CZECH TECHNICAL UNIVERSITY IN PRAGUE

CERN openlab

CERN

Ensemble Model for Detector Simulations

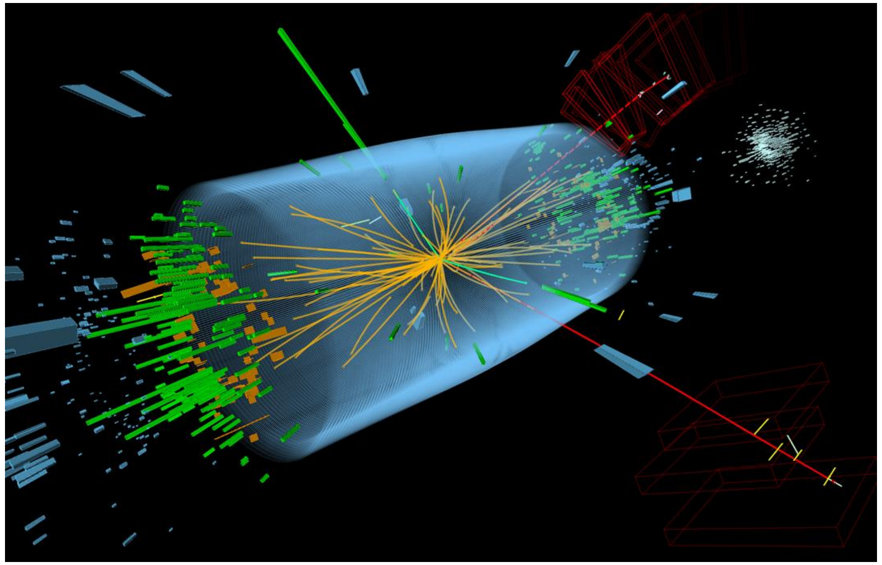
Stochastic and Physical Monitoring Systems

Rumburk 2022

Kristina Jaruskova (CERN, Czech Technical University)

June 2022

Detector Simulations



- Detector simulations - Monte Carlo-based tools (**GEANT4**)
  - Representation of the theory
  - Algorithm tuning
  - Standard tool – Monte-Carlo based algorithms
    - Currently: Geant4 simulation package

- **WLCG (Worldwide LHC Computing Grid)**
  - Global infrastructure of computing resources (data storage, analysis, …)
  - 100 centers in more than 40 countries

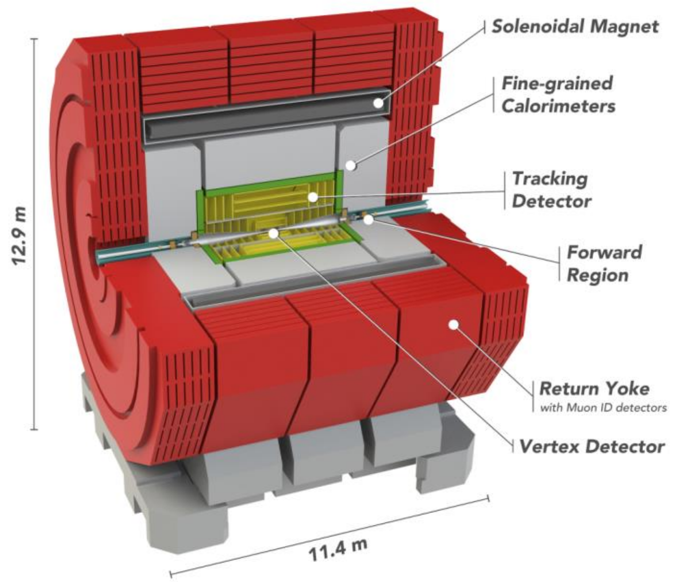
- MC tools are computationally intensive
  - **50 %** of the WLCG resources used for simulations[1]

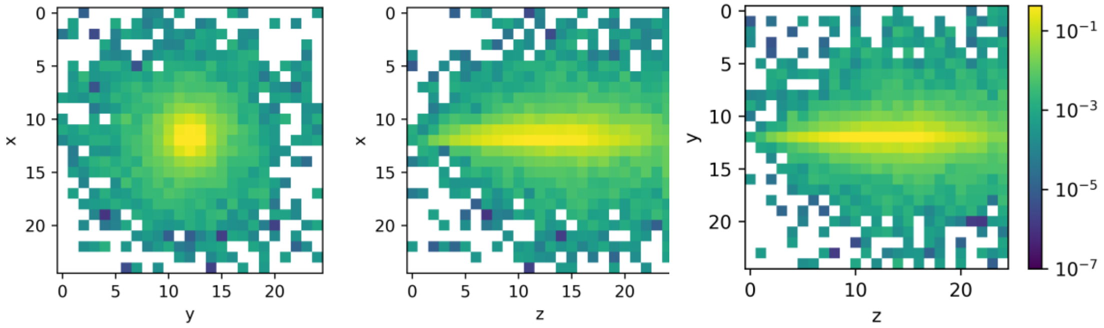
1. Zaborowska A. Geant4 fast and full simulation for Future Circular Collider studies. 2017

