$\text{HelicityAngle}_{B^+,K^+}$	$\lvert\cos x\rvert < 0.8$
	$\text{HelicityAngle}_{K^{*+},\gamma}$	$\lvert\cos x\rvert < 0.8$

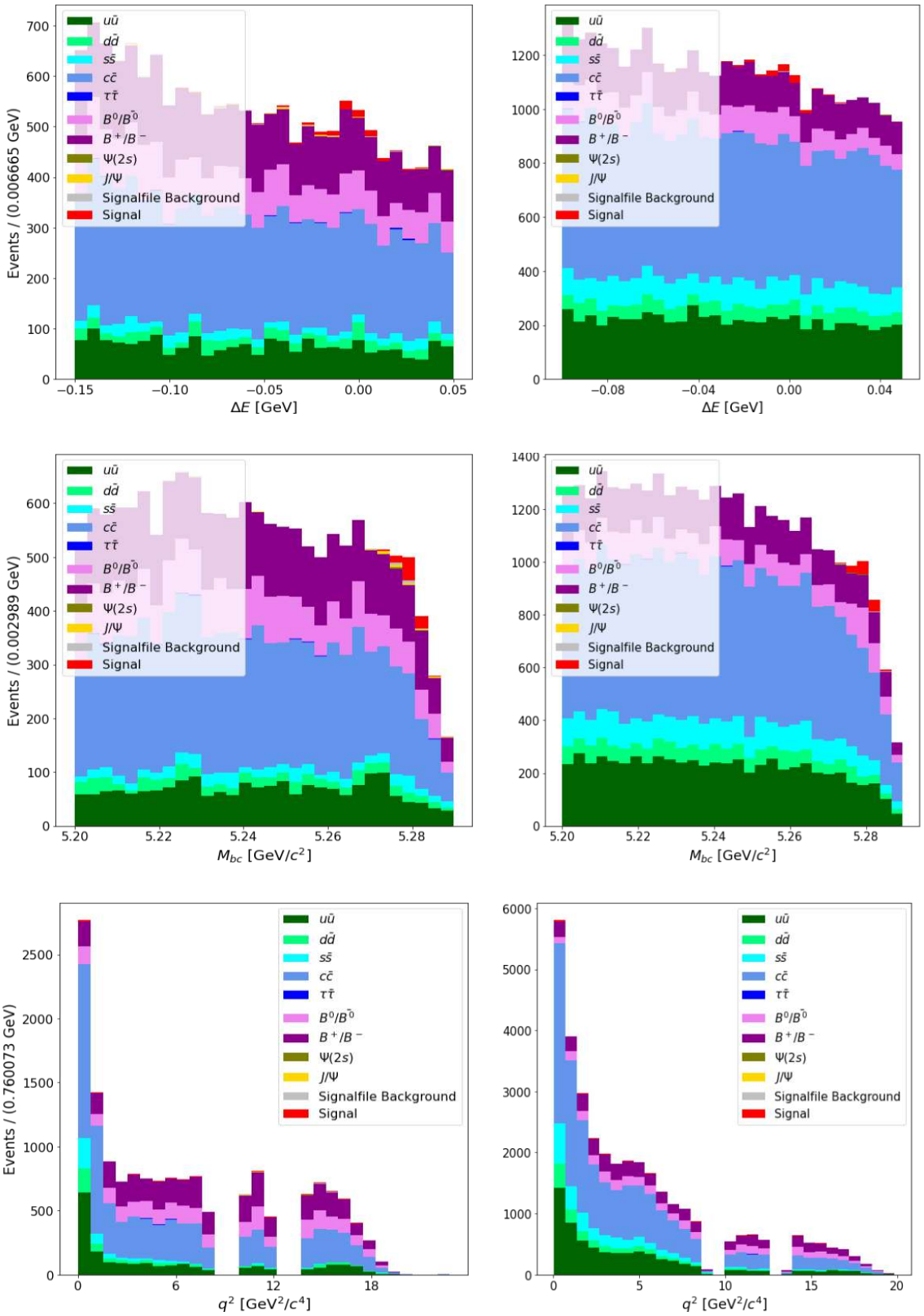

Figure 6.4: Plots of ΔE, M_{bc} and q^2 for the channels $B^0 \to (K^{*0} \to K^+\pi^-)e^+e^-$ (left side) and $B^0 \to (K^{*0} \to K^+\pi^-)\mu^+\mu^-$ (right side) after all preselection cuts before the BDT.

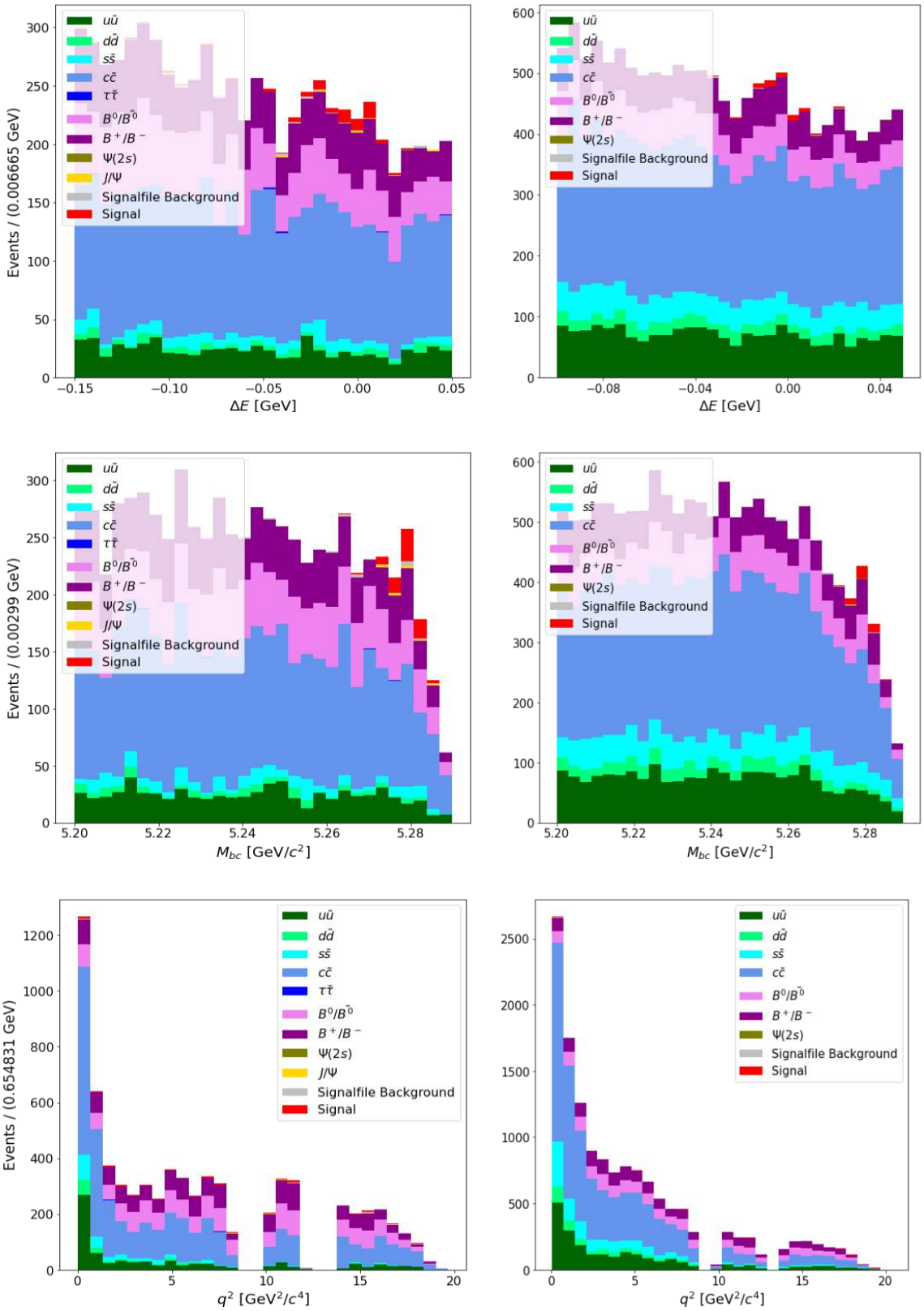

Figure 6.5: Plots of ΔE, M_{bc} and q^2 for the channels $B^+ \to (K^{*+} \to K_s^0\pi^+)e^+e^-$ (left side) and $B^+ \to (K^{*+} \to K_s^0\pi^+)\mu^+\mu^-$ (right side) after all preselection cuts before the BDT.

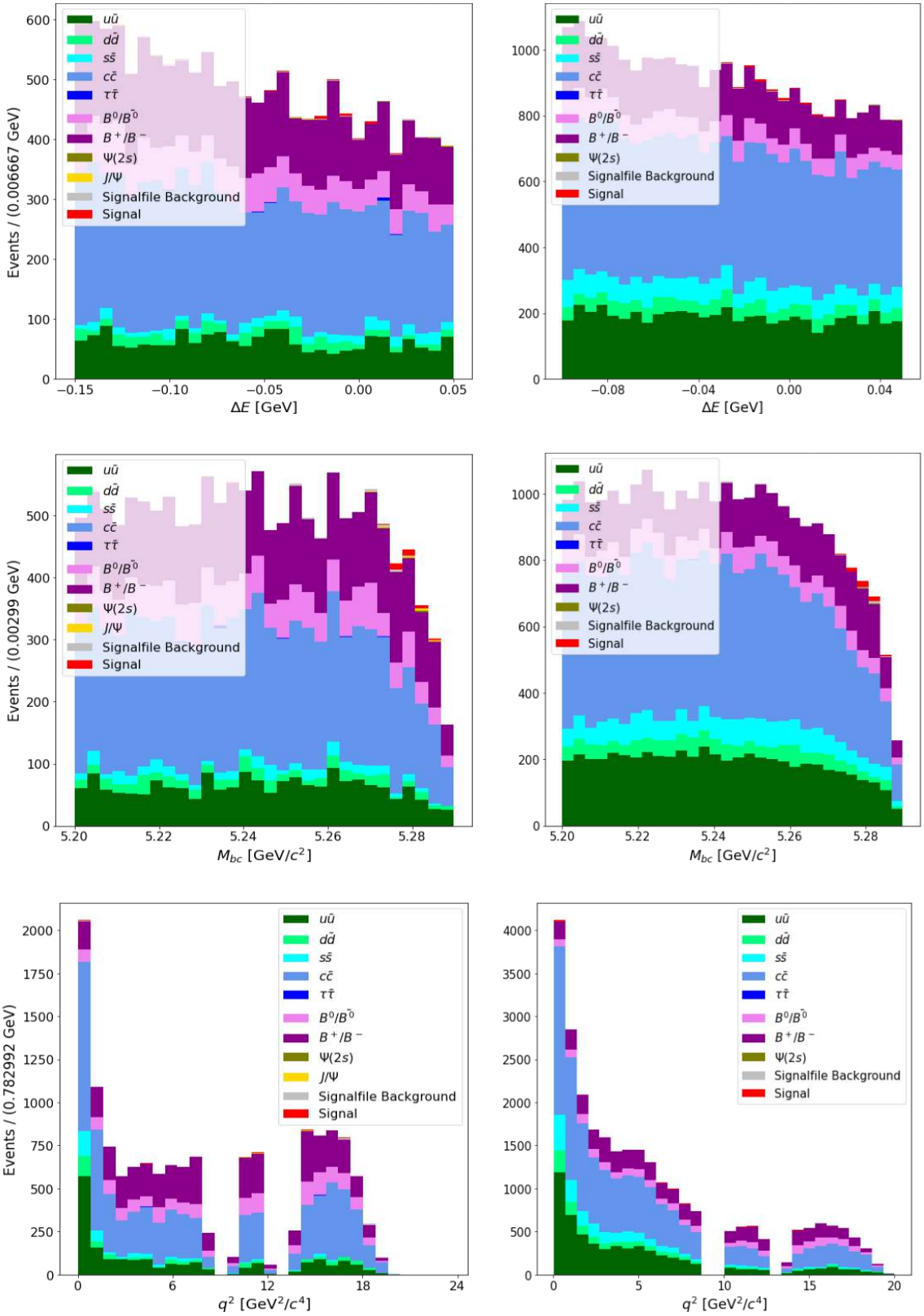

Figure 6.6: Plots of ΔE, M_{bc} and q^2 for the channels $B^+ \to (K^{*+} \to K^+\pi^0)e^+e^-$ (left side) and $B^+ \to (K^{*+} \to K^+\pi^0)\mu^+\mu^-$ (right side) after all preselection cuts before the BDT.

6.8. Background suppression

After the preselection cuts presented in the last chapters, background levels are still very high. One could easily tighten the cuts further and introduce new ones to improve the purity

$$P = \frac{N_{\text{signal}}}{N_{\text{signal}} + N_{\text{background}}} \tag{6.21}$$

of the sample, but this would be nowhere near as effective as using a multivariate method and lose much more efficiency in the process. This is why the preselection cuts are intentionally left lose, leaving the hard work to the BDT, whose setup will be described in this chapter (see section 4 for a general introduction).

The variables used for training the BDT where determined with the correlation analysis procedure described in section 6.5. This leaves the following variables that satisfy the condition dcor < 0.1 for M_{bc} and the three angular variables θ_ℓ, θ_k and ϕ on the combined background for all six decay channels separately.

- M_{K^*}
- ΔE
- dr and dz of B
- $\cos\theta$ of K and π, where θ is the polar angle
- momentum components p_x, p_y, p_z of K^* and both leptons
- dr, dx, dy, dz and θ of both leptons, where θ is the polar angle
- χ^2_{Prob}, the probability of the B vertex fit having been successful
- CLEO Cone Thrusts (CCT): Variables that are based on the sum of the absolute values of the momenta of all particles within angular sectors around the thrust axis in intervals of 10 degrees, resulting in 9 concentric cones. The thrust axis $\vec{T}$ is defined as the unit vector along the thrust T for a set of N particles with momenta p_i, with

$$T = \frac{\sum_{i=1}^{N} |\vec{T} * \vec{p}_i|}{\sum_{i=1}^{N} |\vec{p}_i|} \tag{6.22}$$

- $B_{3,\text{thrust}}$, also called third harmonic moment along the thrust axis, is the third coefficient of the expansion of the event in spherical harmonics around the thrust axis $\vec{T}$. The coefficients are defined as

$$B_l = \sum_{i=1}^{N} \frac{p_i P_l \cos \alpha_i}{\sqrt{s}}, \tag{6.23}$$

  where P_l is the Legendre polynomial of order l, α_i is the angle between momentum of particle i and the chosen axis, and s the respective Mandelstam variable.

- $\delta x_{\ell\ell}, \delta y_{\ell\ell}, \delta z_{\ell\ell}$: separation between the two leptons along direction i

- $E_{\text{vis}}(ROE)$: total sum of energy of trucks and clusters of ROE

- $E_{\text{extra}}(ROE)$: energy from clusters in the ECL that is not associated with the candidate event

- $p(ROE)$: total momentum of unused tracks in CMS frame.

As input to the BDT, the background samples for training described in chapter 6.2 were used, after application of the preselection cuts. Further, the full signal samples with 10 million events for electrons and 5 million events for muons were used, as well after the application of the preselection cuts. The signal samples were not sampled down, but rather weighted such that the effective ratio is 50/50 using *XGBoost*'s parameter *scale_pos_weight*. Of the combined signal and background sample, 70% was used for the actual training, and 30% as a test set for early stopping with the metric *aucPR* (*area under curve Precision Recall*). Furthermore, the data set is split up at $q^2 = 10\,\text{GeV}^2/\text{c}^4$ into two regions of low q^2 and high q^2 respectively, which produced much better results than just a single bin. Increasing the number to four to match the q^2 bins used in the angular analysis did not result in any improvement. Additionally, the following three multi-step BDTs were tested:

- Individual training of continuum background + signal, and $B\bar{B}$ background + signal, then combining the results

- First BDT with continuum + signal only, then apply the output of the first BDT on the whole data set and train again

- Just a single BDT for the whole data set

The second method consistently produced the best results. A comparison of the results for four different combinations of the options described (including different sized q^2 bins) that were still in consideration after the latest update of the data samples is given at the end of the chapter in table 6.6.

For *XGBoost*, a histogram-based approach was used, approximating the data set as histograms to accommodate for the large amount of data. The exact approach increased training times by more than an order of magnitude, and did not yield any gain in performance. Training stops when the *aucPR* evaluation did not improve for 100 new trees in a row on the test set, while dropping the last 100 trees. A lot of experimenting was done with respect to hyperparameter tuning. The algorithm *Bayesian Optimize* from the *scikit-optimize* package produced good results, but was dropped again, since optimizing 24 BDTs (6 channels $\times$ 2 steps $\times$ 2 q^2 bins) proved way too computationally expensive for the slight improvements gained, considering the very regular updates of the data demanded by a still in-progress analysis. In the end, the parameters were optimized by hand, such that the BDT parameters produce consistently good results for all channels. The learning rate was chosen to be 0.05, at a maximum tree depth of 8 for the first BDT and a maximum tree depth of 10 for the second BDT. It might become worthwhile to re-investigate *scikit-optimize* when the full angular analysis is finished and it is certain that no re-training will be necessary in order to squeeze out the very last drops of performance.

The trained models are then applied on $1\,\mathrm{ab}^{-1}$ of independently generated validation data. The model's output is a probability-like number between 0 and 1 for every data point, where 0 indicates a background candidate and 1 a signal candidate (see fig 6.7). The working point is chosen such that the figure of merit

$$\mathrm{FOM} = \frac{N_{\mathrm{signal}}}{\sqrt{N_{\mathrm{signal}} + N_{\mathrm{background}}}} \tag{6.24}$$

is maximized. The FOM is the metric of choice in this study to evaluate the performance, and will also be used to compare the results to Belle I. All FOMs given in this thesis, also including the Belle I results, refer to the interval $M_{bc} = [5.27, 5.29]\,\mathrm{Gev}/c^2$ only, which is the region where signal is expected. The maximization is done for each individual channel and q^2 bin, resulting in $2D$ problems that are solved using

the *Basin-Hopping* algorithm (*scipy.optimize.basinhopping*), which combines a global stepping algorithm with a local minimization (or maximization as in this case) at each step [25]. 100 iterations at a step-size of 0.05 proved to be a reasonable middle-way between accuracy and computational expense.

