Simulating Noise Waveforms in the Belle II Calorimeter Using GANs

Alexandre Beaubien, PhD student — alexandrebeaubien@uvic.ca

University of Victoria

The Belle II collaboration

2023-06-19

University of Victoria

Belle II

The Belle II International Collaboration

~1200 collaborators, ~600 authors

- ~500 students, ~450 “Physicists”, ~230 technical staff

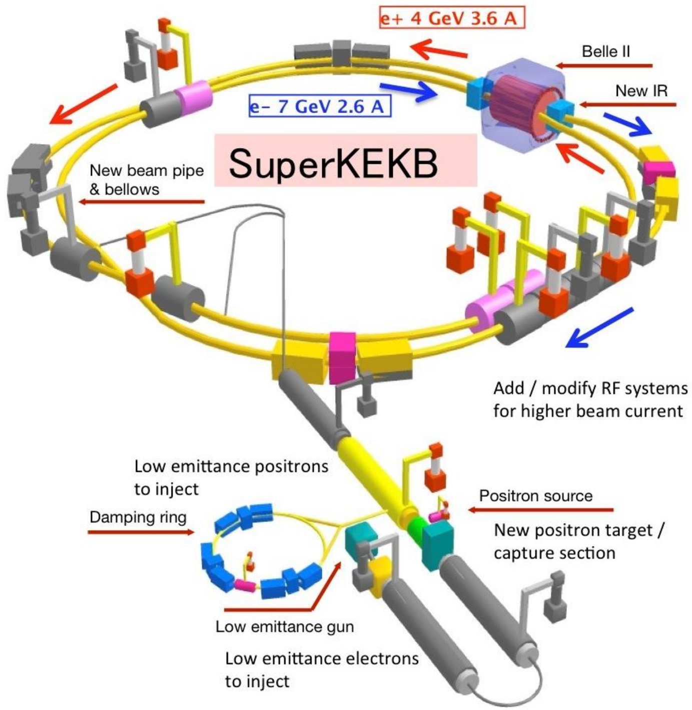
123 Institutions
27 Countries

Collider Experiment at SuperKEKB

See Savino Longo's talk Wednesday 10:30

SuperKEKB is an e^+e^- collider operating at $\sqrt{s} = 10.58$ GeV (4 & 7 GeV beams respectively).

As of 2020, it is the highest luminosity collider *ever*!



Luminosity ∝ intensity of beam.

$$N = \sigma L$$

N = Number of collisions
σ = Cross section
L = Luminosity

More luminosity
= more events
= more physics*

The Belle II Detector

General-purpose detector — Built like an onion around interaction point (IP)

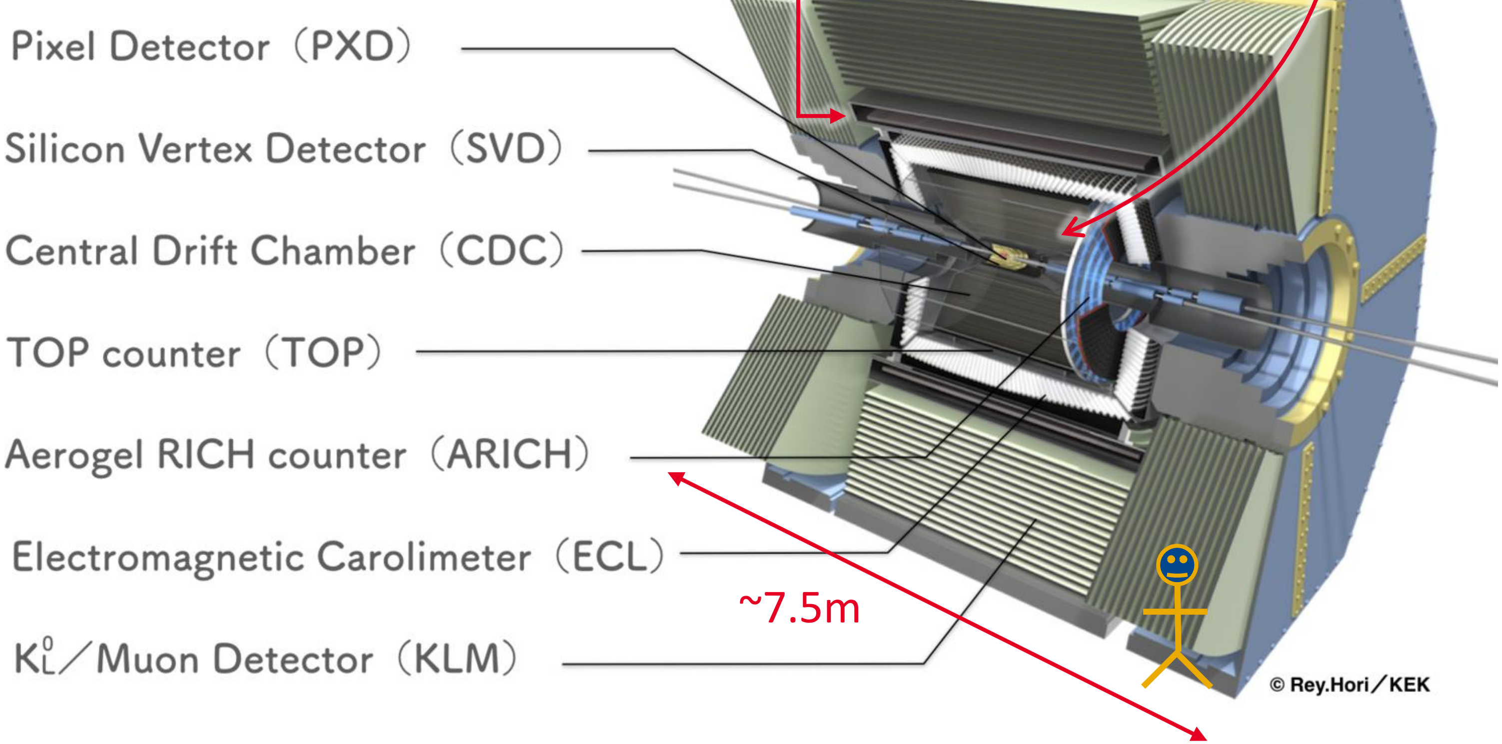
7 sub-detectors

Also, a 1.5T magnet!

