EPJ Web of Conferences **214**, 06032 (2019) https://doi.org/10.1051/epjconf/201921406032
CHEP 2018

The Belle II flavor tagger

Fernando Abudinén[1,2,*] for the Belle II Analysis Software Group

[1]Ludwig-Maximilians-Universität München, Excellence Cluster Universe, Boltzmannstr. 2, 85748 Garching, Germany.
[2]Max-Planck-Institut für Physik, Föhringer Ring 6, 80805 Munich, Germany.

Abstract.
Belle II is a particle-physics experiment at the intensity frontier focused on probing non Standard Model physics through precision measurements of quark-flavor and τ-lepton dynamics. Determining the flavor of neutral B mesons, i.e. their quark composition, is a crucial task which is addressed using flavor tagging algorithms. Due to the novel high-luminosity conditions and the increased beam backgrounds at Belle II, an improved flavor tagging algorithm had to be developed to ensure the success of the Belle II physics program.
The new Belle II flavor tagger exploits the flavor-specific signatures of B^0 decays employing boosted decision trees and neural networks. It identifies B^0-decay products providing flavor-specific signatures and combines the information from all possible signatures into a final output. The algorithm has been validated by comparing its performance on simulated events with its performance on collision events collected by the predecessor experiment Belle.
To explore the advantages of state-of-the-art deep-learning techniques, the Belle II collaboration developed a deep-learning-based flavor tagger. This algorithm tags the flavor of B^0 mesons without identifying flavor specific signatures using a deep-learning neural network. The validation on Belle data of this algorithm is currently ongoing.

1 Introduction

The Belle II experiment is located at the SuperKEKB energy-asymmetric electron-positron collider in Tsukuba, Japan. The experiment aims at constraining non Standard Model dynamics by probing processes that could receive contributions from new virtual heavy particles. Such contributions potentially cause small deviations from the SM predictions. To enhance the sensitivity to possible small deviations, a large data sample is required. With an unprecedented design luminosity of $8 \cdot 10^{35}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, about 40 times larger than its predecessor, SuperKEKB will lead ultimately to a Belle II data set about fifty times larger than the one collected by the predecessor experiment Belle.

A major part of the physics program is focused on the study of B-meson processes and in particular on the study of CP violation and flavor mixing in neutral B-meson decays. A neutral B meson is composed of a heavy b quark and a light d quark; either the heavy or the light one is an antiquark. At SuperKEKB, B mesons are produced in e^+e^- collisions at a center-of-mass energy of 10.58 GeV corresponding to the $\Upsilon(4\mathrm{S})$ resonance. The $\Upsilon(4\mathrm{S})$ decays almost

*e-mail: fernando.abudinen@ts.infn.it

half of the times into a neutral $B^0\overline{B}^0$ pair. Most measurements of *CP* violation and of mixing of neutral B mesons require the full reconstruction of the decay of one of the two neutral B mesons (signal side) and the determination of the flavor of the accompanying neutral B meson (tag side). The second task is referred to as flavor tagging and is accomplished using dedicated algorithms called flavor taggers.

Many decay modes of neutral B mesons provide flavor signatures through flavor-specific final states. Flavor signatures are characteristics of the decay products that are correlated with the flavor of the neutral B mesons, i.e. with the charge sign of the b quark composing it. For example, in $\overline{B}^0 \to D^{*+}l^-\overline{\nu}_l$ decays, a negatively charged lepton tags unambiguously a $\overline{B}^0$, which contains a negatively charged b, while a positively charged lepton tags a B^0, which contains a positively charged $\bar{b}$.

The Belle II category-based flavor tagger exploits the information provided by the different flavor signatures by sorting them into thirteen tagging categories [1]. The algorithm is a two-step process. In the first step, a multivariate analysis is carried out for each category, to identify the B^0-decay products providing flavor signatures. In the second step, the information of all categories is then combined to determine the tag-side B^0 flavor. The algorithm was developed with a dedicated optimization to exploit the capabilities of the new Belle II detector and the new Belle II reconstruction algorithms; it is designed to cope with the new challenging high-luminosity conditions and the increased beam backgrounds at SuperKEKB.

To explore the advantages of deep-learning multivariate methods, the Belle II collaboration also developed a deep-learning flavor tagger [2]. This algorithm determines the B^0 flavor in a single step, that is without pre-identifying the B^0 decay products. In the following, the Belle II flavor taggers and their performance will be discussed, with emphasis on the category-based approach which has been fully validated.

