Generative adversarial networks in particle physics

Sergey Mokhnenko, Denis Derkach, Artem Maevskiy,
Fedor Ratnikov, Alexander Rogachev
HSE University, Moscow, Russia

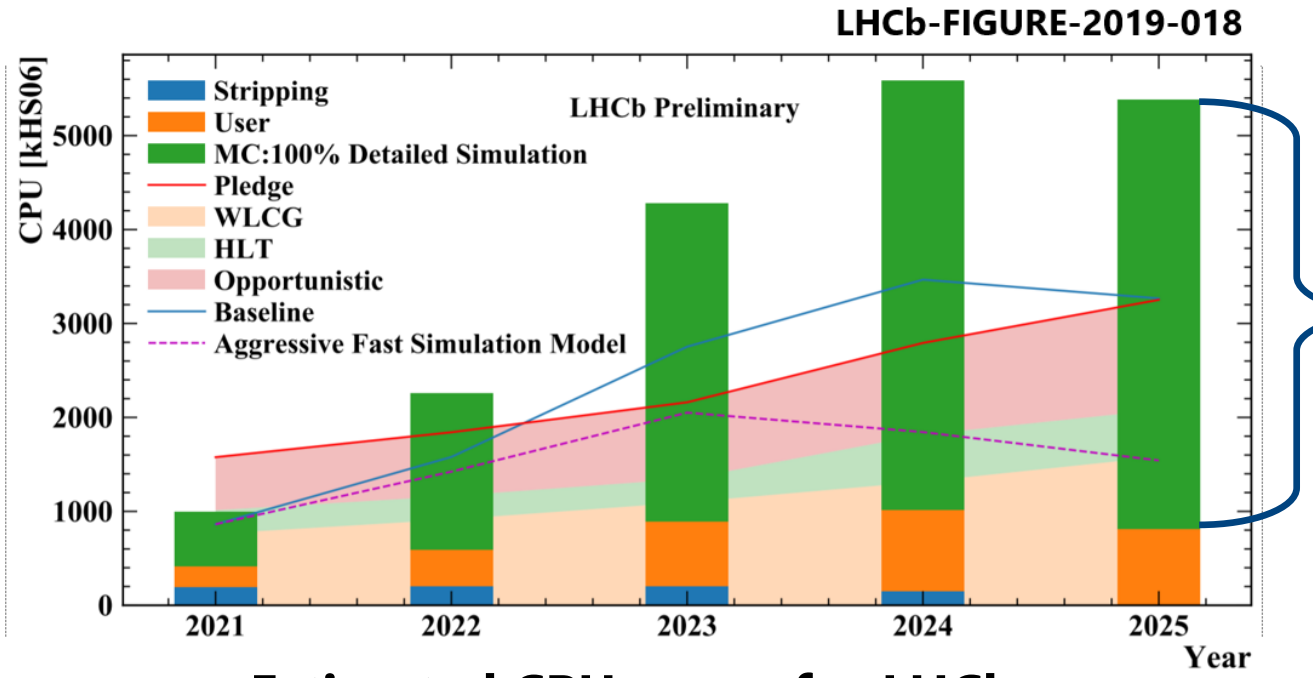
The 6th international conference on particle physics and astrophysics
29 November 2022 - 2 December 2022

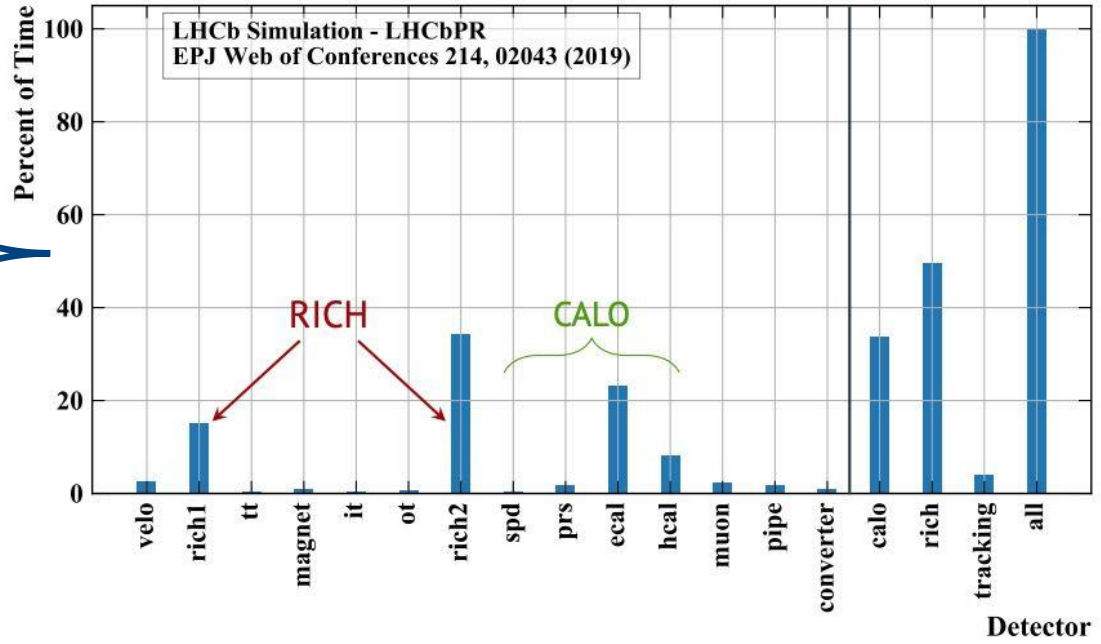
NATIONAL RESEARCH UNIVERSITY

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

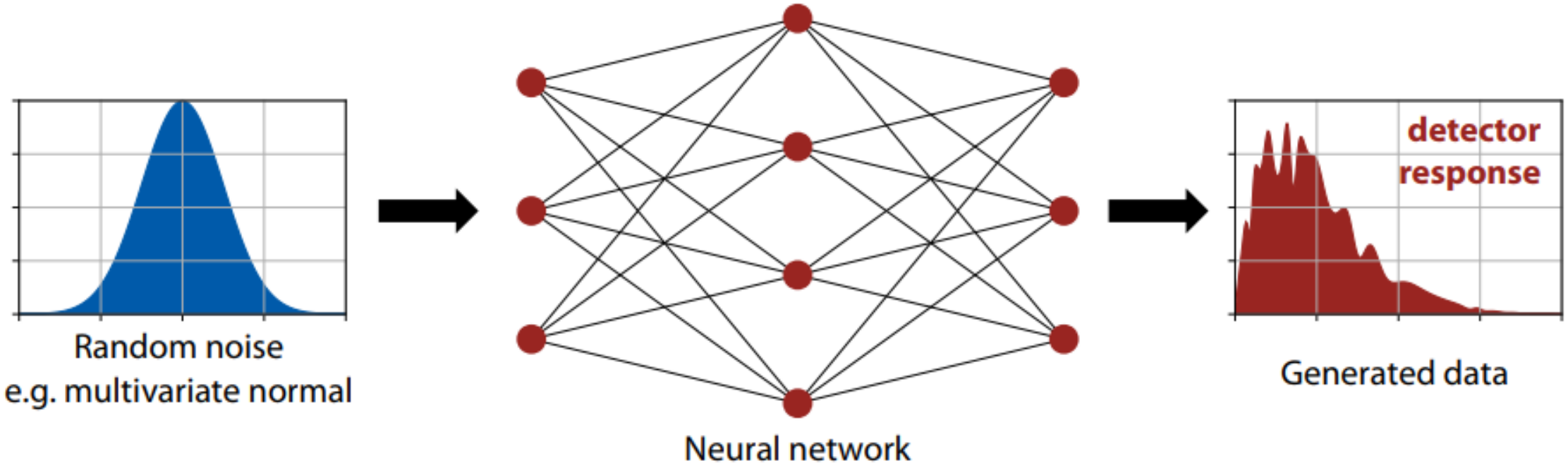




Estimated CPU usage for LHCb

- 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?

