Noname manuscript No.
(will be inserted by the editor)

B2BII

Data conversion from Belle to Belle II

Moritz Gelb[1] · Thomas Keck[1] · Markus Prim[1] · Hulya Atmacan[2] · Jochen Gemmler[1] · Ryosuke Itoh[3] · Bastian Kronenbitter[1⋆] · Thomas Kuhr[4] · Matic Lubej[5] · Felix Metzner[1] · Chanseok Park[6] · Seokhee Park[6] · Christian Pulvermacher[1⋆] · Martin Ritter[4] · Anze Zupanc[5⋆]

Received: date / Accepted: date

Abstract We describe the conversion of simulated and recorded data by the Belle experiment to the Belle II format with the software package `b2bii`. It is part of the Belle II Analysis Software Framework. This allows the validation of the analysis software and the improvement of analyses based on the recorded Belle dataset using newly developed analysis tools.

Keywords b2bii, belle, belle2, hep , basf2, data preservation

1 Introduction

The Belle experiment recorded a dataset of approximately $1\,\mathrm{ab}^{-1}$ during its runtime; mainly at the $\Upsilon(4S)$ resonance. The physics program was very successful with milestones such as the measurement of mixing-induced CPV in $B^0 \to J/\Psi K^0_S$ decays leading to the Noble Prize for Kobayashi and Maskawa in 2008 [1], the precise measurement of the CKM matrix elements [2], and the discovery of tetra quarks [3].
Its successor, the Belle II experiment, will soon start to record the first collisions. To allow for the envisaged 40-times higher peak luminosity, the collider and detector were upgraded. In addition, the Belle II Analysis Software Framework (`BASF2`) [4] was developed from scratch. An thorough validation of the software is necessary to ensure the integrity of upcoming analyses.
In this article we describe the software package (`b2bii`) based on [5], which converts simulated and recorded Belle events into the Belle II format.

1.1 The Belle & Belle II Detector

The design of the Belle II detector resembles it predecessor. Each individual sub-detector is upgraded with a modern version of itself. For a detailed description of the Belle and Belle II detectors see reference [6] and [7], respectively.

Going outwards from the interaction point (IP) the Belle detector consisted of a four layer silicon strip detector (SVD), a central drift chamber (CDC), an Aerogel Cherenkov counter (ACC), a time-of-flight (TOF) detector system, an electromagnetic calorimeter (ECL), a superconducting solenoid which provided a homogeneous magnetic field of 1.5 T, and a return yoke, which was instrumented with glass-electrode resistive plate counters for K_L and muon detection (KLM).

Going outwards from the IP the Belle II detector consists of a two layer pixel detector (PXD), a four layer silicon strip detector (SVD), a central drift chamber (CDC), a proximity-focusing Aerogel ring-imaging Cherenkov detector (ARICH), a time-of-propagation counter (TOP), an electromagnetic calorimeter (ECL), a superconducting solenoid which provides a homogeneous magnetic field of 1.5 T, and a return yoke, which is instrumented with glass-electrode resistive plate coun-

T. Keck
Karlsruhe Institute of Technology
Institute of Experimental Particle Physics
Wolfgang-Gaede-Str. 1
76131 Karlsruhe
E-mail: thomas.keck2@kit.edu

[1] Karlsruhe Institute of Technology
[2] University of South Carolina
[3] High Energy Accelerator Research Organization (KEK)
[4] Ludwig Maximilians University Munich
[5] University of Ljubljana
[6] Yonsei University
[⋆] Research was performed while the author was affiliated with the corresponding institute

ters in the barrel region and scintillator strip in the end-caps for K_L and muon detection (KLM).

1.2 Recorded Belle Data

Most of the Belle data was recorded at the center-of-mass energy of the $\Upsilon(4S)$ resonance. In addition, data was also recorded at the $\Upsilon(1S)$, $\Upsilon(2S)$, $\Upsilon(3S)$ and $\Upsilon(5S)$ resonances. Moreover, off-resonance data, mostly used to study non-resonant background processes, was recorded.