Particle Position & Tracks

Particle Type

Particle Energy



~7m

The Belle II Detector

General-purpose detector — Built like an onion around interaction point (IP)

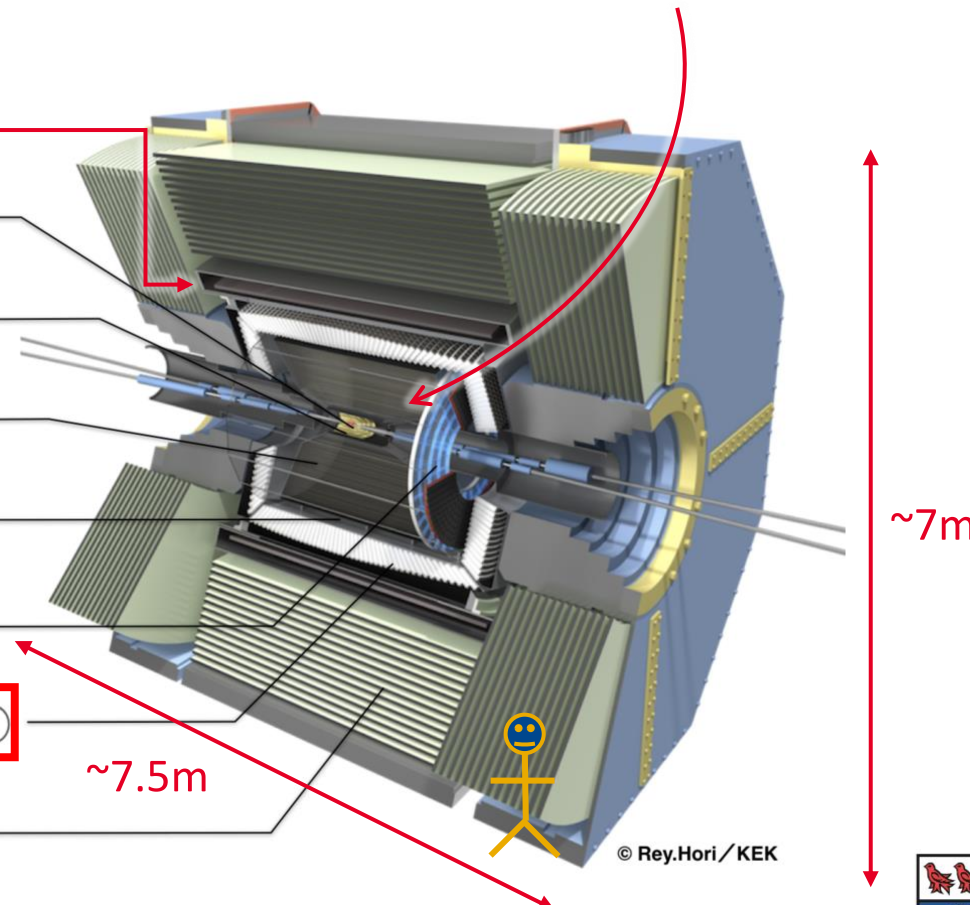
7 sub-detectors

Also, a 1.5T magnet!

Particle Position & Tracks
- Pixel Detector (PXD)
- Silicon Vertex Detector (SVD)
- Central Drift Chamber (CDC)

Particle Type
- TOP counter (TOP)
- Aerogel RICH counter (ARICH)

Particle Energy
- Electromagnetic Carolimeter (ECL)
- K_L^0／Muon Detector (KLM)

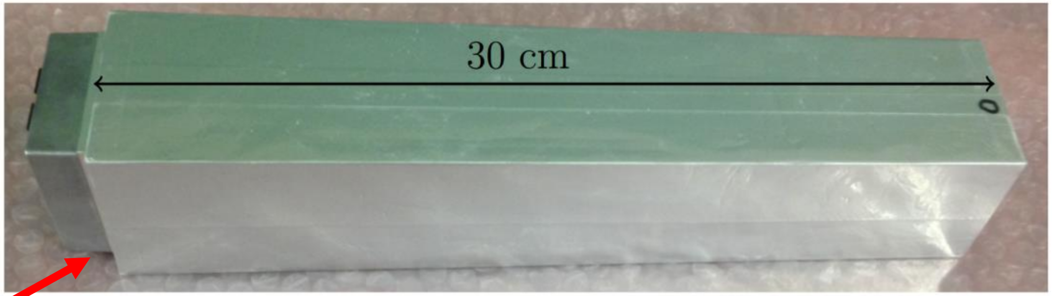


This project is interested in the ECL waveforms

Electromagnetic Calorimeter (ECL)

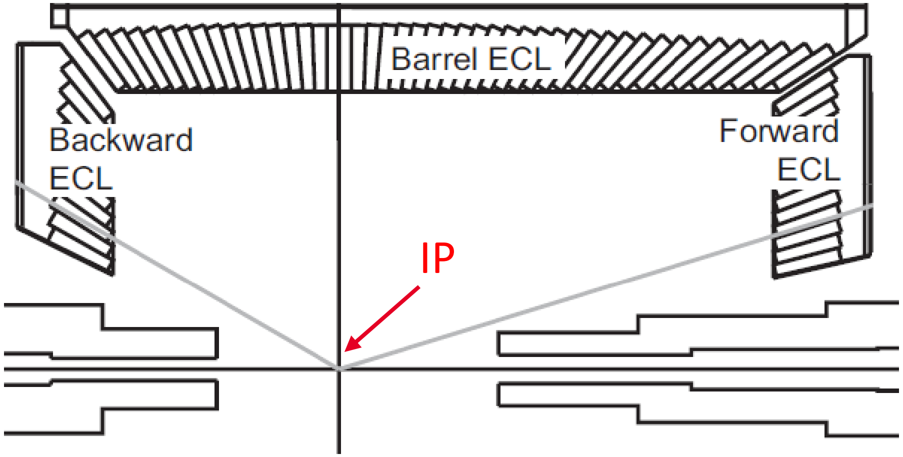
ECL is responsible for:

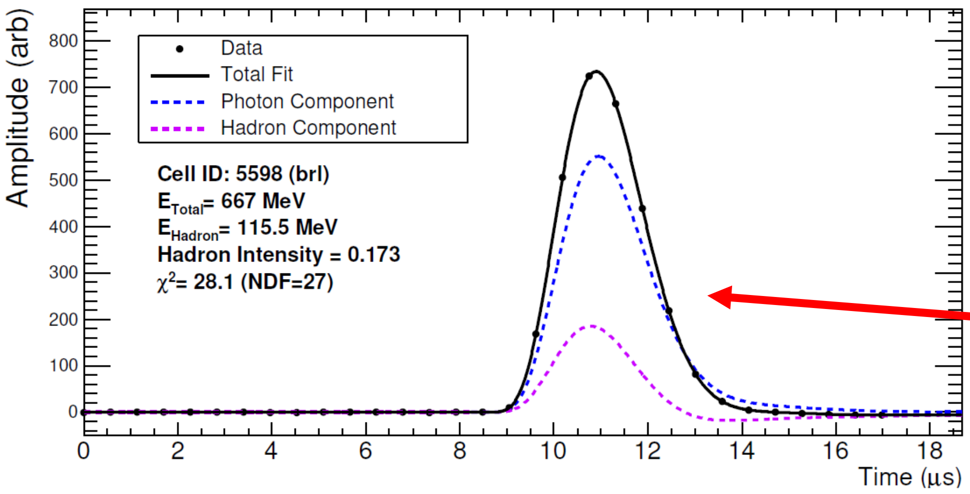
- **Measuring particle's energies** through energy depositions.



Composed of **8736 CsI(Tl) crystals** arranged in a cylinder around the IP.

Upper profile of the ECL





Crystal PMT measurements are **digitized** in **31-length waveforms** and **fit**.

