Neural Generative Modeling of the Time Projection Chamber responses at the MPD detector

S. Mokhnenko, A. Maevskiy, F. Ratnikov, V. Riabov, A. Zinchenko
HSE University, Moscow, Russia

The use of new methods for processing data of a physical experiment.
Application of machine learning methods on the NICA complex.
28 August - 29 August 2023

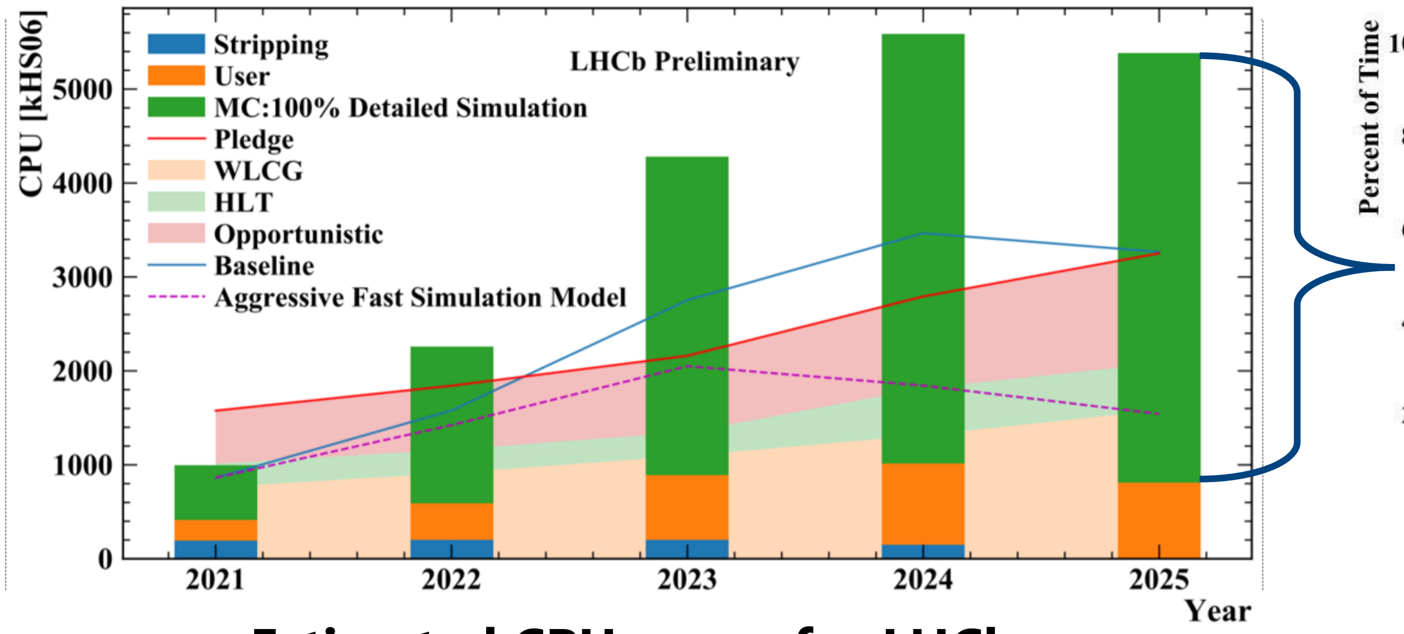
NATIONAL RESEARCH UNIVERSITY

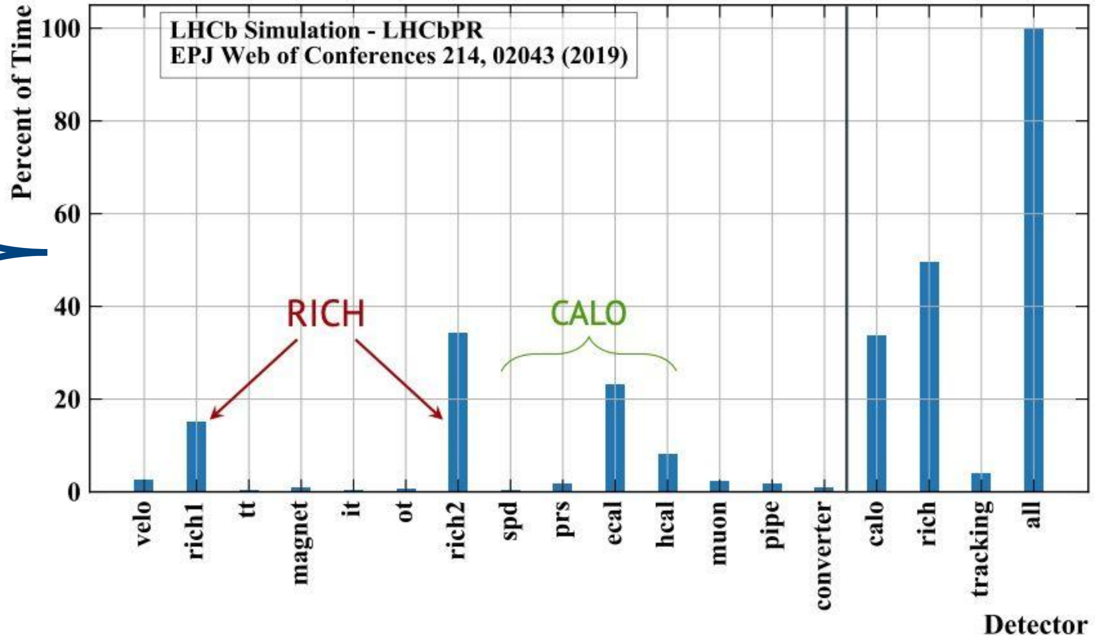
LAMBDA • HSE

Fast simulation problem

- Simulation is an important component in high-energy physics.
- The amount of computation is growing faster than the speed of the processors.
- This problem will get worse with increasing luminosity

LHCb-FIGURE-2019-018

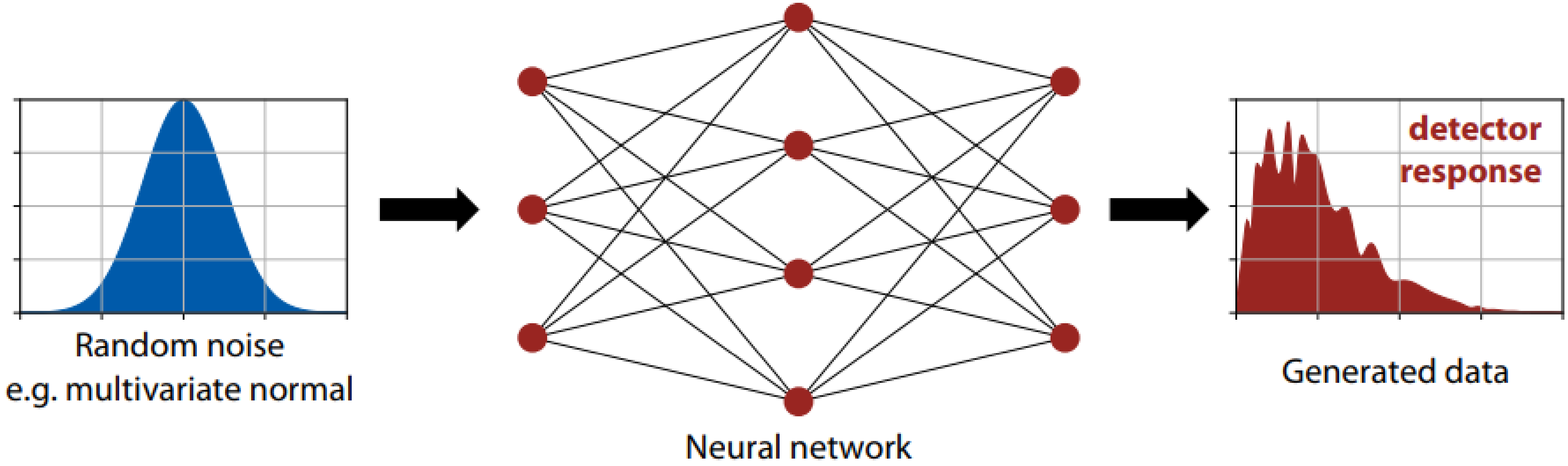




Estimated CPU usage for LHCb

- Several approaches are available: parametric, pre-simulated library, ...
- Generative machine learning models combine the two approaches and allow one to build a parametric model from an existing pre-simulated library.

