The A(i)DAPT program
AI for Data Analysis and Preservation

Tommaso Vittorini

on behalf of A(i)DAPT Working Group

A(i)DAPT

AI for Data Analysis and PreservaTion

- 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

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

A(i)DAPT
AI for Data Analysis and PreservaTion

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)

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

Exlusive reactions: 2 → 3

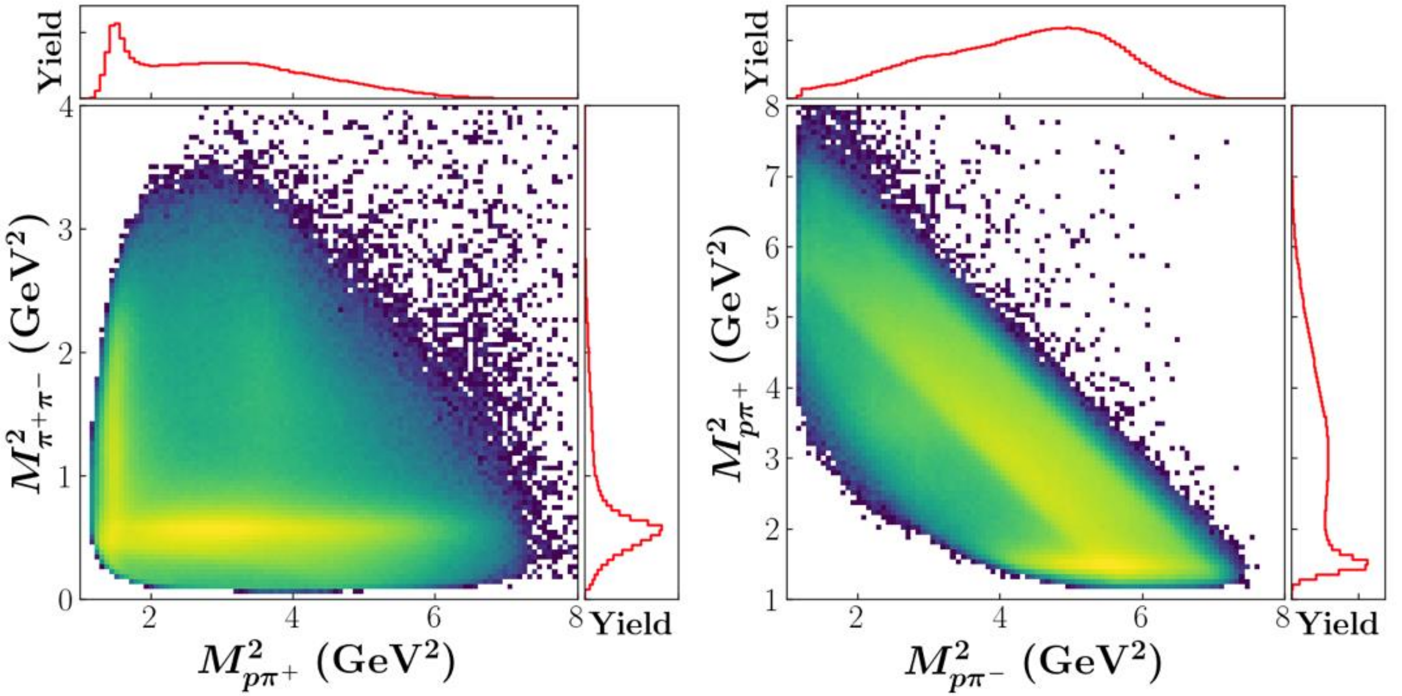
$\gamma p \rightarrow \pi^+\pi^- p$ (unpolarized)

- Initial state: Fully known
- Final state: 3x3 indipendent variables
- Indipendent variables: (3x3) – 4 = 5 (E_γ fixed)
- Many possible choices, such as $M^2_{\pi\pi}, M^2_{p\pi}\ \theta_\pi, \alpha, \phi$

CLAS g11 2π photoproduction

- $E_\gamma = (3\ - 3.8)\ GeV$
- Dataset analyses on $\gamma p \rightarrow p\pi^+(\pi^-)$ with small contamination from $\gamma p \rightarrow p\pi^+$(more than a single missing π^-)
- Complicated dynamics due to the overlap of $(\mathrm{p}\pi)$ to form Δ baryon resnoances and $(\pi\pi)$ to form meson resonances

$$\frac{d\sigma\ (\gamma\ p \rightarrow p\,\pi^+\pi^-)}{dM_{\pi\pi}\ dM_{p\pi}\ d\cos(\theta_\pi)\ d\alpha\ d\phi}$$



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

Credit: Y.Alanazi Awadh, , P..Ambrozewicz, G. Costantini A.Hiller Blin, E. Isupov, T. Jeske, Y.Li, L.Marsicano W. Menlnitchouk, V.Mokeev, N.Sato, A.Szczepaniak, T.Viducic

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?

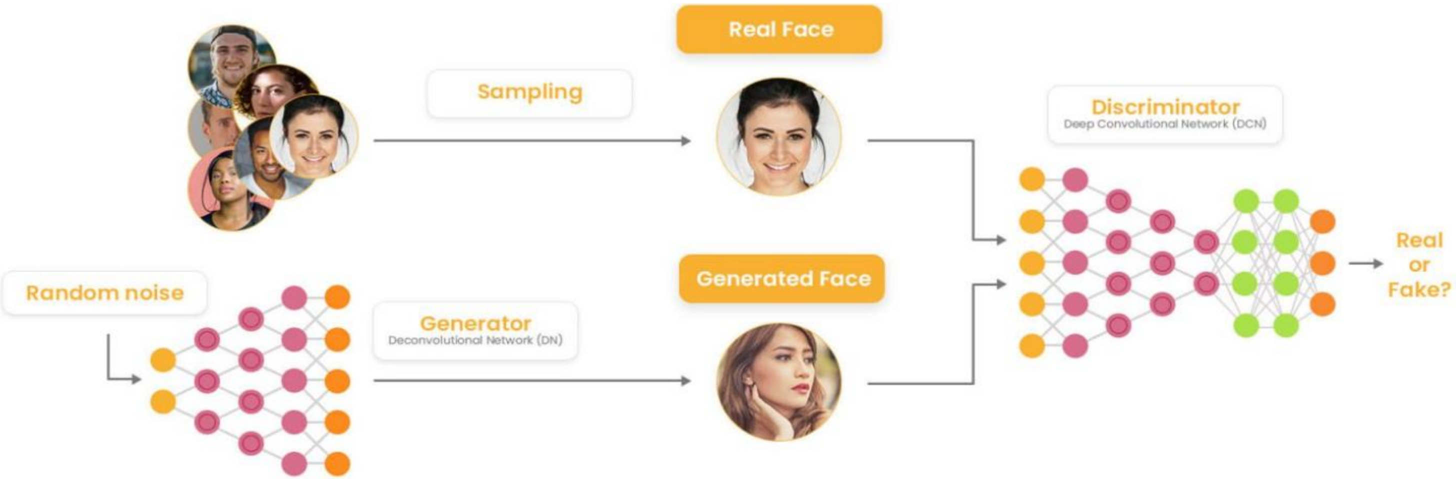
- Interpolation: deal with the holes in the phase space
- Extrapolation: extend our coverage from the borders of measured regions

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

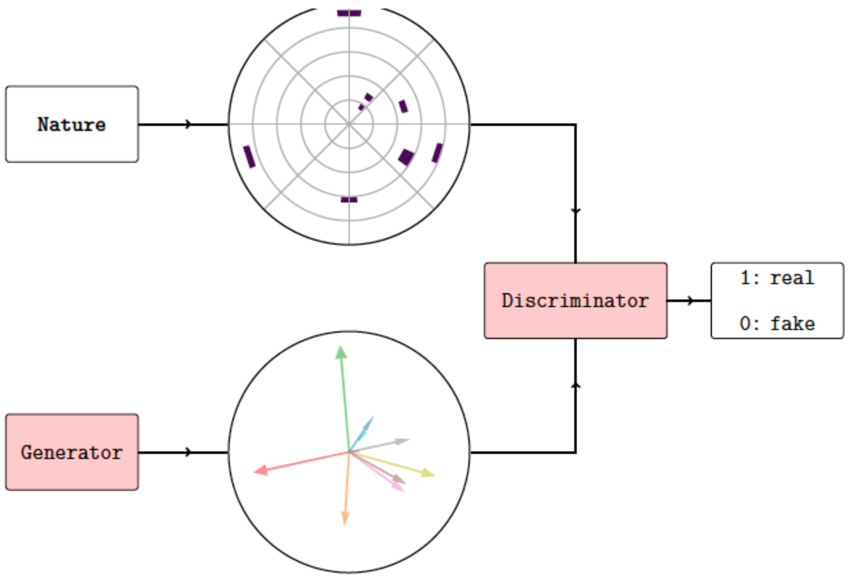
- Spikes may be concealed behind the detector resolution
- Measurements could be extended to unphysical regions

- 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 (GANs)



