Analysis Specific Filters for Selective Background Monte Carlo Simulations at Belle II

Masters's project - AG Kuhr

Luca Schinnerl

February 2, 2022

Overview

- Introduction and motivation.
- Training data production.
- Technical backgrounds.
- Analysis specific filters.
- Conclusion.

Motivation

Motivation - The Belle II experiment

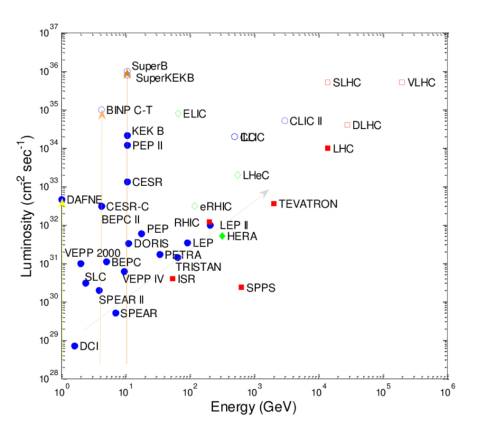


Figure: Luminosity versus energy of colliders.

The goal of Belle II is to obtain extremely high integrated luminosity:

- Fifty times as much data as the Belle experiment.
- For rare processes, a strong statistical knowledge of the background is required to distinguish the signal from background.

Motivation - Monte Carlo Based Experiment

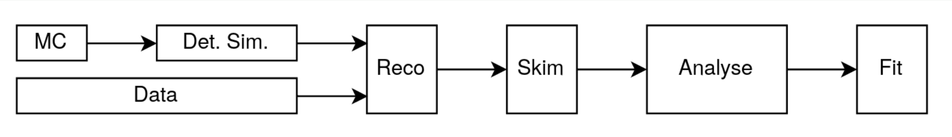


Figure: Usual data flow within a Monte Carlo based experiment. Monte Carlo simulation refers to the output of the first three stages.

The data flow at the Belle II experiment is comprised of two parts:

- Theory simulation which is comprised of event generation and detector simulation.
- Real life running of the experiment.

Motivation - Description of the Problem

Problem

Simulation of the required amount of data is infeasible!

Solution

Add a filter to discared irrelevant events before the detector simulation.

Motivation - Solution

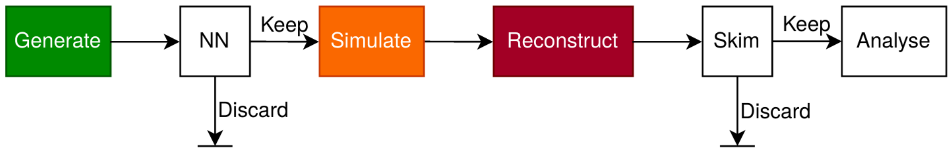


Figure: Selective "smart" background MC simulation.

- Inject a NN between event gerneration and detector simulation which decides to discard or keep an event.
- Event generation takes a much smaller fraction of computing time than detector simulation.
- Most events are discarded by skimming.

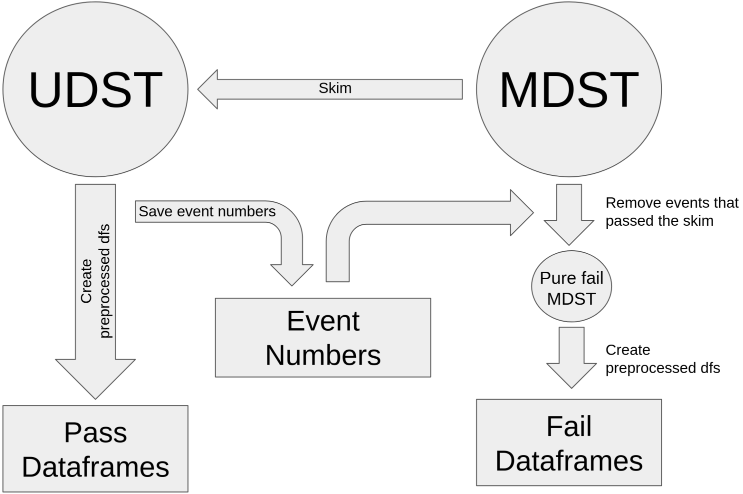
Motivation

Goal of the Master's Project.