Detector Simulations

- High-Luminosity LHC (2027) - upgrade
  - Approx. **150x** more data[1] → MC simulations will be intractable

- Calorimeter simulations
  - Energy depositions of entering particles
  - Most time demanding step in simulation process
    - ATLAS – **70 %** simulation time spent on calorimeters[2]
  - Crucial to find faster alternatives → **generative deep learning**
    - Generative adversarial networks (GANs)
    - Variational autoencoders
  - Preliminary results – up to 10^6x speedup[1]
    - Not accurate enough yet
  - CERN openlab – models for CLIC ECAL simulations
    - Compact Linear Collider (CLIC)
    - Image size: 25x25x25 cells → interpreted as a 3D grayscale image

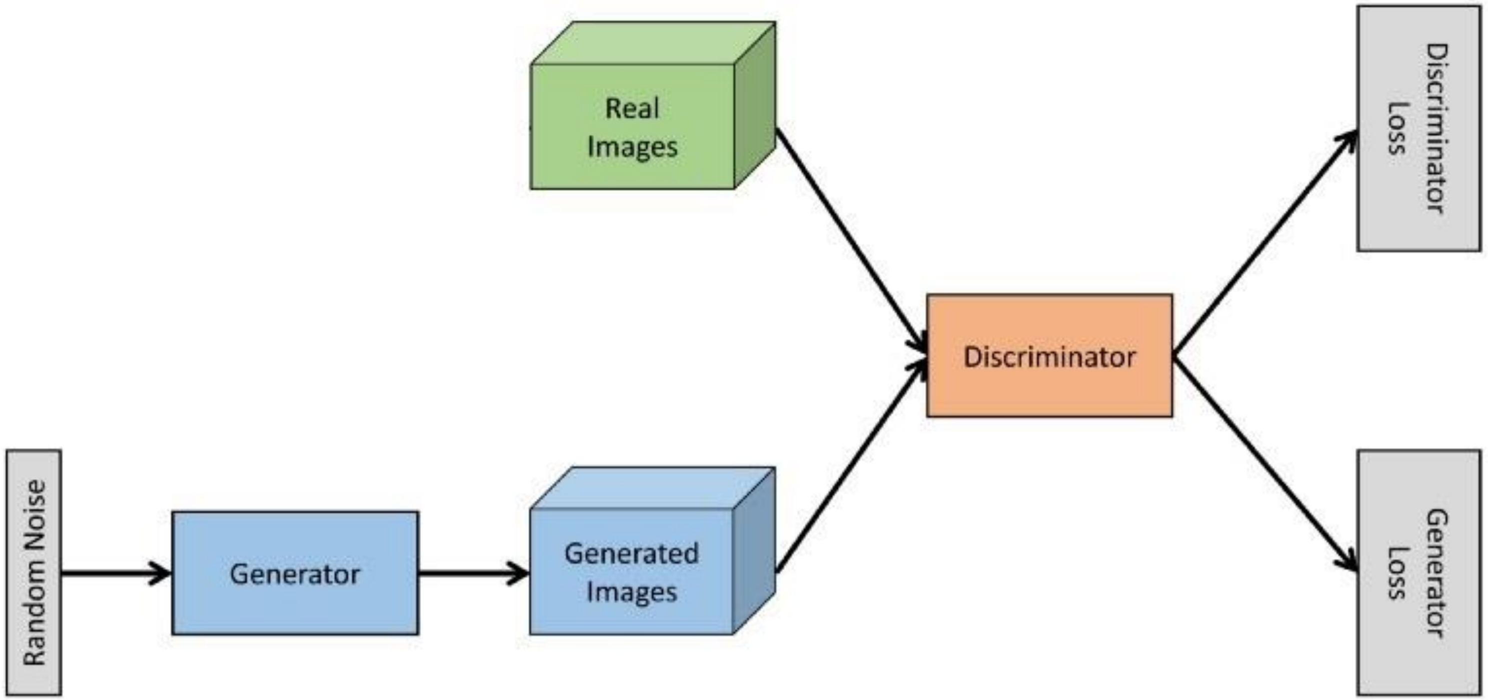




1. Albrecht J. et al. A Roadmap for HEP Software and Computing R&D for the 2020s. 2019
2. Zaborowska A. Geant4 fast and full simulation for Future Circular Collider studies. 2017