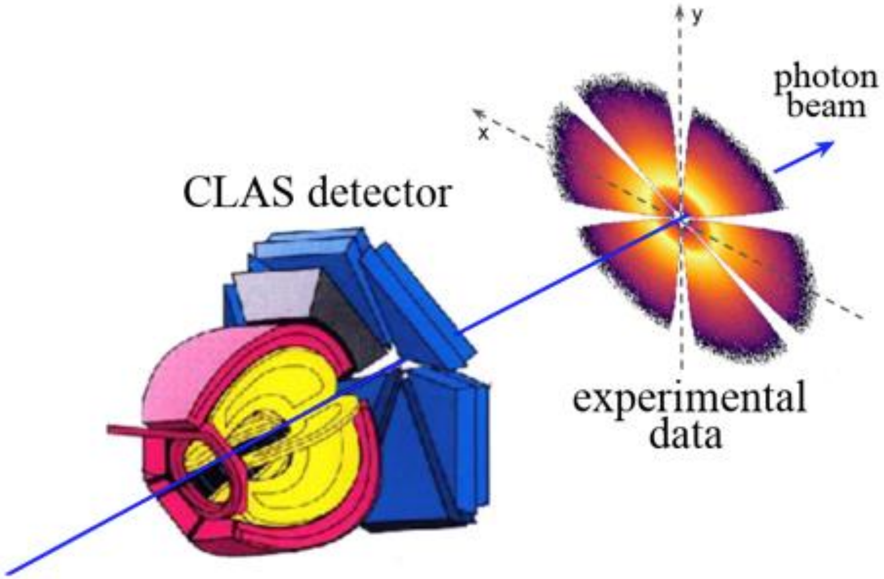
- Generative model based on the competition between two Neural Networks: Generator vs Discriminator
  - **Generator** produces synthetic data which progressively reproduce realistic data and the **Discriminator** has to distinguish between synthetic and realistic data
  - **Generator** be used to retain high dimensional correlations (detector proxies)
  - **Generator** can be used to provide highly realistic pseudo-data in an extremely fast way



Multi-d cross-section: exclusive 2π photoproduction

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

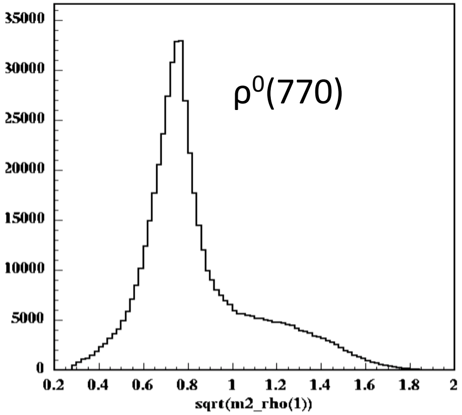
M. Battaglieri *et al.* (CLAS Collaboration) Phys. Rev. D 80, 072005



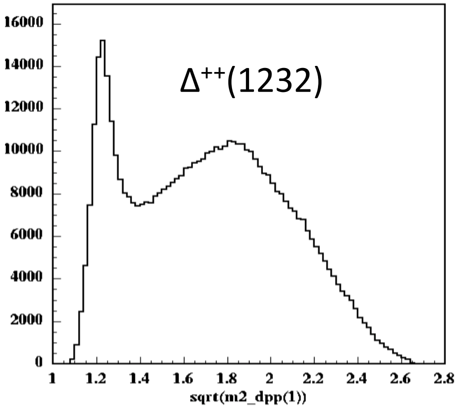
CLAS g11 kinematics

- 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 $\mathrm{E}_\gamma = (3-4)\mathrm{GeV}$ reaction dynamics are dominated by ρ^0 photproduction through $\gamma p \rightarrow p\rho^0$ and Δ^{++} resonance excitation through $\gamma p \rightarrow \Delta^{++}\pi^-$

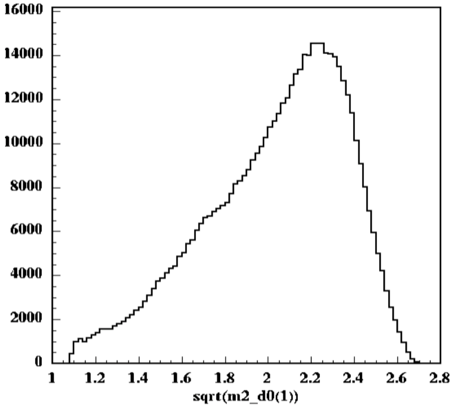
M(ππ)



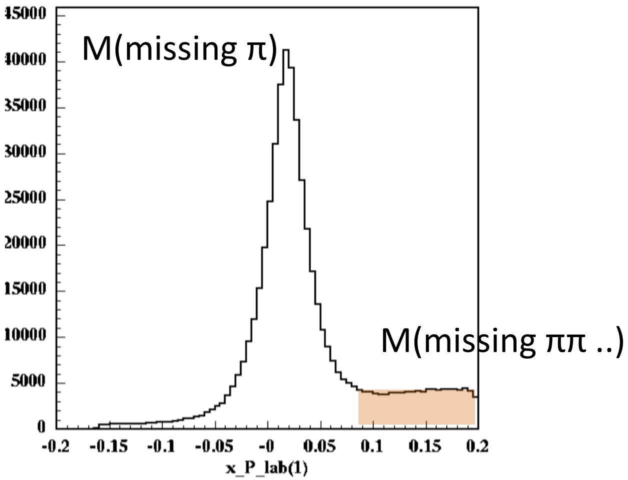
M(pπ+)



M(pπ-)



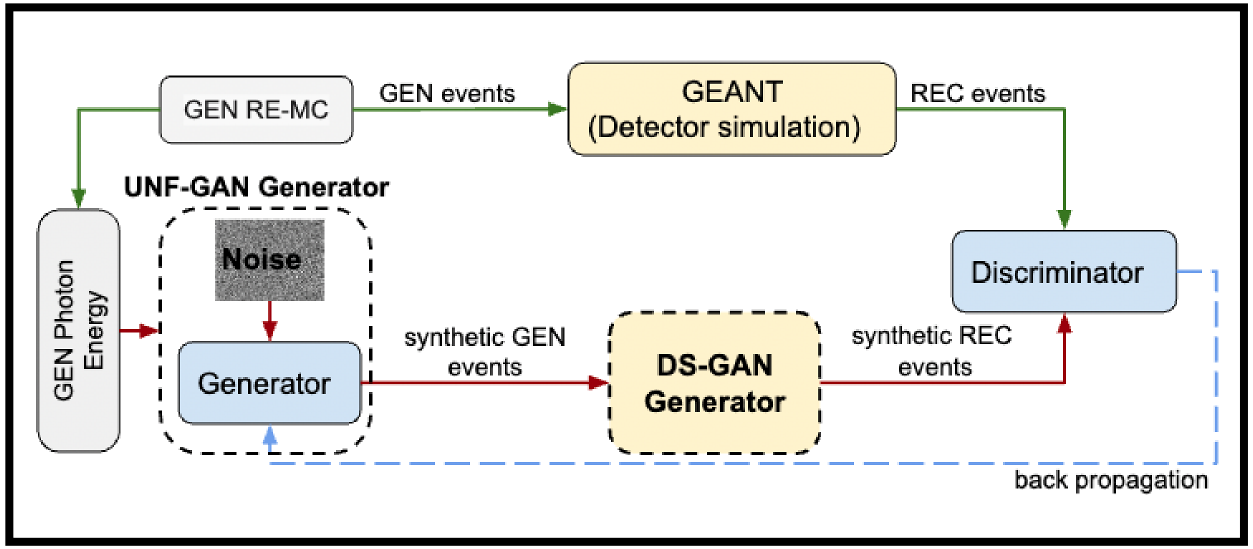
M(γ p → p π+ X)