Motivation for GAN Simulation

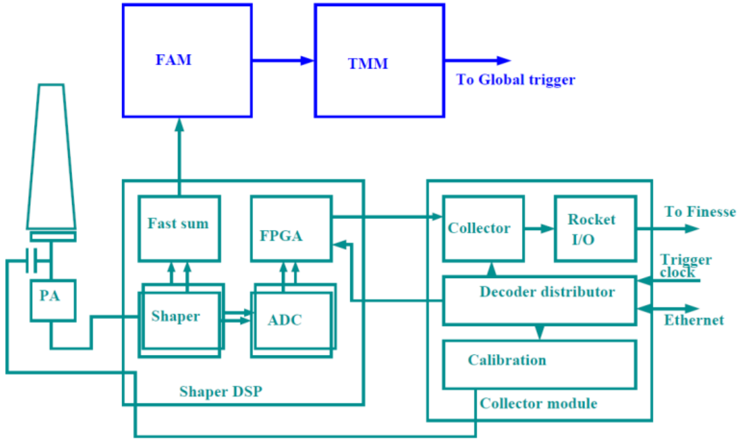
With great luminosity comes great responsibilities (background)

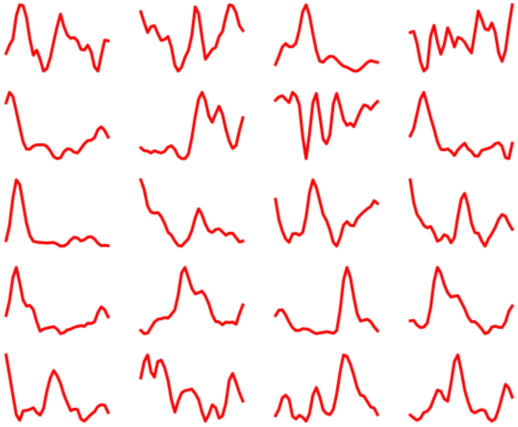
↳ How to accurately represent background during simulation?

Take random snapshots of the detector (triggers)
→ overlay them onto simulation (BGOverlay)!