The raw data coming from the detector was calibrated, reconstructed and stored on tape using PANTHER-based data summary tape (DST) files. PANTHER is a custom serialization format [8]. After each experiment the calibration constants were recomputed by detector experts or computed directly from data, and stored in the Belle Condition Database, based on PostgreSQL. Finally, the data of the completed experiment was reprocessed and stored in a compact form called mDST files, a reduced and compressed form of the data summary tape files. The reconstruction and the processing of the mDST files is handled by the Belle AnalySis Framework (BASF) [9]. Different types of events were simulated using the EvtGen [10] and GEANT3 [11] packages, and reconstructed with the same software as was used for the detector data.

1.3 Anticipated Belle II Data

By 2025, Belle II will record $50\,\mathrm{ab}^{-1}$ of data, which corresponds to 50 times the integrated luminosity of Belle. The same software framework is used in online data taking and offline reconstruction, Monte Carlo production, and physics analysis. After time-dependent calibration parameters are determined, the raw data is reconstructed and stored at the KEK computing center[1]. The time-dependent calibration parameters are stored in the Belle II Condition Database [12] [13]. Monte Carlo production and reconstruction will be distributed to data centers around the world. The reconstructed information is stored in ROOT-based [14] mDST files.

1.4 Data Processing Levels

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

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

Converting the raw-data is in principle possible, but the differences between the Belle and Belle II detector render this a difficult and ill-defined task. While this would allow for the validation of the Belle II reconstruction software (e.g. the track finding and fitting algorithms) on Belle data, this would be only of limited use due to the differences between the Belle and the Belle II detector, the vastly different expected background, and the availability of events recorded by Belle II from cosmic runs.

The Belle to Belle II dataset conversion converts the Belle mDST data, which contains mostly detector independent objects like tracks and calorimeter clusters, into the new mDST format used by BASF2. This enables the validation of the Belle II analysis software, and (re-)production of Belle measurements using the improved software.

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

The Belle experiment provides a large amount of Monte Carlo simulated events, which can be processed using b2bii. However, the production of additional Monte Carlo simulated Belle events still requires BASF and is not part of b2bii.

2 Method

The software responsible for reading in the old PANTHER data-format and representing the data in memory was isolated, cleaned up and compiled into a new library named belle_legacy. A new package was introduced in BASF2 called b2bii (Belle to Belle II). It contains three BASF2 modules developed with the help of the belle_legacy library. The conversion process is visualized in Figure 1.

[1] Other computing centers will store additional copies of the raw data.

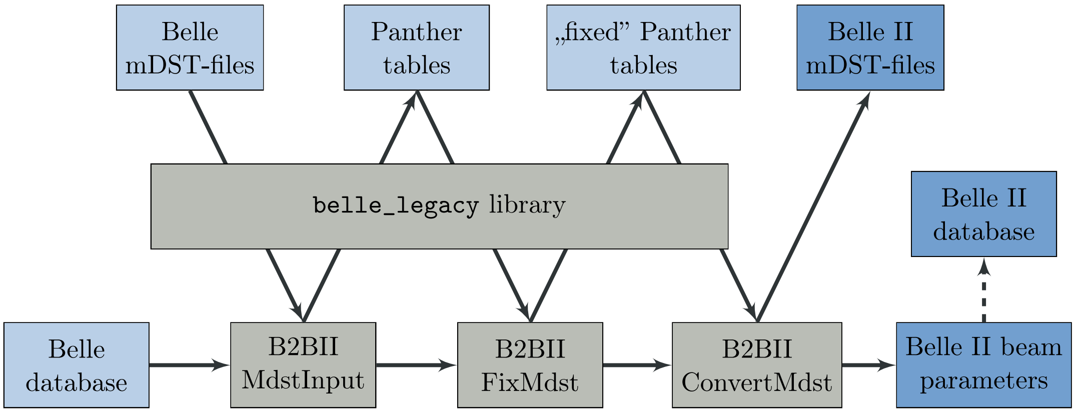


Fig. 1: Schematic view of the conversion process of Belle (light blue) to Belle II (blue) mDST files using the BASF2 modules (gray) provided by the b2bii package and the original Belle software provided by the belle_legacy library (gray).

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

The B2BIIFixMdst module applies various calibration factors onto the PANTHER tables, for instance on the beam-energy, the momenta and error matrices of the fitted tracks, the energy deposition in the ECL, and the particle identification information of the CDC and TOF. It also performs standard cuts to ensure that the selection of the detector data and simulated events is identical to the one obtained with BASF. Finally, π^0 candidates are reconstructed from the γ particle objects and the corrected ECL clusters. An equivalent module named FixMdst was already used by BASF.

