# Toward a generative modeling analysis of CLAS exclusive 2π photoproduction

Tareq Alghamdi

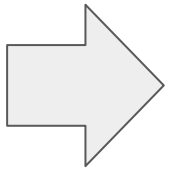
Computer Science Department

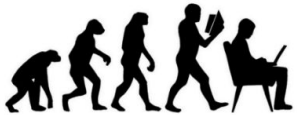
Old Dominion University

OLD DOMINION
UNIVERSITY

- 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).

**Can AI support NP/HEP experiments to extract physics from data in a more efficient way?**






Collaborative effort (regular meeting)
- ML experts (ODU, Jlab)
- Experimentalists (Jlab Hall-B)
- Theorists (JPAC, JAM)

**Develop AI – supported procedures to:**
- **Accurately fit data in multiD space**
- **Unfold detector effects**
- **Compare synthetic (AI-generated) to experimental data**
- **Quantify the uncertainty (UQ)**

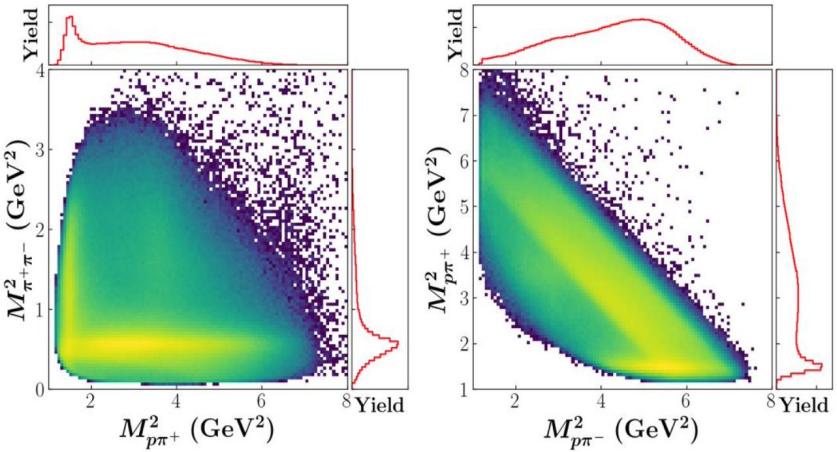
OLD DOMINION UNIVERSITY Jefferson Lab JPAC Theory JAM Computer Science Experiment

# Exclusive reactions: $2 \rightarrow 3$

- $\gamma p \rightarrow \pi^+\pi^- p$ (unpolarized)

- Initial state: Fully known
- Final state: 3x3 independent variables
- Independent variables: (3x3) – 4 = 5 ($E\gamma$ fixed)
- Many possible choices, such as $M2\ \pi\pi$, $M2p\pi$ , $\theta\ \pi$, $\alpha$, $\phi$

CLAS g11 $2\pi$ photoproduction




- $E\gamma$ =(3 −3.8)$GeV$
- Dataset analysis on $\gamma p \rightarrow p\pi^+(\pi^-)$ with small contamination from $\gamma p \rightarrow p\pi^+$ (more than a single missing $\pi^-$)
- Complicated dynamics due to the overlap of (p$\pi$) to form Δ baryon resonances and ($\pi\pi$) to form meson resonances

AI could provide a new way to look at data and to extract observables and physics interpretation


Alghamdi T, Alanazi Y, Battaglieri M, Golda AV, Blin AH, Isupov EL, Li Y, Marsicano L, Melnitchouk W, Mokeev VI, Montaña G. **Toward a generative modeling analysis of CLAS exclusive 2 π photoproduction.** Physical Review D. 2023 Nov 21;108(9):094030.

# Detector unfolding

• Detector effects make measured observables (detector-level) different from the ‘true’ observables (vertex level).

**Acceptance:** Any measurement can access only a limited portion of the phase space. What can we say about these unmeasured regions?

- ➢ Interpolation: deal with the holes in the phase space
- ➢ Extrapolation: extend our coverage from the borders of measured regions

**Resolution**: Any measurement has an experimental resolution that may modify cover-up effects that we’re looking for

- ➢ Spikes may be concealed behind the detector resolution
- ➢ Measurements could be extended to unphysical regions

• Mitigation strategy:

- ➢ Acceptance: *‘Fiducial volumes’ to exclude unmeasured regions and extend the covered measured of the phase space*
- ➢ **Resolution: *build and validate ML-models to unfold resolution effects***

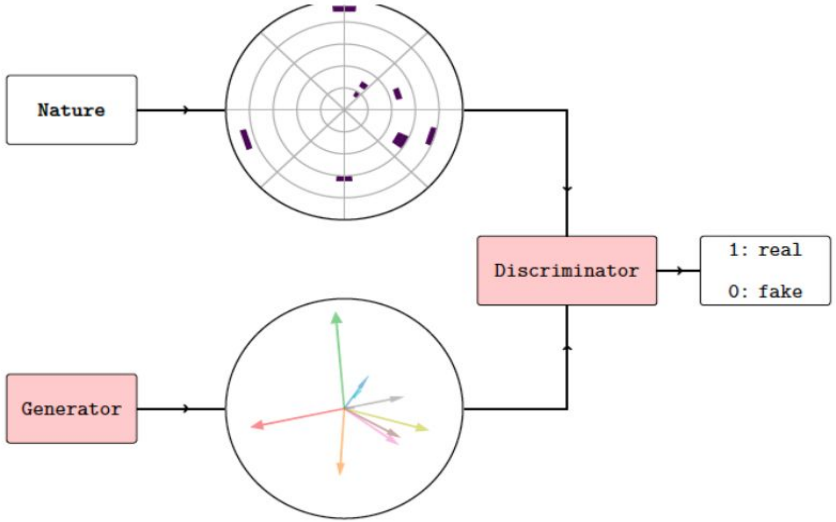
# Generative Adversarial Networks

• A generative model built upon the competition between two neural networks: the **Generator** and the **Discriminator**.

• Discriminator is trained to discern real (nature) from synthetic (generated) events.
• Generator is trained to create events as close as the nature, in order to fool the discriminator.

• The Generator is employed to preserve high-dimensional correlations, acting as detector proxies.

• Generator can be used to provide highly realistic pseudo-data in an extremely fast way.



## GAN variations

- There are many variations of GANs, for example:
  - **Conditional GAN (CGAN):**
    - Condition the model on labels.
  - Deep Convolutional GAN (DCGAN):
    - Uses deep CNN layers.
  - Wasserstein GAN (WGAN)
    - Replaces the discriminator by a critic and uses Wasserstein loss.
  - **Least Squares GAN (LS-GAN):**
    - Replaces the loss function for the discriminator from binary cross entropy to a least squares loss.

# 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**

# $2\pi$ 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 $2\pi$ photoproduction model (RE-MC GEN pseudodata)

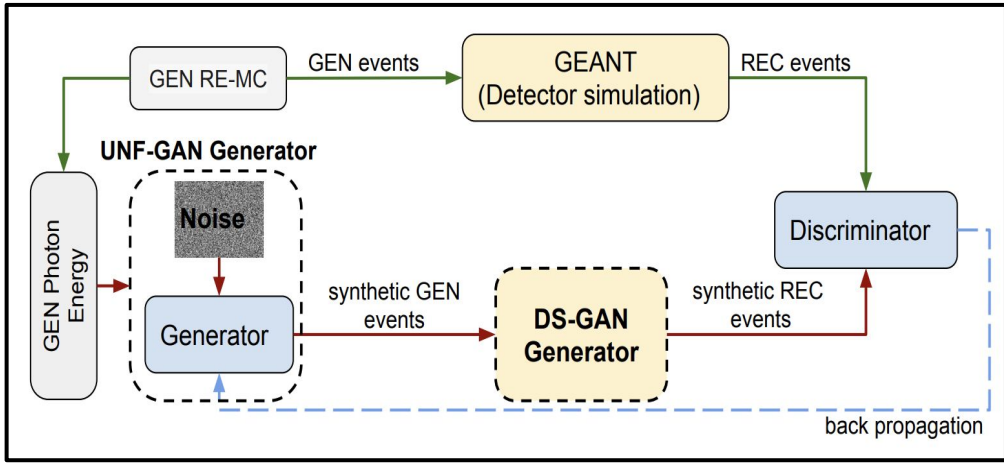
2. Apply detector effects (acceptance and resolution) via GSIM-GEANT (RE-MC REC pseudodata)

3. Deploy a secondary GAN (DS-GAN) to learn detector effects using an independent 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. Future work: Replace RE-MC REC pseudo data with CLAS data in the training to unfold the vertex-level experimental distributions*

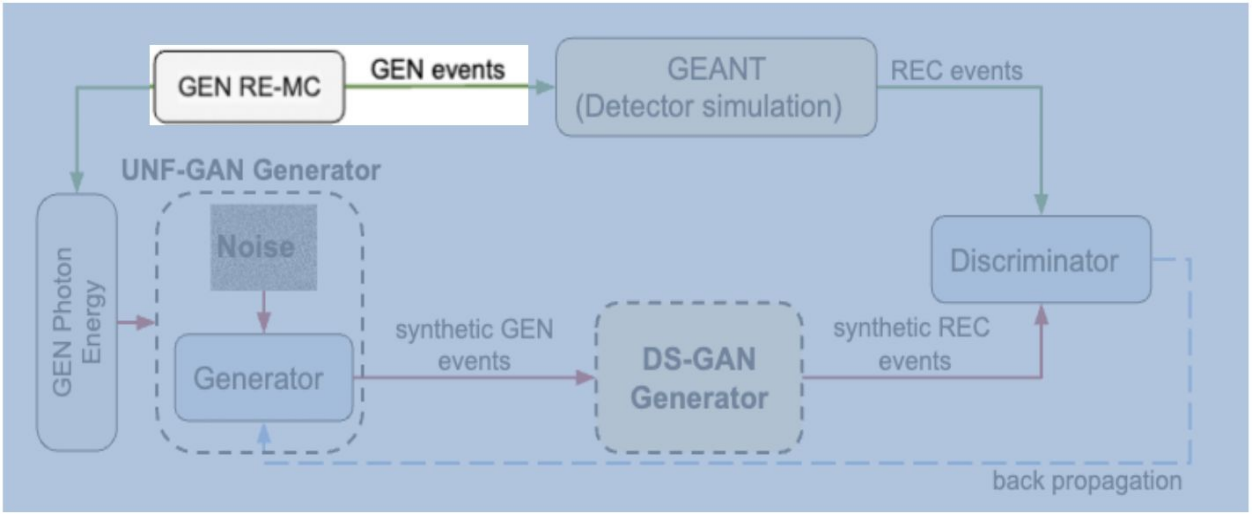



Alghamdi T, Alanazi Y, Battaglieri M, Golda AV, Blin AH, Isupov EL, Li Y, Marsicano L, Melnitchouk W, Mokeev VI, Montaña G. Toward a generative modeling analysis of CLAS exclusive 2 π photoproduction. Physical Review D. 2023 Nov 21;108(9):094030.