Problem: requires too much data & bandwidth

ECL crystal electronics



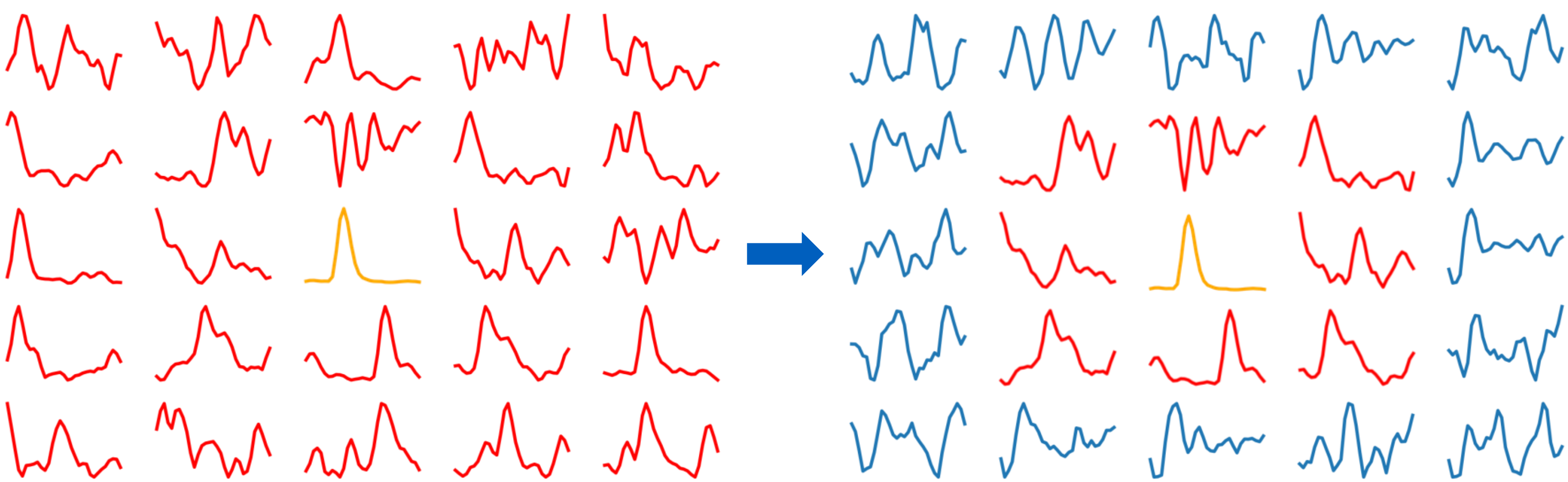


General Simulation Strategy

Only keep interesting waveforms and simulate the rest

→ “high energy”, correlated background

e.g. a high energy waveform and its neighbours are saved



Data (real) noise waveforms

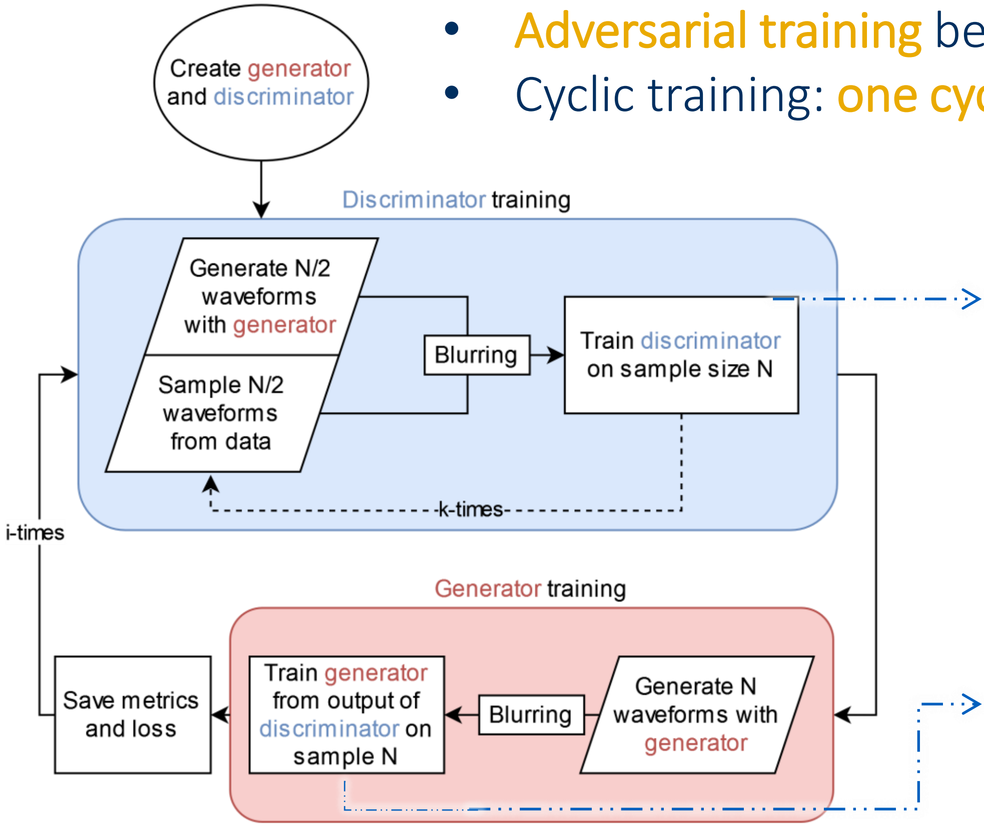
(partly) Simulated noise waveforms

Same technology as Midjourney, DALL-E*

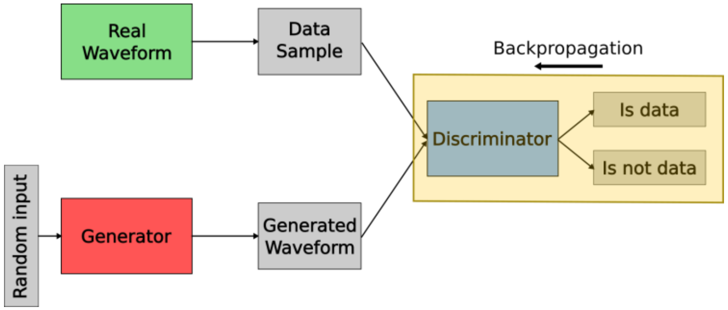
Simulate Using a GAN

GANs are Generative Adversarial trained neural Networks.

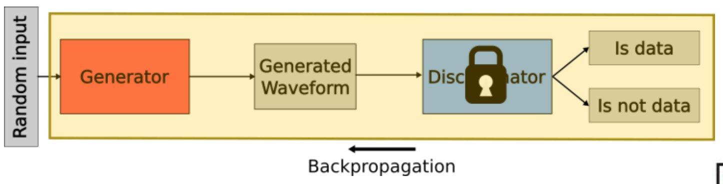
- Generator generates ECL waveforms
- Discriminator distinguishes real from generated waveforms
- Adversarial training between the networks
- Cyclic training: one cycle is called an epoch



Training Framework



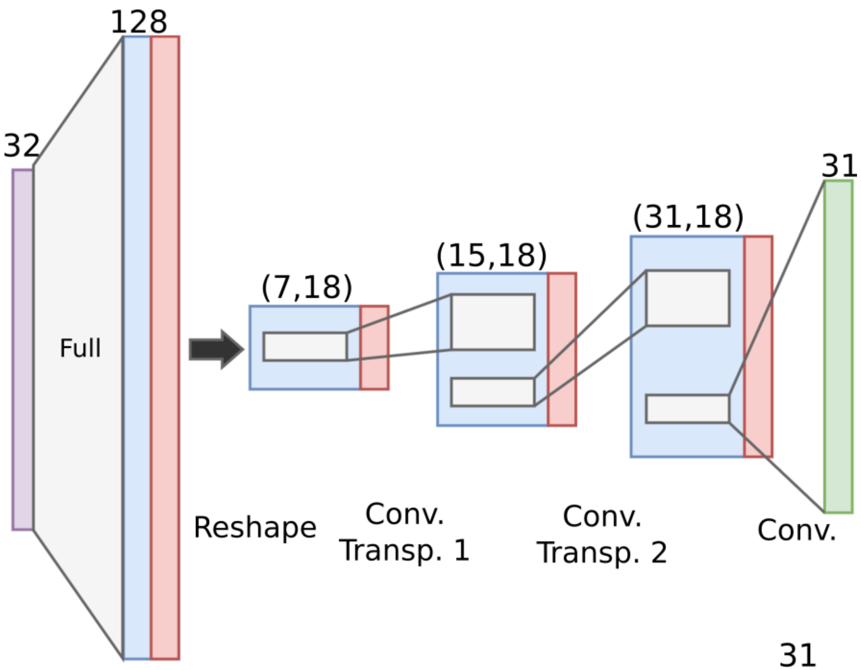
Training the discriminator



Training the generator

The Neural Networks

Purple: input vectors
Blue: neuron layers
Red: leaky ReLU activations
Yellow: dropout layers
Green: output layers

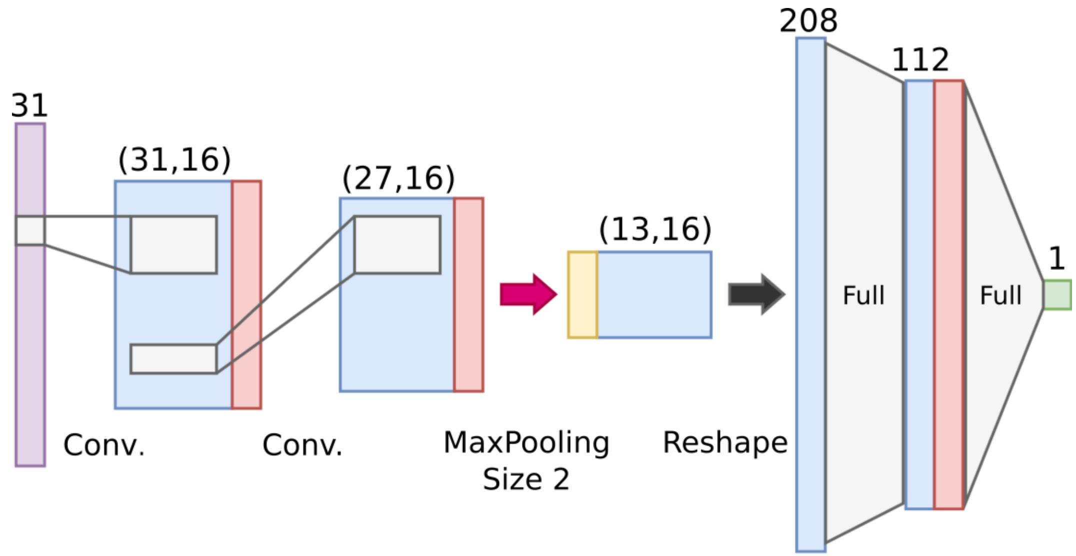


<u>Generator</u>:
1D deep **convolutional neural network** (D-CNN) used to generate waveforms

Random Input → Waveforms

<u>Discriminator:</u>
Another **1D D-CNN** to evaluate (data or generated waveform?)

