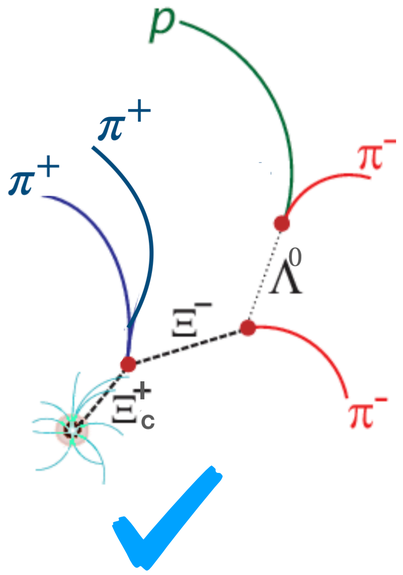
Towards Cost-Effective HEP Simulations Using GAN-Based Data Augmentation

Anisa Khatun on behalf of ALICE Collaboration
HUN-REN Wigner Research Centre for Physics

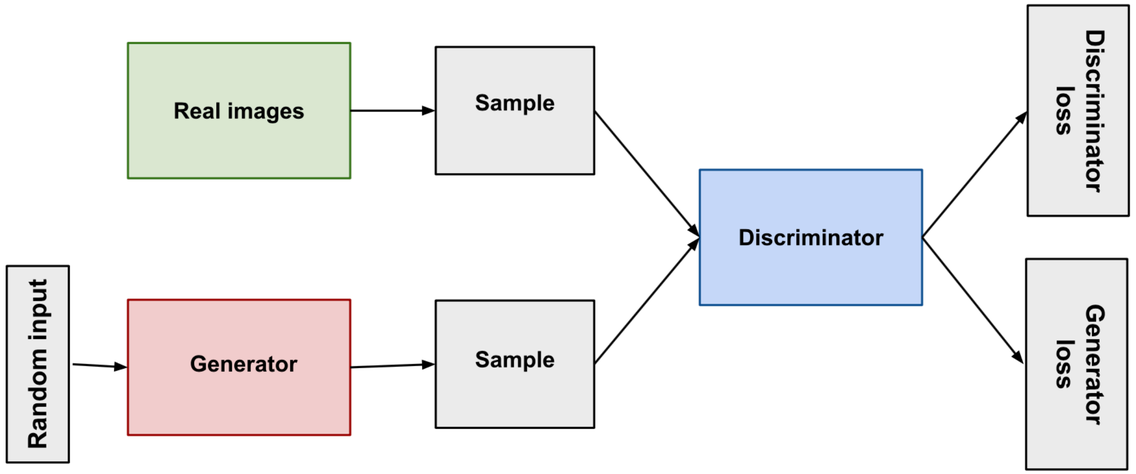
GPU Day 2026

28 - 05 - 2026



Introduction to Generative Adversarial Networks (GANs)

- A class of machine learning frameworks (wikipedia).
- Initially developed by Ian Goodfellow et al. arXiv:1406.2661 [stat.ML].



- Comprise two parts :
  - The **generator** learns to generate plausible data.
  - The **discriminator** learns to distinguish the generator's fake data from real data.

https://developers.google.com/machine-learning/gan/gan_structure

Example GANs uses in HEP :

1. Detector fast simulation

- **CaloGAN**: EM shower generation, orders-of-magnitude faster than Geant4.

2. Tracking / Hit generation

- GANs as surrogate models for silicon detector hits and pattern recognition.
- Used to generate distorted hit patterns & emulate hit correlations.

3. Event-level / Kinematic feature generation

- GANs for top-quark, Higgs, and dijet kinematics.
- GAN-assisted event reweighting and generative unfolding approaches.

4. ALICE-specific work

- TPC cluster GANs.
- PID response generation.
- Generative calorimeter response

5. ML Training & Data Augmentation

- Using GANs to produce synthetic signal samples for ML classifiers.
- Helps balance datasets and reduce statistical fluctuations.

MC bottleneck and reconstruction-level gap

Rare-signal analyses require large MC statistics

- Complex topologies → low signal efficiency
- MC generation is computationally expensive

Existing ML fast simulation focuses on:

- Detector response
- Calorimeter simulation
- Hit generation

Missing workflow layer

☑ **Reconstruction-level MC augmentation for rare signals**

☑ **Target: Address the analysis-workflow bottleneck for rare signals**

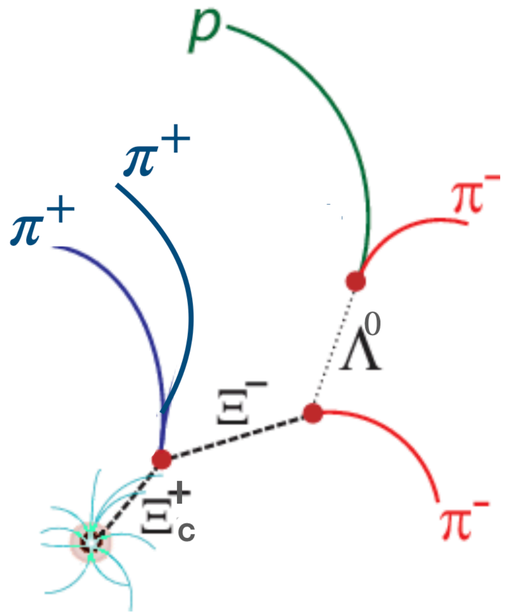
Huge collision sample

↓

Tiny rare-signal yield

↓

Need large MC statistics



Benchmark

Our approach and workflow

Key points

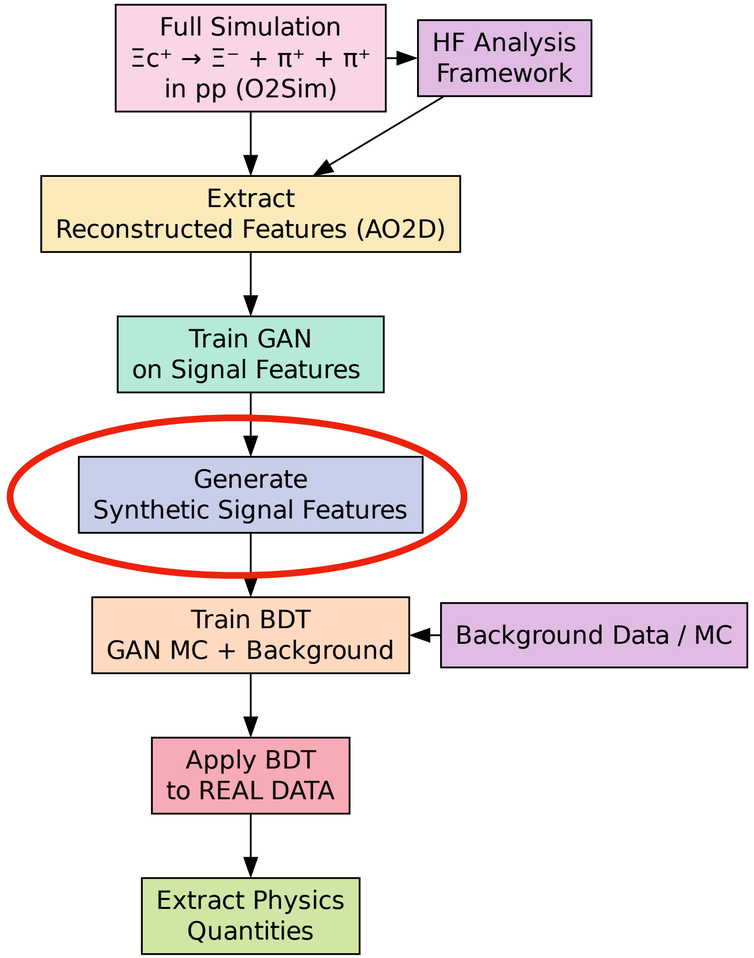
- Does not replace detector simulation
- Not a classifier — acts as a generative surrogate
- Generative augmentation between simulation and analysis

GAN-based reconstruction-level augmentation

- Learns reconstructed feature distributions from MC
- Generates statistically consistent synthetic samples
- Uses generated samples in downstream workflows

☑ Acts as a surrogate for reconstructed-level MC

☑ Increases effective MC statistics without full simulation



Poster

arXiv:2602.12088 [hep-ex]

GAN setup

Data setup:

- 8 Ξ_c^+ decay-topology features from ALICE MC

Model architecture:

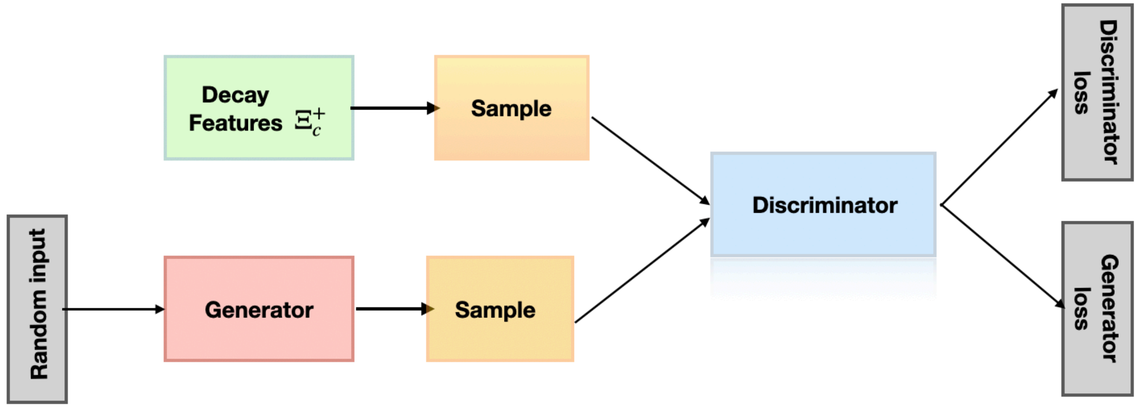
- **Generator**: Noise → fully connected neural network (3 dense layers)
- **Discriminator**: 2 dense layers of neural net

Training Strategy:

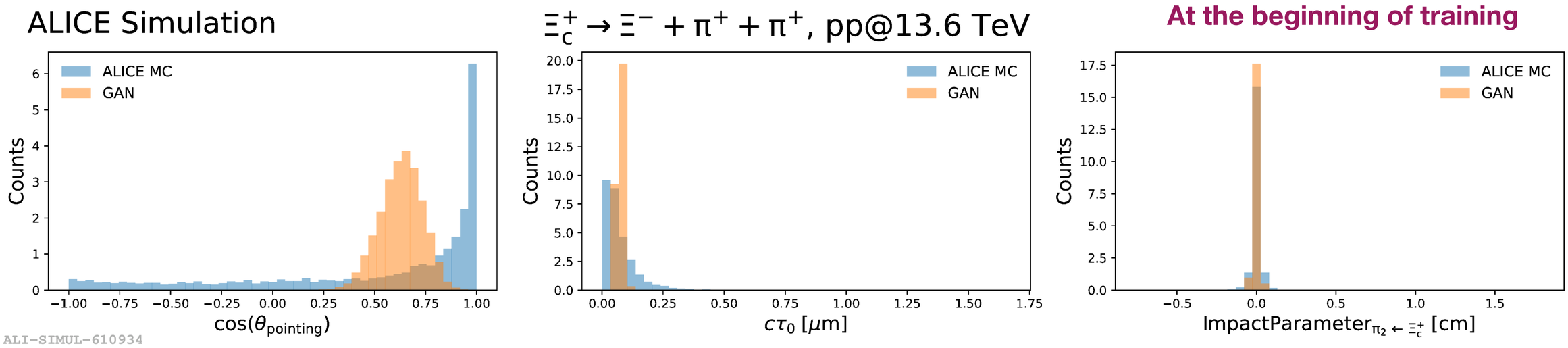
Changed features with asymmetric distributions: pre-processing

- $cos\theta \rightarrow \theta$
- Decaylength -> log(1+ Decaylength) (log1p)

- Alternating training:
  - **Discriminator** learns to distinguish real from fake samples
  - **Generator** learns to fool the discriminator and match distributions (KS penalty)
- Label smoothing, batch normalization ReLU/LeakyReLU
- Additional constraints KS-based loss, range penalty



