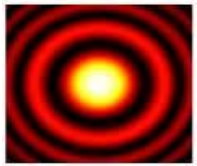
Surrogate Model



QuantOm Collaboration

Jefferson Lab

Argonne NATIONAL LABORATORY

QCD at the Femtoscale in the Era of Big Data

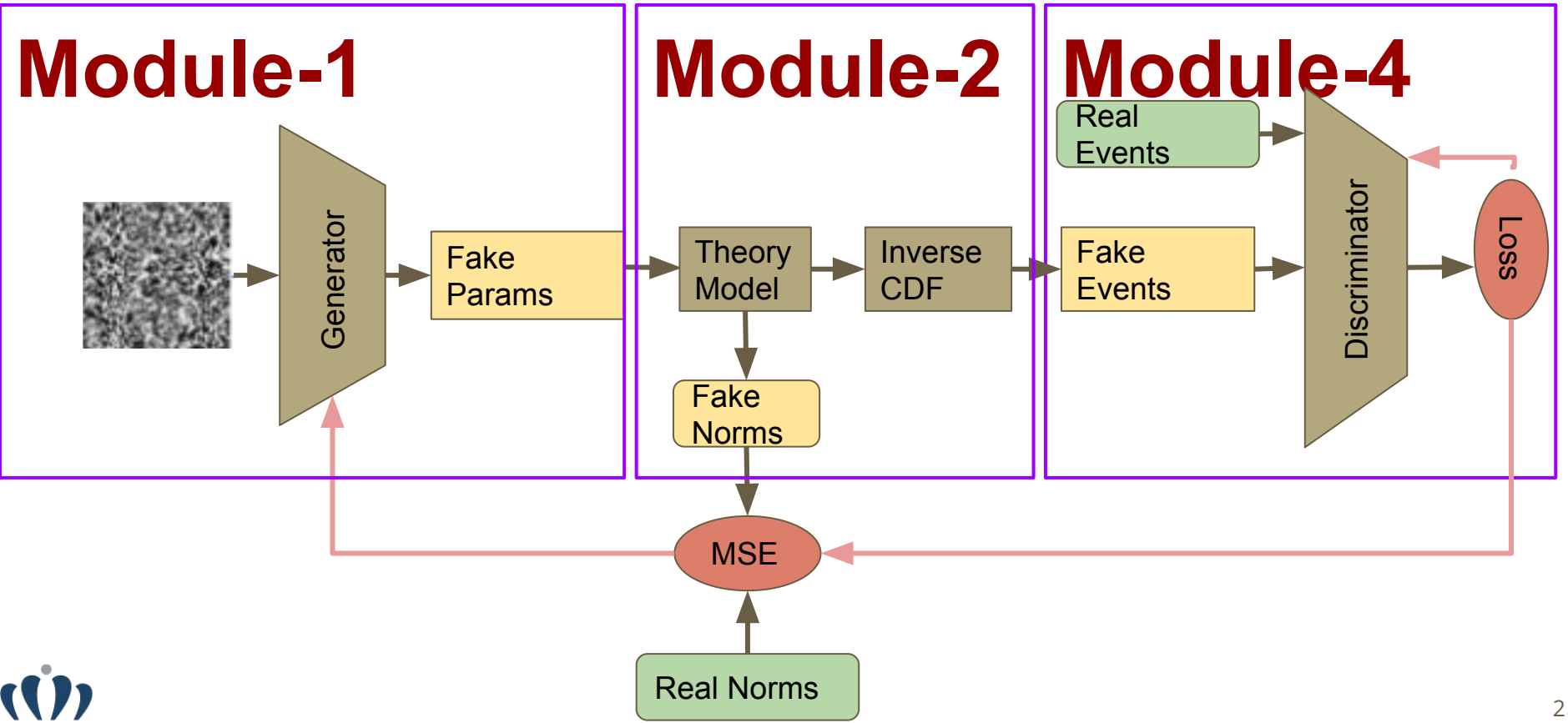
INT Workshop

Tareq Alghamdi

Seattle, Washington

June 28, 2024

Current Workflow implementation



Introduction

The goal is infer the parameters based on the observed physics events.

- In this application, the PMDs for a simplified version of DIS on protons and neutrons are given by

$$\begin{aligned} \boldsymbol{\sigma}_1(x;p) &= 4u(x;p) + d(x;p), \\ \boldsymbol{\sigma}_2(x;p) &= u(x;p) + 4d(x;p)\,. \end{aligned} \tag{1}$$

Where σ1 and σ2 are cross sections (un-normalized probability distributions) and u(x) and d(x) are the universal 1D QCFs called up- and down-quark PDFs.

- QCFs for the proxy problem with two channels are defined by:

$$\begin{aligned} u(x;p) &= N_u x^{a_u}(1-x)^{b_u} \\ d(x;p) &= N_d x^{a_d}(1-x)^{b_d} \end{aligned} \tag{2}$$

Where $x \in (0,1)$ and the parameter vector {Nu, au, bu, Nd, ad, bd} is undetermined.

- We observe events {σp, σn} generated by model (2) and filtered through cross-sections defined in (1) for defined values of the shape parameters.
- From these QCFs, we create PMDs. Then, sampling the PMDs generates the observable physics events.
- These events then serve as the proxy ”experimental events” within the workflow.
- In this study, we consider the parameters p_{true}=[2.1875,−0.5,3,1.09375,−0.5,4] as the ground truth of a control test case, within the parameter bounds of

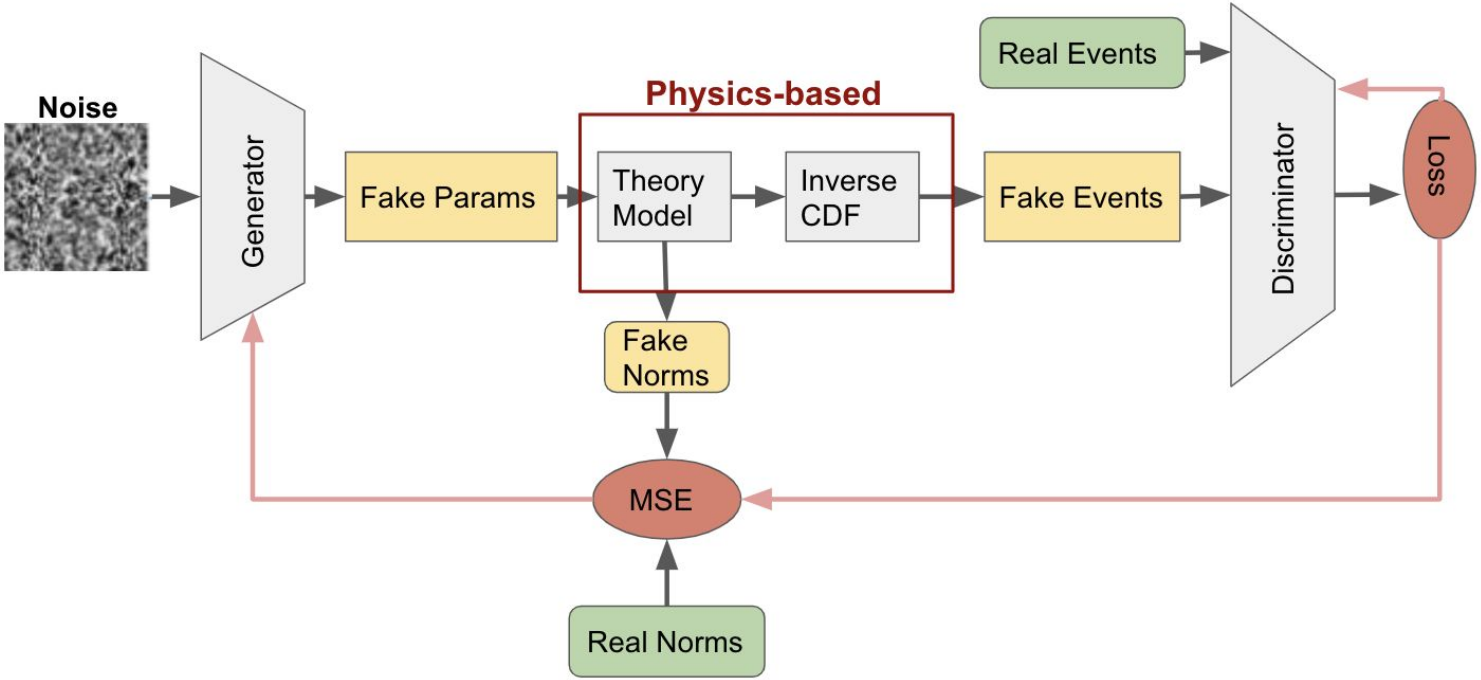
$$[(10^{-5},3),(-1,1),(10^{-5},5),(10^{-5},3),(-1,1),(10^{-5},5)].$$

Motivation:

- Inverse Cumulative Density Function (CDF) can be used to perform 1-dimensional sampling that is differentiable.
- Traditional sampling does not allow backpropagation and gradient flow through it for neural network.
- Traditional samplers require sampling the physics-based QCF functions that are computationally costly to evaluate.

Why do we use a Generative Adversarial Network(GAN) as a surrogate?

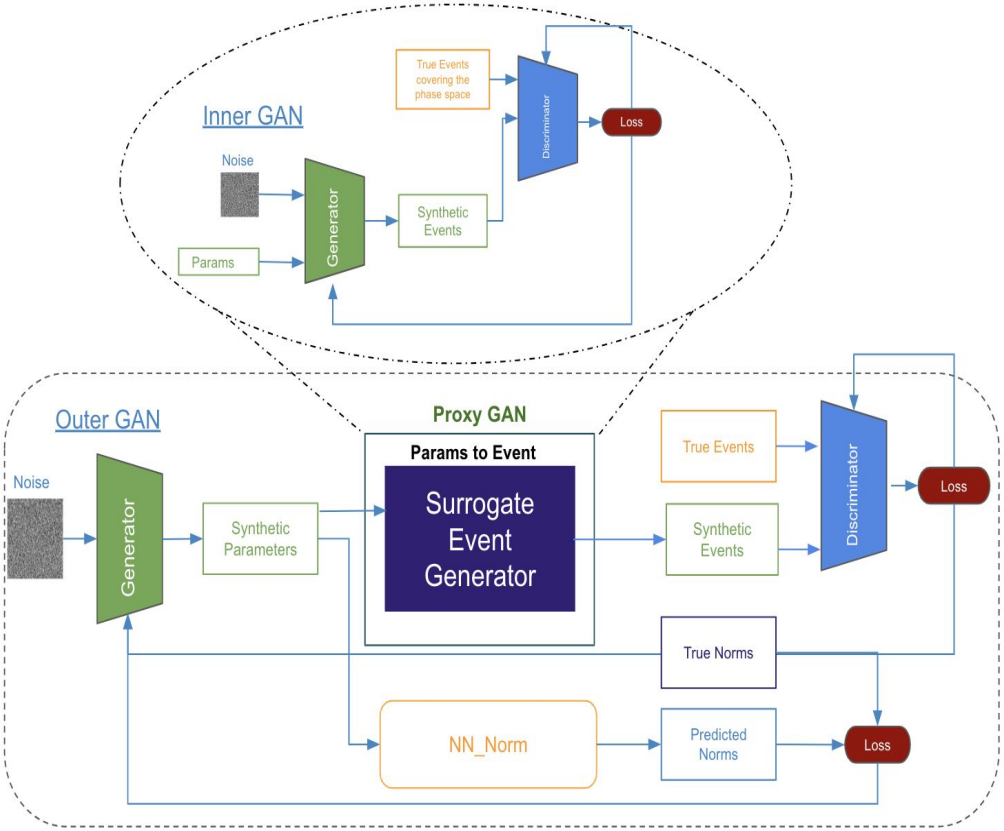
- GAN possess super resolution power due to discrimination generation competition and backpropagation.
- Backpropagation enables the GAN to learn to mimic realistic datasets
- Eventually, GAN will be able to learn the posterior of the parameters that produce a given distribution of the observables.
- The purpose of the surrogate model is to replace the expensive QCF sampling with computationally efficient surrogate neural networks evaluations.


Noise
Generator
Fake Params
Physics-based
Theory Model
Inverse CDF
Fake Events
Real Events
Discriminator
Loss
Fake Norms
MSE
Real Norms

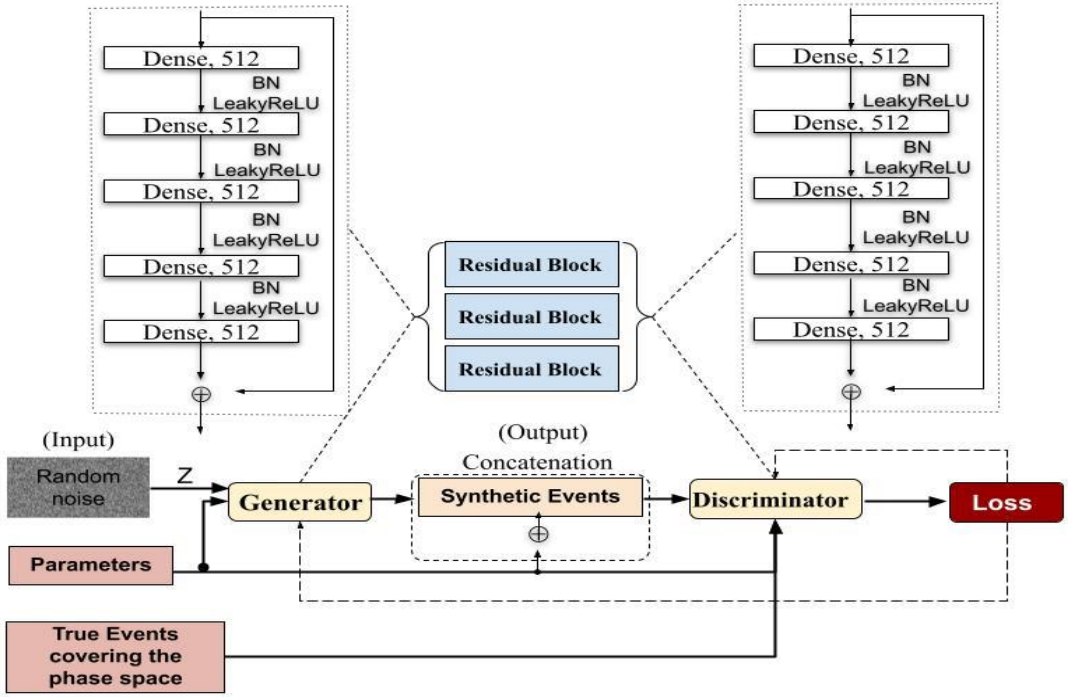
Proposed Method

- We proposed an end-to-end machine learning framework that addresses the challenge of utilizing event-level data to make inferences on QCF parameters.

- The proposed method follows the structure of adversarial learning without relying on any specific underlying physics theory. This approach allows the framework to learn solely from the data.
- It consists of a conditional GAN-based surrogate event generator responsible for generating synthetic event samples from parameterized QCFs, and an outer-GAN that performs the inverse mapping from the observed events back to the parameter space.
- The utilization of a discriminator helps guide the updates of the parameter generator based on event-level data.

