Karlsruhe Institute of Technology

Bundesministerium für Bildung und Forschung

ErUM-Data IDT

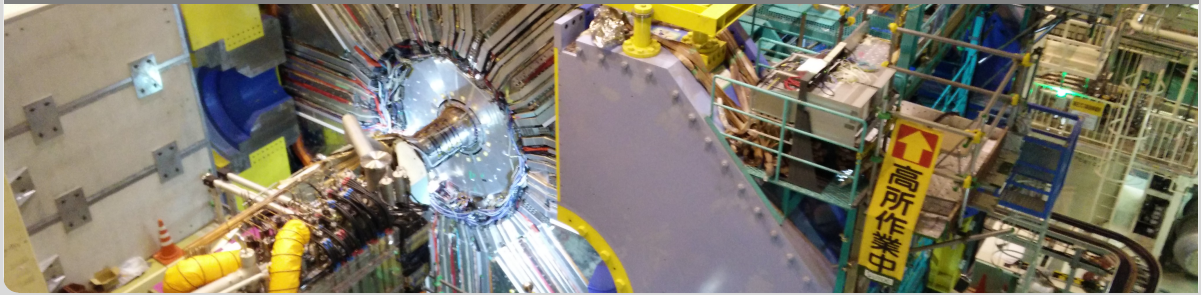
Belle II

Generator level selection dataset

Whole graph classification

James Kahn (KIT), Yannick Bross (LMU) | 8th April 2020

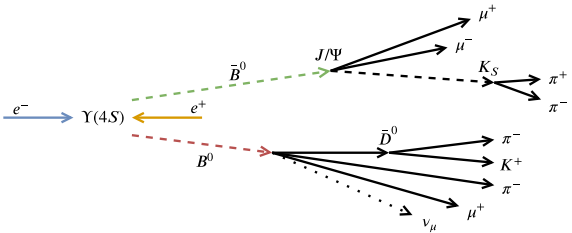
INSTITUT FÜR EXPERIMENTELLE TEILCHENPHYSIK (ETP)



Overview

Selective background Monte Carlo simulation project (CHEP 2019, ErUM collab. meeting)

- Categories:
  - Whole graph classification
  - Bias mitigation/quantification
- Use generator level info to predict usefulness of events
- Usefulness definition depends on physics WG



Dataset goal

Maximise simulation speedup while keeping biases to an acceptable level.

Dataset (in preparation)

Project used 300k events (250/50 split) from three WGs:

- Hadronic $B^{0/+}$ meson reconstruction ($\sim$ 5%)
- Time-dependent CP violation ($\sim$ 0.2%)

Model input

- Fixed number of properties per particle
- Adjacency matrix describing relations
- Particle ID (PDG code)
- Production momentum/position
- Mass/energy/charge

Labels

One binary label per graph

Bias observables

- Event-level kinematics (whole graph vars)
- Physics motivated – specialised for WGs
- Derived from reconstructed particles – can't be used as network input

Performance metrics

Relative simulation speedup

- Interested in how much faster we can simulate one event.
- Calculation is simple[1] but a function of TPR, FPR, and source dataset.
  - Non-differentiable
- Will vary for each WG

Bias quantification

- Yannick Bross (LMU) handling this.
- High-level observables describing whole graph
- KL divergence between all true and true positives
- Can (should) be used during training

Performance judgement: Highest speedup with $\mathrm{KL}_{\mathrm{div}} < X$ across all bias observables.

[1] Assuming trivial inference time

Questions

1. Is there a template we can use? Kaggle?
2. Graph data format?
   - Feature and adjacency matrices (what we used so far)?
   - Feature and edge list (PyTorch Geometric)?
   - ...
3. Do we provide code examples for loading data? Also for calculating metrics (differs for each dataset)?
4. Features normalised?
5. Dataset size limit ($\mathcal{O}(\text{GB})$)?