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

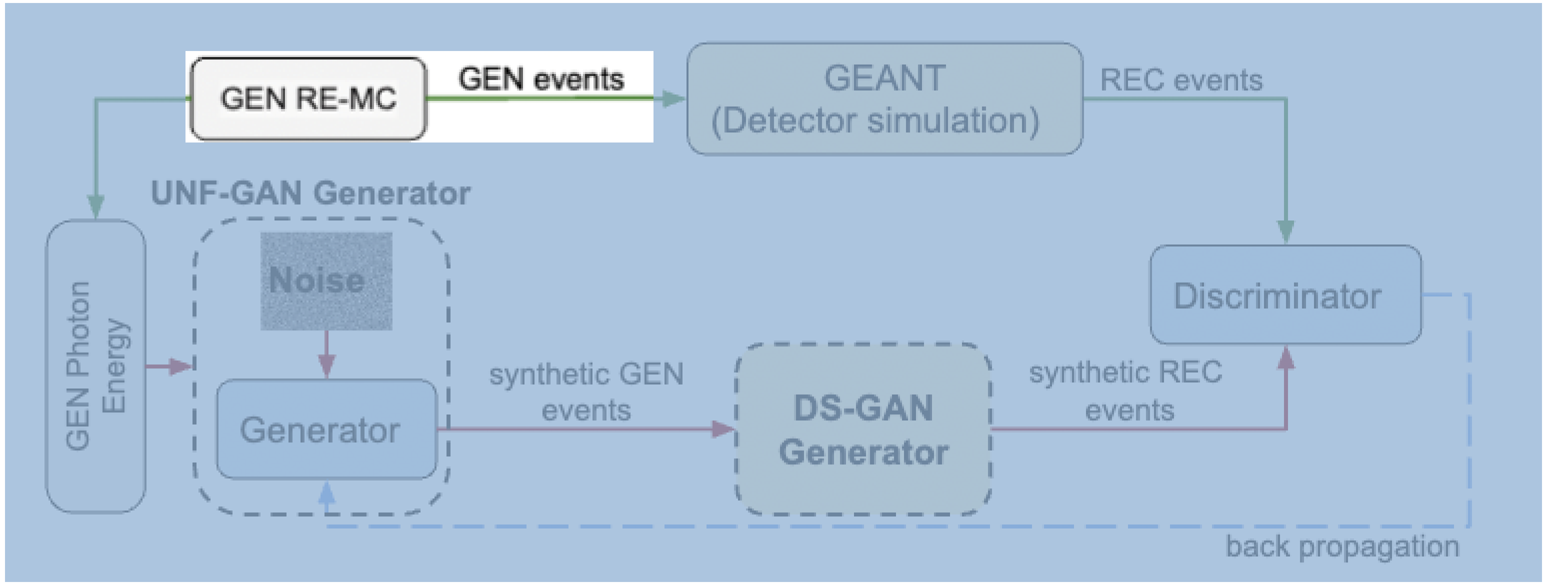


Credit: T.Alghamdi, M.Battaglieri, A.Golda, A. Hiller Blin, L.Marsicano, W.Melnitchouk, G.Montaña, E.Isupov, Y.Li, V.Mokeev, A.Pilloni, N.Sato, A.Szczepaniak, T.Vittorini, Y.Alanazi *arXiv:2307.04450*

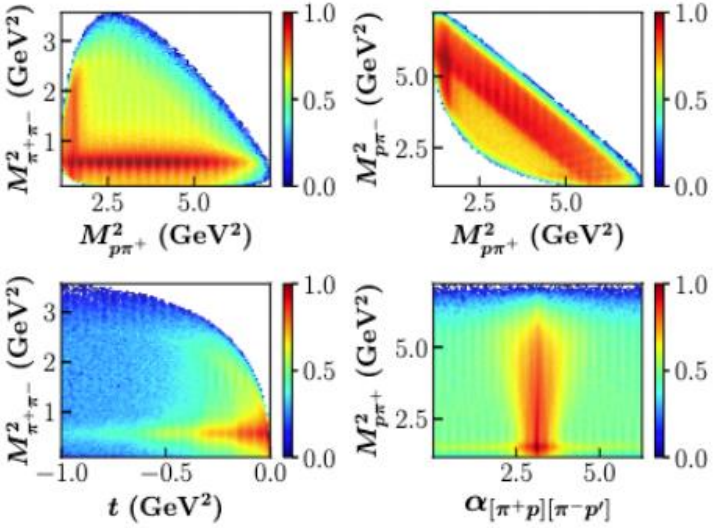
2π photoproduction closure test

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)