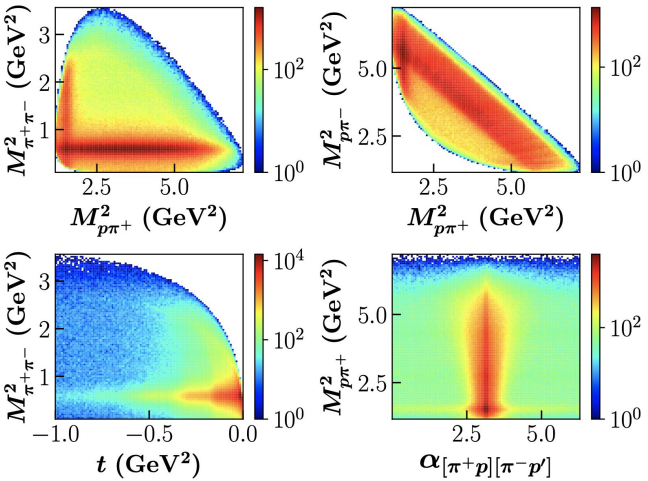
# $2\pi$ photoproduction closure test

1. Generate events with a (realistic) Monte Carlo $2\pi$ 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($\varrho^0$, $\Delta^{++}$, $\Delta^0$ , a contact term)