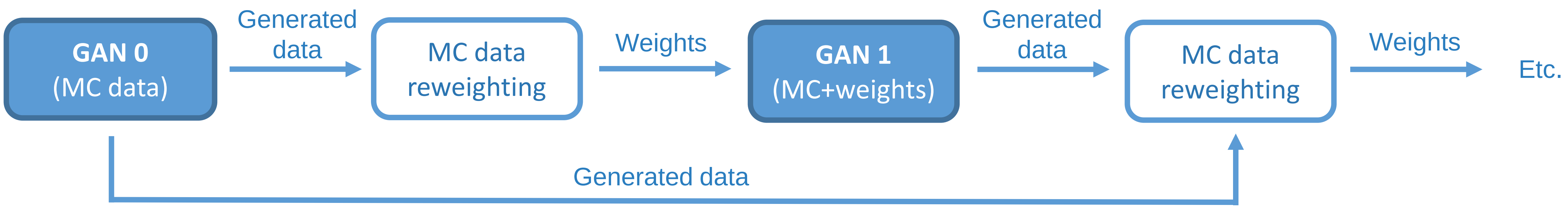
GAN

Generative Adversarial Network



GAN ensemble

- Ensemble techniques in ML/DL
  - Combining predictions from multiple models – reduce variance and generalization error
  - Improvement in different tasks – classification, regression
- **AdaGAN**[3]
  - Ensemble of multiple GANs (generator-discriminator pairs) trained sequentially
  - MC training data – reweighted before training the next GAN
  - Addressing the issue of **missing modes – next GAN focuses on week spots of the previous ones**
    - Poorly reproduced training samples – high weight
    - Well reproduced training samples – low weight