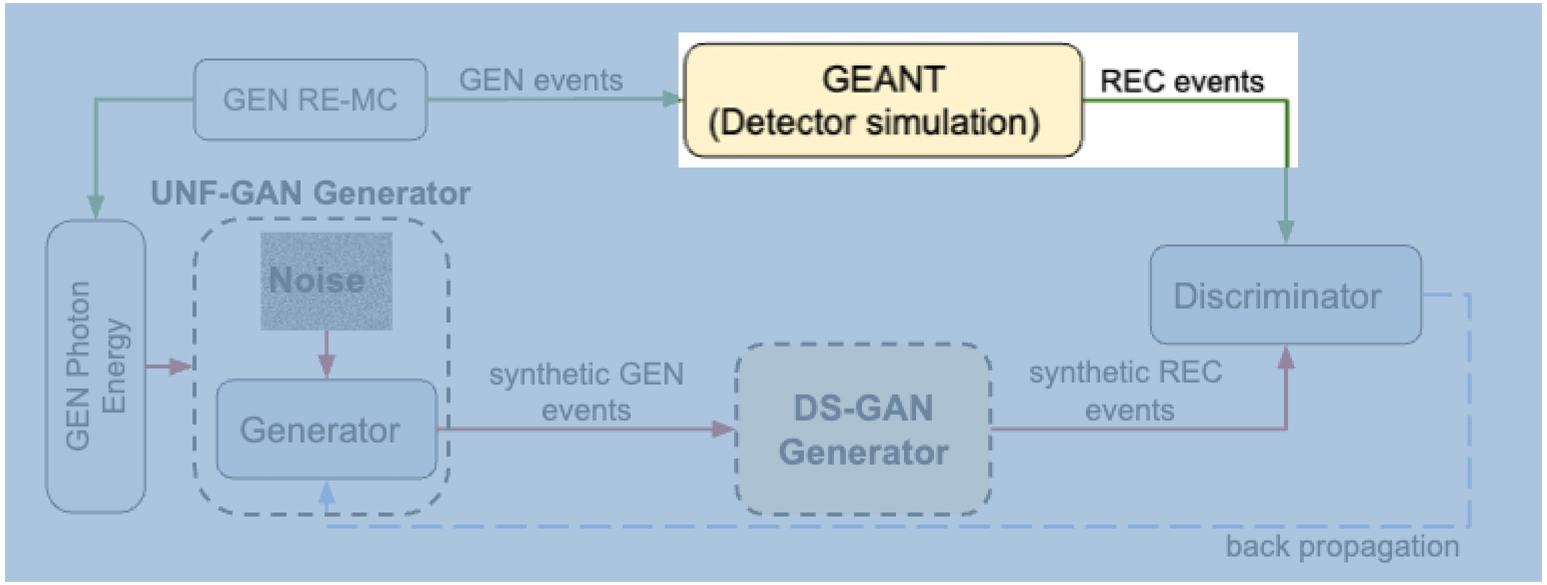
GEN events

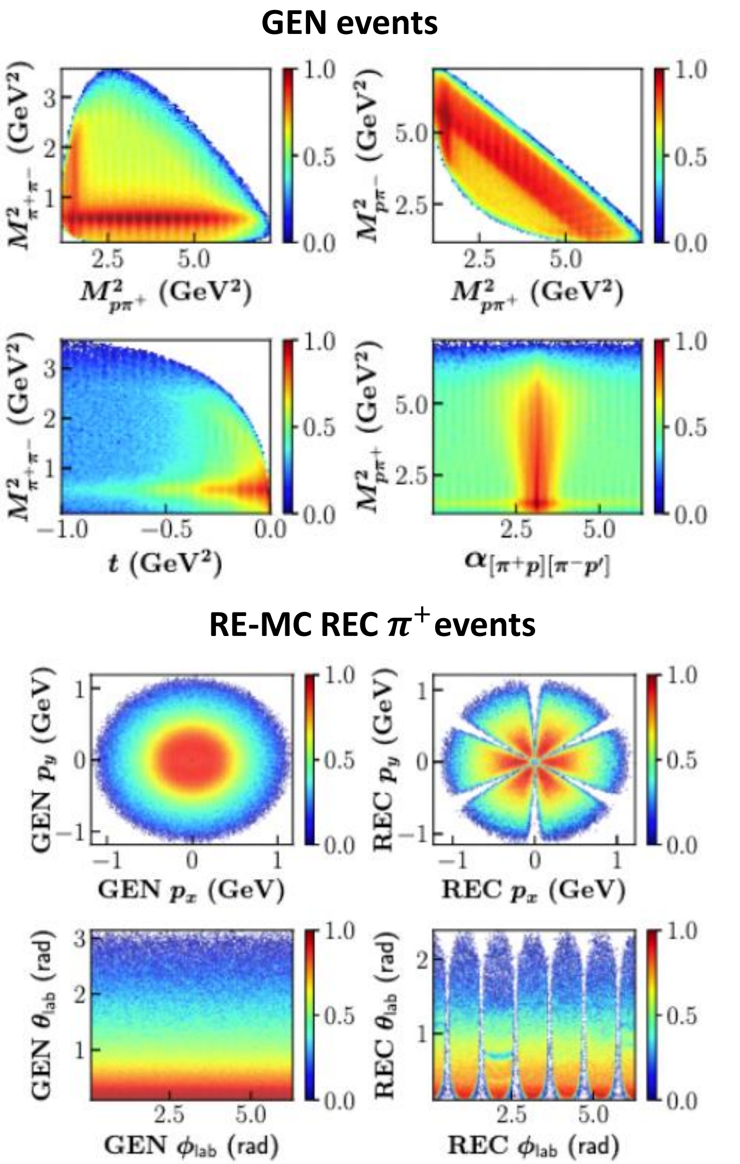


2π photoproduction closure test

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

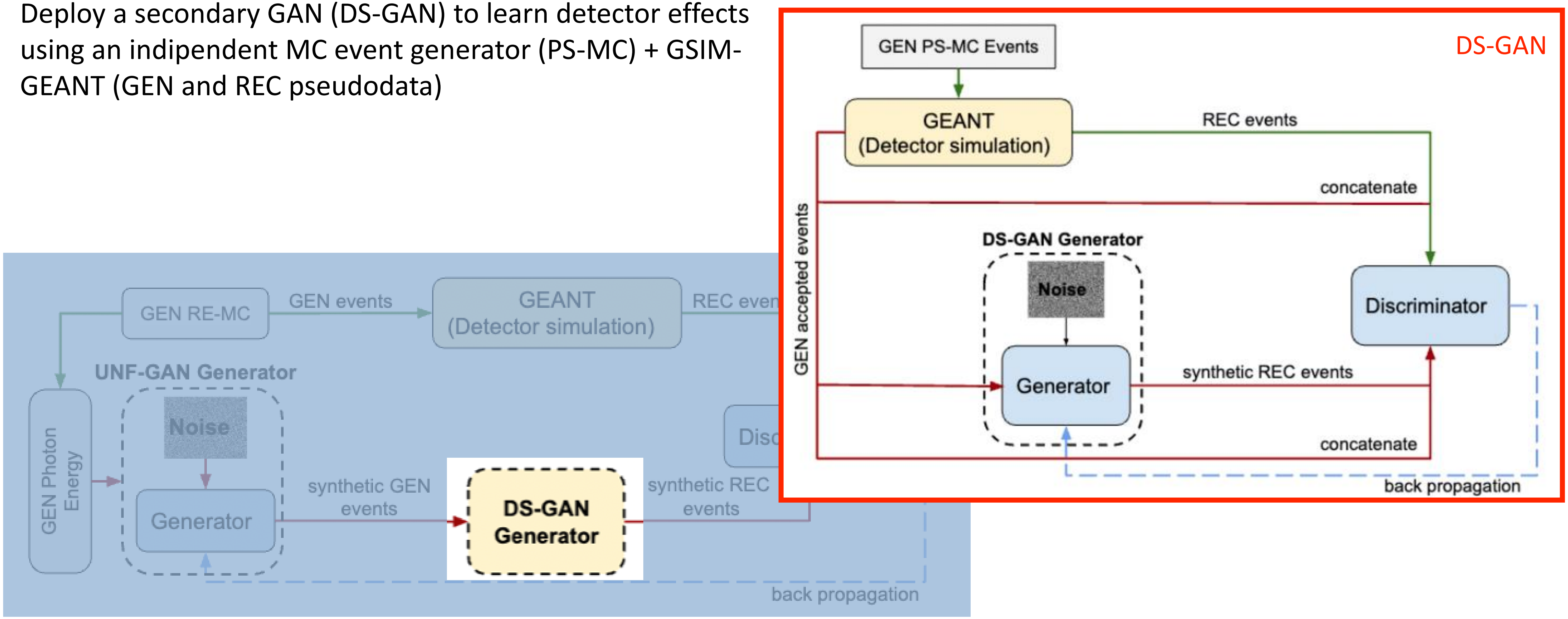
- GSIM: detector simulation package to simulate CLAS detector effects based on GEANT3





2π photoproduction closure test

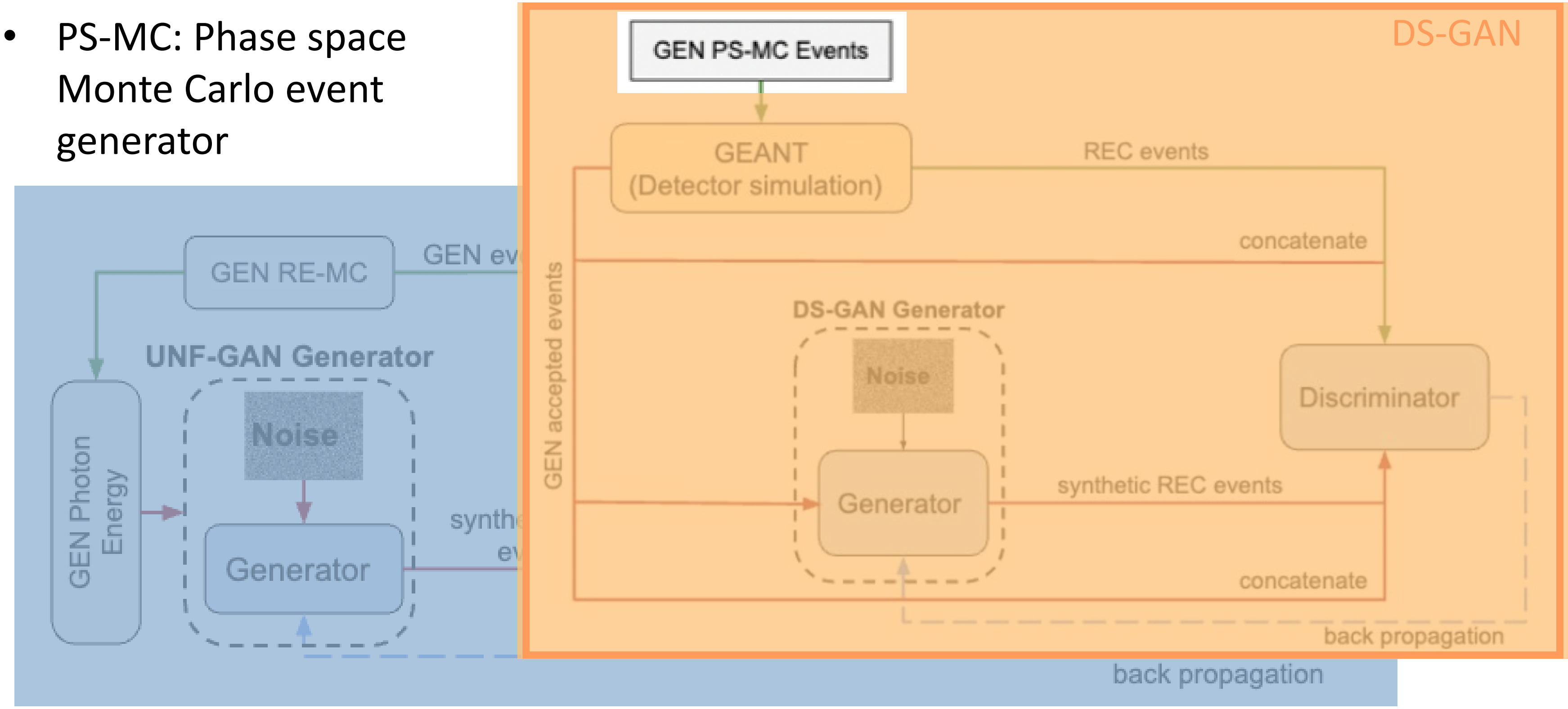
3. Deploy a secondary GAN (DS-GAN) to learn detector effects using an indipendent MC event generator (PS-MC) + GSIM-GEANT (GEN and REC pseudodata)