How can a neural network generate data?

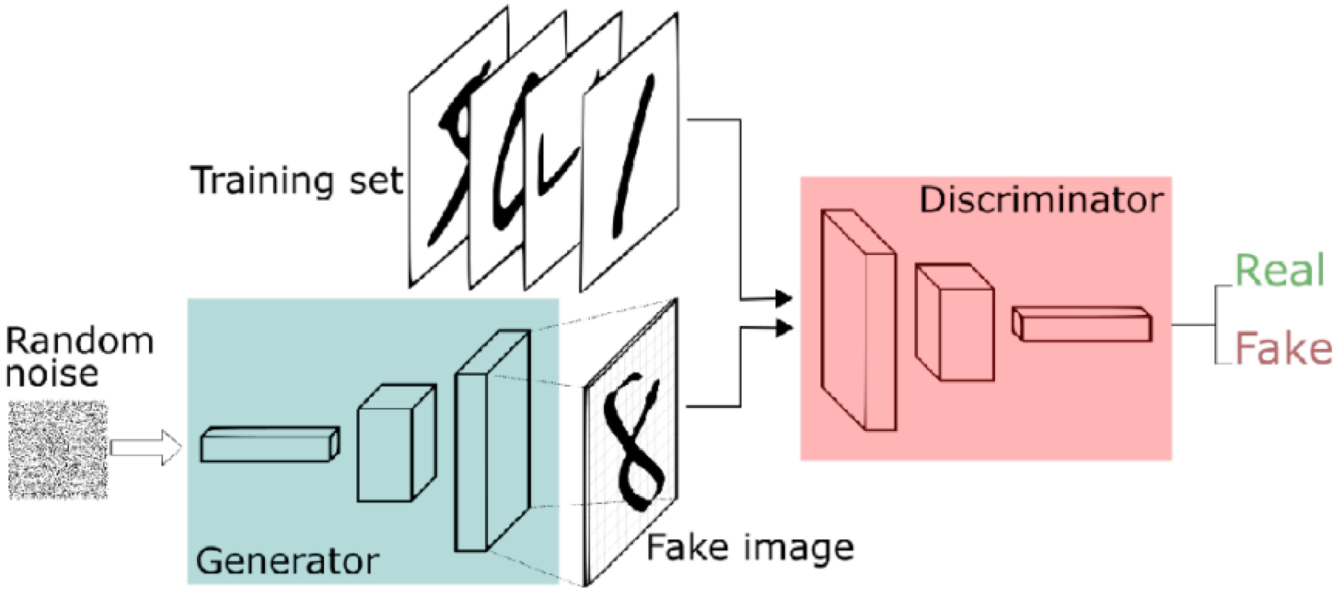


- The task of the generative model is to construct events that correspond to some probability distribution.
- Generating a sample is fast as well-developed and effective industrial ML methods are used.

Generative adversarial networks (GANs)

- There are different approaches to generative models in ML
- Generative adversarial networks (GANs) offer the fastest sampling

- GANs consist of two neural networks: **generator** is trained to creates samples, **discriminator** is trained to distinguishes true samples from those created by generator
- As a result, generator and discriminator dynamically improve each other



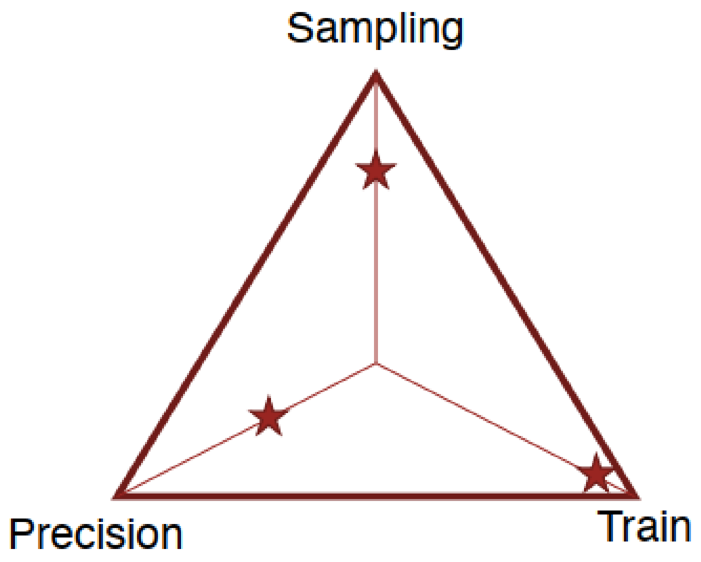
Comparison GANs with traditional methods

- GANs sampling is much faster than direct Geant4
  - Geant4 is accurate and reliable.
  - Geant4 is still considered as a reference
- GANs are flexible comparing to rigid parametric models.
- GANs produce nice smooth distributions comparing to discrete distributions produced by library

- However, making GANs to really work, requires care of some typical problems, which we are going discuss in a moment.

Generative models characteristics

- Fast Sampling
  - much faster than detailed Geant4
  - models can get complicated
- Very Fast training
  - retrain can be done very fast
  - train process still should be periodically controlled
- Good Precision
  - complicated models can be quite precise
  - precision is controlled by train sample statistics