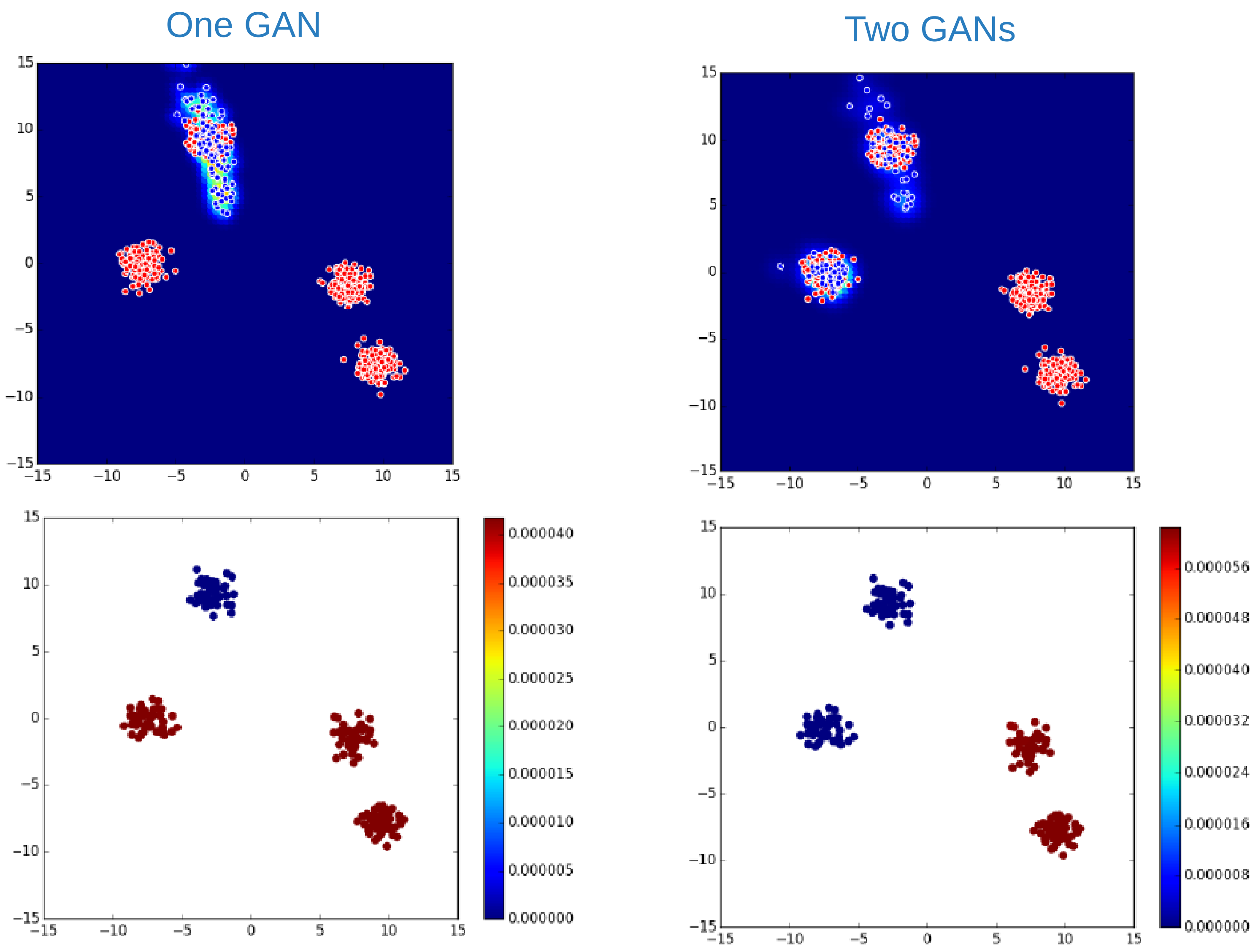


3. I. Tolstikhin et al. AdaGAN: Boosting generative models (https://arxiv.org/abs/1701.02386)

- **Toy example: mixture of Gauss clusters**

(red) training data
(blue) generated samples

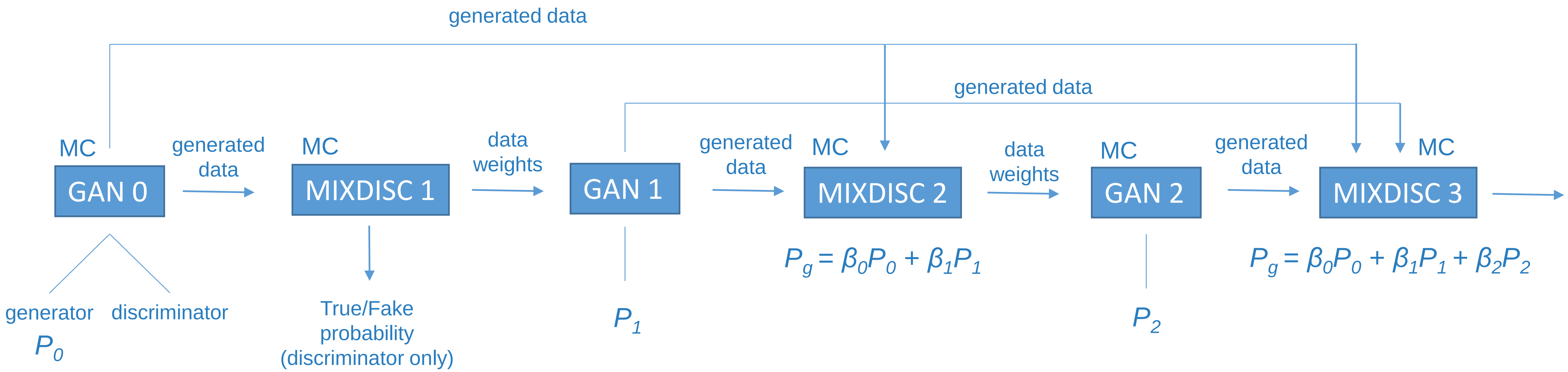
Weights assigned to training data



Ensemble training details

- MC = Monte Carlo training data
- GAN 0: trained on MC training data
- MIXDISC: trained on MC data and data from GAN 0 (1 : 1 ratio)
- GAN 1: trained on weighted MC data (weights from MIXDISC 1)
- MIXDISC 2: trained on MC data (no weights) and data from GAN 0 + GAN 1 (MC : generated = 1 : 1)
- GAN 2: trained on weighted MC data (weights from MIXDISC 2)

etc.



Data weights

- Main idea: minimize Jensen-Shannon divergence between data distribution P_d and the ensemble distribution P_g with the next GAN distribution Q

$$\min_{Q \in \mathbb{P}} D_{JS}((1-\beta)P_g + \beta Q \parallel P_d)$$

- In practice: any improvement on the J-S div. is enough
- Formula:

$$w_i = \frac{p_i}{\beta}\left(\lambda^* - (1-\beta)\frac{1 - D(X_i)}{D(X_i)}\right)_+$$

$p_i = 1/N$ … empirical distribution of the training data
λ^* … normalization factor

- If $D(X_i) \sim 1$ → MIXDISC is certain it is training sample → high weight
- If $D(X_i) \sim 0{,}5$ → MIXDISC is confused → well represented in generated dataset → low weight