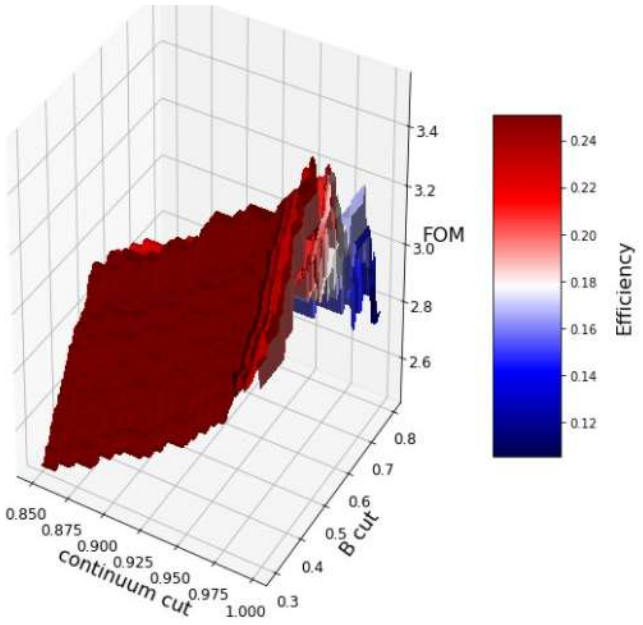

Figure 6.7: Example plot for the selection of BDT working points. The x and y axis show the two BDT probability-like outputs, while the z axis shows the FOM. The algorithm *Basin-Hopping* is used to search for the global FOM maximum. The color indicates the efficiency.

The final step is to eliminate the excess candidates, such that only one candidate per event remains. Best candidate is considered to be the one with the highest $Output_{\mathrm{BDT1}} \times Output_{\mathrm{BDT2}}$.

Figures 6.8 to 6.10 show the resulting ΔE distributions in the top rows, M_{bc} in the middle rows and q^2 in the bottom rows for the six channels. Again the left sides show the e^+e^- channels and the right sides the corresponding $\mu^+\mu^-$ channels. By comparing with the corresponding plots in figures 6.4 to 6.6, one can see that the overall distribution in M_{bc} stay roughly similar, which indicates little to no unwanted correlation. This is confirmed with the distance correlation dcor, following chapter 5.2. The only variable showing some correlation with the BDT output is θ_ℓ, with dcor of up to 0.15 (see table A.7 in the appendix). Whether this turns out to be

problematic will have to be evaluated after developing the angular analysis.

While the difference in BDT working points between the two q^2 regions might be very drastic for some channels such as $B^+ \to (K^{*+} \to K^+\pi^0)e^+e^-$, it does not matter as P_5' will be calculated for the q^2 bins individually anyway. Tables 6.2 to 6.5 show the corresponding key figures, such as N_{signal}, $N_{\text{background}}$, ϵ, FOM and P. The error of the efficiency is calculated as

$$\epsilon_{\text{error}} = \frac{N_{\text{signal, reconstructed}}(N_{\text{signal, generated}} - N_{\text{signal, reconstructed}})}{N^3_{\text{signal, generated}}}. \tag{6.25}$$

For this evaluation, the K^* subdecays $K^{*+} \to K^+\pi^0$ and $K^{*+} \to K^0_s\pi^+$ are added together, since the signal yields of the individual channels are too low to be useful, and to further allow for a comparison to the Belle I results by Wehle et al., given in the same tables where available.

The key figures are evaluated for the full q^2 range and the four q^2 bins $q^2_1 = [0.1, 4.0], q^2_2 = [4.0, 8.0], q^2_3 = [x, x], q^2_4 = [x, 19.0]$, where P_5' will be calculated on. Note that q^2_2 is of particular importance, as it is this region where the strongest discrepancies to the Standard Model where found in previous studies (see chapter 2.4).

In addition to the original $1\,\text{ab}^{-1}$, the results were also scaled down to $711\,\text{fb}^{-1}$ and $350\,\text{fb}^{-1}$. $711\,\text{fb}^{-1}$ is the amount of data used in the Belle I analysis, which allows for a direct comparison of the performance, at least as far as MC data is concerned. $350\,\text{fb}^{-1}$ is roughly the amount of Belle II data expected before the next shutdown and therefore demonstrates what is actually feasible within the foreseeable future. The scaled down results need to be taken with a grain of salt though, as the BDT working points were optimized for $1\,\text{ab}^{-1}$, and might be slightly inaccurate when scaled down to lower integrated luminosities. When comparing the FOMs to Belle 1, there are significant differences to be observed between the individual channels. While the B^+ channels perform very well, being roughly equivalent or even better than Belle I even at $350\,\text{fb}^{-1}$, B^0 does not perform as well. At $711\,\text{fb}^{-1}$, the FOM can just keep up with Belle I, but this is nowhere near the case when scaled to $350\,\text{fb}^{-1}$.

For channel $B^0 \to (K^{*0} \to K^+\pi^-)e^+e^-$, additional comparison for different BDT setups have been made. 2 q^2 bin 2 steps, which was eventually used, performed best at a FOM of 4.99, followed by 1 q^2 bin 2 steps with a FOM of 4.91, 2 bins 1 step at 4.86 and 1 bin 1 step at 4.4. The corresponding data can be found in table 6.6.

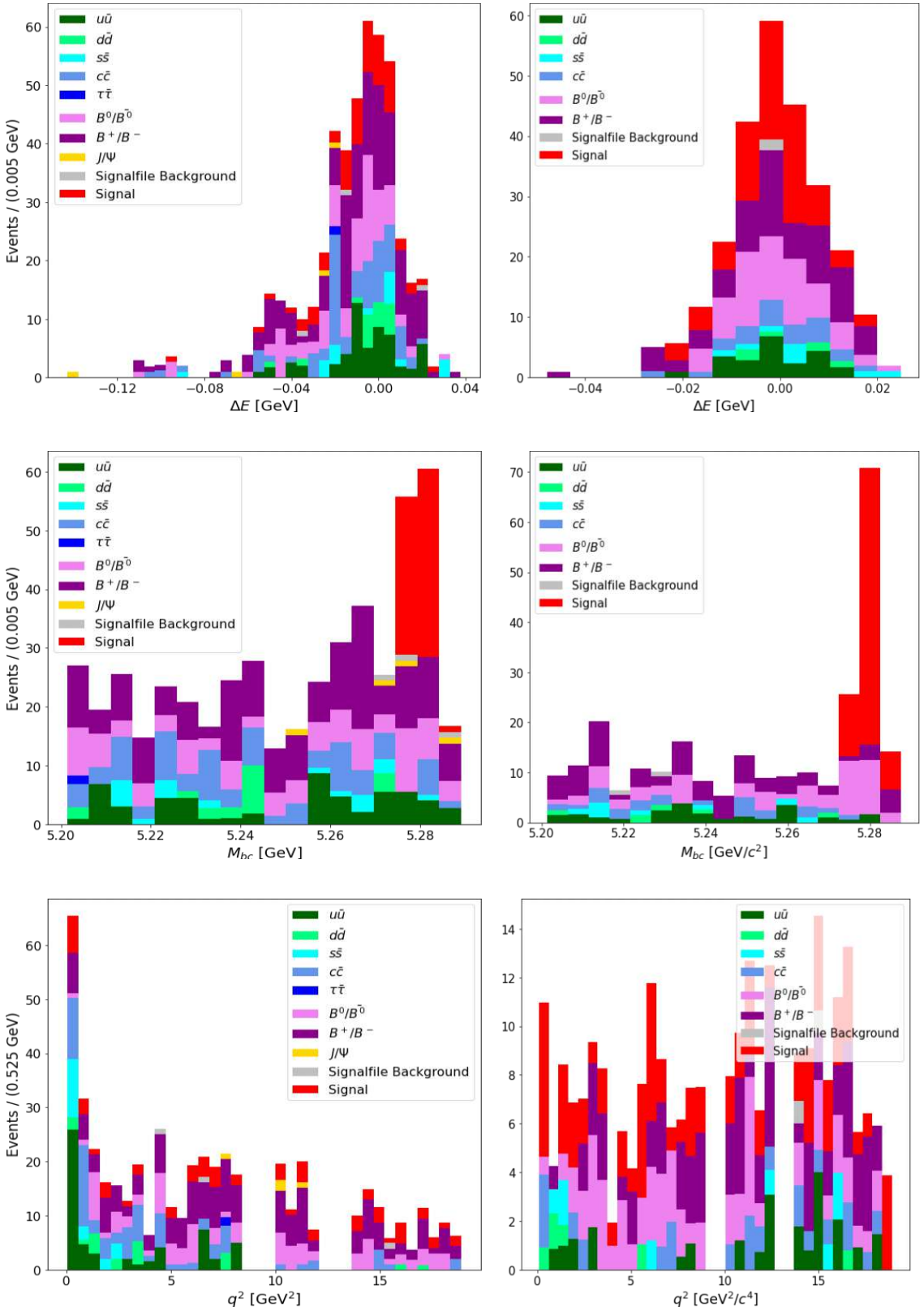

Figure 6.8: Plots of ΔE, M_{bc} and q^2 for the channels $B^0 \to (K^{*0} \to K^+\pi^-)e^+e^-$ (left side) and $B^0 \to (K^{*0} \to K^+\pi^-)\mu^+\mu^-$ (right side) after the BDTs and the removal of excess events.

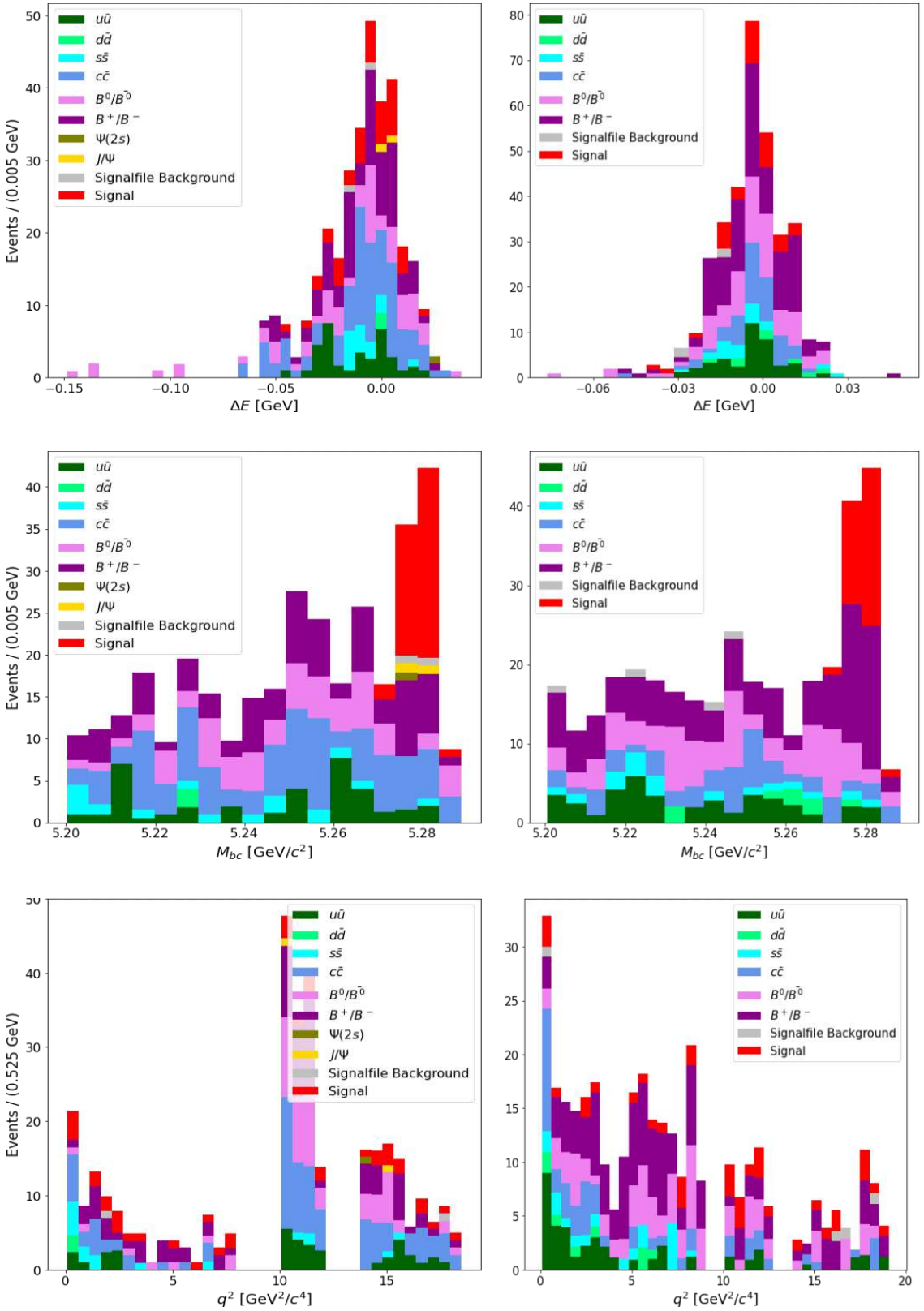

Figure 6.9: Plots of ΔE, M_{bc} and q^2 for the channels $B^+ \to (K^{*+} \to K^0_s\pi^+)e^+e^-$ (left side) and $B^+ \to (K^{*+} \to K^0_s\pi^+)\mu^+\mu^-$ (right side) after the BDTs and the removal of excess events.

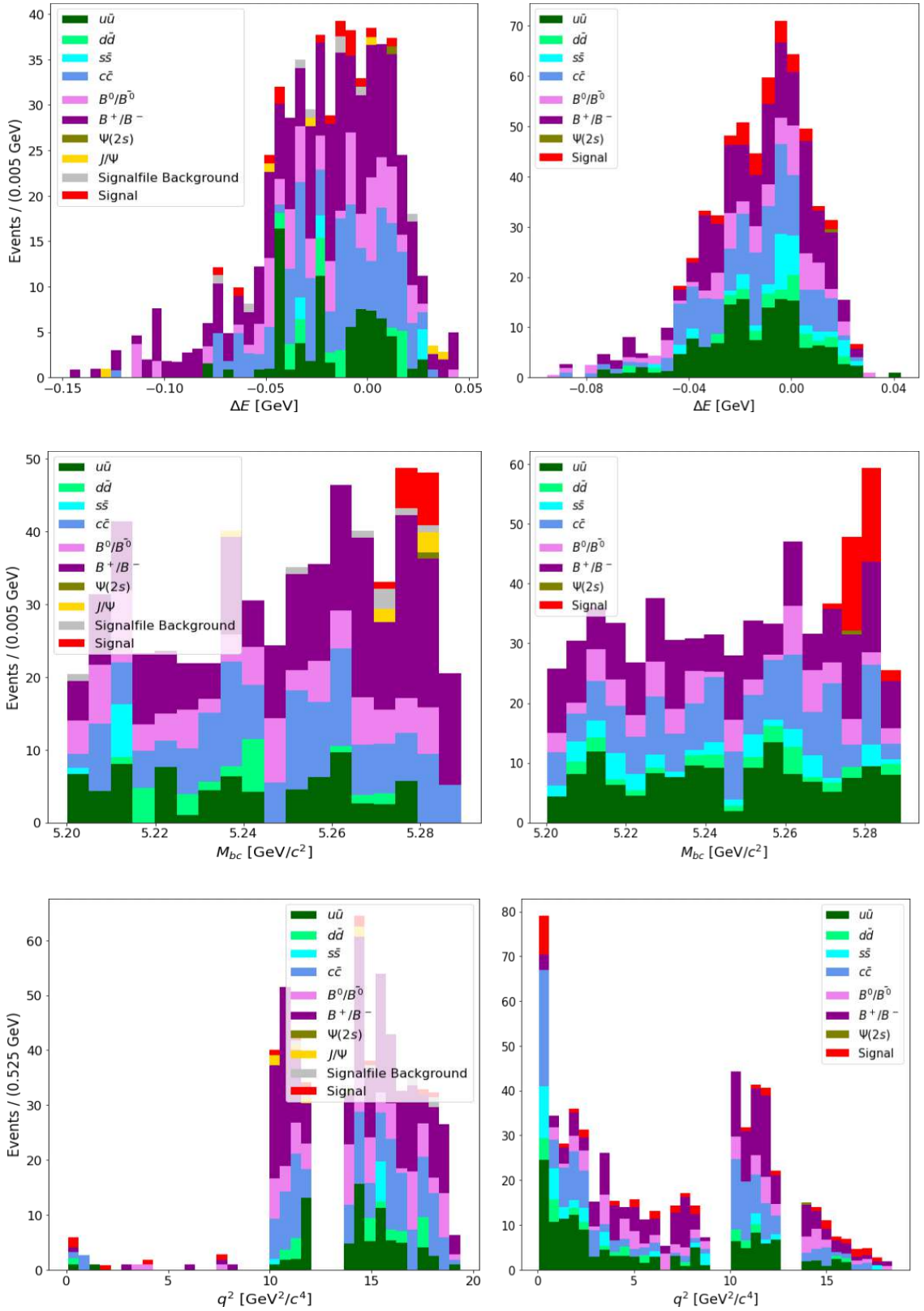

Figure 6.10: Plots of ΔE, M_{bc} and q^2 for the channels $B^+ \to (K^{*+} \to K^+\pi^0)e^+e^-$ (left side) and $B^+ \to (K^{*+} \to K^+\pi^0)\mu^+\mu^-$ (right side) after the BDTs and the removal of excess events.

	q_1^2	q_2^2	q_3^2	q_4^2	q_{full}^2
$N_{\text{sig, 1 ab}^{-1}}$	13.44	13.48	9.84	16.6	60.2
$N_{\text{sig, 711 fb}^{-1}}$	9.56	9.58	7.00	11.80	42.80
$N_{\text{sig, 350 fb}^{-1}}$	4.71	4.72	3.44	5.81	21.07
$N_{\text{sig, Belle I}}$	-	-	-	-	42
$N_{\text{bkg, 1 ab}^{-1}}$	11.65	2.92	50.53	100.83	179.38
$N_{\text{bkg, 711 fb}^{-1}}$	8.29	2.07	35.93	71.69	127.54
$N_{\text{bkg, 350 fb}^{-1}}$	4.08	1.02	17.69	35.29	62.78
$N_{\text{bkg, Belle I}}$	-	-	-	-	47
$\epsilon[\%]$	8.77 ± 2.08	4.31 ± 1.62	9.77 ± 2.52	10.23 ± 2.33	$4.67 \pm 0.62\%$
$\epsilon_{\text{Belle I}}[\%]$	4.712	5.08	2.823	2.37	2.94
$P[\%]$	29.02	30.39	44.87	48.48	38.68
$P_{\text{Belle I}}[\%]$	-	-	-	-	47.19
$\text{FOM}_{1\,\text{ab}^{-1}}$	1.98	2.02	2.10	2.84	4.83
$\text{FOM}_{711\,\text{fb}^{-1}}$	1.67	1.71	1.77	2.39	4.07
$\text{FOM}_{350\,\text{fb}^{-1}}$	1.17	1.20	1.24	1.68	2.85
$\text{FOM}_{\text{Belle I}}$	-	-	-	-	4.45

Table 6.2: Overview of the performance of the channel $B^0 \to K^{*0}e^+e^-$ after the BDT. The data is shown for the original $1\,\text{ab}^{-1}$, and also scaled down for ease of comparison with Belle I's results by Wehle et al., which are included where available.