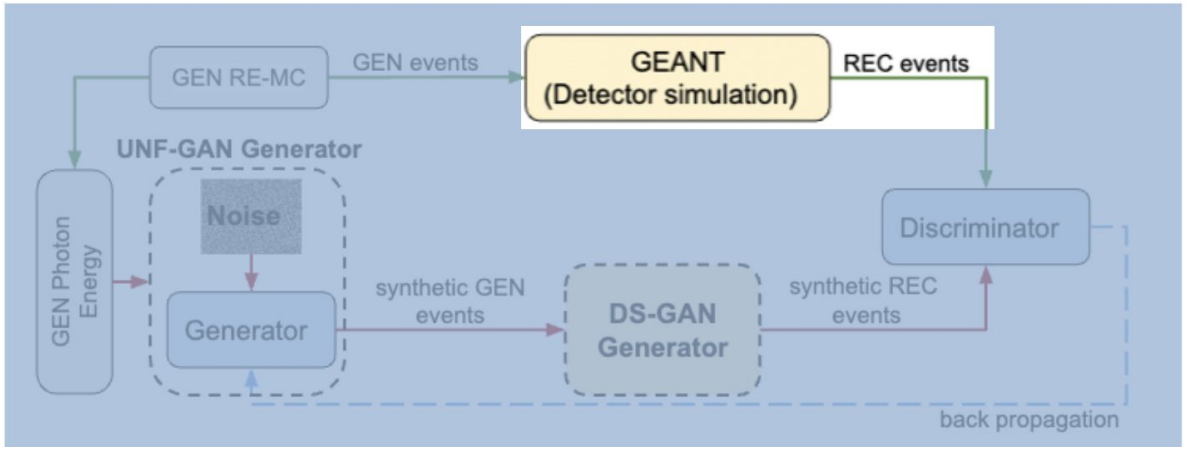

GEN events

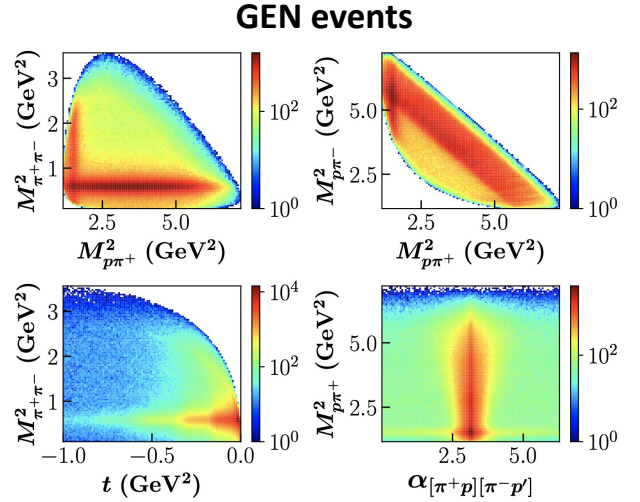


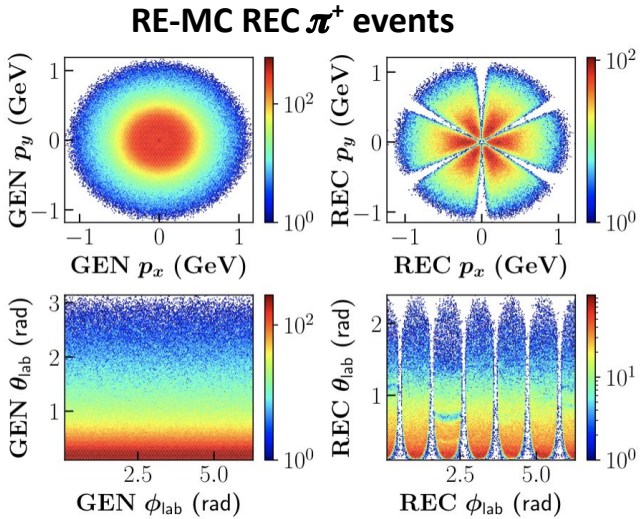
# $2\pi$ photoproduction closure test

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

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

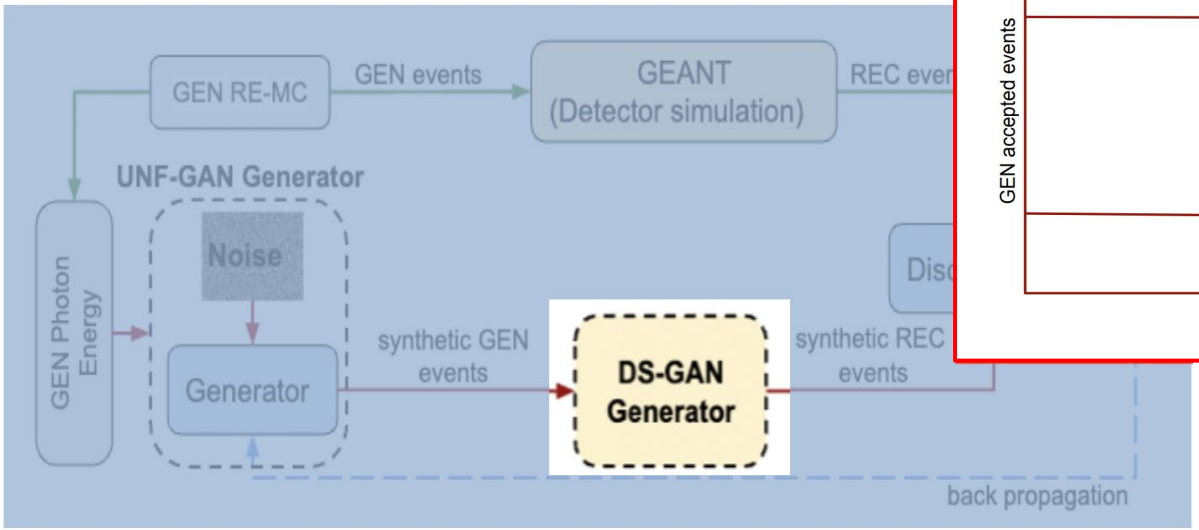


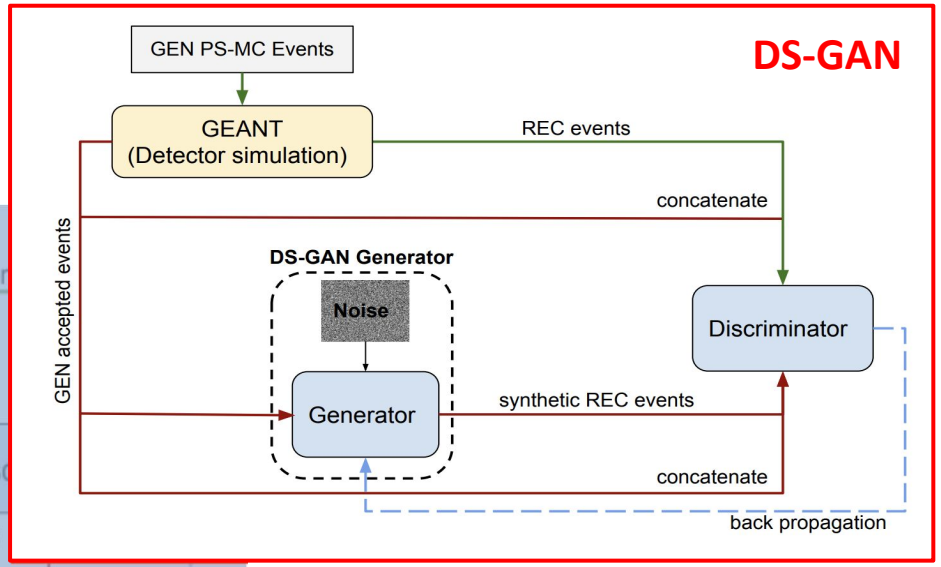



# $2\pi$ photoproduction closure test

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

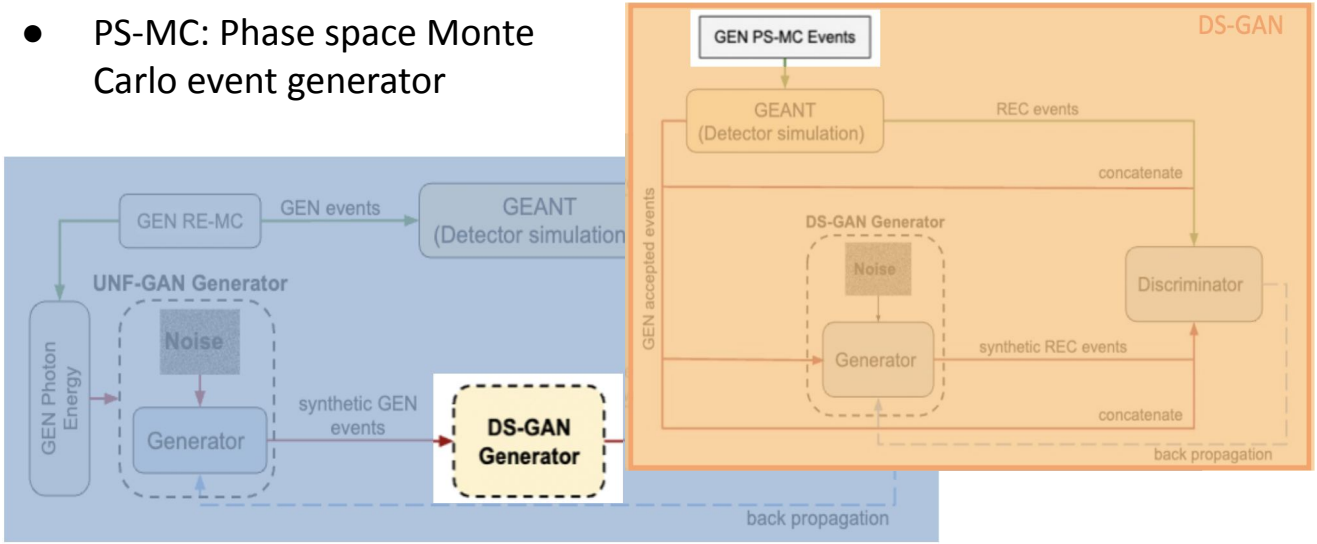


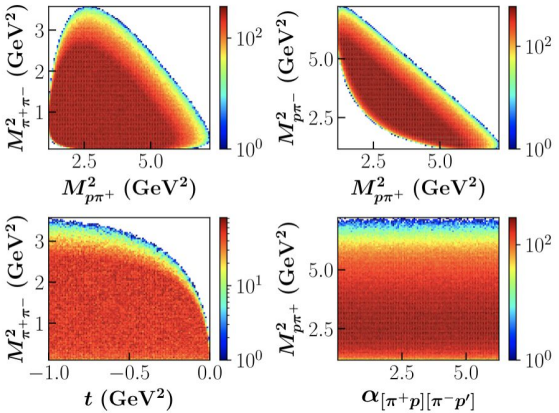
# $2\pi$ photoproduction closure test

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

- PS-MC: Phase space Monte Carlo event generator



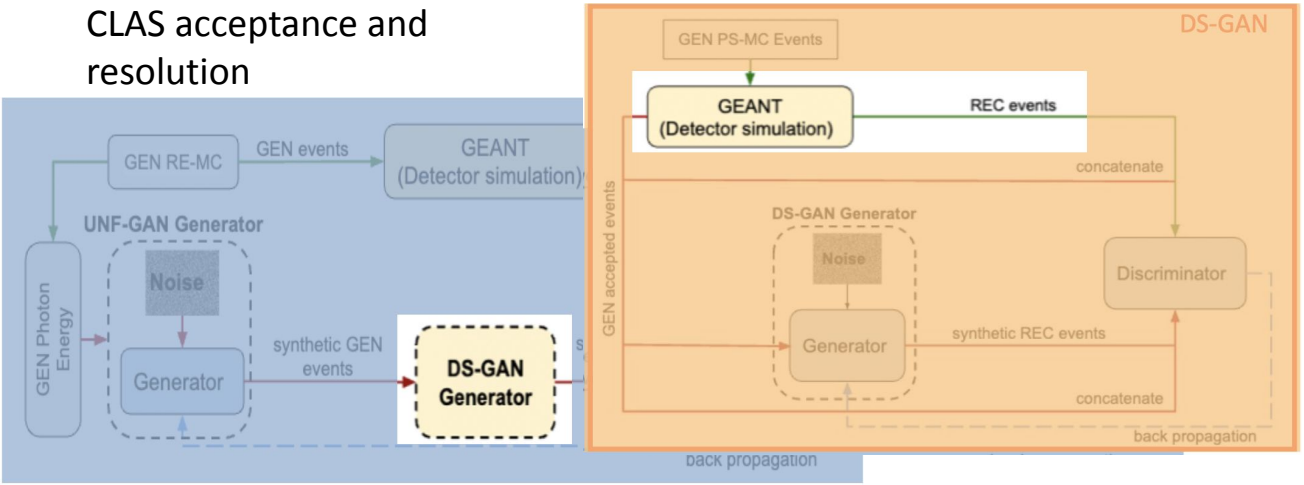

**PS-MC GEN events**



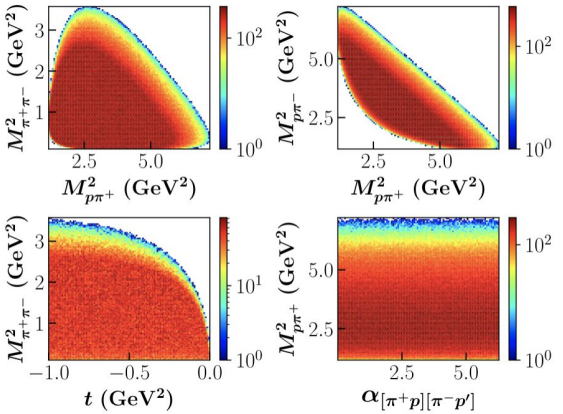
# $2\pi$ photoproduction closure test

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

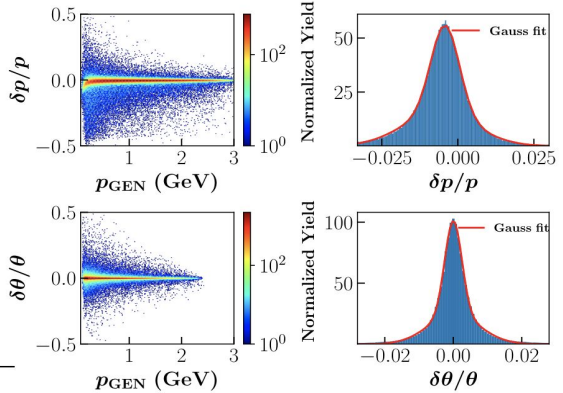
- GSIM-GEANT to simulate CLAS acceptance and resolution




PS-MC GEN events




CLAS resolution on $\pi^+$ kin. variables



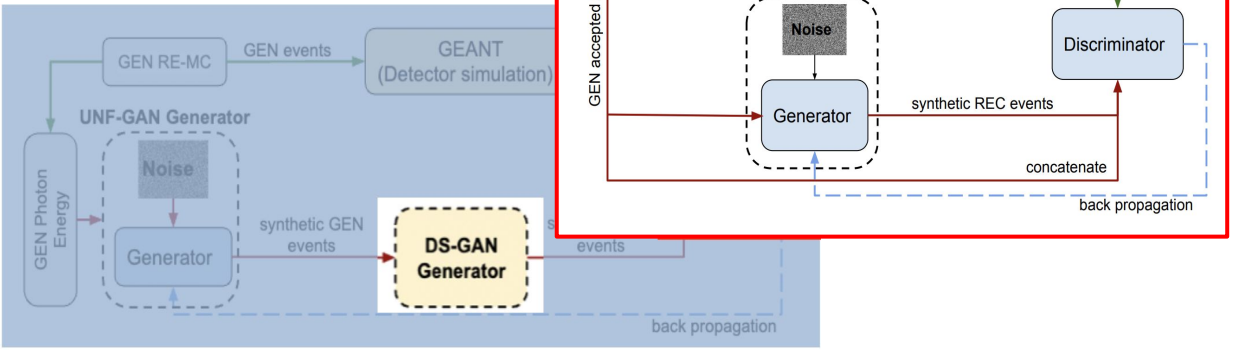

Another way to check the smearing is to consider all momenta $P = \sqrt{p^2_x + p^2_y + p^2_z}$

# DS-GAN results

- Uncertainty quantification via pull calculation: Bootstrap with 20 independently trained GANs

- Pull calculation for each bin:

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

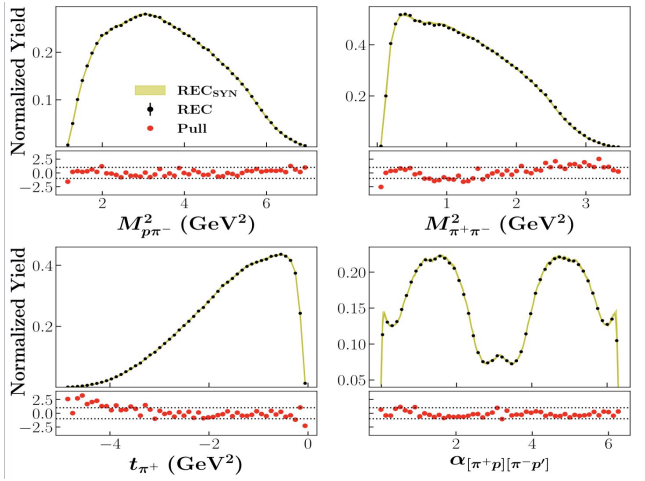




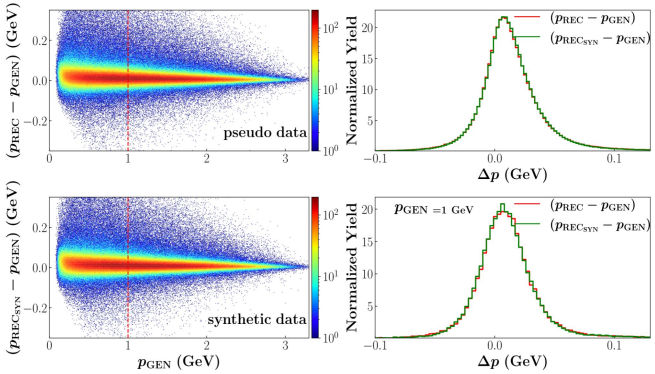


DS-GAN learned the CLAS detector effects!