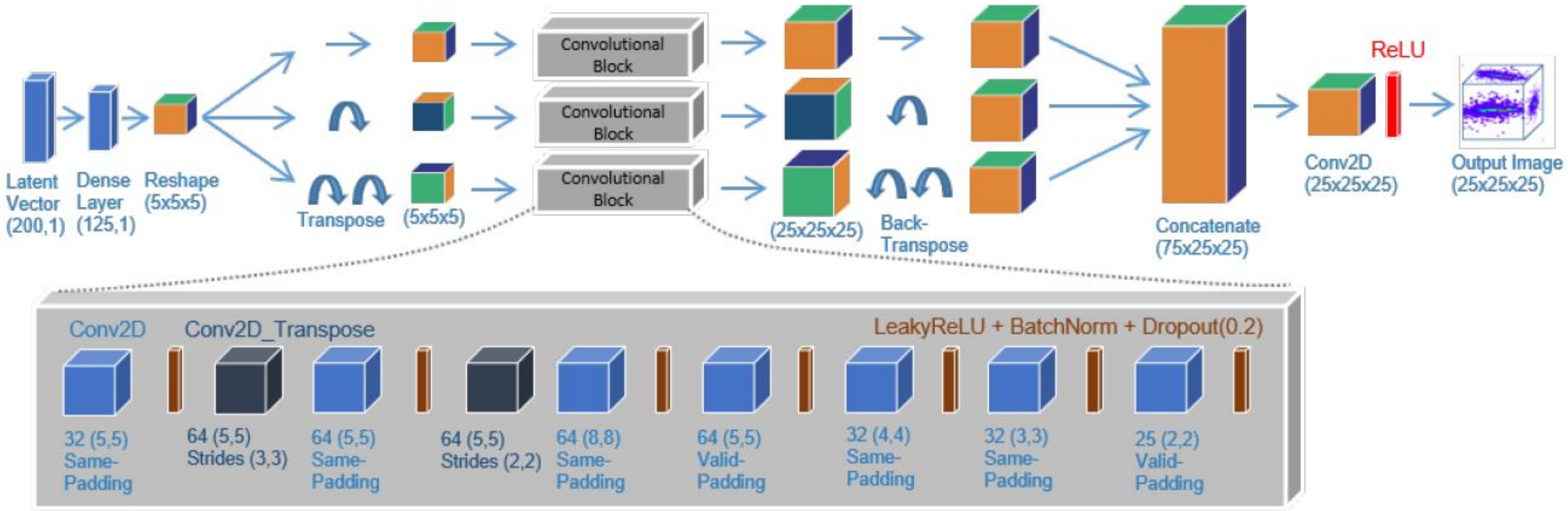
CERN openlab

Conv2D GAN as an ensemble

- Conv2D GAN[4] as a baseline model
  - Uses 2D convolutions only
  - Trained on 200 000 MC samples
- T = 10 GANs were trained with equal component weights β_j creating a mixture of distributions

$$P_g = \sum_{j=0}^{T-1} \frac{1}{T} P_j$$

- Conv2D generator architecture:



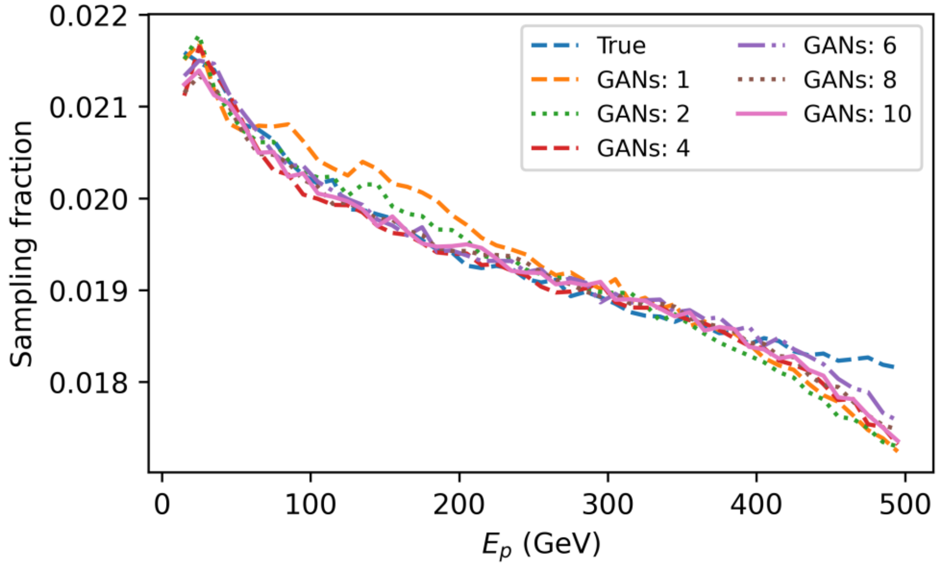
4. Rehm F. Physics Validation of Novel Convolutional 2D Architectures for Speeding Up High Energy Physics Simulations. 2021