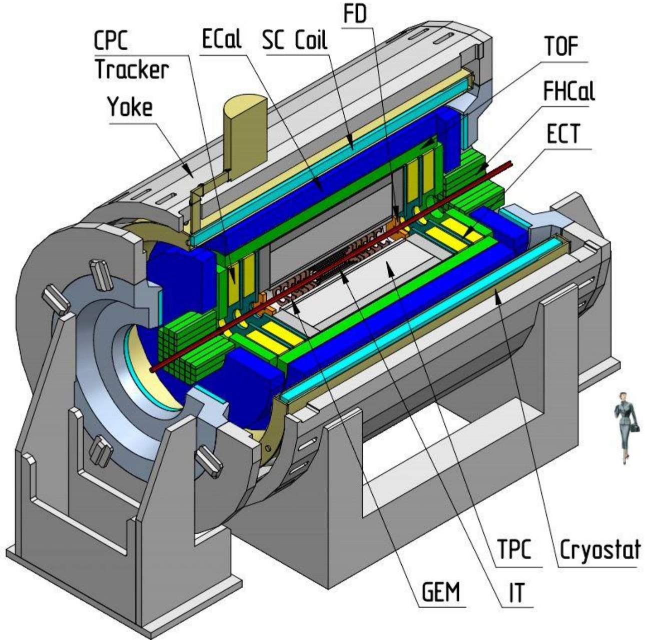


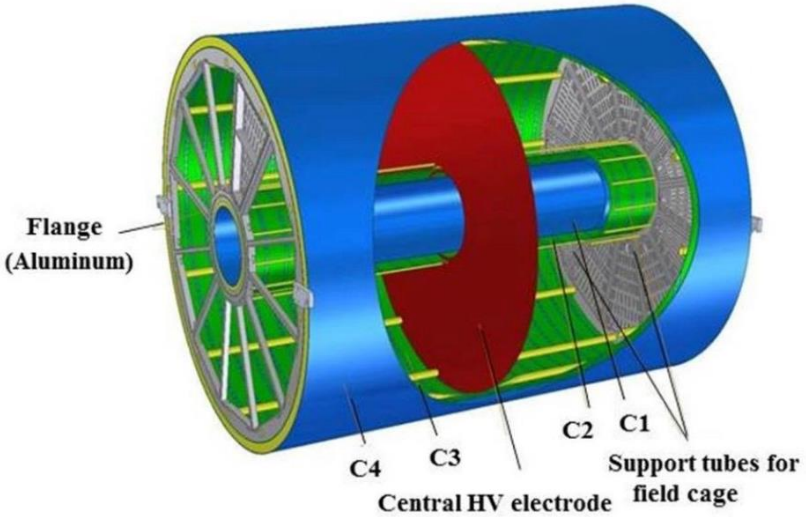
Possible approaches

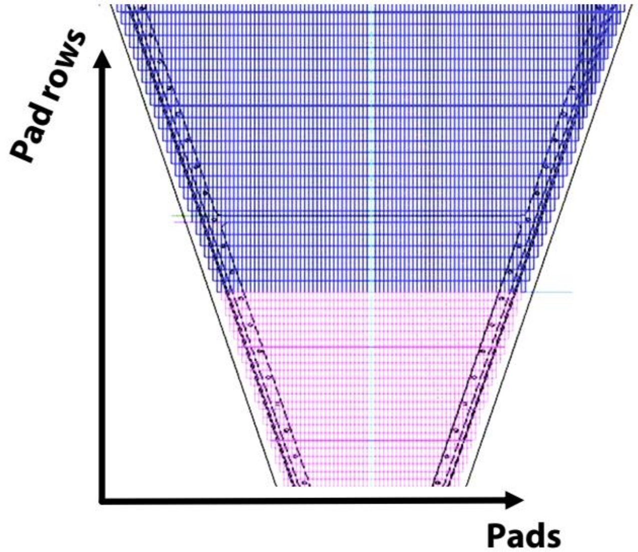
- GANs can be used to sample:
  - Raw signal images from the detector
  - High-level reconstruction results
- GANs can be trained using:
  - Real data
  - Simulated data
- GANs can be used to simulate
  - Whole detector
  - Individual sub-detectors

GAN for NICA
Multi-Purpose Detector

Time projection chamber







3968 pads * 12 sectors * 2 endcaps = 95232 total pads

Problem statement

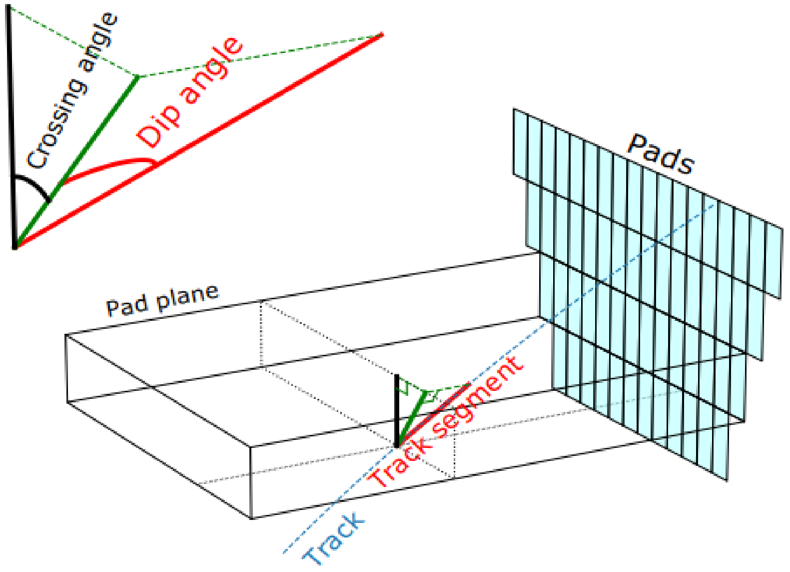
Main goal is fast generation of the signal for Multi-Purpose Detector in Time projection chamber

Train sample:

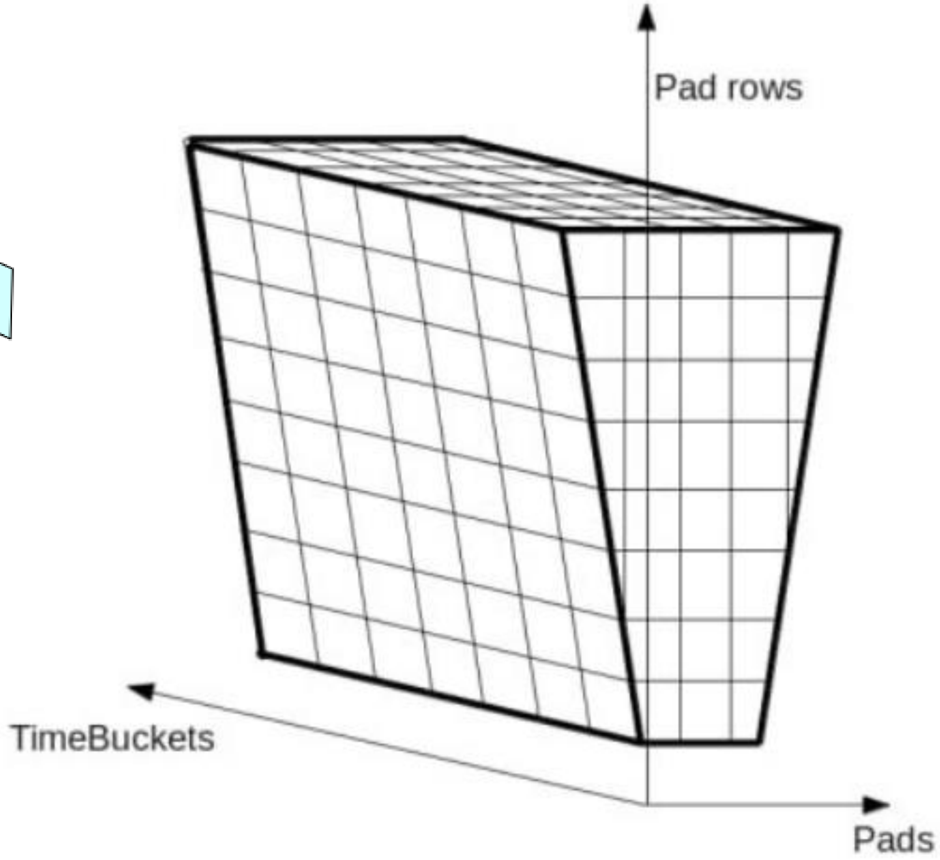
- Simulated data for pion

Input:

- 2 angles (θ, ϕ)
- 3 coordinates per track segment



Track segment parameters



Output:

- $95\,232 \cdot 310$ elements (pads x time buckets)
- Conditioned on the track parameters for the whole event

Dimensionality reduction

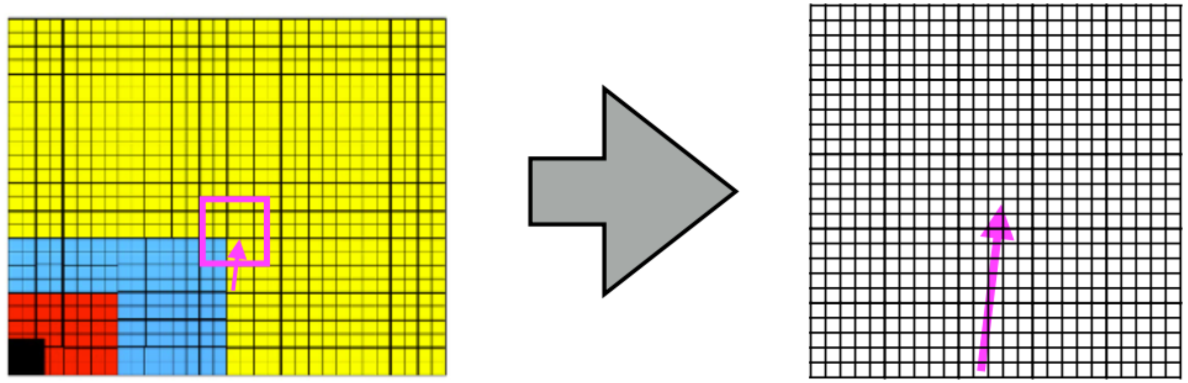
We can hardly build generative model for the full detector

- many channels - high dimensional objects.