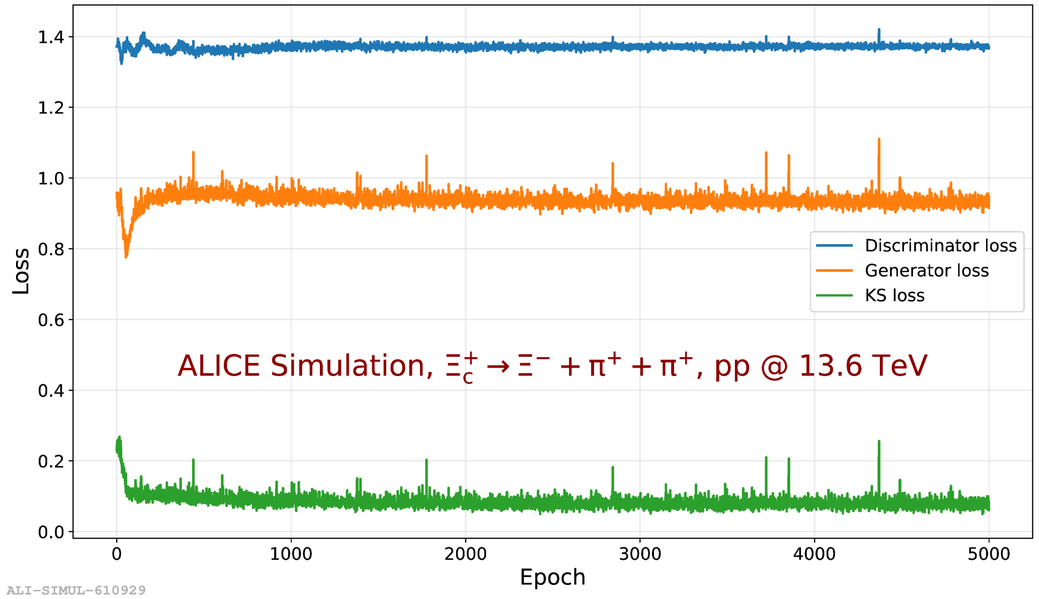
GAN training and validation



After 2k epoch

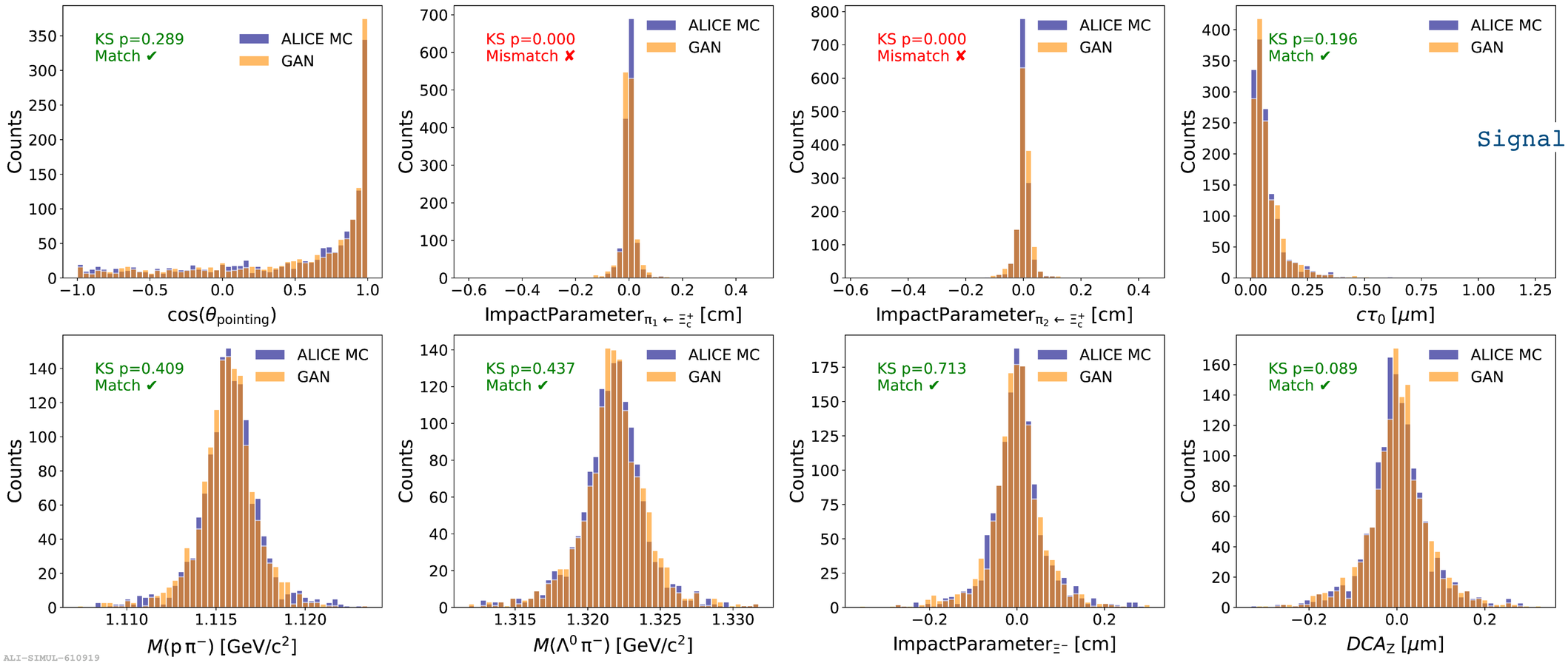
Validation strategy:

- Comparison of real vs. generated feature distributions
- KS test: statistical check of distribution agreement
- Feature correlations
- BDT response validation
- Full analysis chain



Feature-level validation

ALICE Simulation, $\Xi_c^+ \to \Xi^- + \pi^+ + \pi^+$, pp @ 13.6 TeV



ALI-SIMUL-610919

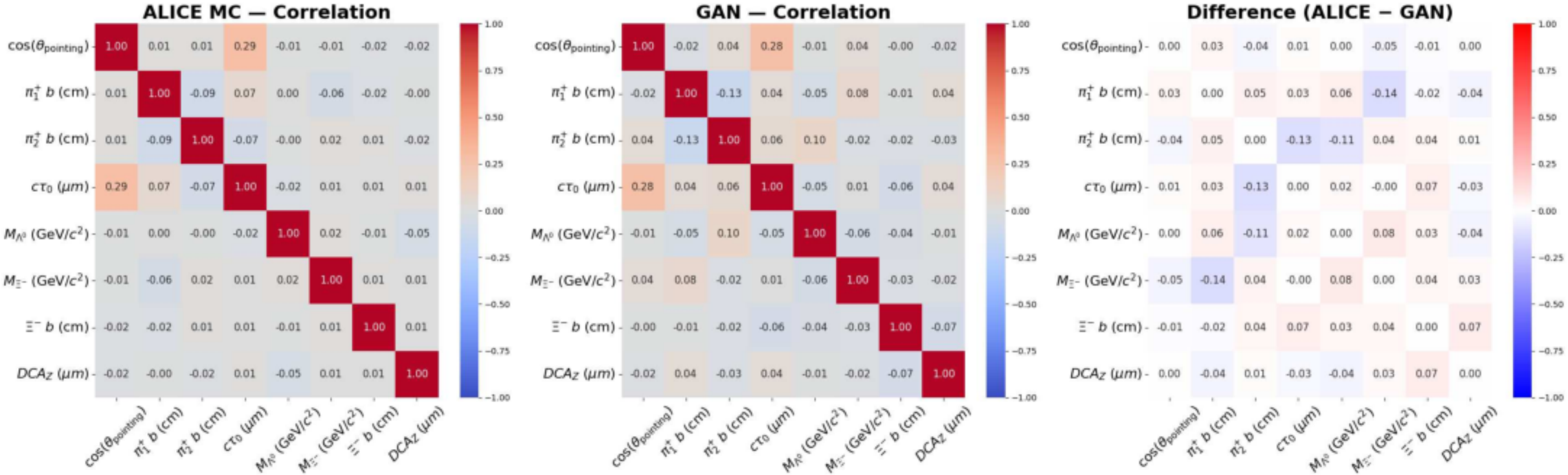
Signal count*: 531548

*Signal count is same for ALICE MC and GAN MC

***KS *p*-value per feature: derived from the statistic D = max |CDF_real – CDF_gen|, the maximum distance between the two CDFs.**