The B2BIIConvertMdst module converts the information stored in the Belle PANTHER tables into the corresponding Belle II ROOT objects. The beam-energy and IP-profile is collected in the BASF2 BeamParameters object and stored in the condition Database of Belle II.

2.1 Data Formats

The Belle data format is based on a custom serialization format called PANTHER. It consists of tables compressed by the zlib library. The table formats are defined by ASCII header files. Each table consists of multiple rows, called entries. The index of each entry can be used to relate entries to other entries. For instance, to express a mother–daughter relationship between particles (a particle which decays inside the detector is called the mother of its decay-products, which are the daughters). BASF processes the data event-by-event, meaning the in-memory representation of the PANTHER tables contain only a single event at a given point. After each event the tables are flushed and filled with the next event.

The Belle II data format is based on ROOT. The ROOT framework takes care of serialization including potential migrations to ensure backward-compatibility. Conceptually we distinguish different types of ROOT objects. **Array objects** are the equivalent to the tables used by PANTHER. The entries of different array objects can be connected by adding so-called relations, which are stored in a separate array object. The relations allow the expression of many-to-many connections between arbitrary entries of the array objects. For instance the relation between a track and the associated clusters: this allows the analyst to easily access all clusters which are associated to a given track. **Single objects** are used to store the remaining event-wise information. For instance the meta data of each event, or particle lists created by the analyst. A particle list is a list of Particle array entries used to organize the reconstruction of decay-chains in BASF2. BASF2 processes the data event-by-event, meaning the in-memory representation of most ROOT objects contain only a single event at a given point. After each event the objects are filled with the next event. Some ROOT objects are persistent in the sense that they are only stored and loaded once per file. For instance, the meta data of each file or statistics of the execution time and memory consumption used for profiling.

2.2 Implementation Details

The detailed matching between PANTHER tables and corresponding BASF2 data-objects is shown in Figure 2. In the following we describe the conversion process in detail for future reference.

2.2.1 Event Information

Event information like the beam energy and position of the IP are loaded from the Belle condition database and stored in `BeamParameters` objects that can be uploaded to the Belle II condition database. The `BeamParameters` of the entire detector data was converted and uploaded. The `BeamParameters` of simulated events are only stored on the local machine.

The description of the magnetic field differs between Belle and Belle II. The conversion uses a magnetic field map which is consistent with the track parametrization in Belle data.

2.2.2 Monte Carlo

The Monte Carlo information of Belle is stored in the so-called `Gen_hepevt` table. It contains the four momenta of the generated particles and the indices of the mother and all daughter particles. The table is converted into an array of `MCParticle` objects, which contains the same information. Consequently, the fine-grained unified Monte Carlo matching algorithm of `BASF2` can be used, and problems contained in algorithms used by `BASF` are avoided [15, sec. 4.3].

The `Gen_hepevt` table includes special entries for a common mother of beam-background particles (PDG code 911) and for virtual photons (PDG code 0). These entries are ignored during the conversion, because there are no corresponding concepts in Belle II. For instance, in `BASF2` beam-background is indicated by a motherless Monte Carlo particle.

The original Belle software does not provide Monte Carlo information for KLM clusters, following the approach of [16, sec. 5.2] true K_L^0 are matched to the closest reconstructed Monte Carlo K_L^0 within ±15 degrees in both θ and ϕ.

Furthermore, unlike Belle II simulated events, the Belle simulated events do not provide information on the differentiation between photons generated directly by the fundamental matrix-element calculated by the Monte Carlo generator `EvtGen` (hereinafter referred to as gamma) and photons generated afterwards for instance by `PHOTOS` [17] or the simulation (hereinafter referred to as final state radiation) (see [18, Appendix C]). Often a reconstructed particle which misses final state radiation is considered as signal, whereas a reconstruction with a missing gamma is considered as wrong. A simple heuristic is applied to distinguish the two cases: Photons from a decay $M \to AB...\gamma$ are flagged as final state radiation, and photons from a decay $M \to A\gamma$ are flagged as gammas. In particular photons from $\pi^0 \to \gamma\gamma$ and $\mathrm{D}^* \to \mathrm{D}\gamma$ are considered gammas. Other cases like $\mathrm{B} \to \mu\nu\gamma$ are regarded by the heuristic as final state radiation and have to be treated by the analyst themself[2].