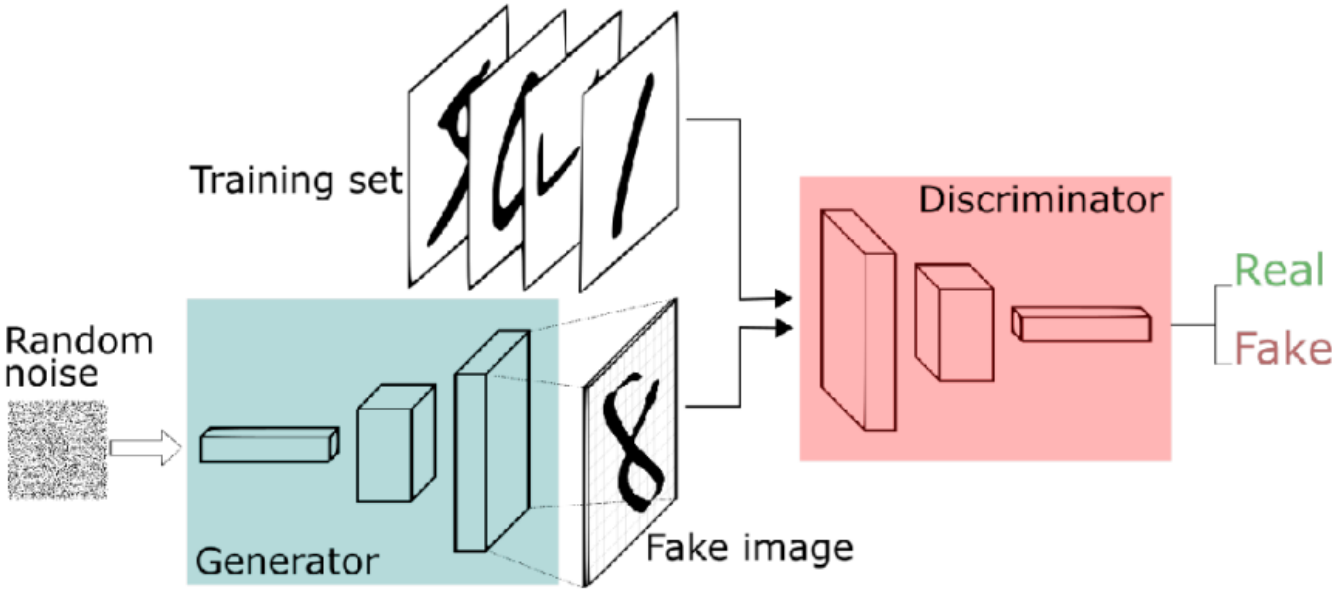


- The task of the generative model is to construct events that correspond to some probability distribution.
- Generating a sample is fast as well-developed and effective industrial ML methods are used.

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



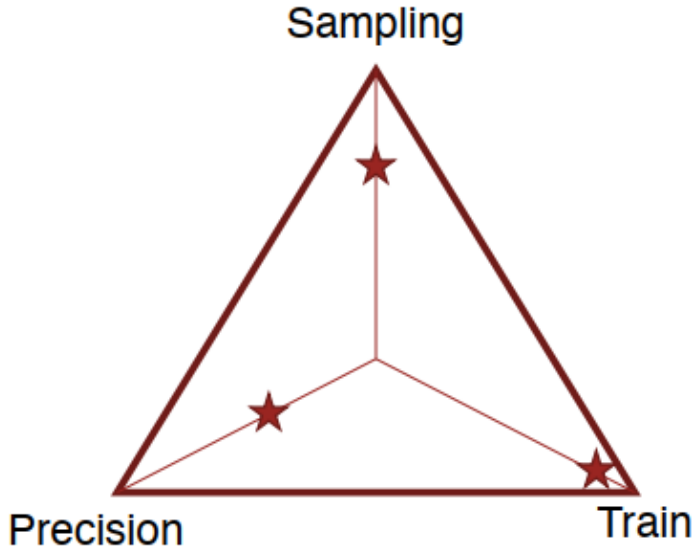
Comparison GANs with traditional methods

- 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



Dimensionality reduction

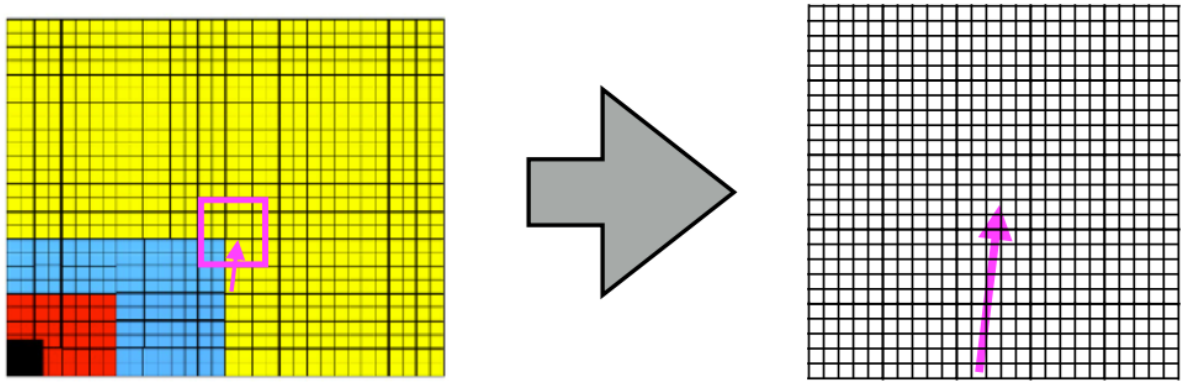
We can hardly build generative model for the full detector

- many channels - high dimensional objects.

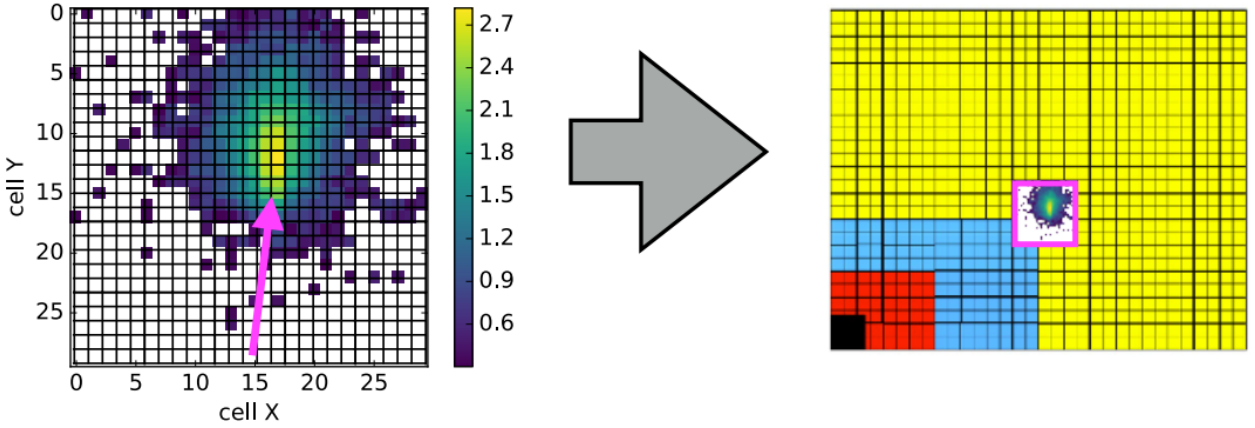
Response of the impact particle is usually local