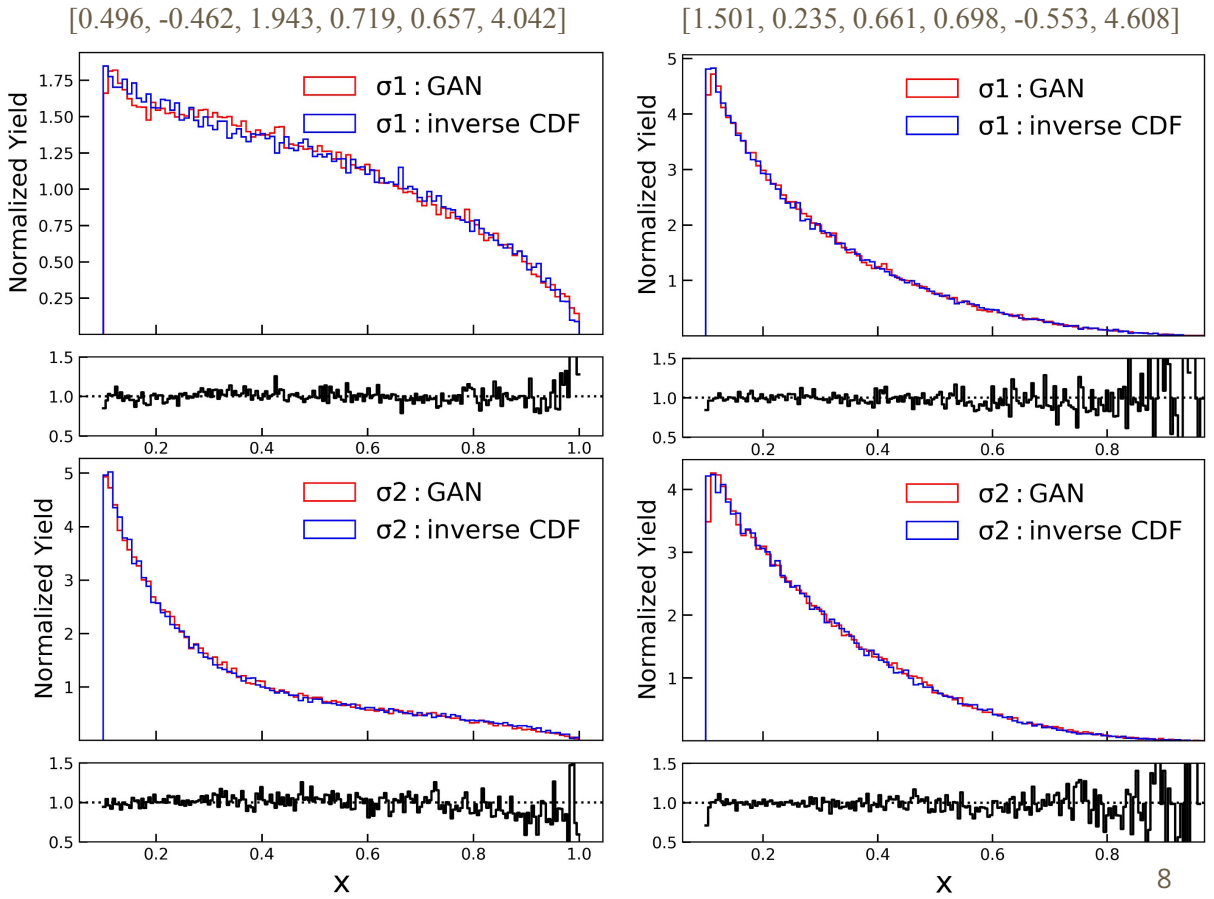


Surrogate Event Generator Architecture



Surrogate Event Generator Results

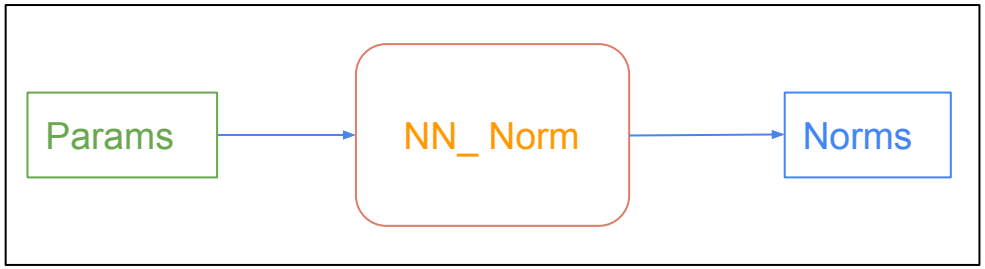
- Surrogate event generator mapping parameters to the distribution of event samples for randomly generated parameters.
- The ratio of the Proxy surrogate event generator model to the truth yields is shown at the bottom of each panel.



Neural Network to estimate normalization constants

- While the physics-based model can generate events and calculate the normalization constants, the surrogate GAN, as an emulated sampler, can only generate synthetic events, but is not able to estimate the normalization constants.
- We build a neural network to estimate the normalization constants.
- The mapping between the parameters {N_u, a_u,b_u, N_d, a_d, b_d}, and the normalization constants is a simple regression problem.

Simple Neural Network



For example:

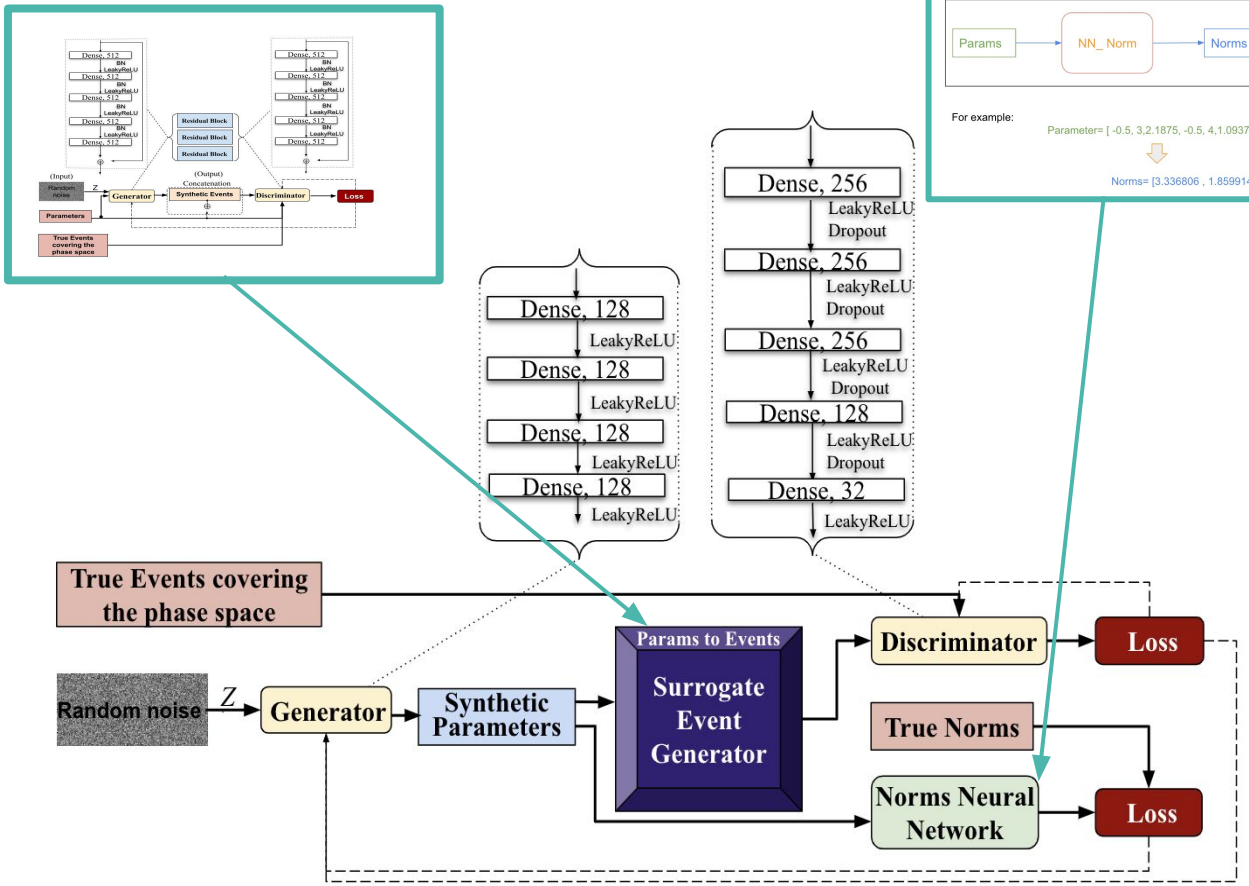
Parameter= [-0.5, 3,2.1875, -0.5, 4,1.09375]

Norms= [3.336806 , 1.8599147]

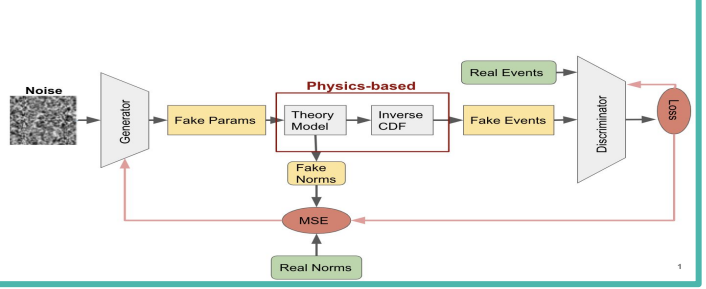
GAN-based Event-level Inverse Mapper(GEIM) Architecture

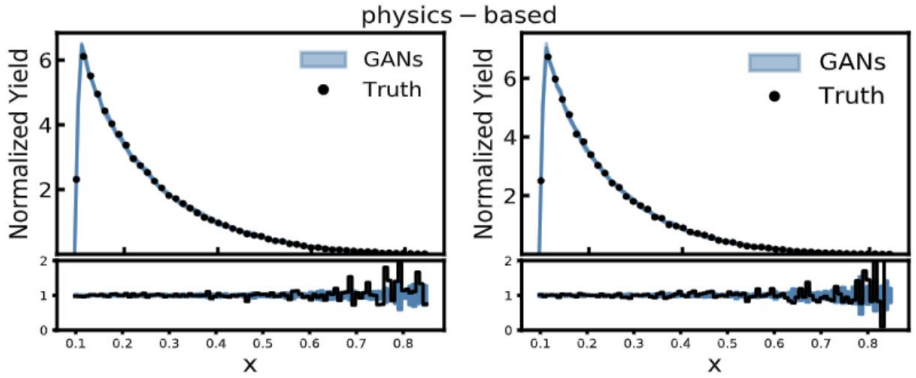
- It is designed to address the inverse problem of femtoscale imaging in QCD.

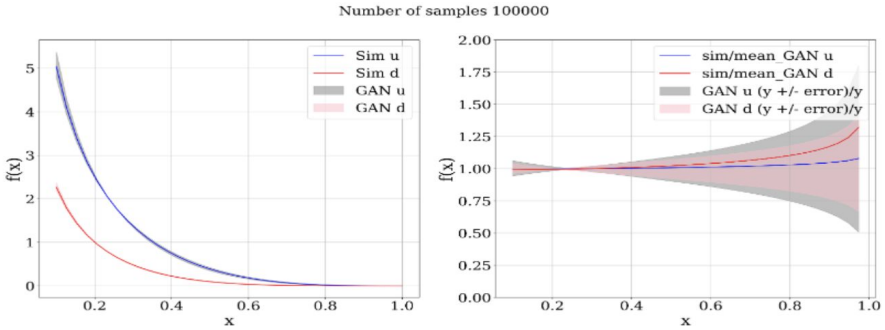
- GEIM consists of two GANs:
  - The conditional GAN-based surrogate event generator, which replaces the physics-based QCF model to generate synthetic events,
  - And the outer-GAN, which performs the backward mapping to derive the parameter distributions.

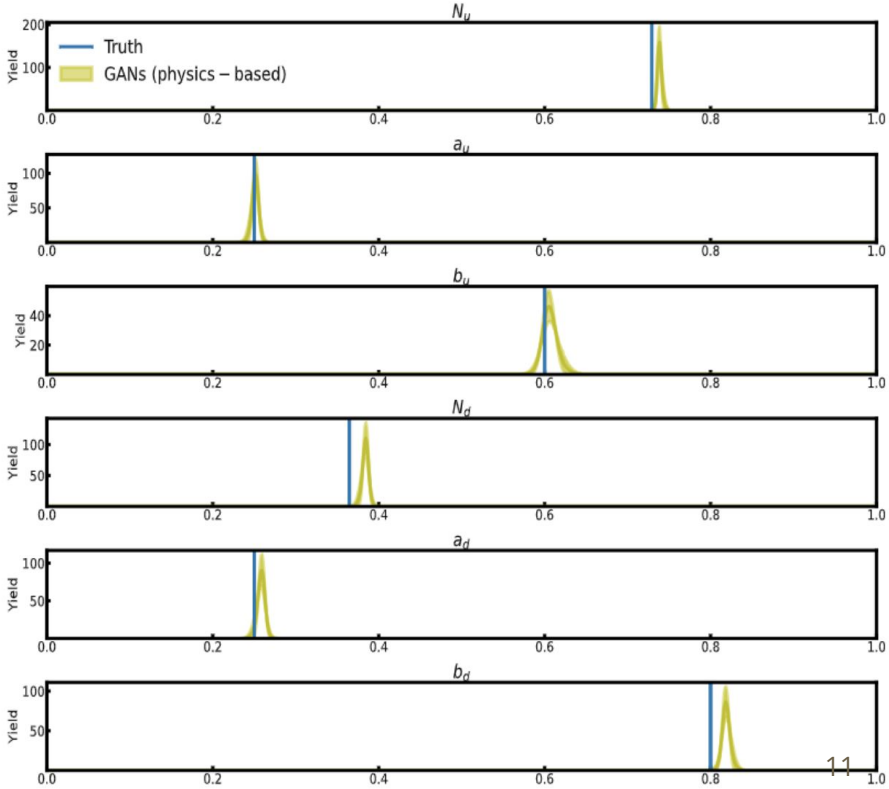


GEIM using a Physics-based model
On a control parameter set as true= [2.1875, -0.5, 3, 1.09375, -0.5, 4] as the test