Introduce analysis specific filters to increase the achieved speedup with filtering.

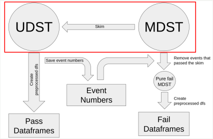
Training Data Production

Data Production - Preprocessed Dataframes



LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Data Production - Step 1: Centrally Commissioned Skim



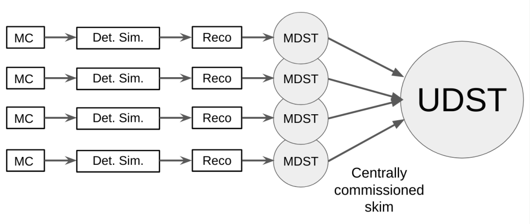
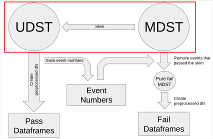


Figure: Schematic illustration of the central skimming at the Belle II experiment.

- Practical to commission a loose central skim.
- Generally: UDSTs labeled are pass, left over MDSTs are fail events.
- Around 5% of events usually survive this process depending which centrally commissioned skim is considered.

Data Production - Step 1: Centrally Commissioned Skim



FEI Hadronic B^0 skim

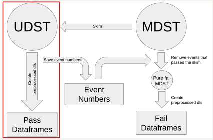
- **nTracks** < 13
- $M_{bc} > 5.24 GeV/c^2$
- $|\Delta E| < 0.2 GeV$
- **sigProb** > 0.001
- **nParticles(B0:feiHadronic)** > 0

retention rate $= 6.21\%$

Data used in this study:

- Mixed ($\Upsilon(4S) \to B^0\bar{B}^0$) samples from MC14.
- Generally: 900k training, 100k validation and 500k testing.
- 50:50 mix between pass and fail data for training.

Data Production - Step 2: Create Pass-Preprocessed dfs



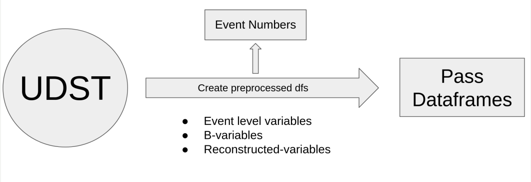


Figure: Schematic illustration of the creation of preprocessed pass dataframes.

- MC particle record (actual training data).
- Event level variables (event shape and kinematics).
- B-variables from FEI (M_{bc}, ΔE and signal probability).
- Variables from a reconstruction, here $\Upsilon(4S)$ with a $B \to K^{(*)}\nu\bar{\nu}$ decay (number of particles, rest of event).
- Label $=$ True.

Data Produciton - Step 3: Create Fail-Preprocessed dfs

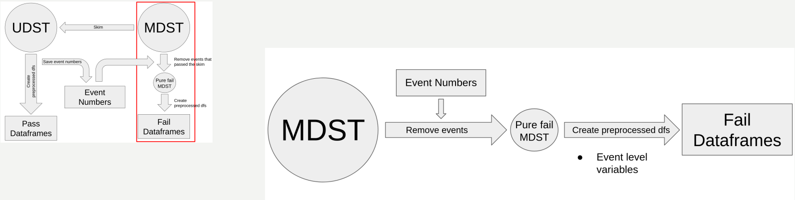
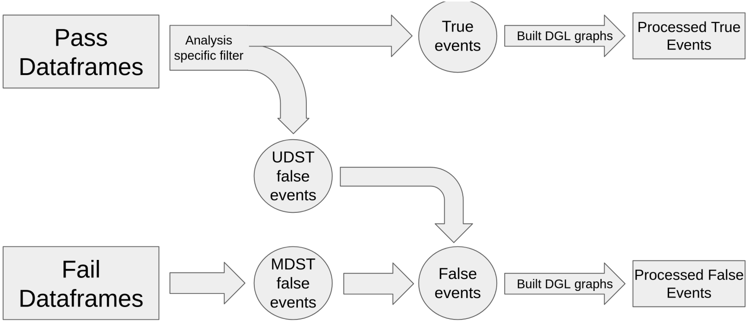


Figure: Schematic illustration of the creation of preprocessed fail dataframes.