More q^2 bins where not tested for the same data set with the same features, but it did not result in any significant improvement in previous tests.

	q_1^2	q_2^2	q_3^2	q_4^2	q_{full}^2
$N_{\text{sig, } 1\,\text{ab}^{-1}}$	16.32	17.82	10.35	23.19	75.21
$N_{\text{sig, } 711\,\text{fb}^{-1}}$	11.60	12.67	7.36	16.49	53.48
$N_{\text{sig, } 350\,\text{fb}^{-1}}$	5.71	6.24	3.62	8.12	26.32
$N_{\text{sig, Belle I}}$	-	-	-	-	46
$N_{\text{bkg, } 1\,\text{ab}^{-1}}$	9.16	13.15	1.98	10.96	39.0
$N_{\text{bkg, } 711\,\text{fb}^{-1}}$	6.51	9.35	1.41	7.80	27.73
$N_{\text{bkg, } 350\,\text{fb}^{-1}}$	3.21	4.60	0.69	3.84	13.65
$N_{\text{bkg, Belle I}}$	-	-	-	-	21
$\epsilon[\%]$	10.95 ± 2.56	14.37 ± 3.15	9.95 ± 2.93	16.92 ± 3.2	11.33 ± 1.23
$\epsilon_{\text{Belle I}}[\%]$	2.56	3.15	2.93	3.20	1.23
$P[\%]$	64.04	57.54	83.93	67.69	65.85
$P_{\text{Belle I}}[\%]$	-	-	-	-	68.66
$\text{FOM}_{1\,\text{ab}^{-1}}$	3.23	3.20	2.95	3.97	7.04
$\text{FOM}_{711\,\text{fb}^{-1}}$	2.73	2.70	2.48	3.35	5.93
$\text{FOM}_{350\,\text{fb}^{-1}}$	1.91	1.89	1.74	2.35	4.16
$\text{FOM}_{\text{Belle I}}$	-	-	-	-	5.62

Table 6.3: Overview of the performance of the channel $B^0 \to K^{*0}\mu^+\mu^-$ after the BDT. The data is shown for the original $1\,\text{ab}^{-1}$, and also scaled down for ease of comparison with Belle I's results by Wehle et al., which are included where available.

	q_1^2	q_2^2	q_3^2	q_4^2	q_{full}^2
$N_{\text{sig, } 1\,\text{ab}^{-1}}$	16.22	6.76	13.59	17.29	53.86
$N_{\text{sig, } 711\,\text{fb}^{-1}}$	11.53	4.81	9.66	12.29	38.30
$N_{\text{sig, } 350\,\text{fb}^{-1}}$	5.68	2.37	4.76	6.05	18.85
$N_{\text{sig, Belle I}}$	-	-	-	-	35
$N_{\text{bkg, } 1\,\text{ab}^{-1}}$	11.65	2.92	50.53	100.83	179.38
$N_{\text{bkg, } 711\,\text{fb}^{-1}}$	8.29	2.07	35.93	71.69	127.54
$N_{\text{bkg, } 350\,\text{fb}^{-1}}$	4.08	1.02	17.69	35.29	62.78
$N_{\text{bkg, Belle I}}$	-	-	-	-	176
$\epsilon[\%]$	8.77 ± 2.08	4.31 ± 1.62	9.77 ± 2.52	10.23 ± 2.33	4.67 ± 0.62
$\epsilon_{\text{Belle I}}[\%]$	4.712	5.08	2.823	2.437	2.94
$P[\%]$	58.19	69.87	21.19	14.64	23.09
$P_{\text{Belle I}}[\%]$	-	-	-	-	16.59
$\text{FOM}_{1\,\text{ab}^{-1}}$	3.07	2.17	1.70	1.59	3.53
$\text{FOM}_{711\,\text{fb}^{-1}}$	2.59	1.83	1.43	1.34	2.97
$\text{FOM}_{350\,\text{fb}^{-1}}$	1.82	1.29	1.00	0.94	2.09
$\text{FOM}_{\text{Belle I}}$	-	-	-	-	2.41

Table 6.4: Overview of the performance of the channel $B^+ \to K^{*+}e^+e^-$ after the BDT. The data is shown for the original $1\,\text{ab}^{-1}$, and also scaled down for ease of comparison with Belle I's results by Wehle et al., which are included where available.

	q_1^2	q_2^2	q_3^2	q_4^2	q_{full}^2
$N_{\text{sig, } 1\,\text{ab}^{-1}}$	16.01	12.57	15.07	16.49	69.14
$N_{\text{sig, } 711\,\text{fb}^{-1}}$	11.38	8.94	10.71	11.73	49.16
$N_{\text{sig, } 350\,\text{fb}^{-1}}$	5.60	4.40	5.27	5.77	24.20
$N_{\text{sig, Belle I}}$	-	-	-	-	26
$N_{\text{bkg, } 1\,\text{ab}^{-1}}$	70.58	53.85	35.22	20.13	203.91
$N_{\text{bkg, } 711\,\text{fb}^{-1}}$	50.19	38.29	25.04	14.31	144.98
$N_{\text{bkg, } 350\,\text{fb}^{-1}}$	24.70	18.85	12.33	7.5	71.37
$N_{\text{bkg, Belle I}}$	-	-	-	-	199
$\epsilon[\%]$	9.47 ± 2.25	8.61 ± 2.32	10.69 ± 2.6	10.24 ± 2.39	7.84 ± 0.91
$\epsilon_{\text{Belle I}}[\%]$	3.658	5.16	5.78	3.52	2.94
$P[\%]$	18.49	18.93	29.96	45.03	25.32
$P_{\text{Belle I}}[\%]$	-	-	-	-	38.01
$\text{FOM}_{1\,\text{ab}^{-1}}$	1.72	1.54	2.12	2.73	4.18
$\text{FOM}_{711\,\text{fb}^{-1}}$	1.45	1.30	1.79	2.30	3.53
$\text{FOM}_{350\,\text{fb}^{-1}}$	1.02	0.91	1.26	1.61	2.48
$\text{FOM}_{\text{Belle I}}$	-	-	-	-	2.14

Table 6.5: Overview of the performance of the channel $B^+ \to K^{*+}\mu^+\mu^-$ after the BDT. The data is shown for the original $1\,\text{ab}^{-1}$, and also scaled down for ease of comparison with Belle I's results by Wehle et al., which are included where available.

	2 bins, 2 steps	1 bin, 2 steps	2 bins, 1 step	1 bin, 1 step
N_{signal}	58.32	45.66	56.60	58.27
$N_{\text{background}}$	77.96	40.64	78.88	114.62
ϵ	4.37%	6.39%	7.92%	8.15%
P	47.19%	52.91%	41.77%	33.70%
FOM	5.0	4.91	4.86	4.43

Table 6.6: Comparisons between different BDT training setups, in order of performance from best to worst, for the channel $B^0 \to (K^{*0} \to K^+\pi^-)e^+e^-$. When optimizing for maximum FOM, 2 bins 2 steps performs best. Attempts with more bins did not result in any further improvement.

6.9. Signal yields

In this chapter, a fitting procedure is developed in order to model signal and background shapes on M_{bc} and extract the BR's for the four $B \to K^*\ell^+\ell^-$ channels from the area below the signal curve. In principle, one could of course also calculate the BR's with the information gained in the last chapter, assuming that the ratio between N_{signal} and $N_{\text{background}}$ stays the same on real data. This is a fairly strong assumption though, especially due to the low statistics available in this study, and other systematic errors might come into play as well. Modeling the shapes on the other hand allows for floating curve parameters, which can to some extent accommodate for these problems, as will be shown below. Furthermore, the models can be used for statistical checks (see section 6.10).

The fitting procedure is conducted using the packages *iMinuit* and *probfit* [26]. Signal is modeled using a crystalball (CB) function, whose central part around the mean $\bar{x}$ is of a gaussian form, with the standard deviation σ. On the left side, the gaussian transitions into a power law with the free parameter n. The transition point is determined by the parameter α. Signal-like background is also modeled with the help of a CB function, while the continuum and combinatorial backgrounds are modeled with an argus function. The free parameters of the argus function are the kinematic limit of the invariant mass distribution c, the power p and the curvature χ.

The first step is to find the parameters of the signal function by fitting the CB to a pure signal sample containing 200 000 generated events. This sample is also independent of the training data, just like the validation set. Using such a large file instead of the regular $1\,\text{ab}^{-1}$ set allows to model the signal shape much more accurately. In this fit, all four parameters μ, σ, n and α are left floating. The second step is the determination of the parameters for the two background curves. Here, MC truth is used to eliminate all signal from the post-BDT data sets gained last chapter, resulting in a pure background sample. The signal-like CB is fixed to the parameters gained in the true signal fit, apart from σ and the normalization factor $N_{\text{peaking background}}$. For the argus, c is another well known parameter and set to 5.29, while p is set to 0.5 using trial and error such that the fit converges. This leaves only χ and N_{argus} floating, resulting in a fit with four free parameters in total. The normalization factors $N_{\text{peaking background}}$ and N_{argus} correspond to the number of

particles below the respective curve. The final fit is then performed on the full data sample, using the parameters determined in the last two fits. An identical fit with the same parameters will eventually also be applied on real data. The free parameters are the total background $N_{\text{background}} = N_{\text{argus}} + N_{\text{peaking background}}$, χ, N_{signal} and σ_{signal}. These floating parameters accommodate for possible differences between MC and real data. The resulting plots can be seen in figures 6.11 to 6.14. As a cross-check, linear pearson correlations were calculated for the parameters of each individual fit to ensure that no parameter is redundant and everything is working properly. No serious correlations ($>\sim 0.8$) were found. For the correlation tables, refer to the appendix in section B.

The BR's of the four $B \to K^* \ell^+ \ell^-$ decays can then be calculated using

$$\text{BR} = \frac{N_{\text{signal}}}{2\frac{2}{3}\epsilon_{200\,000} I \sigma_B}, \tag{6.26}$$

where N_{signal} is retrieved from the final fit, $\epsilon_{200\,000}$ is the efficiency calculated from the 200 000 events file, I is the integrated luminosity and σ the cross section of either B^0 or B^+. The factor 1/2 results from the 2 B mesons in $\Upsilon(4S) \to B\bar{B}$, and the factor $\frac{1}{2/3}$ takes the not reconstructed subdecays of K^* into account (see chapter 6.3). The error is calculated using

$$\text{BR}_{\text{error}} = \frac{N_{\text{signal, error}}}{2\frac{2}{3}\epsilon_{200\,000} I \sigma_B}, \tag{6.27}$$

where $N_{\text{signal, error}}$ is the statistical error from the fitting procedure. Further systematic errors are not taken into account at this point of the study. Table 6.7 shows the thereby gained BR's compared with the reference values given by PDG, which are also the values that were used to generate the MC data in the first place. Therefore, the good agreement comes as no surprise, but it is a good cross-check nevertheless. While it would be possible to apply the machinery presented to real data as-is, this is not wanted before the whole angular analysis is finished. Doing so might lead to information from real data subconsciously leaking into the analysis, introducing biases, as already mentioned.

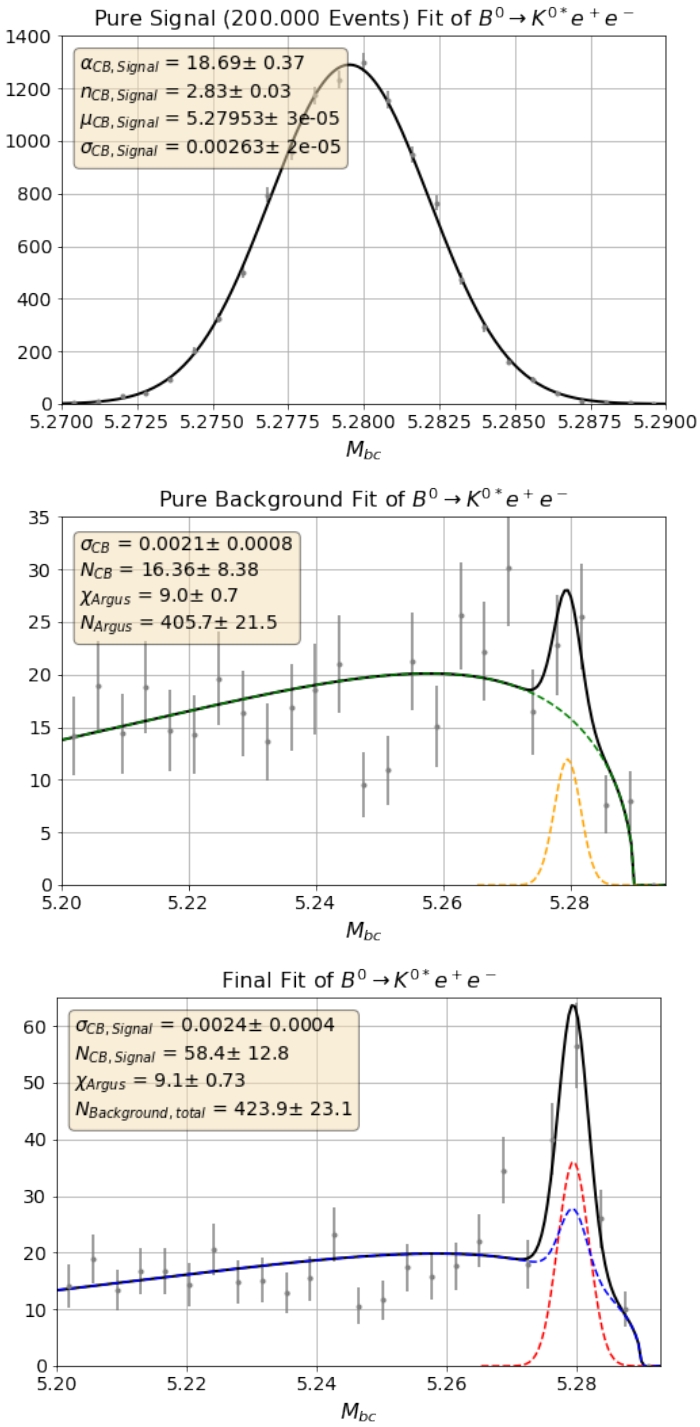

Figure 6.11: Fit plots of $B^0 \to K^{*0} e^+ e^-$. The yellow box in each plot shows the floating parameters and corresponding values of each fit. The first two fits are applied on pure signal and background respectively to find the parameters of the curves, while the last fit is applied on the full data.

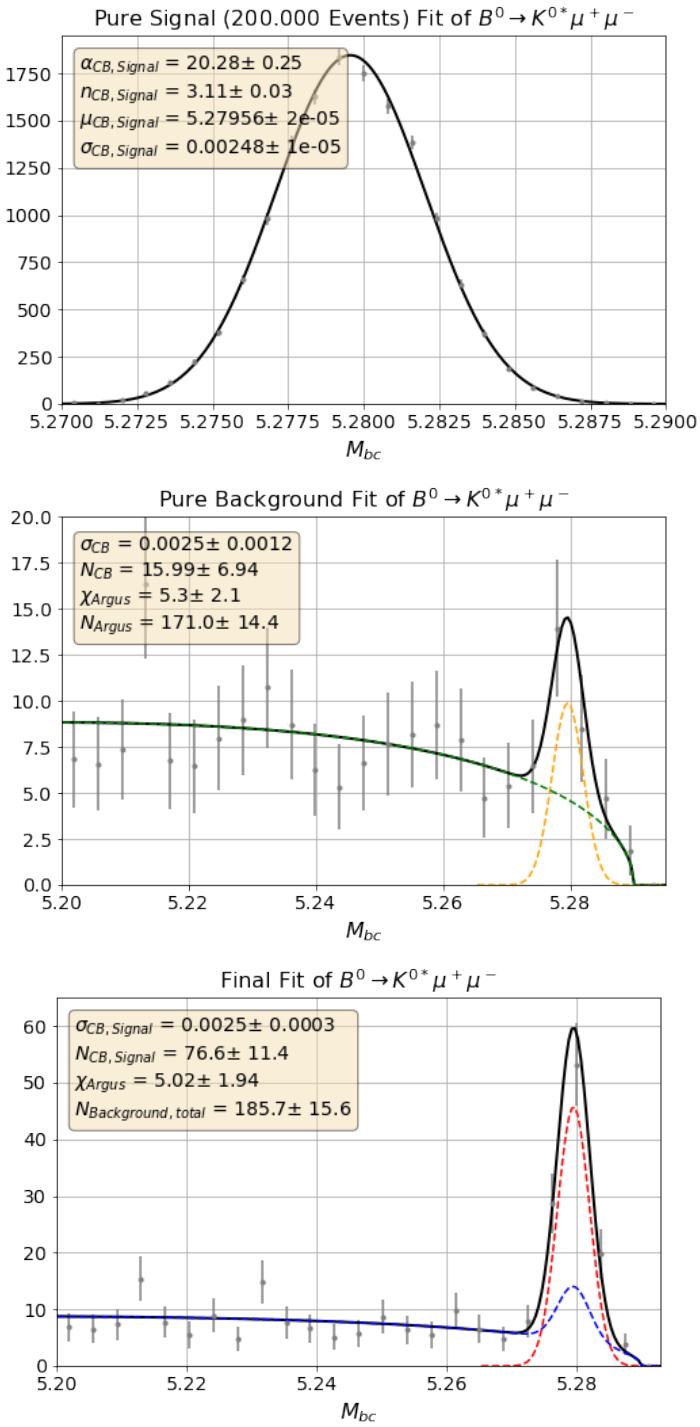

Figure 6.12: Fit plots of $B^0 \to K^{*0}\mu^+\mu^-$. The yellow box in each plot shows the floating parameters and corresponding values of each fit. The first two fits are applied on pure signal and background respectively to find the parameters of the curves, while the last fit is applied on the full data.