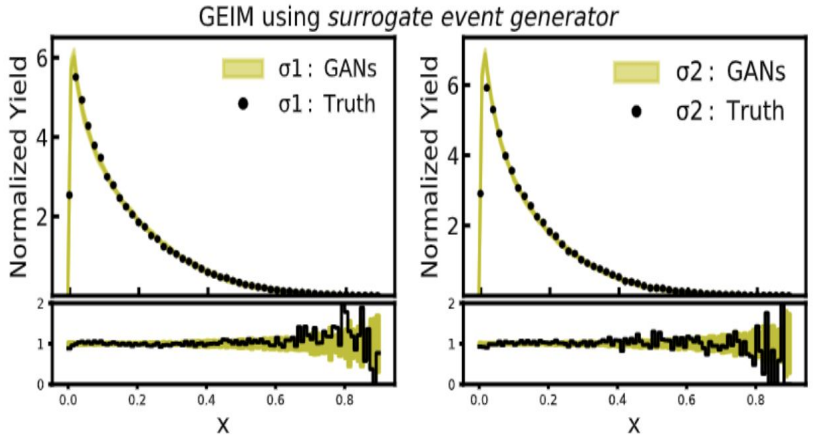


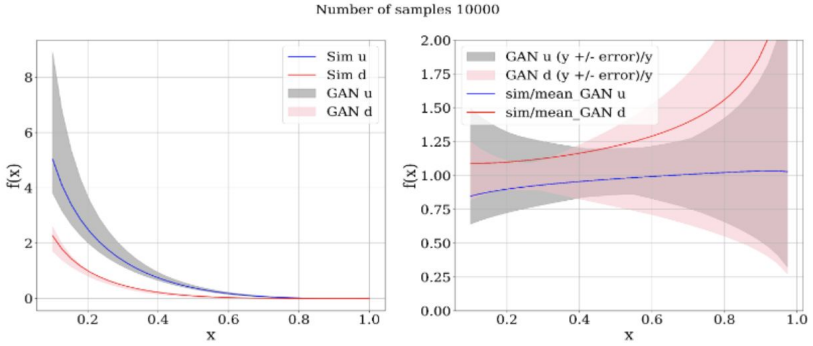


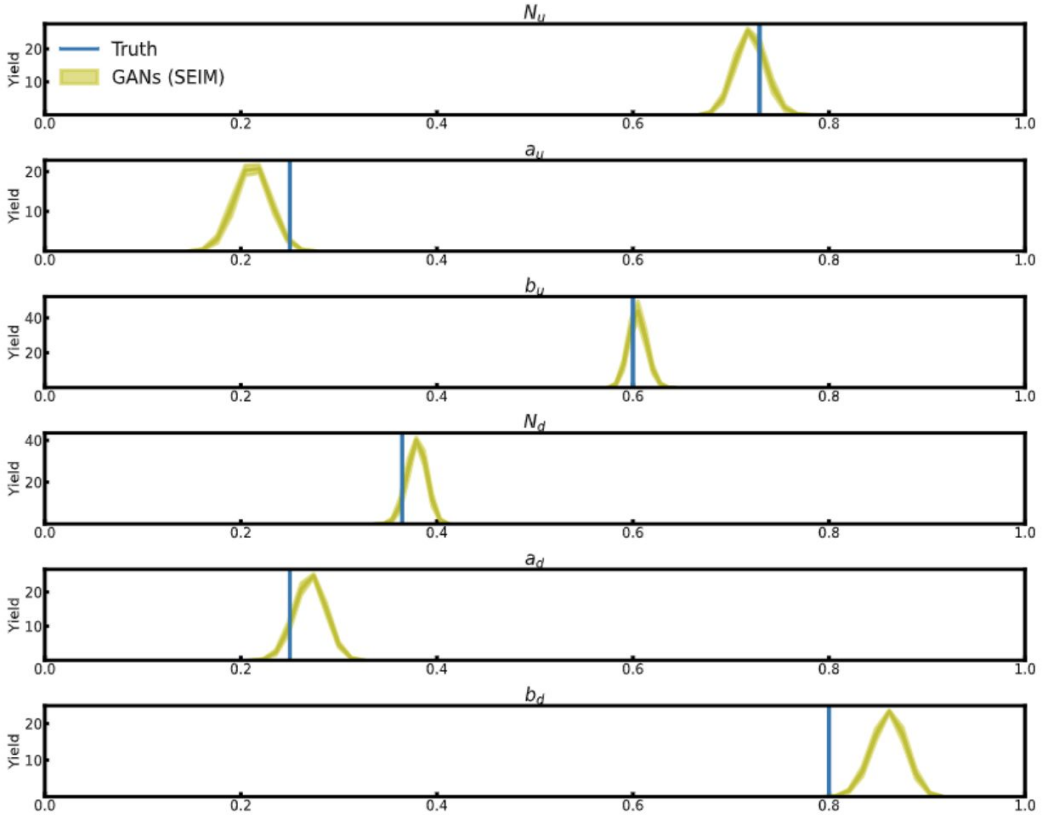




GEIM using a surrogate event generator
On a control parameter set as true= [2.1875, -0.5, 3, 1.09375, -0.5, 4] as the test





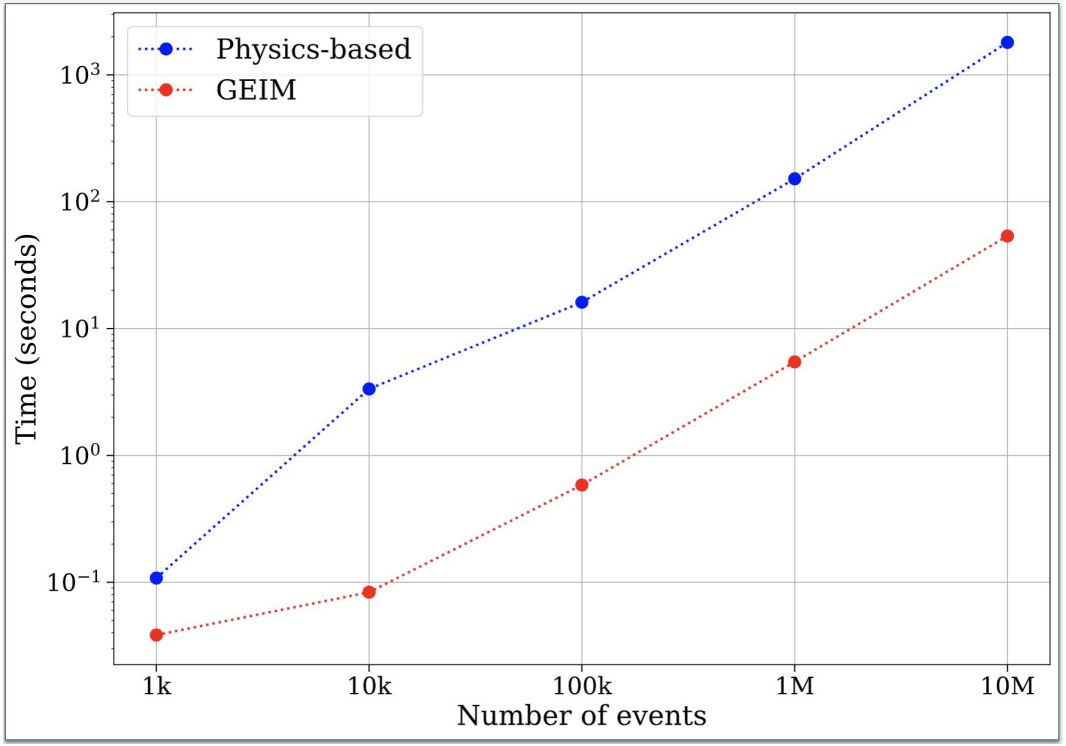


- It leads to wider derived parameter distributions and this is due to the error generated by the surrogate event generator to approximate the real physics-based probability density function, which increases the uncertainty in the inverse inference.

How does the computational efficiency of the surrogate event generator's neural networks compare to that of physics-based models?

- Comparing the computational times of generating 1^3 to 10^7 events using the inverse CDF method to sample the physics-based QCFs and the surrogate event generator.
- More cost-efficient than inverse CDF sampling of the physics-based model, which requires evaluating complex physics-based distribution functions.
- Resulting in approximately 20 times faster performance.

Computational time comparison

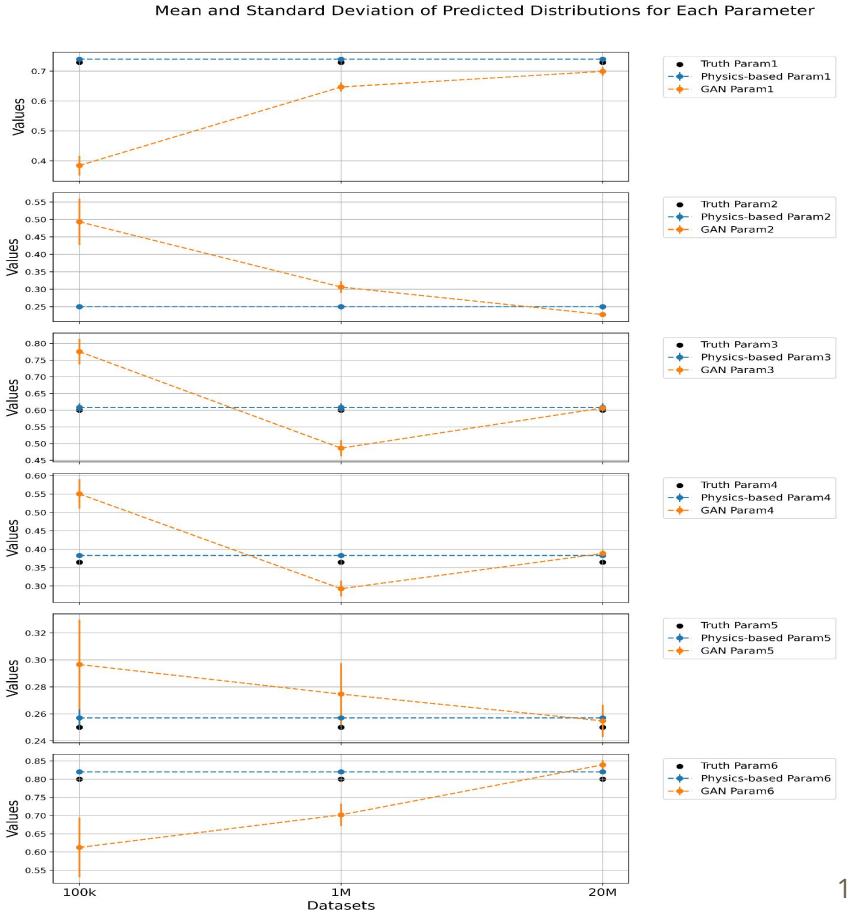


Comparison using different sizes of events while training the inner GAN (Inner GAN vs. inverse CDF)

- The plots compare three sets of values for each parameter:
  - Truth (black dots): The actual values of the parameters.
  - Physics-based model (blue dashed lines): Values predicted by the physics-based inverse CDF sampling model.
  - GAN model (orange dashed lines): Values predicted by the inner GAN model.

- The inner GAN is trained with 100k, 1M, and 20M events.
- The pre-trained inner GAN is used within the outer model to predict the values of the six parameters.
- These predictions are compared to the truth values and those from the physics-based model.

- The GAN model's performance improves with more training data, showing a reduction in the mean and standard deviation error.
- Training the GAN model with a substantial amount of data allows it to closely approximate the true parameters.



Drawbacks of this approach:

- We need to know the range of the parameters to train the Surrogate event generator.

- Training the Surrogate event generator in our approach requires a substantial dataset, specifically 20 million samples. This high sample requirement results in significant computational cost and resource consumption, which can be a limiting factor in terms of both time and available computing resources.

  If we scale the problem to higher dimensions, the data requirements would increase exponentially.

- **While this approach works for a simple 1-D deep inelastic scattering problem, a key question remains:**

  Will it work if we scale the problem to higher dimensions??

Moving Forward:

Event Generators

FOLDING & UNFOLDING THE DETECTOR SMEARING EFFECTS

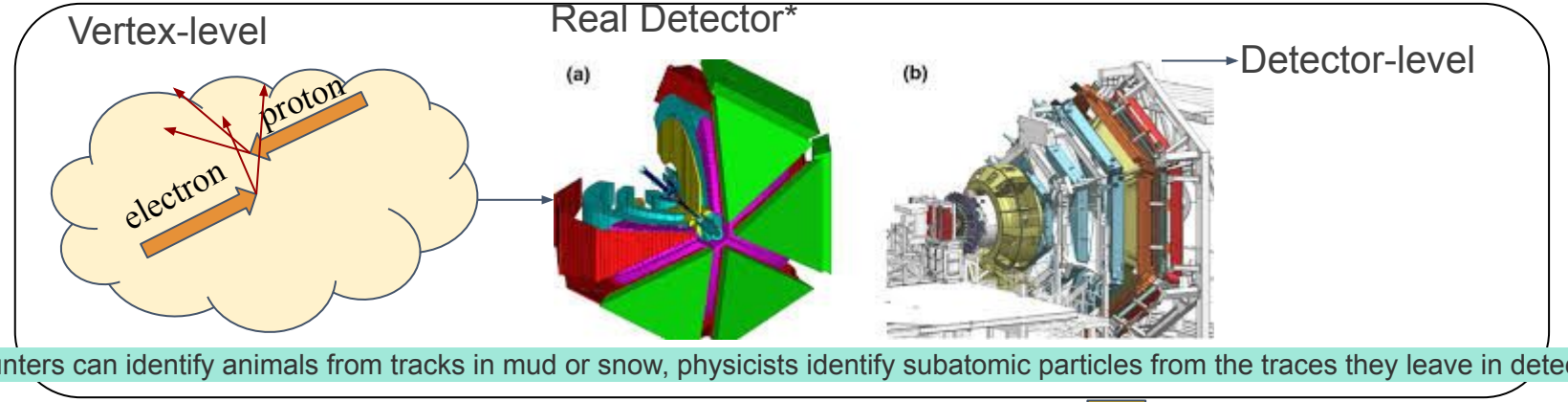
QCD at the Femtoscale in the Era of Big Data

INT Workshop

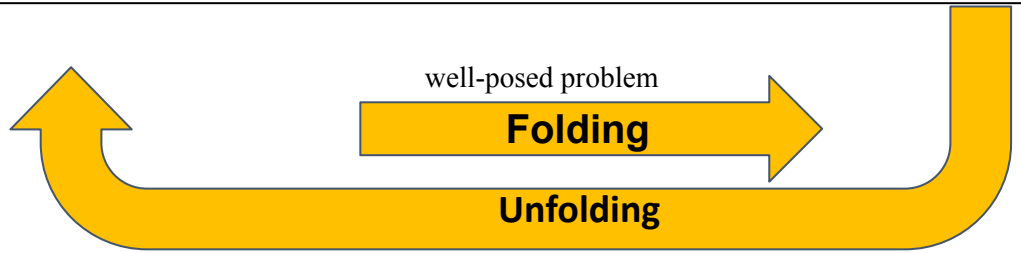
Tareq Alghamdi Seattle, Washington June 28, 2024

Particle collision experiment:

- Data collected by NP/HEP experiments are (always) affected by the detector's effects
- Before starting physics analysis the detector's effect unfolding is required
- Traditional observables may not be adequate to extract physics in multidimensional space (multi-particles in the final state)
- At High-Intensity frontiers, data sets are large and difficult to manipulate/preserve



Just as hunters can identify animals from tracks in mud or snow, physicists identify subatomic particles from the traces they leave in detectors[1]



ill-posed problem

*CLAS detector at Jefferson Lab

[1]CERN Accelerating science

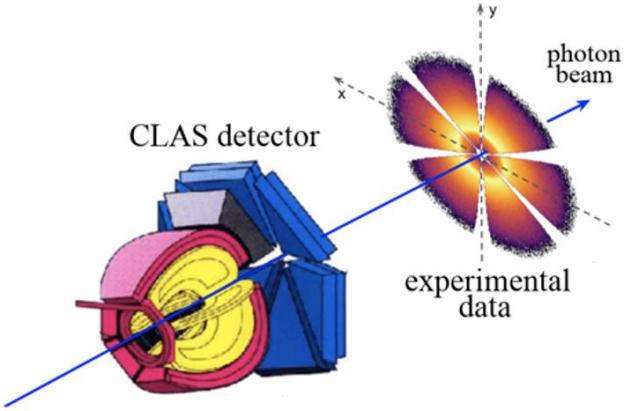
Main components:

- This work has two main components:
  - Simulating the smearing detector effects using ML tools.
    - Folding GAN ⟶ Detector-Simulation (**DS-GAN**) ⟶ **MC-Phase Space pseudodata**

  - Building a ML-based event generator framework to reconstruct vertex-level events.
    - Unfolding GAN ⟶ **UNF-GAN** ⟶ **MC-Realistic pseudodata**