Good agreement observed between GAN-generated and MC feature distributions

Feature correlation matrices

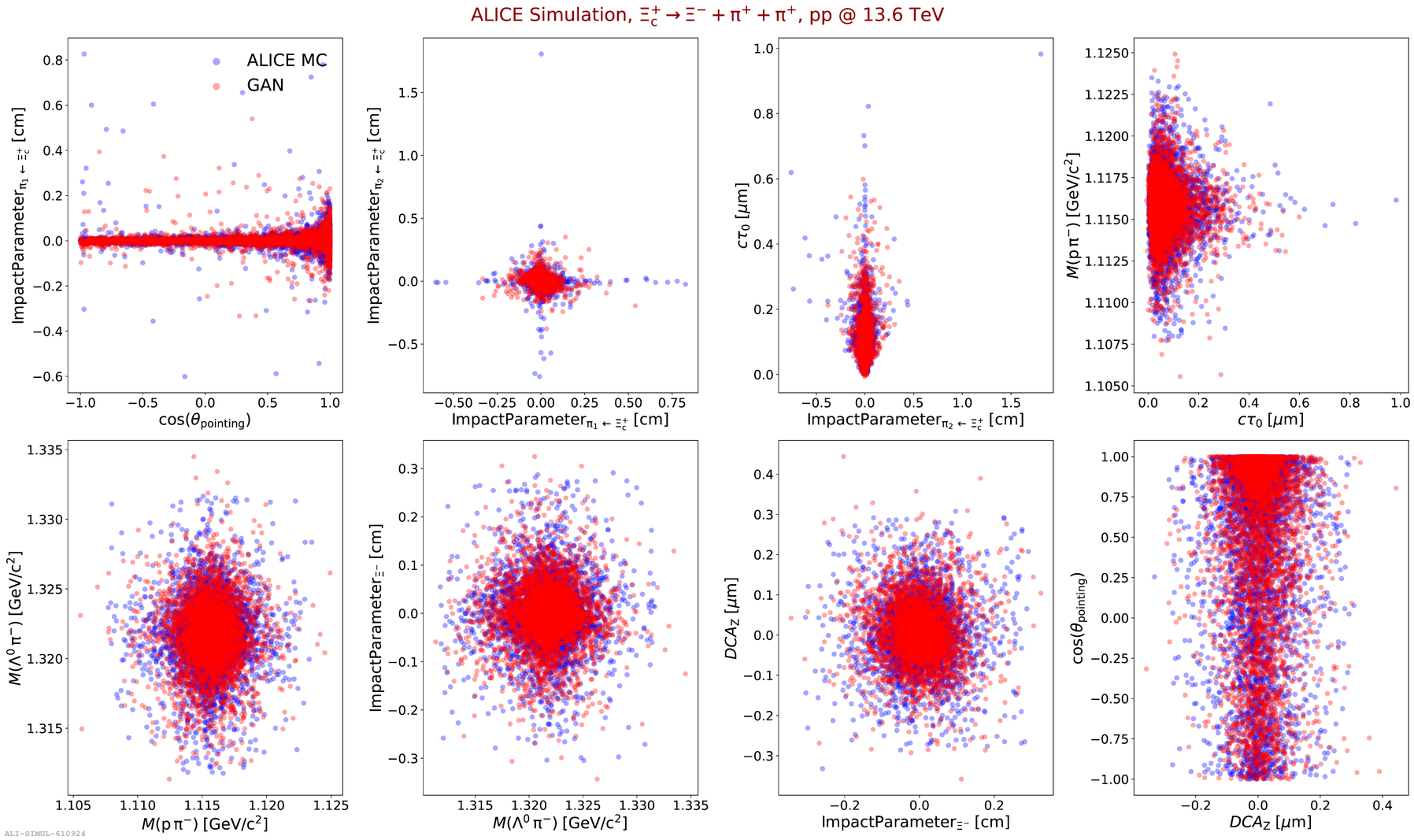


ALI-SIMUL-632300

- Pearson coefficients computed for all feature pairs
- Residual heatmaps are shown
- Correlation coefficients agree within ±0.05