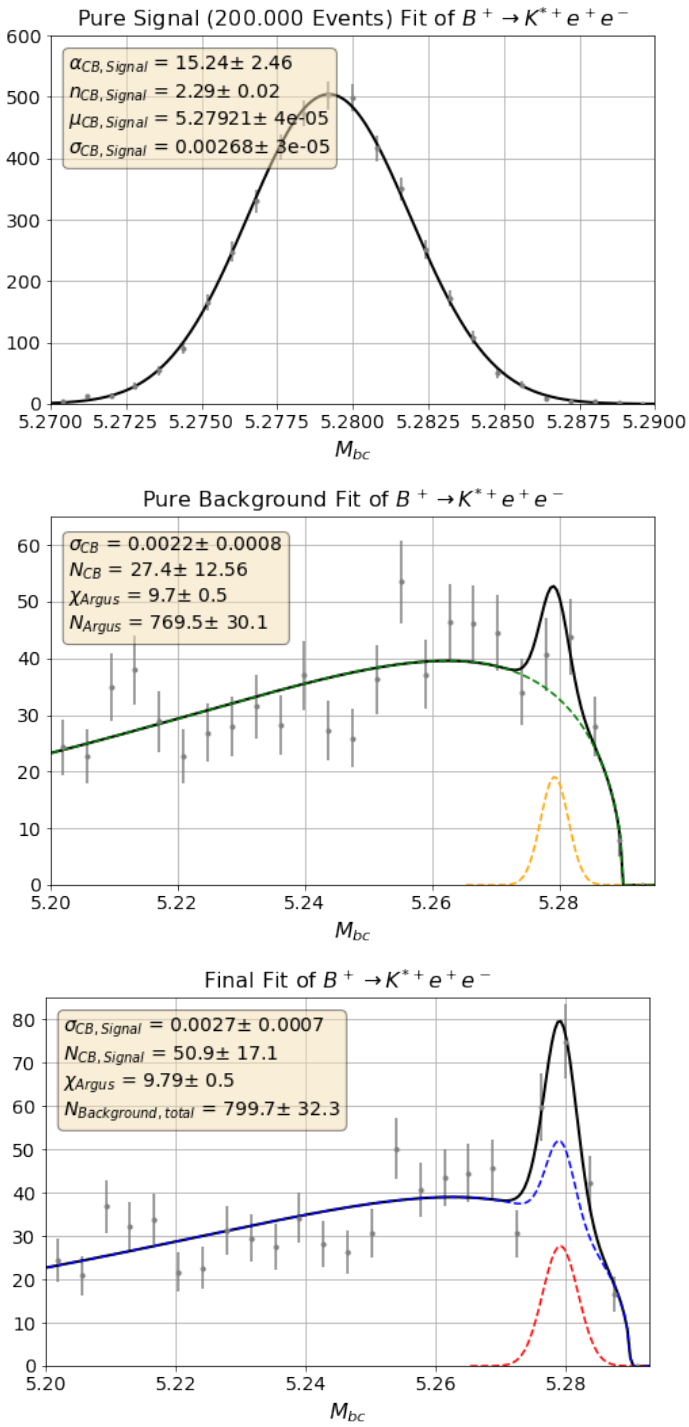

Figure 6.13: Fit plots of $B^+ \to K^{*+}e^+e^-$. The yellow box in each plot shows the floating parameters and corresponding values of each fit. The first two fits are applied on pure signal and background respectively to find the parameters of the curves, while the last fit is applied on the full data.

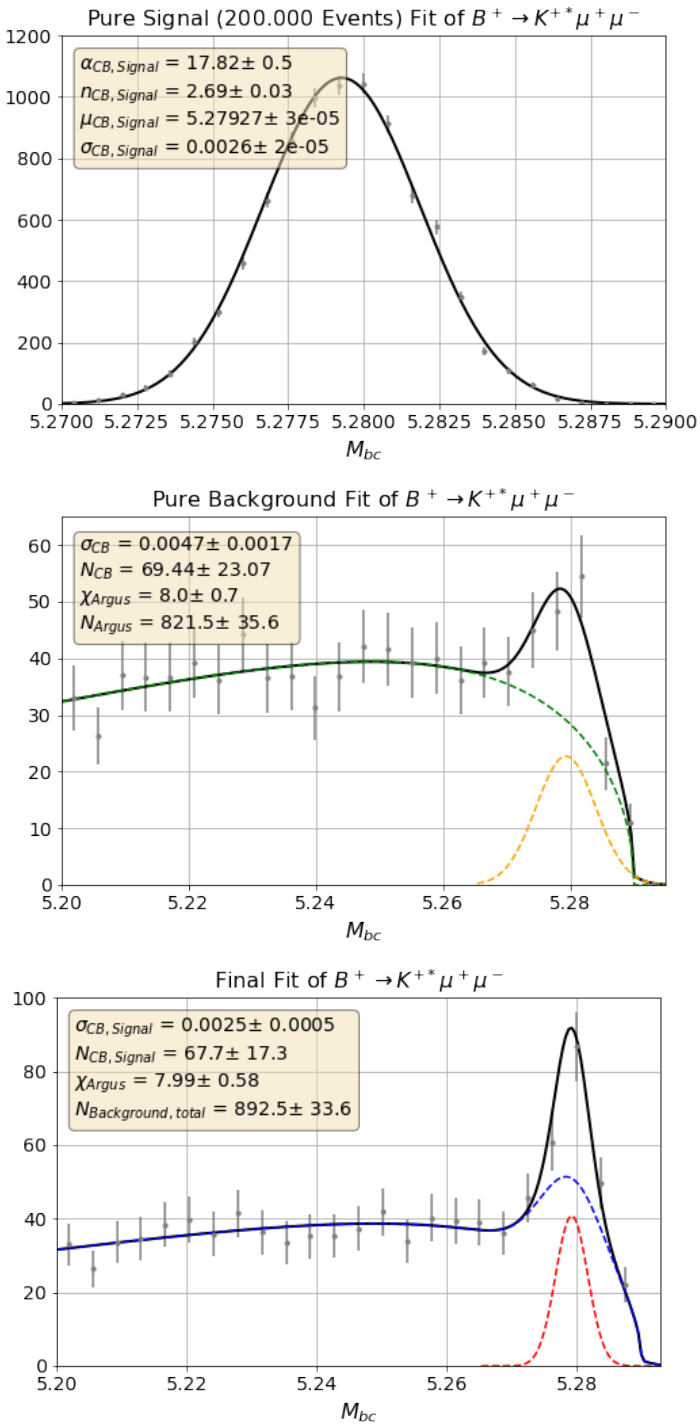

Figure 6.14: Fit plots of $B^+ \to K^{*+}\mu^+\mu^-$. The yellow box in each plot shows the floating parameters and corresponding values of each fit. The first two fits are applied on pure signal and background respectively to find the parameters of the curves, while the last fit is applied on the full data.

	$B^0 \to K^{*0}e^+e^-$	$B^0 \to K^{*0}\mu^+\mu^-$	$B^+ \to K^{*+}e^+e^-$	$B^+ \to K^{*+}\mu^+\mu^-$
BR$[10^{-6}]$	1.03±0.23	1.0±0.15	1.4±0.47	0.91±0.23
$\text{BR}_{\text{PDG}}[10^{-6}]$	$1.03^{+0.19}_{-0.17}$	0.94±0.5	$1.55^{+0.40}_{-0.31}$	0.96±1.0
N_{sig}	58.3±12.8	76.6±11.4	50.9±17.1	67.7±17.3
$N_{\text{sig, true}}$	60.2	75.2	53.9	69.1
$N_{\text{bkg, full}}$	423.9±23.1	185.7±15.6	799.7±32.3	892.6±33.6
$N_{\text{bkg, sig region}}$	93.9±5.9	37.6±4.3	186.4±8.9	206.1±8.4
$\epsilon[\%]$	8.2±1.8	11.5±1.7	4.4±1.5	7.7± 2.0
$P[\%]$	38.3	67.0	21.5	24.7
FOM	4.73	7.17	3.31	4.09

Table 6.7: An overview of the results from the fitting procedure for the four $B \to K\ell^+\ell^-$ channels and a comparison to the true generated signal values and the PDG BR's that were used to generate the data. The results are in good agreement.

6.10. Toy study

As briefly mentioned last chapter, it is to be expected that MC and real data will not correspond to each other perfectly, be it due to statistical fluctuations or other biases. Therefore it is important to ensure that the fitting procedure can deal with such differences at least to some degree. The stability of the fitter with respect to statistical fluctuations can best be checked by conducting a so-called toy study, which is the content of this chapter and concludes the current state of the analysis.

10 000 toy data samples per $B \to K^* \ell^+ \ell^-$ channel are generated from the models gained in the fitting procedure, using the same parameters. The values for N_{signal}, N_{argus} and $N_{\text{peaking background}}$ are fluctuated around their original value following a poisson distribution. The final fit with the floating parameters N_{signal}, $N_{\text{background}}$, σ_{signal} and χ_{argus}, is then applied on every single toy data sample and the BR is calculated for every single fit. To check whether the fitter is stable and not biased, pull distributions are calculated for all parameters floating in the fit and the BR, with the pull of a variable x being defined as

$$Pull(x) = \frac{x_{\text{fit}} - x_{\text{true}}}{x_{\text{fit, error}}}, \tag{6.28}$$

where x_{fit} is the value gained in the fit of an individual toy sample, $x_{\text{fit, error}}$ the statistical uncertainty of the fit, and x_{true} the value of the model the sample was generated from. Assuming that the fluctuations lead to no biases in the fitter and that the uncertainty is estimated correctly, the pull distribution will have a perfect gaussian shape in the limit of infinite data points, with a mean of 0 and a width of 1. Deviations in the mean indicate biases, whilst deviations in the standard deviation result from an either overestimated or underestimated statistical error. In figures 6.15 to 6.18, all pull plots for all channels are shown. N_{signal}, $N_{\text{background}}$, χ and BR show very good agreement with the expected values for all channels. A discrepancy of ~ 0.1 in the mean can be observed for σ_{signal} though, as well for all channels. This systematic uncertainty can be attributed to the fact that N_{argus} and $N_{\text{peaking background}}$ are fluctuated independently, but the only free parameters in the fit are $N_{\text{background}}$ and χ. Allowing more background parameters such as $N_{\text{peaking background}}$ to float would not work though because then the fitter could not distinguish peaking background from signal anymore. The mean of the BR plots agree very well with the BR's calculated in section 6.9, as they should.

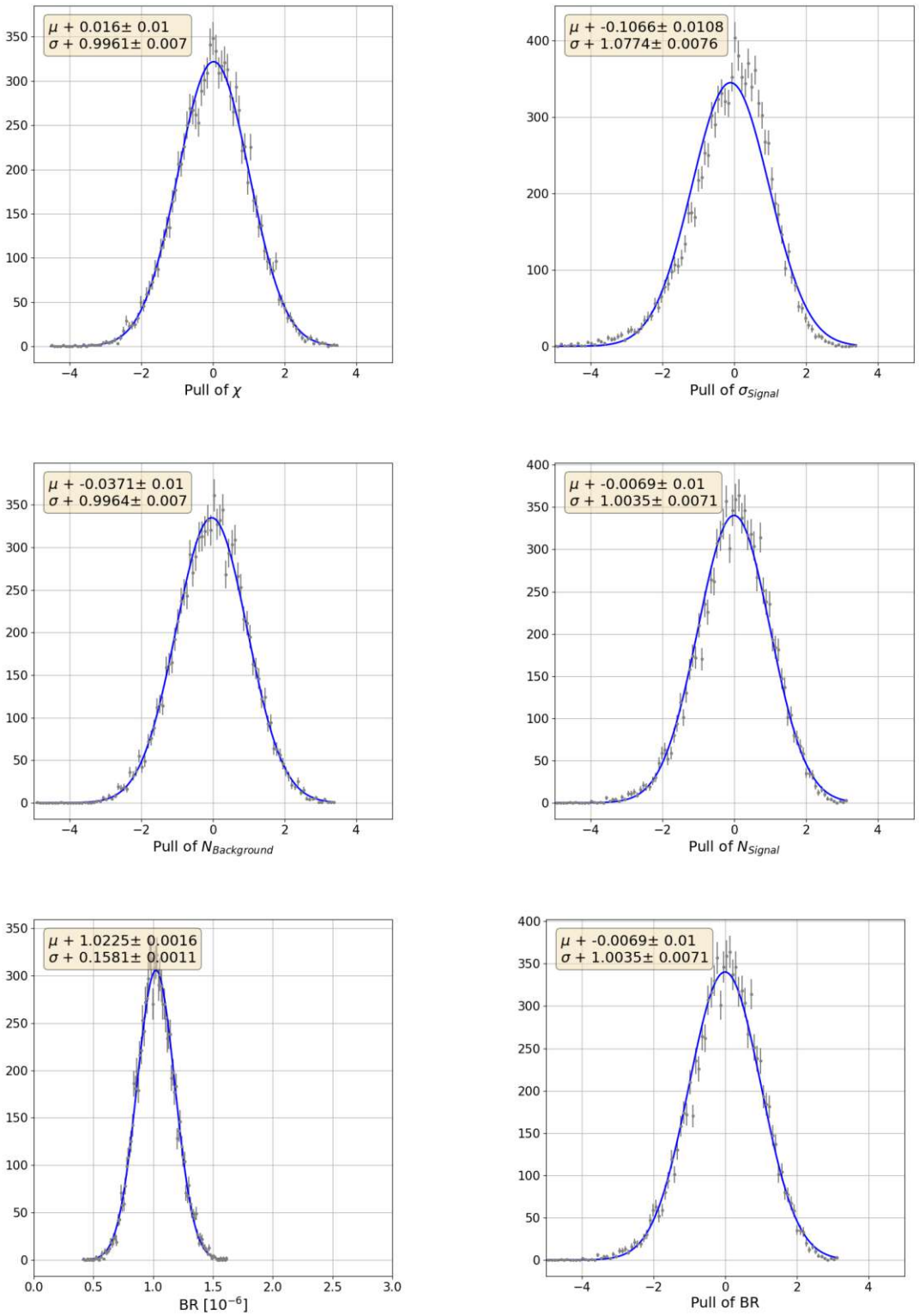

Figure 6.15: Pulls plots for the decay $B^0 \to K^{*0}e^+e^-$. The first four plots show the pulls of all four floating variables in the fit. Bottom left shows the BR distribution, and bottom right the corresponding pull distribution.

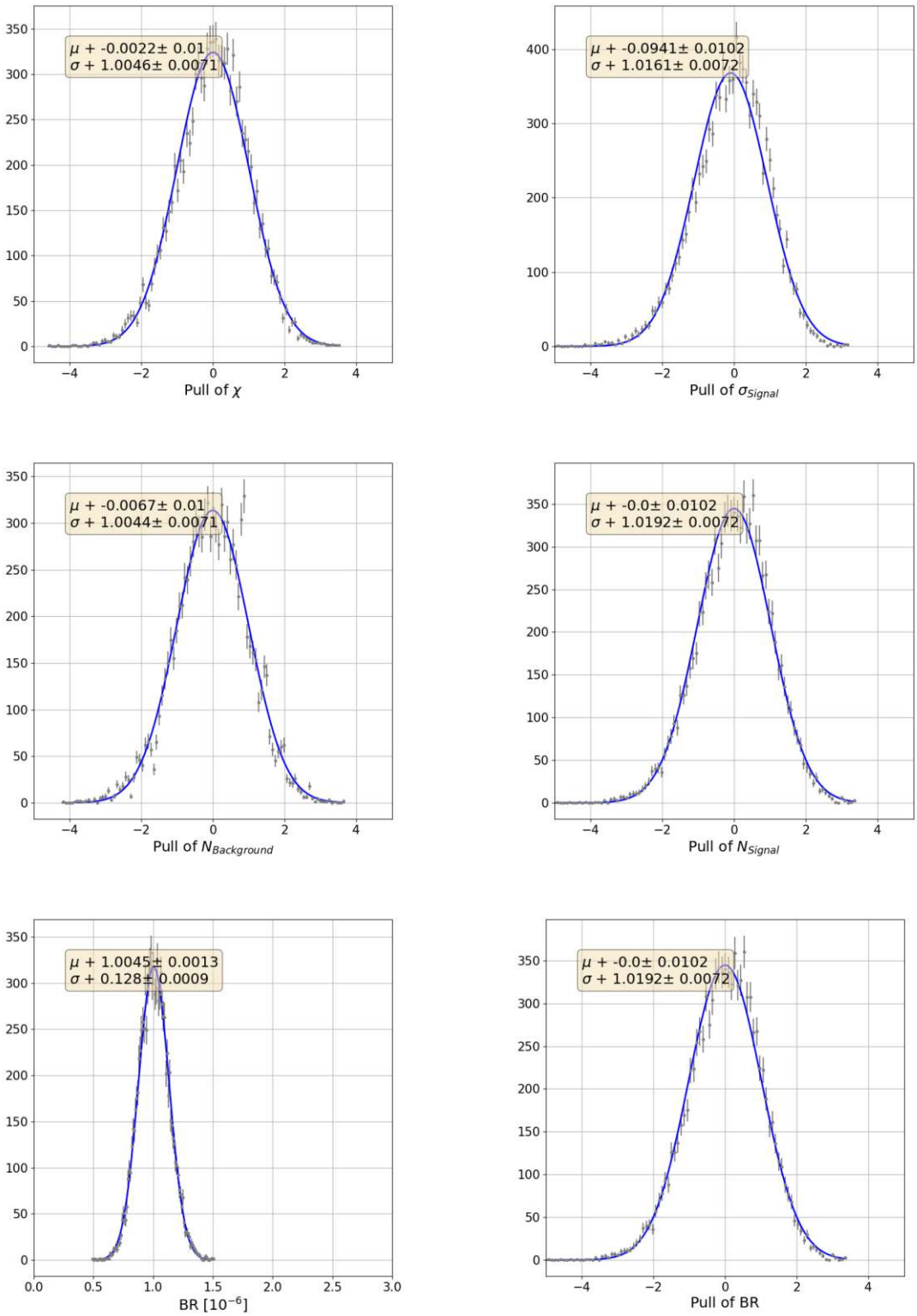

Figure 6.16: Pulls plots for the decay $B^0 \to K^{*0}\mu^+\mu^-$. The first four plots show the pulls of all four floating variables in the fit. Bottom left shows the BR distribution, and bottom right the corresponding pull distribution.

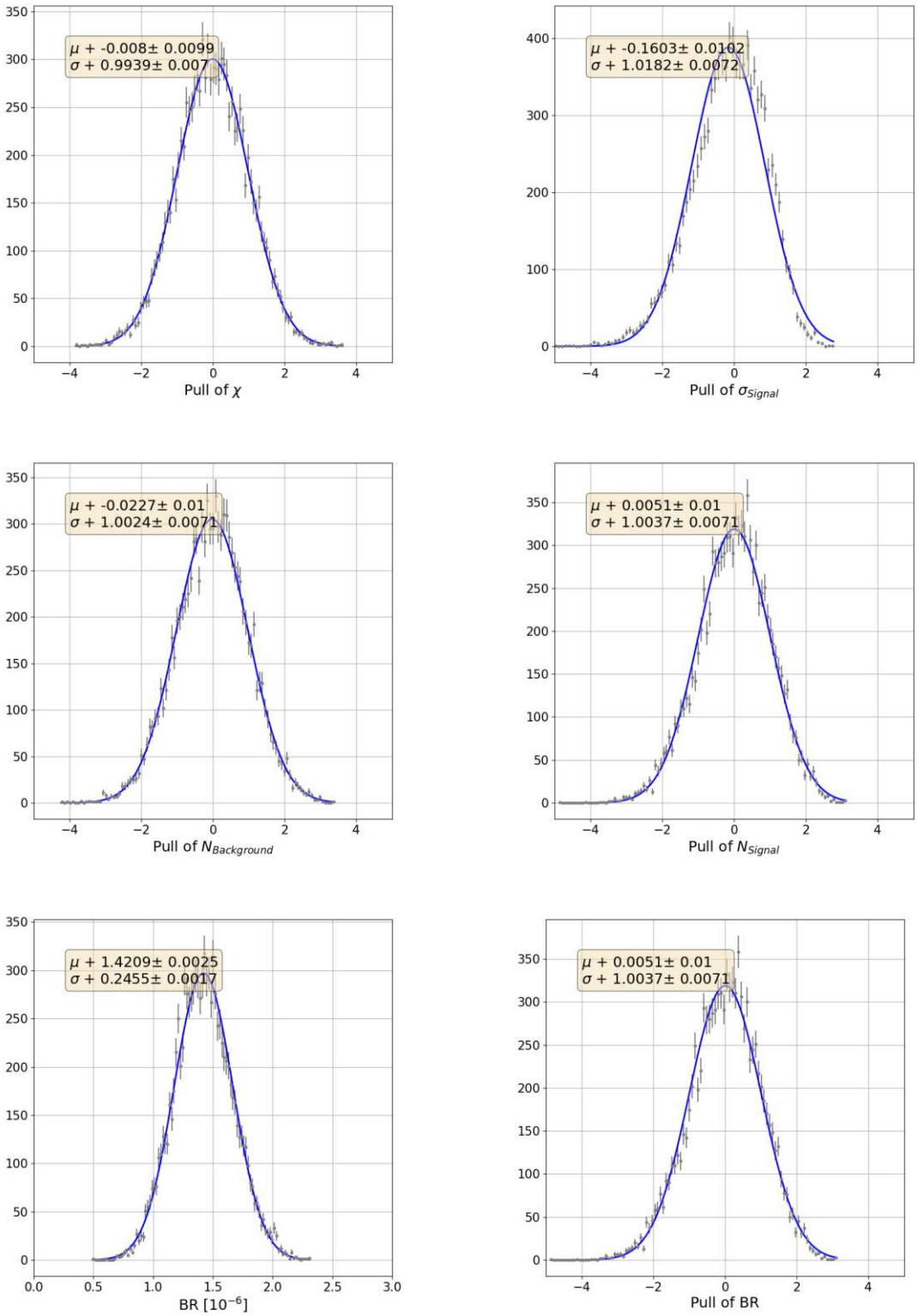

Figure 6.17: Pulls plots for the decay $B^+ \to K^{*+}e^+e^-$. The first four plots show the pulls of all four floating variables in the fit. Bottom left shows the BR distribution, and bottom right the corresponding pull distribution.