2 The new category-based flavor tagger

The category-based flavor tagger relies on flavor-specific decay modes [1]. Each decay mode has a particular decay topology and provides a flavor specific signature. Additional flavor signatures are obtained by combining similar or complementary decay modes. Each flavor signature becomes a category of the algorithm. The particles providing flavor signatures are called targets. Table 1 shows an overview of all thirteen categories together with the underlying decay modes and the corresponding targets.

The algorithm uses discriminating input variables to identify the targets among all the particles on the tag side. There are two types of input variables: particle identification (PID) variables, which combine the PID information provided by the various sub-detectors, and kinematic variables, which are sorted into simple and higher-level variables. Simple kinematic variables include for example momenta and impact parameters. Higher-level kinematic variables need global information provided by all particles reconstructed on the tag side. Examples of such variables are the recoil mass and the energy in the direction of the W boson (assuming a semileptonic decay).

In comparison with the previous Belle algorithm, the new Belle II category-based flavor tagger considers more flavor signatures and more input variables, and is based on multivariate methods avoiding the previous cut-based approach [1].

2.1 Algorithm

The Belle II flavor tagging algorithms provide an output value $y \in [-1, 1]$ equivalent to $y = q \cdot r$, where $q = \mathrm{sgn}(y)$ is the flavor of the tag-side B^0 meson (B^0_{tag}), and $r = |y|$ is the

Table 1: Tagging categories and their targets (left) with examples of the considered decay modes (right). Here, p^* stands for momentum in the center-of-mass frame and $\ell^\pm$ for charged leptons (μ^- or e^-).

Categories	Targets for $\overline{B}^0$
Electron	e^-
Intermediate Electron	e^+
Muon	μ^-
Intermediate Muon	μ^+
Kinetic Lepton	ℓ^-
Intermediate Kinetic Lepton	ℓ^+
Kaon	K^-
Kaon-Pion	K^-, π^+
Slow Pion	π^+
Maximum p^*	ℓ^-, π^-
Fast-Slow-Correlated (FSC)	ℓ^-, π^+
Fast Hadron	π^-, K^-
Lambda	Λ

Underlying decay modes

$\overline{B}^0 \to D^{*+}\ \overline{\nu}_\ell\ \ell^-$
$\quad\hookrightarrow D^0\ \pi^+$
$\quad\quad\hookrightarrow X\ K^-$

$\overline{B}^0 \to D^+\ \pi^-\ (K^-)$
$\quad\hookrightarrow K^0\ \nu_\ell\ \ell^+$

$\overline{B}^0 \to \Lambda_c^+\ X^-$
$\quad\hookrightarrow \Lambda\ \pi^+$
$\quad\quad\hookrightarrow p\ \pi^-$

so-called flavor dilution factor. A dilution factor $r = 0$ corresponds to a fully diluted flavor (no possible distinction between B^0 and $\overline{B}^0$) and a dilution factor $r = 1$ to a perfectly tagged flavor; the probability for a right flavor tag corresponds to $(r + 1)/2$. When $y = -1$, B^0_{tag} is perfectly tagged as $\overline{B}^0$, and when $y = +1$, B^0_{tag} is perfectly tagged as B^0.

The category-based algorithm is organized into a two-step process: event and combiner level. On the event level, individual tag-side tracks are considered. The tag-side tracks are those remaining from the full reconstruction of the signal B meson. For five types of particles, e, μ, K, π and p, an independent list of particle candidates is built by assigning the corresponding PDG mass to the tag-side tracks. Additionally, Λ particles are reconstructed from pairs of proton and pion candidates. The event-level process is performed for each category. Each category considers the list of particle candidates belonging to its own targets.

Figure 1 illustrates the procedure for an example category: a multivariate method gets a specific set of input variables (PID and kinematic variables) exploiting the characteristics of the associated decay mode. Some input variables require information from all reconstructed tracks and all neutral clusters on the tag side. Neutral clusters are clusters in the electromagnetic calorimeter (ECL) and in the instrumented iron yoke (KLM) of the Belle II detector that are not associated with a track (charged particle). Two special categories get information from other categories: the Kaon-Pion category and the Fast-Slow-Correlated (FSC) category.

Summing the input variables for all categories yields a total number $n_{\text{input}} = 220$. Some variables are used multiple times for the same candidates in different categories. To save computing time, each variable is calculated only once for each candidate, reducing the number of variables to be calculated to $n_{\text{unique}} = 108$.