Sampling fraction

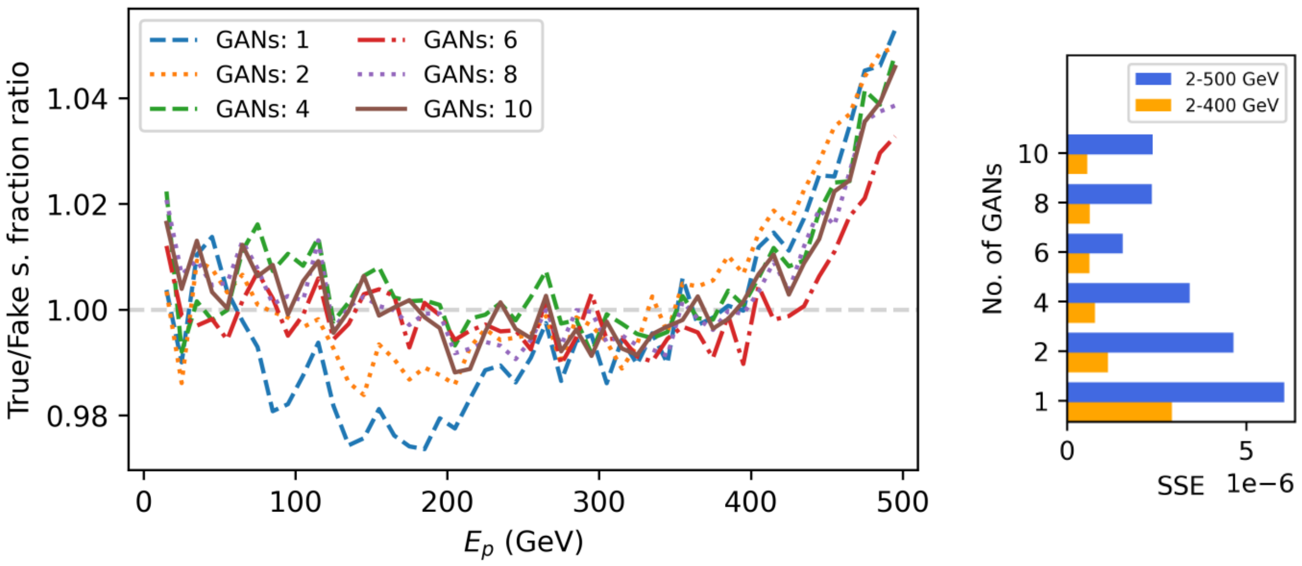
- Visible improvement in sampling fraction – ratio of the total deposited energy to the primary energy of a particle