Waveforms → Boolean



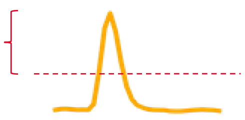
Measuring the Quality of Waveforms

Characterize waveforms using **4 metrics**

For waveform: $\mathbf{S} = S_i = s_1 \ldots s_{31}$

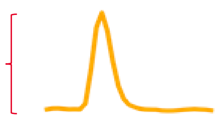
1. ΔA_{avg} (Average Amplitude Difference):

$$\frac{1}{31}\sum_{i=1}^{31}|S_i - \bar{S}|$$



2. ΔA_{max} (Maximum Amplitude Difference):

$$\max(\mathbf{S}) - \min(\mathbf{S})$$



3. f (Power Spectrum, Discrete Fourier Analysis)

4. χ^2 (from straight line):

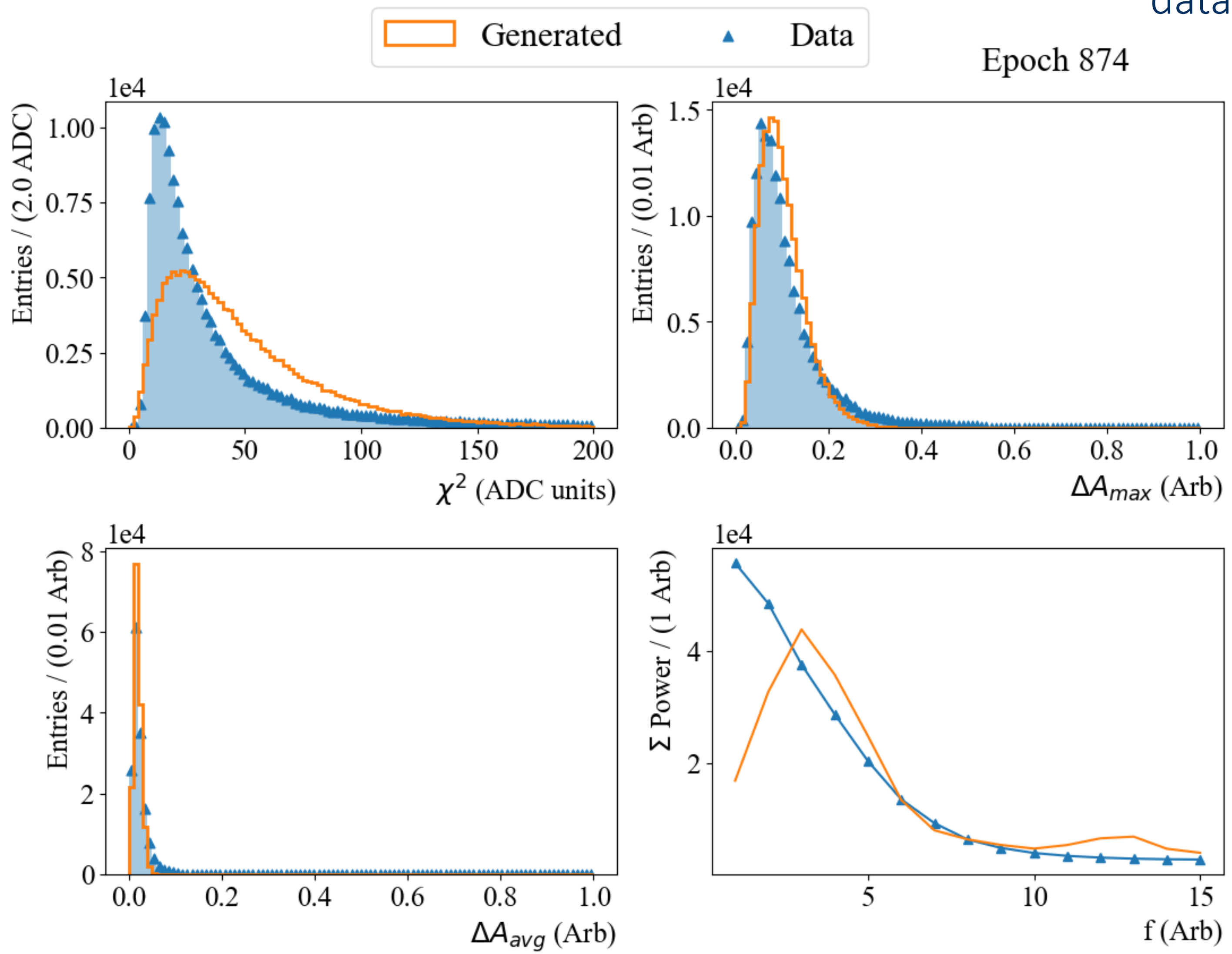
$$\chi^2 = \mathbf{D}\mathbf{C}^{-1}\mathbf{D}$$

$$\mathbf{D} = \mathbf{S} - \mathbf{G}$$

where $\mathbf{C}$ is the autocovariance matrix of $\mathbf{S}$, and $\mathbf{G}$ is a straight line.

GAN Performance

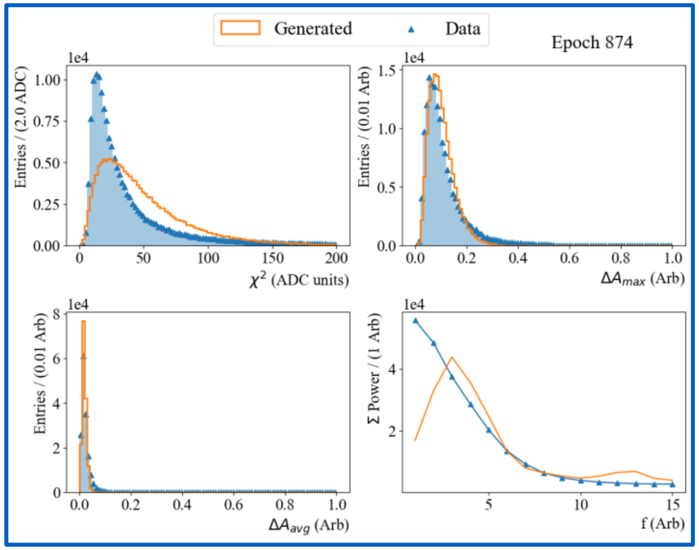
After **every epoch**:
1. Generate waveforms
2. **Calculate** their **metrics**
3. **Compare** metrics of generated & data waveforms