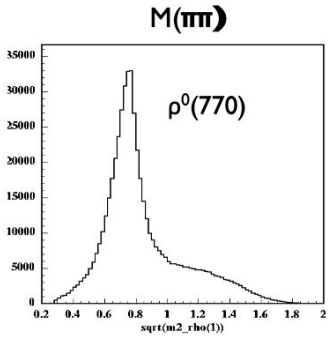
Multi-d cross-section: exclusive 2π photoproduction

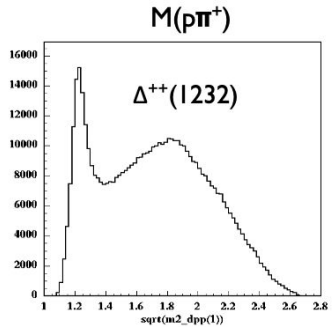
M. Battaglieri *et al.* (CLAS Collaboration) Phys. Rev. Lett. 102, 102001
M. Battaglieri *et al.* (CLAS Collaboration) Phys. Rev. D 80, 072005

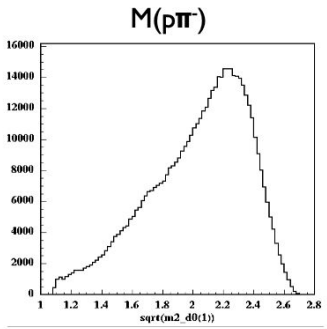


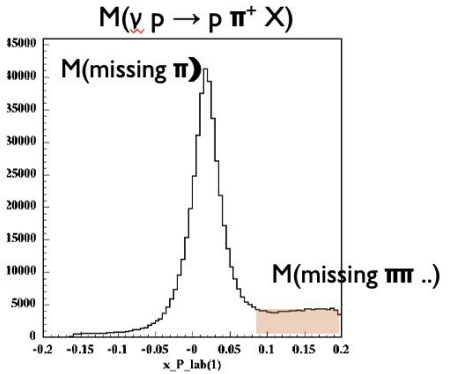
CLAS g11 kinematics

- Dataset used by CLAS Collaboration for many publications
- Fiducial cuts (p, θ, ϕ) as used in published analyses
- Focus on $\gamma p \rightarrow p\pi^+(\pi^-)$
- Final exclusive 2π state identified by missing mass technique (variables are reconstructed by energy/momentum conservation)
- Multi-pion backgound comes from $\gamma p \rightarrow p\omega^0 \rightarrow p\pi^+\pi^-\pi^0$
- At $E_\gamma = (3-4)$GeV reaction dynamics are dominated by ρ^0 photproduction through $\gamma p \rightarrow p\rho^0$ and Δ^{++} resonance excitation through $\gamma p \rightarrow \Delta^{++}\pi^-$







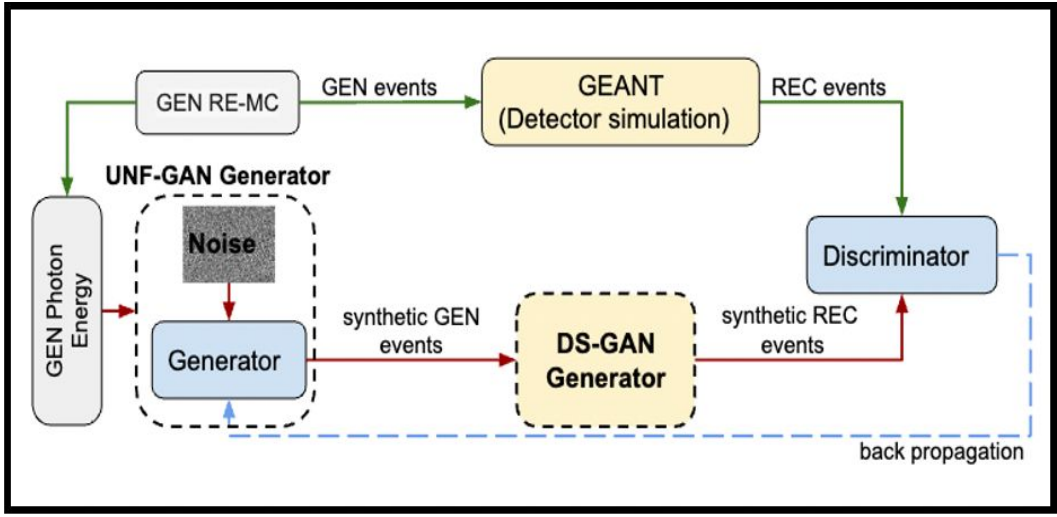


2π photoproduction closure test

- CLOSURE TEST:

  Demonstrate that GANs reproduce 'true' multi-d correlations, unfolding CLAS detector effects, comparing vertex-level (GEN) events with GAN GEN SYNT events, trained at detector-level and unfolded with a (GAN-based) detector proxy

1. Generate events with a (realistic) Monte Carlo 2π photoproduction model (RE-MC GEN pseudodata)
2. Apply detector effects (acceptance and resolution) via GSIM-GEANT (RE-MC REC pseudodata)
3. Deploy a secondary GAN (DS-GAN) to learn detector effects using an indipendent MC event generator (PS-MC) + GSIM-GEANT (GEN and REC pseudodata)
4. Deploy the unfolding GAN (UNF-GAN) that includes the DS-GAN, and train it with RE-MC REC pseudodata
5. Compare UNF-GAN GEN SYNT data to RE-MC GEN pseudodata
6. Replace RE-MC REC pseudo data with CLAS data in the training to unfold the vertex-level experimental distributions

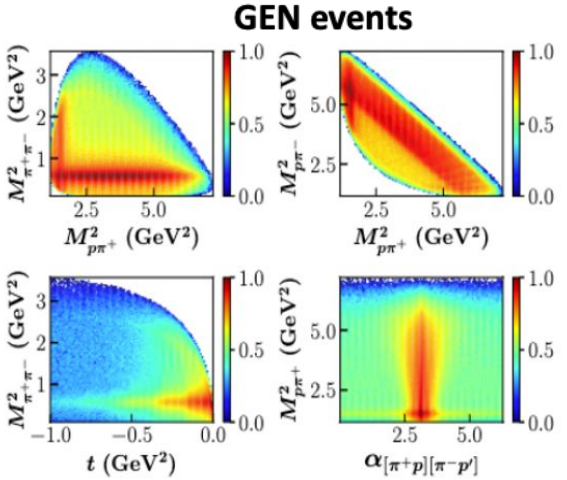


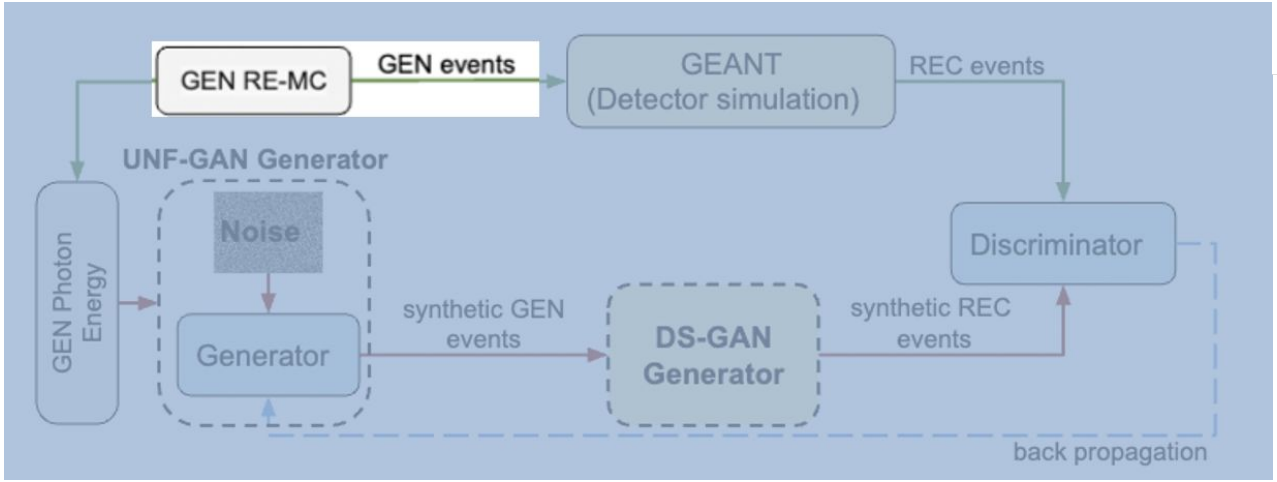
T. Alghamdi, M. Battaglieri, Y. Li, N.Sato, A.Szczepaniak, and et al.
"Toward a generative modeling analysis of CLAS exclusive 2π photoproduction."
Phys. Rev. D 108, 094030 – Published 21 November 2023

2π photoproduction closure test

1. Generate events with a (realistic) Monte Carlo 2π photoproduction model (RE-MC GEN pseudodata)

- RE-MC realistic Monte Carlo event generator to mimic real data. Includes measured cross-sections, angular distributions and decay of dominant mechanisms ($\rho^0, \Delta^{++}, \Delta^0$ + a contact term)

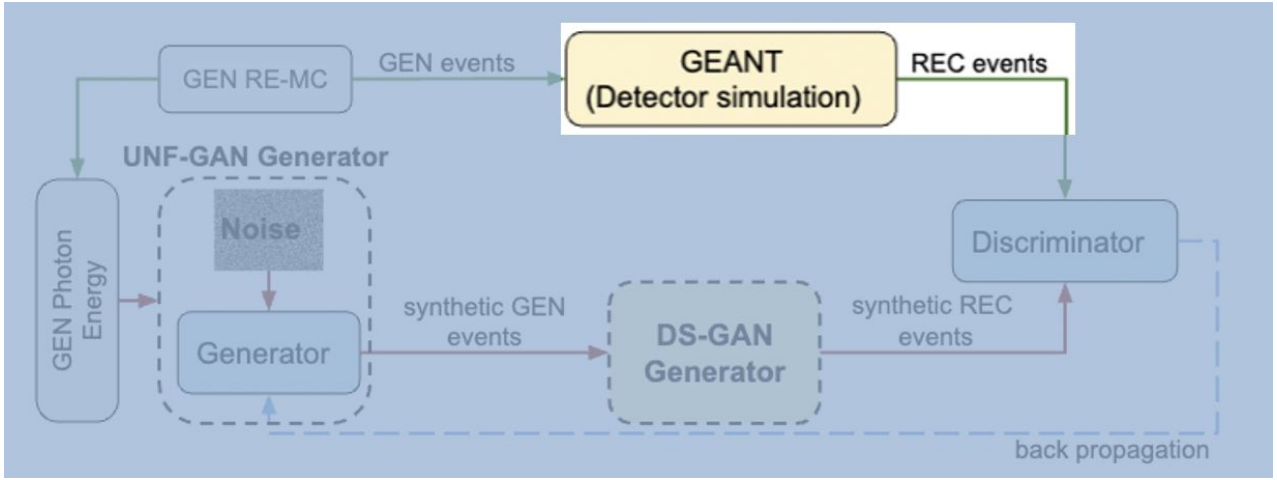




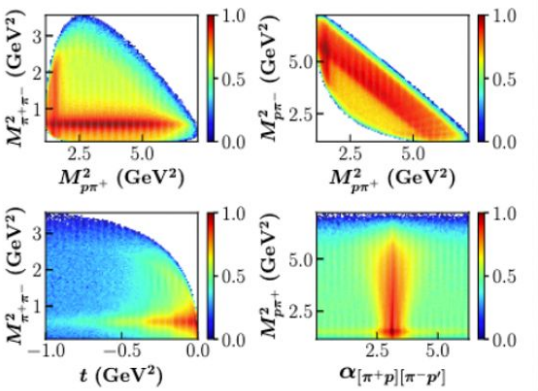
2π photoproduction closure test

2. Apply detector effects (acceptance and resolution) via GISM-GEANT (RE-MC REC pseudodata)

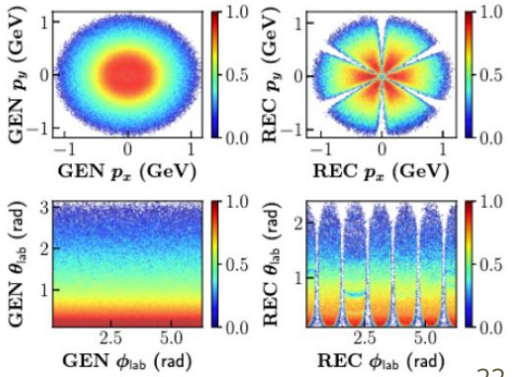
- GSIM: detector simulation package to simulate CLAS detector effects based on GEANT3



GEN events

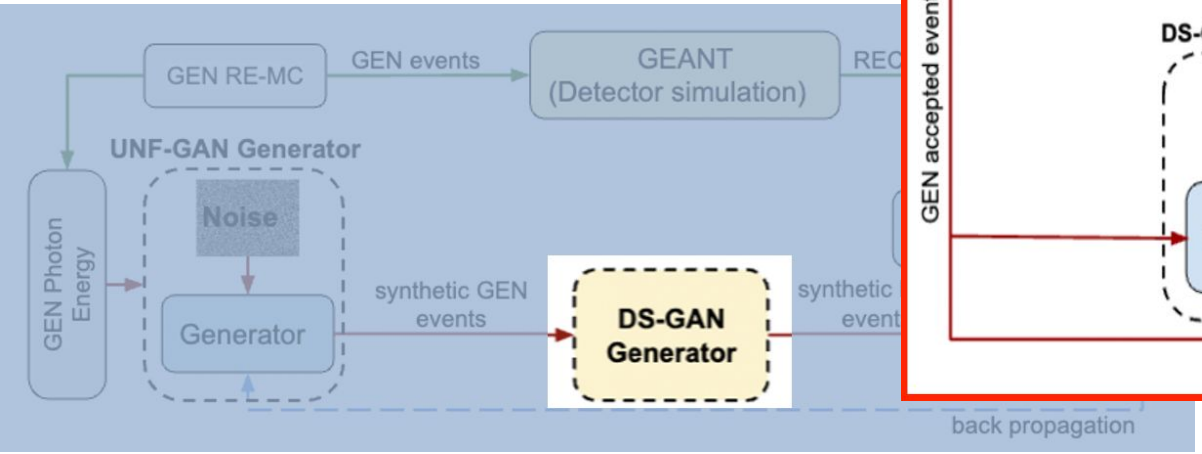


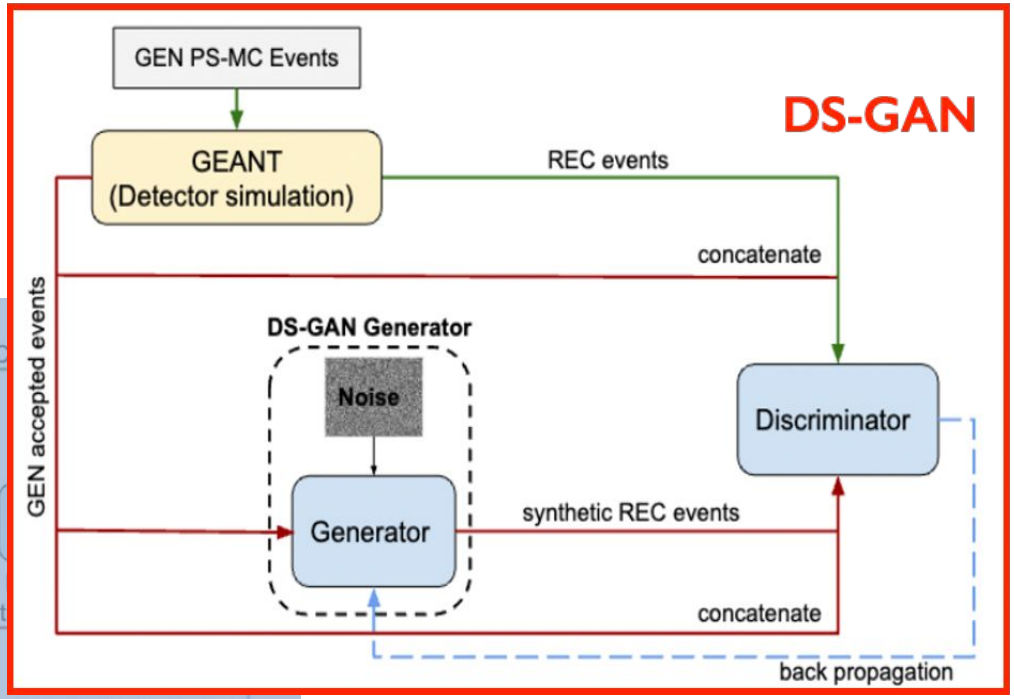
RE-MC REC π^+ events



2π photoproduction closure test

3. Deploy a secondary GAN (DS-GAN) to learn detector effects using an indipendent MC event generator (PS-MC) + GSIM-GEANT (GEN and REC pseudodata)