The official Belle Monte Carlo campaigns produced ten times the real integrated luminosity in $\mathrm{B}\overline{\mathrm{B}}$ events and six times that in continuum events, however some inconsistencies were encountered during the development of the conversion software, which were fixed if possible: The Monte Carlo campaign deleted the 8 leftmost bits of the 32 bit long PDG codes during the Monte Carlo simulation[3]. During the conversion these corrupted PDG codes are restored by matching their lower 24 bit to known PDG codes. In the official Belle Monte Carlo campaign from 2010 for $\mathrm{B} \to \mathrm{u}\ell\nu$ and other rare B decays, the mass of almost all MC particles is set to zero, which can lead to wrong results if this quantity is used during the analysis. However, this information is redundant since the correct mass of the MC particles can be calculated using either the PDG values or the MC four-momenta.

2.2.3 Tracks

The track reconstruction output of `BASF` is stored in the so-called `Mdst_charged` and `Mdst_trk` tables. They contain the 5D track parametrization for up to five different final state particle hypotheses. The track parametrization is transformed and stored into `Track` and associated `TrackFitResult` array objects. The transformation is unique but non-trivial because the two experiments employ different 5D track parameterizations and conventions for the reference point of the track.

2.2.4 ECL Clusters

The output of the ECL cluster algorithm of Belle is stored in the so-called `Mdst_ecl` and `Mdst_ecl_aux` tables. They contain information about the energy, position and shape of the clusters. The ECL information is converted and stored in the `ECLCluster` array object. Information is mapped to the corresponding representation, e.g. the energy and position of the clusters with the

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

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

corresponding covariance matrix and shower variables, such as the $E9E25$ ratio (which stored in the field for the $E9E21$ ratio as this is the one now used in Belle II). Advanced shower variables like Zernike moments were not available for Belle and are therefore not set. In addition two ParticleList objects are created containing the γ and π^0 candidates, which were created by B2BIIFixMdst earlier. The lists are named gamma:mdst and pi0:mdst, respectively. The ParticleList objects provide a fast and easy access to the possible γ and π^0 candidates, used by the analyst during their analysis.

2.2.5 KLM Clusters

The output of the KLM cluster algorithm of Belle is stored in the so-called Mdst_klm_cluster and Mdst_klong tables. The KLM information is converted and stored in the KLMCluster array object. In addition a ParticleList is filled containing K^0_L candidates. The list is named K_L0:mdst.

2.2.6 V0 Objects

A V0 object is a pair of tracks with a common vertex usually outside of the beam pipe. Such a signature indicates the decay of a particle with a relatively long lifetime like a K^0_S. The output of the V0 Finder of Belle is stored in the so-called Mdst_vee_daughters and Mdst_vee tables. Additional information is created on-the-fly by the nisKsFinder, which provides quality information. The V0 information is directly transformed into ParticleList objects containing candidates for γ, K^0_S and Λ. The lists are named gamma:v0mdst, K_S0:mdst and Lambda0:mdst, respectively. The additional quality information is stored in the ExtraInfo field of the Particle array entries under the keys goodKs, ksnbVLike, ksnbNoLam and ksnbStandard.

2.2.7 PID Information

The PID information provided by the different Belle sub-detectors is stored in the so-called kid_acc, Mdst_tof. kid_cdc, eid and Mdst_klm_mu_ex tables. It is mapped to similar Belle II sub-detectors, so that the physical meaning of the information is partially preserved. In particular the Belle time-of-flight (TOF) and Aerogel Cherenkov counter (ACC) detectors are mapped to the Belle II time-of-propagation (TOP) and Aerogel ring imaging Cherenkov (ARICH) detectors, respectively. The converted information is stored in the PIDLikelihood array object.