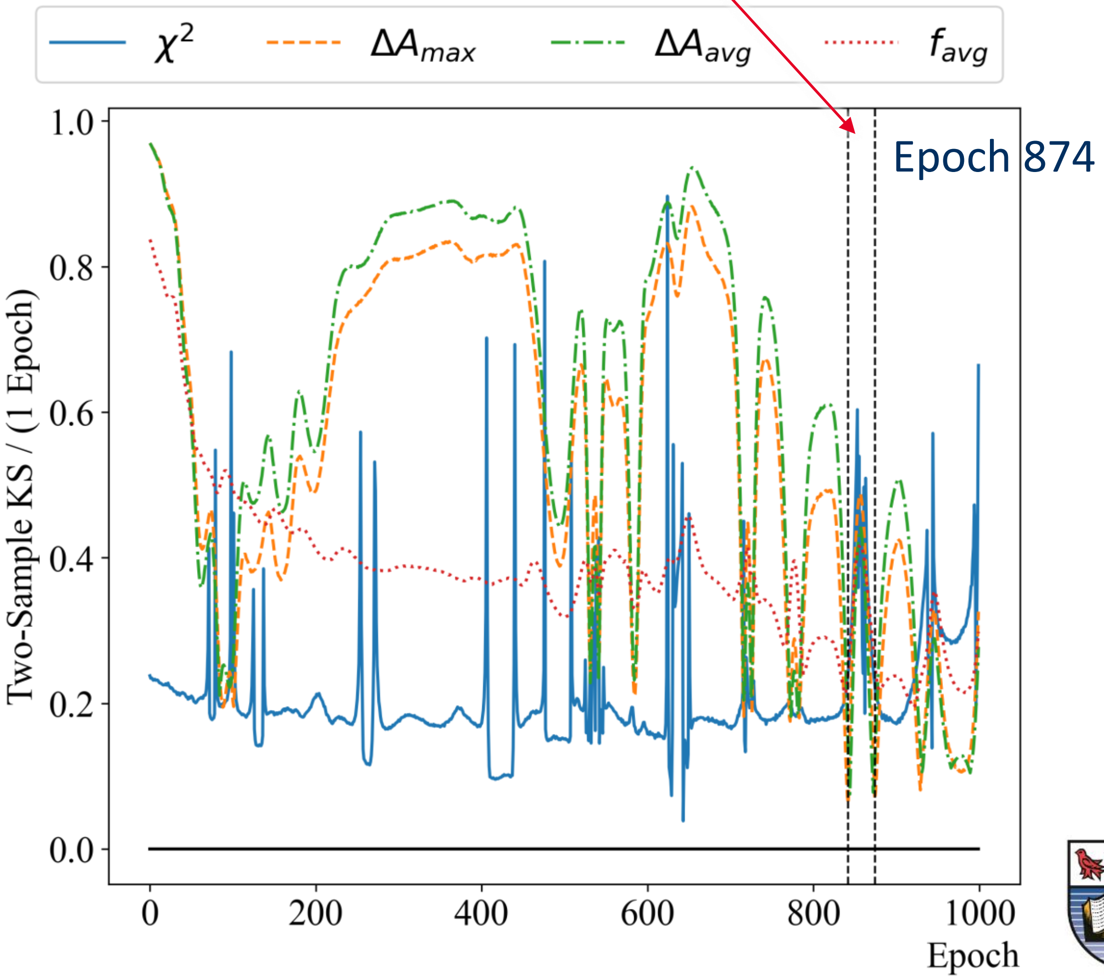


Quantifying GAN Performance

Quantify using a two-sample Kolmogorov-Smirnov (KS2) test. The best model is currently selected as the average lowest KS2 result.

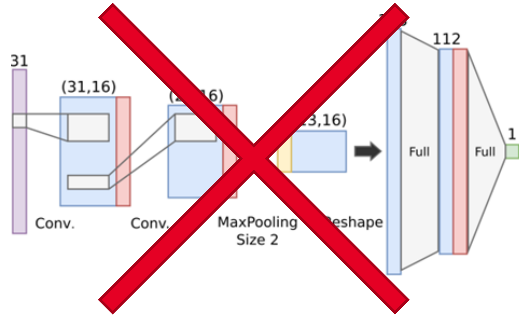
This is the epoch when the generated waveforms are most similar to data.



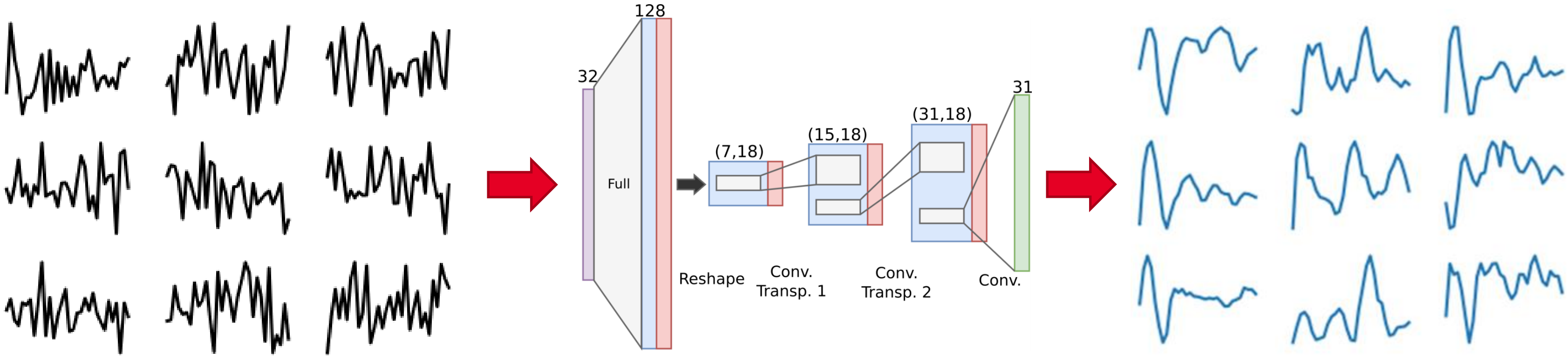


Congratulations, you now have a trained and ready to use GAN!