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

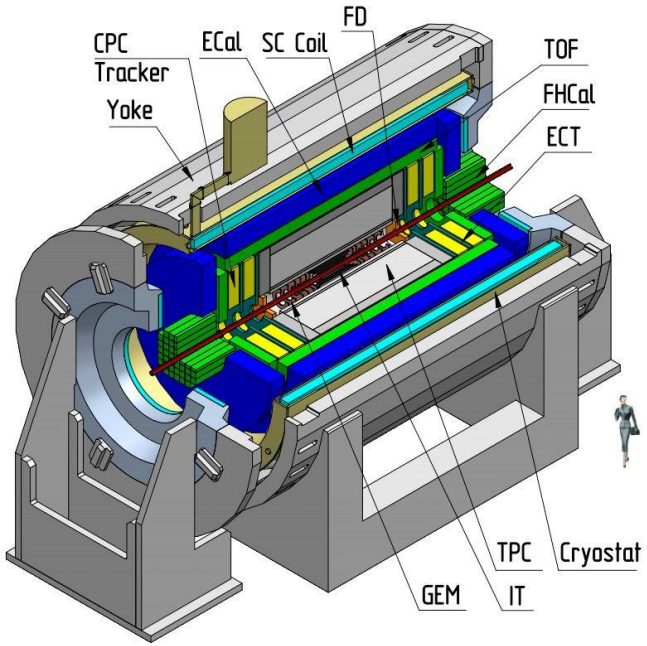
Global -> local ML

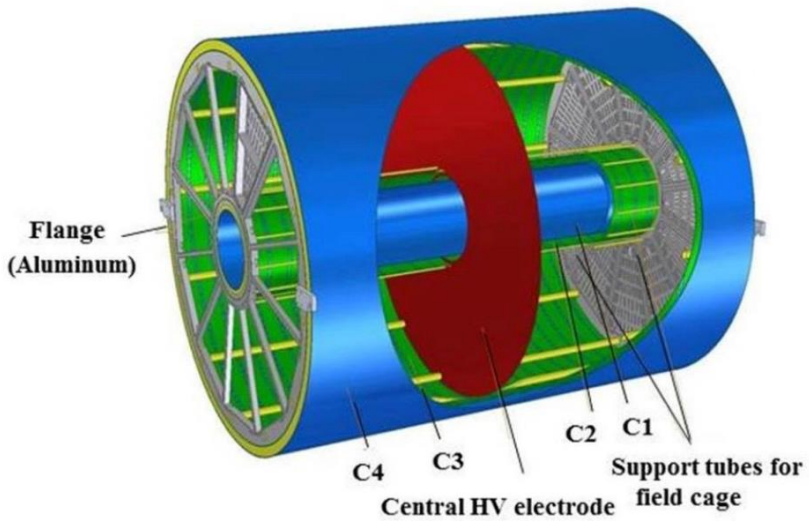


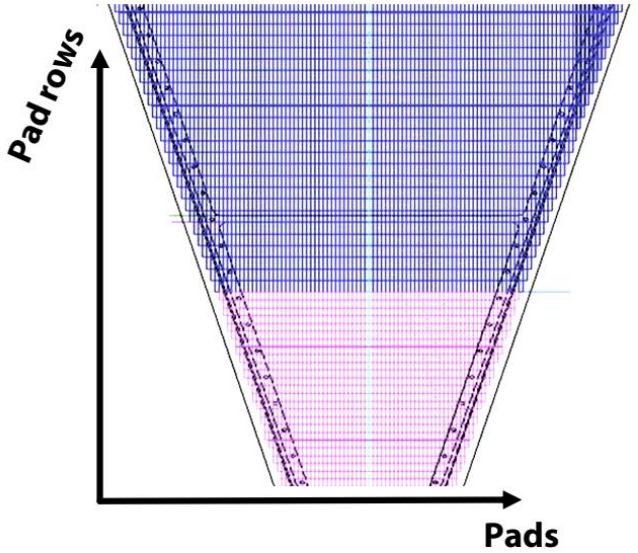
local ML -> global



Time projection chamber



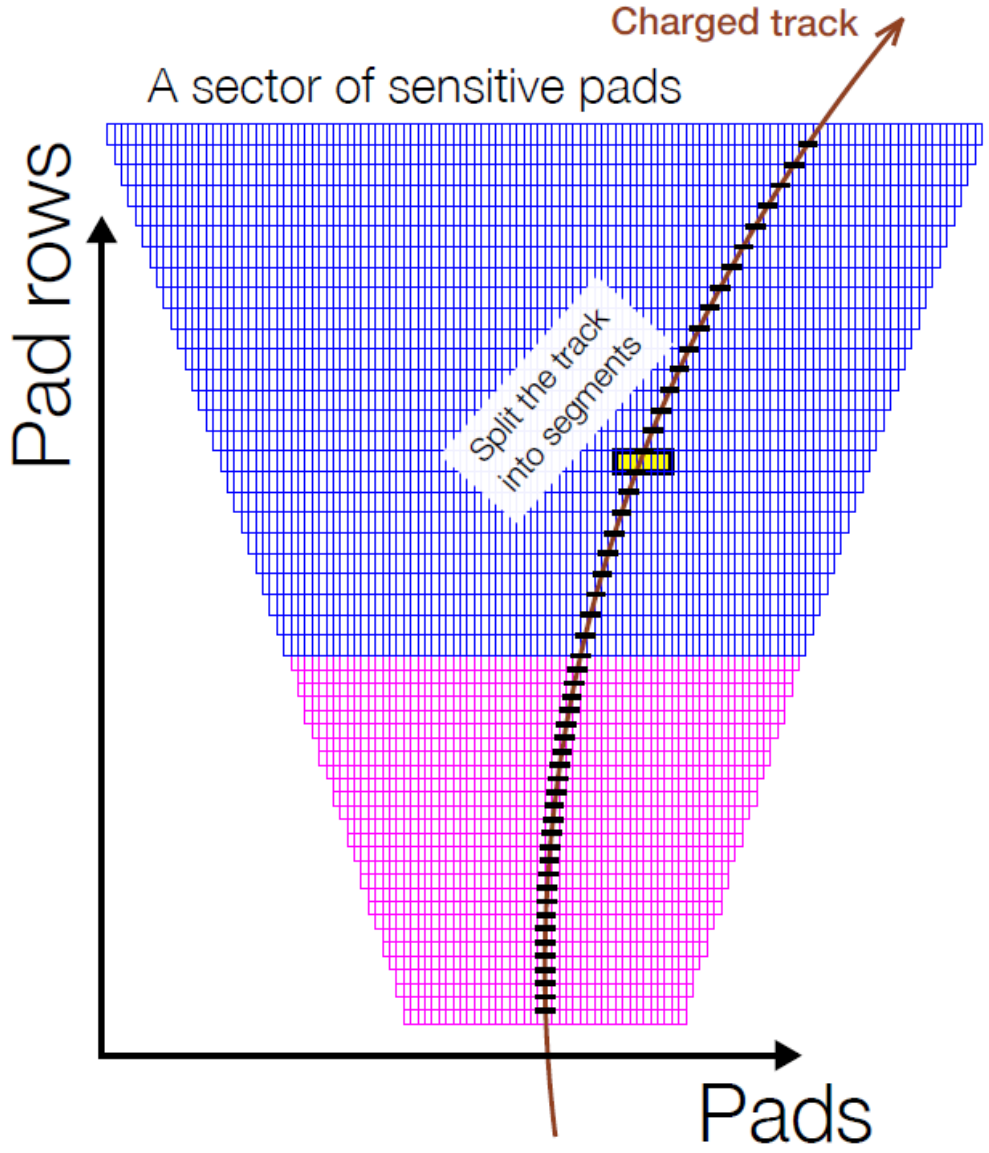




3968 pads * 12 sectors * 2 endcaps = 95232 total pads

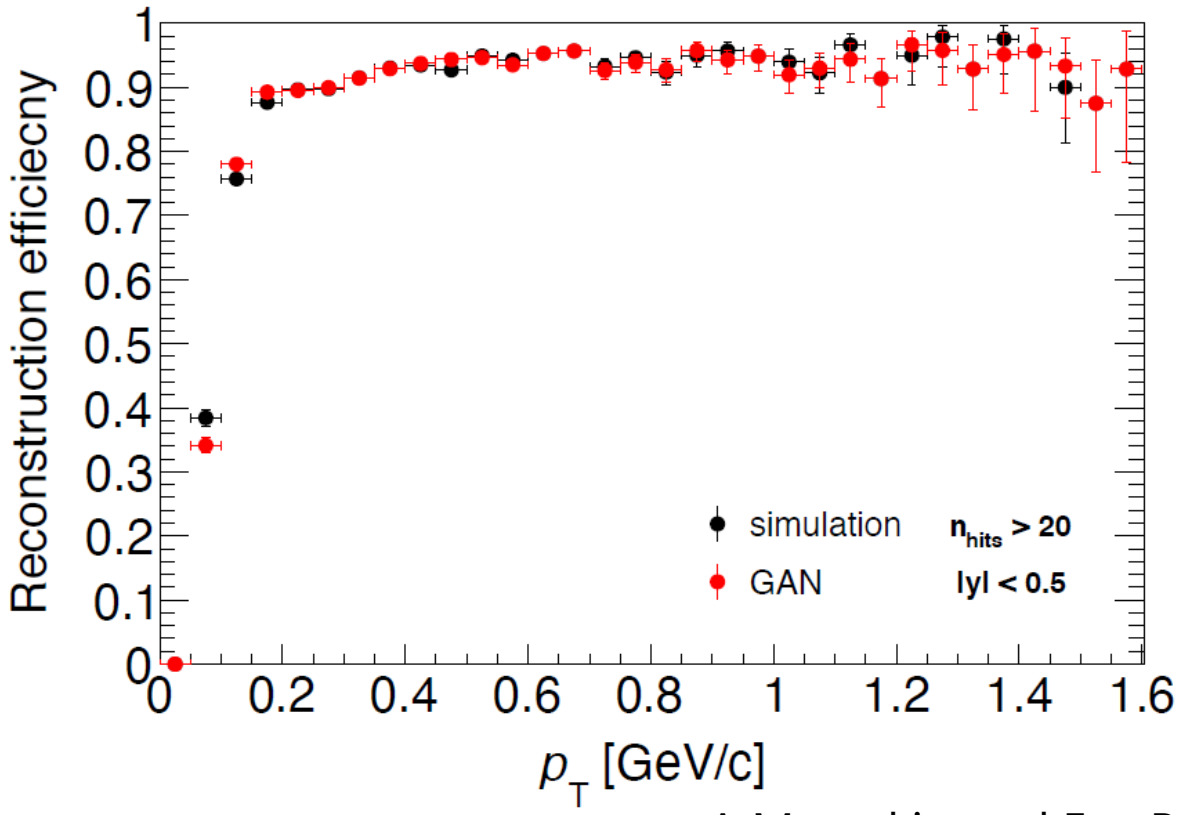
Assumptions for fast simulation

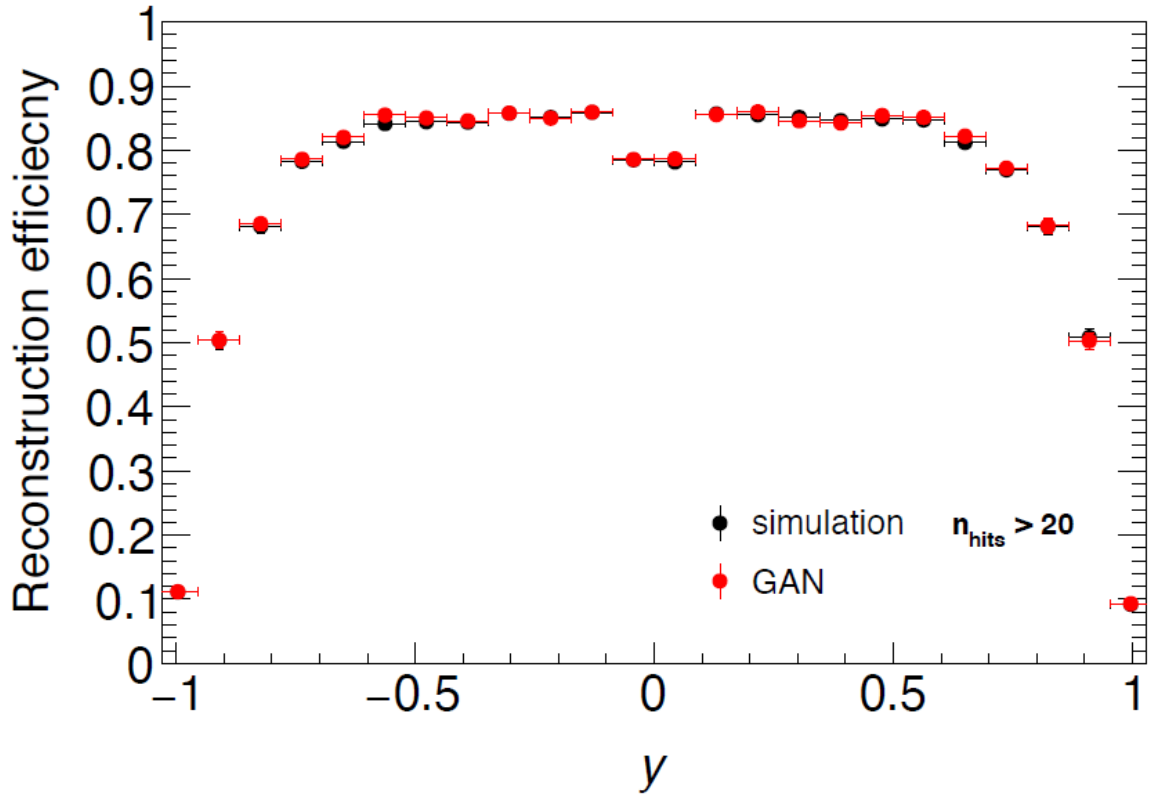
- Factorizing the pad rows
  - dividing tracks to segments, each contributing to a particular pad row
  - can model such contributions independently!
- Signal localization (both position & time)