The event-level multivariate method assigns to each particle candidate an output value $y_{\text{cat}} \in [0, 1]$ corresponding to the probability of being the target of the corresponding cate-

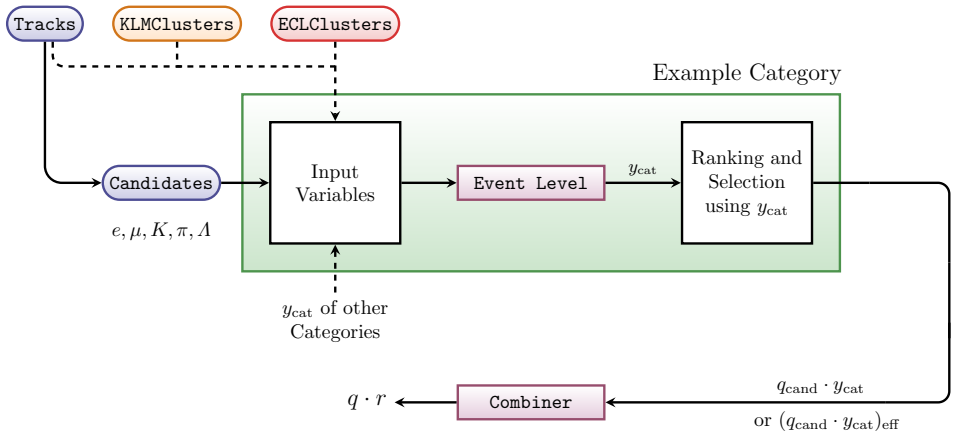

Figure 1: Procedure for each single category (green box): the candidates correspond to the reconstructed tracks for a specific mass hypothesis. Some of the input variables consider all reconstructed tracks and all neutral ECL and KLM clusters on the tag side. The magenta boxes represent multivariate methods: y_{cat} is the output of the event level. The output of the combiner is equivalent to the product $q \cdot r$.

gory providing the right flavor tag. Within each category, the particle candidates are ranked according to the values of y_{cat}. The candidate with the highest y_{cat} is selected as target. Only for the Maximum p^* category, the target is the candidate with the largest momentum in the $\Upsilon(4S)$ frame.

All categories contribute to the final tag. This improves the performance of the flavor tagger since the B^0_{tag} decay may offer more than one flavor-specific signature. The combiner corresponds to a multivariate method that receives thirteen input values, i.e. one input value from each category, and outputs $y = q \cdot r$. Each input value is the product $q_{\text{cand}} \cdot y_{\text{cat}}$ of each category, where the charge q_{cand} and the probability y_{cat} correspond to the particle candidate selected as target. For two special cases, the Kaon and the Lambda categories, the input is the effective product $(q_{\text{cand}} \cdot y_{\text{cat}})_{\text{eff}}$ of the three particles with the highest y_{cat} value. This improves by about 7% the performance of the flavor tagger. For the Lambda category, q_{cand} corresponds to the B^0 flavor tagged by the Λ candidate, i.e. $q_\Lambda = -1(+1)$ for $\Lambda(\bar{\Lambda})$.

The multivariate method chosen for the event and the combiner level is the Fast Boosted Decision Tree (FBDT), a stochastic gradient-boosted decision tree developed especially for Belle II [3]. It incorporates several mechanisms for regularization and is optimized to save computing resources during its application and during its training procedure. To cross-check the result of the FBDT, an independent multivariate method, a Multi-Layer Perceptron (MLP), is employed for the combiner level. The implementation of the MLP is based on the open-source library FANN [4]. For both combiner level methods, the input values are identical. The flavor tagger provides the output of both the FBDT and the MLP combiners.

To avoid biases from statistical correlations, the algorithm is trained using two statistically independent MC samples: one sample for the event level, and one sample for the combiner level. At each training step, one half of the sample is used as training sample and the other half as a test (validation) sample for an unbiased evaluation of the performance against over-

fitting. The event level is trained first and each category is trained independently. The FBDT and the MLP combiners are trained afterward.

Over-fitting is checked for each of the multivariate methods by comparing the distribution of the output from the training sample with that from the test sample. The output on the training sample has to be statistically compatible with that of the testing sample. For the MLP, the validation sample is used actively for the stopping criterion during the training procedure [5, 6].