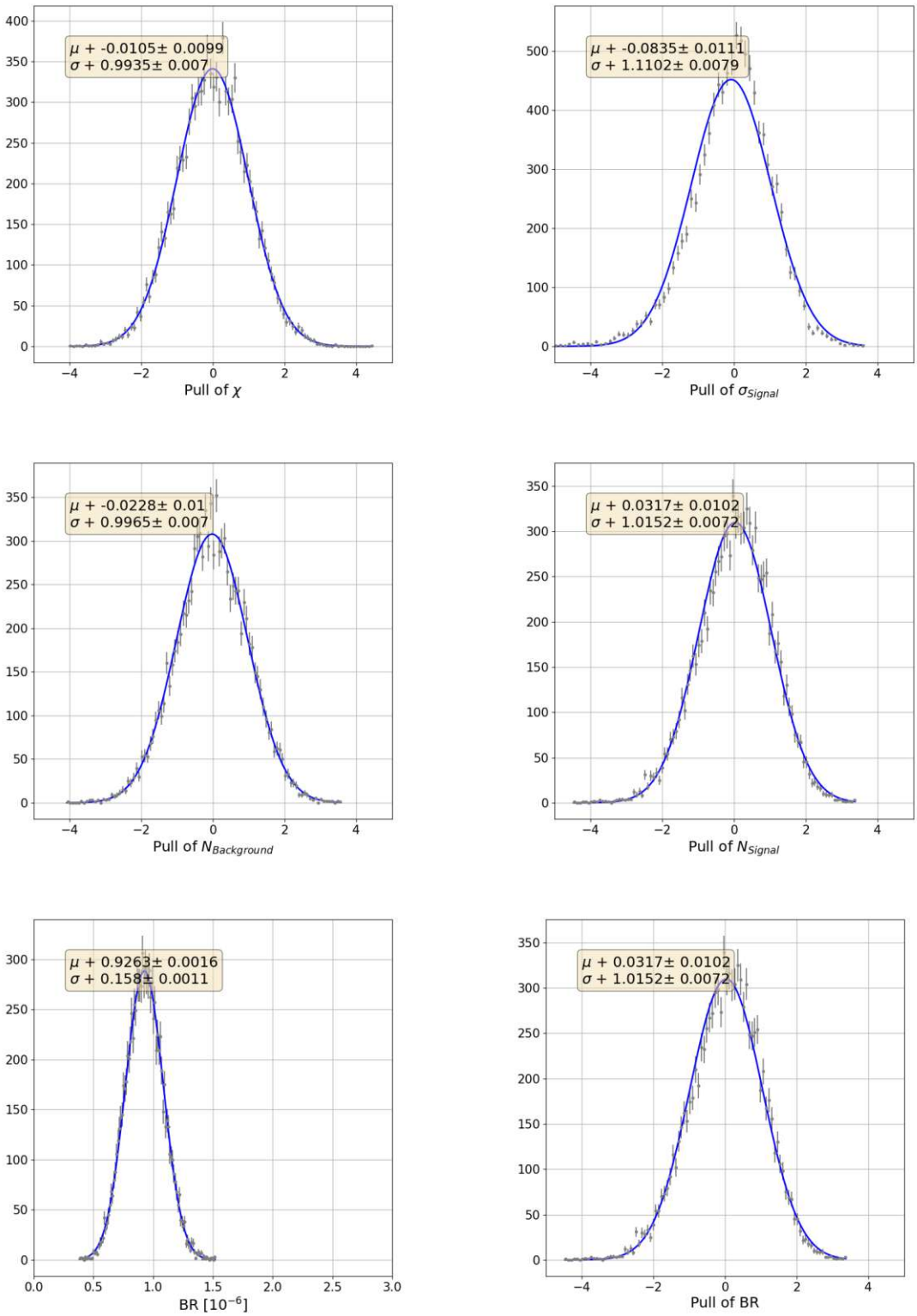

Figure 6.18: Pulls plots for the decay $B^+ \to K^{*+}\mu^+\mu^-$. The first four plots show the pulls of all four floating variables in the fit. Bottom left shows the BR distribution, and bottom right the corresponding pull distribution.

6.11. Discussion and outlook

With the toy study, all major steps performed in the analysis so far have been presented. Should it be decided to continue the analysis as planned, the next steps are further cross checks using the vetoed J/Ψ and $\Psi(2S)$ decays. These channels have orders of magnitude more events than the signal of interest while having very similar properties, allowing the conduction of further statistical tests and sanity checks. For example, one could test the accuracy of the efficiency estimation, since the BR's of these decays are known at a much higher precision due to their statistical abundance than the BR's of the channels that are the prime subject of this analysis. Furthermore, one can use J/Ψ and $\Psi(2S)$ to achieve a first comparison between MC and real data and see if everything works out as expected, without needing to unblind the regions of data that will be used to extract P_5'. After that, a $3D$ fitting procedure needs to be developed for the angular variables θ_ℓ, θ_k and ϕ, similarly to the $1D$ M_{bc} fit in this thesis, which is why correlations with those three variables have already been dealt with. This $3D$ fit will then be used to extract the angular observables of interest using the formulas presented in section 2.3. Last but not least, a thorough study of the systematic errors needs to be done, including further toy studies for the $3D$ fit such as the one in this thesis.

To evaluate the current performance of the analysis, it is best to compare the results to the Belle I analysis by Wehle et al. Unfortunately, for the signal yields gained by the M_{bc} fit no suitable Belle I data (peaking backgrounds were not taken into account on their part, thereby skewing the yields and rendering comparisons useless) is available, but the MC information post-BDT gives a clear enough picture already. As shown in chapter 6.8 and tables 6.2 to 6.5, the FOM of this analysis for the B^0 channels is quite similar to Belle I when scaling to the same integrated luminosity, and much better for B^+. Unfortunately, COVID-related problems lead to major delays in data taking, resulting in only about $350\,\mathrm{fb}^{-1}$ available right now, with a maintenance shutdown of the collider lasting more than half a year starting at the time of writing. This is not even half the data Belle I had available for their analysis, which necessitated a bigger improvement than was possible over Belle I with respect to reconstruction and BDT performance in order to produce competitive results without waiting for many years. Therefore, when scaling the sample down to $350\,\mathrm{fb}^{-1}$, the FOM for B^+ can only just compete with Belle I's results, which is,

while still a very promising result long-term, not what would have been needed for a successful short- to mid-term P_5' measurement with the current situation being what it is. The B^0 is even farther away from that.

The possibly last option remaining to improve on Belle I's result within the next ~ 2 years would be to combine the currently available $\sim 350\,\mathrm{fb}^{-1}$ Belle II data with the $711\,\mathrm{fb}^{-1}$ Belle I data, resulting in $1061\,\mathrm{fb}^{-1}$ in total. Such an endeavor would be quite complex though, both with respect to technical details and properly assessing systematic errors at the end of the study.

References

[1] S. Wehle, "Angular analysis of $B \to K^* \ell\ell$ and search for $B^+ \to K^+ \tau\tau$ at the Belle experiment", Universität Hamburg, Diss., 2016, Dr. Universität Hamburg, Hamburg, 2016, p. 209. DOI: 10.3204/PUBDB-2016-03770. [Online]. Available: https://bib-pubdb1.desy.de/record/308592.

[2] R. Aaij, C. Abellán Beteta, B. Adeva, M. Adinolfi, A. Affolder, Z. Ajaltouni, S. Akar, J. Albrecht, F. Alessio, and et al., "Measurement of CP-averaged observables in the $B^0 \to K^{*0}\mu^+\mu^-$ decay", *Physical Review Letters*, vol. 125, no. 1, Jun. 2020. DOI: 10.1103/physrevlett.125.011802. [Online]. Available: https://doi.org/10.1103/PhysRevLett.125.011802.

[3] M. Thomson, *Modern particle physics*. New York: Cambridge University Press, 2013, ISBN: 978-1-107-03426-6.

[4] P. Zyla *et al.*, "Review of particle physics", *PTEP*, vol. 2020, no. 8, p. 083C01, 2020. DOI: 10.1093/ptep/ptaa104.

[5] W. Altmannshofer, P. Ball, A. Bharucha, A. J. Buras, D. M. Straub, and M. Wick, "Symmetries and asymmetries of $B \to K^* \mu^+ \mu^-$ decays in the Standard Model and beyond", *Journal of High Energy Physics*, vol. 2009, no. 01, pp. 019–019, Jan. 2009. DOI: 10.1088/1126-6708/2009/01/019. [Online]. Available: https://doi.org/10.1088/1126-6708/2009/01/019.

[6] R. Aaij, C. Abellán Beteta, B. Adeva, M. Adinolfi, A. Affolder, Z. Ajaltouni, S. Akar, J. Albrecht, F. Alessio, and et al., "Angular analysis of the $B^0 \to K^{*0}\mu^+\mu^-$ decay using 3 fb^{-1} of integrated luminosity", *Journal of High Energy Physics*, vol. 2016, no. 2, Feb. 2016, ISSN: 1029-8479. DOI: 10.1007/jhep02(2016)104. [Online]. Available: http://dx.doi.org/10.1007/JHEP02(2016)104.

[7] R. Aaij, C. A. Beteta, B. Adeva, M. Adinolfi, C. Adrover, A. Affolder, Z. Ajaltouni, J. Albrecht, F. Alessio, and et al., "Differential branching fraction and angular analysis of the decay $B^0 \to K^{*0}\mu^+\mu^-$ decay using 3 fb^{-1}", *Journal of High Energy Physics*, vol. 2013, no. 8, Aug. 2013, ISSN: 1029-8479. DOI: 10.1007/jhep08(2013)131. [Online]. Available: http://dx.doi.org/10.1007/JHEP08(2013)131.

[8] R. Aaij, C. A. Beteta, T. Ackernley, B. Adeva, M. Adinolfi, H. Afsharnia, C. A. Aidala, and A. et. al., "Angular analysis of the $B^+K^{*+}\mu^-\mu^-$ decay", *Phys. Rev. Lett.*, vol. 126, p. 161 802, 16 Apr. 2021. DOI: 10.1103/PhysRevLett.126.161802. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.126.161802.

[9] E. Kou, P. Urquijo, W. Altmannshofer, F. Beaujean, G. Bell, M. Beneke, I. I. Bigi, F. Bishara, M. Blanke, C. Bobeth, and et al., "The Belle II physics book", *Progress of Theoretical and Experimental Physics*, vol. 2019, no. 12, Dec. 2019, ISSN: 2050-3911. DOI: 10.1093/ptep/ptz106. [Online]. Available: http://dx.doi.org/10.1093/ptep/ptz106.

[10] M. Tanabashi, K. Hagiwara, K. Hikasa, K. Nakamura, Y. Sumino, F. Takahashi, J. Tanaka, and K. e. a. Agashe, "Review of particle physics", *Phys. Rev. D*, vol. 98, p. 030 001, 3 Aug. 2018. DOI: 10.1103/PhysRevD.98.030001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[11] K. Akai, K. Furukawa, and H. Koiso, "SuperKEKB collider", *Nuclear instruments and methods in physics research section A: accelerators, spectrometers, detectors and associated equipment*, vol. 907, pp. 188–199, 2018, Advances in Instrumentation and Experimental Methods (Special Issue in Honour of Kai Siegbahn), ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2018.08.017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900218309616.

[12] B. II. (2022). "SuperKEKB and Belle II", [Online]. Available: https://www.belle2.org/project/super_kekb_and_belle_ii (visited on 08/10/2022).

[13] T. Abe, I. Adachi, K. Adamczyk, S. Ahn, H. Aihara, K. Akai, M. Aloi, L. Andricek, and K. e. a. Aoki, "Belle II technical design report", 2010. DOI: 10.48550/arxiv.1011.0352. [Online]. Available: https://arxiv.org/abs/1011.0352.

[14] C. Schwanda, K. Adamczyk, H. Aihara, T. Aziz, S. Bacher, and S. e. a. Bahinipati, "Performance of the Belle II Silicon Vertex Detector", *PoS*, vol. Vertex2019, p. 014, 2020. DOI: 10.22323/1.373.0014.

[15] V. Bertacchi, T. Bilka, N. Braun, G. Casarosa, L. Corona, and S. C. et al., "Track finding at Belle II", *Computer Physics Communications*, vol. 259, p. 107 610, Feb. 2021. DOI: 10.1016/j.cpc.2020.107610. [Online]. Available: https://doi.org/10.1016/j.cpc.2020.107610.

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, San Francisco, California, USA: ACM, 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785.

[17] C. Li. (2022). "A gentle introduction to gradient boosting", [Online]. Available: http://www.chengli.io/tutorials/gradient_boosting.pdf (visited on 08/12/2022).

[18] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.", *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: 10.1214/aos/1013203451. [Online]. Available: https://doi.org/10.1214/aos/1013203451.

[19] D. Boigelot. (2007). "An example of the linear correlation of x and y for various distributions of (x,y) pairs.", [Online]. Available: https://commons.wikimedia.org/wiki/File:Correlation_examples.png (visited on 08/15/2022).

[20] G. J. Szekely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances", *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007. DOI: 10.1214/009053607000000505. [Online]. Available: https://doi.org/10.1214/009053607000000505.

[21] D. Edelmann, T. F. Móri, and G. J. Székely, "On relationships between the Pearson and the distance correlation coefficients", *Statistics & Probability Letters*, vol. 169, p. 108 960, 2021, ISSN: 0167-7152. DOI: https://doi.org/10.1016/j.spl.2020.108960. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167715220302637.

[22] D. Boigelot. (2012). "An example of the distance correlation of x and y for various distributions of (x,y) pairs", [Online]. Available: https://commons.

wikimedia.org/wiki/File:Distance_Correlation_Examples.svg (visited on 08/15/2022).

[23] S. Choudhury, "$B \to K^* \ell\ell$ rediscovery at Belle II", Presentation at 39th B2GM, Jun. 2021.

[24] G. C. Fox and S. Wolfram, "Observables for the analysis of event shapes in e^+e^- annihilation and other processes", *Phys. Rev. Lett.*, vol. 41, pp. 1581–1585, 23 Dec. 1978. DOI: 10.1103/PhysRevLett.41.1581. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.41.1581.

[25] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, and D. e. a. Cournapeau, "SciPy 1.0: Fundamental algorithms for scientific computing in Python", *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[26] H. Dembinski and P. O. et al., "Scikit-hep/iminuit", Dec. 2020. DOI: 10.5281/zenodo.3949207. [Online]. Available: https://doi.org/10.5281/zenodo.3949207.

List of Figures

List of Tables

Appendices

A. Distance correlation tables

The following tables A.1 to A.6 show the distance correlations dcor for the six decay channels of interest after the skim cuts plus the unavoidable PID, J/Ψ, and $\Psi(2S)$ cuts (see chapter 6.1 for more details). dcor is evaluated on pure background samples for all variables that are considered for training. Only those with dcor < 0.1 for $M_{bc}, \theta_\ell, \theta_k$ and ϕ_B in all channels are used in the BDT. Table A.7 shows dcor of the BDT outputs with the same four variables.

Table A.1: Distance correlations for the channel $B^0 \to (K^{*0} \to K^+\pi^-)e^+e^-$.