2.2.8 Relations

Finally, some of the created array entries are related to one another (see Figure 2). Hence, BASF2 relations are created: from the ECLCluster entries to the MCParticle and Track entries which are responsible for the creation of the cluster; similarly from the KLMCluster entries to the ECLCluster and Track entries; from the Track entries to the MCParticle entries that created it; and from the Track entries to the corresponding PIDLikelihood entries. Additional relations are created between the Particle entries in the created ParticleList objects and the corresponding MCParticle and PIDLikelihood entries. The links between TrackFitResult to Track; Track to Particle; and ECLCluster and KLMCluster to Particle are not represented by relations in BASF2.

3 Discussion

In order to ensure the correctness of the conversion, a study was performed with 200 000 recorded data and 600 000 simulated Belle events at the center of mass energy of the $\Upsilon(4\mathrm{S})$ resonance.

The events were processed with the old BASF framework and more than 360 quantities; for instance kinematic quantities like four-momenta, Monte Carlo information, PID information and beam-parameters; were extracted from the PANTHER tables shown in Figure 2. The complete list of extracted quantities can be found in [5, app. A]. Afterwards the events were processed a second time with the new BASF2 software using the b2bii conversion and the same quantities were extracted.

3.1 Observed Differences

Most quantities do not differ at all between the original Belle Software and the converted data using b2bii.

The observed differences between BASF and b2bii were further investigated and either corrected or ensured to be harmless. Minor differences occur due to small shifts caused by numerical imprecision leading to the migration of events between adjacent bins, especially for values near zero, and differences in the treatment of special floating point values such as infinity and NaN (Not a Number) leading to migration from the overflow/underflow bin to the bin including zero in rare cases (see Figure 3a).

Further differences are found: in the PDG codes of the MCParticle object due to the recovery of the full 32 bit as mentioned above; the number of daughters