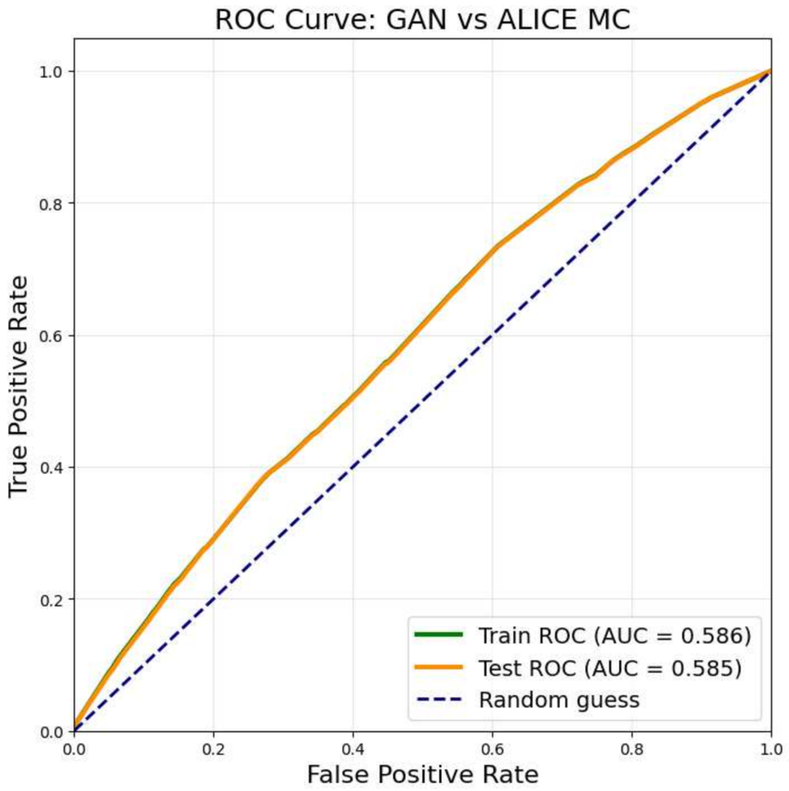
Correlation structure between GAN-generated and MC samples is largely preserved