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
E_B	0.15	0.02	0.02	0.02
E_{K^*}	0.03	0.06	0.05	0.02
E_K	0.02	0.04	0.56	0.02
M_K	0.00	0.00	0.00	0.00
$\cos\theta_K$	0.02	0.02	0.07	0.02
dr_K	0.03	0.05	0.13	0.02
dz_K	0.02	0.05	0.06	0.02
$\lvert\vec{p}_K\rvert$	0.02	0.04	0.56	0.02
$p_{t,K}$	0.02	0.05	0.54	0.02
$p_{x,K}$	0.02	0.03	0.20	0.02
$p_{y,K}$	0.01	0.02	0.20	0.01
$p_{z,K}$	0.02	0.03	0.21	0.02
$\lvert\vec{p}_{K,\text{ cms frame}}\rvert$	0.02	0.05	0.59	0.02
M_{K^*}	0.02	0.02	0.03	0.02
$\lvert\vec{p}_{K^*}\rvert$	0.03	0.06	0.05	0.02
E_π	0.03	0.03	0.79	0.01
M_π	0.00	0.00	0.00	0.00
$\cos\theta_\pi$	0.03	0.02	0.04	0.01
dr_π	0.01	0.03	0.22	0.02
dz_π	0.02	0.04	0.11	0.02
$\lvert\vec{p}_\pi\rvert$	0.03	0.03	0.79	0.01
$p_{t,\pi}$	0.03	0.04	0.77	0.02
$p_{x,\pi}$	0.02	0.02	0.34	0.02
$p_{y,\pi}$	0.02	0.03	0.34	0.02

continued on next page

Table A.1 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$p_{z,\pi}$	0.02	0.02	0.36	0.01
$\lvert\vec{p}_{\pi,\text{ cms frame}}\rvert$	0.04	0.04	0.81	0.01
p_{t,K^*}	0.04	0.07	0.06	0.02
p_{x,K^*}	0.02	0.03	0.02	0.02
p_{y,K^*}	0.02	0.03	0.03	0.02
p_{z,K^*}	0.01	0.03	0.04	0.02
M_{bc}	1.00	0.02	0.03	0.01
aplanarity	0.08	0.10	0.09	0.05
χ^2_{prob}	0.02	0.04	0.02	0.01
CCT_0	0.08	0.13	0.05	0.01
CCT_1	0.02	0.07	0.03	0.01
CCT_2	0.04	0.03	0.05	0.02
CCT_3	0.03	0.05	0.02	0.01
CCT_4	0.01	0.05	0.03	0.02
CCT_5	0.01	0.07	0.03	0.01
CCT_6	0.02	0.07	0.02	0.01
CCT_7	0.02	0.05	0.02	0.01
CCT_8	0.03	0.03	0.02	0.01
$\cos\theta_B$	0.51	0.02	0.02	0.01
ΔE	0.02	0.01	0.02	0.01
$\delta x_{\ell^+\ell^-}$	0.02	0.05	0.02	0.02
$\delta y_{\ell^+\ell^-}$	0.02	0.04	0.02	0.02
$\delta z_{\ell^+\ell^-}$	0.02	0.04	0.02	0.02
dr	0.02	0.07	0.02	0.02
dr	0.02	0.07	0.02	0.02
dz	0.02	0.05	0.02	0.02
dz	0.02	0.05	0.02	0.02
R_1	0.07	0.06	0.03	0.02
R_2	0.03	0.18	0.03	0.01
R_3	0.08	0.20	0.02	0.01
R_4	0.07	0.13	0.06	0.02
$B_{0,\text{thrust}}$	0.06	0.08	0.02	0.02
$B_{1,\text{thrust}}$	0.02	0.02	0.02	0.02
$B_{2,\text{thrust}}$	0.04	0.19	0.02	0.02
$B_{3,\text{thrust}}$	0.02	0.06	0.03	0.01
$B_{4,\text{thrust}}$	0.07	0.20	0.03	0.02
E_{ℓ^-}	0.02	0.14	0.02	0.02
dr_{ℓ^-}	0.02	0.03	0.02	0.01

continued on next page

Table A.1 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$dx_{\ell-}$	0.02	0.03	0.02	0.02
$dy_{\ell-}$	0.02	0.03	0.02	0.02
$dz_{\ell-}$	0.03	0.06	0.02	0.02
$p_{\ell-}$	0.02	0.14	0.02	0.02
$\text{pValue}_{\ell-}$	0.02	0.02	0.02	0.02
$p_{t,\ell-}$	0.03	0.14	0.02	0.02
$p_{x,\ell-}$	0.02	0.04	0.02	0.02
$p_{y,\ell-}$	0.02	0.05	0.01	0.02
$p_{z,\ell-}$	0.02	0.04	0.02	0.01
$\theta_{\ell-}$	0.02	0.03	0.02	0.01
$\lvert\vec{p}_{\ell-,\text{CMS frame}}\rvert$	0.04	0.17	0.02	0.02
$E_{\ell+}$	0.03	0.14	0.03	0.03
$dr_{\ell+}$	0.02	0.05	0.02	0.01
$dx_{\ell+}$	0.02	0.04	0.02	0.02
$dy_{\ell+}$	0.02	0.04	0.02	0.02
$dz_{\ell+}$	0.02	0.05	0.02	0.02
$p_{\ell+}$	0.03	0.14	0.03	0.03
$\text{pValue}_{\ell+}$	0.02	0.02	0.02	0.02
$p_{t,\ell+}$	0.03	0.13	0.02	0.02
$p_{x,\ell+}$	0.02	0.04	0.02	0.01
$p_{y,\ell+}$	0.02	0.04	0.02	0.02
$p_{z,\ell+}$	0.02	0.05	0.03	0.02
$\theta_{\ell+}$	0.02	0.02	0.02	0.02
$\lvert\vec{p}_{\ell+,\text{CMS frame}}\rvert$	0.04	0.17	0.03	0.03
M_{Recoil}	0.34	0.02	0.02	0.01
$M_{\text{event, missing}}$	0.05	0.08	0.02	0.02
$\lvert\vec{p_B}\rvert$	0.25	0.02	0.02	0.01
$\lvert\vec{p}_{\text{Recoil}}\rvert$	0.23	0.02	0.01	0.01
$\theta_{\text{missing } p}$	0.52	0.02	0.02	0.01
ϕ_B	0.01	0.02	0.01	1.00
$p_{t,B}$	0.54	0.02	0.03	0.01
$p_{x,B}$	0.22	0.01	0.02	0.01
$p_{y,B}$	0.21	0.02	0.02	0.02
$p_{z,B}$	0.20	0.02	0.02	0.01
$E(ROE)$	0.06	0.08	0.02	0.02
$E_{\text{extra}}(ROE)$	0.05	0.08	0.02	0.01
$M(ROE)$	0.07	0.09	0.02	0.02
$\lvert\vec{p}(ROE)\rvert$	0.04	0.02	0.02	0.01

continued on next page

Table A.1 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$p_t(ROE)$	0.13	0.04	0.04	0.02
sphericity	0.08	0.26	0.06	0.02
θ_k	0.03	0.04	1.00	0.01
θ_ℓ	0.02	1.00	0.04	0.02
thrust_{event}	0.02	0.18	0.03	0.01
thrust_B	0.05	0.28	0.06	0.02
$E_{\ell^+\ell^-}$	0.03	0.02	0.02	0.01

Table A.2: Distance correlations for the channel $B^0 \to (K^{*0} \to K^+\pi^-)\mu^+\mu^-$.

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
E_B	0.18	0.01	0.03	0.02
E_{K^*}	0.04	0.11	0.04	0.01
E_K	0.02	0.06	0.67	0.02
M_K	0.00	0.00	0.00	0.00
$\cos\theta_K$	0.01	0.03	0.06	0.01
dr_K	0.02	0.05	0.13	0.03
dz_K	0.02	0.05	0.05	0.02
$\lvert\vec{p}_K\rvert$	0.02	0.06	0.67	0.02
$p_{t,K}$	0.02	0.05	0.65	0.02
$p_{x,K}$	0.02	0.02	0.22	0.01
$p_{y,K}$	0.01	0.02	0.22	0.01
$p_{z,K}$	0.01	0.04	0.24	0.02
$\lvert\vec{p}_{K,\text{ cms frame}}\rvert$	0.02	0.06	0.73	0.02
M_{K^*}	0.01	0.02	0.03	0.01
$\lvert\vec{p}_{K^*}\rvert$	0.04	0.11	0.04	0.01
E_π	0.03	0.06	0.84	0.01
M_π	0.00	0.00	0.00	0.00
$\cos\theta_\pi$	0.01	0.02	0.03	0.02
dr_π	0.02	0.03	0.19	0.01
dz_π	0.03	0.04	0.10	0.02
$\lvert\vec{p}_\pi\rvert$	0.03	0.06	0.84	0.01
$p_{t,\pi}$	0.03	0.06	0.83	0.01
$p_{x,\pi}$	0.02	0.03	0.35	0.02
$p_{y,\pi}$	0.01	0.03	0.35	0.02
$p_{z,\pi}$	0.01	0.03	0.38	0.02
$\lvert\vec{p}_{\pi,\text{ cms frame}}\rvert$	0.03	0.06	0.87	0.01
p_{t,K^*}	0.03	0.10	0.04	0.02

continued on next page

Table A.2 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
p_{x,K^*}	0.02	0.03	0.02	0.01
p_{y,K^*}	0.01	0.03	0.02	0.01
p_{z,K^*}	0.01	0.04	0.04	0.02
M_{bc}	1.00	0.02	0.02	0.02
aplanarity	0.06	0.09	0.11	0.05
χ^2_{prob}	0.02	0.09	0.02	0.02
CCT_0	0.06	0.10	0.07	0.02
CCT_1	0.04	0.05	0.04	0.02
CCT_2	0.03	0.05	0.06	0.03
CCT_3	0.02	0.05	0.04	0.02
CCT_4	0.02	0.04	0.04	0.01
CCT_5	0.01	0.04	0.03	0.01
CCT_6	0.03	0.04	0.02	0.02
CCT_7	0.03	0.03	0.02	0.02
CCT_8	0.04	0.03	0.02	0.01
$\cos\theta_B$	0.51	0.02	0.02	0.02
ΔE	0.01	0.01	0.03	0.02
$\delta x_{\ell^+\ell^-}$	0.02	0.04	0.03	0.02
$\delta y_{\ell^+\ell^-}$	0.02	0.03	0.02	0.02
$\delta z_{\ell^+\ell^-}$	0.02	0.04	0.02	0.02
dr	0.03	0.06	0.03	0.01
dr	0.03	0.06	0.03	0.01
dz	0.02	0.05	0.02	0.02
dz	0.02	0.05	0.02	0.02
R_1	0.05	0.07	0.01	0.01
R_2	0.02	0.12	0.04	0.01
R_3	0.07	0.12	0.01	0.02
R_4	0.07	0.10	0.07	0.02
$B_{0,\text{thrust}}$	0.02	0.12	0.02	0.02
$B_{1,\text{thrust}}$	0.02	0.03	0.01	0.01
$B_{2,\text{thrust}}$	0.02	0.15	0.04	0.02
$B_{3,\text{thrust}}$	0.01	0.04	0.01	0.02
$B_{4,\text{thrust}}$	0.04	0.16	0.05	0.02
E_{ℓ^-}	0.02	0.12	0.02	0.04
dr_{ℓ^-}	0.03	0.03	0.02	0.01
dx_{ℓ^-}	0.02	0.03	0.02	0.02
dy_{ℓ^-}	0.02	0.02	0.02	0.02
dz_{ℓ^-}	0.03	0.04	0.02	0.02

continued on next page

Table A.2 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$p_{\ell-}$	0.02	0.12	0.02	0.04
$\text{pValue}_{\ell-}$	0.02	0.02	0.01	0.02
$p_{t,\ell-}$	0.02	0.11	0.02	0.03
$p_{x,\ell-}$	0.01	0.04	0.02	0.01
$p_{y,\ell-}$	0.02	0.03	0.02	0.02
$p_{z,\ell-}$	0.01	0.05	0.02	0.03
$\theta_{\ell-}$	0.01	0.04	0.03	0.02
$\lvert\vec{p}_{\ell-,\text{CMS frame}}\rvert$	0.03	0.14	0.02	0.04
$E_{\ell+}$	0.03	0.13	0.02	0.03
$dr_{\ell+}$	0.01	0.05	0.02	0.01
$dx_{\ell+}$	0.02	0.04	0.02	0.02
$dy_{\ell+}$	0.02	0.04	0.03	0.02
$dz_{\ell+}$	0.02	0.05	0.02	0.02
$p_{\ell+}$	0.03	0.13	0.02	0.03
$\text{pValue}_{\ell+}$	0.02	0.04	0.02	0.02
$p_{t,\ell+}$	0.03	0.13	0.02	0.03
$p_{x,\ell+}$	0.03	0.04	0.01	0.02
$p_{y,\ell+}$	0.02	0.05	0.01	0.02
$p_{z,\ell+}$	0.02	0.05	0.02	0.02
$\theta_{\ell+}$	0.02	0.03	0.02	0.02
$\lvert\vec{p}_{\ell+,\text{CMS frame}}\rvert$	0.04	0.14	0.02	0.03
M_{Recoil}	0.44	0.01	0.03	0.02
$M_{\text{event, missing}}$	0.01	0.09	0.02	0.01
$\lvert\vec{p_B}\rvert$	0.27	0.01	0.03	0.02
$\lvert\vec{p}_{\text{Recoil}}\rvert$	0.24	0.02	0.02	0.02
$\theta_{\text{missing } p}$	0.52	0.02	0.03	0.02
ϕ_B	0.02	0.02	0.02	1.00
$p_{t,B}$	0.55	0.02	0.01	0.02
$p_{x,B}$	0.22	0.02	0.02	0.01
$p_{y,B}$	0.22	0.01	0.01	0.02
$p_{z,B}$	0.22	0.02	0.03	0.02
$E(ROE)$	0.04	0.11	0.02	0.02
$E_{\text{extra}}(ROE)$	0.02	0.10	0.02	0.02
$M(ROE)$	0.05	0.11	0.03	0.02
$\lvert\vec{p}(ROE)\rvert$	0.06	0.04	0.02	0.03
$p_t(ROE)$	0.15	0.03	0.02	0.02
sphericity	0.04	0.20	0.10	0.02
θ_k	0.02	0.04	1.00	0.02

continued on next page

Table A.2 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
θ_ℓ	0.02	1.00	0.04	0.02
thrust_{event}	0.04	0.12	0.03	0.02
thrust_B	0.04	0.19	0.09	0.02

Table A.3: Distance correlations for the channel $B^+ \to (K^{*+} \to K^0_s\pi^+)e^+e^-$.

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
E_B	0.16	0.01	0.02	0.01
E_{K^*}	0.03	0.05	0.04	0.03
E_K	0.02	0.03	0.58	0.02
M_K	0.02	0.02	0.04	0.02
dr_K	0.01	0.02	0.24	0.02
dz_K	0.02	0.02	0.14	0.02
$\lvert\vec{p}_K\rvert$	0.02	0.03	0.58	0.02
$p_{t,K}$	0.01	0.03	0.56	0.02
$p_{x,K}$	0.01	0.02	0.20	0.02
$p_{y,K}$	0.01	0.02	0.20	0.01
$p_{z,K}$	0.02	0.02	0.21	0.02
$\lvert\vec{p}_{K,\text{ cms frame}}\rvert$	0.02	0.03	0.62	0.02
M_{K^*}	0.02	0.03	0.02	0.02
$\lvert\vec{p}_{K^*}\rvert$	0.03	0.05	0.04	0.03
E_π	0.03	0.04	0.79	0.03
M_π	0.00	0.00	0.00	0.00
dr_π	0.02	0.02	0.21	0.01
dz_π	0.02	0.04	0.11	0.02
$\lvert\vec{p}_\pi\rvert$	0.03	0.04	0.79	0.03
$p_{t,\pi}$	0.03	0.04	0.78	0.03
$p_{x,\pi}$	0.02	0.03	0.35	0.03
$p_{y,\pi}$	0.02	0.02	0.35	0.02
$p_{z,\pi}$	0.02	0.02	0.36	0.02
$\lvert\vec{p}_{\pi,\text{ cms frame}}\rvert$	0.03	0.04	0.81	0.03
p_{t,K^*}	0.02	0.05	0.04	0.04
p_{x,K^*}	0.01	0.02	0.02	0.02
p_{y,K^*}	0.01	0.02	0.02	0.01
p_{z,K^*}	0.03	0.02	0.04	0.02
M_{bc}	1.00	0.03	0.02	0.01
aplanarity	0.05	0.10	0.05	0.04
χ^2_{prob}	0.01	0.02	0.04	0.02

continued on next page

Table A.3 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
CCT_0	0.06	0.11	0.05	0.01
CCT_1	0.02	0.08	0.05	0.02
CCT_2	0.03	0.03	0.05	0.03
CCT_3	0.04	0.06	0.02	0.01
CCT_4	0.02	0.06	0.01	0.02
CCT_5	0.02	0.05	0.01	0.02
CCT_6	0.02	0.05	0.02	0.01
CCT_7	0.03	0.03	0.02	0.01
CCT_8	0.04	0.03	0.01	0.01
$\cos\theta_B$	0.50	0.02	0.02	0.02
ΔE	0.02	0.02	0.01	0.02
$\delta x_{\ell^+\ell^-}$	0.02	0.03	0.02	0.02
$\delta y_{\ell^+\ell^-}$	0.02	0.03	0.02	0.02
$\delta z_{\ell^+\ell^-}$	0.02	0.02	0.02	0.02
dr	0.03	0.04	0.07	0.02
dr	0.03	0.04	0.07	0.02
dz	0.02	0.05	0.03	0.02
dz	0.02	0.05	0.03	0.02
R_1	0.04	0.04	0.22	0.01
R_2	0.02	0.15	0.05	0.01
R_3	0.06	0.10	0.14	0.01
R_4	0.06	0.11	0.06	0.02
$B_{0,\text{thrust}}$	0.04	0.06	0.14	0.01
$B_{1,\text{thrust}}$	0.01	0.02	0.08	0.01
$B_{2,\text{thrust}}$	0.03	0.15	0.11	0.01
$B_{3,\text{thrust}}$	0.02	0.01	0.04	0.01
$B_{4,\text{thrust}}$	0.07	0.17	0.08	0.01
E_{ℓ^-}	0.04	0.12	0.02	0.06
dr_{ℓ^-}	0.02	0.03	0.02	0.02
dx_{ℓ^-}	0.02	0.03	0.02	0.02
dy_{ℓ^-}	0.03	0.03	0.02	0.02
dz_{ℓ^-}	0.02	0.04	0.02	0.02
p_{ℓ^-}	0.04	0.12	0.02	0.06
pValue_{ℓ^-}	0.01	0.02	0.02	0.01
p_{t,ℓ^-}	0.03	0.12	0.02	0.05
p_{x,ℓ^-}	0.02	0.03	0.01	0.03
p_{y,ℓ^-}	0.02	0.03	0.01	0.02
p_{z,ℓ^-}	0.01	0.03	0.02	0.03

continued on next page

Table A.3 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
θ_{ℓ^-}	0.01	0.02	0.02	0.01
$\lvert\vec{p}_{\ell^-,\text{CMS frame}}\rvert$	0.05	0.15	0.02	0.06
E_{ℓ^+}	0.04	0.12	0.03	0.04
dr_{ℓ^+}	0.01	0.03	0.02	0.02
dx_{ℓ^+}	0.02	0.03	0.02	0.02
dy_{ℓ^+}	0.02	0.03	0.02	0.02
dz_{ℓ^+}	0.02	0.04	0.02	0.02
p_{ℓ^+}	0.04	0.12	0.03	0.04
pValue_{ℓ^+}	0.01	0.02	0.02	0.02
p_{t,ℓ^+}	0.03	0.12	0.02	0.04
p_{x,ℓ^+}	0.02	0.03	0.02	0.02
p_{y,ℓ^+}	0.02	0.03	0.01	0.02
p_{z,ℓ^+}	0.02	0.03	0.03	0.02
θ_{ℓ^+}	0.01	0.02	0.03	0.01
$\lvert\vec{p}_{\ell^+,\text{CMS frame}}\rvert$	0.04	0.14	0.02	0.04
M_{Recoil}	0.35	0.02	0.01	0.02
$M_{\text{event, missing}}$	0.04	0.07	0.10	0.01
$\lvert\vec{p_B}\rvert$	0.26	0.02	0.03	0.02
$\lvert\vec{p}_{\text{Recoil}}\rvert$	0.24	0.02	0.03	0.02
$\theta_{\text{missing } p}$	0.52	0.02	0.02	0.01
ϕ_B	0.01	0.02	0.02	1.00
$p_{t,B}$	0.53	0.02	0.01	0.01
$p_{x,B}$	0.22	0.01	0.02	0.01
$p_{y,B}$	0.22	0.01	0.01	0.02
$p_{z,B}$	0.21	0.02	0.03	0.02
$E(ROE)$	0.06	0.08	0.02	0.02
$E_{\text{extra}}(ROE)$	0.05	0.06	0.02	0.02
$M(ROE)$	0.07	0.08	0.02	0.02
$\lvert\vec{p}(ROE)\rvert$	0.04	0.02	0.02	0.02
$p_t(ROE)$	0.13	0.05	0.02	0.02
sphericity	0.08	0.24	0.04	0.01
θ_k	0.02	0.03	1.00	0.02
θ_ℓ	0.03	1.00	0.03	0.02
$\text{thrust}_{\text{Event}}$	0.02	0.14	0.05	0.01
thrust_B	0.08	0.23	0.03	0.01

Table A.4: Distance correlations for the channel $B^+ \to (K^{*+} \to K_s^0\pi^+)\mu^+\mu^-$.