- Remove events that previously appeared when creating pass dataframes.
- Only include MC particle record and event level variables (event shape and kinematics).
- Label = False.

Data Production - Analysis Specific Filters



Data Production - Analysis Specific Filters

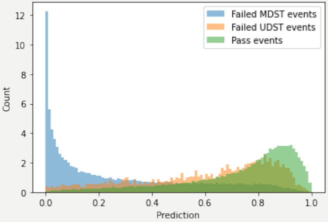


Figure: Example prediction with separated false events.

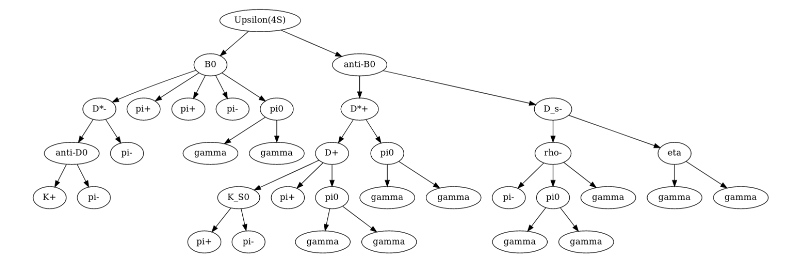
Important to preserve the fractions of UDST false events to MDST false events, as UDST false events are harder to distinguish from true events.

- Cuts lead to two types of false events: Events that previously passed the skim (UDST false events)and events that have not (MDST false events).
- UDST false events are a lot more similar to true events than MDST false events, which may reduce the performance.
- UDST false rate:

$$\frac{\#\text{UDST false eventes}}{\#\text{UDST false eventes} + \#\text{MDST false eventes}}$$

Technical background

Tec. Background - Graph Structure Data



Node attribute: PDG ID, 4-vector components, Vertex positions, Decay times.

- Data is graph structured with:
  - Nodes = Particles
  - Node attributes = Particle properties
  - Edges = Parent-daugther relations (decays)
  - Graph type = Tree
- Usage of graph neural networks very useful.

Tec. Background - Graph Attention Networks

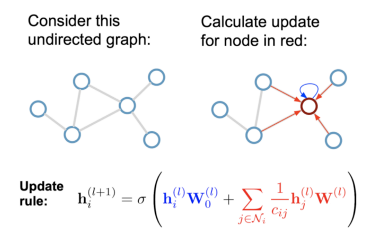


Figure: Graph Convolutional Network.

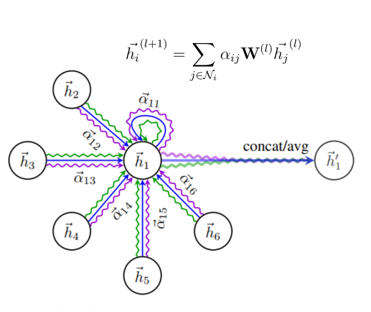
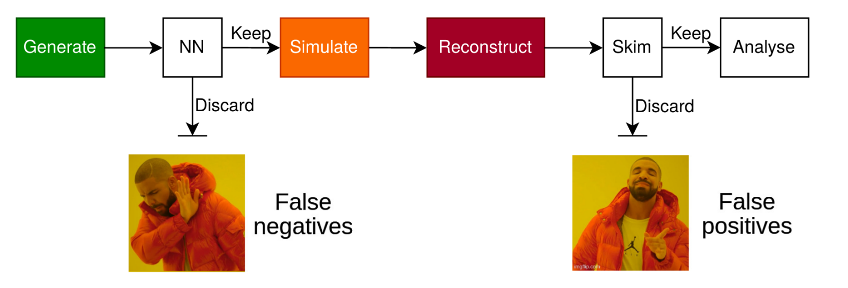


Figure: Graph Attention Network.