Pairwise feature correlation scatter plots



GAN successfully reproduces pairwise feature correlations present in MC

ML-based validation with BDT

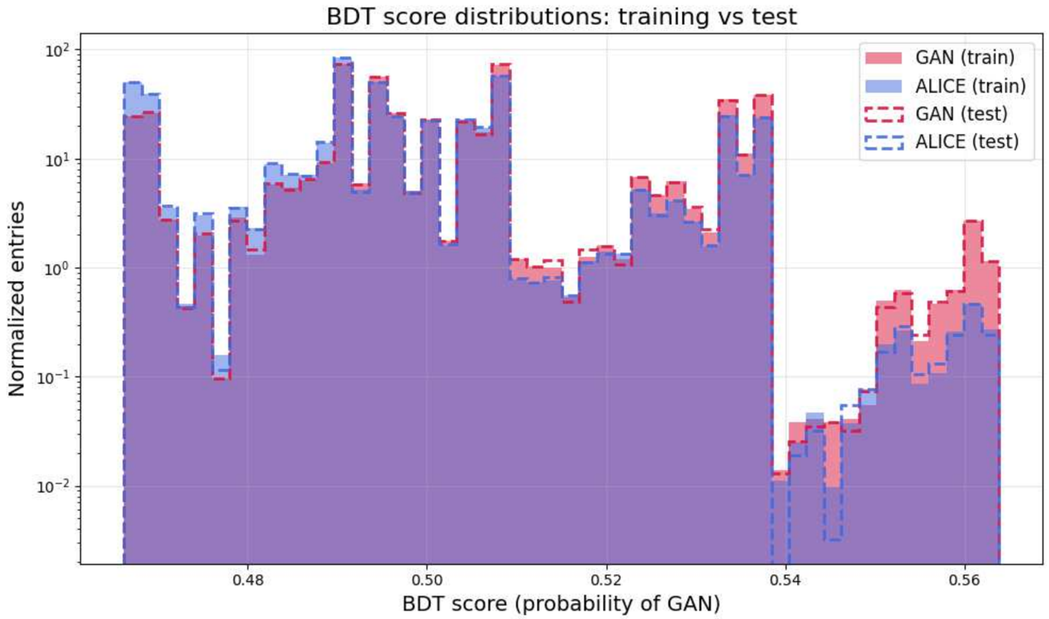


ALI-SIMUL-632310

BDT can tell how close GAN sample is to the ALICE MC

Signal: ALICE MC
Background: GAN Sample

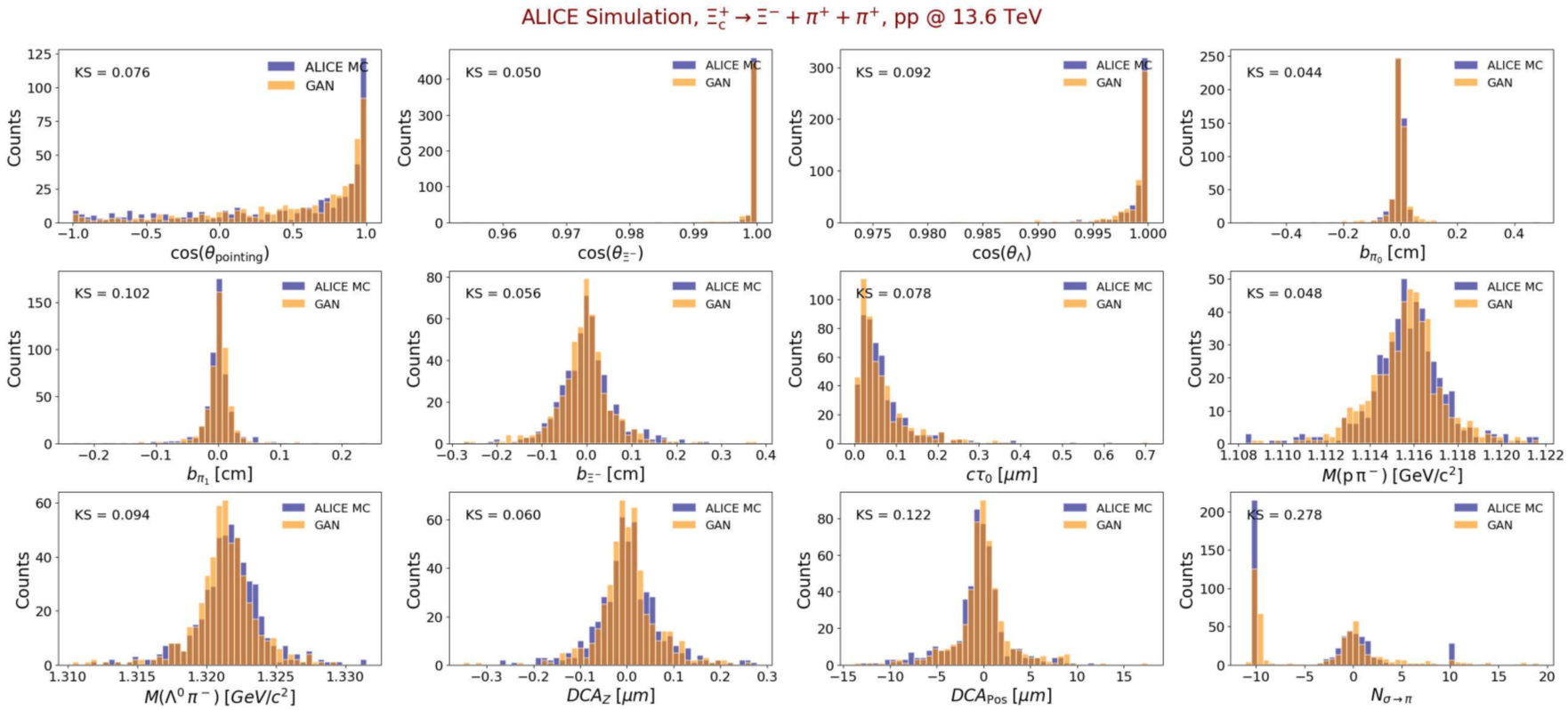
Both GAN and ALICE distributions are overlapping and peak roughly at 0.5.



ALI-SIMUL-632315

- Train BDT (binary classifier) on GAN and ALICE MC
- Comparable ROC performance
- AUC differs from 0.5 by ~ 8 - 9%, indicating GAN reproduces the ALICE MC distributions at > 90% level
- GAN and MC score distributions show strong overlap

Feature-space extension



ALI-SIMUL-632305

- KS distances summarized using mean and maximum statistics across features
- Mean KS = 0.072
- Max KS = 0.220

☑ Demonstrates scalability toward higher-dimensional feature spaces

☑ Minimal architectural changes required when extending to higher feature dimensionality

Typical challenge

Increasing feature dimensionality usually requires:

- larger datasets
- increased model complexity

Our approach

- Additional reconstructed observables introduced without architectural changes:
- GAN architecture unchanged
- Underlying MC statistics unchanged
- Only feature-aware preprocessing adapted

Metric	Good	Excellent
Mean KS	<0.06	<0.035
Max KS	<0.15	<0.08

Computational impact

Full MC production

- Requested events: **800M**
- Resources:
  - **~24 days @ 10k CPUs**
  - **~38 TB output**
  - CPU efficiency: **69%**

Effective reconstructed statistics

- Real data sample:
  - **86M events**
- Reconstructed signal sample used for GAN:
  - **531k signal candidates**

800M MC events
↓
531k reconstructed signal candidates
↓
Train GAN once
↓
Generate arbitrarily large synthetic samples

GAN approach

Trained on CERN SWAN

- 4 CPUs
- 16 GB RAM

Training time

- **8 features:** ~10–15 min
- **12 features:** ~20–30 min

Generation

- Synthetic sample generation:
  - **< 1 minute for 86M-scale synthetic samples**

☑ Typically requires MC statistics **4–5× larger than data**
☑ **800M simulated events required to obtain O(10^5) usable signal samples**

☑ Decouples reconstruction-level statistics from full simulation cost

☑ Enables scalable MC augmentation for rare-signal workflows

Rare-signal scaling challenge

- Current benchmark Ξ_c^+:
  - O(10^3) reconstructed candidates from 86M events
- Future rare/exotic channels:
  - potentially O(10–10^2) candidates from O(10^9) events

Generative augmentation becomes increasingly relevant as signal rarity increases !