Sampling fraction comparison

- Data split up into energy bins of 10 GeV



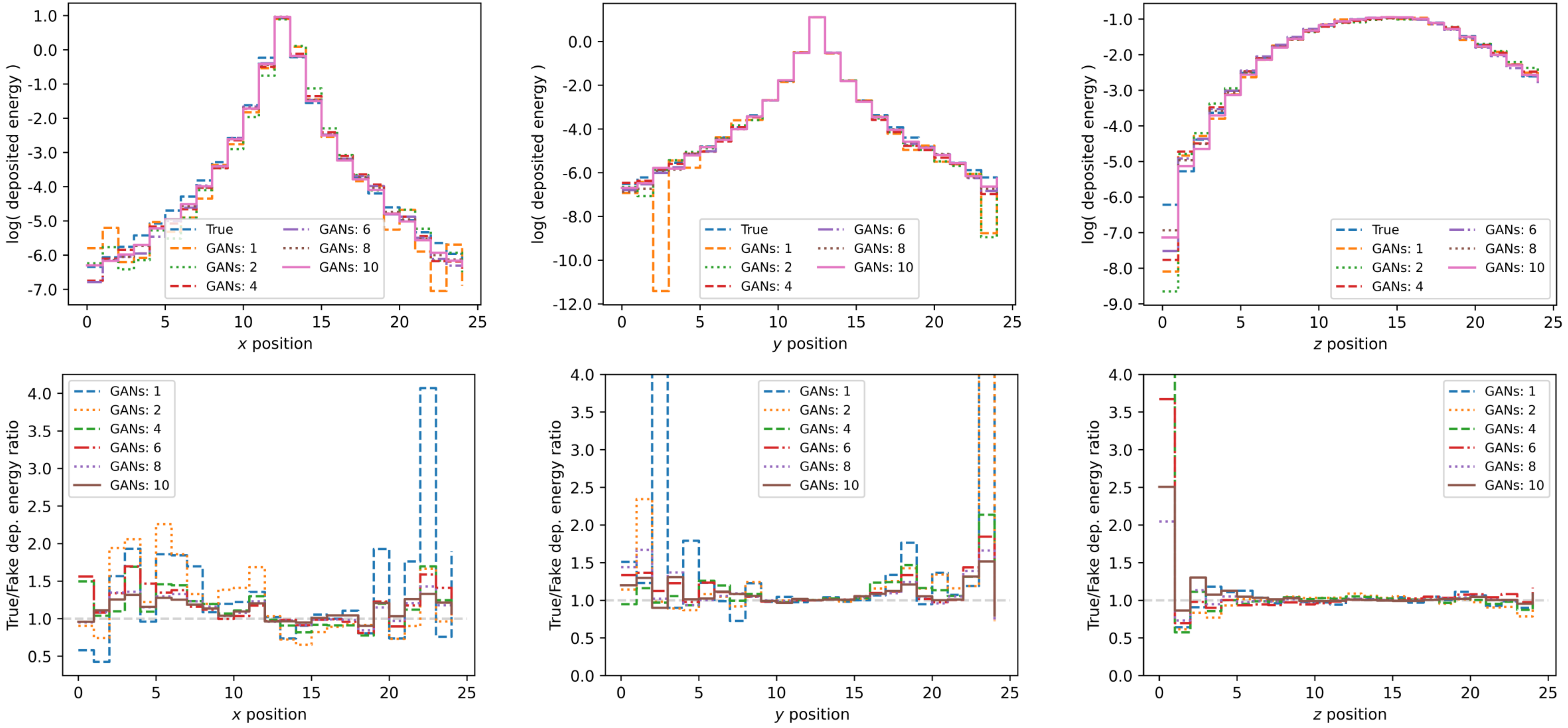
Ratio of the true samp. fraction to the samp. fraction of the generated data



CERN openlab

Shower shapes

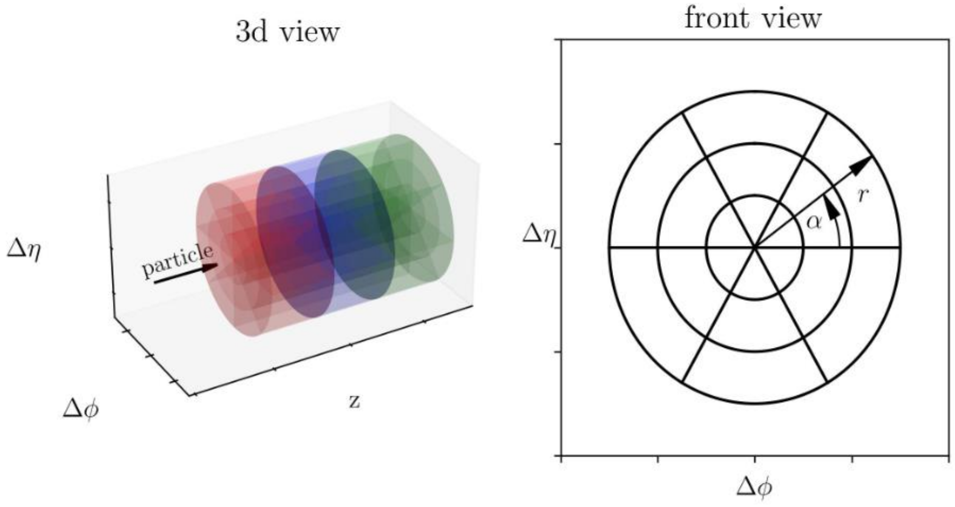
- Improvement in average shower shapes around the edges



Thank you.