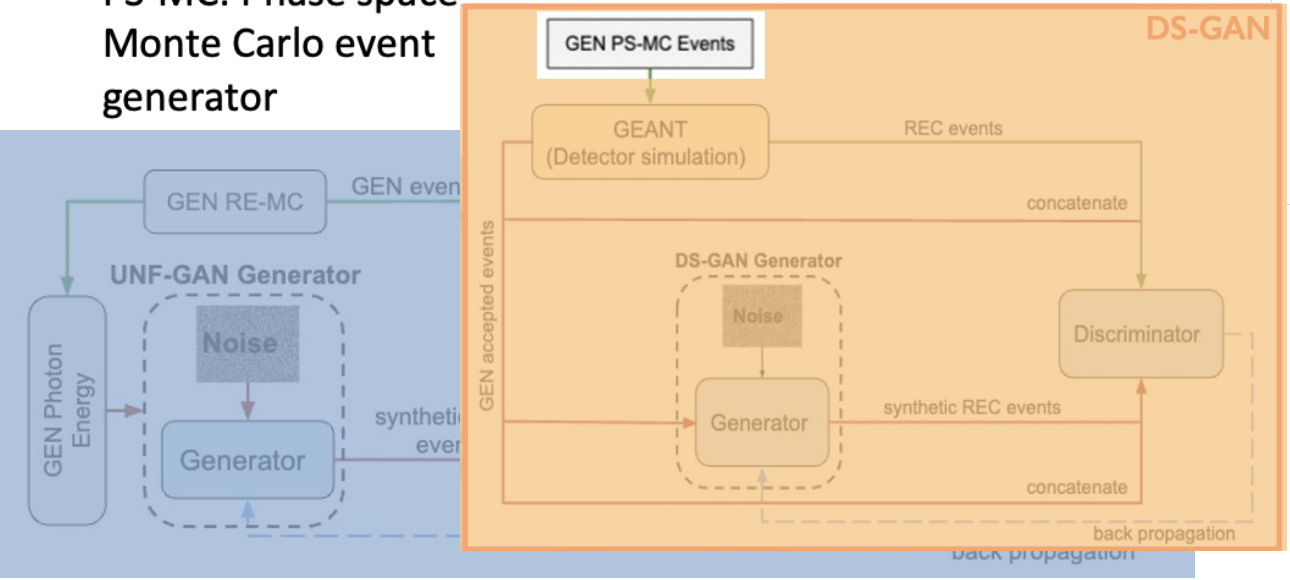




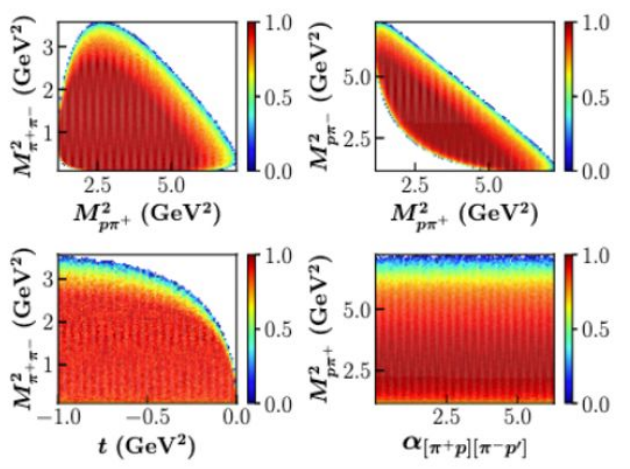
2π photoproduction closure test

3. Deploy a secondary GAN (DS-GAN) to learn detector effects using an indipendent MC event generator (PS-MC) + GSIM-GEANT (GEN and REC pseudodata)

- PS-MC: Phase space Monte Carlo event generator



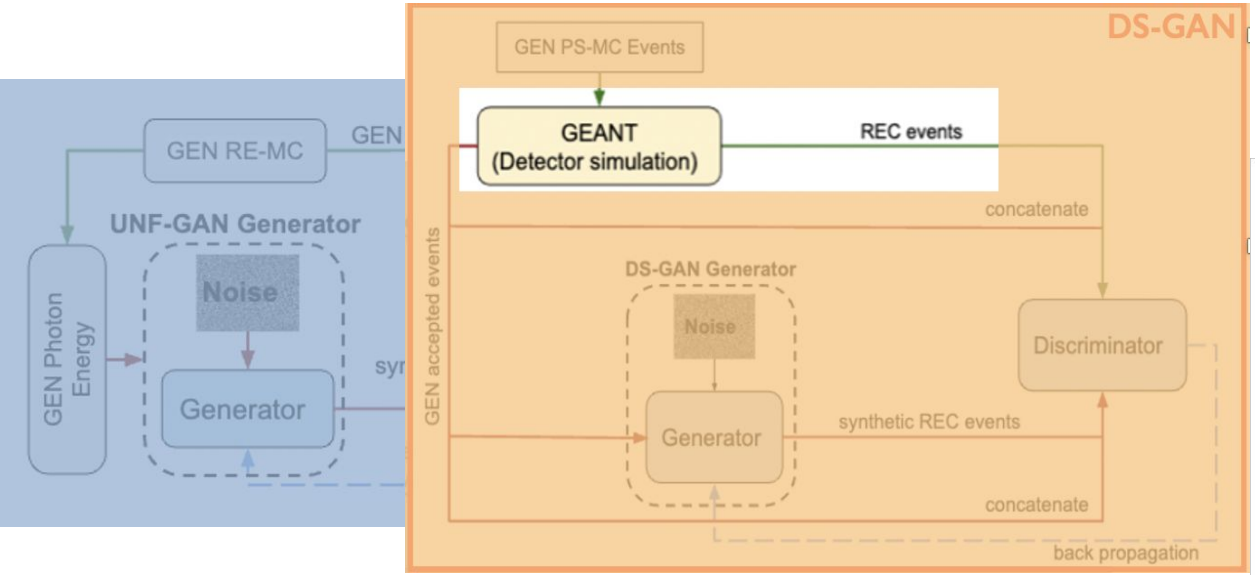
PS-MC GEN events



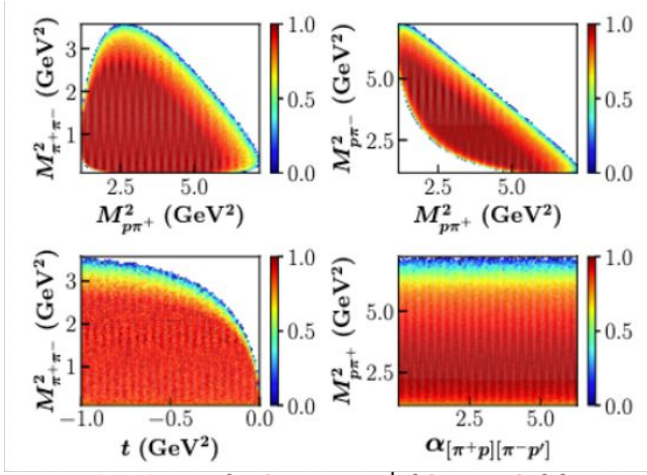
2π photoproduction closure test

3. Deploy a secondary GAN (DS-GAN) to learn detector effects using an indipendent MC event generator (PS-MC) + GSIM-GEANT (GEN and REC pseudodata)

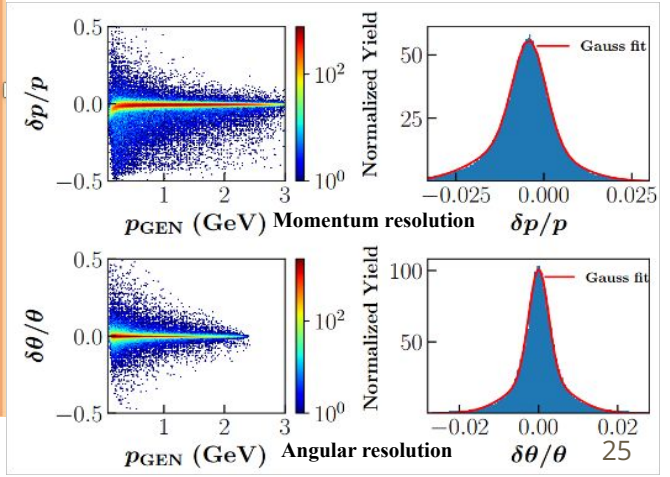
- GSIM-GEANT to simulate CLAS acceptance and resolution



PS-MC GEN events



CLAS resolution on π^+ kin. variables

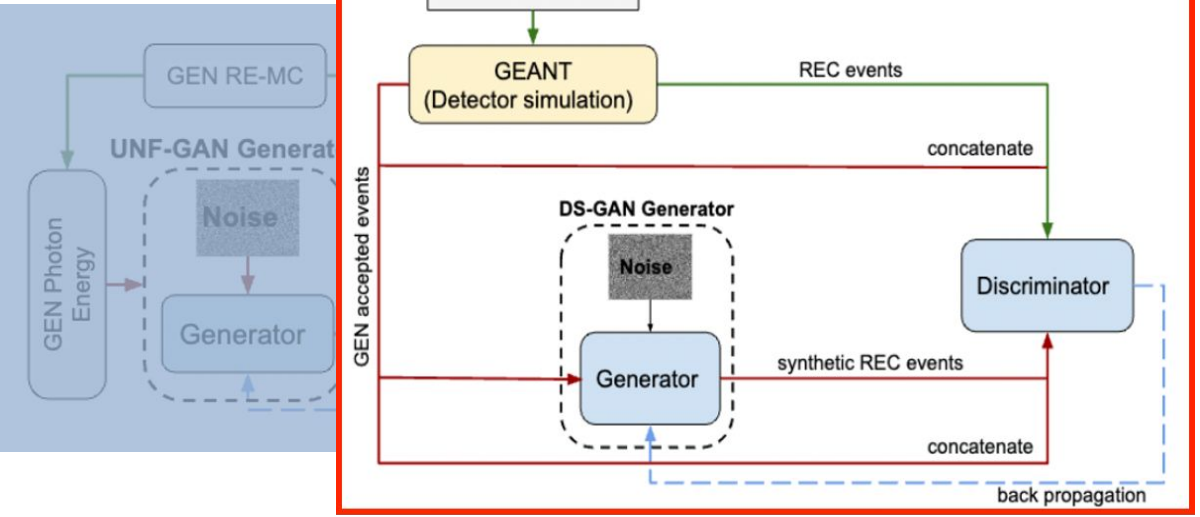


DS-GAN Results

3. Deploy a secondary GAN (DS-GAN) to learn detector effects using an indipendent MC event generator (PS-MC) + GSIM-GEANT (GEN and REC pseudodata)

- Pull calculation for each bin:

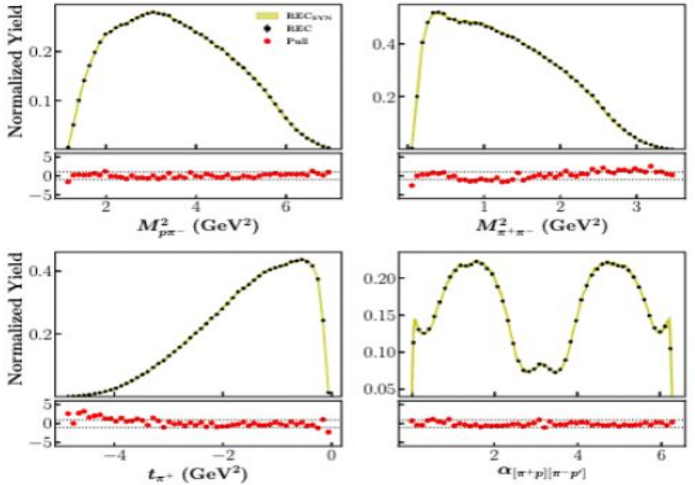
$$\frac{\mu_{\mathrm{SYN}} - \mu_{\mathrm{pseudodata}}}{\sqrt{\sigma^2_{\mathrm{SYN}} + \sigma^2_{\mathrm{pseudodata}}}}$$

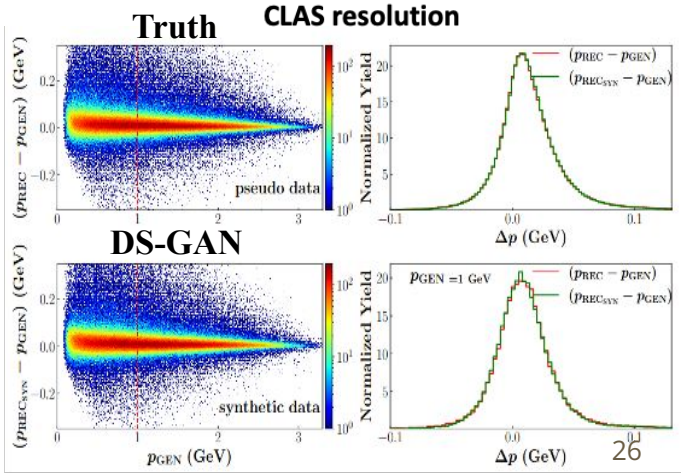


Uncertainty quantification via **pull** calculation: Bootstrap with 20 indipendently trained GANs

DS-GAN learned the CLAS detector effects!