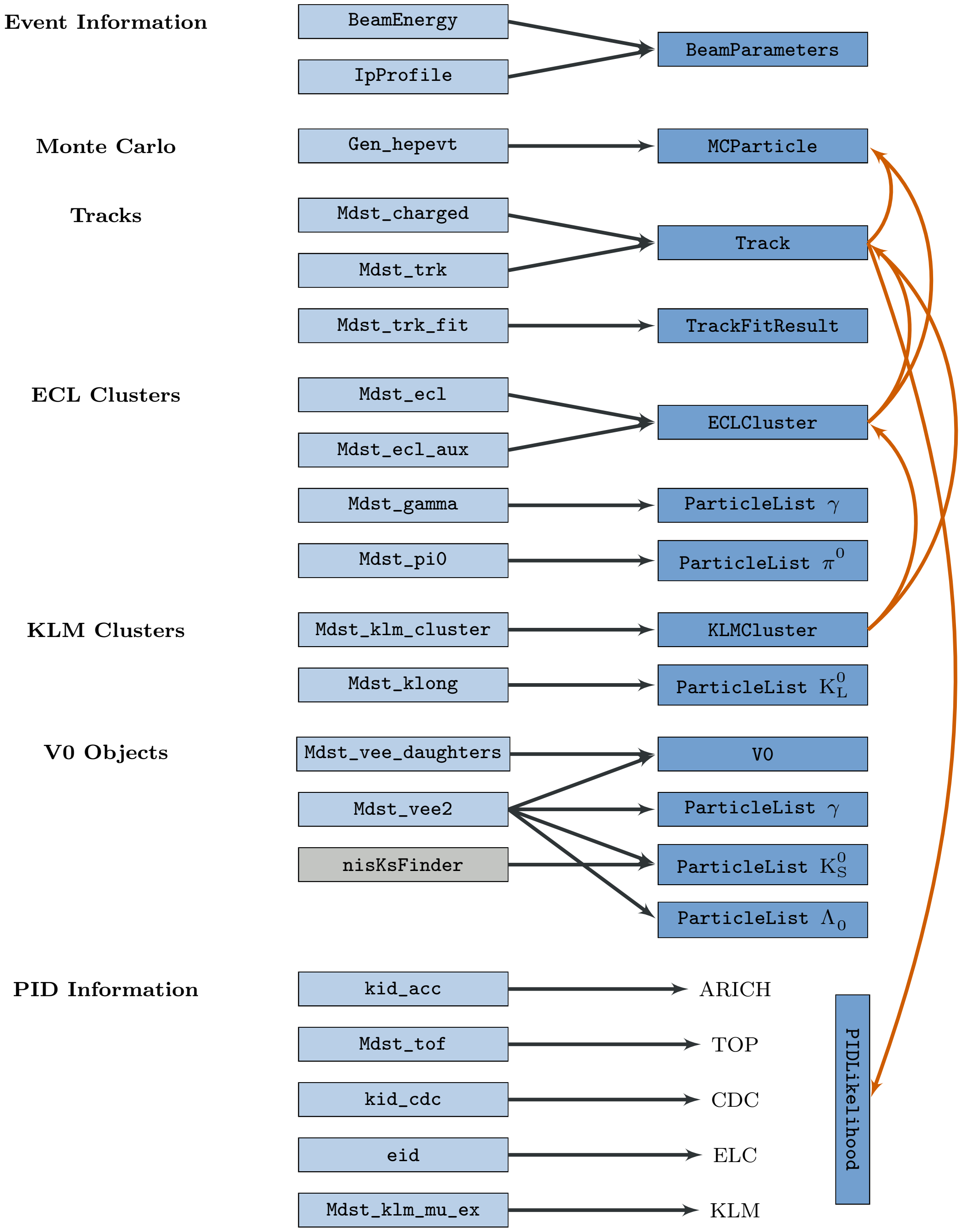
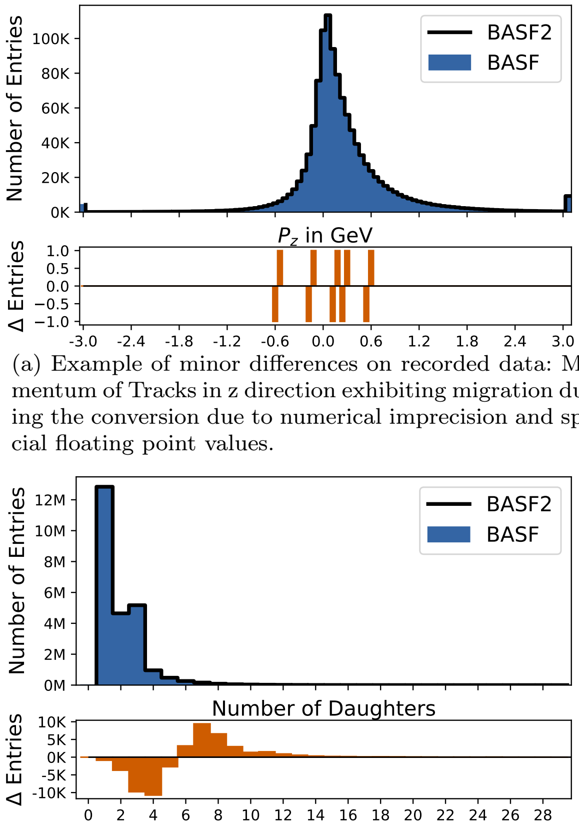


Fig. 2: Matching of the Belle PANTHER Tables (light blue) to the Belle II ROOT objects (blue) and relations (orange).

of the MCParticle object due to the unconverted virtual photons occurring in nuclear interactions between the hadronic final state particles and the detector material (see Figure 3b); and in all kinematic quantities of $V0$ and π^0 objects after the mass-constrained vertex fit caused by different software employed to fit the vertices.

4 Conclusion

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



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

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

Fig. 3: Comparison of `BASF` (Belle) and `b2bii` (Belle II). The leftmost (rightmost) bin represents the underflow (overflow) bin. The upper plots show the superimposed Belle (each component is shown individually) and Belle II Monitoring Histograms (the total number of entries is shown as a black line). The lower plots show the differences between Belle and Belle II, hence a positive (negative) difference means there are less (more) entries for the respective bin in the Belle II Monitoring Histogram.

In order to validate `BASF2` on a global level, physics analyses have been performed and compared to results published by the Belle collaboration [5, 19–23]. Other measurements using the `b2bii` conversion are in preparation.