**MC REC pseudodata vs. DS-GAN synthetic data**



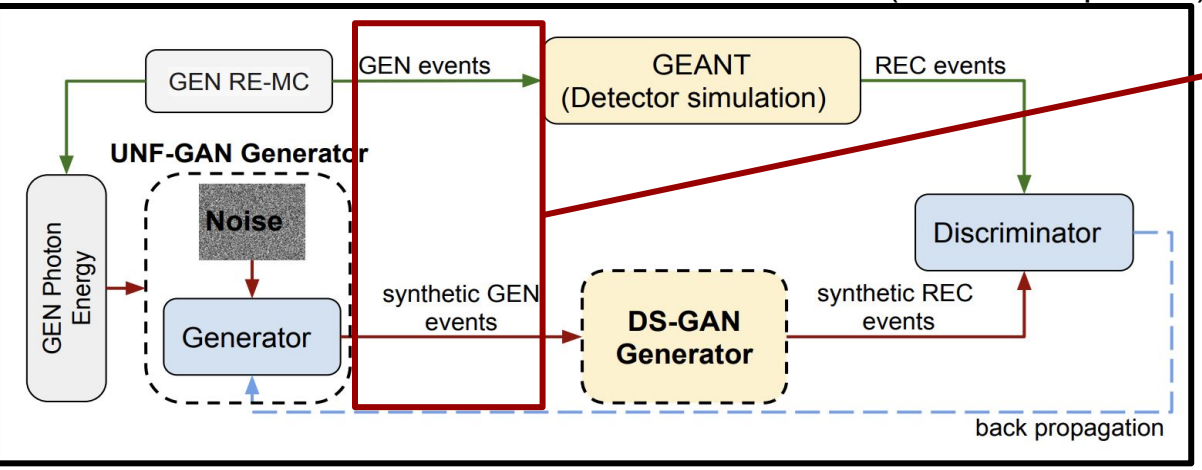

**CLAS resolution**



# UNF-GAN results

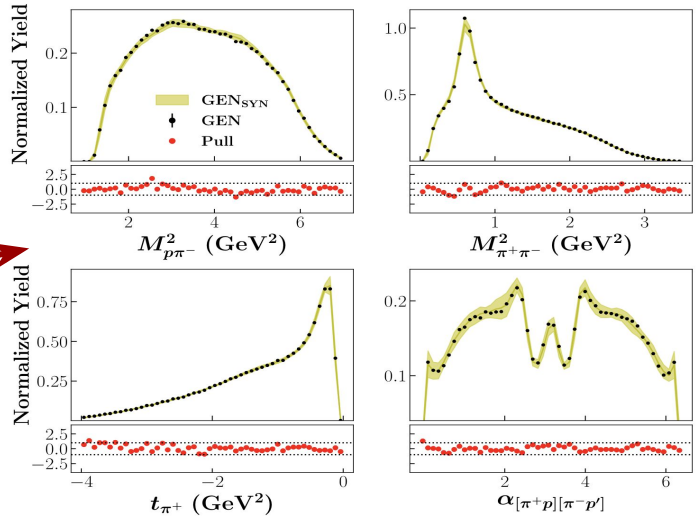
4. Deploy the unfolding GAN (UNF-GAN) that includes the DS- GAN, and train it with RE-MC REC pseudodata

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




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

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



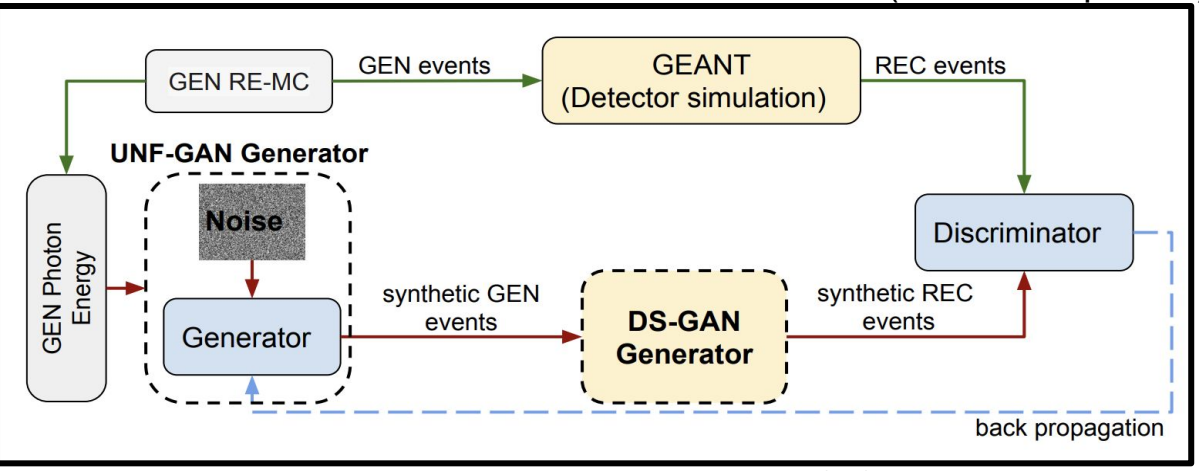

Systematic of the full procedure (two-GANs) estimated by bootstrap with 20+20 independently trained GANs

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

# UNF-GAN results

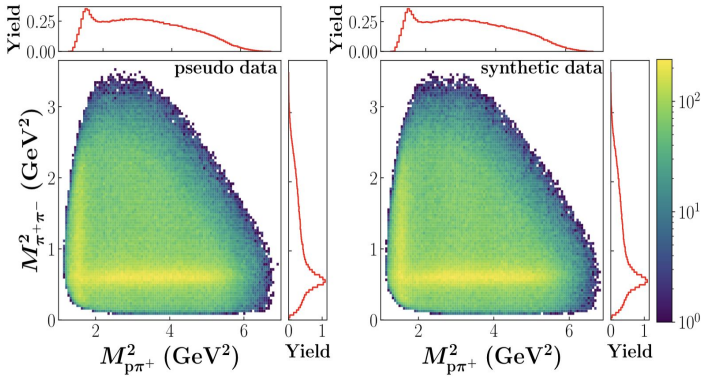
4. Deploy the unfolding GAN (UNF-GAN) that includes the DS- GAN, and train it with RE-MC REC pseudodata

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

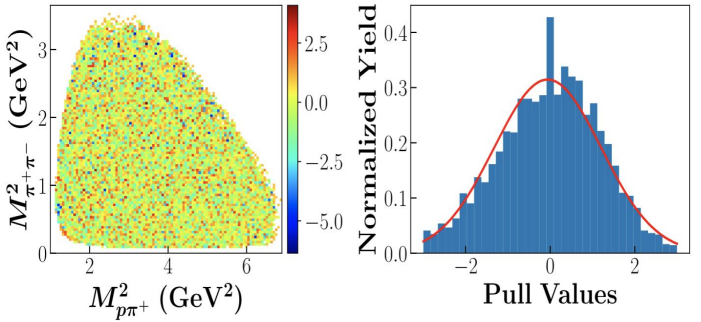



5. Compare UNF-GAN GEN SYNT to RE-MC GEN pseudodata

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




**2D pull**

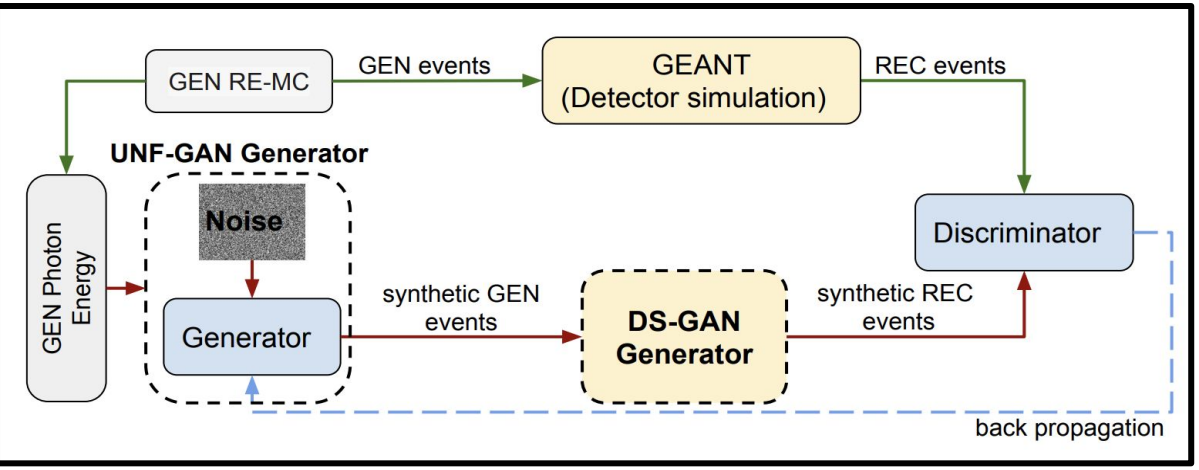



Good agreement (±1$\sigma$) for vertex-level training variables!

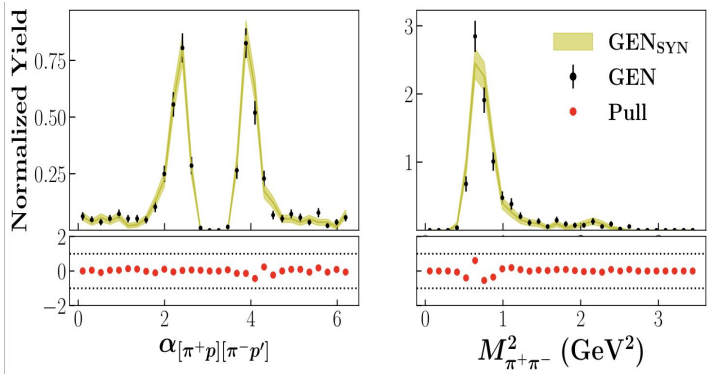
# UNF-GAN results

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

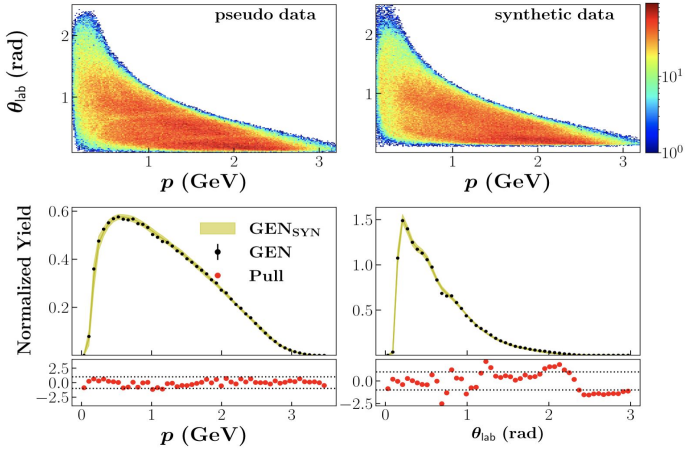
- UNF-GAN trained with REC-MC pseudodata (experimental data proxy)
- DS-GAN used to unfold CLAS detector effects (within acceptance)




