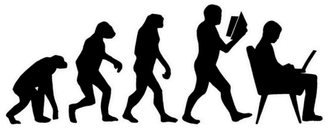
QNP2022 - The 9th International Conference on Quarks and Nuclear Physics

5-9 September 2022

A(i)DAPT
AI for Data Analysis and Preservation

M.Battaglieri (INFN)

on behalf of A(i)DAPT Working Group

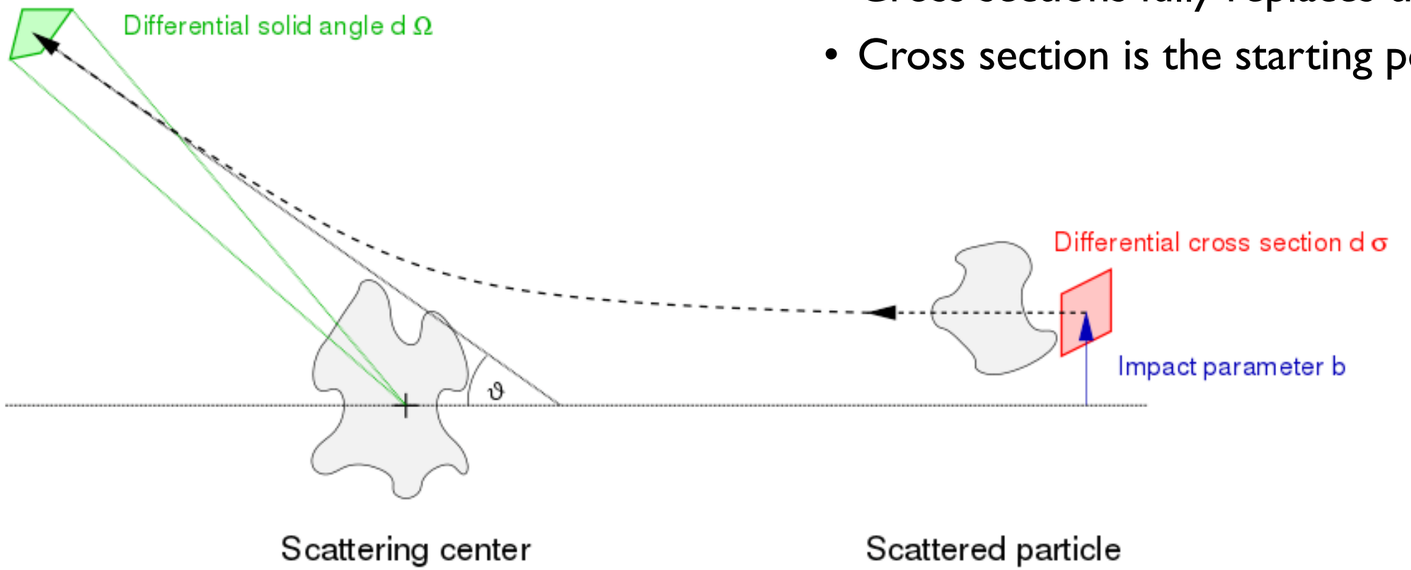


A(i)DAPT
AI for DAta PreservaTion

The cross section in particle physics

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = (2\pi)^4 m_i m_f \frac{p_f}{p_i} |T_{fi}|^2$$

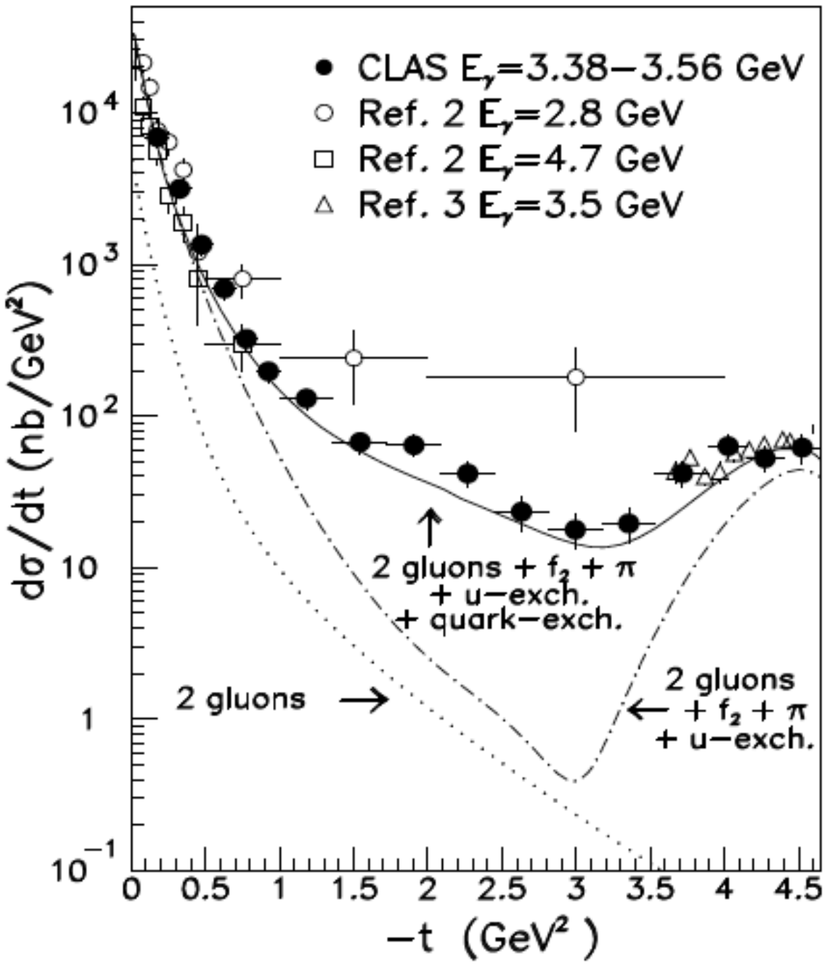
- The *cross section* is related to the transition probability between an initial to a final state
- In case of scattering, cross sections provides information about the elementary interaction
- Cross section is expressed as squared sum of scattering amplitudes (complex functions) depending on the kinematic Lorentz-invariant of the problem and embedding the interaction properties
- It is derived by measuring the momentum distributions of reaction particle (at different CM energy)
- Correlations between particles in the final state reflects the underlying dynamics
- Cross sections fully replaces the 4-mom data sample in a compact and efficient way
- Cross section is the starting point for any higher level physics analysis



- Traditional approach: particles (4-momenta) measured into the detector, extract the relevant observables, extract physics mechanisms
- Cross section **preserves** this information as replacement for the original particle-by-particle scattering information

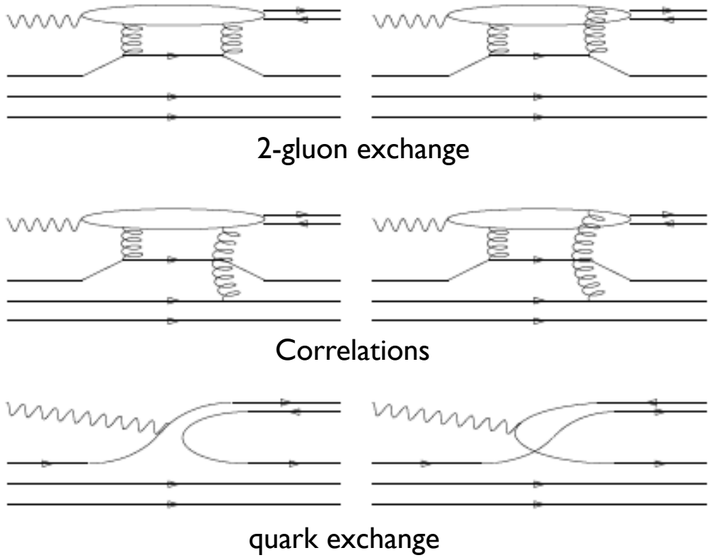
Exclusive reactions: 2 → 2

M. Battaglieri *et al.* (CLAS Collaboration) Phys. Rev. Lett. 90, 022002



JLab-CLAS g6 ω photo production at large momentum transfer

$\gamma p \rightarrow p \omega$



2 → 2 scattering (no polarisation)

- Initial state: known
- Final state: 2 x 3
- Parameters: (2 x 3) - 4 = 2
- Possible choice: *-t* and ϕ
- *t*he physics depends only on one variable (-t)

- It worked (and still works!) well if limited to channels with a single variable
- Xsec, Polarization observables, angular distribution, decay matrix, …

Exclusive reactions: 2 → 3

M. Battaglieri *et al.* (CLAS Collaboration) Phys. Rev. Lett. 90, 022002

2 → 3 scattering (no polarization)
- Initial state: known
- Final state: 3 x 3
- Parameters: (3 x 3) - 4 = 5
- Possible choice: $M^2_{\pi\pi}$, $M^2_{p\pi}$ θ_π, α, ϕ

- It does not work (in practice) when you have several independent variables: multi-particle final states (spectroscopy) or multi-variable correlations (SIDIS)
- In the integration to reduce to 1-dim all correlations are lost

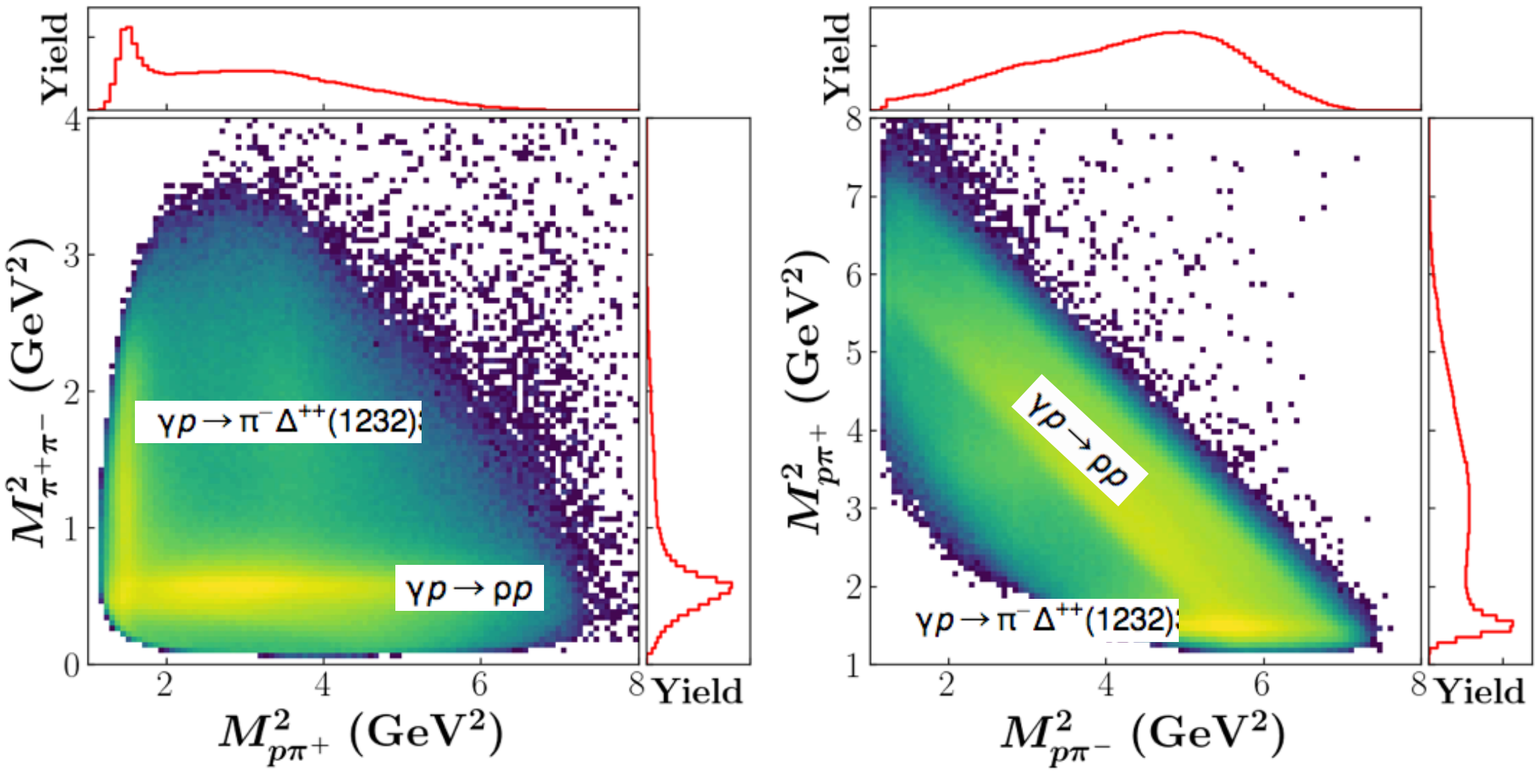
AI may provide a new way to look at data and extract observables and physics interpretation (on event by event base)

CLAS g11 2π photo production

- E_γ = (3.0 - 3.8) GeV

  $\gamma\, p \rightarrow p\, \pi^+\, \pi^-$ exclusive reaction
  - data set analyses so far $\gamma\, p \rightarrow p\, \pi^+$ (π^-) + small contamination of $\gamma\, p \rightarrow p\, \pi^+$ (more than a missing π^-)
  - complicated dynamic for the overlap of ($p\pi$) to form Δ baryon resonances and ($\pi\pi$) to form meson resonances

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



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 'true' observables (vertex-level)**
  - **Acceptance**: any measurements only access a limited region of the phase space.