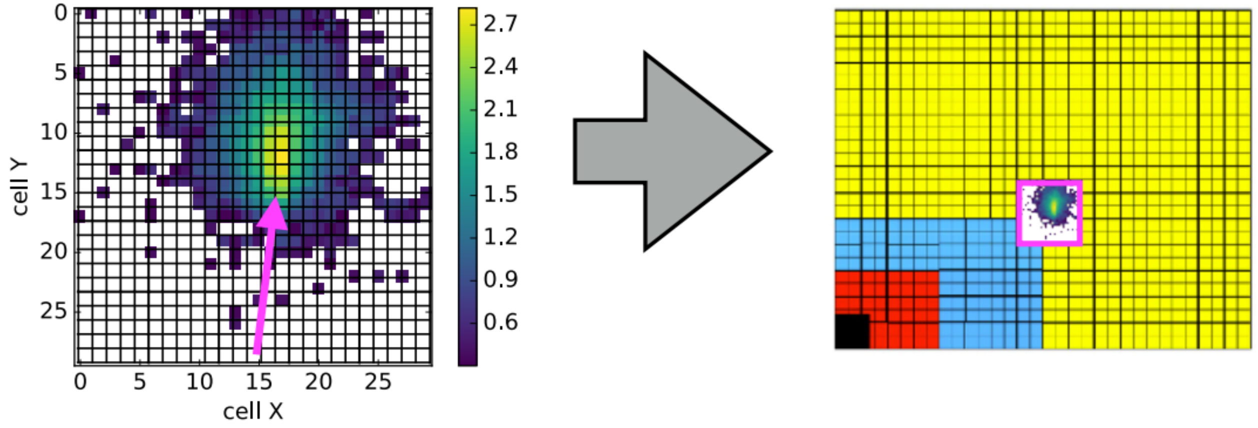
Response of the impact particle is usually local

- can limit generated object to the local area of the response

Global -> local ML

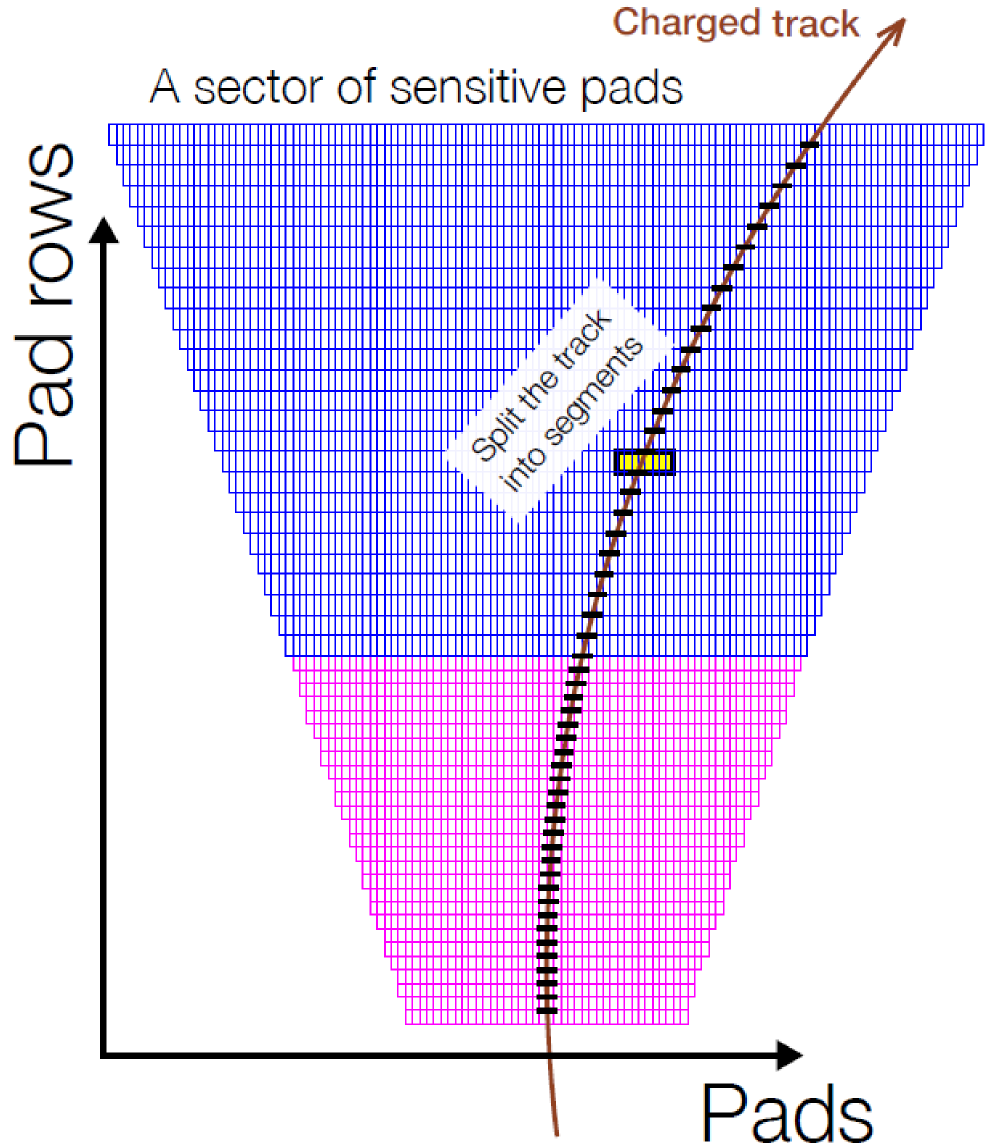


local ML -> global



Assumptions for fast simulation

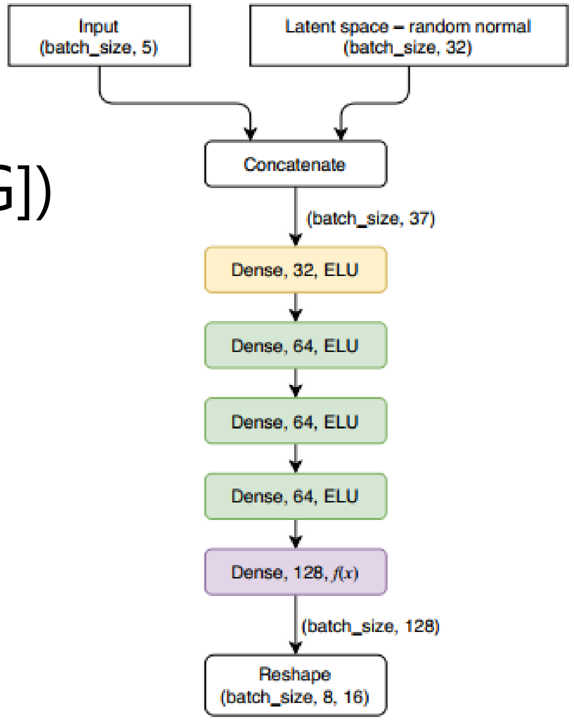
- Factorizing the pad rows
  - dividing tracks to segments, each contributing to a particular pad row
  - can model such contributions independently!
- Signal localization (both position & time)
  - model only a small area instead of the full row
  - model only a few time buckets
- Target dimensionality:
  8 pads x 16 time buckets
  (instead of original 95 232 * 310)



Model architecture

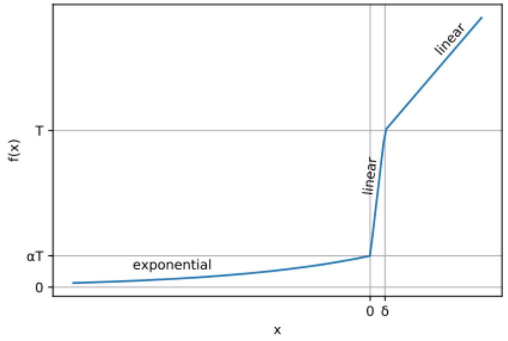
- Model: WGAN-GP (arXiv:1704.00028 [cs.LG])
- Generator:
  - Fully connected
  - ELU activations, custom output layer activation
  - 5 layers
- Discriminator:
  - Deep convolutional NN
  - ELU activations
  - Dropout layers
- Optimization: RMSprop, learning rate exponential decay