**Distribution in 4d bins**



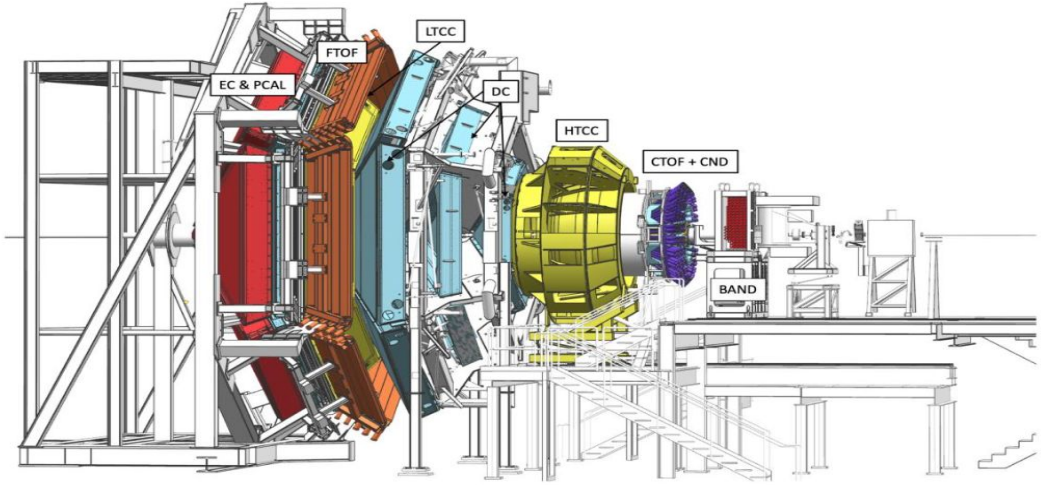

**RE-MC GEN pseudodata vs. UNF-GAN SYN data**




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

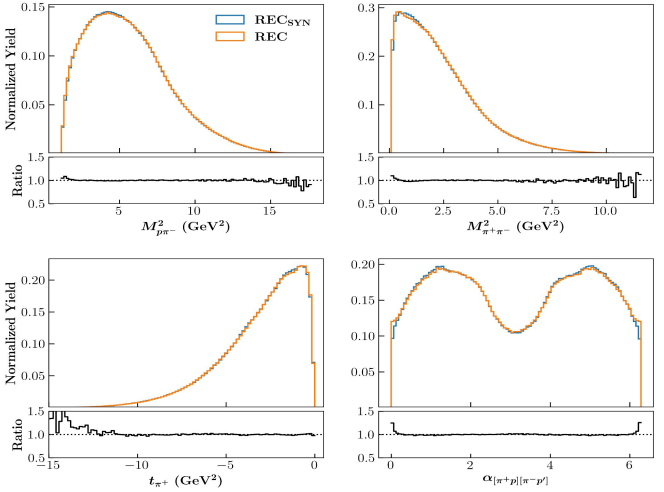
# CLAS12 Application

- Working towards the application of the developed machinery to CLAS12 pseudodata
- If this procedure works well on CLAS and CLAS12 data the architecture robustness is guaranteed
- We can put together in a coherent way information from different kinematic regions

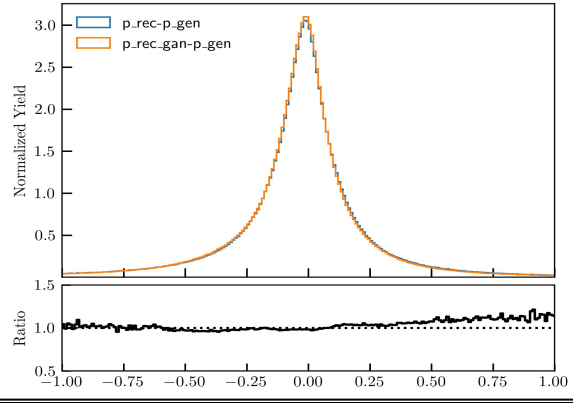



## Preliminary results

REC SYN vs REC pseudodata training variables




CLAS12 resolution



# Summary

- 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 an alternative compact and efficient form
- The uncertainty quantification of the entire procedure was assessed by combining a bootstrap for the two GANs
- For more information about this work, please refer to: https://journals.aps.org/prd/abstract/10.1103/PhysRevD.108.094030

**We are working on:**

- Quantifying the systematic error introduced by the detector acceptance
- Implementing this architecture into JLAB software in order to make it easily available to everyone
- Make this procedure an efficient way to analyze CLAS12 $2\pi$ data

*While there is still progress to be made in efficiently using AI to extract physics from data, we are moving in the right direction!*

# ACKNOWLEDGMENTS

A(i)DAPT
AI for Data Analysis and PreservaTion

**Collaborators:**
Y. Li (ODU), M. Battaglieri (JLab/INFN), N. Sato (JLab), A. Pilloni (INFN), W. Melnitchouk (JLab), L. Bibrzycki (AGH), V. Mokeev (JLab), L. Marsicano (INFN), E. Isupov (SINP), Y.Alanazi(JLab), A. Hiller (JLab), A. Golda (JLab), A. Szczepaniak(JLab), T. Vittorini (INFN)

- We thank J. Qiu for helpful discussions.
- We thank the CLAS Collaboration for providing the two-pion dataset this work is based on.
- This work was supported by the Jefferson Lab LDRD project No. LDRD19-13 and

No. LDRD20-18, and in part by the U.S. Department of Energy contract DE-AC05-06OR23177, under which Jefferson Science Associates, LLC, manages and operates Jefferson Lab. ANHB is supported by the DFG through the Research Unit FOR 2926 (project number 409651613).

- T.Alghamdi was supported by a Ph.D. scholarship from Al-Baha University, Saudi Arabia.
- The work of N.Sato was supported by the DOE, Office of Science, Office of Nuclear Physics in the Early Career Program. This work contributes to the aims of the U.S. Department of Energy ExoHad Topical Collaboration, contract DE-SC0023598.

# Thank you!

# Backup slides

# Event generators:

- **Monte Carlo Event Generators:**
  - Pythia
  - Herwig
  - Others

- **Machine Learning based event generator:**
  - Generative Adversarial Networks (GANs)
  - Variational Autoencoders (VAEs)
  - Normalizing Flows (NFs)

# Motivation:

**Why do we need machine learning?**

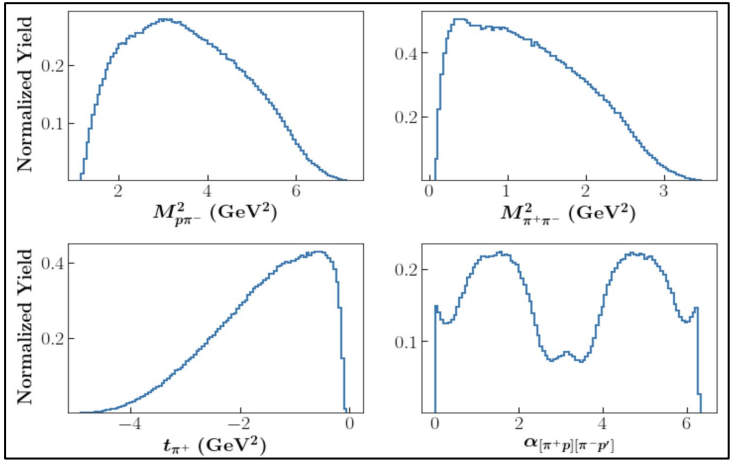
- Machine learning based has several attractive advantages over Monte-Carlo based, for example:
  - Machine learning based eliminates the theoretical assumptions.
    - Capture more complex and realistic correlations.
    - The ability to capture a wide range of correlations present in the training data
  - Improved speed and efficiency
    - Can generate millions of events within seconds.
  - Reduced need for multiple repositories.
    - Once the MLEG model is trained, it can serve as a single generator for producing events.

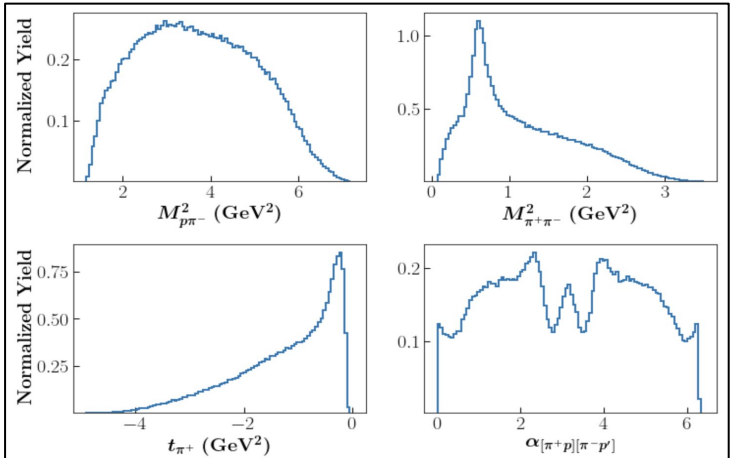
# Data descriptions: PS-MC GEN, RE- MC GEN Events:

- **Data is generated using two different Monte Carlo event generators: PS-MC and RE-MC**

- **Folding GAN:**
  - **PS-MC: Phase-Space Monte Carlo dataset**

- **Unfolding GAN**
  - **RE-MC: Realistic Monte Carlo dataset**



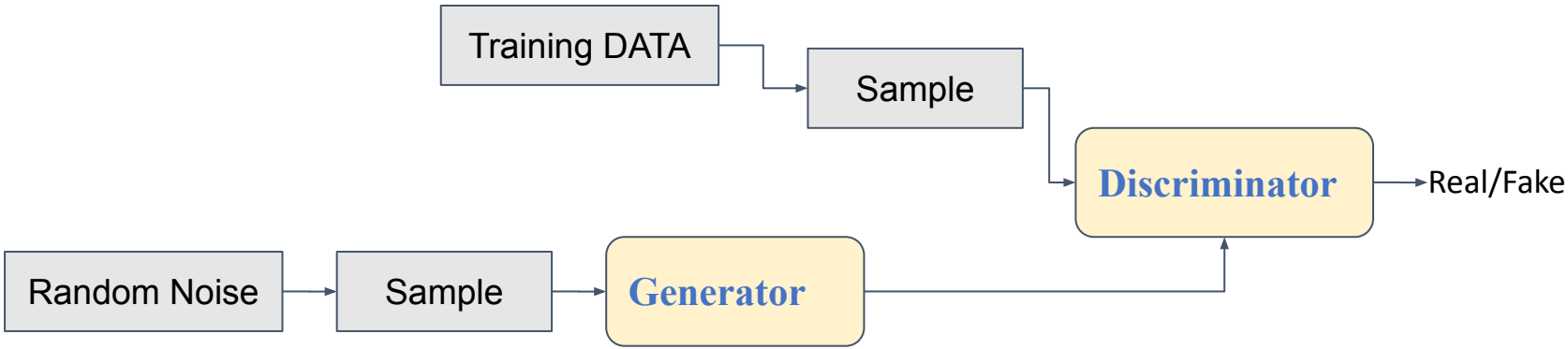

*GEANT is a system of detector description and simulation tools that help physicists in such studies.