  - model only a small area instead of the full row
  - model only a few time buckets
- Target dimensionality: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



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

Another dimensionality reduction approach

The detector may be too complex to fully simulate. For example, accurate modeling of Cherenkov detectors would include:

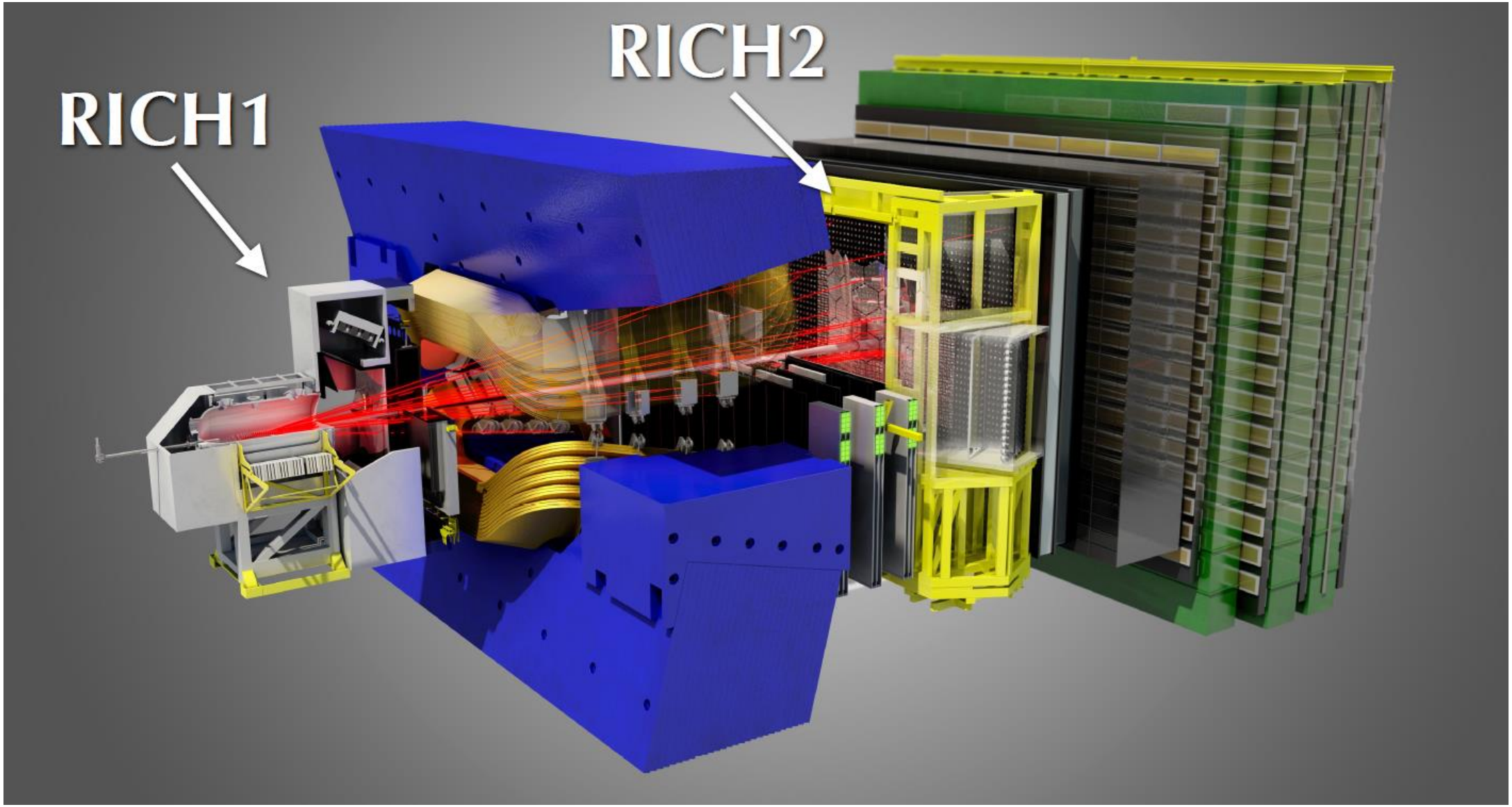
- tracing the particles through the radiators
- Cherenkov light generation
- photon propagation, reflection, refraction and scattering
- Photon Detector (photo-cathode + silicon pixel) simulation