- Main limitation of plain GCN: equal contribution from each neighbour during aggregation.
- A GAT is able to learn weights for neighbour aggregation from features of adjacent nodes.
- Global attention pooling: learn weights for aggregation into global features from node features.

Tec. Background - The Problem with Naive Filtering



- **False positives** are not too problematic, these are thrown away later by running the skim.
- **False negatives** can introduce bias (we can't get them back).

Problem

Naive smart background simulation may produce bias!

Tec. Background - The Sampling Method

Bias mitigation

Sample events with the NN output.

- Use NN output as **probability to keep event**.
- Weights events by inverse probability like importance sampling with $w = \frac{1}{p_{NNfilter}}$.
- No bias by construction.
- Use speedup as loss function for training.

Metric to optimise (loss function) is the **Speedup**:
How much faster can the same number of events (in terms of effective sample size) be produced compared to brute force simulation.

Effective sample size: $\frac{(\sum_i w_i)^2}{\sum_i w_i^2}$

Tec. Background - The Speedup Metric

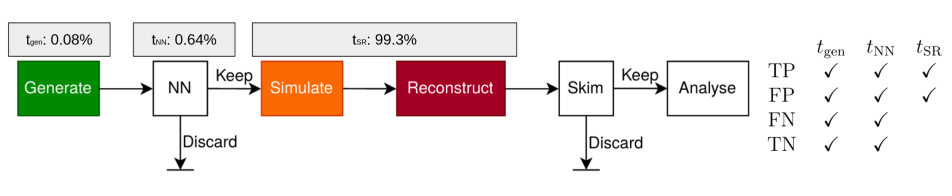


Figure: Processing times of each step in the MC simulation.

The ratio between time consumption of the whole workflow with and without the NN filters for producing **the same effective sample size**:

$$R_0 = \frac{t_{\text{no\_filter}}}{t_{\text{filter}}}$$

Analysis Specific Filters for $B \to K^{(*)} \nu\bar{\nu}$

Analysis Specific Filter - No Additional Cuts

When no analysis specific filters are specified and only the commissioned FEI hadronic skim is considered, the speedup is limited.

Cut

- **nTracks** < 13
- $M_{bc} > 5.24GeV/c^2$
- $|\Delta E| < 0.2GeV$
- **sigProb** > 0.001
- **nParticles(B0:feiHadronic)** > 0

retention rate $= 6.21\%$

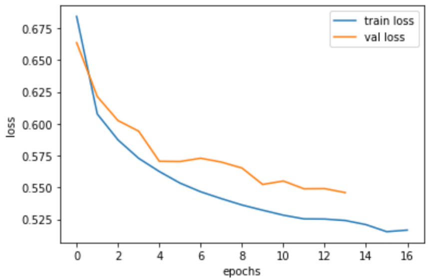


Figure: Loss per epoch, with speedup as loss function.

Analysis Specific Filter - No Additional Cuts

Generally speaking, bias of filtered data is as expected within statistical uncertainty:

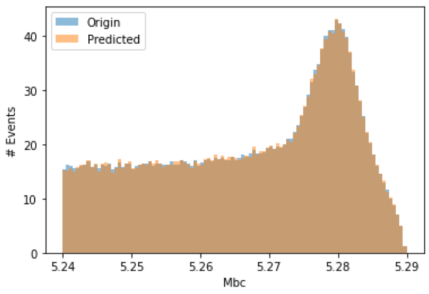


Figure: Predicted and original M_{bc}.

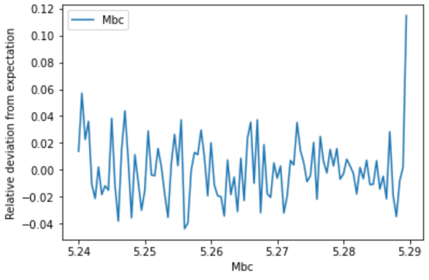


Figure: Difference between the original and predicted M_{bc} using the sampling method.