Generator

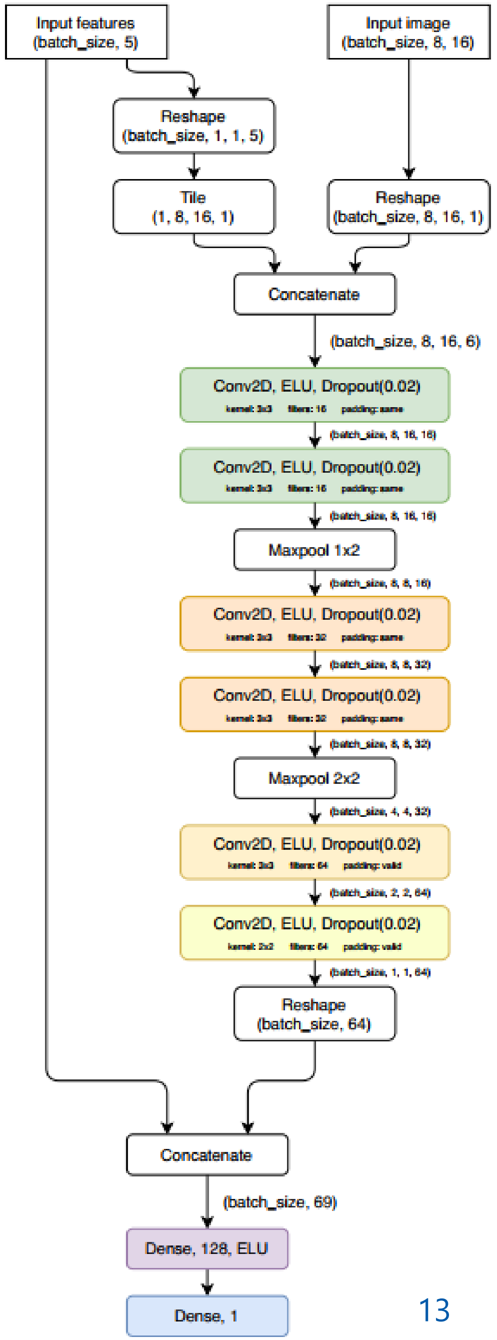


Custom activation

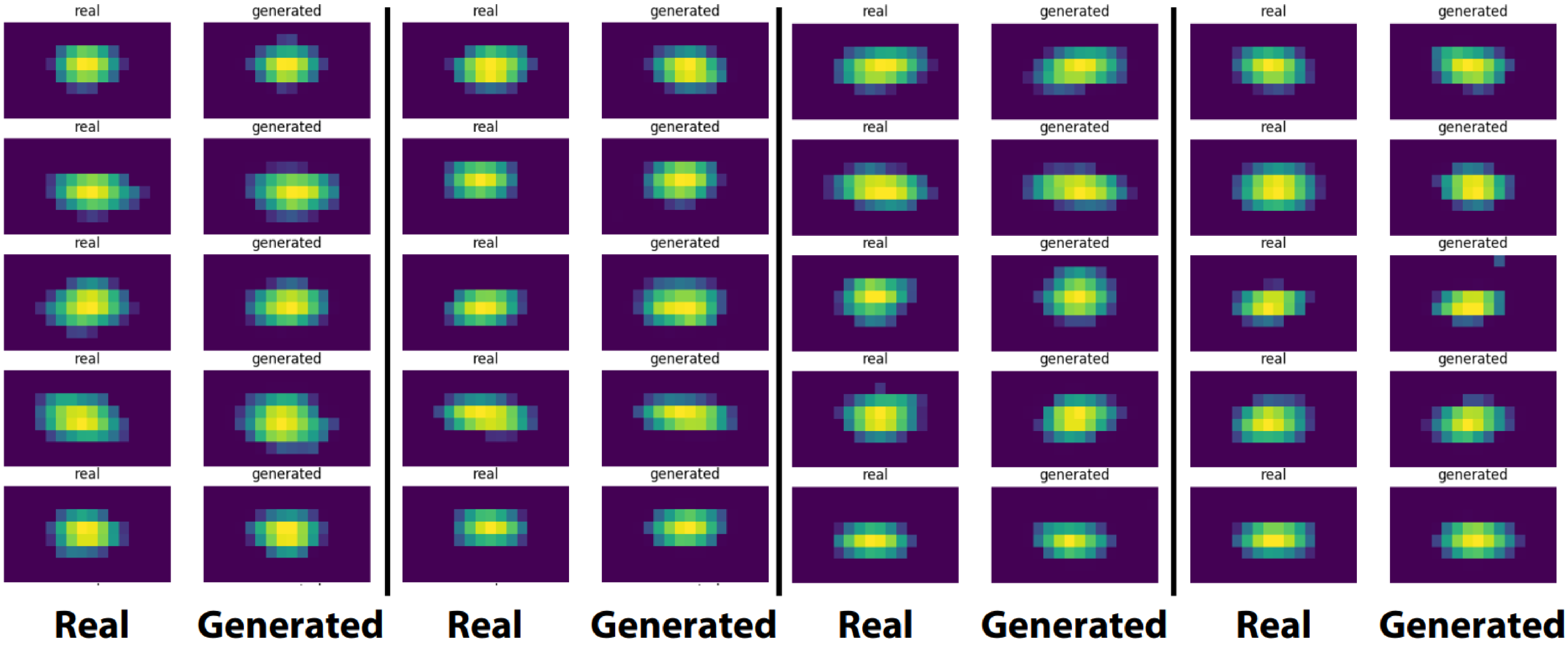
$$f(x) = \begin{cases} \alpha T e^{x} & x \leq 0 \\ T\left(\alpha + (1-\alpha)\frac{x}{\delta}\right) & 0 \leq x \leq \delta \\ T - \delta + x & \delta \leq x, \end{cases}$$



Discriminator



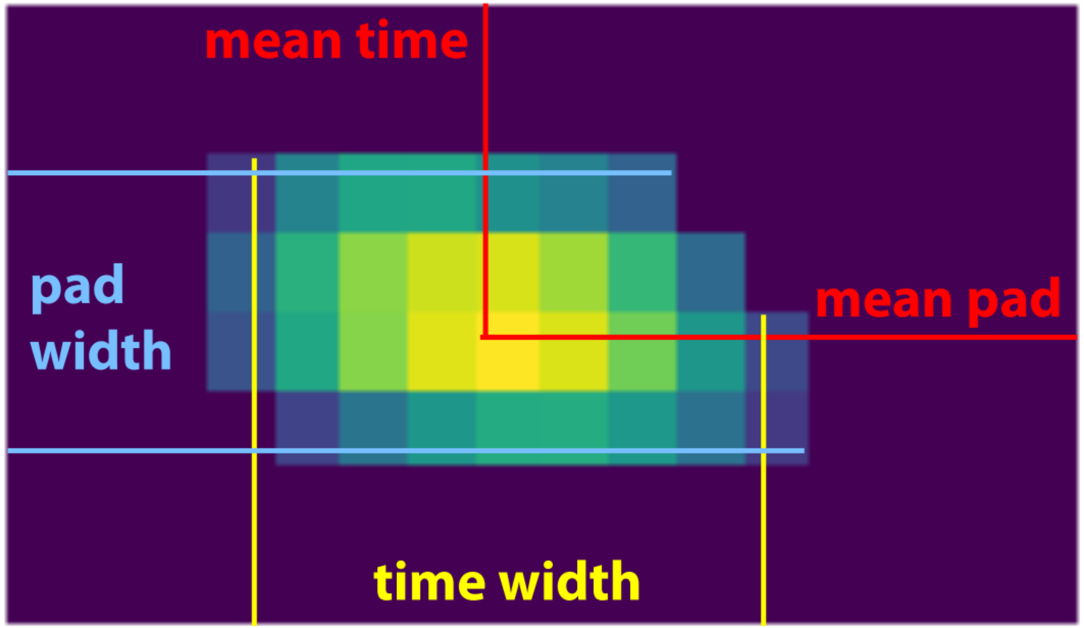
Raw pad responses



A Maevskiy et al Eur. Phys. J. C 81, 599 (2021)