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
E_B	0.18	0.01	0.01	0.02
E_{K^*}	0.03	0.11	0.02	0.02
E_K	0.02	0.06	0.66	0.01
M_K	0.02	0.02	0.03	0.02
$\cos\theta_K$	0.03	0.04	0.07	0.02
dr_K	0.02	0.02	0.26	0.01
dz_K	0.03	0.03	0.14	0.02
$\lvert\vec{p}_K\rvert$	0.02	0.06	0.66	0.01
$p_{t,K}$	0.02	0.05	0.65	0.01
$p_{x,K}$	0.02	0.02	0.22	0.01
$p_{y,K}$	0.01	0.02	0.21	0.03
$p_{z,K}$	0.02	0.04	0.23	0.02
$\lvert\vec{p}_{K,\text{ cms frame}}\rvert$	0.02	0.06	0.72	0.01
M_{K^*}	0.01	0.03	0.03	0.01
$\lvert\vec{p}_{K^*}\rvert$	0.03	0.11	0.02	0.02
E_π	0.02	0.05	0.83	0.02
M_π	0.00	0.00	0.00	0.00
$\cos\theta_\pi$	0.02	0.03	0.05	0.02
dr_π	0.02	0.03	0.22	0.01
dz_π	0.02	0.03	0.11	0.02
$\lvert\vec{p}_\pi\rvert$	0.02	0.05	0.83	0.02
$p_{t,\pi}$	0.02	0.04	0.82	0.02
$p_{x,\pi}$	0.02	0.02	0.34	0.02
$p_{y,\pi}$	0.02	0.02	0.36	0.03
$p_{z,\pi}$	0.03	0.03	0.38	0.02
$\lvert\vec{p}_{\pi,\text{ cms frame}}\rvert$	0.02	0.05	0.86	0.02
p_{t,K^*}	0.03	0.10	0.02	0.02
p_{x,K^*}	0.02	0.03	0.01	0.02
p_{y,K^*}	0.01	0.03	0.02	0.03
p_{z,K^*}	0.03	0.05	0.04	0.02
M_{bc}	1.00	0.02	0.01	0.01
aplanarity	0.04	0.11	0.05	0.04
χ^2_{prob}	0.01	0.02	0.02	0.02
CCT_0	0.05	0.09	0.07	0.02
CCT_1	0.02	0.06	0.07	0.02
CCT_2	0.03	0.04	0.04	0.02
CCT_3	0.02	0.04	0.03	0.02

continued on next page

Table A.4 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
CCT_4	0.02	0.03	0.02	0.02
CCT_5	0.01	0.04	0.02	0.01
CCT_6	0.02	0.04	0.01	0.01
CCT_7	0.03	0.03	0.02	0.03
CCT_8	0.01	0.03	0.02	0.01
$\cos\theta_B$	0.50	0.02	0.02	0.02
ΔE	0.02	0.01	0.01	0.02
$\delta x_{\ell^+\ell^-}$	0.01	0.03	0.02	0.02
$\delta y_{\ell^+\ell^-}$	0.02	0.03	0.02	0.01
$\delta z_{\ell^+\ell^-}$	0.02	0.03	0.02	0.02
dr	0.03	0.03	0.03	0.02
dr	0.03	0.03	0.03	0.02
dz	0.02	0.04	0.02	0.02
dz	0.02	0.04	0.02	0.02
R_1	0.02	0.04	0.25	0.01
R_2	0.02	0.10	0.10	0.02
R_3	0.04	0.05	0.17	0.03
R_4	0.04	0.09	0.08	0.03
$B_{0,\text{thrust}}$	0.02	0.06	0.17	0.02
$B_{1,\text{thrust}}$	0.02	0.01	0.10	0.02
$B_{2,\text{thrust}}$	0.02	0.11	0.16	0.01
$B_{3,\text{thrust}}$	0.01	0.03	0.05	0.02
$B_{4,\text{thrust}}$	0.03	0.13	0.13	0.02
E_{ℓ^-}	0.03	0.12	0.01	0.02
dr_{ℓ^-}	0.01	0.02	0.01	0.01
dx_{ℓ^-}	0.02	0.02	0.02	0.02
dy_{ℓ^-}	0.02	0.02	0.02	0.02
dz_{ℓ^-}	0.02	0.03	0.02	0.01
p_{ℓ^-}	0.03	0.12	0.01	0.02
pValue_{ℓ^-}	0.01	0.02	0.01	0.01
p_{t,ℓ^-}	0.04	0.11	0.02	0.01
p_{x,ℓ^-}	0.02	0.04	0.02	0.01
p_{y,ℓ^-}	0.02	0.04	0.03	0.02
p_{z,ℓ^-}	0.02	0.03	0.02	0.02
θ_{ℓ^-}	0.02	0.03	0.01	0.02
$\lvert\vec{p}_{\ell^-,\text{CMS frame}}\rvert$	0.04	0.14	0.02	0.02
E_{ℓ^+}	0.02	0.12	0.02	0.02
dr_{ℓ^+}	0.03	0.03	0.01	0.02

continued on next page

Table A.4 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$dx_{\ell+}$	0.02	0.03	0.02	0.02
$dy_{\ell+}$	0.02	0.03	0.02	0.02
$dz_{\ell+}$	0.03	0.04	0.02	0.02
$p_{\ell+}$	0.02	0.12	0.02	0.02
$\text{pValue}_{\ell+}$	0.01	0.02	0.01	0.01
$p_{t,\ell+}$	0.02	0.11	0.02	0.03
$p_{x,\ell+}$	0.02	0.04	0.03	0.02
$p_{y,\ell+}$	0.02	0.04	0.02	0.02
$p_{z,\ell+}$	0.01	0.04	0.03	0.02
$\theta_{\ell+}$	0.02	0.03	0.04	0.02
$\lvert\vec{p}_{\ell+,\text{CMS frame}}\rvert$	0.03	0.13	0.02	0.02
M_{Recoil}	0.44	0.01	0.01	0.02
$M_{\text{event, missing}}$	0.02	0.06	0.12	0.02
$\lvert\vec{p_B}\rvert$	0.26	0.02	0.01	0.02
$\lvert\vec{p}_{\text{Recoil}}\rvert$	0.24	0.01	0.01	0.02
$\theta_{\text{missing } p}$	0.53	0.01	0.01	0.02
ϕ_B	0.01	0.02	0.01	1.00
$p_{t,B}$	0.53	0.02	0.02	0.01
$p_{x,B}$	0.22	0.02	0.02	0.02
$p_{y,B}$	0.21	0.01	0.01	0.01
$p_{z,B}$	0.21	0.01	0.01	0.02
$E(ROE)$	0.04	0.06	0.03	0.03
$E_{\text{extra}}(ROE)$	0.03	0.06	0.02	0.02
$M(ROE)$	0.05	0.06	0.03	0.03
$\lvert\vec{p}(ROE)\rvert$	0.07	0.01	0.02	0.02
$p_t(ROE)$	0.15	0.03	0.03	0.02
sphericity	0.04	0.21	0.06	0.01
θ_k	0.01	0.02	1.00	0.01
θ_ℓ	0.02	1.00	0.02	0.02
thrust_{event}	0.02	0.10	0.08	0.01
thrust_B	0.06	0.17	0.03	0.02

Table A.5: Distance correlations for the channel $B^+ \to (K^{*+} \to K^+\pi^0)e^+e^-$.

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
E_B	0.16	0.02	0.02	0.01
E_{K^*}	0.03	0.07	0.04	0.05
E_K	0.02	0.07	0.49	0.04

continued on next page

Table A.5 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
M_K	0.01	0.00	0.00	0.00
$\cos\theta_K$	0.03	0.02	0.06	0.01
dr_K	0.02	0.03	0.14	0.02
dz_K	0.03	0.06	0.06	0.02
$\lvert\vec{p}_K\rvert$	0.02	0.07	0.49	0.04
$p_{t,K}$	0.02	0.07	0.47	0.03
$p_{x,K}$	0.03	0.03	0.19	0.02
$p_{y,K}$	0.02	0.04	0.18	0.02
$p_{z,K}$	0.03	0.04	0.20	0.02
$\lvert\vec{p}_{K,\text{ cms frame}}\rvert$	0.02	0.07	0.52	0.03
M_{K^*}	0.02	0.03	0.07	0.02
$\lvert\vec{p}_{K^*}\rvert$	0.03	0.07	0.05	0.05
E_π	0.04	0.04	0.77	0.04
M_π	0.01	0.01	0.08	0.02
$\cos\theta_\pi$	0.02	0.03	0.10	0.02
dr_π	0.00	0.00	0.00	0.00
dz_π	0.00	0.00	0.00	0.00
$\lvert\vec{p}_\pi\rvert$	0.04	0.04	0.77	0.04
$p_{t,\pi}$	0.03	0.04	0.75	0.04
$p_{x,\pi}$	0.03	0.03	0.34	0.03
$p_{y,\pi}$	0.02	0.02	0.33	0.02
$p_{z,\pi}$	0.03	0.03	0.37	0.02
$\lvert\vec{p}_{\pi,\text{ cms frame}}\rvert$	0.04	0.05	0.79	0.04
p_{t,K^*}	0.03	0.07	0.04	0.05
p_{x,K^*}	0.03	0.03	0.02	0.03
p_{y,K^*}	0.02	0.03	0.02	0.03
p_{z,K^*}	0.03	0.03	0.03	0.02
M_{bc}	1.00	0.03	0.03	0.02
aplanarity	0.07	0.09	0.07	0.06
χ^2_{prob}	0.04	0.03	0.02	0.01
CCT_0	0.07	0.13	0.05	0.02
CCT_1	0.02	0.07	0.03	0.02
CCT_2	0.08	0.04	0.04	0.02
CCT_3	0.01	0.04	0.03	0.02
CCT_4	0.02	0.06	0.02	0.01
CCT_5	0.03	0.07	0.05	0.02
CCT_6	0.03	0.09	0.03	0.01
CCT_7	0.02	0.05	0.02	0.02

continued on next page

Table A.5 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
CCT_8	0.03	0.02	0.04	0.02
$\cos\theta_B$	0.51	0.02	0.02	0.01
ΔE	0.02	0.02	0.01	0.01
$\delta x_{\ell^+\ell^-}$	0.02	0.04	0.02	0.02
$\delta y_{\ell^+\ell^-}$	0.02	0.04	0.02	0.02
$\delta z_{\ell^+\ell^-}$	0.02	0.03	0.02	0.02
dr	0.02	0.07	0.03	0.01
dr	0.02	0.07	0.03	0.01
dz	0.02	0.06	0.02	0.01
dz	0.02	0.06	0.02	0.01
R_1	0.07	0.07	0.02	0.02
R_2	0.03	0.19	0.05	0.02
R_3	0.09	0.19	0.03	0.02
R_4	0.06	0.11	0.04	0.03
$B_{0,\text{thrust}}$	0.05	0.09	0.06	0.02
$B_{1,\text{thrust}}$	0.02	0.02	0.02	0.02
$B_{2,\text{thrust}}$	0.03	0.20	0.02	0.02
$B_{3,\text{thrust}}$	0.01	0.05	0.02	0.02
$B_{4,\text{thrust}}$	0.05	0.20	0.04	0.02
E_{ℓ^-}	0.03	0.11	0.02	0.03
dr_{ℓ^-}	0.02	0.04	0.02	0.02
dx_{ℓ^-}	0.02	0.04	0.02	0.02
dy_{ℓ^-}	0.02	0.04	0.02	0.02
dz_{ℓ^-}	0.02	0.06	0.02	0.01
p_{ℓ^-}	0.03	0.11	0.02	0.03
pValue_{ℓ^-}	0.01	0.02	0.01	0.02
p_{t,ℓ^-}	0.03	0.12	0.02	0.04
p_{x,ℓ^-}	0.01	0.03	0.02	0.02
p_{y,ℓ^-}	0.02	0.04	0.02	0.02
p_{z,ℓ^-}	0.01	0.03	0.01	0.02
θ_{ℓ^-}	0.01	0.02	0.01	0.02
$\lvert\vec{p}_{\ell^-,\text{CMS frame}}\rvert$	0.05	0.15	0.03	0.04
E_{ℓ^+}	0.03	0.12	0.03	0.02
dr_{ℓ^+}	0.01	0.05	0.01	0.01
dx_{ℓ^+}	0.02	0.04	0.02	0.02
dy_{ℓ^+}	0.02	0.04	0.02	0.02
dz_{ℓ^+}	0.02	0.05	0.02	0.02
p_{ℓ^+}	0.03	0.12	0.03	0.02

continued on next page

Table A.5 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$\text{pValue}_{\ell+}$	0.01	0.01	0.02	0.01
$p_{t,\ell+}$	0.03	0.12	0.03	0.03
$p_{x,\ell+}$	0.02	0.04	0.02	0.02
$p_{y,\ell+}$	0.02	0.03	0.02	0.01
$p_{z,\ell+}$	0.03	0.03	0.01	0.02
$\theta_{\ell+}$	0.03	0.03	0.01	0.01
$\lvert\vec{p}_{\ell+,\text{CMS frame}}\rvert$	0.04	0.14	0.03	0.03
M_{Recoil}	0.33	0.02	0.01	0.01
$M_{\text{event, missing}}$	0.03	0.08	0.07	0.02
$\lvert\vec{p_B}\rvert$	0.26	0.02	0.02	0.01
$\lvert\vec{p}_{\text{Recoil}}\rvert$	0.24	0.02	0.02	0.02
$\theta_{\text{missing } p}$	0.54	0.02	0.02	0.02
ϕ_B	0.02	0.02	0.03	1.00
$p_{t,B}$	0.54	0.02	0.01	0.02
$p_{x,B}$	0.22	0.02	0.01	0.02
$p_{y,B}$	0.22	0.02	0.02	0.02
$p_{z,B}$	0.22	0.02	0.02	0.01
$E(ROE)$	0.07	0.09	0.04	0.02
$E_{\text{extra}}(ROE)$	0.05	0.07	0.05	0.02
$M(ROE)$	0.08	0.10	0.04	0.01
$\lvert\vec{p}(ROE)\rvert$	0.05	0.04	0.02	0.02
$p_t(ROE)$	0.14	0.07	0.02	0.01
sphericity	0.07	0.25	0.06	0.01
θ_k	0.03	0.08	1.00	0.03
θ_ℓ	0.03	1.00	0.08	0.02
thrust_{event}	0.02	0.19	0.04	0.01
thrust_B	0.06	0.28	0.06	0.02

Table A.6: Distance correlations for the channel $B^+ \to (K^{*+} \to K^+\pi^0)\mu^+\mu^-$.

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
E_B	0.18	0.02	0.02	0.02
E_{K^*}	0.02	0.10	0.04	0.03
E_K	0.03	0.05	0.61	0.02
M_K	0.00	0.00	0.00	0.00
$\cos\theta_K$	0.02	0.03	0.04	0.02
dr_K	0.02	0.06	0.09	0.02
dz_K	0.02	0.05	0.04	0.02

continued on next page

Table A.6 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$\lvert\vec{p}_K\rvert$	0.03	0.05	0.60	0.02
$p_{t,K}$	0.03	0.05	0.58	0.02
$p_{x,K}$	0.01	0.02	0.21	0.01
$p_{y,K}$	0.03	0.02	0.20	0.01
$p_{z,K}$	0.02	0.03	0.22	0.02
$\lvert\vec{p}_{K,\text{ cms frame}}\rvert$	0.04	0.07	0.65	0.02
M_{K^*}	0.03	0.01	0.06	0.01
$\lvert\vec{p}_{K^*}\rvert$	0.02	0.10	0.04	0.03
E_π	0.04	0.06	0.84	0.03
M_π	0.02	0.02	0.07	0.02
$\cos\theta_\pi$	0.01	0.02	0.10	0.01
dr_π	0.00	0.00	0.00	0.00
dz_π	0.00	0.00	0.00	0.00
$\lvert\vec{p}_\pi\rvert$	0.04	0.06	0.84	0.03
$p_{t,\pi}$	0.04	0.05	0.83	0.03
$p_{x,\pi}$	0.02	0.02	0.36	0.02
$p_{y,\pi}$	0.03	0.03	0.37	0.02
$p_{z,\pi}$	0.02	0.04	0.40	0.02
$\lvert\vec{p}_{\pi,\text{ cms frame}}\rvert$	0.04	0.06	0.86	0.03
p_{t,K^*}	0.03	0.09	0.03	0.04
p_{x,K^*}	0.01	0.03	0.02	0.02
p_{y,K^*}	0.02	0.03	0.02	0.02
p_{z,K^*}	0.02	0.04	0.02	0.01
M_{bc}	1.00	0.03	0.04	0.01
aplanarity	0.07	0.08	0.11	0.04
χ^2_{prob}	0.04	0.08	0.07	0.01
CCT_0	0.05	0.09	0.07	0.01
CCT_1	0.02	0.06	0.05	0.02
CCT_2	0.04	0.04	0.06	0.02
CCT_3	0.02	0.04	0.03	0.01
CCT_4	0.02	0.03	0.04	0.01
CCT_5	0.03	0.04	0.06	0.01
CCT_6	0.02	0.05	0.05	0.01
CCT_7	0.02	0.03	0.03	0.02
CCT_8	0.02	0.02	0.05	0.01
$\cos\theta_B$	0.52	0.02	0.02	0.01
ΔE	0.02	0.02	0.02	0.02
$\delta x_{\ell^+\ell^-}$	0.01	0.04	0.03	0.02