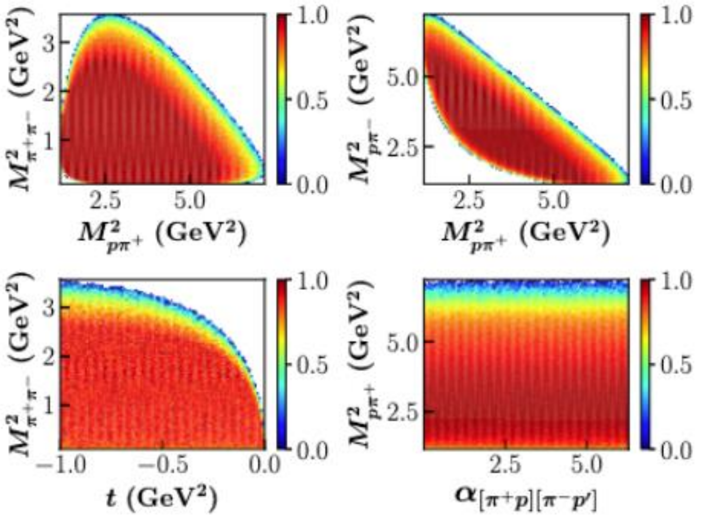
2π photoproduction closure test

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



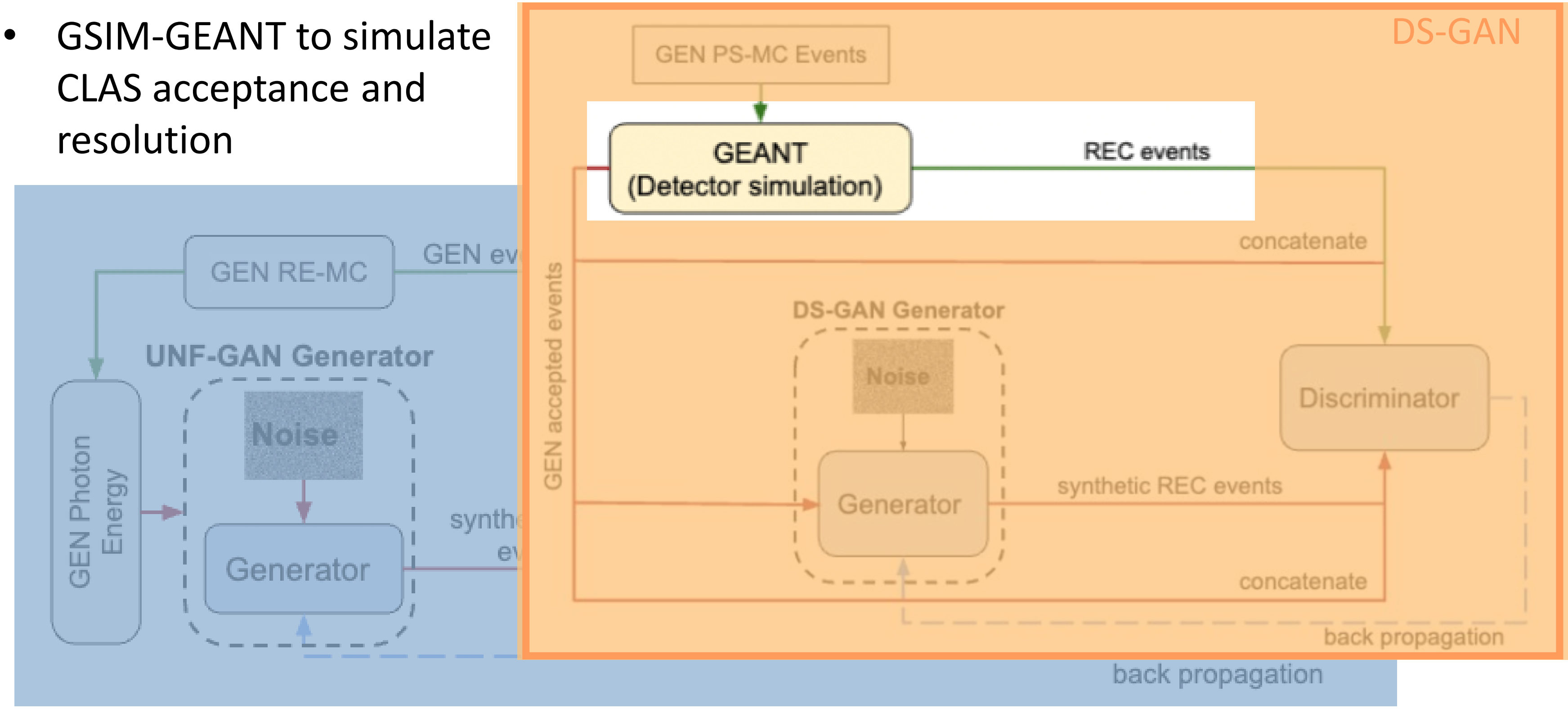
PS-MC GEN events



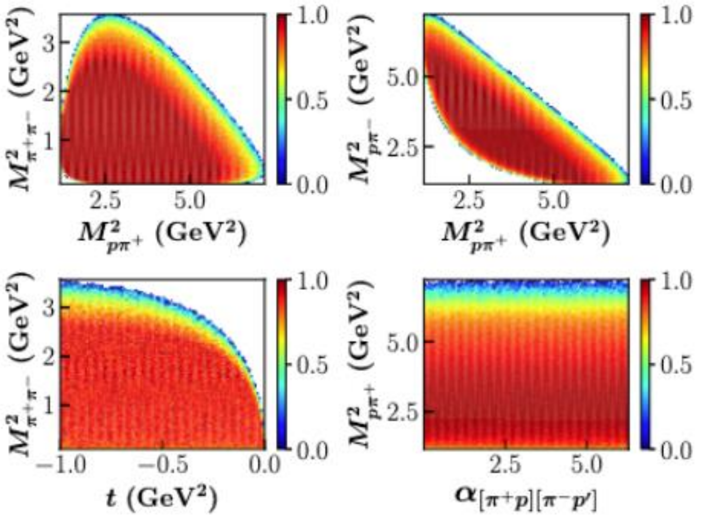
2π photoproduction closure test

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)

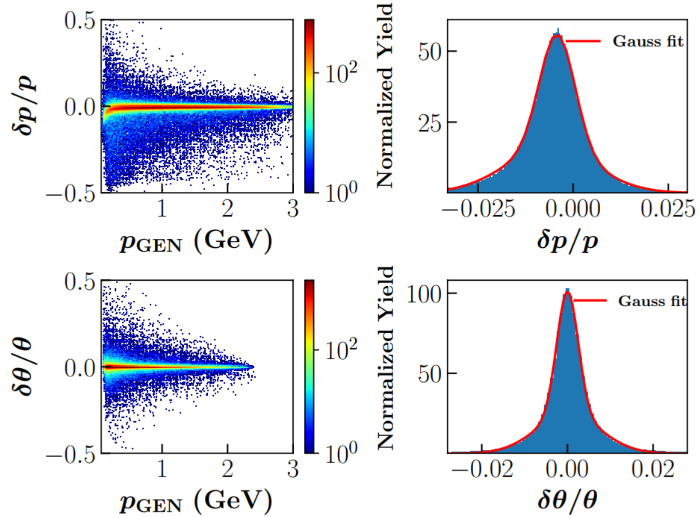
- GSIM-GEANT to simulate CLAS acceptance and resolution