Discard Discriminator

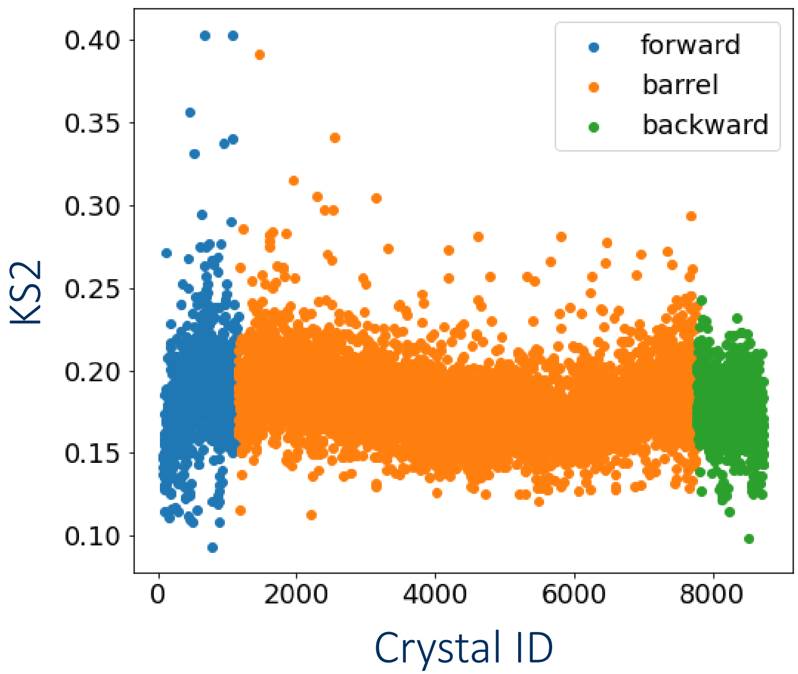


Use Generator at will

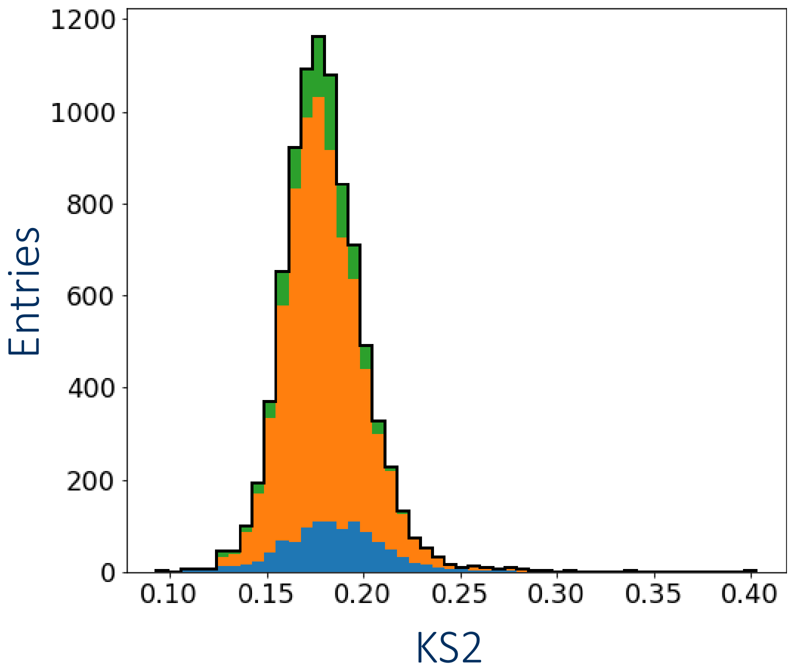


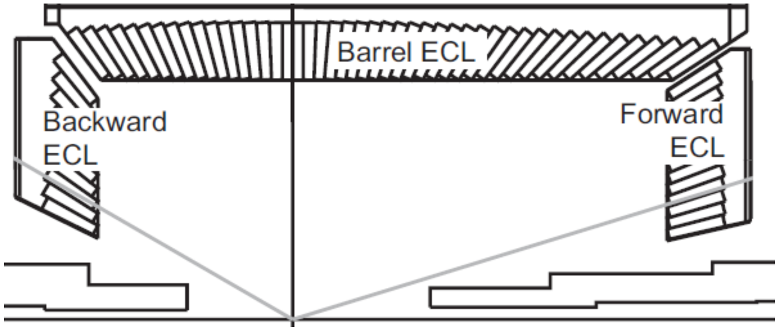
Training the Entire ECL

Best KS2 for each crystal



Best KS2 for each crystal





Average best KS2

- GAN: 0.18
- Autoencoders: 0.21*
- Cov. Matrix Methods: >0.3**

*Matt Forbes, **Alexei Sibidanov – UVic

You Can (Soon) See this Code

The Belle II analysis framework (basf2) is an open-source software. Hundreds of collaborators have contributed to it in one capacity or another:

https://github.com/belle2/basf2

Note: the public code is a few months behind the current main branch for development purposes.

It includes many advanced computing techniques

e.g. A.I., machine learning, etc.

Used for Physics tools, computing tools, …

You can find tools like this GAN on there!

Summary

- New and world-class machine learning frameworks are being implemented into the Belle II analysis framework.
- A GAN is developed to generate ECL waveforms, which are used during simulation to reproduce background.
- A robust and extensive system of metrics is used to grade the model performances.

Validation and testing is ongoing.

Contact:

Alexandre Beaubien — alexandrebeaubien@uvic.ca

Extra

“off the shelf” Hyper-Parameters

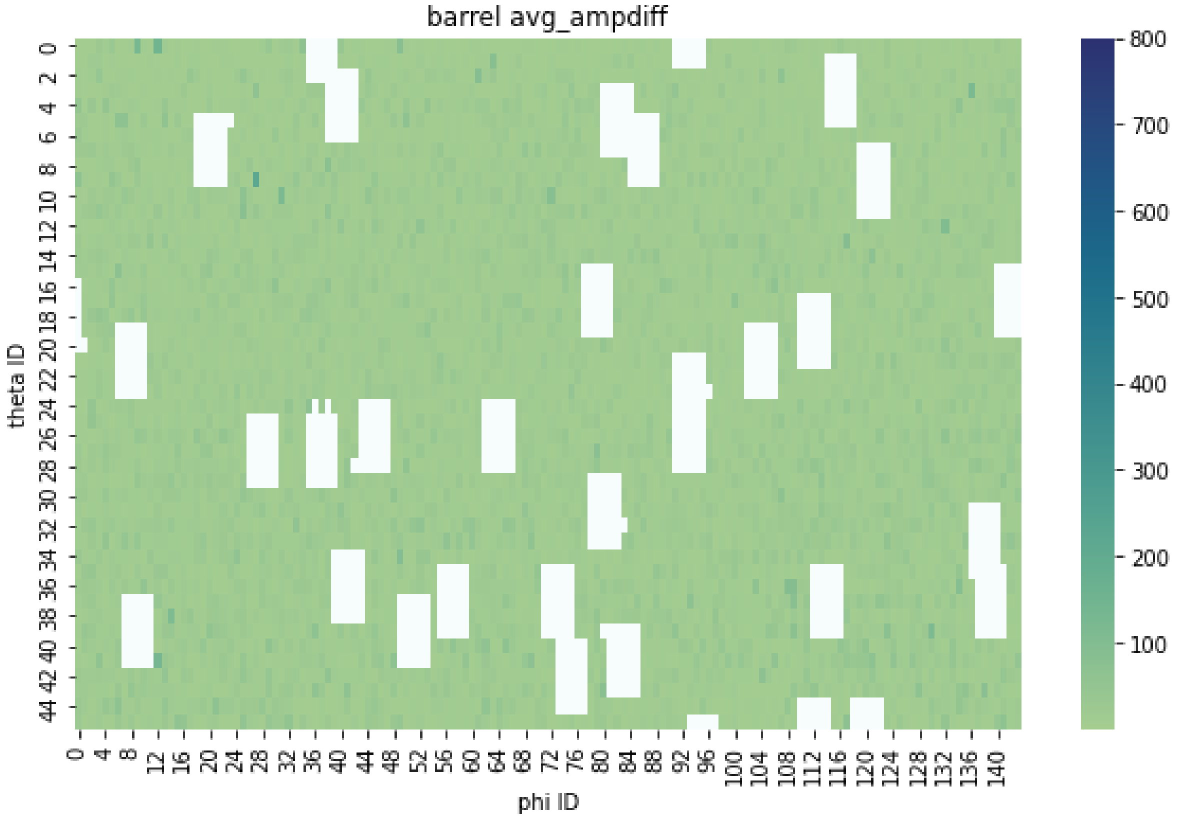
- **1000** training loops (**epochs**):
- Discriminator to Generator **training ratio** = 3
  - Train discriminator 3 times (three batches)
  - Train generator 1 time (one batch)
- Use **batch sizes** of **128**
  - Discriminator batch: 64 generated waveforms, 64 real waveforms
  - Generator batch: 128 generated waveforms
- Use **noisy and smooth of target labels**
  - Target label randomly reversed (0 -> 1 or 1 -> 0): 5% prob.
  - Target label randomly smoothed over (1 -> 0.7 to 1.2 or 0 -> 0 to 0.3)

Selection Criteria for Interesting Waveforms

From Thomas Grammatico – UVic

Select waveforms:

1. Deposit E>10 MeV in crystal
2. Are a local maximum in a 5x5 region
3. Then, keep all waveforms in a 5x5 region around seed crystal
4. Then, keep all waveforms in a 7x7 region with E>5 MeV



Visualization of the barrel 2D waveform map left after selection criteria are applied.