MC REC pseudodata vs. DS-GAN synthetic data

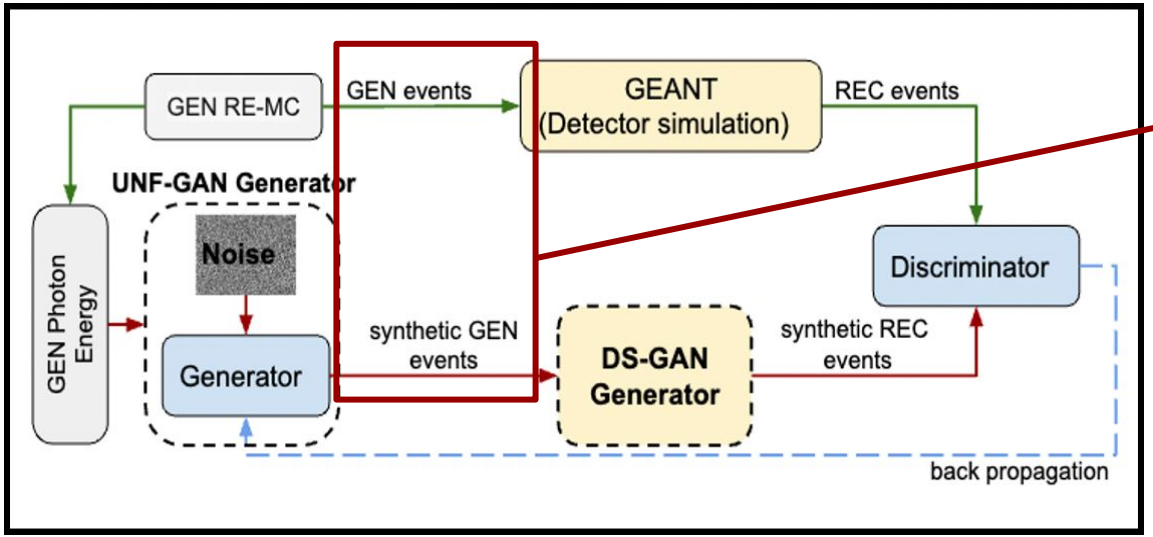




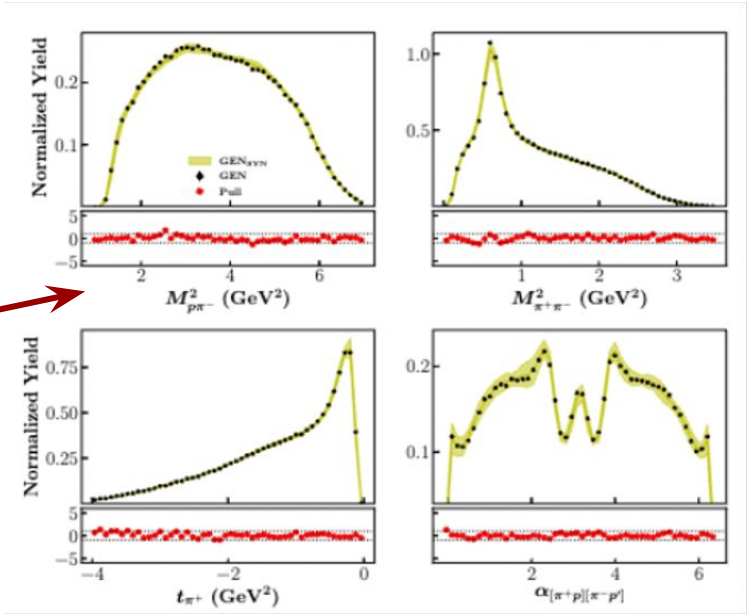
UNF-GAN results

4. Deploy the unfolding GAN (UNF-GAN) that includes the DS-GAN, and train it with RE-MC REC pseudodata

- UNF-GAN trained with REC-MC pseudodata (experimental data proxy)
- DS-GAN used to unfold CLAS detector effects (within acceptance)



RE-MC GEN pseudodata vs. UNF-GAN SYN data



- Systematic of the full procedure (two-GANs) estimated by bootstrap with 20+20 independently trained GANs

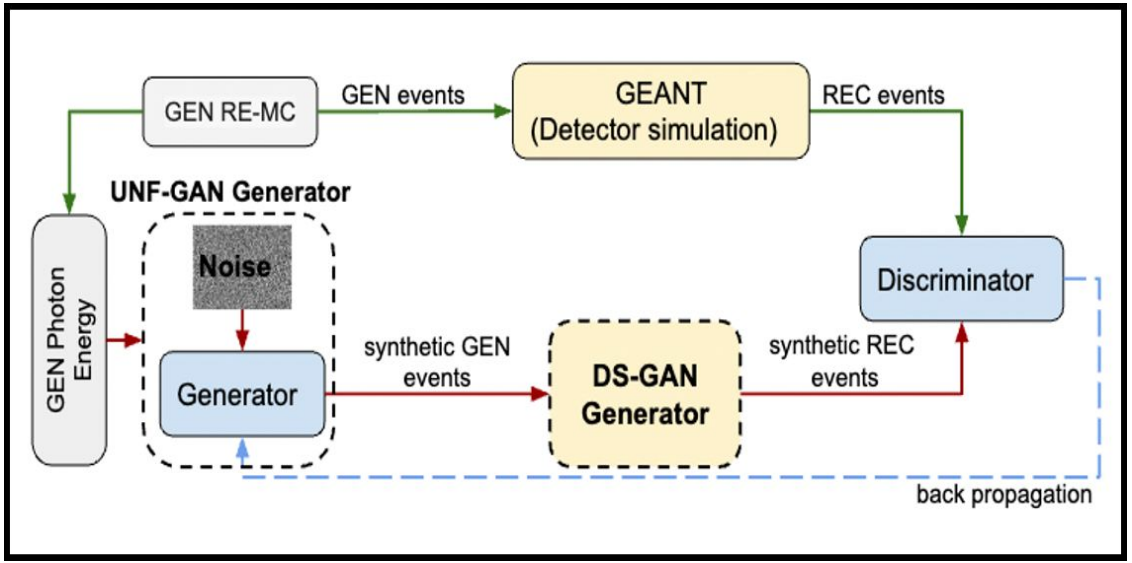
5. Compare UNF-GAN GEN SYNT to RE-MC GEN pseudodata

Good agreement ($\pm 1\sigma$) for vertex-level training variables!

2π photoproduction closure test

4. Deploy the unfolding GAN (UNF-GAN) that includes the DS-GAN and train it with RE-MC REC pseudodata

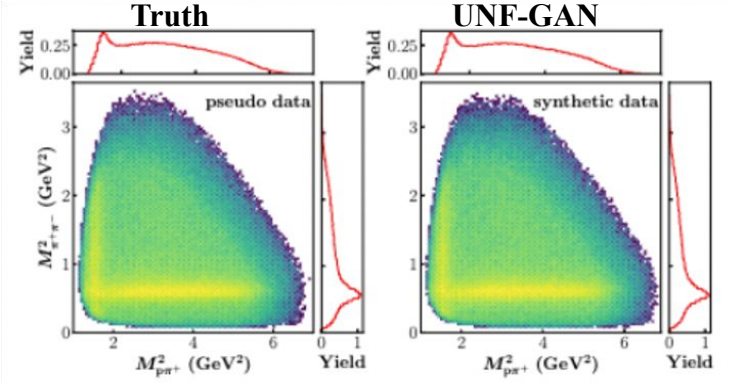
- UNF-GAN trained with REC-MC pseudodata (experimental data proxy)
- DS-GAN used to unfold CLAS detector effects (within acceptance)



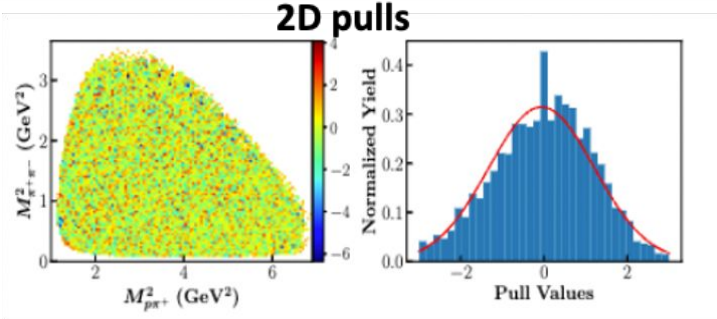
5. Compare UNF-GAN GEN SYNT to RE-MC GEN pseudodata

A key point of this closure test is to demonstrate that synthetic data maintain the correlations of original pseudodata.

RE-MC GEN pseudodata vs. UNF-GAN SYN data



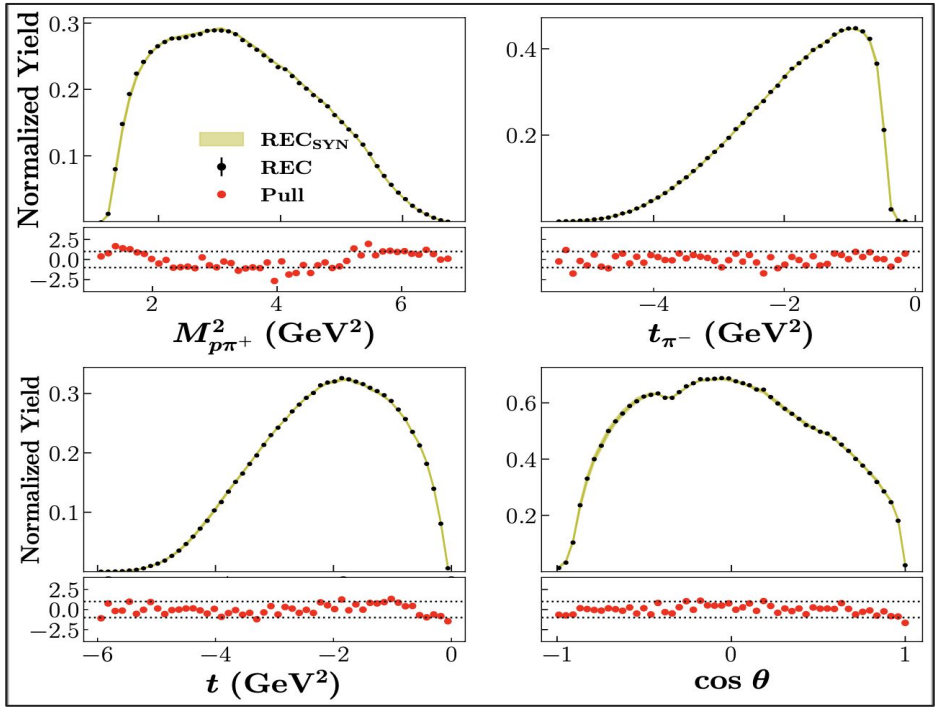
2D pulls



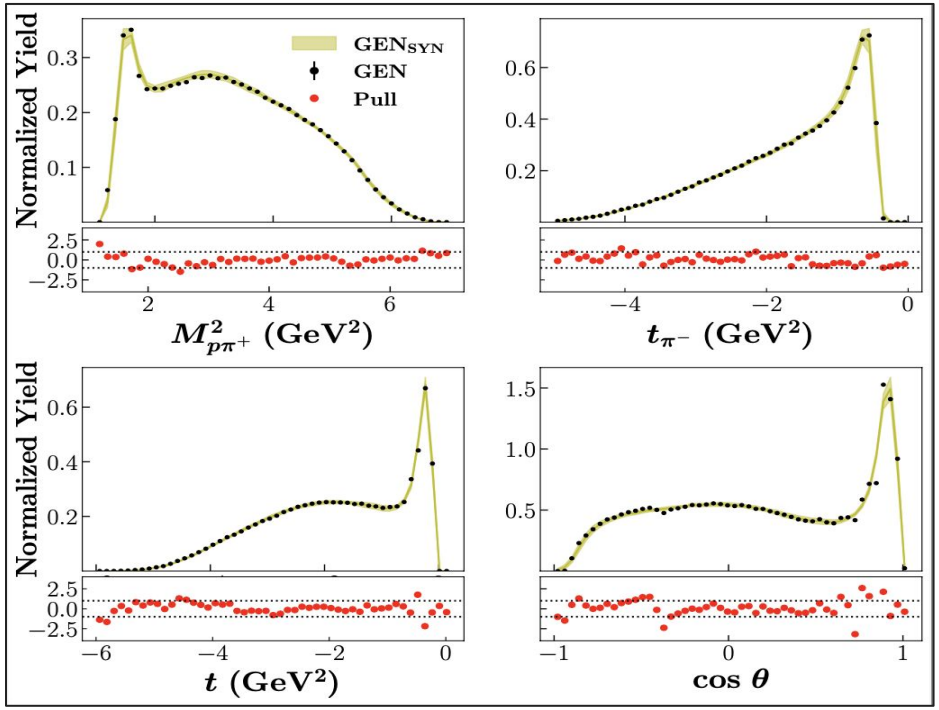
Good agreement ($\pm 1\sigma$) for 2D distributions (correlations)

Derived variables (not used in the training)

DS-GAN Results:



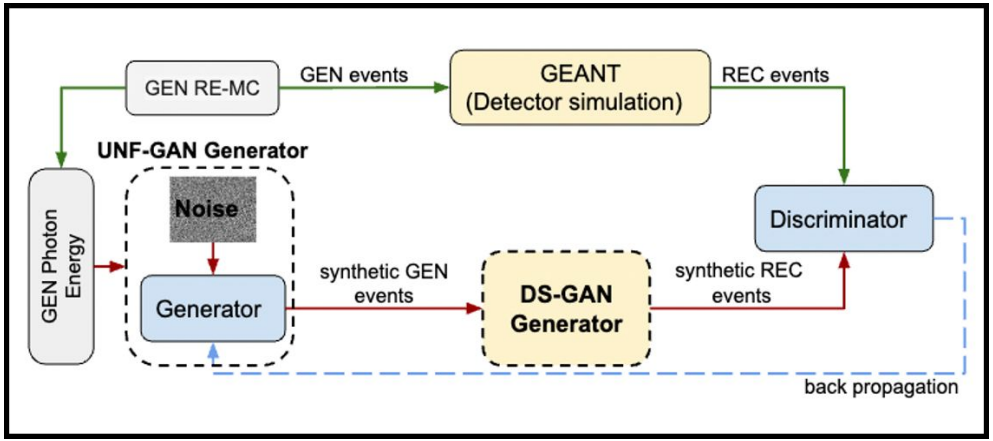
UNF-GAN Results:

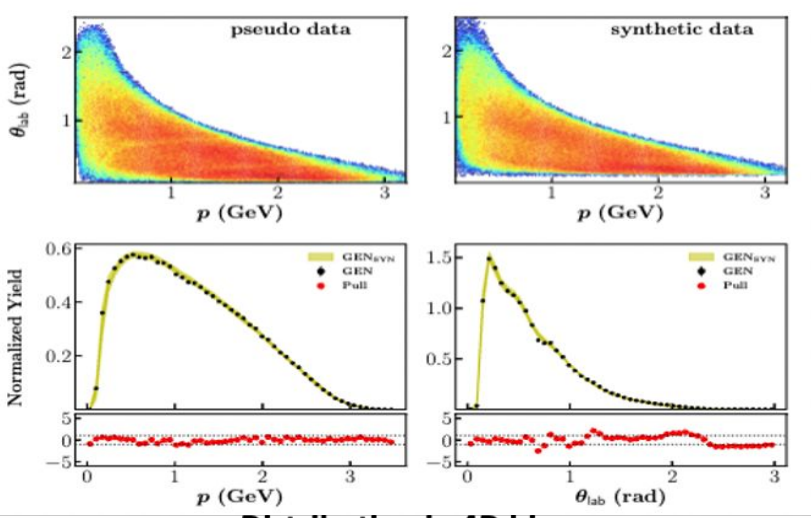


- The comparison is extended to other physics-relevant distributions not used in the training and derived from the four training variables.
- The agreement, quantified by the pull distributions shown at the bottom of each plot, is remarkable, in both cases, with most of the points lying within ±1σ. This indicates that the DS-GAN is indeed able to learn the CLAS detector effects.