PS-MC GEN events



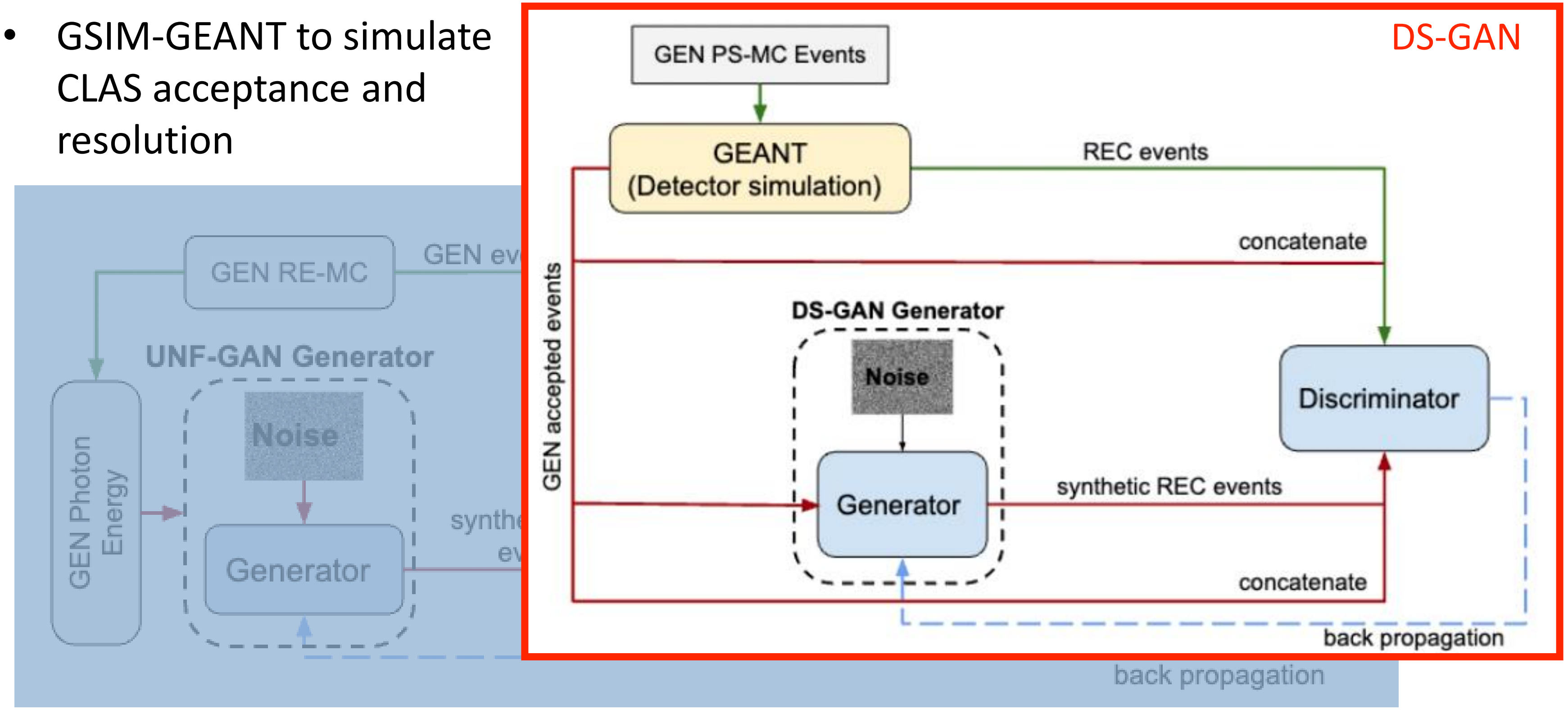
CLAS resolution on π^+ kin. variables



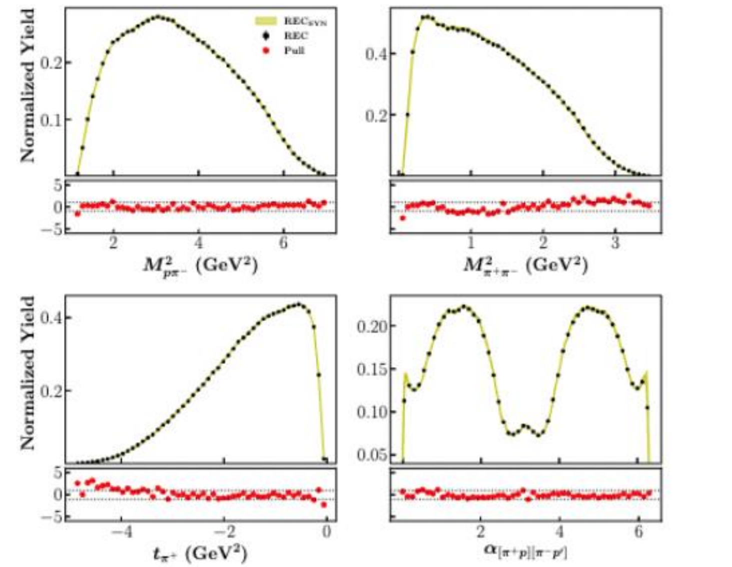
2π photoproduction closure test

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)

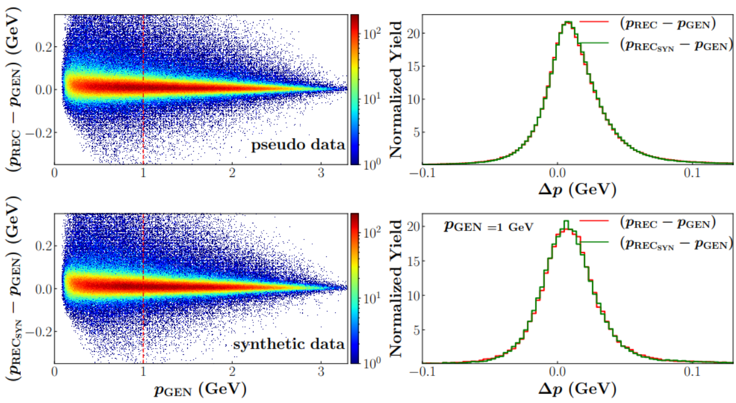
- GSIM-GEANT to simulate CLAS acceptance and resolution



MC REC pseudodata vs. DS-GAN synthetic data



CLAS resolution

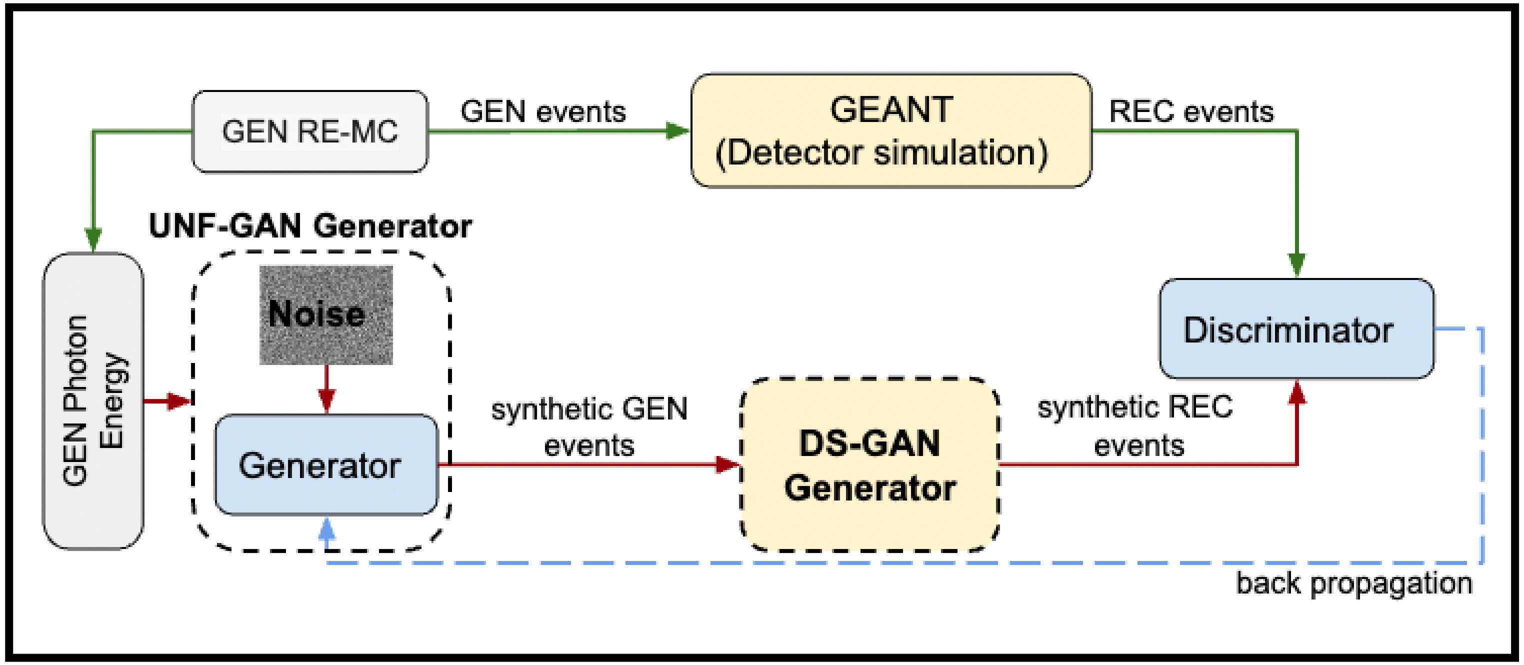


Uncertainty quantification via **pull** calculation: Bootstrap with 20 indipendently trained GANs

DS-GAN learned the CLAS detector effects!

2π photoproduction closure test

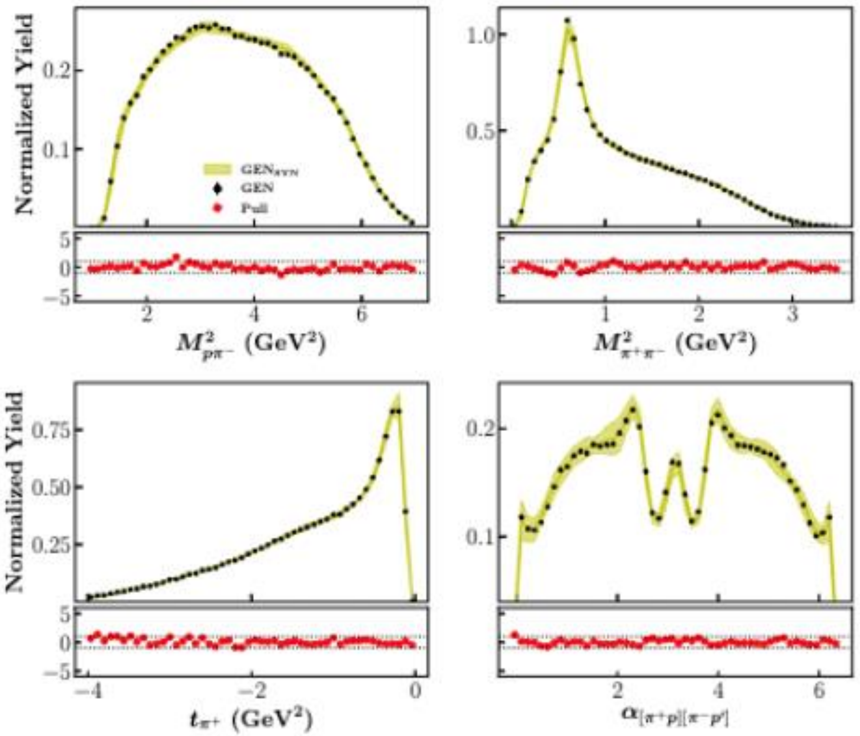
- UNF-GAN trained with REC-MC pseudodata (experimental data proxy)
- DS-GAN used to unfold CLAS detector effects (within acceptance)