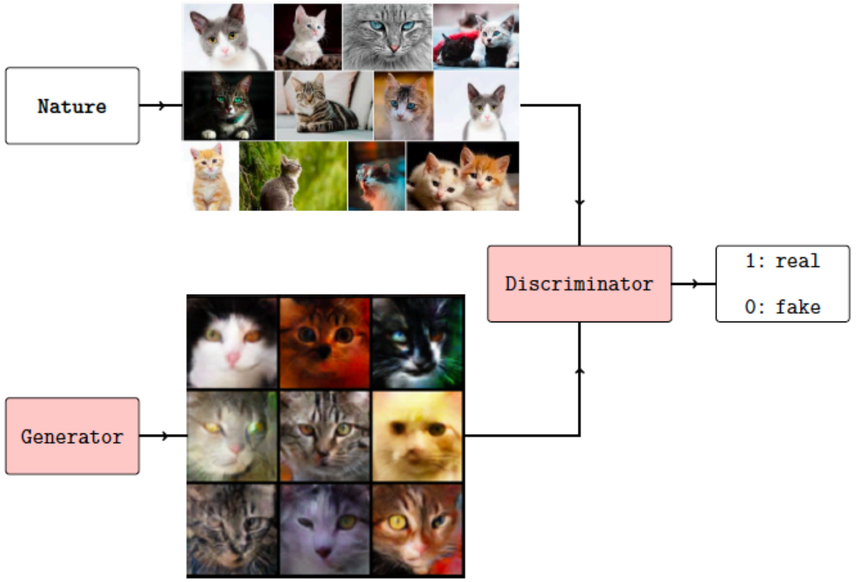
    How to recover the unmeasured region?
    - Interpolation: holes in the phase space
    - Extrapolation: border of the accessible phase space
  - **Resolution**: any measurements introduce an experimental resolution that may hide or washout the effect searched for
    - A spike could be not resolved, the measurement may extend in an unphysical region (e.g. negative squared missing mass)
- **For both effects, one needs to quantify the systematic errors introduced to the vertex-level observables**

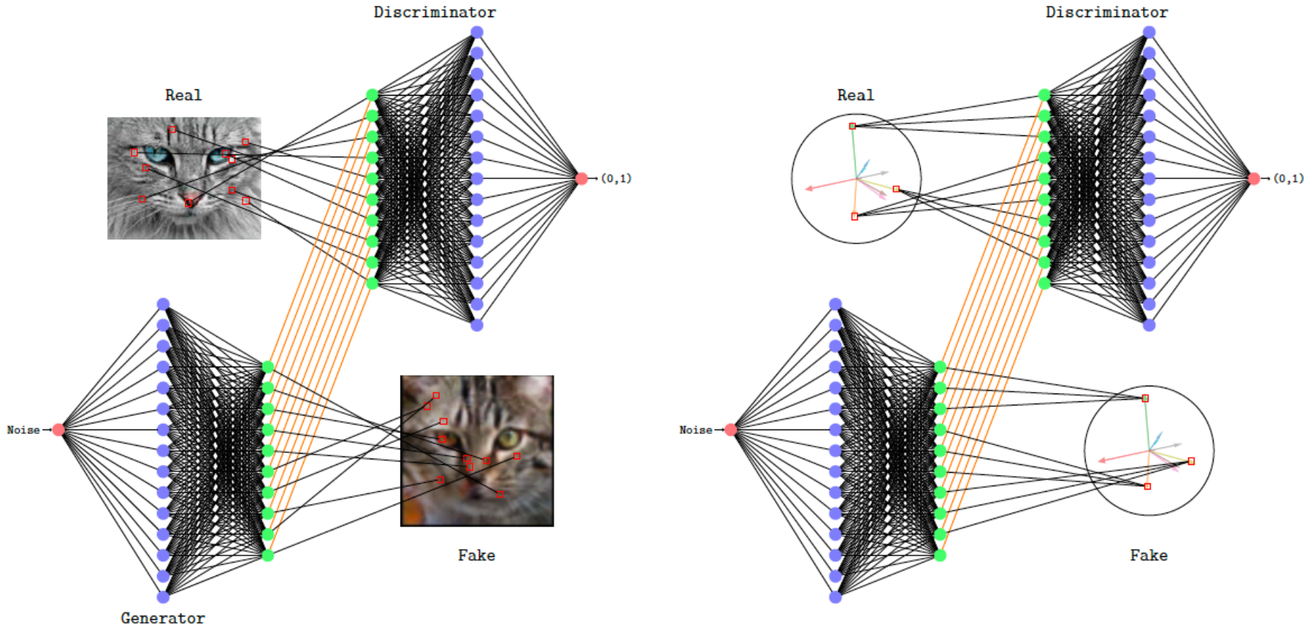
- **Mitigation strategy:**
  - Acceptance: '*fiducial volumes*' to exclude unmeasured regions verifying the training convergence
  - Resolution: closure test with a reasonable model of the detector using a detector proxy (parametric or GEANT-based)

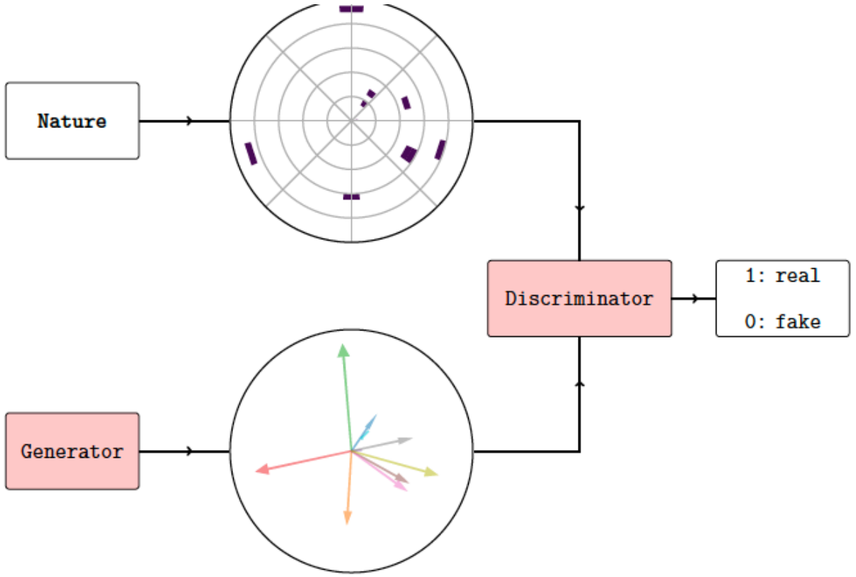
https://doi.org/10.24963/ijcai.2021/588

Generative Adversarial Network (GANs)

- The colored boxes are built using NNs
- Discriminator is trained to output "real" for Nature samples
- Generator is trained to fool the discriminator
- The Generator can be used as data compression tool
- Typical size for the Generator: O(1MB) - to be compared to NP/HEP experiments data set O(1GB/TB)
- Simple to distribute instead of stored events on tapes





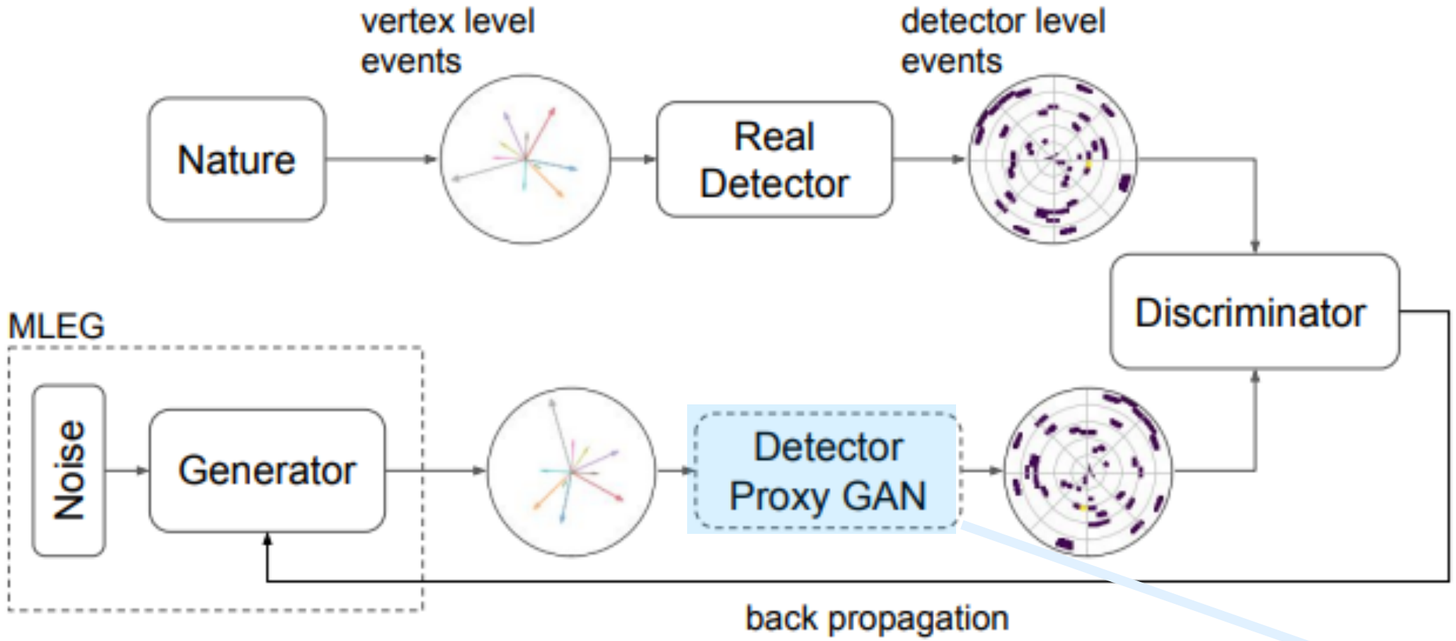


Credit: Y. Li, N.Sato

A simple study case: DIS scattering

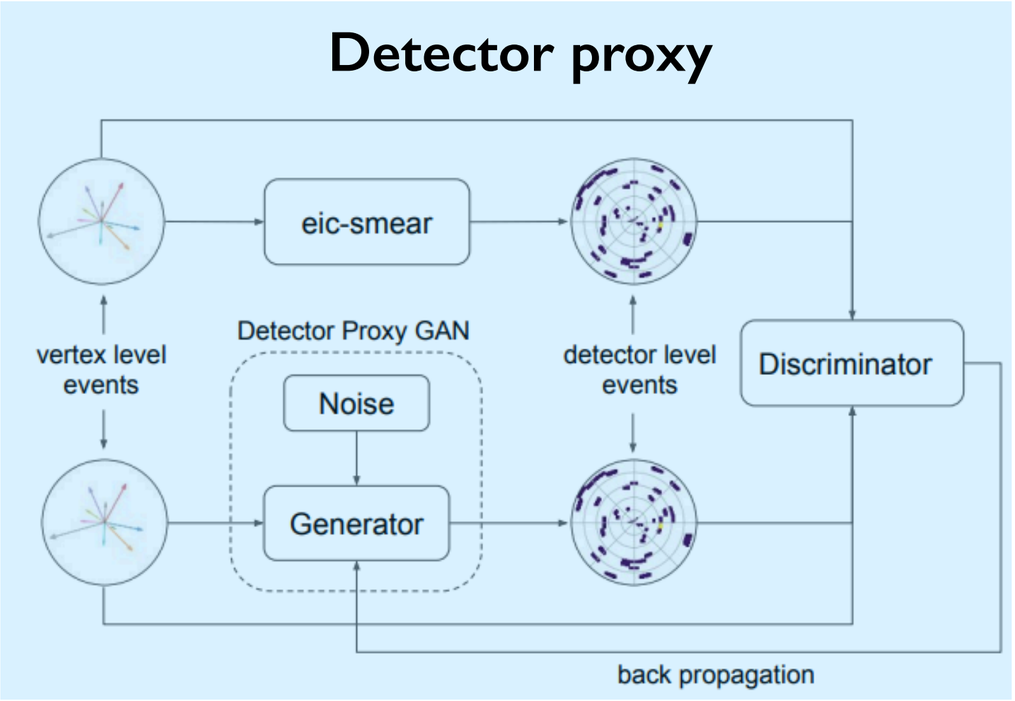
https://arxiv.org/abs/2008.03151v2

ML Event Generator GAN scheme



- 100-d white noise entered at 0, unit standard dev.
- Generator: 5 hidden layers / 512 neurone per layer, ReLU activation function. Last layer connected to 2 neurons output to generate v_1 and v_2 variables
- Discriminator: same NN architecture as for the generator
- Detector proxy: similar architecture
- Least Squares GAN (LSGAN)
- Trained adversarially for 100000 epochs (pass through the training data set)
- Adam's optimizer