2.2 Performance and validation

The performance of the category-based flavor tagger is evaluated using simulated Belle II events, simulated Belle events and Belle collision data within the Belle II software framework. The simulated events used for training and testing correspond to $B^0\overline{B}^0$ pairs in which one meson (B^0_{sig}) decays to $J/\psi K^0_S$ while B^0_{tag} decays to any possible final state according to the known rates. Only events where the benchmark decay channel $B^0_{\text{sig}} \to J/\psi \left[\to \mu^+\mu^-\right] K^0_S \left[\to \pi^+\pi^-\right]$ could be fully reconstructed and correctly matched with the simulated decay chain are selected for training and testing. After the selection, the size of the Belle II and the Belle training samples is respectively about 1.3 and 1 million simulated events (same size for the event and the combiner level), and the size of the testing samples is about 2.6 and 2 million MC events. Tests with larger training samples showed no considerable improvement of the performance. After training and testing the algorithm with different samples, the evaluation of the results showed that it is mandatory to generate the training sample without built-in *CP* violation to avoid that the algorithm learns *CP* asymmetries on the tag side [1].

Figure 2 (left) shows the output $y = q \cdot r$ for the default FBDT combiner on simulated Belle II events. The dashed blue and the solid red curves are the output distributions for events with true $\overline{B}^0$ and true B^0 mesons on the tag side. The dotted black curve shows the total distribution. Figure 2 (right) shows a linearity check between the true dilution r_{MC} determined using the simulated Monte Carlo information and the mean $\langle r \rangle$ of the dilution provided by the combiners. The dilution determined using simulated information is defined as $r_{\text{MC}} = |1 - 2w_{\text{MC}}|$, where the w_{MC} is the fraction of wrongly tagged events determined by comparing the truth with the FBDT combiner output. The mean dilution $\langle r \rangle$ of the FBDT combiner output is simply the mean of $|q \cdot r|$ for each r bin. The linearity verifies the equiva-

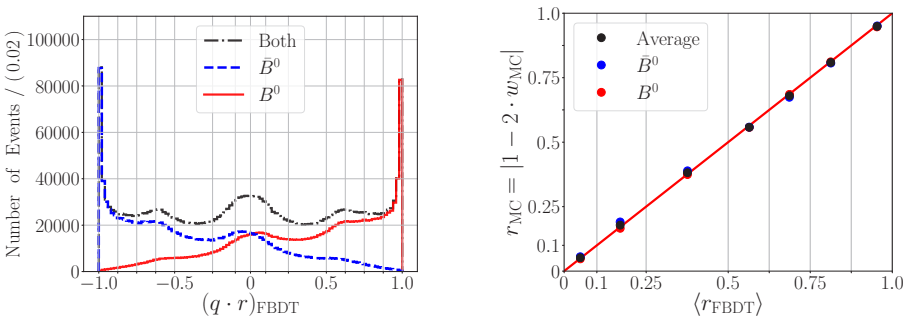

Figure 2: Performance of the default FBDT combiner on Belle II MC without simulated background: (left) combiner output, (right) correlations between $r_{\text{MC}} = |1 - 2w_{\text{MC}}|$ and $\langle |q \cdot r| \rangle$ for each r bin.

lence in average between the output $q \cdot r$ and the product $q_{\mathrm{MC}} \cdot r_{\mathrm{MC}}$. The linearity is checked for events with true B^0 and true $\overline{B}^0$ and in average for both kinds of events.

An important outcome is that the simulated events to which none of the thirteen categories can be attributed (about 7.2% of the events) have no further discrimination power between B^0 and $\overline{B}^0$, showing that the algorithm is complete [1].

The validation on Belle data is performed on a set of $B^0\overline{B}^0$ pairs, where the same benchmark decay channel is reconstructed on the signal side. The events are obtained from the full Belle data sample corresponding to $772 \cdot 10^6$ $B\overline{B}$ pairs. Belle collision events are reconstructed in the same way as simulated Belle events [1]. To analyze Belle data and Belle simulation, the event format of Belle (based on PANTHER tables) is converted into the Belle II format (based on ROOT objects) as described in [7].

Figure 3 (left) shows the results using Belle data and Belle simulation by superimposing the normalized $q \cdot r$ output distributions. Within the uncertainties, the shapes of the normalized $q \cdot r$ distributions for Belle data and Belle simulation show good agreement. Figure 3 (right) shows the linearity check for simulated Belle events.