# What are Generative Adversarial Networks (GANs)?

- GAN* is a class of machine learning frameworks that typically consists of two adversarial NNs: a Generator & a Discriminator.

- **The Generator takes random noise as inputs and is trained to generate synthetic samples from the problem domain.**

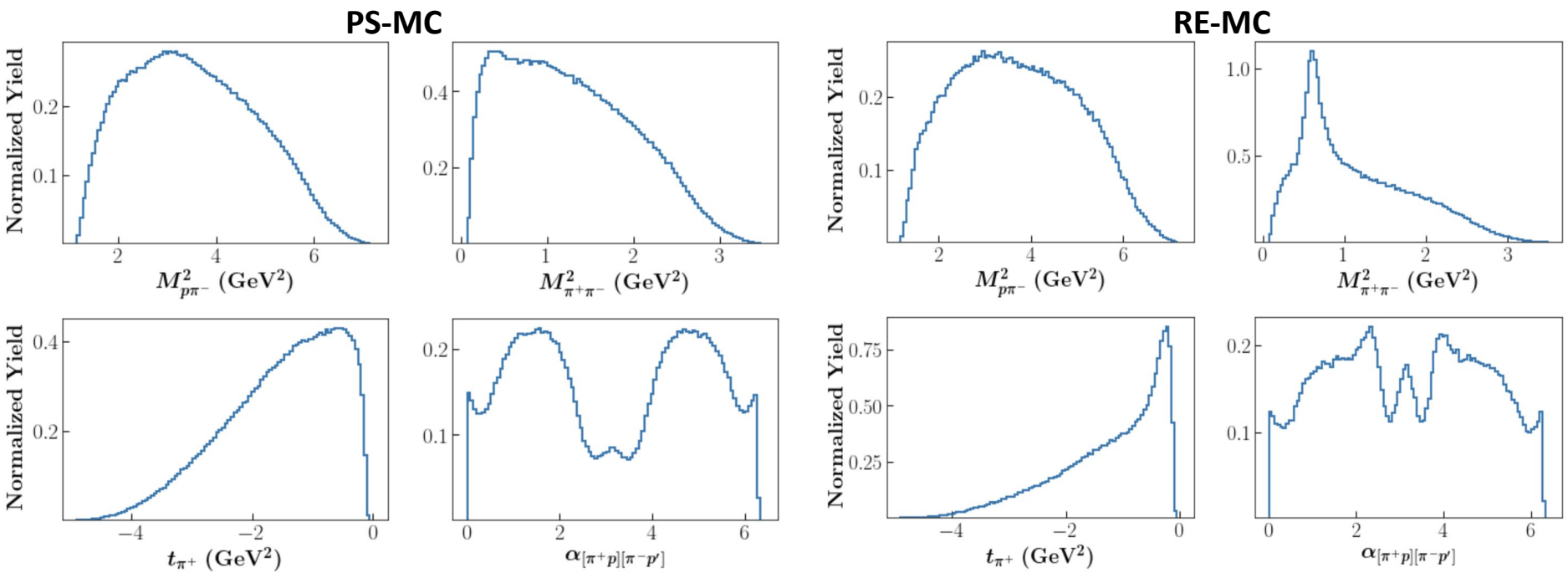
- **The Discriminator is trained to distinguish between samples drawn from the training data and those drawn from the Generator.**




** GAN was developed in 2014 by Ian Goodfellow, et al.*

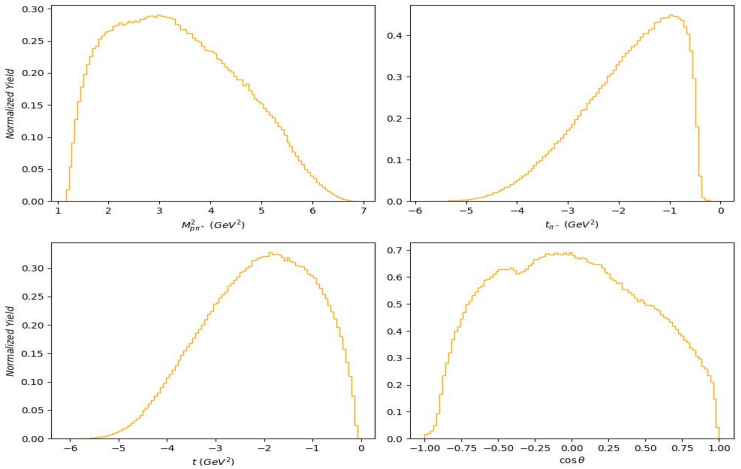
# Simulating particle collision events can be challenging:

The distributions of events exhibit narrow peaks, holes, and steep edges, which pose difficulty to precisely learn physical laws.



# Challenge:

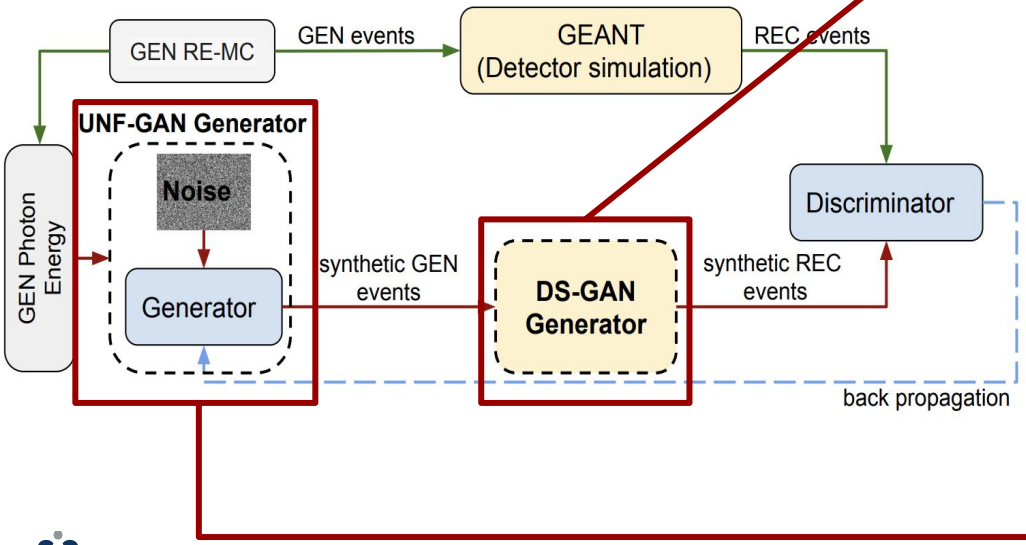
**Derived quantities not used in the training**

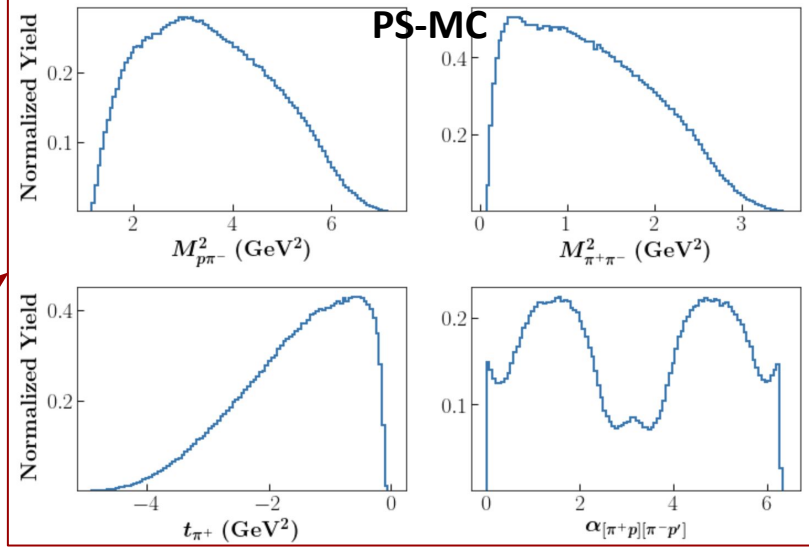


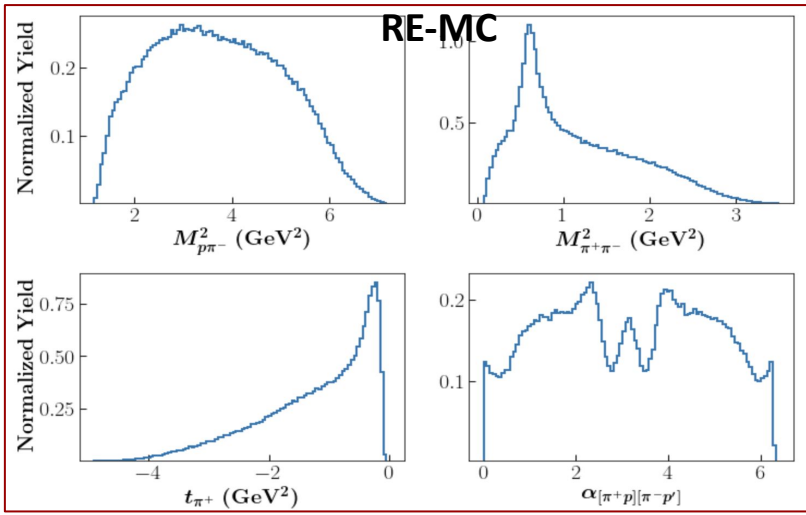

- We need to model the event features and the correlations precisely for the nature of particle reactions to be faithfully replicated.
- For example, we need to examine:
  - Calculate the derived physics variables that we do **NOT train on** such as: *M^2pπ+, t, tπ-,cosθ*
  - Transfer the results of training invariants variables to another space (CM then to Lab frame)
  - Calculate the momentum resolution $p = (\sqrt{p^2}_x, p^2_y, p^2_z)$**(NOT train features)**
  - Calculate the $pT = (\sqrt{p_x} * p_x + p_y, p_y)$ **(NOT train features)**

# Overall Closure Test:

**PS-MC:** Phase-Space Monte Carlo event generator(1M events)
**RE-MC:** Realistic Monte Carlo event generator (400K events)