Detector proxy



- *eic-smear*: parametric smearing routine for the Electron Ion Collider detectors (no GEANT-based simulations)
- Parameters tuned to reproduce ZEUS/H1 detectors
- Full 4π acceptance

A simple study case: DIS scattering

I) GAN training w/o detector effects

Pseudo-data sample (JAM)

- Inclusive electron DIS generated at E_{CM}=318.2 GeV (HERA kinematics)
- 2-dim differential cross section $d\sigma/dxdQ^2$
- Lorentz boosted from CM to Lab (+ uniform azimuthal angle)
- To reduce violation of momentum conservation on the edge of the phase space due to smearing effects, electron momentum is replaced by new variables:

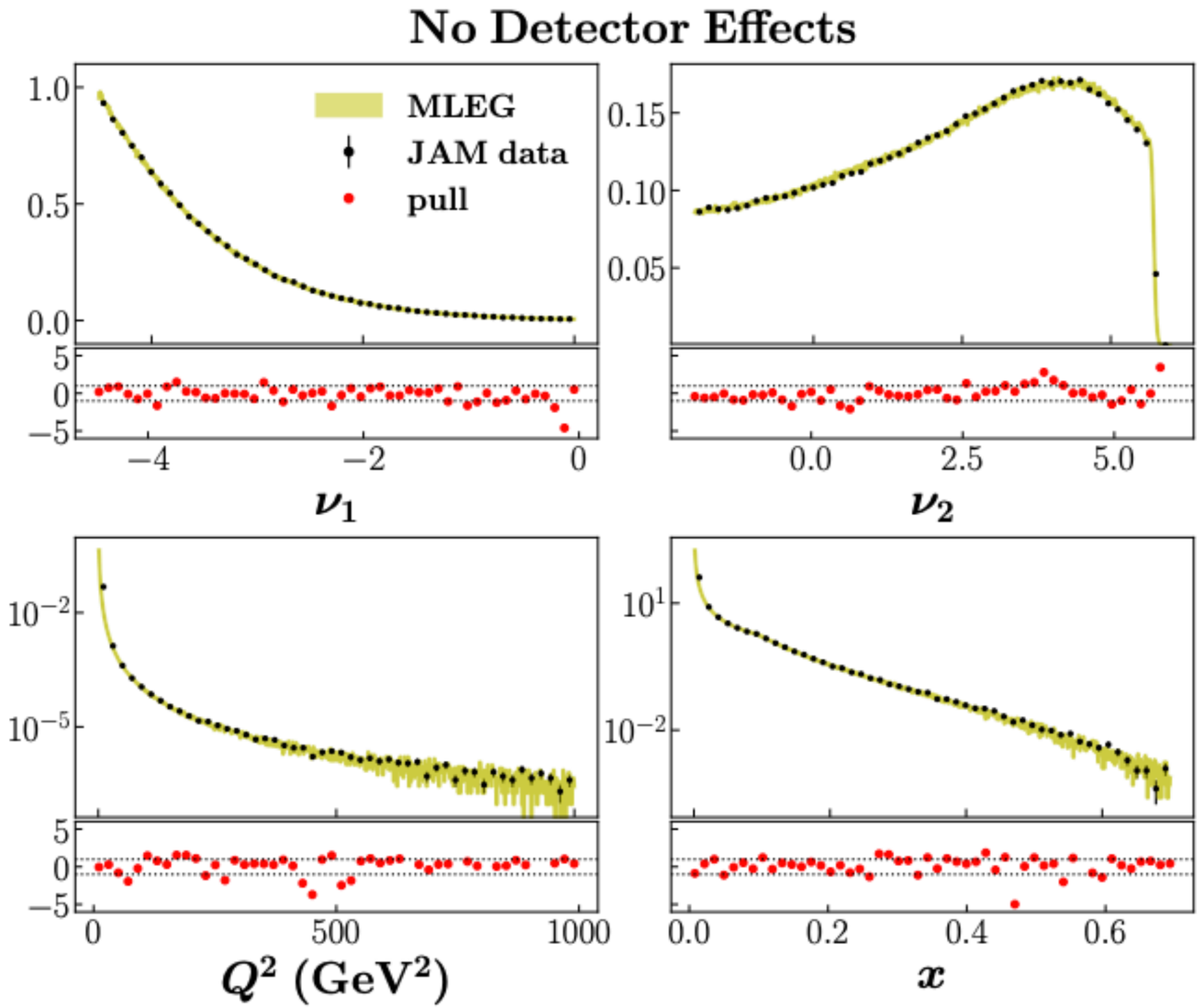
$$\nu_1 = \ln\left((k'_0 - k'_z)/1\,\mathrm{GeV}\right),$$
$$\nu_2 = \ln\left((2E_e - k'_0 - k'_z)/1\,\mathrm{GeV}\right),$$

Uncertainty Quantification via *pull* calculation

- Metric: *pull*

$$\mathrm{pull} = \frac{\mathrm{E}\left[\mathcal{P}(\mathcal{O}|\mathrm{bin})\right]_{\mathrm{GAN}} - \mathrm{E}\left[\mathcal{P}(\mathcal{O}|\mathrm{bin})\right]_{\mathrm{JAM}}}{\sqrt{\mathrm{V}\left[\mathcal{P}(\mathcal{O}|\mathrm{bin})\right]_{\mathrm{GAN}} + \mathrm{V}\left[\mathcal{P}(\mathcal{O}|\mathrm{bin})\right]_{\mathrm{JAM}}}}$$

- *Bootstrap with* 10 independently trained GANs



A simple study case: DIS scattering

I) GAN training w/o detector effects

Pseudo-data sample (JAM)

- Inclusive electron DIS generated at E_{CM}=318.2 GeV (HERA kinematics)
- 2-dim differential cross section dσ/dxdQ2
- Lorentz boosted from CM to Lab (+ uniform azimuthal angle)
- To reduce violation of momentum conservation on the edge of the phase space due to smearing effect, electron momentum is replaced by new variables:

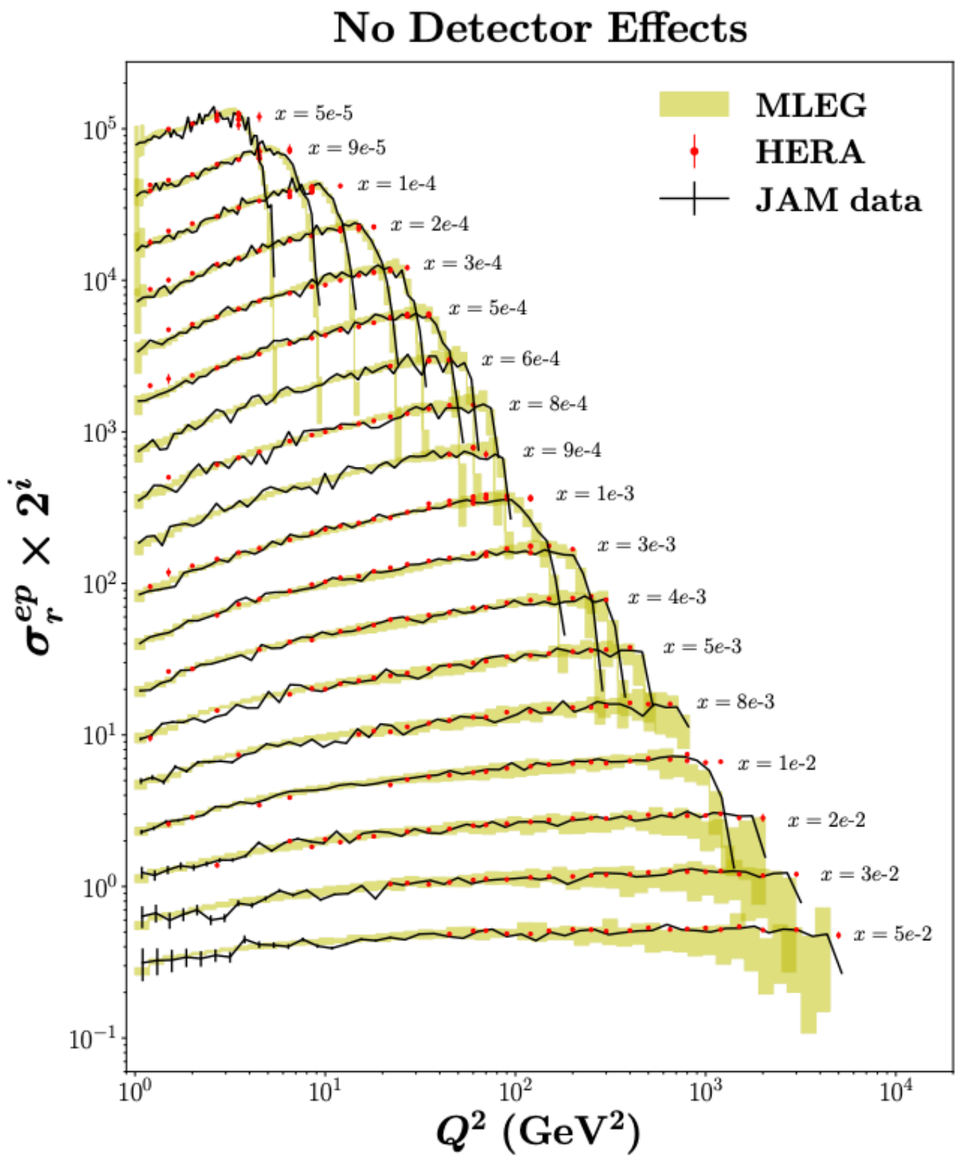
$$\nu_1 = \ln\left((k'_0 - k'_z)/1\,\mathrm{GeV}\right),$$
$$\nu_2 = \ln\left((2E_e - k'_0 - k'_z)/1\,\mathrm{GeV}\right),$$

Uncertainty Quantification via *pull* calculation

- Metric: *pull*

$$\mathrm{pull} = \frac{\mathrm{E}\left[\mathcal{P}(\mathcal{O}|\mathrm{bin})\right]_{\mathrm{GAN}} - \mathrm{E}\left[\mathcal{P}(\mathcal{O}|\mathrm{bin})\right]_{\mathrm{JAM}}}{\sqrt{\mathrm{V}\left[\mathcal{P}(\mathcal{O}|\mathrm{bin})\right]_{\mathrm{GAN}} + \mathrm{V}\left[\mathcal{P}(\mathcal{O}|\mathrm{bin})\right]_{\mathrm{JAM}}}}$$

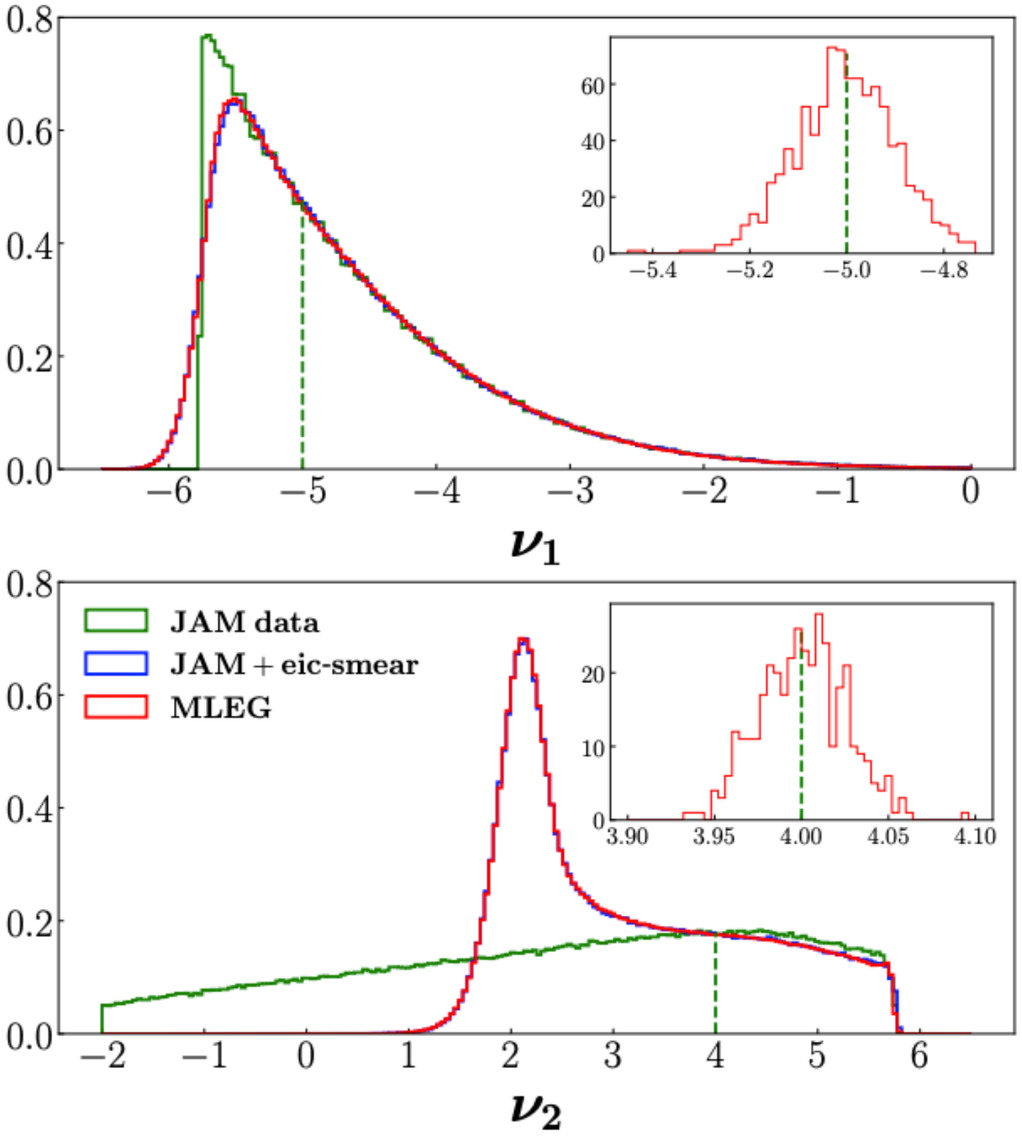
- *Bootstrap with* 10 independently trained GAN