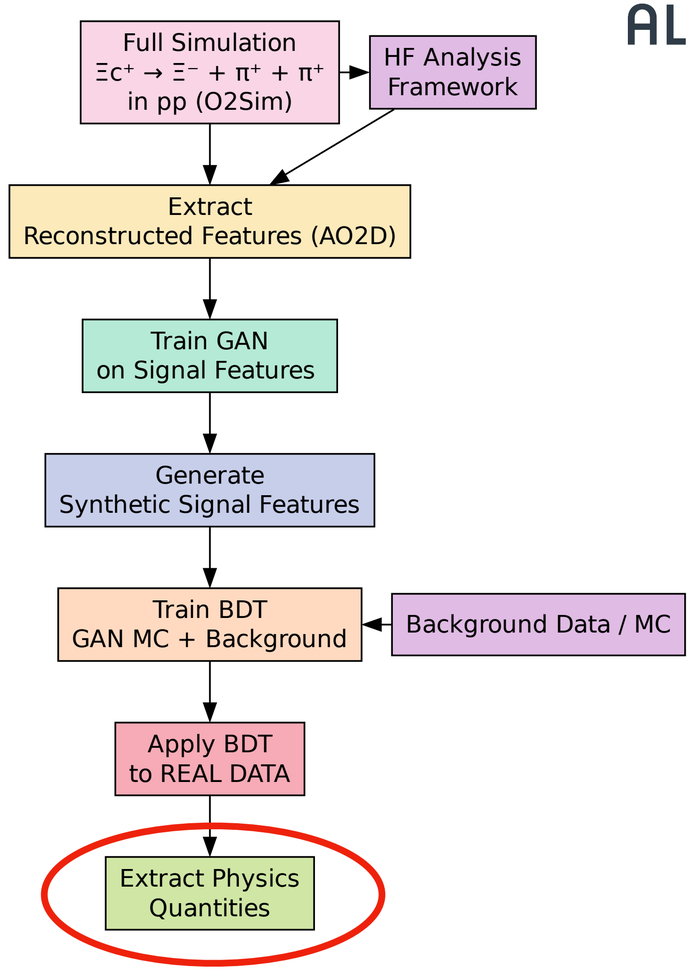
Current scope and ongoing validation

Current validated scope:

- Feature-level agreement
- Correlation preservation
- Higher-dimensional feature extension
- Computational scaling studies

Ongoing studies:

- Full downstream ML validation
- Signal extraction studies within HF analysis framework
- Evaluation of efficiency propagation



Current work focuses on reconstruction-level MC augmentation rather than detector-response simulation

Summary

Conclusions

- GAN-based reconstructed-level augmentation is feasible, preserving:
  - feature distributions
  - correlations
- GAN architecture scales to higher-dimensional feature spaces without modification

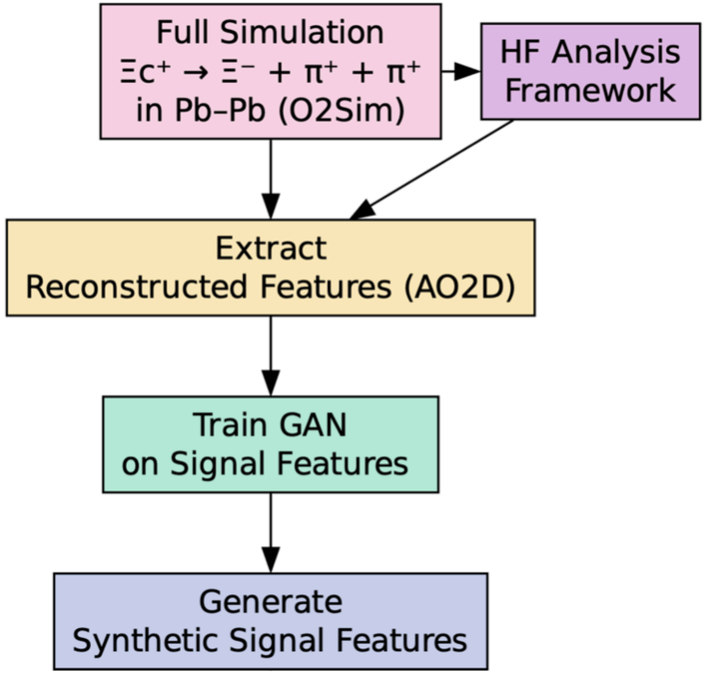
Impact

- Reduces dependence on large-scale MC production
- Enables scalable augmentation for ML workflows

Outlook

- Extension to larger feature spaces
- Application to heavy-ion collision environments
- Run 4 / ALICE 3 applications
- Integration into O2Physics / Hyperloop

Central Pb–Pb collisions



Strong potential for GAN-based augmentation in future HEP workflows, ready to be explored!

Thank you!

Acknowledgement:
NKFIH NKKP ADVANCED 25-153456
NEMZ KI-2022-00058
2025-1.1.5-NEMZ KI-2025-00005
2024-1.2.5-TET-2024-00022
and
Wigner Scientific Computing Laboratory