# Detector Simulation (DS-GAN) using PS-MC pseudodata:

**Training variables:**

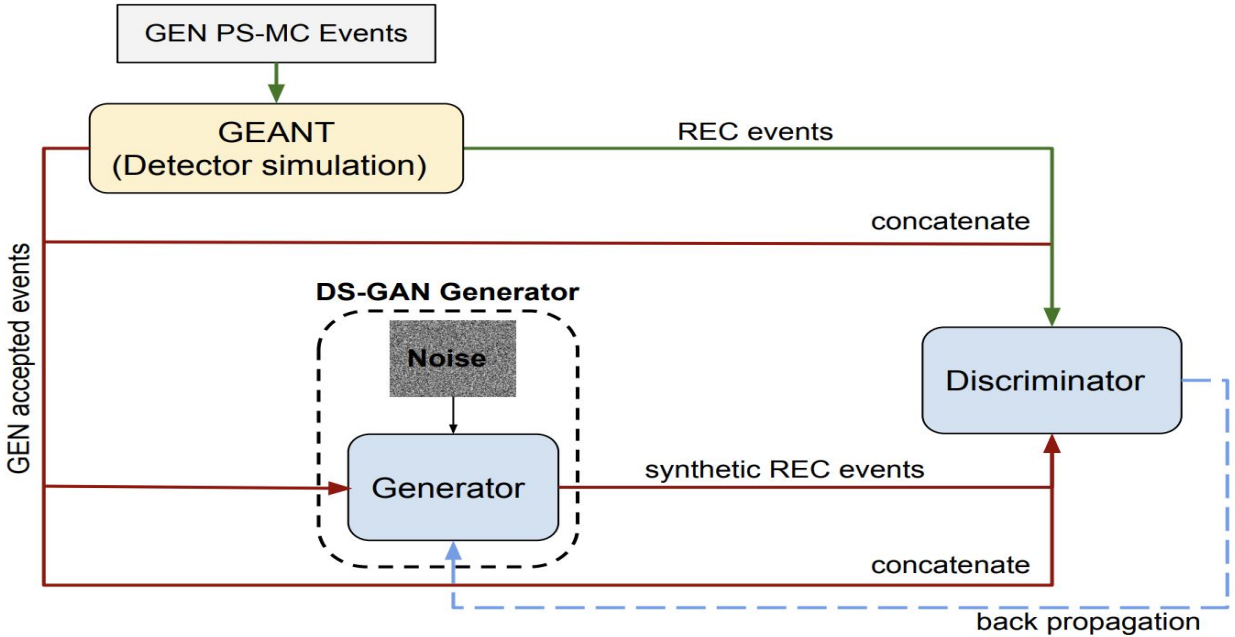
**Gamma**

***M^2 pπ-***

***M^2 π+π-***

***tπ+***

**Angle *α*[*π*+p][*π*-p']**

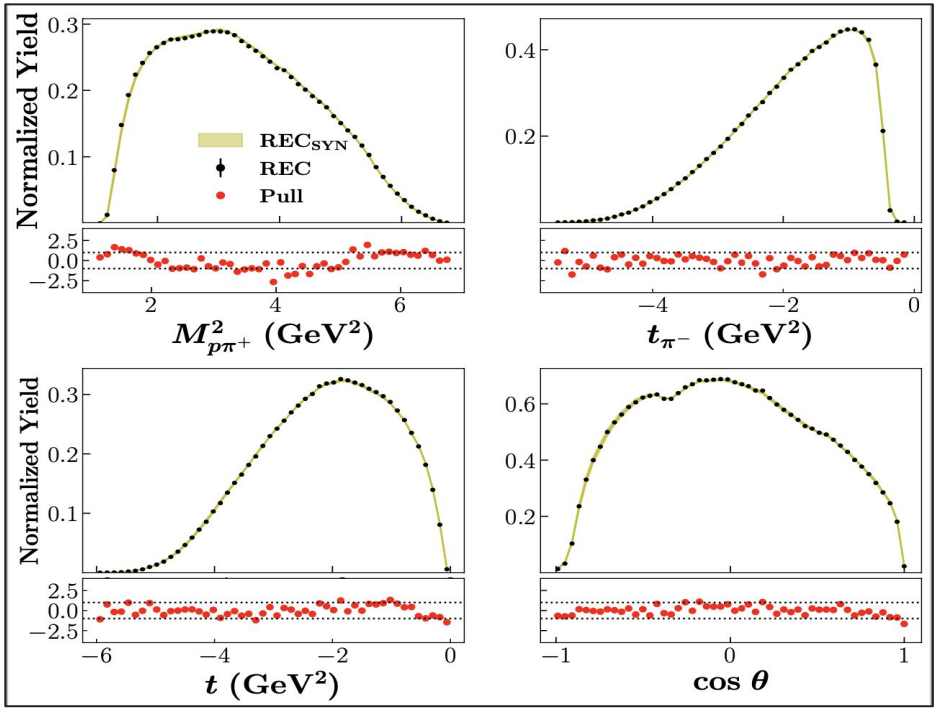



Illustrative view of the ML detector simulation (DS-GAN), where the GAN generator converts input GEN vertex-level events features and noise to REC detector-level events. The training is performed on PS-MC pseudodata passed through the GEANT simulation. Synthetic REC and REC pseudodata are concatenated with GEN PS-MC events and fed to the discriminator.

*GEANT is a system of detector description and simulation tools that help physicists in such studies.

# DS-GAN Results:

**Derived variables (not used in the training)**




**DS-GAN learned the CLAS detector effects!**

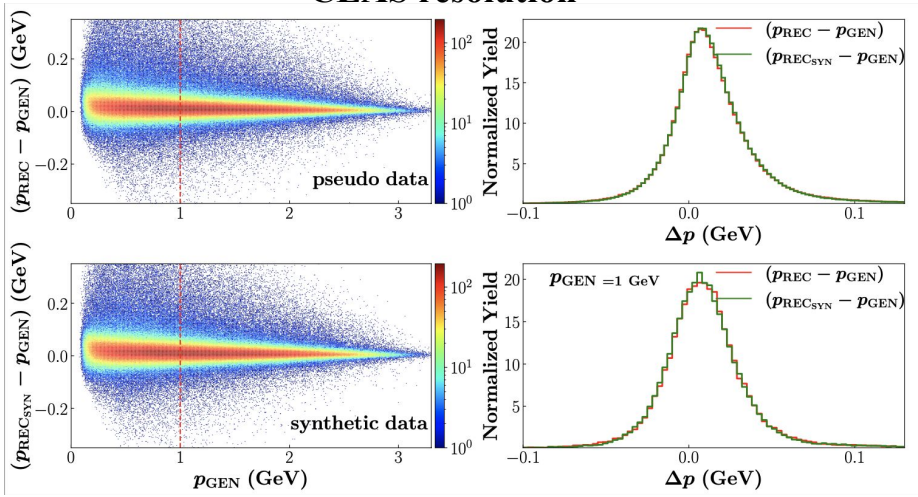
# DS-GAN Results:

Another way to check the smearing is to consider all momenta $P = \sqrt{p^2{}_x + p^2{}_y + p^2{}_z}$

**CLAS resolution**

**Truth**

**DS-GAN**




We define the energy of detector-level as, $p_{rec} = \sqrt{p_x^2 + p_y^2 + p_z^2}$

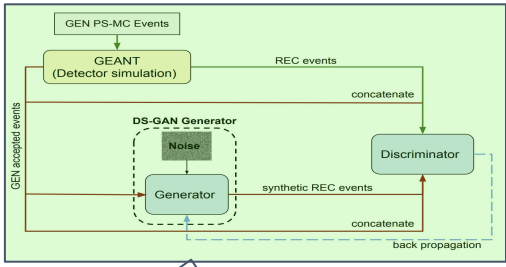
We define the energy of vertex-level as, $p_{gen} = \sqrt{p_x^2 + p_y^2 + p_z^2}$

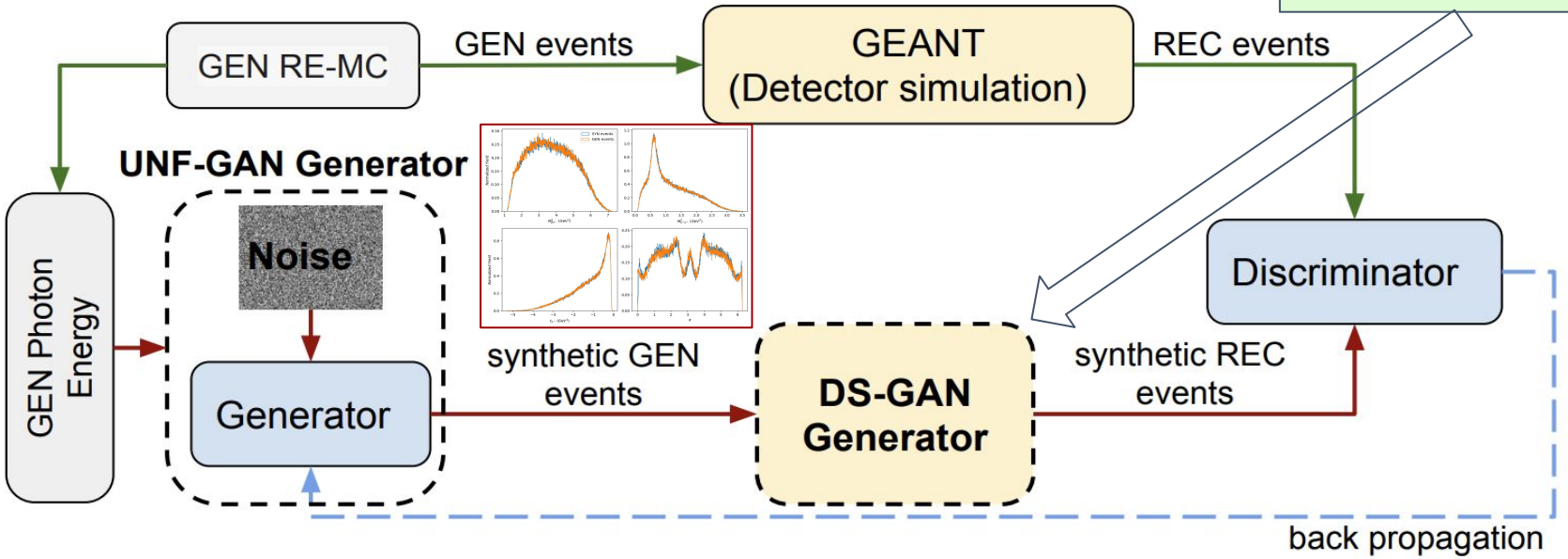
Then we plot the difference between $p_{rec}$ and $p_{gen}$ as a function of $p_{gen}$