1.84

Analysis Specific Filter - Cuts on B-Variables from FEI

Increasing the cut on M_{bc} and **sigProb** is motivated by increasing the efficiencies on the reconstructed B_{tag}.

Cut

- **nTracks** < 13
- $M_{bc} > 5.265GeV/c^2$
- $|\Delta E| < 0.2GeV$
- **sigProb** > 0.005
- **nParticles(B0:feiHadronic)** > 0

retention rate $= 2.28\%$

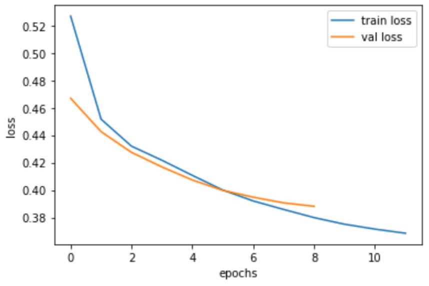


Figure: Loss per epoch, with speedup as loss function.

2.55

Analysis Specific Filter - Require a $\Upsilon(4S)$

Next step is to require that a $\Upsilon(4S)$ with a $B \to K^{(*)}\nu\bar{\nu}$ decay could be reconstructed.

Cut

- **nTracks** < 13
- $M_{bc} > 5.265GeV/c^2$
- $|\Delta E| < 0.2GeV$
- **sigProb** > 0.005
- **nParticles(B0:feiHadronic)** > 0
- **nParticles(Upsilon(4S):reconstructed)** > 0

retention rate $= 0.45\%$

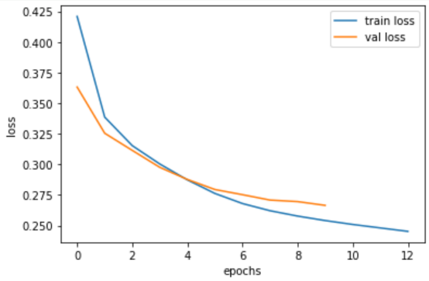
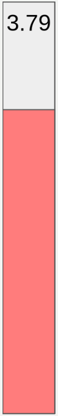


Figure: Loss per epoch, with speedup as loss function.



Analysis Specific Filter - No Tracks in the ROE

This cut requires that the rest of event has no charge tracks. This ensures that all particles from the primary physics event were used in the $\Upsilon(4S)$ reconstruction.

Cut

- **nTracks** < 13
- $M_{bc} > 5.265 GeV/c^2$
- $|\Delta E| < 0.2 GeV$
- **sigProb** > 0.005
- **nParticlesInList(Upsilon(4S):reconstructed)** > 0
- **ROE(Tracks)** $= 0$

retention rate $= 0.30\%$

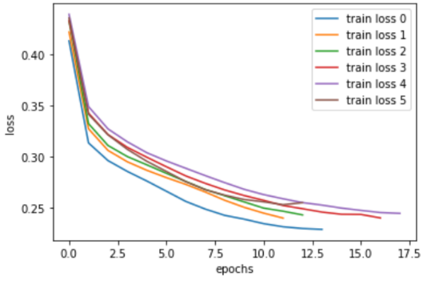


Figure: Loss per epoch with different number of tracks in ROE allowed. Loss function: Speedup.

4.21

Conclusion and Outlook

Conclusion:

- A flexible and fast training data production is essential for analysis specific filters
- Smart background simulation has the potential to significantly improve simulation speeds.
- Introducing analysis specific filters can refine this process.
- For the process $B \to K^{(*)}\nu\bar{\nu}$, the rate of the speedup was raised from less to than 2, more than 4.

Outlook:

- Upgrade the training data production to handle other MC samples to analyse these, such as the quark continuum.
- Train the network with event weights and an uneven pass/fail rate, so that no events have to be discarded for training.

Thank you

for your attention