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

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

References

1. K. Abe et al. Observation of mixing-induced CP violation in the neutral B meson system. *Phys. Rev. D*, 66, Aug 2002. doi: 10.1103/PhysRevD.66.032007.
2. Y. Amhis et al. Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2016. *Eur. Phys. J.*, C77(12), 2017. doi: 10.1140/epjc/s10052-017-5058-4. URL `https://arxiv.org/abs/1612.07233`.
3. S.-K. Choi et al. Observation of a Resonancelike Structure in the $\pi^{+-}\psi^{'}$ Mass Distribution in Exclusive $B \to K\pi^{+-}\psi^{'}$ Decays. *Phys. Rev. Lett.*, 100, Apr 2008. doi: 10.1103/PhysRevLett.100.142001.
4. A. Moll. The software framework of the Belle II experiment. *J. Phys. Conf. Ser.*, 331, 2011. doi: 10.1088/1742-6596/331/3/032024.
5. T. Keck. Machine learning algorithms for the Belle II experiment and their validation on Belle data, 2017.
6. A. Abashian, K. Gotow, N. Morgan, and L. Piilonen. The Belle Detector. *Nucl. Instrum. Meth.*, A479(1), 2002. doi: 10.1016/S0168-9002(01)02013-7.
7. T. Abe et al. Belle II Technical Design Report. 2010. URL `https://arxiv.org/abs/1011.0352`.
8. N. Katayama and R. Itoh et al. Belle computing model. *Computer Physics Communications*, 110(1), 1998. doi: 10.1016/S0010-4655(97)00148-3.
9. R. Itoh. BASF - BELLE AnalysiS Framework. In *Proceedings, 9th International Conference on Computing in High-Energy Physics (CHEP 1997)*, 1997. URL `http://www.ifh.de/CHEP97/paper/244.ps`.
10. D. J. Lange. The EvtGen particle decay simulation package. *Nucl. Instrum. Meth.*, A462, 2001. doi: 10.1016/S0168-9002(01)00089-4.
11. R. Brun et al. GEANT3. 1987.
12. M. Ritter, T. Kuhr, and M. Starič. High Level Interface to Conditions Data at Belle II. *J. Phys. Conf. Ser.*, 898(4), 2017. doi: 10.1088/1742-6596/898/4/042046.
13. L. Wood, T. Elsethagen, M. Schram, and E. Stephan. Conditions Database for the Belle II Experiment. *J. Phys. Conf. Ser.*, 898(4), 2017. doi: 10.1088/1742-6596/898/4/042060.
14. R. Brun and F. Rademakers. ROOT — An object oriented data analysis framework. *Nucl. Instrum. Meth.*, 389(1–2), 1997. doi: 10.1016/S0168-9002(97)00048-X.
15. C. Pulvermacher. Analysis Software and Full Event Interpretation for the Belle II Experiment, 2015.

16. K. Hara. Calibration of low momentum K_L^0 efficiency for veto usage in missing E analyses. Belle Note 1228 (internal).
17. E. Barberio, B. van Eijk, and Z. Was. Photos — a universal Monte Carlo for QED radiative corrections in decays. *Computer Physics Communications*, 66(1), 1991. doi: 10.1016/0010-4655(91) 90012-A.
18. A. Ryd et al. EvtGen – A Monte Carlo Generator for B - Physics. URL `http://evtgen.warwick.ac.uk/static/docs/EvtGenGuide.pdf`. User manual.
19. F. Metzner. Analysis of $B^+ \to \ell^+ \nu_\ell \gamma$ decays with the Belle II Analysis Software Framework. Master's thesis, Karlsruhe Institute of Technology, 2016. URL `https://ekp-invenio.physik.uni-karlsruhe.de/record/48845`.
20. S. Kohl. Dalitz analysis of $B^- \to D^+\pi^-\pi^-$. Master's thesis, Karlsruhe Institute of Technology, 2016. URL `https://ekp-invenio.physik.uni-karlsruhe.de/record/48803`.
21. F. Fichter. Search for $B_S \to \phi\pi^0$ Decays at the Belle Experiment. Master's thesis, Karlsruhe Institute of Technology, 2016. URL `https://ekp-invenio.physik.uni-karlsruhe.de/record/48880`.
22. J. Schwab. Calibration of the full event interpretation for the belle and the belle ii experiment. Master's thesis, Karlsruhe Institute of Technology, 2017. URL `https://ekp-invenio.physik.uni-karlsruhe.de/record/48931`.
23. A. Y. Espla. Validation of Belle II analysis framework searching for the decay $B \to \ell\gamma$ in Belle data. Master's thesis, Ludwig-Maximilians-Universität München.