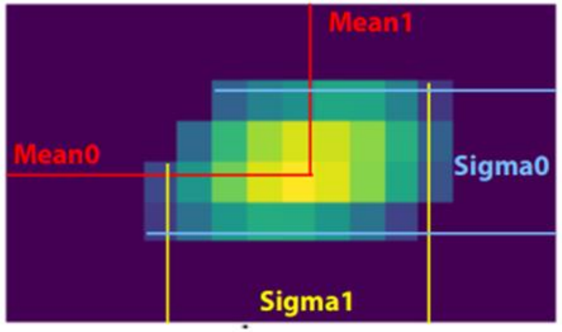
Low-level metrics

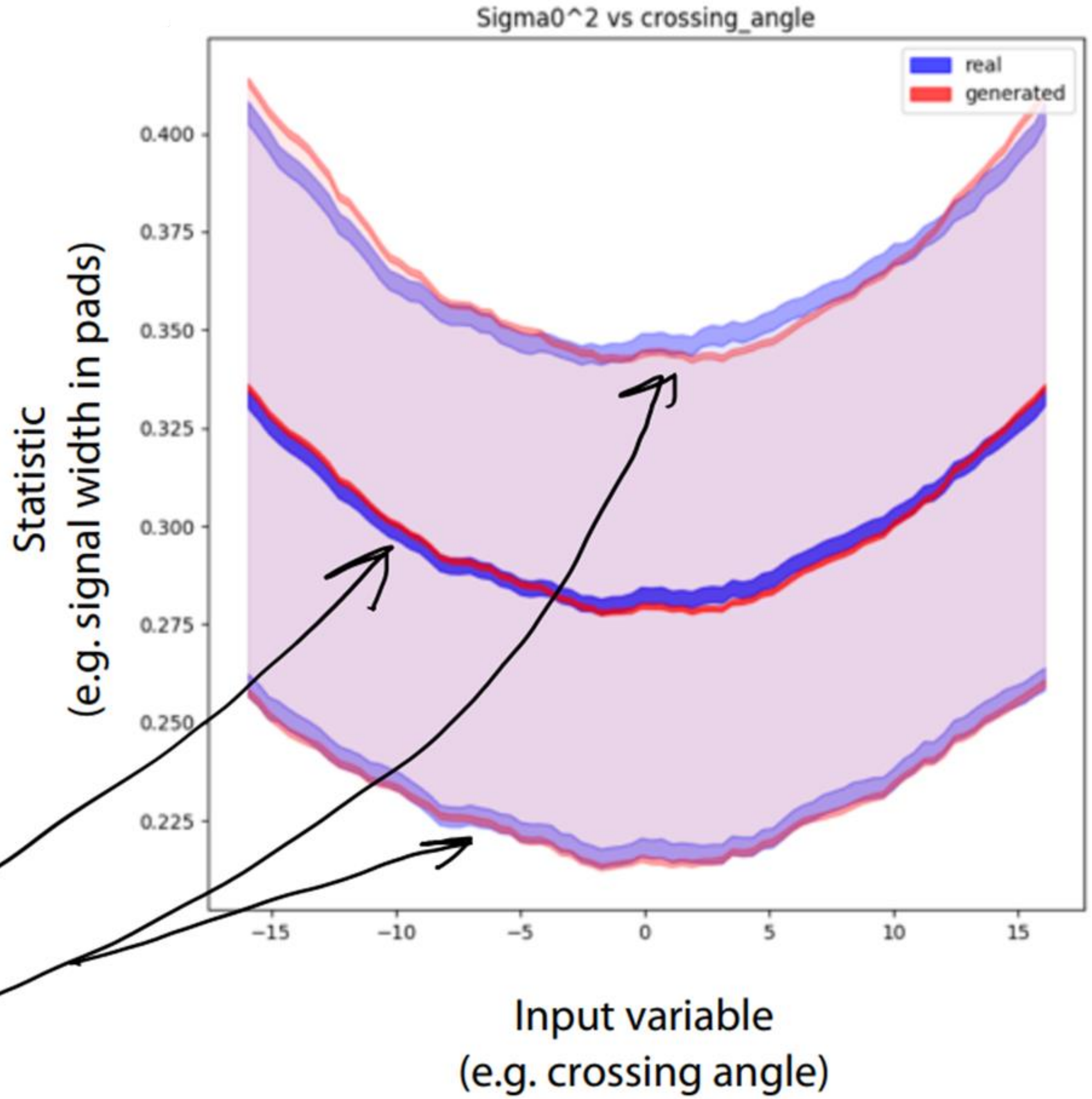
- Start with a simple preliminary metric: we compare the 1st & 2nd order moments of the signal images, i.e.:
  - the location of the signal in pads and time bins
  - the widths of the signal in pads and time bins
- Also looking at the integrated amplitudes
- All this as a functions of track segment parameters (2 angles + 3 coordinates)