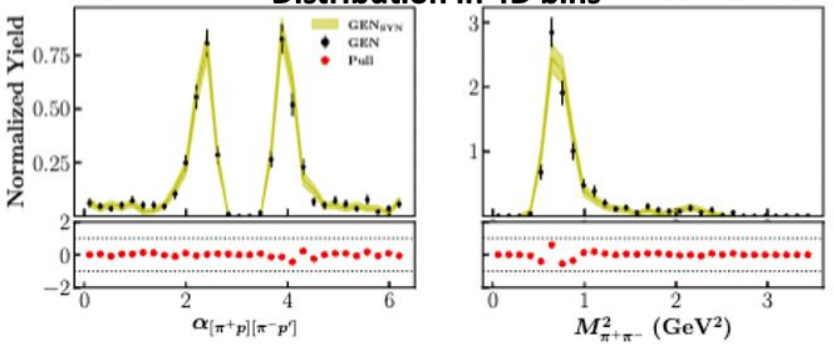
2π photoproduction closure test

- Comparison of GEN and GENSYN for p momentum components and θlab in the laboratory reference frame using RE-MC data.
- We compare 1D distributions in a given bin of the other variables.
- The success of this test shows that correlations underlying the multidifferential cross section are correctly reproduced in the synthetic datasets.





Distribution in 4D bins



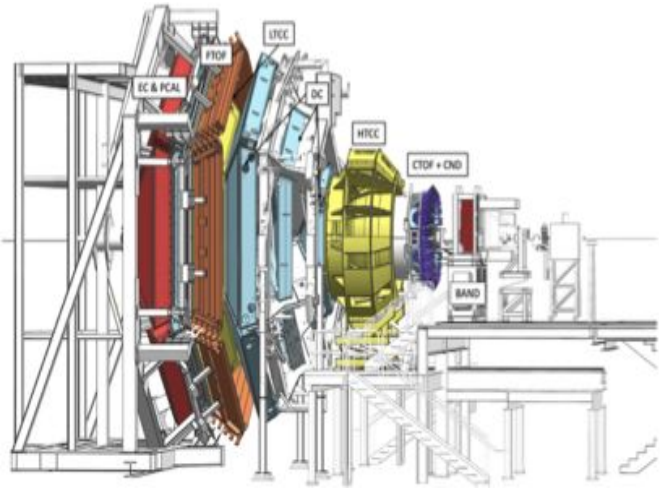
1D histograms for fixed slice of the other variables ($2.55 < W < 2.60$ GeV, $-0.7 < t_{\pi^+} < -0.3$ GeV^2, $2.5 < M^2_{p\pi^-} < 3.3$ GeV^2). Left panel: $\alpha_{[\pi^+p][\pi^-p']}$ distribution for $0.6 < M^2_{\pi^+\pi^-} < 0.9$ GeV^2. Right panel: $M^2_{\pi^+\pi^-}$ distribution for $2.0 < \alpha_{[\pi^+p][\pi^-p']} < 2.5$.

Good agreement ($\pm 1\sigma$) for lab variables and in 4D bins

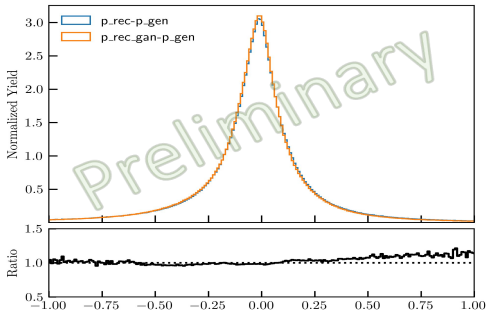
A(i)DAPT
AI for Data Analysis and PreservaTion

Moving forward: CLAS12 application

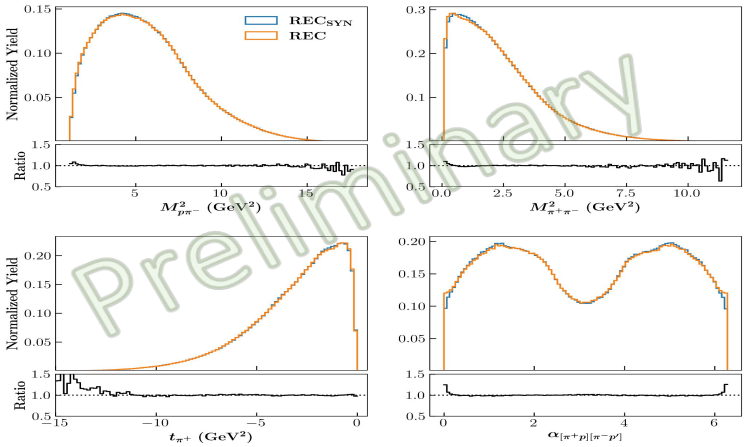
- Working towards the application of the developed machinery to CLAS12 pseudata for the $ep \rightarrow e'p'\pi^+\pi^-$
- If this procedure works well on CLAS and CLAS12 data, the robustness of the architecture is guaranteed.
- Then we can put together in a coherent way information from different kinematic regions

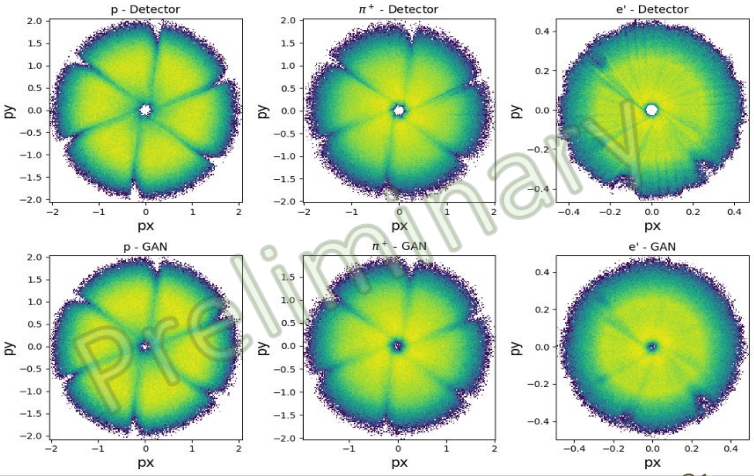


CLAS12 resolution



REC SYN vs REC pseudodata training variables



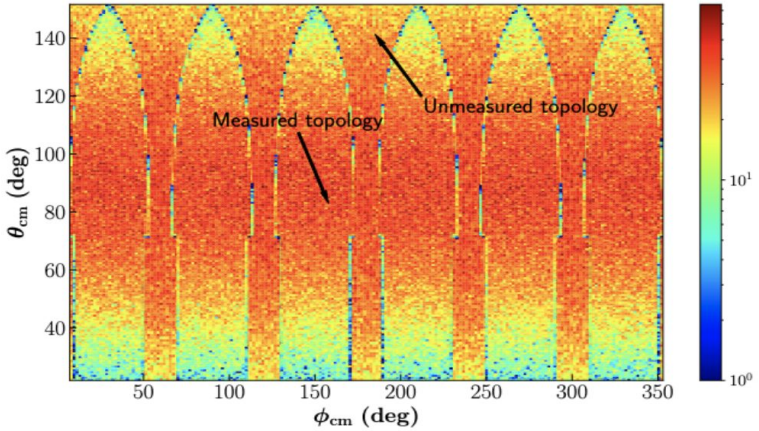


Credit: Derek Glazier, Marco Spreafico

Moving forward: Acceptance Effects

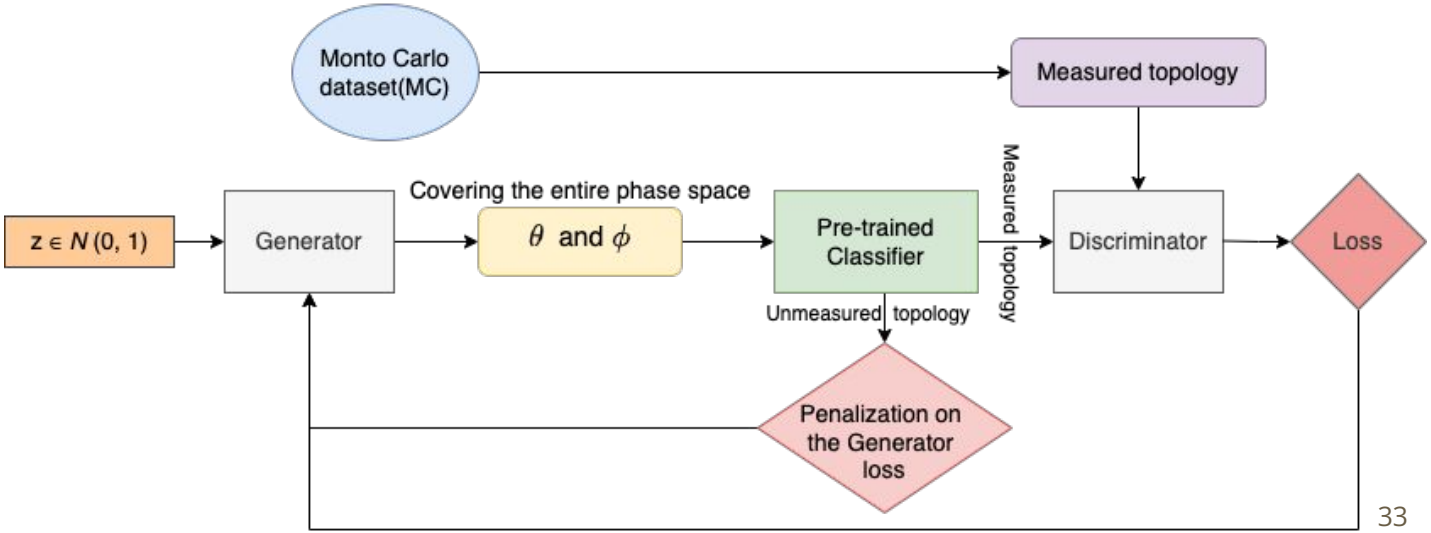
- Acceptance refers to the geometric and efficiency-related limitations of the detector system. Due to the detector's limited area coverage, particles produced in certain directions may not be detected.

- We generate Monte Carlo (MC) pseudo-data based on the photoproduction reaction $\gamma p \to \pi^0 p$ in a kinematic range where $E_{cm} \sim 1.34 GeV$
- Two independent variables (at fixed energy): θ_cm and ϕ_cm, along with their associated topologies.
- Each event is represented as a three-dimensional vector, where the first two elements denote the angles and the third element represents the topological state (0 as unmeasured or 1 as measured).
- The two different classes of events, which can be seen in this plot, have been produced passing the reaction output through a simple proxy of the CLAS detector, which is able to tell if a given event has been measured or not.



Unfold Acceptance Effects in Particle Detectors with Generative AI

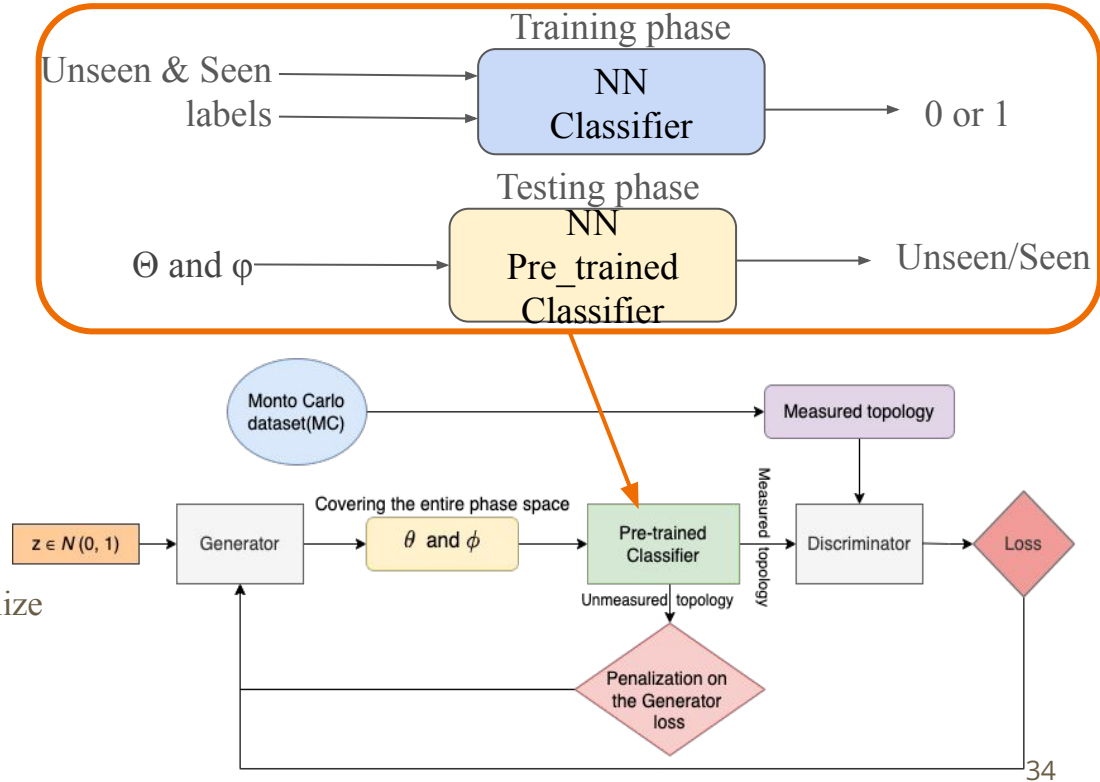
- To address the challenge of acceptance effects in detector systems
- The architecture is designed to learn the underlying physics distribution and generate events in both measured and not measured regions.
- A pre-trained classifier is used to categorize these events according to their topologies.
- A key aspect of our framework is the classifier and the inclusion of a penalty in the loss function, which penalizes the generator to avoid a surplus of events in the undetected region (unseen).
- This approach allows the framework to learn collectively from the detected data (training data) and the penalization term.



Classifier:

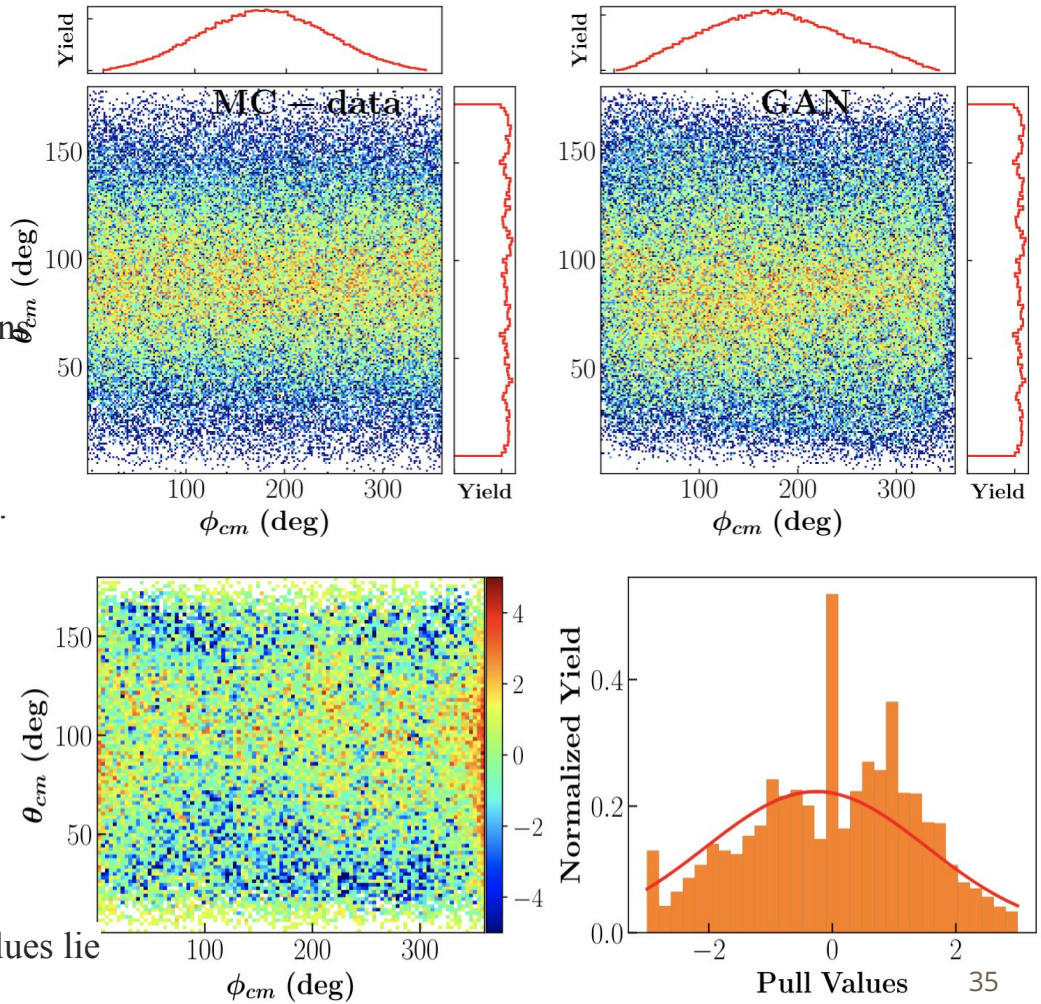
- We trained a binary classifier that can distinguish between different topological states.
- The training dataset consists of two sets:
  - "unseen" data representing events that have not been measured
  - "seen" data representing events that have been measured.
  - The pre-trained classifier will be integrated into our GAN architecture.