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

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

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

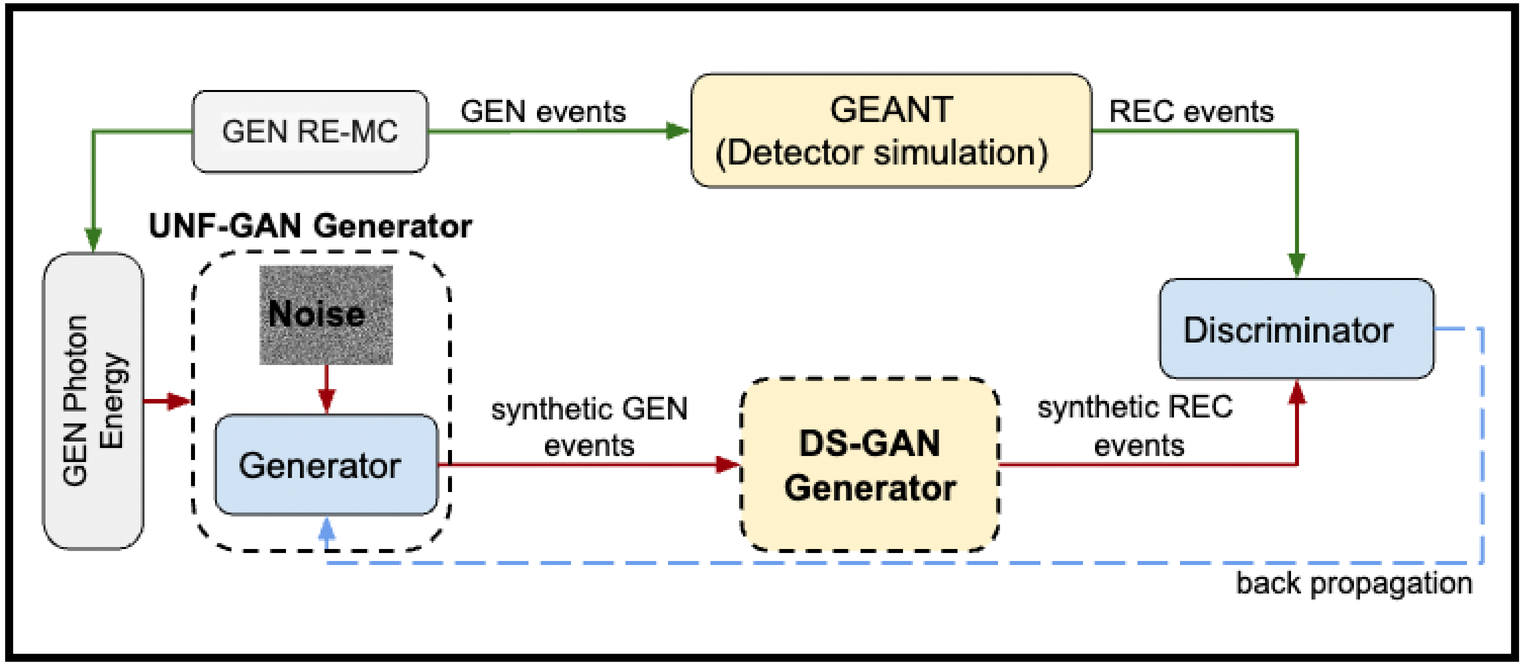


- Systematic of the full procedure (two-GANs) estimated by bootstrap with 20+20 independently trained GANs

2π photoproduction closure test

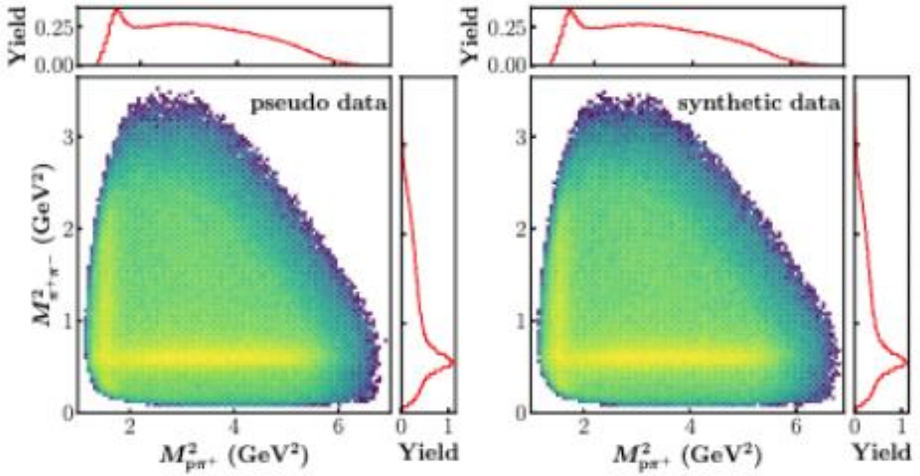
4. Deploy the unfolding GAN (UNF-GAN) that includes the DS-GAN and train it with RE-MC REC pseudodata

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

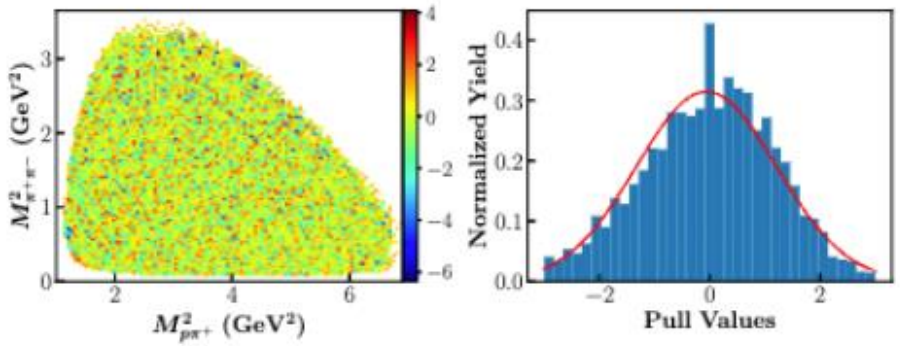


5. Compare UNF-GAN GEN SYNT to RE-MC GEN pseudodata

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



2D pulls

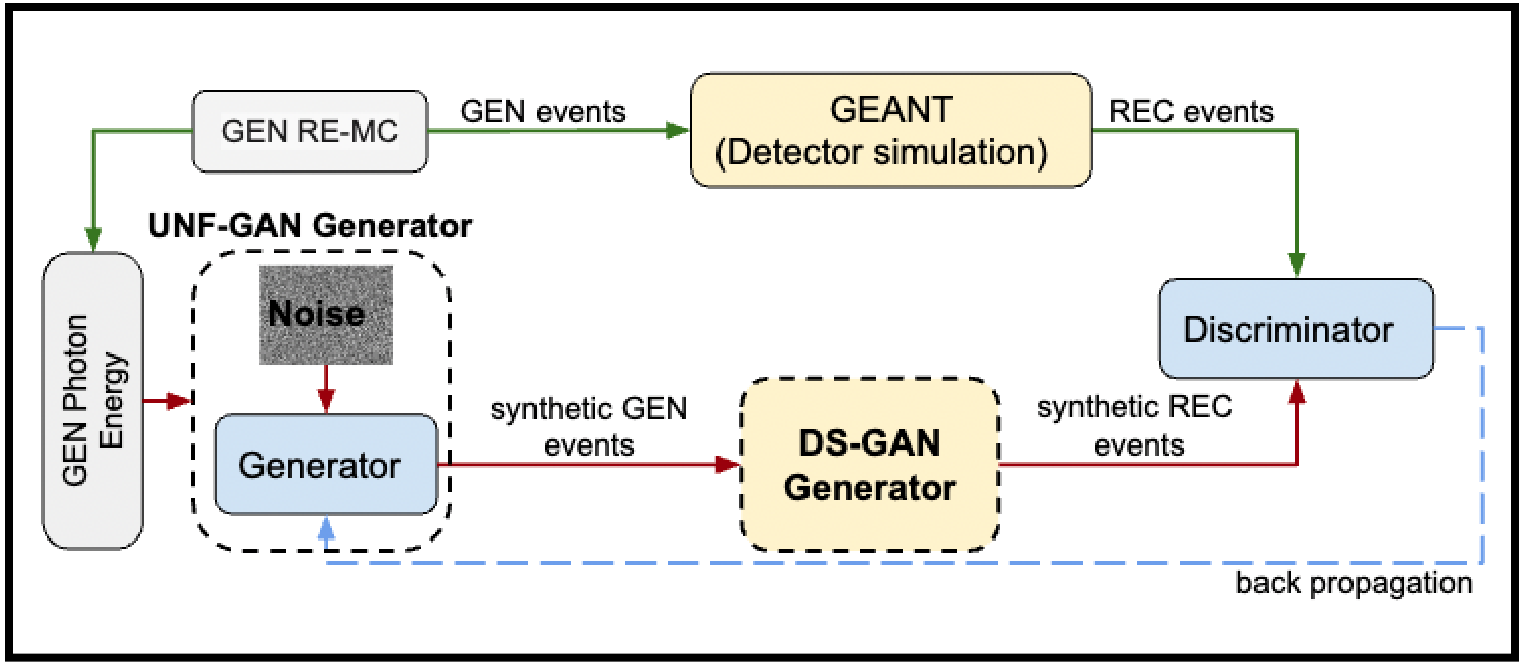


Good agreement ($\pm 1\sigma$) for 2D distributions (correlations)