Low-level metrics - profiles

Widths of the shaded lines correspond to the statistical uncertainties

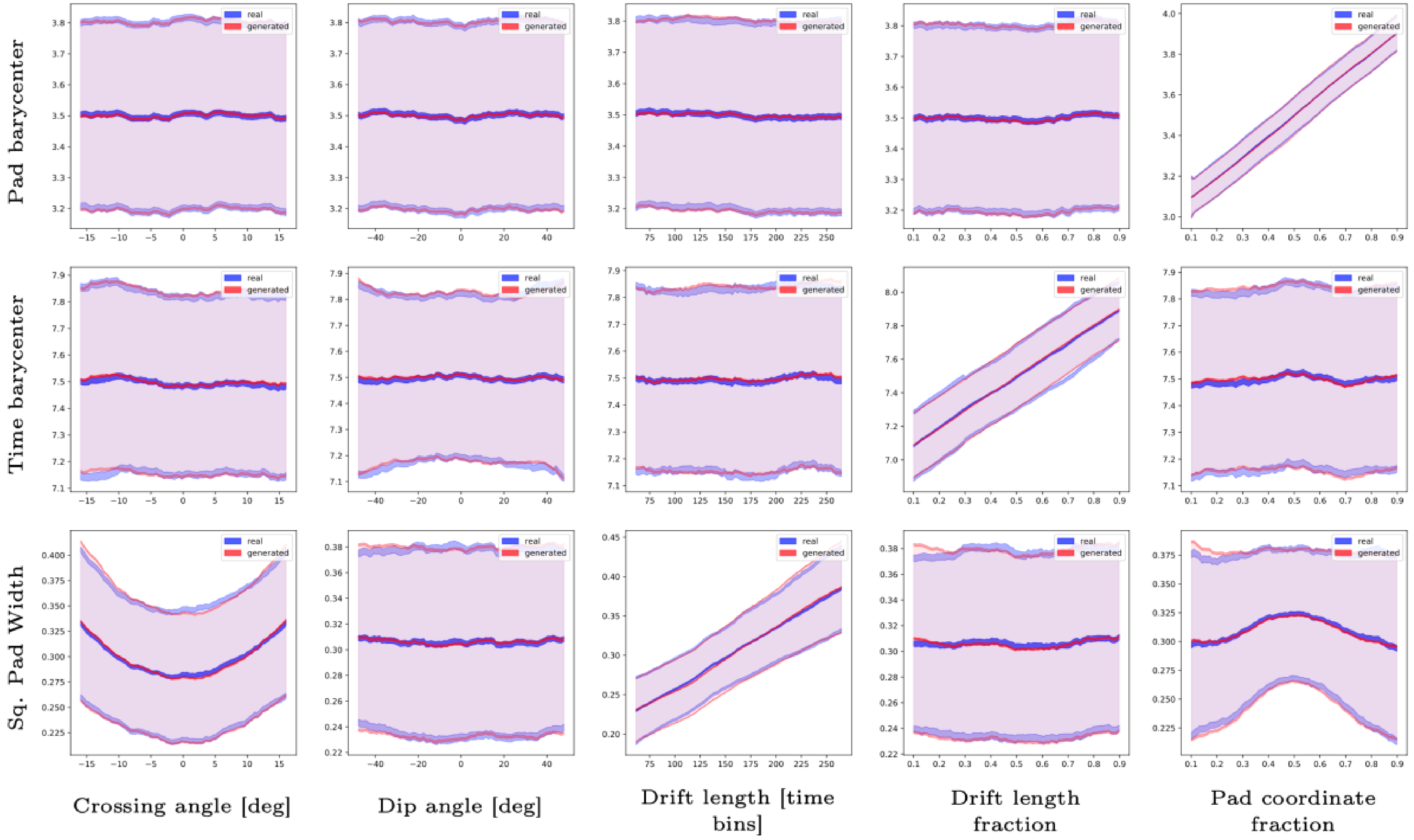




Mean of the statistic

Mean ± 1 standard deviation

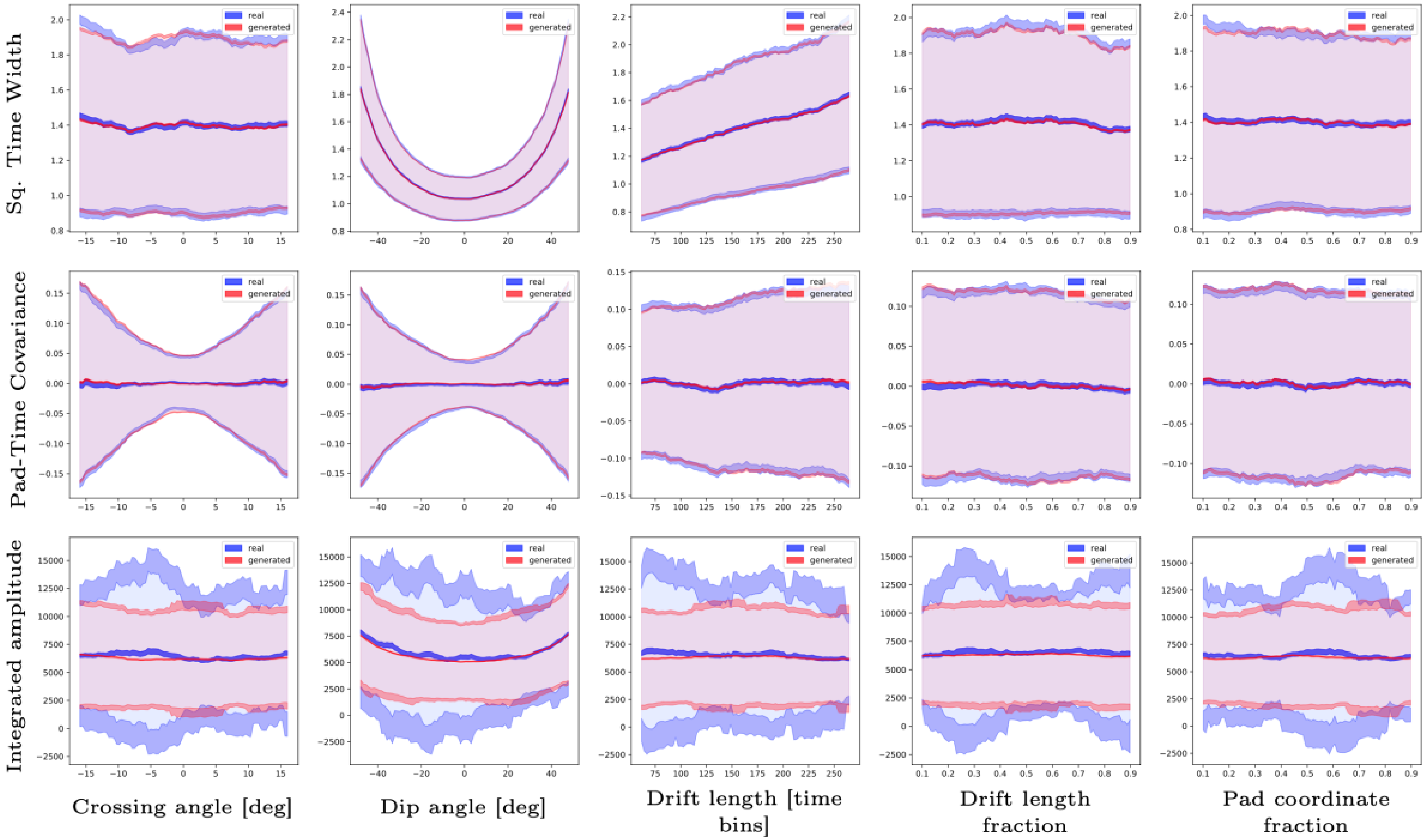
Low-level metrics - profiles



Mostly good agreement

"Real"
Generated

Low-level metrics - profiles



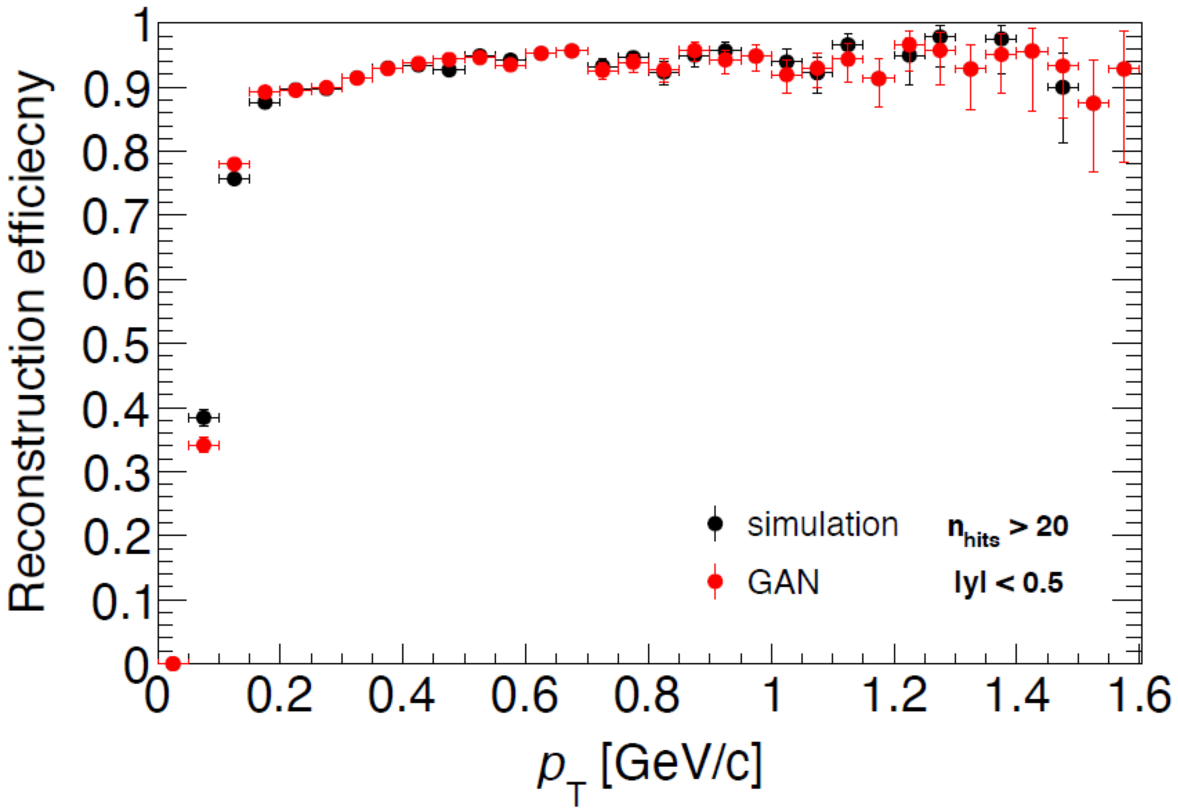
Mostly good agreement

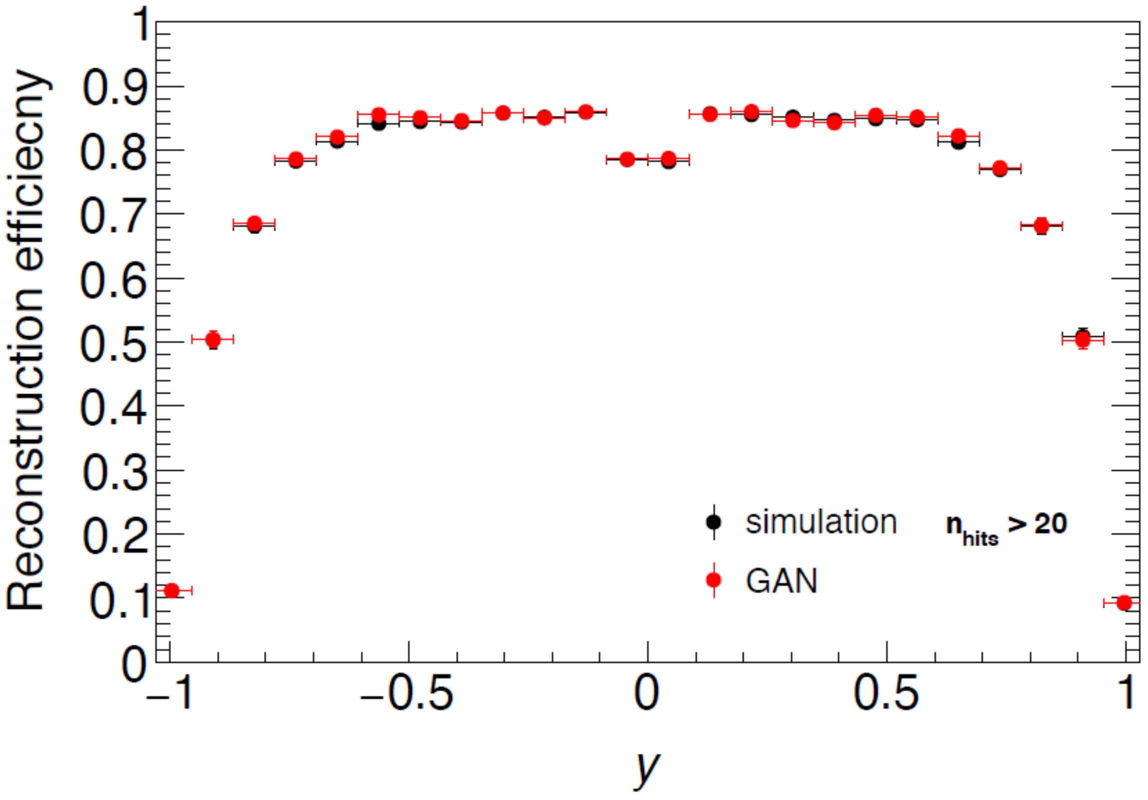
Integrated amplitude can be factorized out and simulated separately from 1st principles

"Real"
Generated

Physics-level model quality metric

At reconstruction level we can consider reconstruction efficiencies





A Maevskiy et al Eur. Phys. J. C 81, 599 (2021)

Agreement looks pretty good. Our assumptions make sense

Conclusion

- Generative adversarial networks may boost simulations of elementary particle detectors by orders of magnitude compared to regular Geant3(4).
- Dimension of problem may be significantly reduced by considering specific structure of detector.
- Our model accelerates the detailed simulation by at least an order of magnitude and
- It is capable of producing detector responses that look authentic in both low- and high-level validation procedures.
- We are currently working on accounting for correlations between rows of pads.

Backup

Library vs Generative Approach

Reference dataset is necessary to train generative model

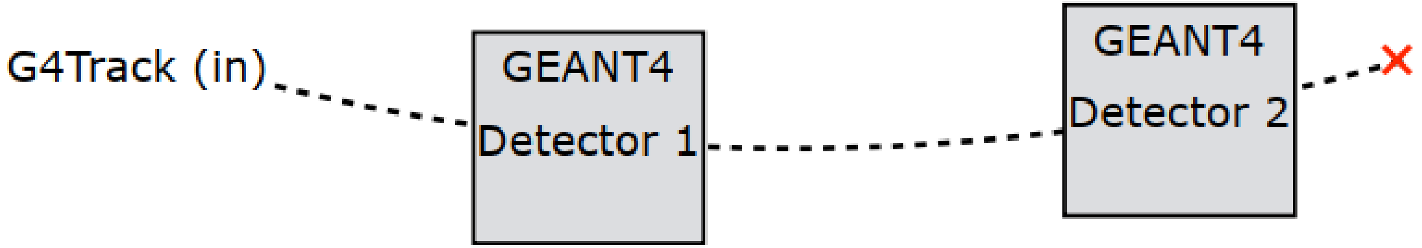
Reference dataset may be used to sample objects directly