continued on next page

Table A.6 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$\delta y_{\ell^+\ell^-}$	0.02	0.04	0.03	0.02
$\delta z_{\ell^+\ell^-}$	0.02	0.03	0.03	0.03
dr	0.02	0.05	0.02	0.02
dr	0.02	0.05	0.02	0.02
dz	0.02	0.05	0.02	0.02
dz	0.02	0.05	0.02	0.02
R_1	0.05	0.07	0.02	0.01
R_2	0.03	0.12	0.07	0.01
R_3	0.09	0.12	0.03	0.02
R_4	0.08	0.09	0.07	0.02
$B_{0,\text{thrust}}$	0.03	0.10	0.05	0.02
$B_{1,\text{thrust}}$	0.02	0.03	0.03	0.01
$B_{2,\text{thrust}}$	0.03	0.15	0.03	0.02
$B_{3,\text{thrust}}$	0.02	0.03	0.02	0.01
$B_{4,\text{thrust}}$	0.05	0.15	0.05	0.01
E_{ℓ^-}	0.03	0.12	0.02	0.02
dr_{ℓ^-}	0.01	0.04	0.02	0.02
dx_{ℓ^-}	0.02	0.04	0.03	0.02
dy_{ℓ^-}	0.02	0.03	0.02	0.02
dz_{ℓ^-}	0.02	0.04	0.03	0.03
p_{ℓ^-}	0.03	0.12	0.02	0.02
pValue_{ℓ^-}	0.02	0.03	0.02	0.01
p_{t,ℓ^-}	0.02	0.11	0.02	0.02
p_{x,ℓ^-}	0.02	0.04	0.02	0.01
p_{y,ℓ^-}	0.02	0.03	0.02	0.01
p_{z,ℓ^-}	0.02	0.03	0.02	0.02
θ_{ℓ^-}	0.02	0.03	0.01	0.02
$\lvert\vec{p}_{\ell^-,\text{CMS frame}}\rvert$	0.03	0.13	0.02	0.02
E_{ℓ^+}	0.02	0.11	0.02	0.02
dr_{ℓ^+}	0.02	0.04	0.04	0.01
dx_{ℓ^+}	0.02	0.04	0.03	0.02
dy_{ℓ^+}	0.02	0.03	0.03	0.02
dz_{ℓ^+}	0.02	0.04	0.03	0.02
p_{ℓ^+}	0.02	0.11	0.02	0.02
pValue_{ℓ^+}	0.01	0.01	0.01	0.01
p_{t,ℓ^+}	0.02	0.10	0.02	0.02
p_{x,ℓ^+}	0.02	0.04	0.02	0.02
p_{y,ℓ^+}	0.02	0.03	0.02	0.01

continued on next page

Table A.6 – continued from previous page

	$\mathbf{M_{bc}}$	θ_ℓ	$\theta_\mathbf{k}$	$\phi_\mathbf{B}$
$p_{z,\ell+}$	0.01	0.03	0.01	0.02
$\theta_{\ell+}$	0.01	0.02	0.01	0.02
$\lvert\vec{p}_{\ell+,\text{CMS frame}}\rvert$	0.04	0.13	0.02	0.02
M_{Recoil}	0.44	0.03	0.03	0.02
$M_{\text{event, missing}}$	0.03	0.08	0.05	0.02
$\lvert\vec{p_B}\rvert$	0.25	0.02	0.02	0.02
$\lvert\vec{p}_{\text{Recoil}}\rvert$	0.25	0.02	0.02	0.02
$\theta_{\text{missing } p}$	0.52	0.02	0.03	0.01
ϕ_B	0.01	0.01	0.02	1.00
$p_{t,B}$	0.54	0.02	0.02	0.01
$p_{x,B}$	0.22	0.02	0.02	0.01
$p_{y,B}$	0.22	0.01	0.01	0.01
$p_{z,B}$	0.21	0.02	0.02	0.02
$E(ROE)$	0.04	0.08	0.05	0.02
$E_{\text{extra}}(ROE)$	0.02	0.08	0.04	0.03
$M(ROE)$	0.05	0.09	0.05	0.02
$\lvert\vec{p}(ROE)\rvert$	0.07	0.03	0.02	0.02
$p_t(ROE)$	0.16	0.03	0.02	0.01
sphericity	0.06	0.20	0.12	0.02
θ_k	0.04	0.04	1.00	0.02
θ_ℓ	0.03	1.00	0.04	0.01
thrust_{event}	0.02	0.12	0.07	0.01
thrust_B	0.06	0.20	0.09	0.02
$E_{\ell+\ell-}$	0.04	0.02	0.02	0.01

Table A.7: Distance correlations of the BDT outputs on the background samples.

$B^0 \to (K^{*0} \to K^+\pi^-)e^+e^-$		
$\text{Output}_{\text{BDT 1}}$	M_{bc}	0.08377708291677033
$\text{Output}_{\text{BDT 1}}$	θ_k	0.09251298347027148
$\text{Output}_{\text{BDT 1}}$	θ_ℓ	0.1502394009476188
$\text{Output}_{\text{BDT 1}}$	ϕ_B	0.035109399027079886
$\text{Output}_{\text{BDT 2}}$	M_{bc}	0.06560019626736993
$\text{Output}_{\text{BDT 2}}$	θ_k	0.036760002768887606
$\text{Output}_{\text{BDT 2}}$	θ_ℓ	0.04493633100888982
$\text{Output}_{\text{BDT 2}}$	ϕ_B	0.037380696213341225
$B^0 \to (K^{*0} \to K^+\pi^-)\mu^+\mu^-$		

continued on next page

Table A.7 – continued from previous page

$\text{Output}_{\text{BDT 1}}$	M_{bc}	0.07969238939821299
$\text{Output}_{\text{BDT 1}}$	θ_k	0.030144400961631287
$\text{Output}_{\text{BDT 1}}$	θ_ℓ	0.09552654089892161
$\text{Output}_{\text{BDT 1}}$	ϕ_B	0.02356460526322155
$\text{Output}_{\text{BDT 2}}$	M_{bc}	0.053189916963236744
$\text{Output}_{\text{BDT 2}}$	θ_k	0.018923866797893308
$\text{Output}_{\text{BDT 2}}$	θ_ℓ	0.06381261111541672
$\text{Output}_{\text{BDT 2}}$	ϕ_B	0.03488528162217495
$B^+ \to (K^{*+} \to K^+\pi^0)e^+e^-$		
$\text{Output}_{\text{BDT 1}}$	M_{bc}	0.06583268826754061
$\text{Output}_{\text{BDT 1}}$	θ_k	0.06525357123543939
$\text{Output}_{\text{BDT 1}}$	θ_ℓ	0.17559421417493626
$\text{Output}_{\text{BDT 1}}$	ϕ_B	0.047183459200694326
$\text{Output}_{\text{BDT 2}}$	M_{bc}	0.07055699079881736
$\text{Output}_{\text{BDT 2}}$	θ_k	0.06765172016357826
$\text{Output}_{\text{BDT 2}}$	θ_ℓ	0.12497815836648081
$\text{Output}_{\text{BDT 2}}$	ϕ_B	0.039465787007005135
$B^+ \to (K^{*+} \to K^0_s\pi^+)e^+e^-$		
$\text{Output}_{\text{BDT 1}}$	M_{bc}	0.0705029447351758
$\text{Output}_{\text{BDT 1}}$	θ_k	0.07747996314938511
$\text{Output}_{\text{BDT 1}}$	θ_ℓ	0.07111369641607514
$\text{Output}_{\text{BDT 1}}$	ϕ_B	0.04722075872393238
$\text{Output}_{\text{BDT 2}}$	M_{bc}	0.07062581051891831
$\text{Output}_{\text{BDT 2}}$	θ_k	0.05658220752892994
$\text{Output}_{\text{BDT 2}}$	θ_ℓ	0.05289536906936217
$\text{Output}_{\text{BDT 2}}$	ϕ_B	0.0677118293277203
$\text{Output}_{\text{BDT 2}}$	ϕ_B	0.03488528162217495
$B^+ \to (K^{*+} \to K^+\pi^0)\mu^+\mu^-$		
$\text{Output}_{\text{BDT 1}}$	M_{bc}	0.043118561891196325
$\text{Output}_{\text{BDT 1}}$	θ_k	0.11440976382812508
$\text{Output}_{\text{BDT 1}}$	θ_ℓ	0.09013674609182583
$\text{Output}_{\text{BDT 1}}$	ϕ_B	0.040677672054097946
$\text{Output}_{\text{BDT 2}}$	M_{bc}	0.027627604880588786
$\text{Output}_{\text{BDT 2}}$	θ_k	0.09338371928876314
$\text{Output}_{\text{BDT 2}}$	θ_ℓ	0.06280938768528271
$\text{Output}_{\text{BDT 2}}$	ϕ_B	0.03240416856252249
$\text{Output}_{\text{BDT 2}}$	ϕ_B	0.03488528162217495
$B^+ \to (K^{*+} \to K^0_s\pi^+)\mu^+\mu^-$		

continued on next page

Table A.7 – continued from previous page

$\text{Output}_{\text{BDT 1}}$	M_{bc}	0.037478522760355085
$\text{Output}_{\text{BDT 1}}$	θ_k	0.030838433411260278
$\text{Output}_{\text{BDT 1}}$	θ_ℓ	0.09103659287380171
$\text{Output}_{\text{BDT 1}}$	ϕ_B	0.039380309768145105
$\text{Output}_{\text{BDT 2}}$	M_{bc}	0.038311354664147194
$\text{Output}_{\text{BDT 2}}$	θ_k	0.0412646979199185
$\text{Output}_{\text{BDT 2}}$	θ_ℓ	0.06722917493349598
$\text{Output}_{\text{BDT 2}}$	ϕ_B	0.03555730058708565

B. Correlation tables from fitting procedure

$\alpha_{\text{CB, sig}}$	$\sigma_{\text{CB, sig}}$	0.0	$N_{\text{CB, bkg}}$	χ_{argus}	0.142
$\alpha_{\text{CB, sig}}$	$\bar{x}$	-0.0	χ_{argus}	$N_{\text{CB, bkg}}$	-0.411
$\alpha_{\text{CB, sig}}$	n_{argus}	0.0	χ_{argus}	$N_{\text{CB, bkg}}$	0.142
$\sigma_{\text{CB, sig}}$	$\alpha_{\text{CB, sig}}$	0.0	$N_{\text{CB, sig}}$	$N_{\text{total, bkg}}$	-0.356
$\sigma_{\text{CB, sig}}$	$\bar{x}$	0.003	$N_{\text{CB, sig}}$	$\sigma_{\text{CB, sig}}$	0.349
$\sigma_{\text{CB, sig}}$	n_{argus}	0.0	$N_{\text{CB, sig}}$	χ_{argus}	-0.399
$\bar{x}$	$\alpha_{\text{CB, sig}}$	-0.0	$N_{\text{total, bkg}}$	$N_{\text{CB, sig}}$	-0.356
$\bar{x}$	$\sigma_{\text{CB, sig}}$	0.003	$N_{\text{total, bkg}}$	$\sigma_{\text{CB, sig}}$	-0.191
$\bar{x}$	n_{argus}	-0.0	$N_{\text{total, bkg}}$	χ_{argus}	0.205
n_{argus}	$\alpha_{\text{CB, sig}}$	0.0	$\sigma_{\text{CB, sig}}$	$N_{\text{CB, sig}}$	0.349
n_{argus}	$\sigma_{\text{CB, sig}}$	0.0	$\sigma_{\text{CB, sig}}$	$N_{\text{total, bkg}}$	-0.191
n_{argus}	$\bar{x}$	-0.0	$\sigma_{\text{CB, sig}}$	χ_{argus}	-0.266
$N_{\text{CB, bkg}}$	$\bar{x}$	-0.296	χ_{argus}	$N_{\text{CB, sig}}$	-0.399
$N_{\text{CB, bkg}}$	χ_{argus}	-0.411	χ_{argus}	$N_{\text{total, bkg}}$	0.205
$N_{\text{CB, bkg}}$	$N_{\text{CB, bkg}}$	-0.296	χ_{argus}	$\sigma_{\text{CB, sig}}$	-0.266

Table B.1: Correlations for all floating variables from the fitting procedure for the channel $B^0 \to K^{*0}e^+e^-$, as described in chapter 6.9.

$\alpha_{\text{CB, sig}}$	$\sigma_{\text{CB, sig}}$	0.0	$N_{\text{CB, bkg}}$	χ_{argus}	0.217
$\alpha_{\text{CB, sig}}$	$\bar{x}$	-0.0	χ_{argus}	$N_{\text{CB, bkg}}$	-0.483
$\alpha_{\text{CB, sig}}$	n_{argus}	0.0	χ_{argus}	$N_{\text{CB, bkg}}$	0.217
$\sigma_{\text{CB, sig}}$	$\alpha_{\text{CB, sig}}$	0.0	$N_{\text{CB, sig}}$	$N_{\text{total, bkg}}$	-0.299
$\sigma_{\text{CB, sig}}$	$\bar{x}$	-0.002	$N_{\text{CB, sig}}$	$\sigma_{\text{CB, sig}}$	0.282
$\sigma_{\text{CB, sig}}$	n_{argus}	0.0	$N_{\text{CB, sig}}$	χ_{argus}	-0.307
$\bar{x}$	$\alpha_{\text{CB, sig}}$	-0.0	$N_{\text{total, bkg}}$	$N_{\text{CB, sig}}$	-0.299
$\bar{x}$	$\sigma_{\text{CB, sig}}$	-0.002	$N_{\text{total, bkg}}$	$\sigma_{\text{CB, sig}}$	-0.199
$\bar{x}$	n_{argus}	-0.0	$N_{\text{total, bkg}}$	χ_{argus}	0.208
n_{argus}	$\alpha_{\text{CB, sig}}$	0.0	$\sigma_{\text{CB, sig}}$	$N_{\text{CB, sig}}$	0.282
n_{argus}	$\sigma_{\text{CB, sig}}$	0.0	$\sigma_{\text{CB, sig}}$	$N_{\text{total, bkg}}$	-0.199
n_{argus}	$\bar{x}$	-0.0	$\sigma_{\text{CB, sig}}$	χ_{argus}	-0.312
$N_{\text{CB, bkg}}$	$\bar{x}$	-0.318	χ_{argus}	$N_{\text{CB, sig}}$	-0.307
$N_{\text{CB, bkg}}$	χ_{argus}	-0.483	χ_{argus}	$N_{\text{total, bkg}}$	0.208
$N_{\text{CB, bkg}}$	$N_{\text{CB, bkg}}$	-0.318	χ_{argus}	$\sigma_{\text{CB, sig}}$	-0.312

Table B.2: Correlations for all floating variables from the fitting procedure for the channel $B^0 \to K^{*0}\mu^+\mu^-$, as described in chapter 6.9.

$\alpha_{\text{CB, sig}}$	$\sigma_{\text{CB, sig}}$	0.0	$N_{\text{CB, bkg}}$	χ_{argus}	0.189
$\alpha_{\text{CB, sig}}$	$\bar{x}$	0.0	χ_{argus}	$N_{\text{CB, bkg}}$	-0.45
$\alpha_{\text{CB, sig}}$	n_{argus}	0.0	χ_{argus}	$N_{\text{CB, bkg}}$	0.189
$\sigma_{\text{CB, sig}}$	$\alpha_{\text{CB, sig}}$	0.0	$N_{\text{CB, sig}}$	$N_{\text{total, bkg}}$	-0.437
$\sigma_{\text{CB, sig}}$	$\bar{x}$	-0.0	$N_{\text{CB, sig}}$	$\sigma_{\text{CB, sig}}$	0.42
$\sigma_{\text{CB, sig}}$	n_{argus}	0.0	$N_{\text{CB, sig}}$	χ_{argus}	-0.486
$\bar{x}$	$\alpha_{\text{CB, sig}}$	0.0	$N_{\text{total, bkg}}$	$N_{\text{CB, sig}}$	-0.437
$\bar{x}$	$\sigma_{\text{CB, sig}}$	-0.0	$N_{\text{total, bkg}}$	$\sigma_{\text{CB, sig}}$	-0.222
$\bar{x}$	n_{argus}	0.0	$N_{\text{total, bkg}}$	χ_{argus}	0.257
n_{argus}	$\alpha_{\text{CB, sig}}$	0.0	$\sigma_{\text{CB, sig}}$	$N_{\text{CB, sig}}$	0.42
n_{argus}	$\sigma_{\text{CB, sig}}$	0.0	$\sigma_{\text{CB, sig}}$	$N_{\text{total, bkg}}$	-0.222
n_{argus}	$\bar{x}$	0.0	$\sigma_{\text{CB, sig}}$	χ_{argus}	-0.297
$N_{\text{CB, bkg}}$	$\bar{x}$	-0.348	χ_{argus}	$N_{\text{CB, sig}}$	-0.486
$N_{\text{CB, bkg}}$	χ_{argus}	-0.45	χ_{argus}	$N_{\text{total, bkg}}$	0.257
$N_{\text{CB, bkg}}$	$N_{\text{CB, bkg}}$	-0.348	χ_{argus}	$\sigma_{\text{CB, sig}}$	-0.297

Table B.3: Correlations for all floating variables from the fitting procedure for the channel $B^+ \to K^{*+}e^+e^-$, as described in chapter 6.9.

$\alpha_{\text{CB, sig}}$	$\sigma_{\text{CB, sig}}$	0.0	$N_{\text{CB, bkg}}$	χ_{argus}	0.421
$\alpha_{\text{CB, sig}}$	$\bar{x}$	0.0	χ_{argus}	$N_{\text{CB, bkg}}$	-0.672
$\alpha_{\text{CB, sig}}$	n_{argus}	-0.0	χ_{argus}	$N_{\text{CB, bkg}}$	0.421
$\sigma_{\text{CB, sig}}$	$\alpha_{\text{CB, sig}}$	0.0	$N_{\text{CB, sig}}$	$N_{\text{total, bkg}}$	-0.4
$\sigma_{\text{CB, sig}}$	$\bar{x}$	-0.004	$N_{\text{CB, sig}}$	$\sigma_{\text{CB, sig}}$	0.458
$\sigma_{\text{CB, sig}}$	n_{argus}	0.0	$N_{\text{CB, sig}}$	χ_{argus}	-0.412
$\bar{x}$	$\alpha_{\text{CB, sig}}$	0.0	$N_{\text{total, bkg}}$	$N_{\text{CB, sig}}$	-0.4
$\bar{x}$	$\sigma_{\text{CB, sig}}$	-0.004	$N_{\text{total, bkg}}$	$\sigma_{\text{CB, sig}}$	-0.237
$\bar{x}$	n_{argus}	0.0	$N_{\text{total, bkg}}$	χ_{argus}	0.213
n_{argus}	$\alpha_{\text{CB, sig}}$	-0.0	$\sigma_{\text{CB, sig}}$	$N_{\text{CB, sig}}$	0.458
n_{argus}	$\sigma_{\text{CB, sig}}$	0.0	$\sigma_{\text{CB, sig}}$	$N_{\text{total, bkg}}$	-0.237
n_{argus}	$\bar{x}$	0.0	$\sigma_{\text{CB, sig}}$	χ_{argus}	-0.266
$N_{\text{CB, bkg}}$	$\bar{x}$	-0.542	χ_{argus}	$N_{\text{CB, sig}}$	-0.412
$N_{\text{CB, bkg}}$	χ_{argus}	-0.672	χ_{argus}	$N_{\text{total, bkg}}$	0.213
$N_{\text{CB, bkg}}$	$N_{\text{CB, bkg}}$	-0.542	χ_{argus}	$\sigma_{\text{CB, sig}}$	-0.266

Table B.4: Correlations for all floating variables from the fitting procedure for the channel $B^+ \to K^{*+}\mu^+\mu^-$, as described in chapter 6.9.