In general, the effective efficiency $\varepsilon_{\mathrm{eff}}$ of a flavor tagger is defined such that the statistical uncertainty on the measured CP asymmetries is proportional to $1/\sqrt{N \cdot \varepsilon_{\mathrm{eff}}}$, where N is the total number of events. The effective efficiency $\varepsilon_{\mathrm{eff}}$ corresponds to the effective reduction of events due to the flavor dilution r. For the calculation of $\varepsilon_{\mathrm{eff}}$, Belle II adopts the r binning applied by the Belle experiment [8].

The default FBDT combiner achieves a slightly better performance than the MLP combiner on all samples. On simulated Belle II events without background, the FBDT combiner reaches an effective efficiency $\varepsilon_{\mathrm{eff}}$ of about 37%, and with simulated background about 34%. On Belle data, the FBDT combiner reaches $(33.5 \pm 0.5\%)$. These figures show a relative 10% improvement over the Belle flavor tagging performance. The improvement corresponds effectively to an enlargement of the data sample by about the same percentage since $N \cdot \varepsilon_{\mathrm{eff}}$ corresponds to the number of effectively tagged events.

An advantage of the category-based approach is that it can be trained to use only primary-lepton categories. Using only primary-lepton categories is the only way to fully eliminate systematic uncertainties caused by the so-called tag-side interference effect [9]. This, however, reduces the effective efficiency by about a factor 3.

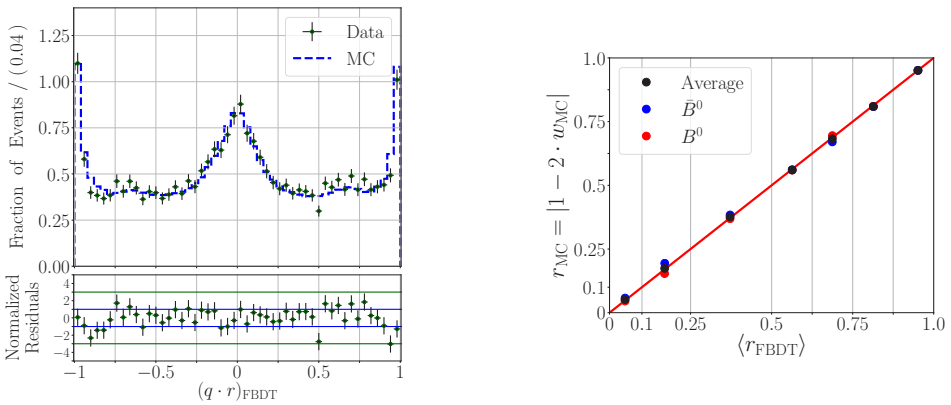

Figure 3: Performance of the FBDT combiner: (left) normalized $q \cdot r$ distributions on Belle data and on Belle MC and (right) correlations between $r_{\mathrm{MC}} = |1 - 2w_{\mathrm{MC}}|$ and $\langle |q \cdot r| \rangle$ on Belle MC.

The effective efficiency on Belle data is calculated assuming the wrong tag fractions obtained from simulated events. The good agreement between the results on Belle data and on Belle simulation and the linearity between the average output dilution and the true MC dilution support this approach. The true wrong tag fractions will be determined on Belle II data once it becomes available by measuring the $B^0 - \overline{B}^0$ mixing.

3 Alternative deep-learning flavor tagger

The deep-learning flavor tagger is based on an MLP with 8 hidden layers [2]. The deep-learning MLP receives as input variables the characteristics of the reconstructed tracks on the tag side, and gives as output the product $q \cdot r$.

The deep-learning MLP is designed to learn the correlations between the characteristics of the tag-side tracks and the flavor of B^0_{tag} avoiding any pre-selection of B^0 decay products. The deep-learning MLP implementation is based on the machine learning library TensorFlow [10].

The deep-learning flavor tagger sorts the tracks on the tag side into two groups, a positive and a negative one, according to particle's electric charge. The algorithm ranks the tracks in each group according to their momenta in the $\Upsilon(4S)$ frame, and selects the top five tracks in each group. If an event contains less than five positive or less than five negative tracks, the algorithm sets the input variables for the missing candidates to zero. For each candidate, the deep-learning MLP receives 14 input variables corresponding to PID variables, simple kinematic variables, the number of hits in the tracking detectors, and the p-value of the track fit. Multiplying the number of input variables by the number of candidates yields 140, corresponding to the number of input nodes.