A simple study case: DIS scattering

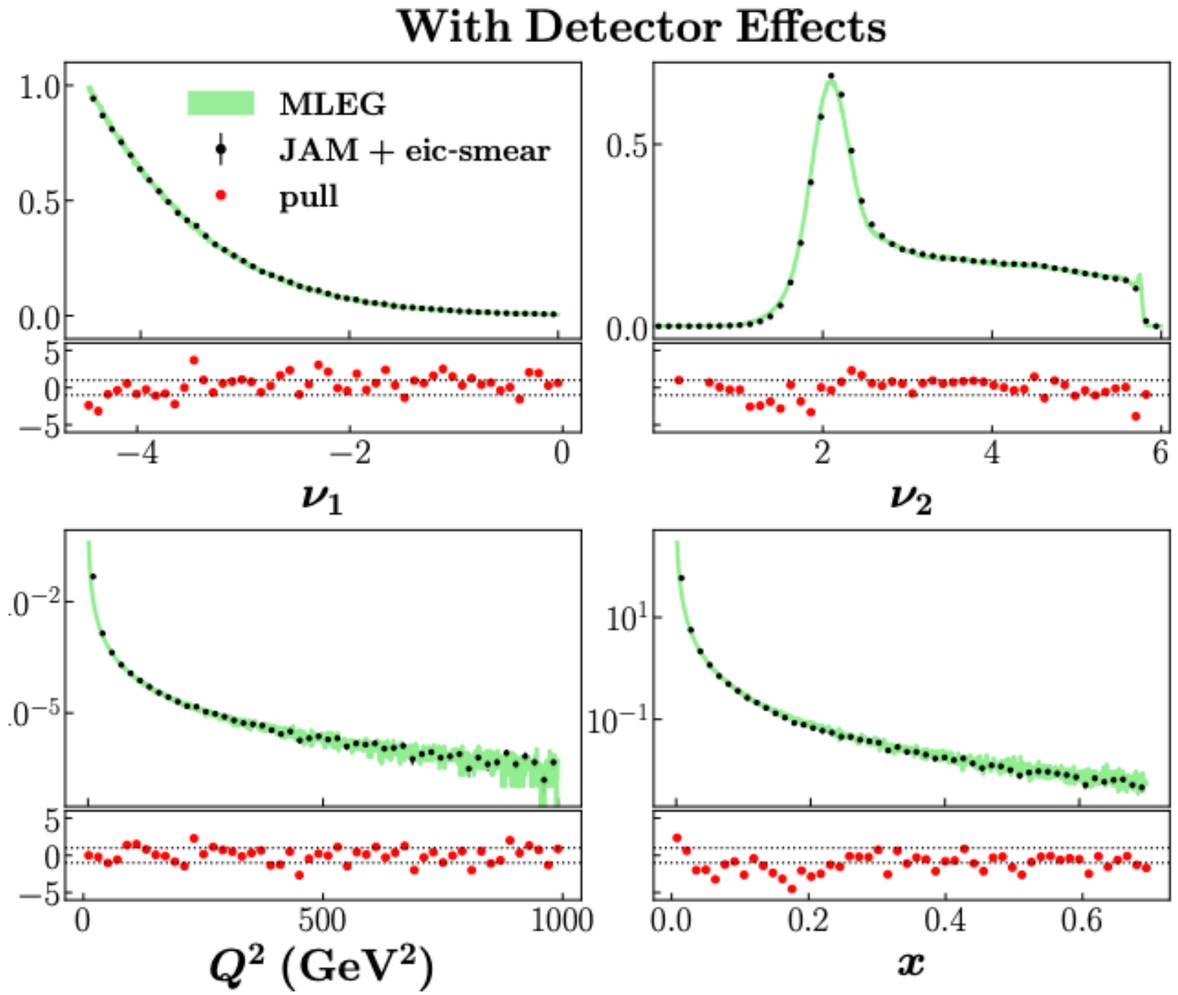
II) GAN training WITH detector effects

- eic-smear introduces significant distortions to the detector level sample in particular on v_2