The neural network will "reproduce" this map and the holes will be filled back with the data saved.

Cost & Validation

Currently, **generating 8736 waveforms** takes **7.7%** the time of a full $\boldsymbol{B\overline{B}}$ **simulation (our flagship decay).**

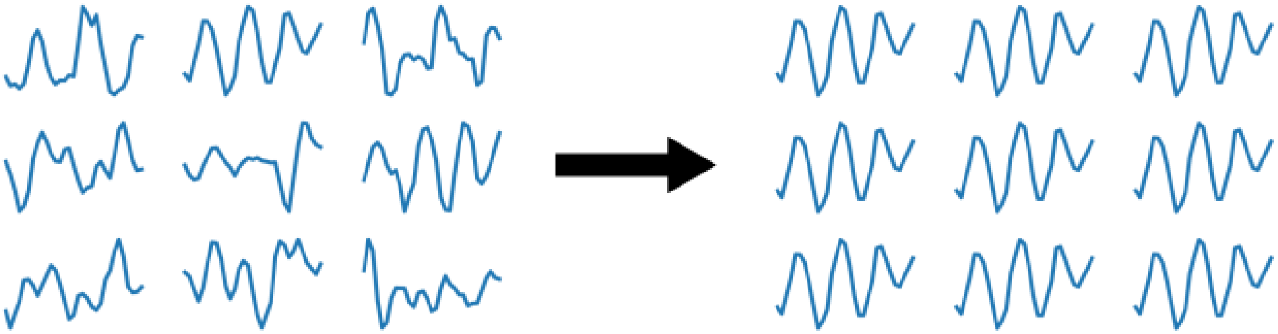
In progress: **validation** of the result by using GAN assisted simulations to perform performance **analyses**, e.g.

- Mass of π_0
- Extra energy in the ECL (E_{extra})
- Timing of energy depositions (clusters)
- Energy distribution of cluster energies

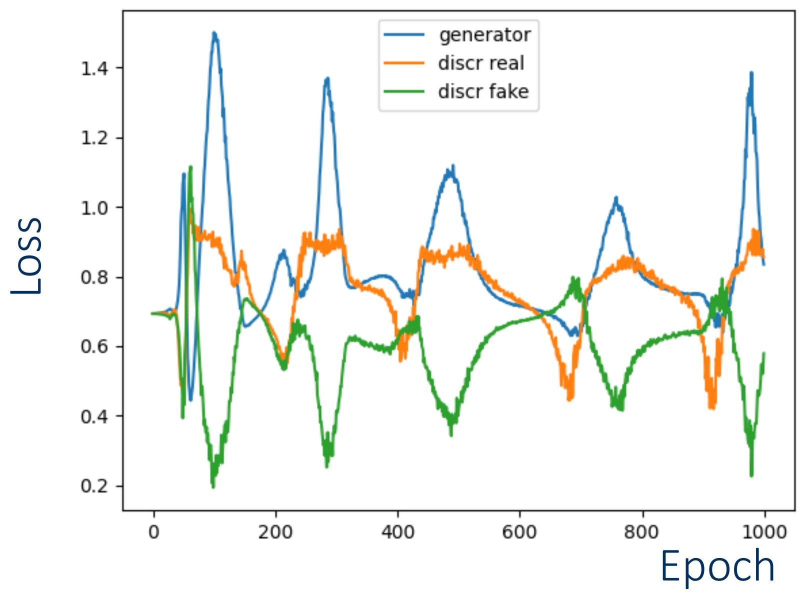
Preliminary analyses looks promising, needs time!

GAN Difficulties

Mode Collapse

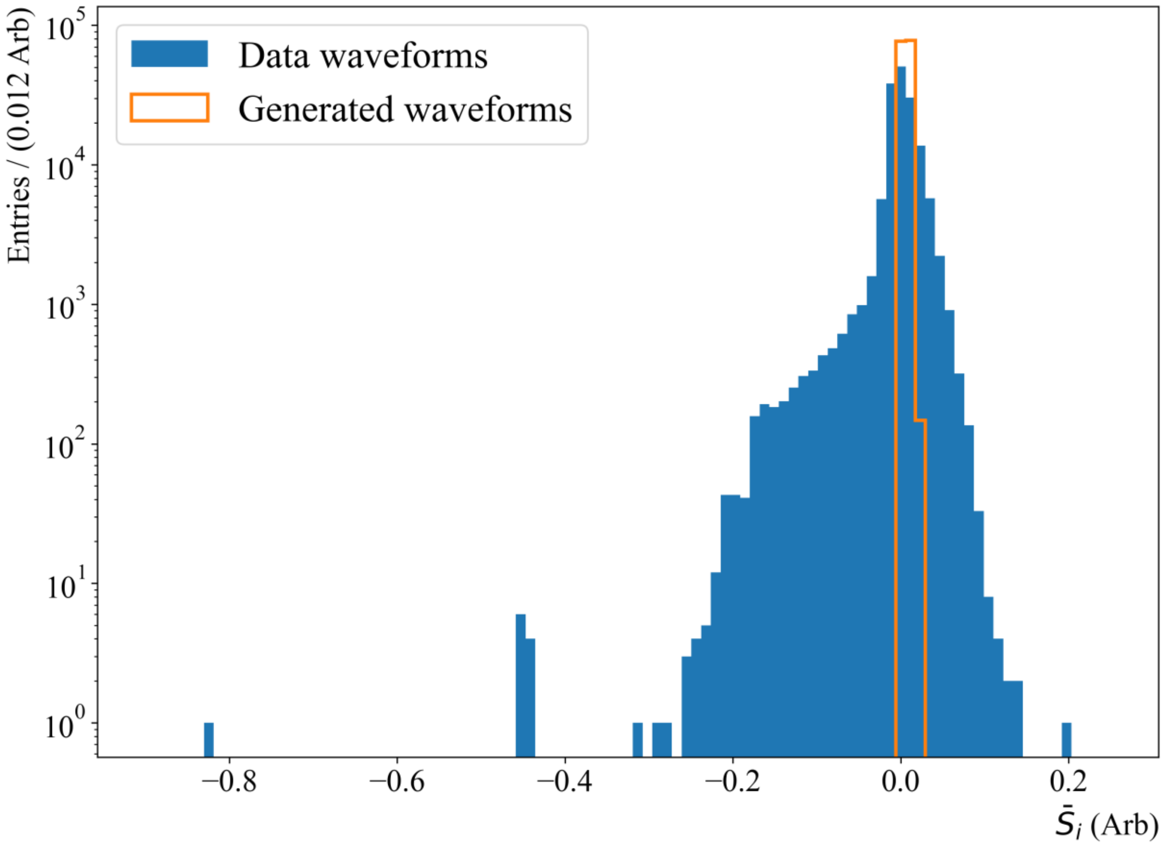


Non-convergence of loss function



Standalone GAN is not Perfect

Waveform average has structure which is not captured



Power Spectrum Unpacked

Each histogram shows the amplitude (power) of a discrete frequency bin (15 in total) for all waveforms in the sample.