CaloChallenge 2022:

- Publicly open challenge from the Geant4 group
- https://calochallenge.github.io/homepage/



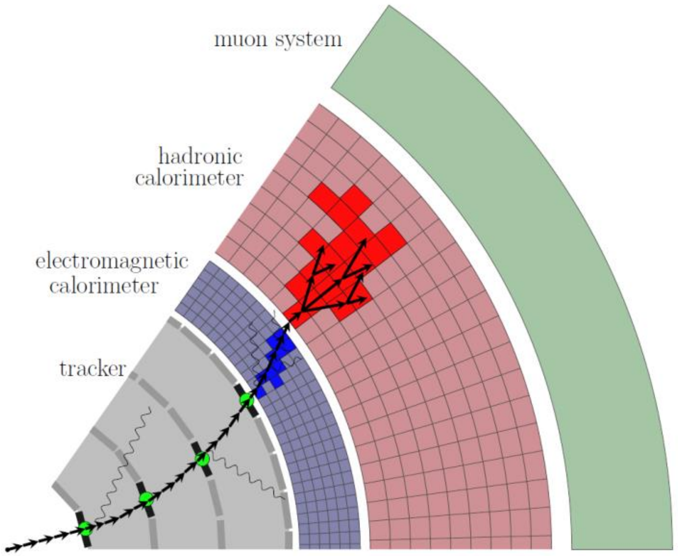
Backup slides.

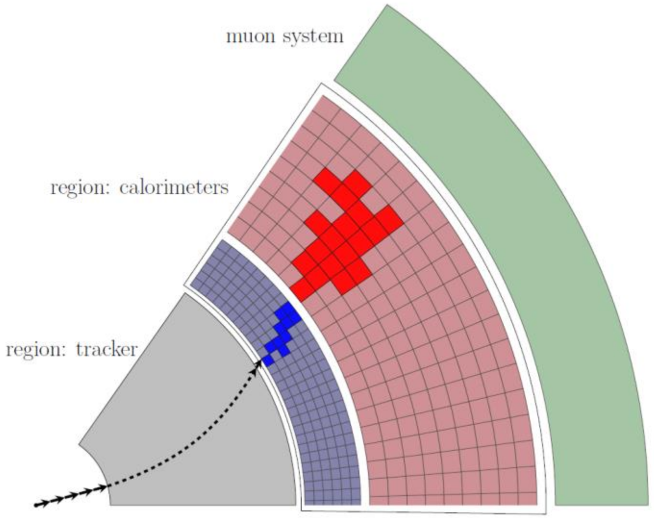
Detector Simulations

- Currently in Geant4:

Full simulation
- Step-by-step
- Many variables (velocity, momentum, direction, angle etc.)
- Time consuming

Fast simulation
- Only the overall response
- Parametrizing, pre-simulated showers
- Approx. 10x - 1000x speedup

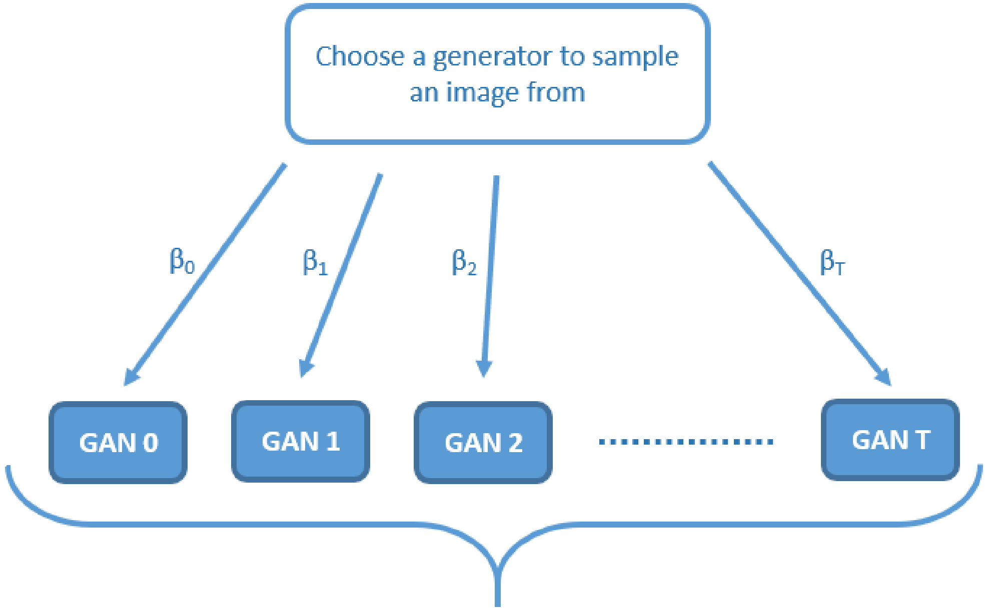




Images: Zaborowska A. Geant4 fast and full simulation for Future Circular Collider studies. 2017

Generating samples from the ensemble

- Component weights: $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_T)$
- Generator distributions: $P_0, P_1, P_2, \ldots, P_T$
- Final distribution P_g: linear mixture of GAN distributions



$$P_g = \beta_0 P_0 + \beta_1 P_1 + \ldots + \beta_T P_T$$