**DS-GAN learned the CLAS detector effects!**

# Unfolding GAN (Unf-GAN) using RE-MC pseudodata:

- UNF-GAN trained with RE-MC REC pseudodata
- DS-GAN used to unfold CLAS detector effects

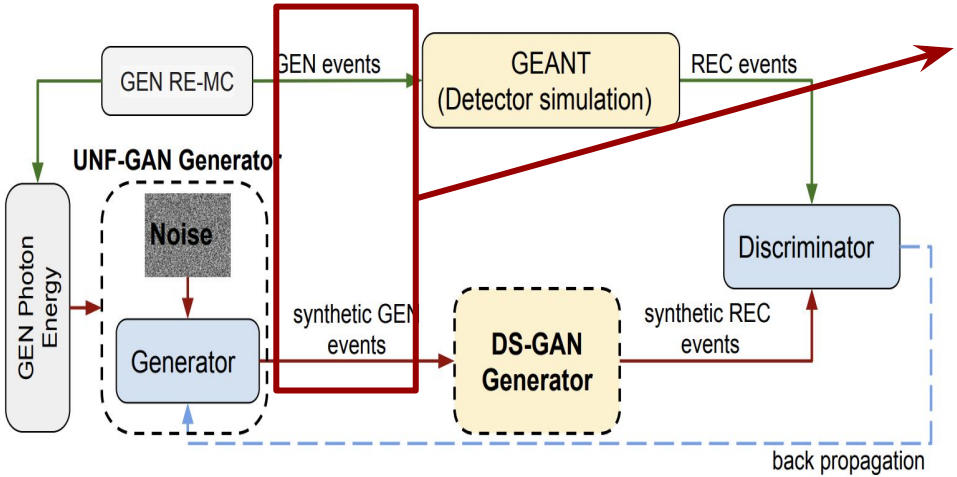


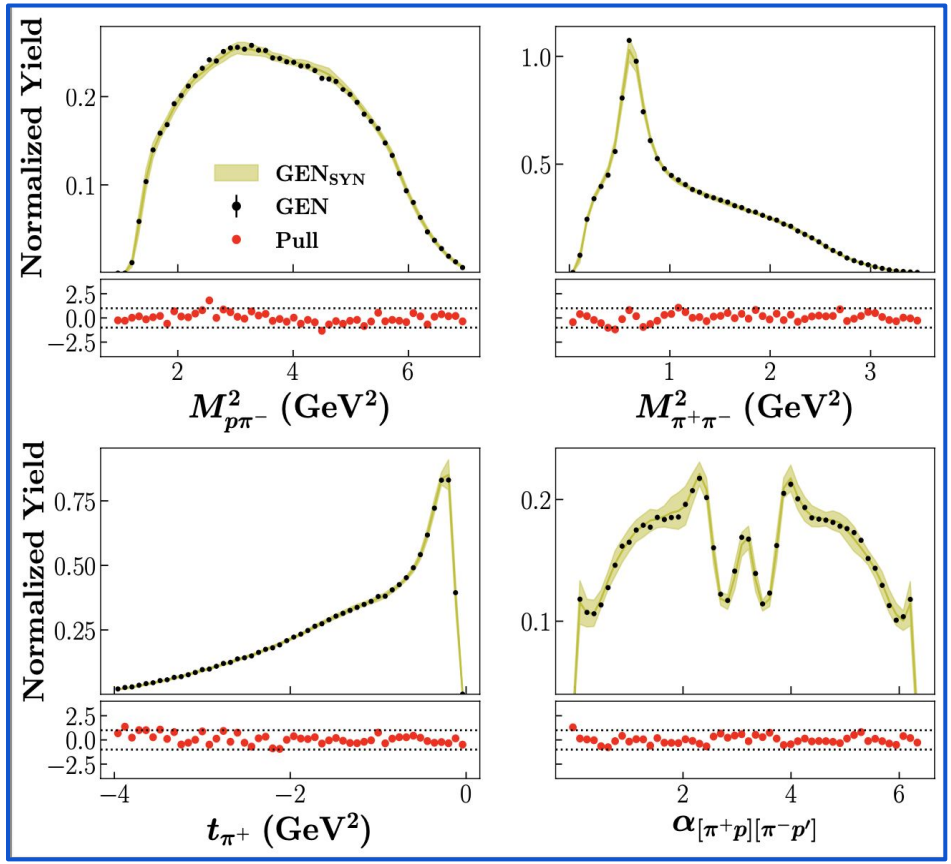




The UNF-GAN utilizes a generator that converts a GEN photon energy and random noise into synthetic GEN event features. These pass through the DS-GAN to incorporate the detector effect, and get converted into synthetic REC event features.

## UNF-GAN Results:

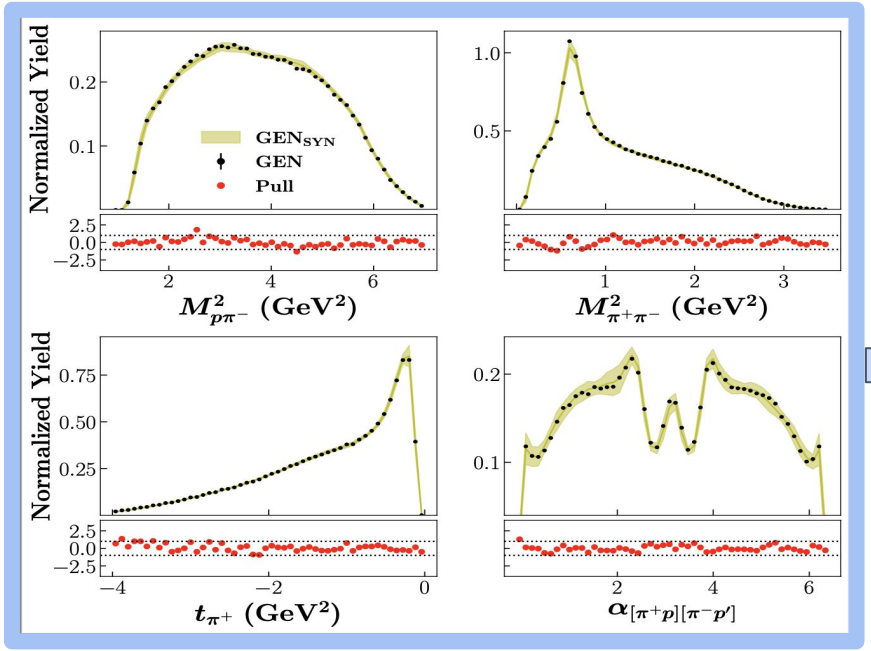
- Using REC RE-MC pseudodata to train the UNF-GAN, and extract the GEN-SYN distributions



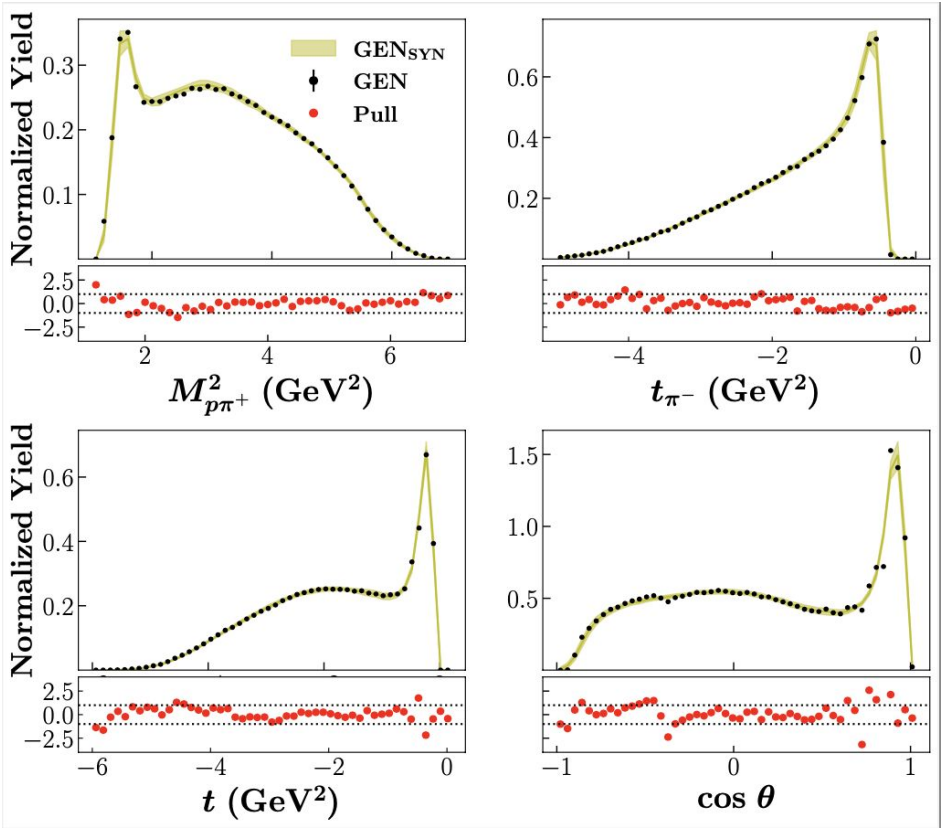




Systematic of the full procedure (two GANs) estimated by bootstrap with 20+20 independently trained GANs

# UNF-GAN Results:




**Derived variables (not used in the training)**



# Unfolding technique using ML:

**Omnifold technique*:**

- A novel approach for unfolding multiple observables simultaneously
- The OmniFold technique incorporates reweighting strategies.
- Using neural networks
- Taking into account correlations between observables and using them to improve the unfolding of each individual observable.
- Works for high-dimensional data

**Andreassen, Anders, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. "OmniFold: A method to simultaneously unfold all observables." Physical review letters 124, no. 18 (2020): 182001.*

**How to GAN away Detector Effects*:**

- A novel method to handle detector effects in high-energy physics experiments
- This approach uses GAN.
- The aim of this study is to simulate the detector response.

**Bellagente, M., Butter, A., Kasieczka, G., Plehn, T., & Winterhalder, R. (2020). How to GAN away detector effects. SciPost Physics, 8(4), 070.*

# Other related work

- In the literature, GANs, VAEs, NFs, and their various improved architectures have been used to simulate physics events from different reactions and training datasets
- Two of them* reported that the LS-GAN yields better performance than other generative models
- Better precision

*Butter, Anja, and Tilman Plehn. "Generative Networks for LHC events." In Artificial intelligence for high energy physics, pp. 191-240. 2022.

*Otten, Sydney, et al. "Event generation and statistical sampling for physics with deep generative models and a density information buffer." Nature communications 12.1 (2021): 2985.