- The GAN will produce Θ and φ angles and utilize the classifier to classify the events into the appropriate topological states.
- A custom generator loss function is employed to penalize the generator based on the difference between the generated and true unmeasured topology event distributions.



Results

- MC-data at the entire phase space Θ and φ distributions are compared to the synthetic data produced by the GAN.
- The histograms above and to the right of each scatter plot provide distributions of θ_{cm} and φ_{cm}, respectively.

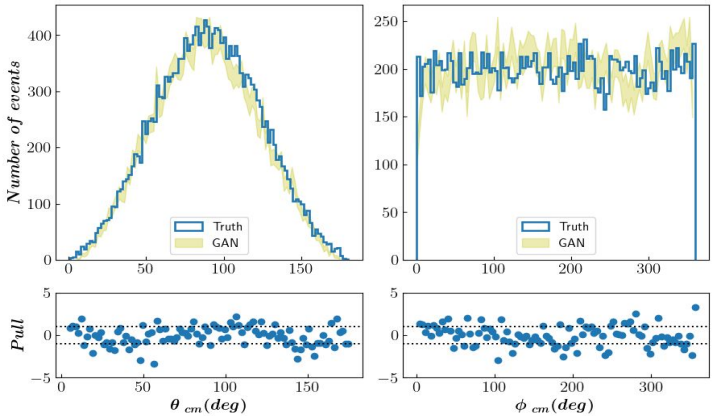


We can see from the 2D pull distributions, most of the values lie within ±2.5σ

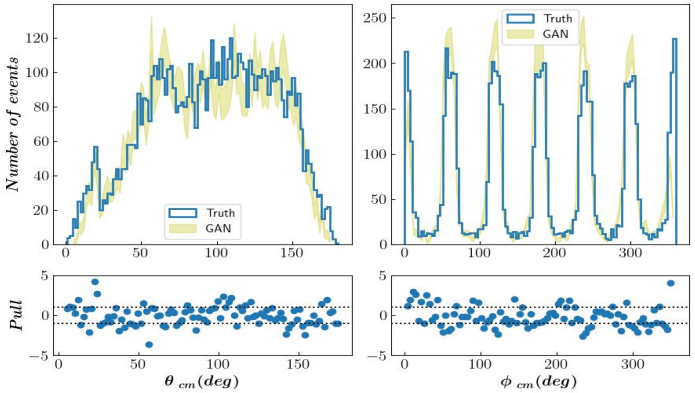
Results:

The comparison between θ_{GAN}, φ_{GAN} and θ_{MC} and φ_{MC} distributions.

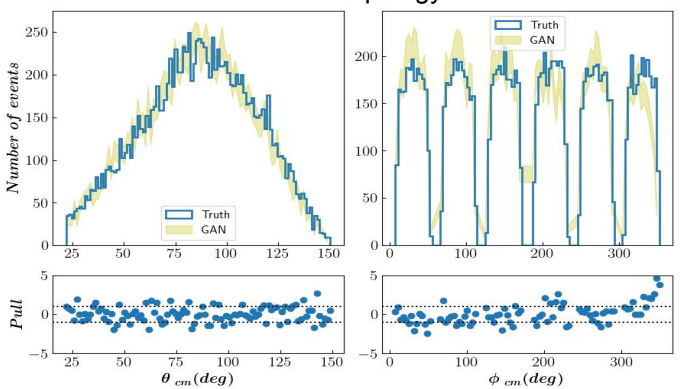
A further comparison has been made between the variables after classifying them into the two topologies.



Unmeasured topology



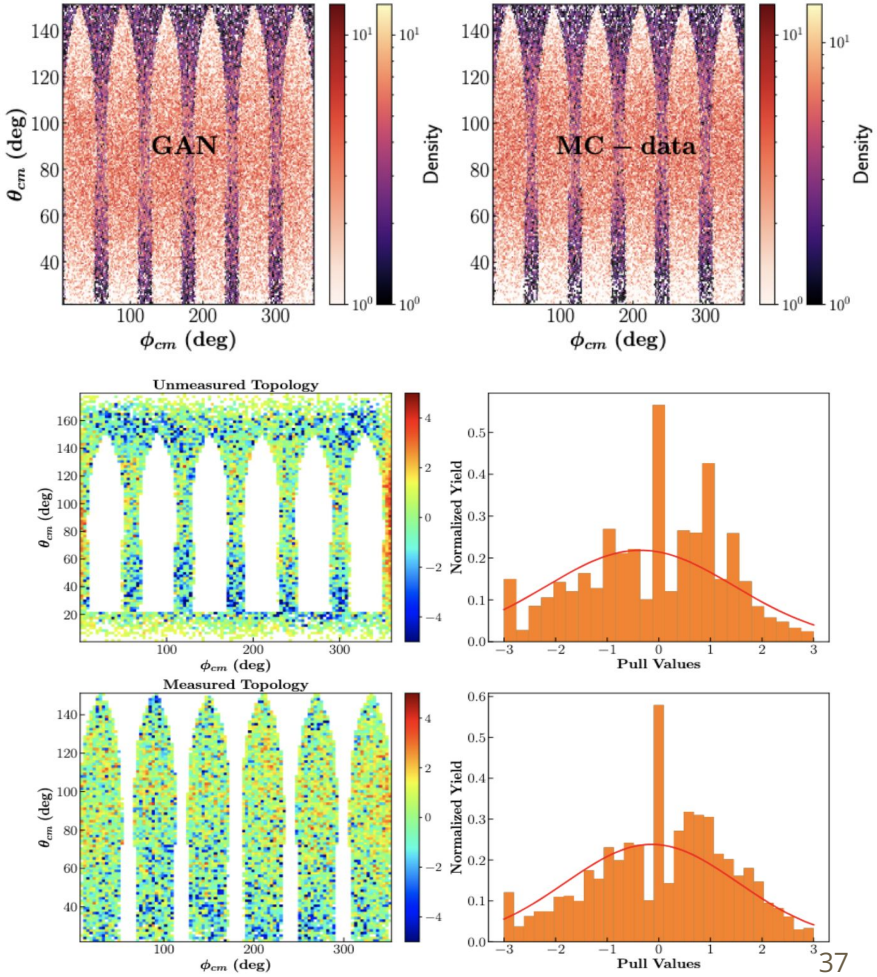
Measured topology



We can see from the pull distributions, most of the values lie within ±2.5σ

Results:

- We show here all topologies, both measured and unmeasured regions after taking the θ_{cm} and φ_{cm} variables generated by GAN and classify them to measured and unmeasured regions.
- Evaluated using the calculated pull values shown here. In the left column, the 2D distributions display the pull values for measured and unmeasured topologies.
- Our approach effectively replicates the overall patterns and features of the true data, including the distinction between measured and unmeasured parts



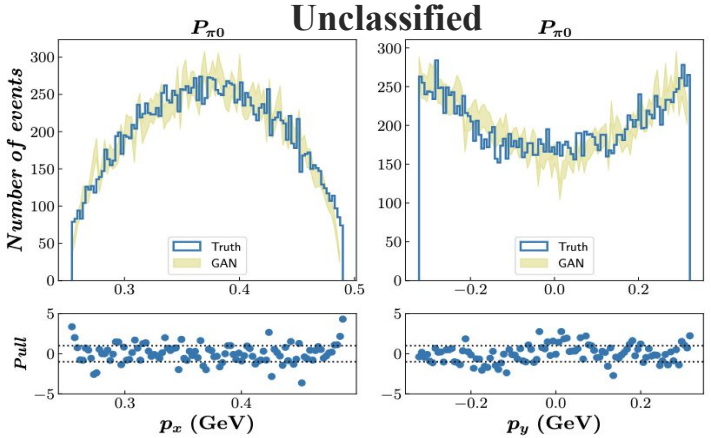
Results:

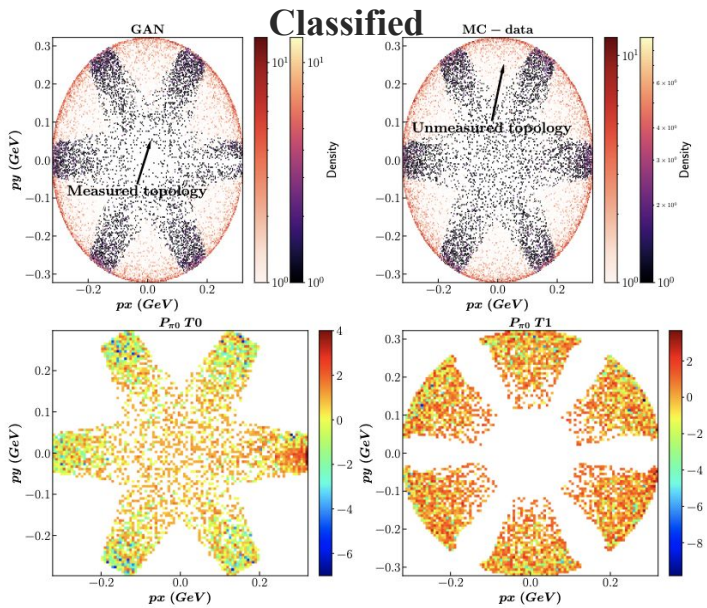
Derived variables (not used in the training)

- The comparison is extended to other physics-relevant distributions not used in the training and derived from the two variables θ and φ, namely (E, p_x, p_y, p_z) for the outgoing π0 in the laboratory frame.

- We applied the transformation on both **unclassified** and **classified** data to obtain the four-momentum components for all topologies.

- Comparing these transformed variables will help us assess how well the GAN preserves the underlying physics and correlations present in the MC data, beyond the initial θcm and φcm variables used during training.
- The good agreement and preservation of correlations remain valid for derived kinematic variables that were not used for training.

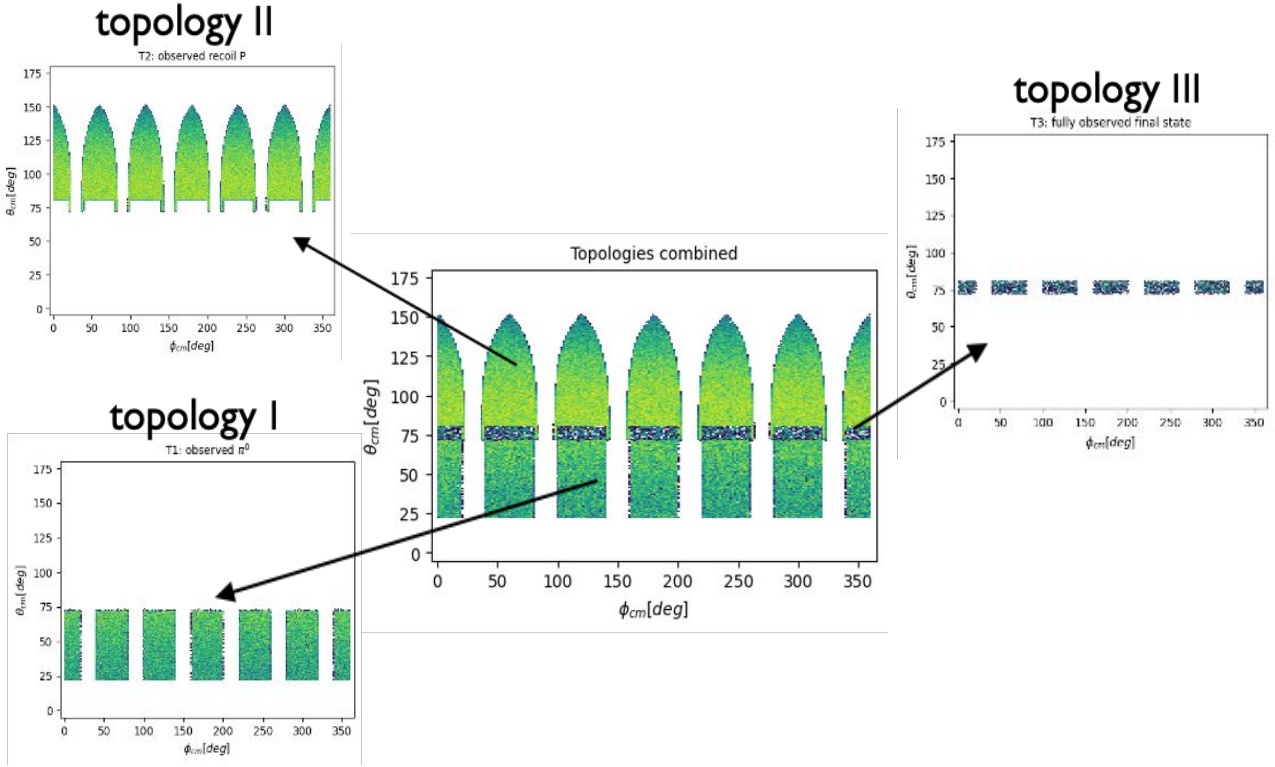




Moving forward: Acceptance(Multiple Topologies)

Work in progress

Build a single GAN that is able to generate in the full phase space according to the correct distribution



Credit: Tommaso Vittorini

Conclusion

• Surrogate Event Generator: Replaces the inverse CDF sampler for physics-based components.
• Surrogate event generator produces events matching the target distribution.
• Outer-GAN: accurately infers the correct parameters from the given events.
• It is more computationally efficient than physics-based samplers.

- We performed a positive closure test on 2pion photoproduction
- We demonstrated that GANs are a viable tool to unfold detector effects (smearing) to generate a synthetic copy of data
- We demonstrated that the original correlations are preserved
- Preserve data in alternative compact and efficient form

- We demonstrated that our approach is an effective tool for correcting detector effects (acceptance).
- We have shown the GAN's ability to recover the distribution of unmeasured topologies, despite being trained solely on measured data.

• The derived variables, such as four-momentum components, has highlighted the GAN's proficiency in preserving underlying physics beyond the variables explicitly used during training.

• We are aiming to develop a single GAN capable of generating events across multiple topologies while also associating each event with the probability of belonging to a specific topology.

Thank you!