

# **Belle II Open Data Proposal**

**Quinn Campagna**

# Belle II Open Data Policy

## Belle II Public Datasets Policy

Approved by the IB on March 3rd, 2023

- Our initial proposal will be to make some subset of the most recently deprecated Belle II MC campaign available for TREASURE
- Justification: participation in TREASURE
- Tokenizing data for use with LLMs (presumably for analysis)

### Specific datasets

A Belle II member or a group of Belle II members may request that a dataset of events measured with the Belle II detector or simulated with the Belle II software is made publicly available. The proponents must provide:

- A justification that discusses benefits and risks of the dataset publication
- A convincing plan that external users will be able to perform analysis with the data
- A documentation of the format and content of the dataset with sufficient information that non Belle II members can use it with a low risk of making mistakes
- At least one example analysis that illustrates the use of the data
- A credible assurance that the required computing resources are provided for a reasonable period of time

The physics coordinator discusses the proposal with the relevant group conveners and decides together with them. Public datasets must be citable, for example by adding them to zenodo.

### General rules

All public datasets must be listed on a dedicated Belle II page and require non-Belle II users to include the following acknowledgment and statement in presentations and publications of results obtained with it: “We acknowledge and thank Belle II for enabling this research using public Belle II data. The responsibility for the result is solely with the authors.”

While the usability of datasets by non-Belle II members is an important criteria in the decision of making them public, datasets are provided as-is with no commitment of any support from Belle II.

Belle II members who want to publish scientific results obtained with public Belle II data must follow the “Belle II paper and authorship types” policy.

# Open Questions

- What is the data format we are “making available” to TREASURE?
  - Is Belle II providing ntuples to TREASURE, or is the HDF5 conversion happening on the Belle II side?
- Does TREASURE use a service to host open data (eg. Zenodo)?
- What amount of data is recommended to properly train the tokenizer (based on ATLAS experience)?