These require significant computing resources

However, such detectors are used only for particle identification.

- It is possible to train a generative model for direct conversion of track kinematics to PID variables (just 5 numbers: RichDLLe, RichDLLμ, RichDLLk, RichDLLp, RichDLLOthers)

Cherenkov detectors at LHCb



Problem statement

Main goal is fast generation of PID parameters (RICH DLLs), given particle type and track characteristics

Train sample:

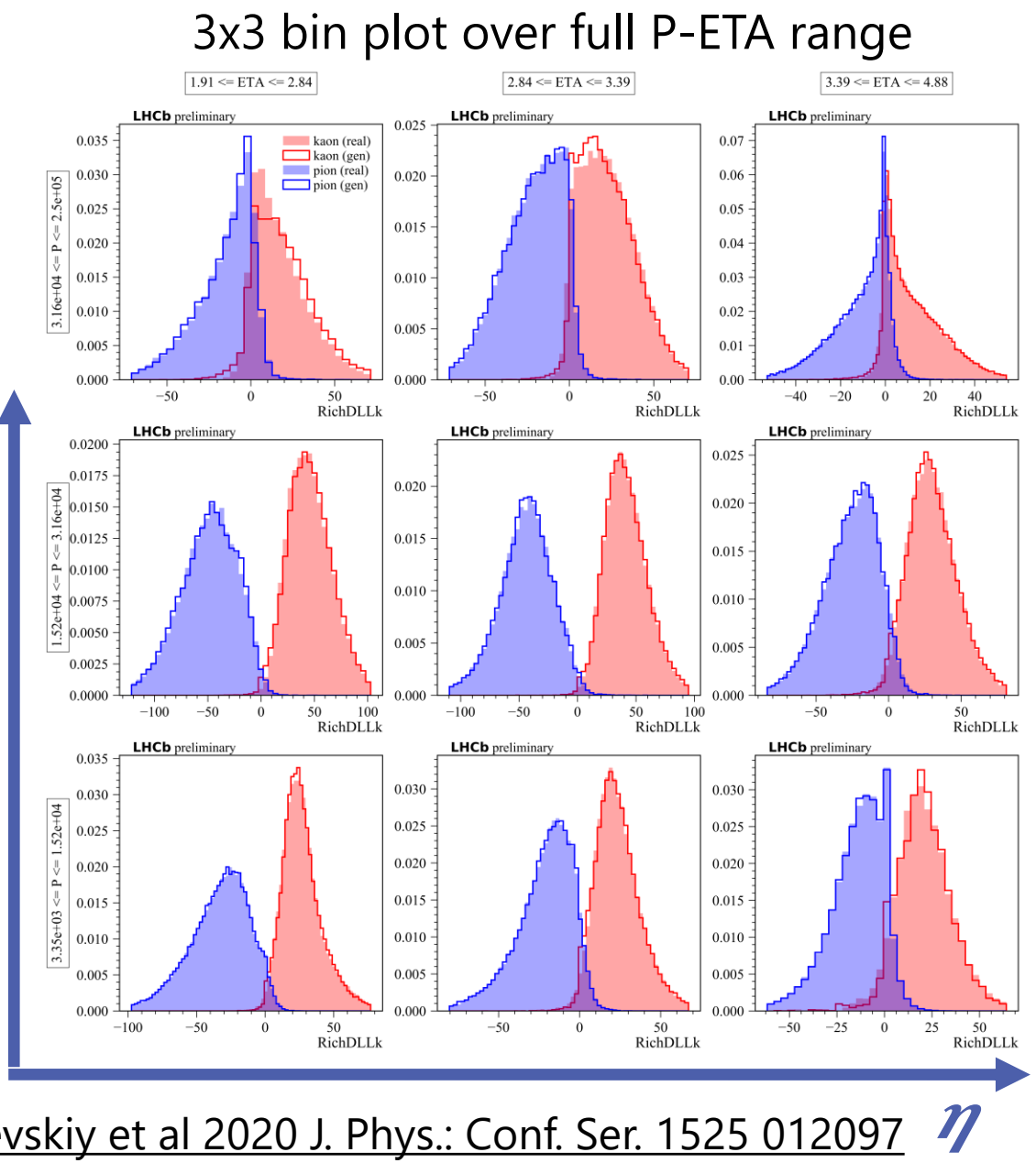
- Geant4 based simulation

Input:

- P – momentum
- η – pseudorapidity
- nSPDHits – number of hits in the Scintillating Pad Detector

Output:

- 5 output variables (RichDLLx, $x \in e, \mu, k, p$, below threshold)

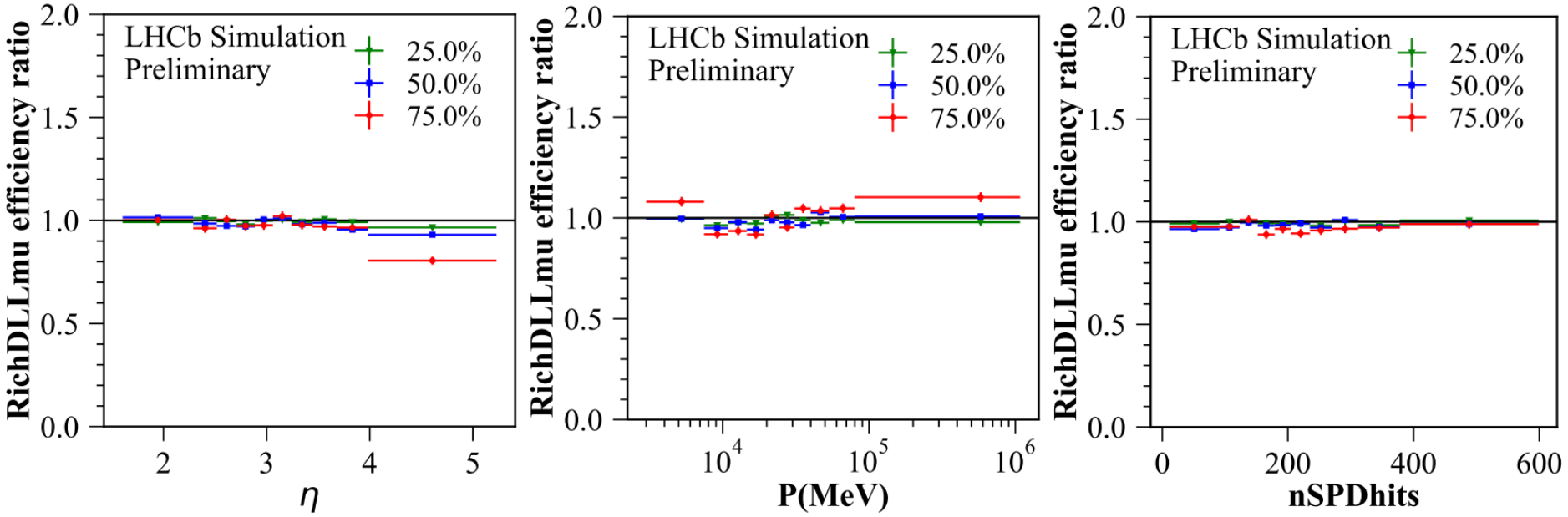


3x3 bin plot over full P-ETA range

A Maevskiy et al 2020 J. Phys.: Conf. Ser. 1525 012097

Model stability

- We want to check that model trained on data samples in limited phase space would generalize to the full phase space.
- We trained GAN on simulated data limited calibrated samples for decays: Inclusive J/ψ and $B^{\pm} \to J/\psi(\mu^+\mu^-)K^{\pm}$. The ratio of efficiencies between GAN predictions and simulated events for decay $B^{\pm} \to K^{*\pm}\mu^+\mu^-$ is presented



S Mokhnenko et al ACAT 2021 arXiv:2204.09947

On a qualitative level, the model demonstrates stability of metrics important for physical analysis.