- approach accommodated by CMS, ATLAS, LHCb
- PRO library approach comparing to generative models
  - aggregated distributions are guaranteed by construction
- PRO generative models comparing to library approach
  - discreetness of events
    - partly compensated by energy scaling
  - speed
    - massive matrix operations vs massive object search
  - size
    - both transient and persistent

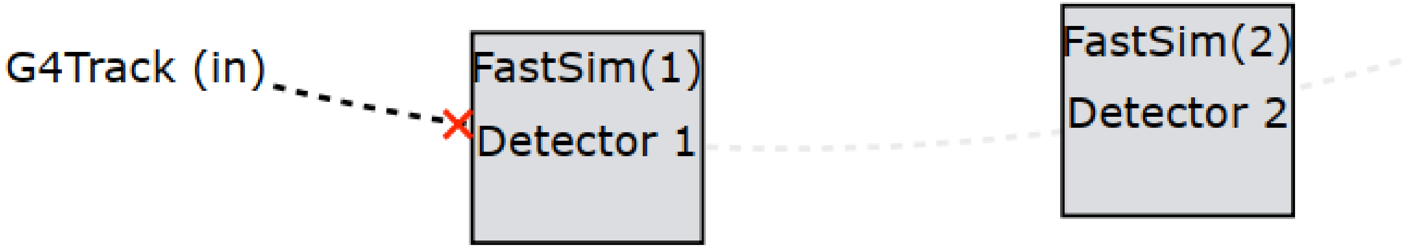
From technical perspective, library-based and ML-based modules have very similar interfaces for both gathering train data and inferencing objects

Operation Scheme

To speed up Geant4 we need to intercept G4Track in front of the detector, generate detector response, fill DetHits structures



$$\text{G4Track(in)} \xrightarrow[\text{GEANT4}]{} \text{Detector1Hits, G4Tracks} \xrightarrow[\text{GEANT4}]{} \text{Detector2Hits}$$



$$\text{G4Track(in)} \xrightarrow[\text{FastSim}]{} \text{Detector2Hits} \oplus \text{Detector2Hits}$$

Evaluation metric

- We measure the **efficiency** of RichDLLx cuts at various quantiles of the RichDLLx distribution:

$$\varepsilon = \frac{number\ of\ tracks\ above\ x\%\ threshold}{total\ number\ of\ tracks}$$

- Do this as a function of the input variables:

$$\varepsilon(P, \eta, nSPDHits)$$

- Calculate the **efficiency ratio** between GAN predictions and simulated events (in bins of a variable):

$$efficiency\, ratio = \frac{\varepsilon_{GAN}}{\varepsilon_{simulated}}$$