2π photoproduction closure test

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

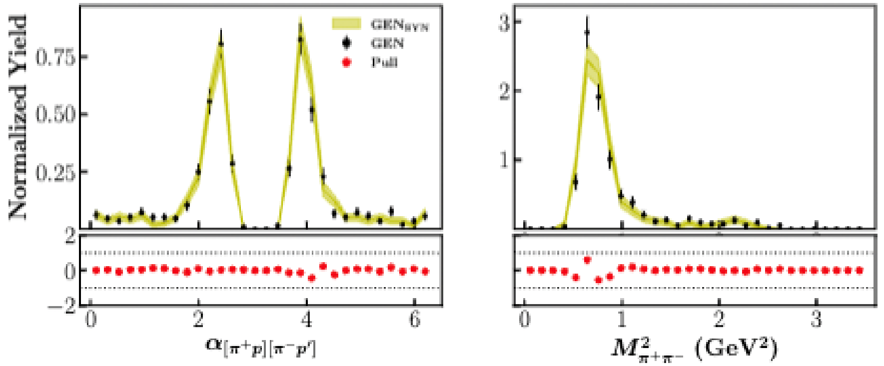
- UNF-GAN trained with REC-MC pseudodata (experimental data proxy)
- DS-GAN used to unfold CLAS detector effects (within acceptance)



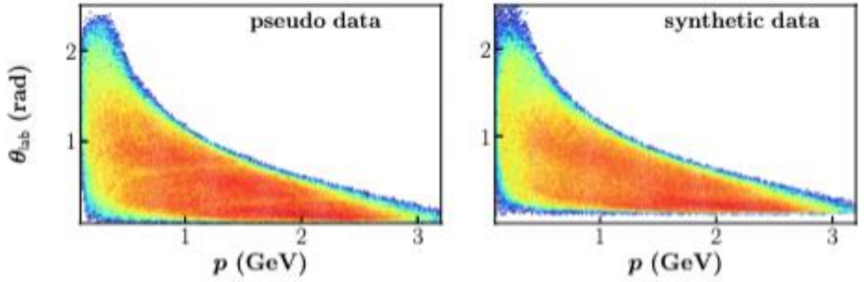
5. Compare UNF-GAN GEN SYNT to RE-MC GEN pseudodata

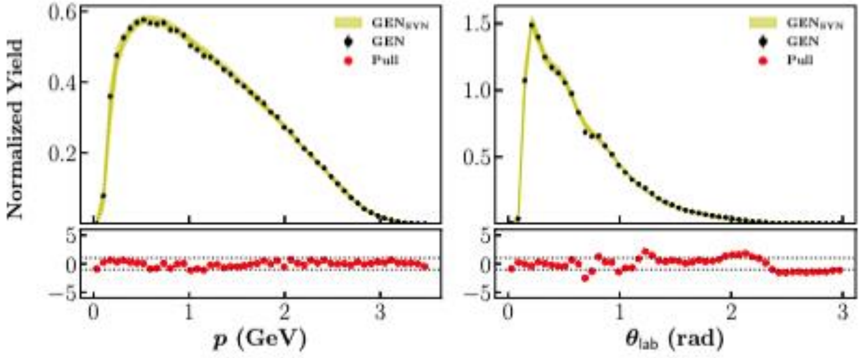
Good agreement ($\pm 1\sigma$) for lab variables and in 4D bins

Distribution in 4D bins



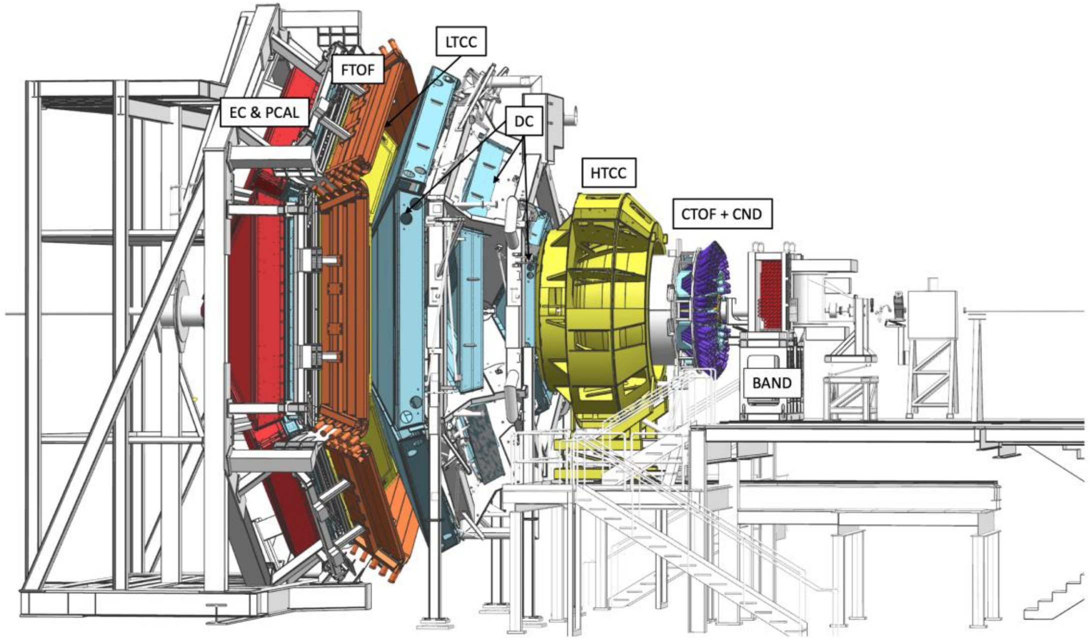
RE-MC GEN pseudodata vs. UNF-GAN SYN data



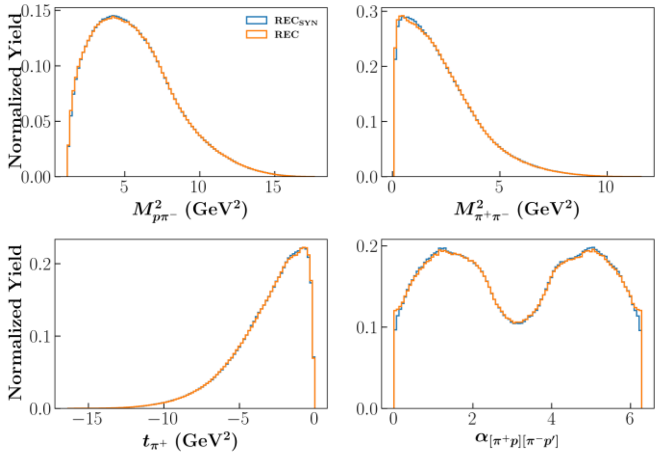


CLAS12 application

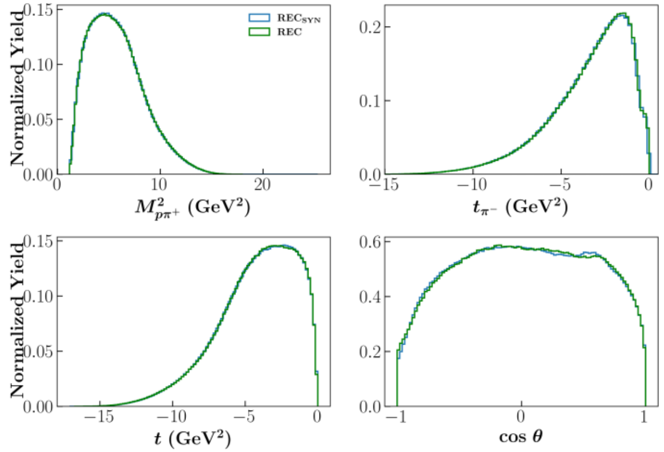
- Working towards the application of the developed machinery to CLAS12 pseudodata
- If this procedure works well on CLAS and CLAS12 data the architecture robustness is guaranteed
- We can put together in a coherent way information from different kinematic regions



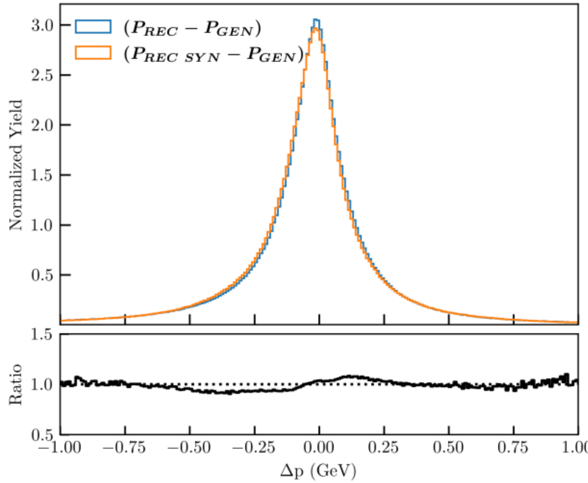
REC SYN vs REC pseudodata training variables



REC SYN vs REC pseudodata derived variables



CLAS12 resolution



Good agreement for training and derived variables

Credit: Derek Glazier, Tareq Alghamdi

Summary

A(I)DAPT program aims to demonstrate a novel way to extract and interpret physics observables

- Multi-step program
- 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 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
- Further verify that this procedure is well defined confronting the results obtained analysing CLAS data with traditional analysis in order to extract a 4D cross-section
- Make this procedure an efficient way to analyse CLAS12 2π data

There is still a long way to go to be able to use AI to extract physics from data in an efficient way, but we are moving towards the right direction!

Thank you!