Fine tune specific metric

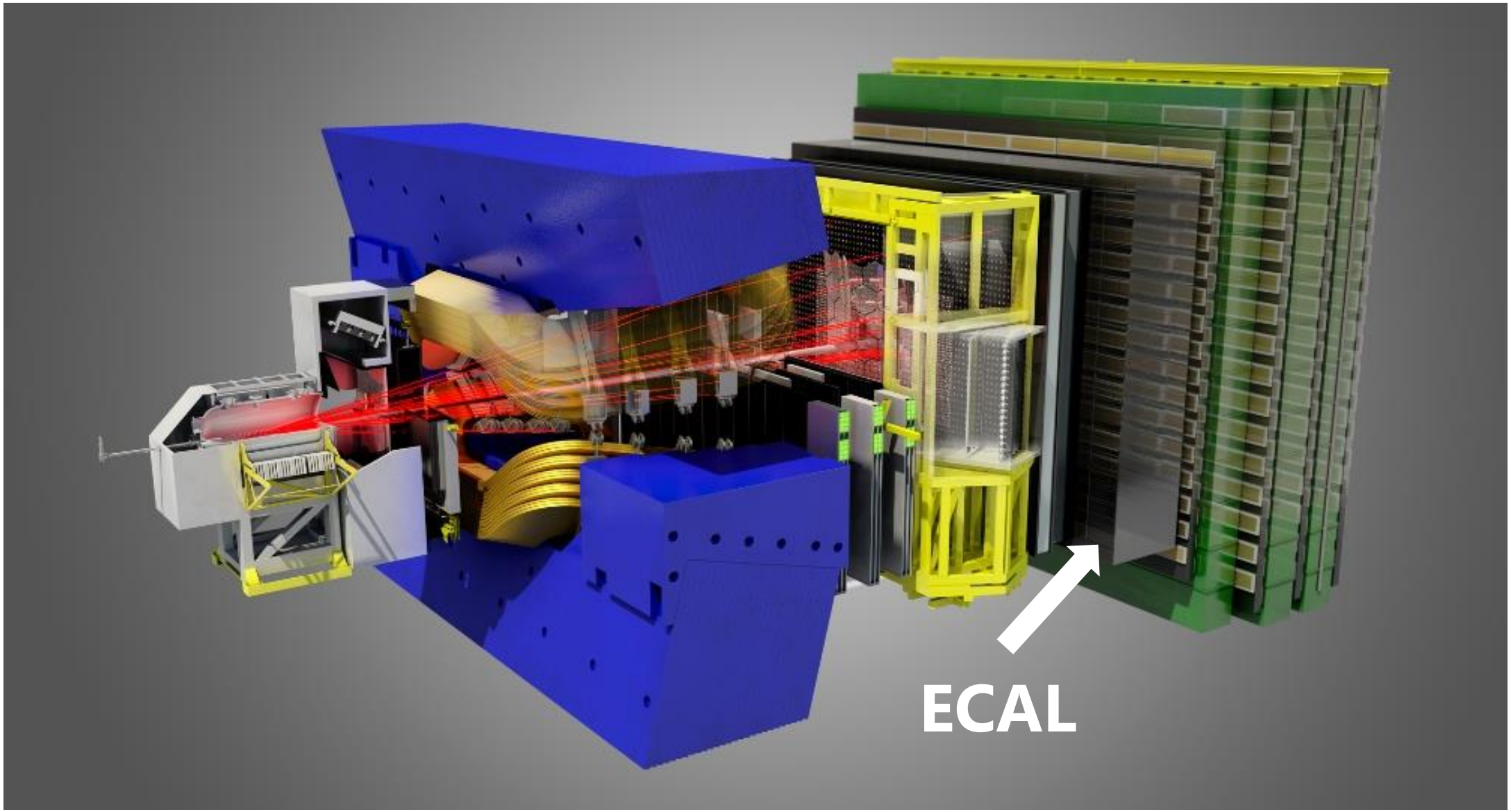
Question:

- How can we enforce generative model to learn specific physics requirements with higher priority?

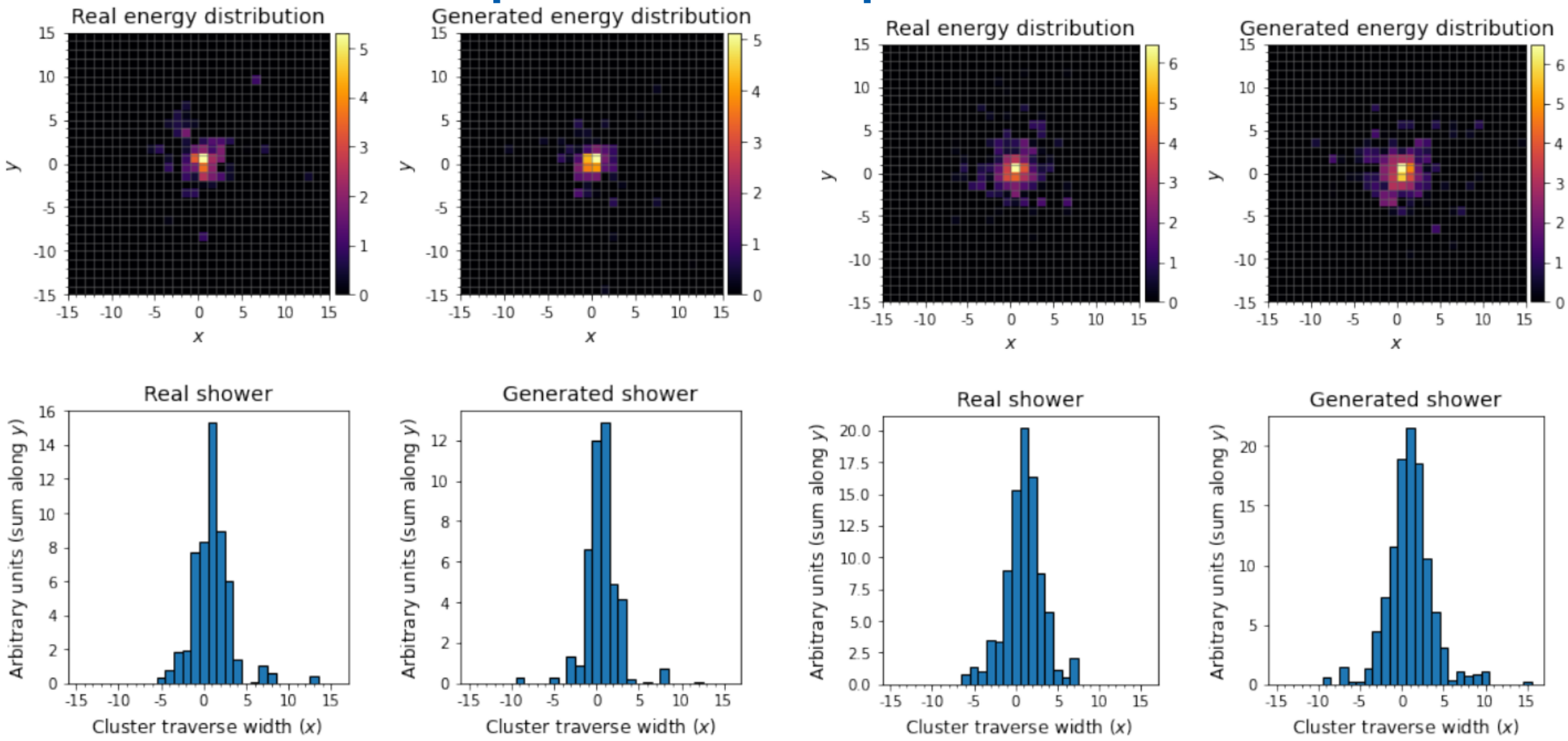
Answer:

- If the target metric is differentiable, you can include it directly in the loss function
- If the target metric is more complex and cannot be expressed as a computational graph:
  - construct an auxiliary surrogate regressor to estimate the target metric for the generated object
  - consider surrogate metric as an object feature
  - train generative model with emphasis on the target feature and the target regressor simultaneously

Electromagnetic Calorimeter at LHCb



Generated samples example

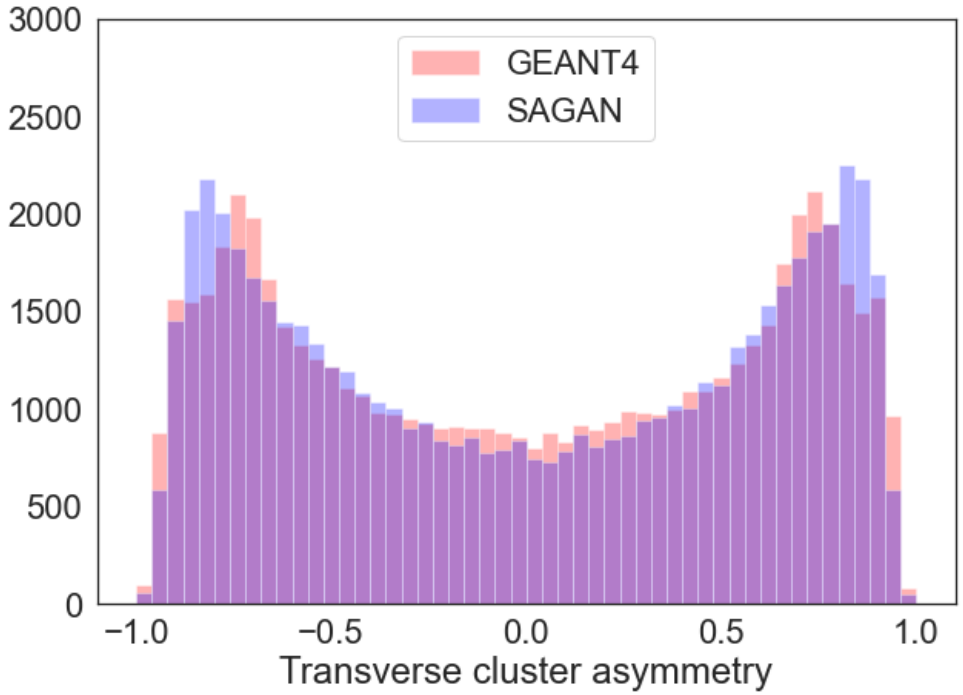


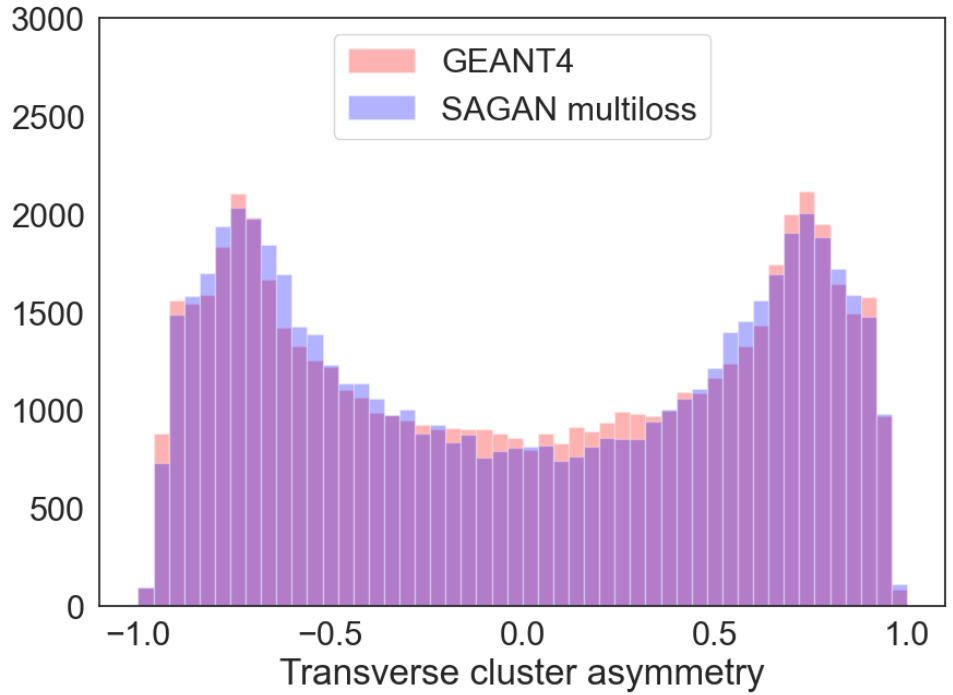
F Ratnikov and A Rogachev EPJ Web Conf., 251 (2021) 0304

An excellent description of the basic features

Asymmetry distribution

- Additional regressor that evaluates asymmetry, the distribution of asymmetry calculated over generated samples improved
- The regressor and GAN are fitted simultaneously





A Rogachev, F Ratnikov ACAT 2021 arXiv:2207.06329

- If some metric is important to us, it can be explicitly taken into account in the model

Conclusion

- Generative adversarial networks may boost simulations of elementary particle detectors by orders of magnitude compared to regular Geant4.
- Dimension of problem may be significantly reduced by considering specific structure of detector.
- High-level detector response may be generated bypassing simulating low-level signal.
- Generalizability models require special care.
- Specific metrics may be enforced by appropriate training procedures.

Backup

Problem statement

Main goal is to generate energy distribution in ECAL.

Train sample:

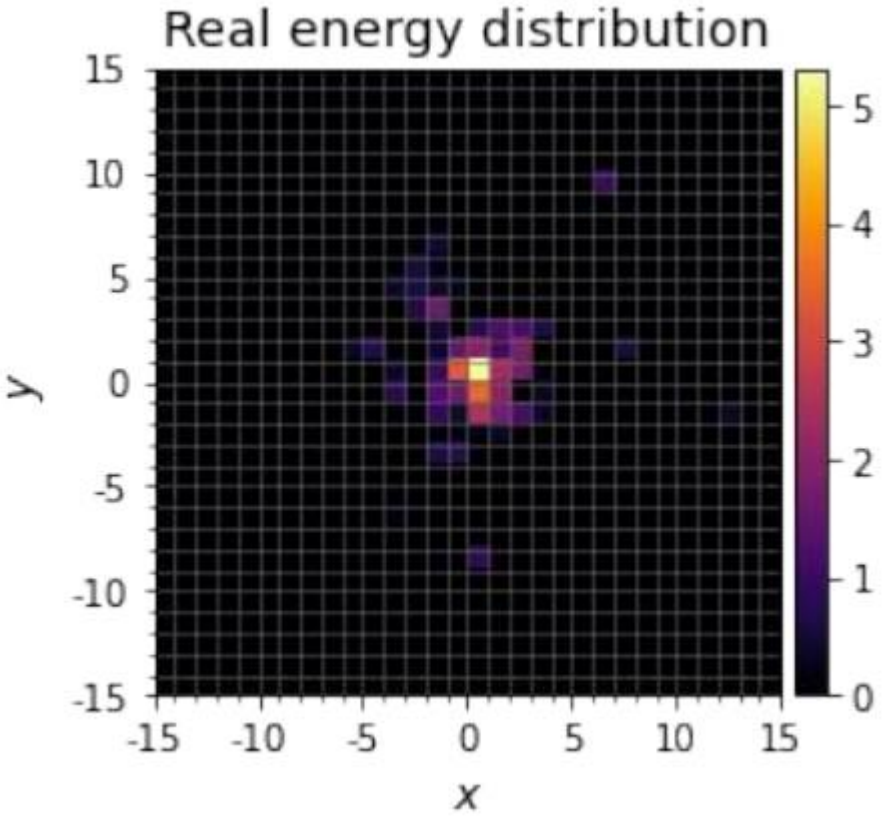
- Geant4 based simulation

Input:

- ParticlePoint (x,y,z) – known starting point location
- ParticleMomentum (p_x, p_y, p_z) –known momentum

Output:

- Consider 20 mm cell to fit both 40 mm and 60 mm cells
- EnergyDeposit – 30×30 energy distribution matrix, shower width < 600 mm



Example log output

Problem statement

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

Train sample:

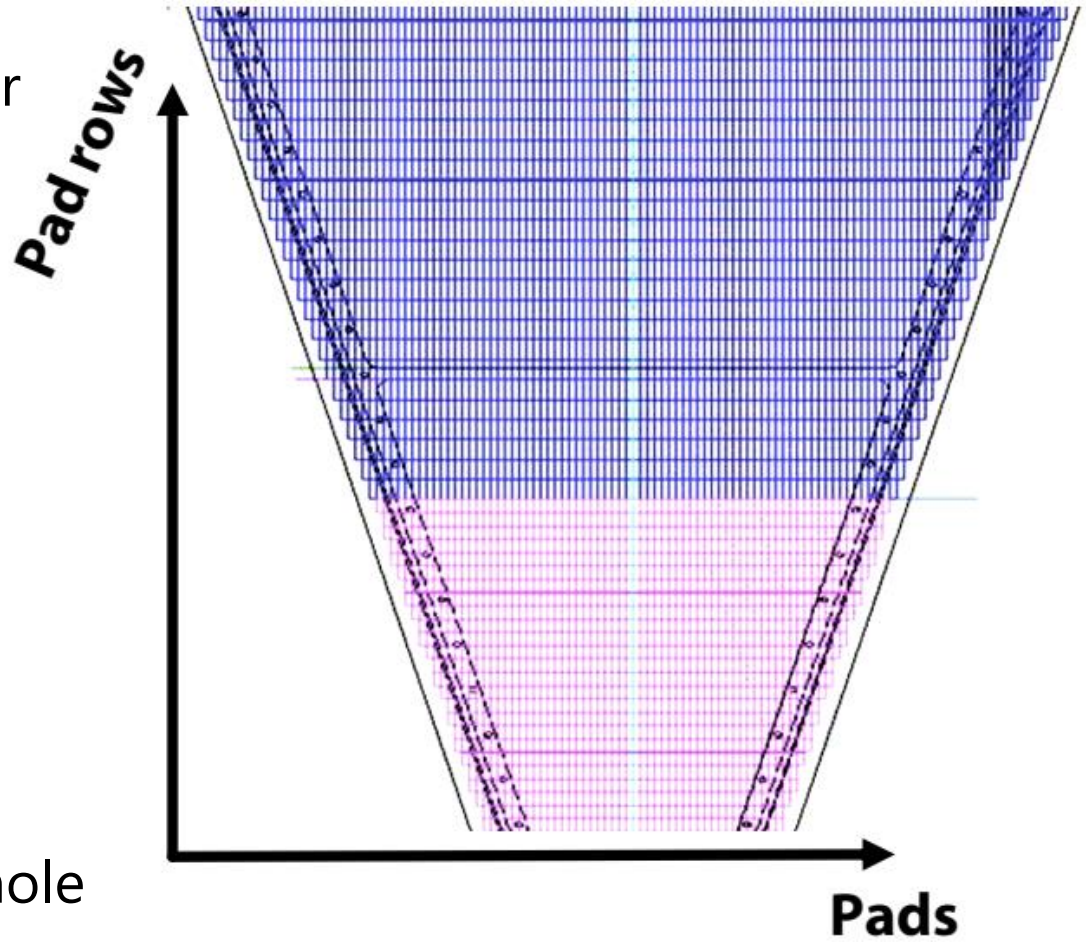
- Simulated data for pions

Input:

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

Output:

- 95 232 · 310 elements (pads x time buckets)
- Conditioned on the track parameters for the whole event



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

GAN for NICA Multi-Purpose Detector

GAN for LHCb Cherenkov detectors

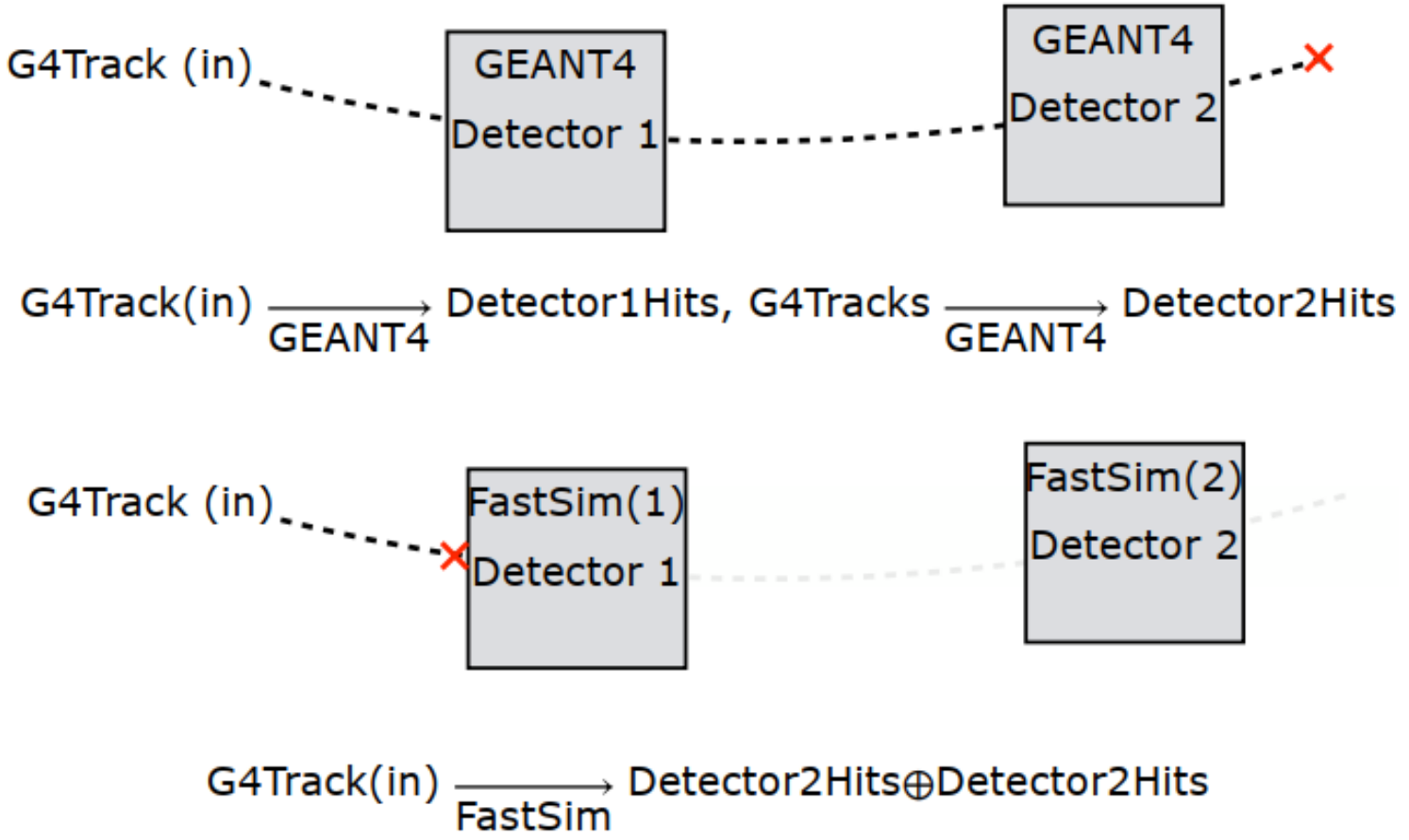
GAN for LHCb Calorimeter

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

Operation Scheme

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



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):

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