The deep-learning algorithm is optimized and evaluated using simulated Belle II samples without background and simulated Belle samples. Table 2 lists the effective efficiencies of the category-based and the deep-learning flavor taggers on simulated Belle II and simulated Belle events. The deep-learning flavor tagger clearly outperforms the category-based one on simulated Belle II events, but not on simulated Belle events on which both flavor taggers reach a similar effective efficiency. Further studies are needed to understand why the improvement in performance is not observed on Belle simulation. A direct comparison of both taggers is still to be performed for different ranges of the dilution factor r. Comparing the output of the individual categories with the output of the deep-learning flavor tagger could test if the MLP of the deep-learning flavor tagger learns physical information or features of the simulation.

Table 2: Effective efficiencies of the category-based flavor tagger and the deep-learning flavor tagger on simulated Belle II events without background and on simulated Belle events. All values are given in percent.

	Belle II MC	Belle MC
Approach	$\varepsilon_{\text{eff}} \pm \delta\varepsilon_{\text{eff}}$	$\varepsilon_{\text{eff}} \pm \delta\varepsilon_{\text{eff}}$
Category-based	36.64 ± 0.05	34.18 ± 0.06
Deep-learning	40.69 ± 0.03	34.42 ± 0.09

The complex MLP architecture requires large training samples. The best results have been obtained with a sample of 55 million events; the tendency shows that the performance could still improve with larger samples.

The MLP complexity and the large training sample sizes call for significant computing resources to train the algorithm and to prepare the training samples. To exploit their parallel

computation capabilities, GPUs are used to train the deep-learning MLP. On a GTX970 GPU, the training procedure for the 8-layers MLP takes about 48 hours. In comparison, the training procedure for the category-based flavor tagger takes about 5 hours running on a single CPU core.

4 Conclusion and outlook

Flavor tagging is crucial for the success of the Belle II physics program. Belle II has deployed a large effort to develop high-performance flavor tagging algorithms that optimally exploit the enhanced capabilities of the experiment and make use of state-of-the-art multivariate methods to cope with the challenging high-luminosity conditions.

The optimization of the category-based algorithm is now complete: the performance evaluation studies show that the algorithm uses the full flavor information provided by flavor-specific B^0-meson decays. The algorithm considers more flavor signatures and input variables than the previous Belle algorithm, and is based on multivariate methods which replaced the cut-based procedure at Belle. As the validation studies on Belle data showed, the new Belle II algorithm reaches an effective efficiency about 10% (relative) higher than the Belle algorithm. Even under the expected conditions of unprecedented high luminosity, the new Belle II category-based flavor tagger is expected to outperform the Belle flavor tagger by about the same percentage. The rise in effective efficiency corresponds effectively to an increase of the data sample (in addition to the increase due to the high luminosity) and will contribute to the sensitivity of the Belle II experiment in searches for non-SM physics [11].

The deep-learning-based flavor tagger is currently being validated on Belle data. Its performance on simulated Belle II events without background is very promising. Tests using simulated Belle II events with background are planned for the near future.

The differences in the performance of the two flavor taggers still need understanding. The goal is to validate and calibrate both flavor taggers. A direct comparison of the two approaches will offer unique possibilities to cross check the results.

The first calibration of the flavor taggers on collision data is planned for summer 2019. The calibration will include a first study of systematic effects; studies on possible new systematic effects due to the novel high-luminosity conditions still have to be performed.

References

[1] F. Abudinén, Ph.D. thesis, LMU Munich (2018), BELLE2-PTHESIS-2018-003
[2] J.F. Gemmler, Master's thesis, KIT (2016), EKP-2016-00033
[3] T. Keck, CoRR **abs/1609.06119** (2016)
[4] S. Nissen, Report, DIKU Copenhagen **31**, 29 (2003)
[5] F. Abudinén, Master's thesis, LMU Munich (2014), BELLE2-MTHESIS-2018-007
[6] S. Pohl, Ph.D. thesis, LMU Munich (2017), BELLE2-PTHESIS-2018-001
[7] T. Keck, Ph.D. thesis, KIT (2017), EKP-2017-00067
[8] A.J. Bevan et al. (Belle, BaBar), Eur. Phys. J. **C74**, 3026 (2014), 1406.6311
[9] O. Long, M. Baak, R.N. Cahn, D.P. Kirkby, Phys. Rev. **D68**, 034010 (2003), hep-ex/0303030
[10] M. Abadi, et al., *TensorFlow: Large-scale machine learning on heterogeneous systems* (2015), software available from tensorflow.org, https://www.tensorflow.org/
[11] E. Kou et al. (2018), 1808.10567