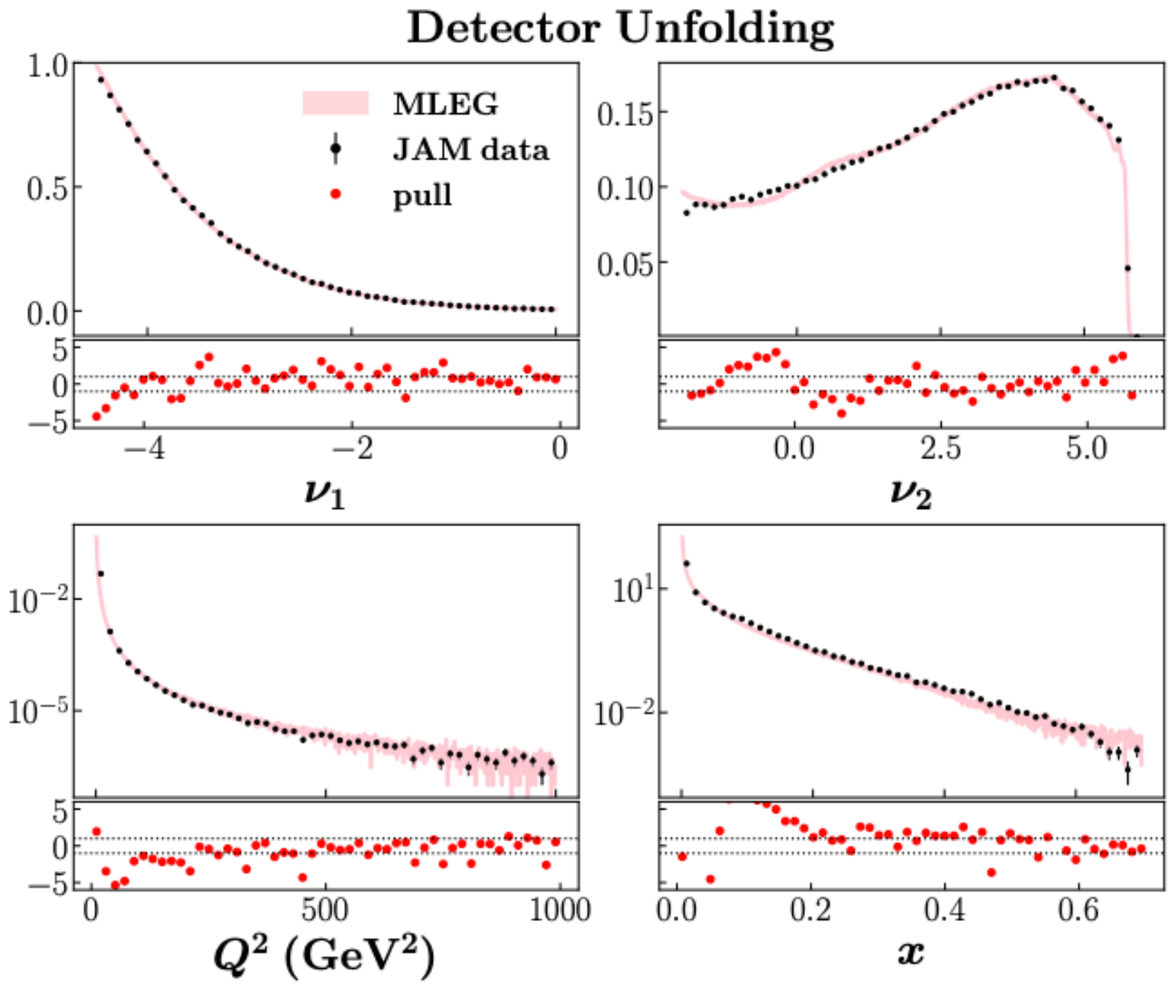


A simple study case: DIS scattering

II) GAN training with detector effects



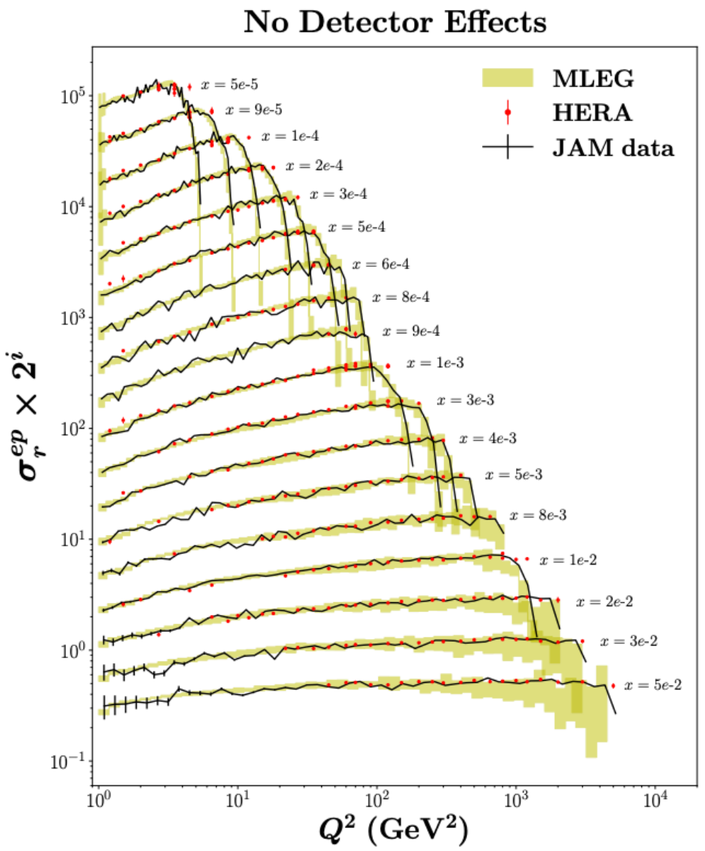
Detector-level

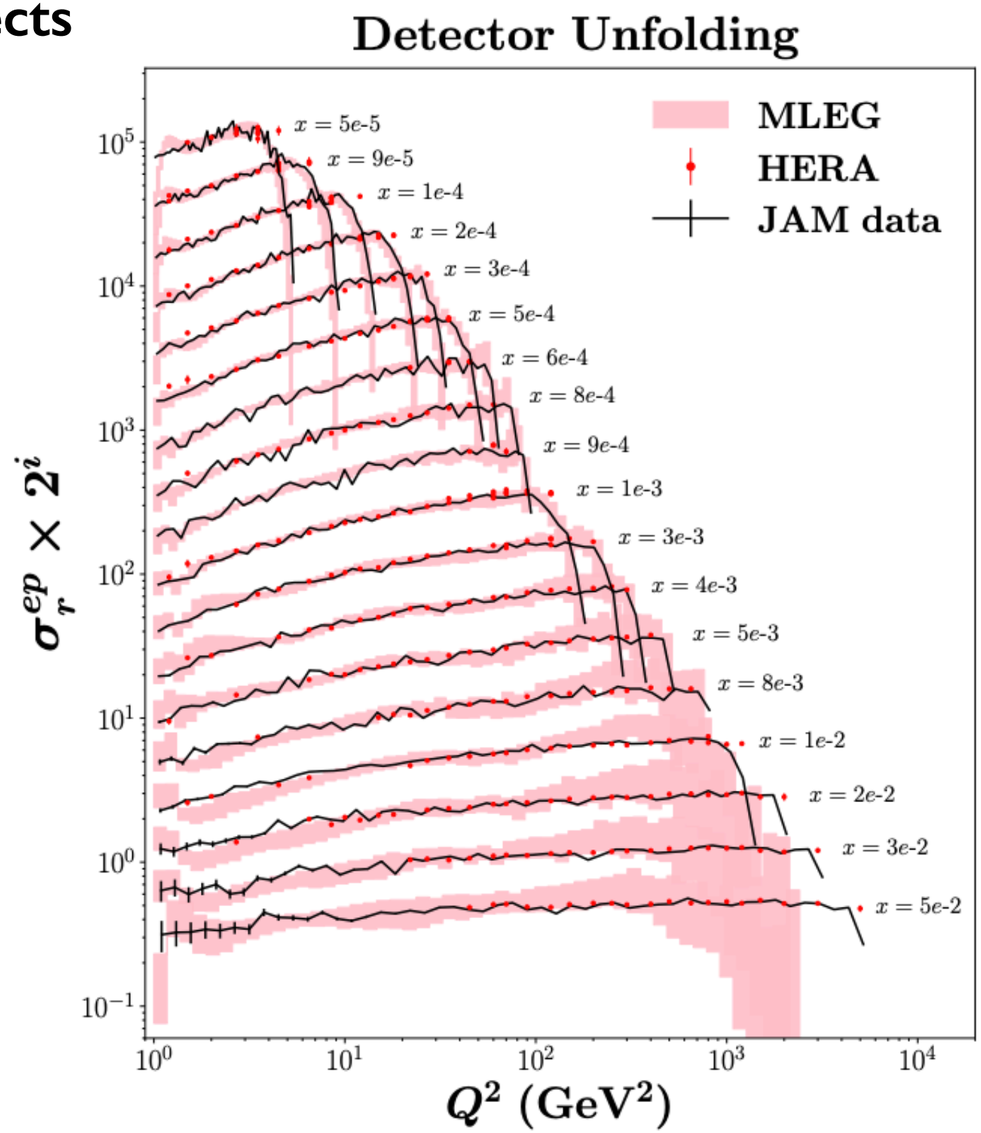


Vertex-level

A simple study case: DIS scattering

II) GAN training with detector effects





Conclusions

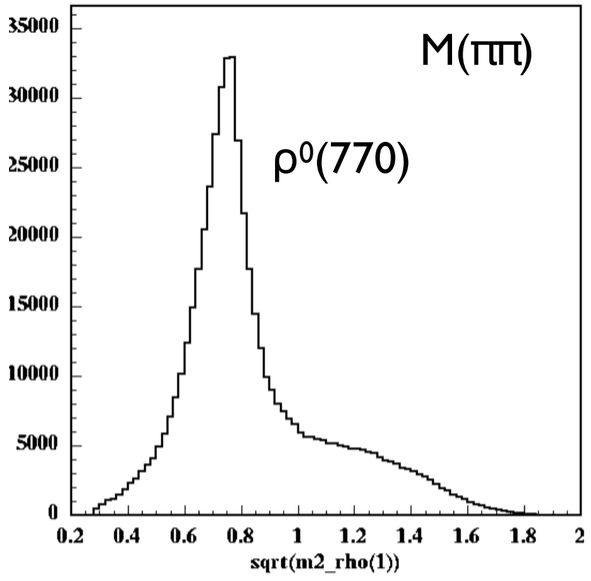
- The 'closure test' was successful
- 'generated' events (trained on generated)
- 'reconstructed' events (trained on reconstructed)
- 'generated' (trained on reconstructed) with larger error bars, in particular on the edge of the phase space
- The PDFs are correctly recovered

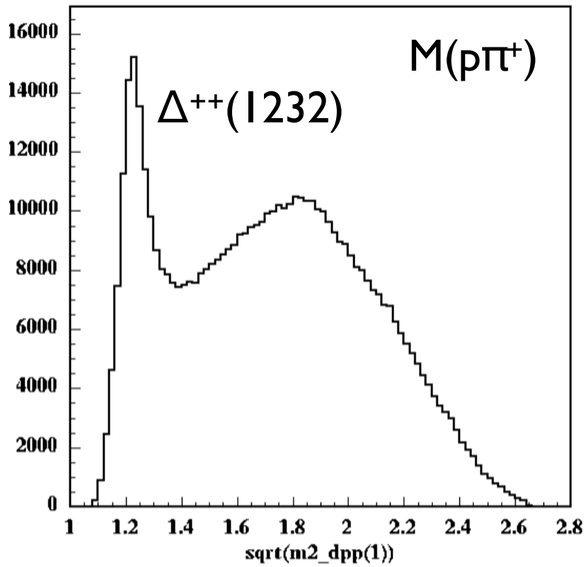
Multi-d xsec: 2π photo production

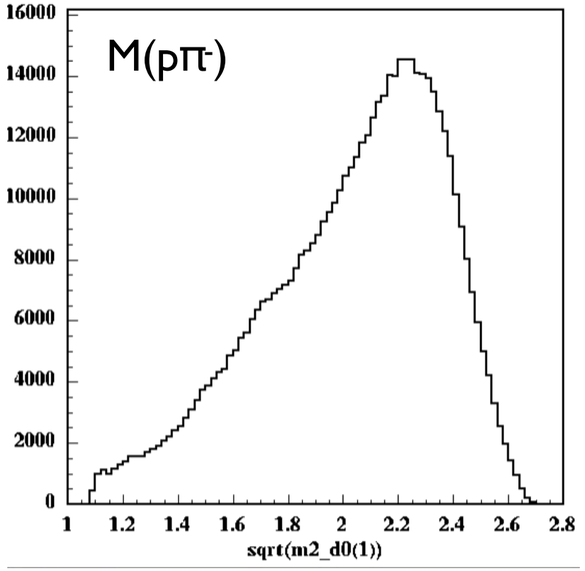
M. Battaglieri *et al.* (CLAS Collaboration) Phys. Rev. Lett. 102, 102001
M. Battaglieri *et al.* (CLAS Collaboration) Phys. Rev. D 80, 072005

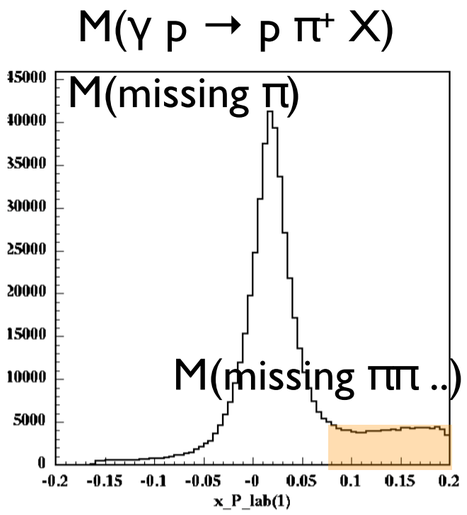
CLAS g11 data set

- Same data set used by CLAS Collaboration for many publications
- Fiducial cuts (p, Θ, ϕ) as used in published analysis
- All four topologies available but only focused on $\gamma\, p \rightarrow p\, \pi^+\, (\pi^-)$
- Final exclusive 2π state identified by missing mass technique (variables reconstructed by energy/momentum conservation)
- Multipion background comes from $\gamma\, p \rightarrow p\, \omega^0 \rightarrow p\, \pi^+\, \pi^-\, \pi^0$
- At E_γ=3-4 GeV reaction dynamics dominated by ρ^0 photo production ($\gamma\, p \rightarrow p\, \rho^0$) and Δ^{++} resonance excitation ($\gamma\, p \rightarrow \Delta^{++}\, \pi^-$)





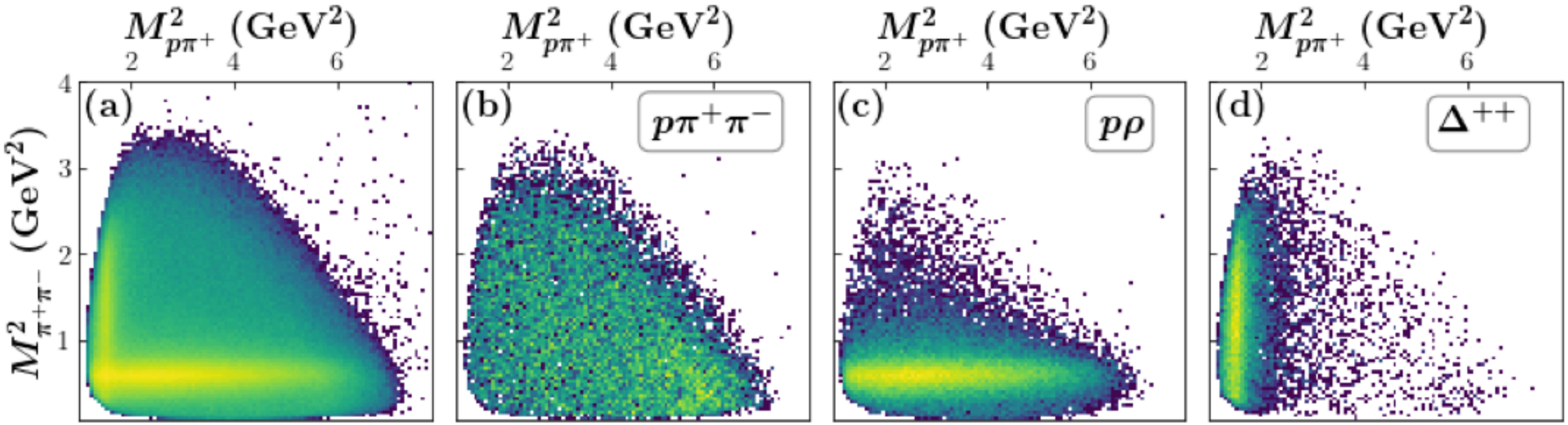




Multi-d xsec: 2π photo production

Pseudo-data event generator

- Two event generators
  - I) REALISTIC used to mimic real data. Includes measured xsec, angular distributions, dominant mechanism (ρ^0 and Δ^{++} channels)
  - II) PHASE SPACE used to train the detector proxy GAN. Based on phase space with uniform distributions
- Reconstructed events are passed through the full analysis chain (GEANT detector simulation + CLAS data reconstruction code)

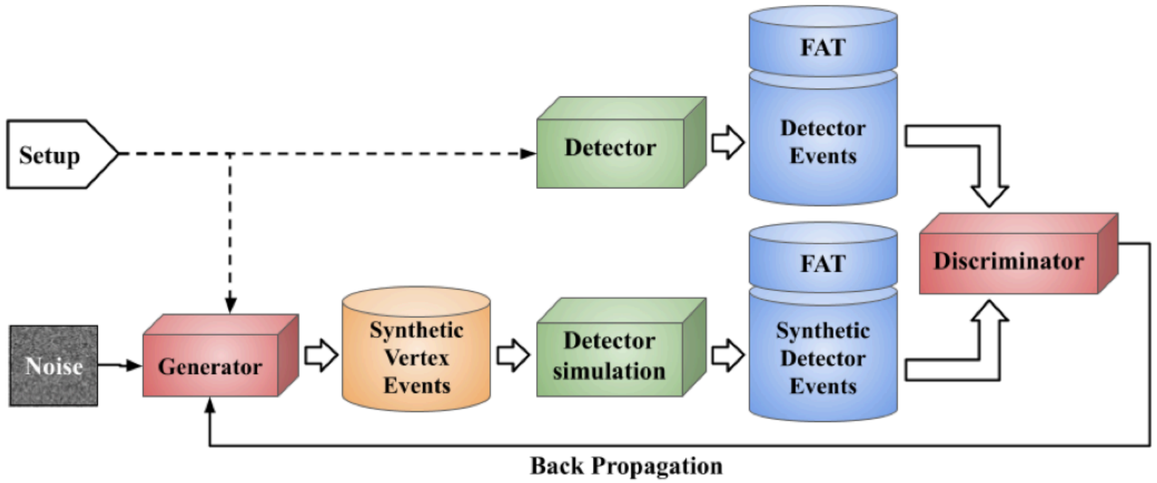


Credit: Y. Alanazi Awadh, , P.. Ambrozewicz, G. Costantini A. Hiller Blin, E. Isupov, T. Jeske, Y. Li, L. Marsicano W. Menlnitchouk, V. Mokeev, N. Sato, A. Szczepaniak, T. Viducic

GAN training on data (no detector unfolding)

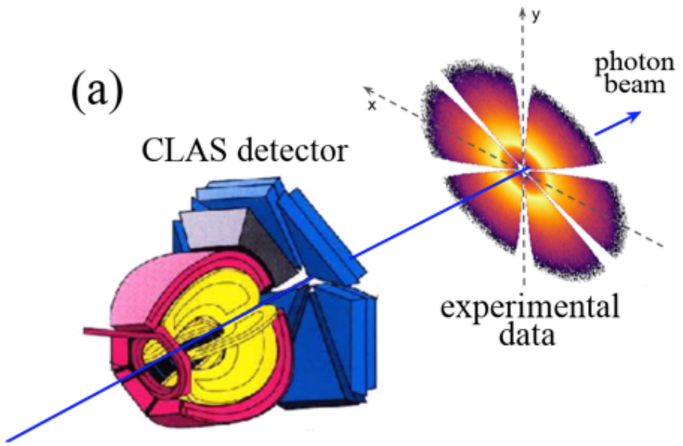
- Train a NN to generate events (synthetic) with the SAME correlations of experimental data
- Replace the xsec with a NN (synthetic data are equivalent to data)
- Light GAN-based event generator can generate any statistics

GAN architecture

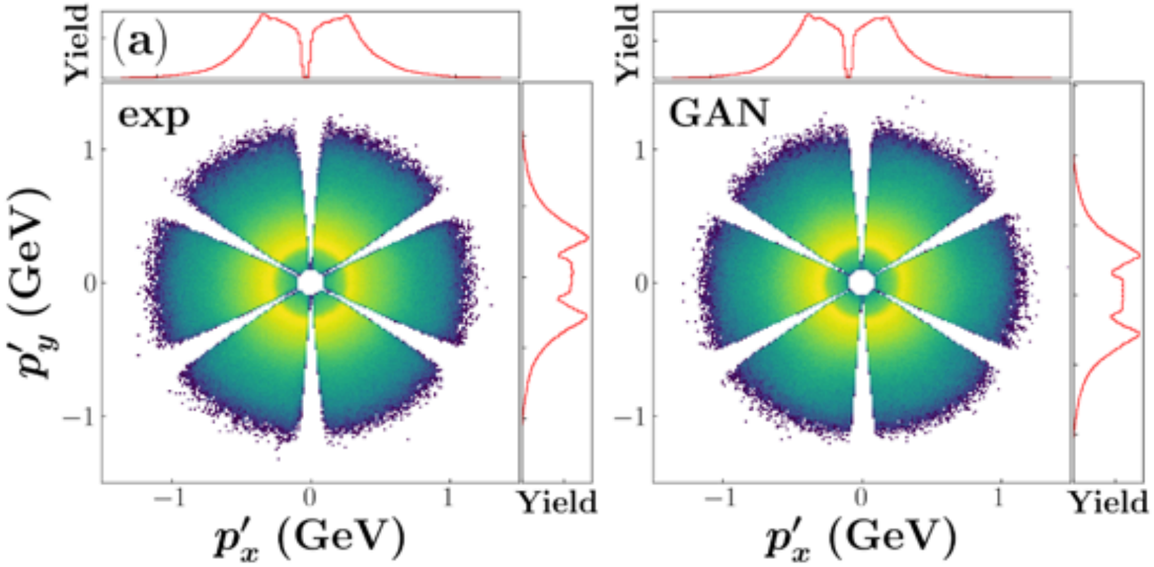


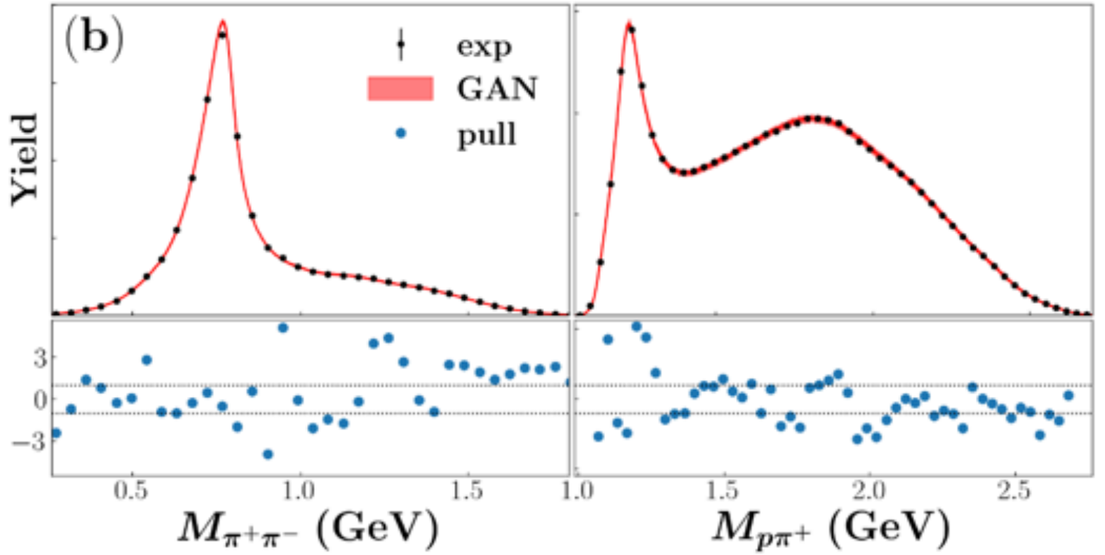
- Tested different options: 4-moms in CM or LAB or using kinematics INVARIANTS
- Implemented a folding/unfolding procedure to take into account detector acceptance/resolution
- Training performed on 10% of the data set to use the others for systematic studies. When optimised the training will be done on the whole data set
- Results are validated by comparing data/synt 1-dim projected and selected bins in 5-d space in LAB, CM and INVARIANTS space
- Error quantified to include statistical (via bootstrap) and systematic (different data samples, different reference systems)
- The final check is performed comparing Y_{LM} moments extracted from data and synt-data (in the assumption that any further physics observables will be extracted from Y_{LM})

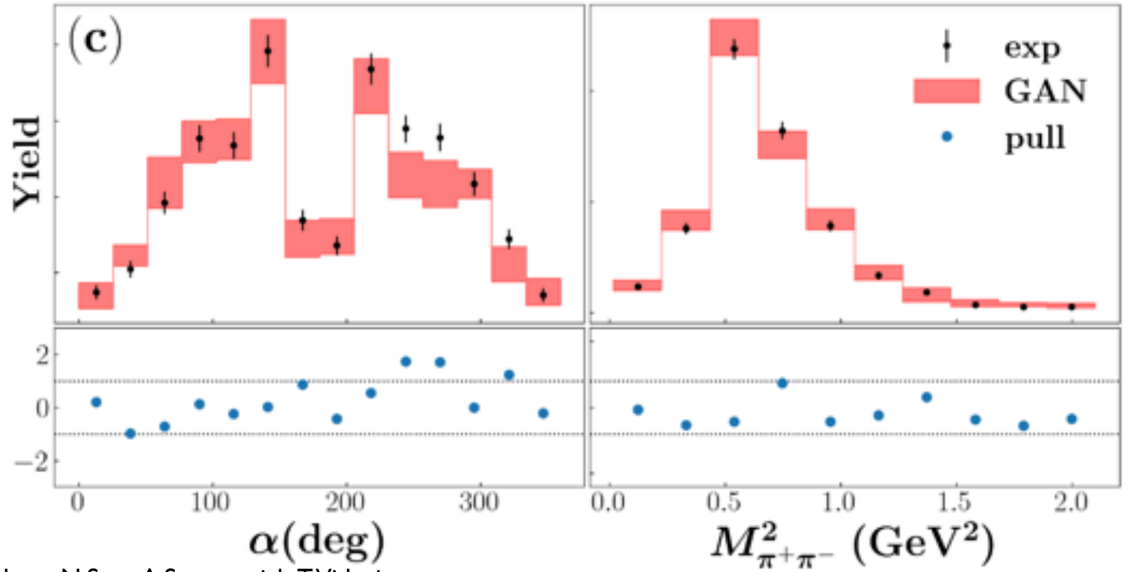
Synthetic vs data (no detector unfolding)



- These results did not correct for detector effects (no unfolding)
- Despite CLAS distortions (mainly acceptance), GAN is able to mimic the data set







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

Moments of the angular distribution (no detector unfolding)

$$\langle Y_{\lambda\mu}\rangle(E_\gamma,t,M)=\frac{1}{\sqrt{4\pi}}\int d\Omega_\pi \frac{d\sigma}{dtdMd\Omega_\pi}Y_{\lambda\mu}(\Omega_\pi)$$

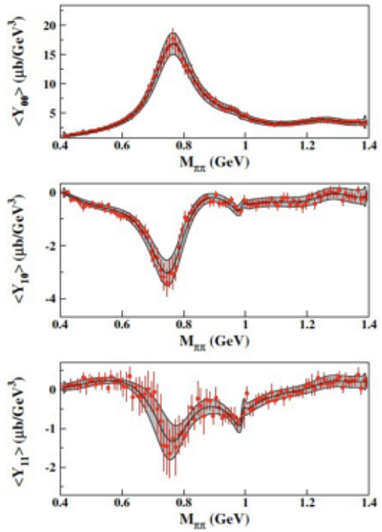
Amplitude parametrization (Dispersion relation)

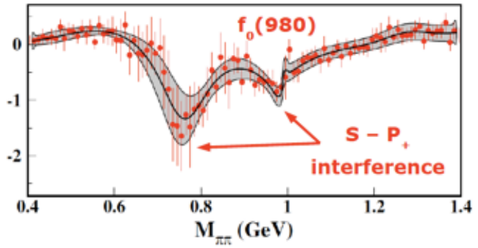
Related to ππ scattering matrix: phase-shift, inelasticity, S-P-D-F amplitude in 0.4 GeV < $M_{\pi\pi}$ < 1.4 GeV

$$a_{lm,I}(s)=\frac{1}{2}[I+S_{lm,I}(s)]\tilde{a}_{lm,I}(s)-\frac{1}{\pi}D^{-1}_{lm,I}(s)PV\int_{s_{th}} ds'\frac{N_{lm,I}(s')p(s')\tilde{a}_{lm,I}(s')}{s'-s}$$

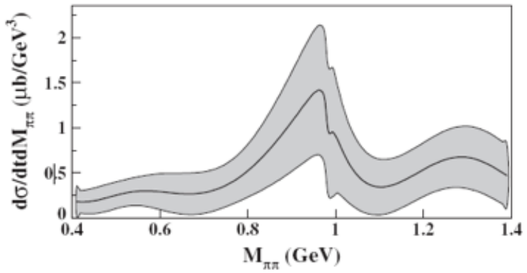
$$\tilde{a}_{lm,I}=[\mathcal{A}+\mathcal{B}s+\mathcal{C}s^2+\cdots][k]$$

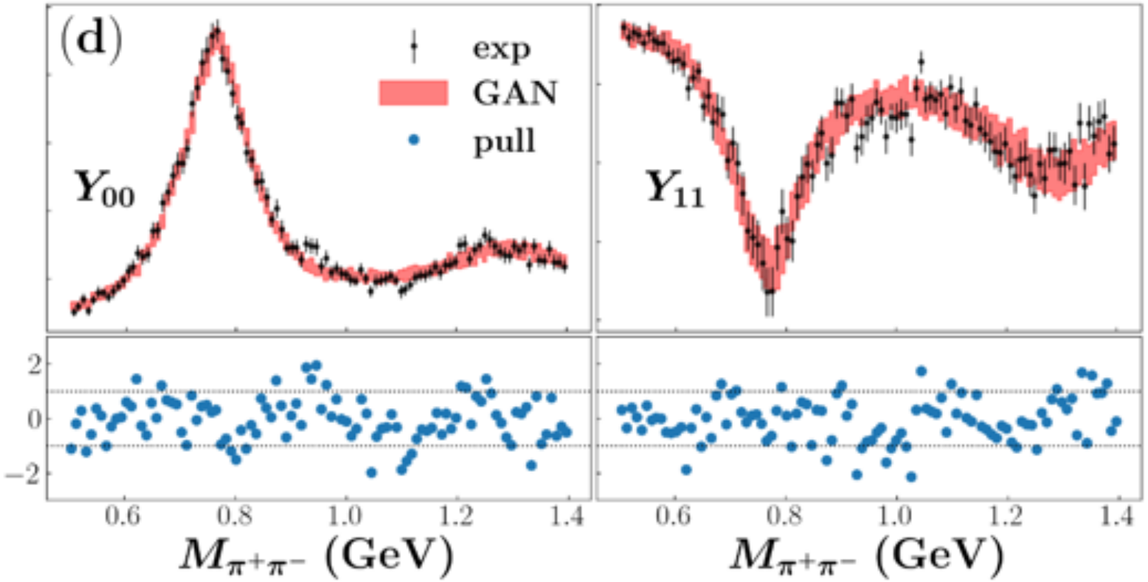
Expanded in a Taylor series: coefficient fit to the experimental moment





First observation of the f0(980) in a photoproduction experiment

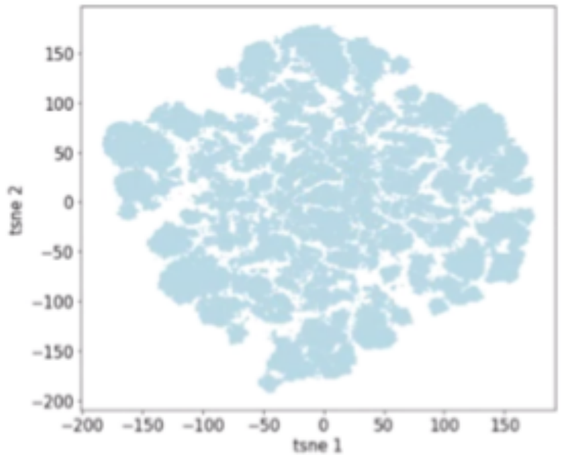




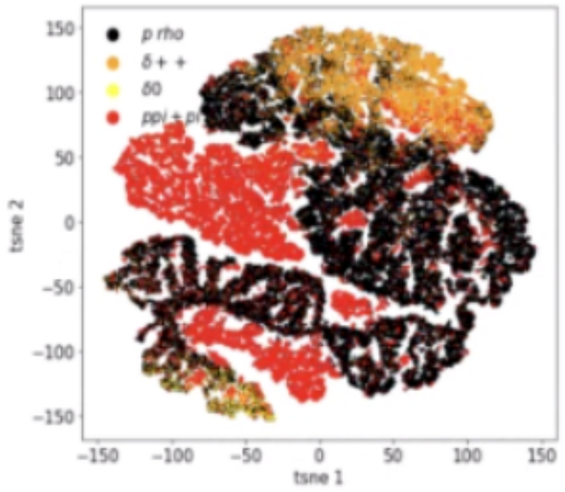
- Moments extracted from synthetic data copy moments extract from data

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

tSNE analysis (no detector unfolding)

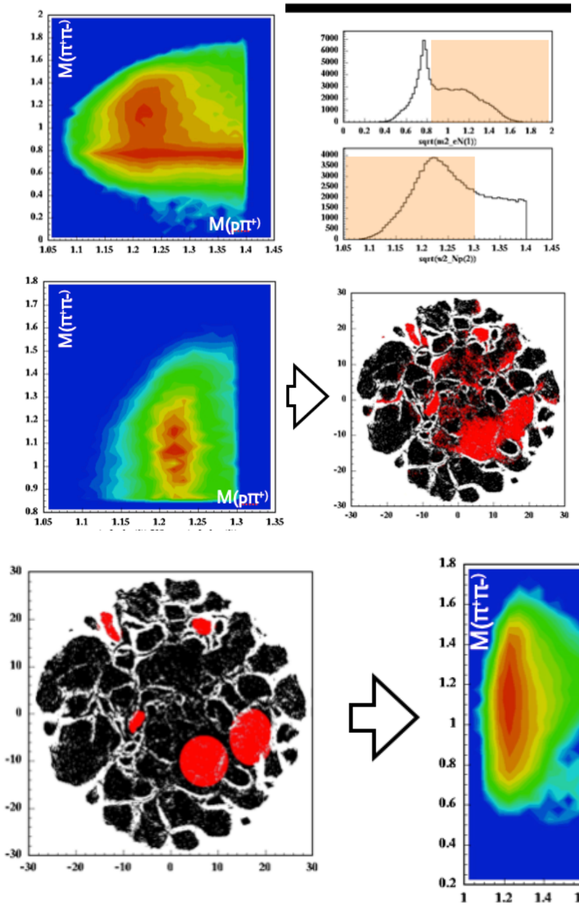


CLAS

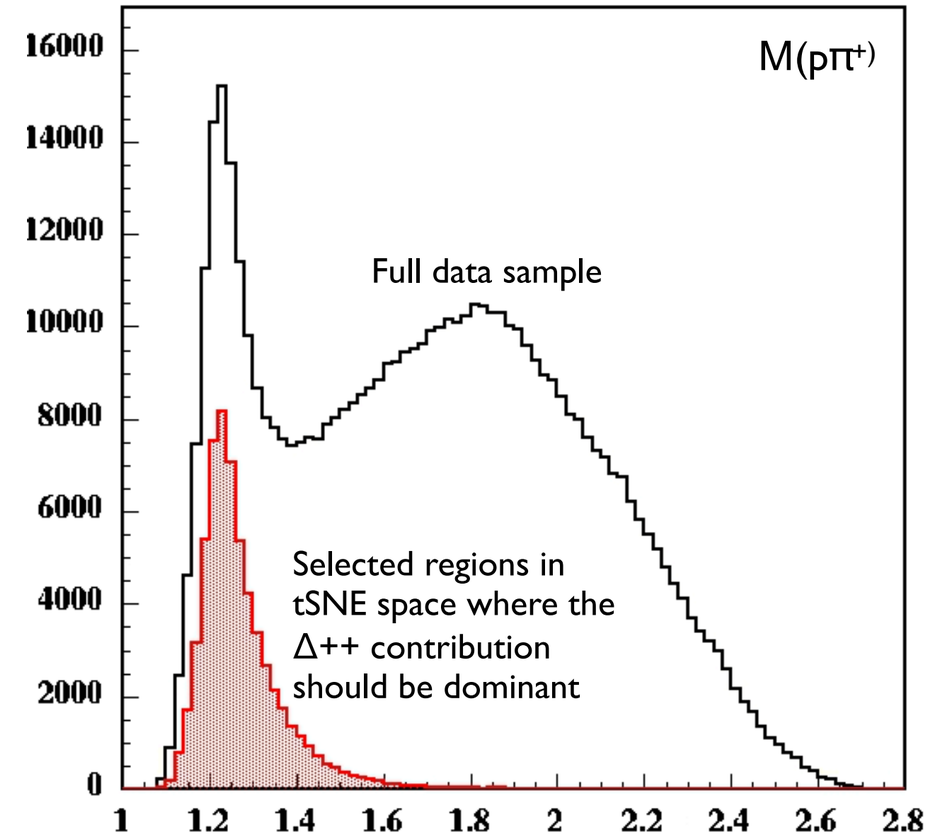


MC-Realistic (gen)

- **t-SNE analysis strategy**
  - explore correlations with multi-pion final states
  - Select the prominent Δ++(1232) peak and check correlations
  - Select areas of t-SNE1.vs.t-SNE2 space to demonstrate meson/baryon systems separation

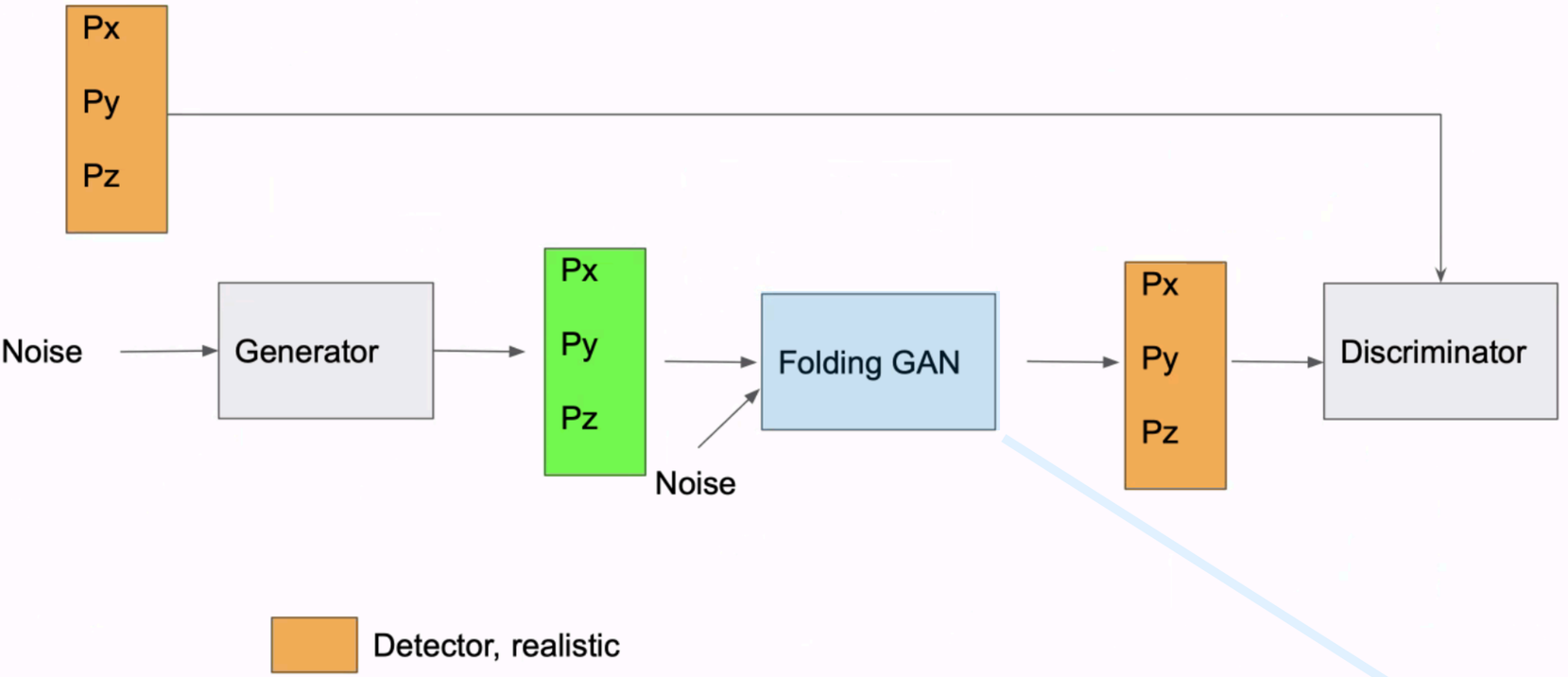


Red regions are associated with the Δ++ selection cuts

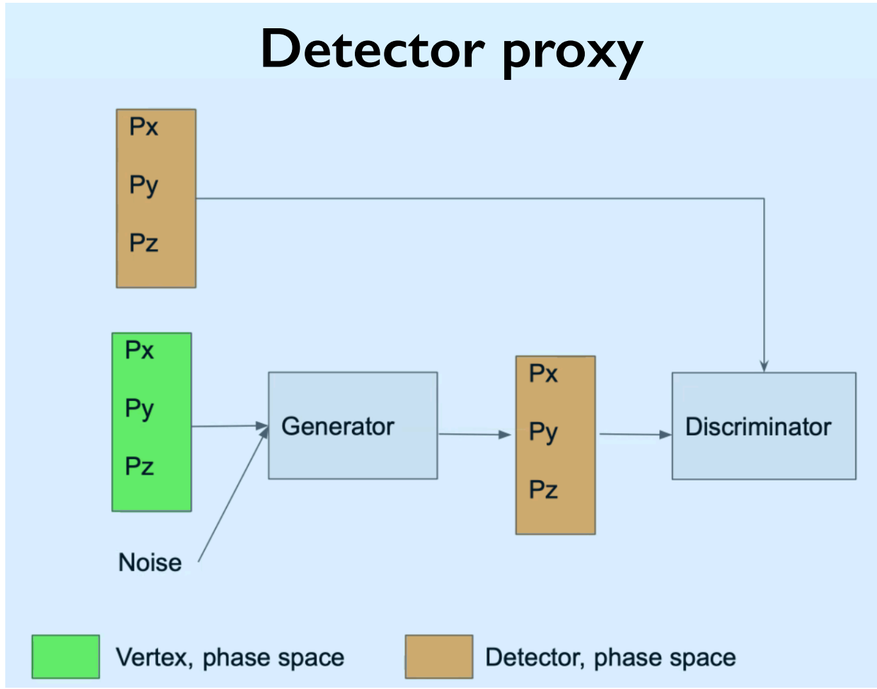


Closure test (using simulations)

ML Event Generator GAN scheme

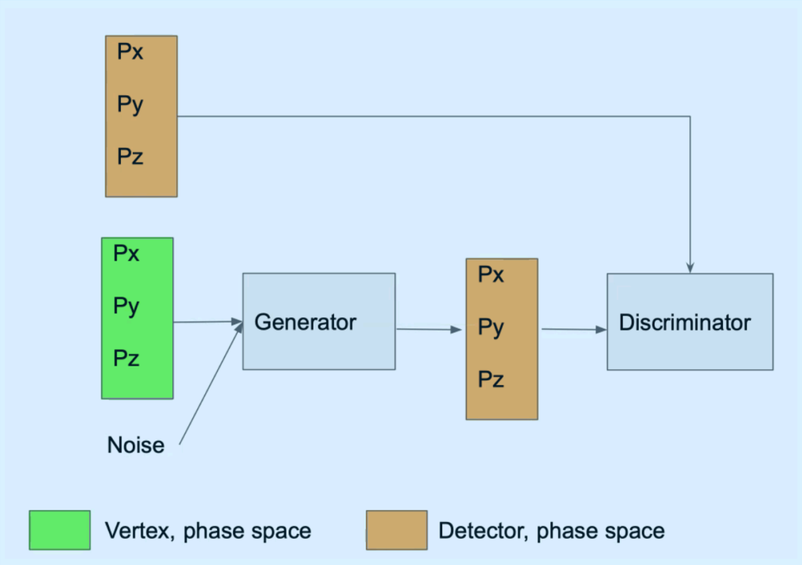


Detector proxy

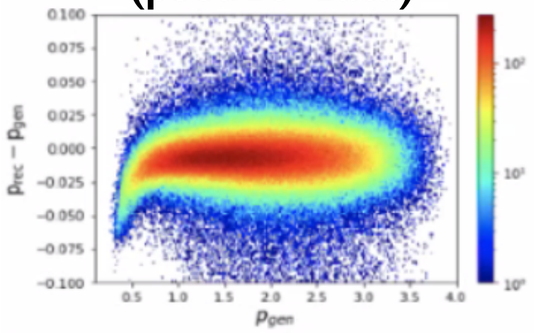


Credit: Y.Alanazi Awadh, T.Alghamdi, Y.Li

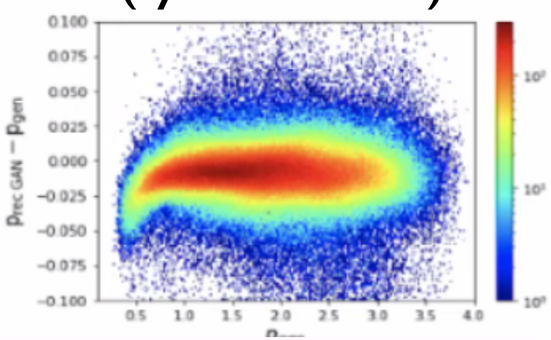
Closure test (using simulations)

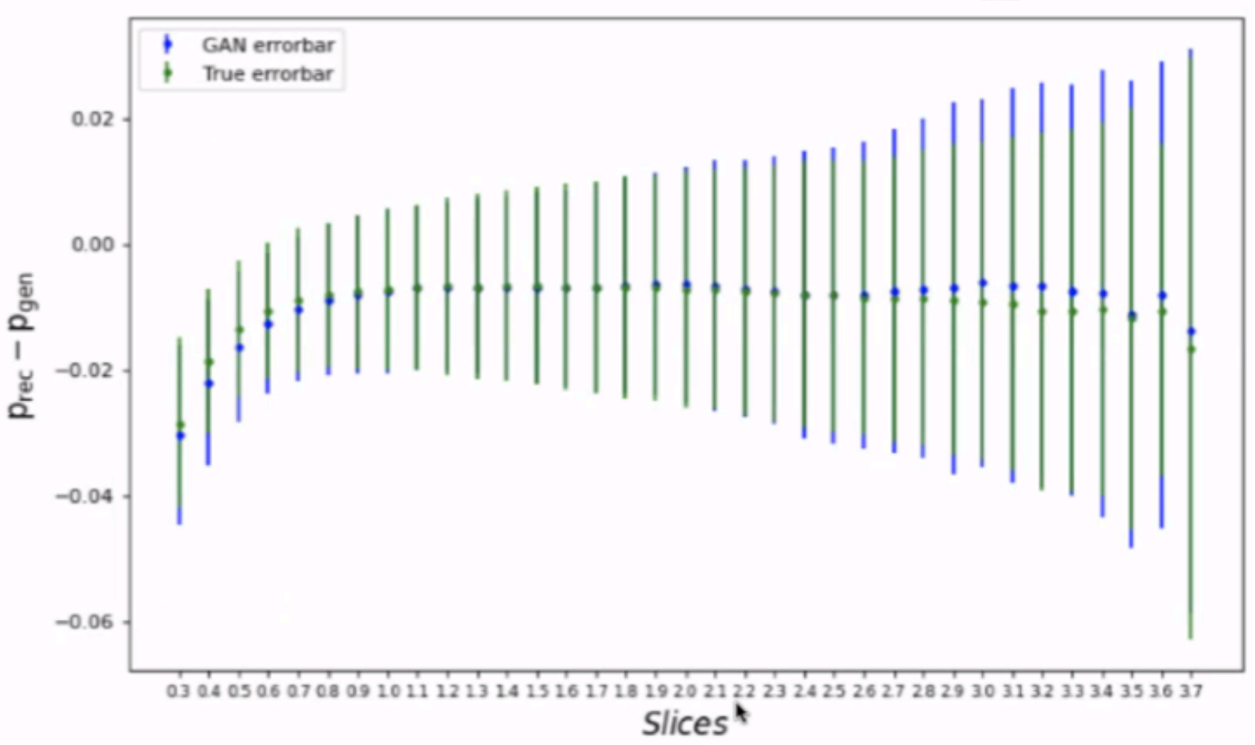


Smearing (pseudo-data)



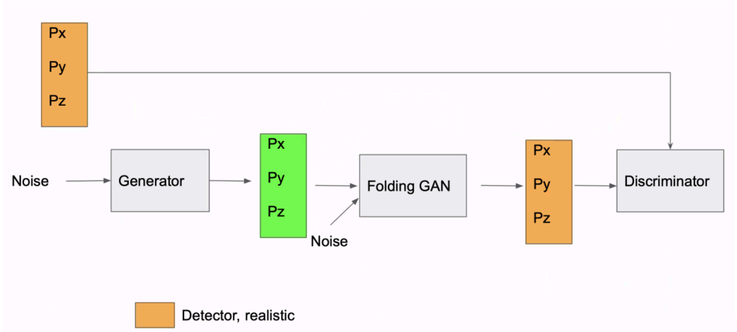
Smearing (synthetic-data)





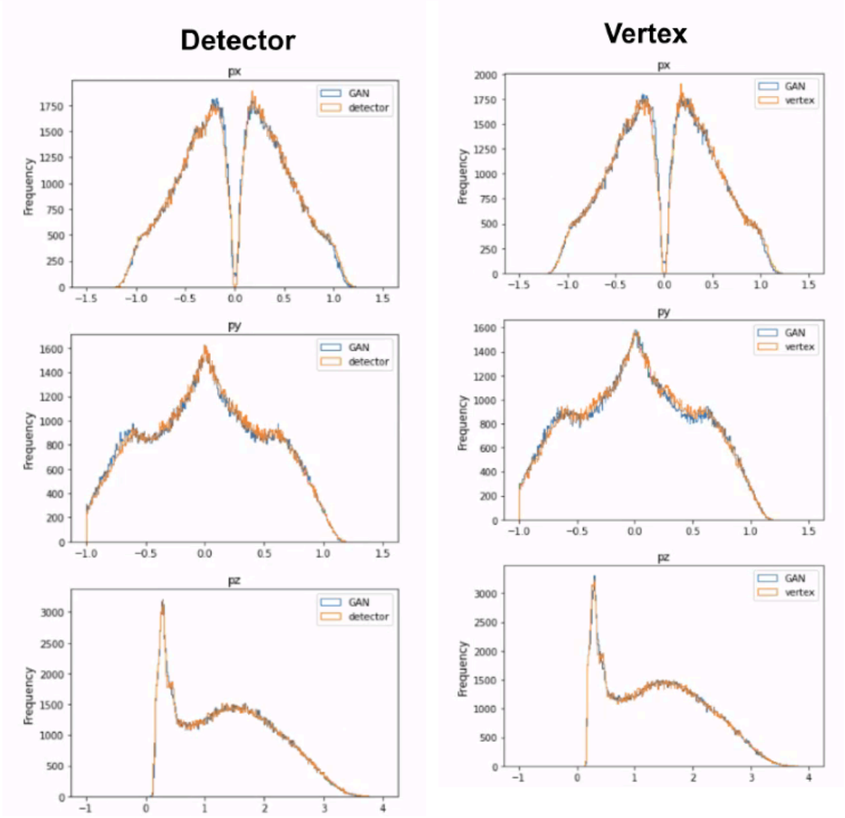
Credit: Y.Alanazi Awadh, T.Alghamdi, Y.Li

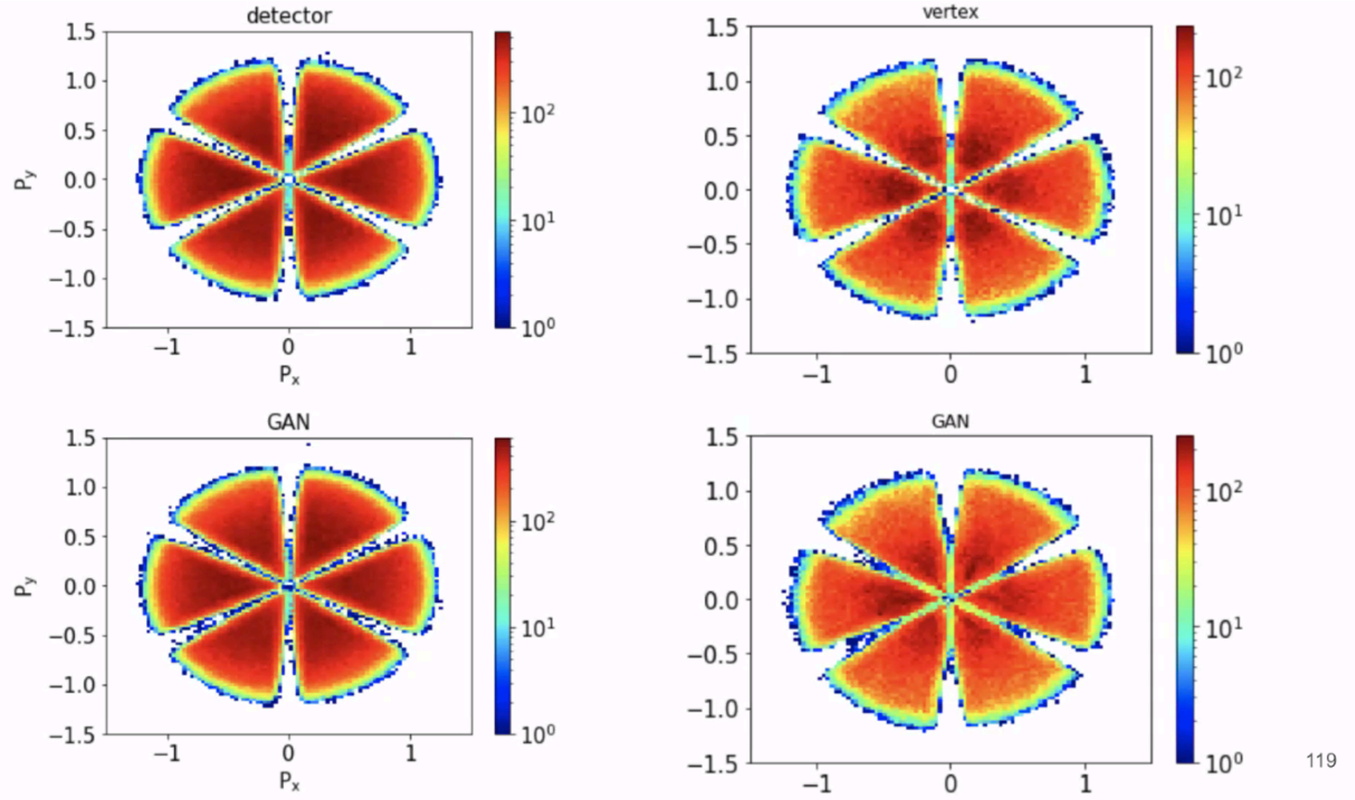
Closure test (using simulations)



Conclusions

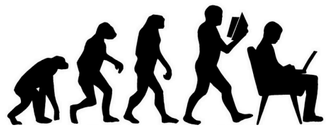
- The 'closure test' is still in progress but preliminary results show:
  - I) the detector effects are correctly reproduced by the unfolding GAN (trained with phase-space MC)
  - II) Realistic MC pseudo-data are reproduced at detector- and vertex level
- Needs to quantify the agreement
- Use trained Unfolding GAN on data to extract information at vertex level





Credit: Y.Alanazi Awadh, T.Alghamdi, Y.Li

ML to access multi-d cross section



A(i)DAPT

AI for DAta PreservaTion

Goals

- Cross section: embed multi-d cross section information (correlation) in a data-trained event generator
- Preserve data in an alternative compact and efficient form (to be applied to current JLab physics program)
- Statistics: use the NN to determine the necessary statistics for a given analysis
- Statistics: overcome statistics limitation exploiting ML super-resolution
- Detector Efficiency: folding/unfolding detector effects to extract physics at vertex level (via sim or data)
- Physics analysis: incorporate Universality (of scattering amplitudes) training the NN with different kinematics of the same final state or different final states (coupled channels)
- Physics analysis: extract from the NN features related to the underlying physics
- Physics analysis: structure the NN to reflect amplitudes properties (poles, cuts, dynamics, …)

- **Collaborative effort**
  - **ML**
  - **Data manipulation**
  - **Validation**
  - **Unfolding**
  - **Theory**
- **Regular weekly meeting**

- develop a procedure to best fit data (ML Group)
- develop a procedure to compare synt-data to data (Validation Group)
- develop a procedure to quantify the error associated to sent-data (Data manipulation Group)
- develop a procedure to take into account the detector effect (Folding/Unfolding Group)
- extract physics form data and sent-data and compare (Theory Group)

- **Wiki page:** https://clasweb.jlab.org/wiki/index.php/A%28I%29DAPT_-_AI_for_Data_Analysis_and_PresevaTion