Building Monte Carlo Event Generators using Generative Adversarial Networks

Yaohang Li

Department of Computer Science
Old Dominion University
Norfolk, Virginia 23529
yaohang@cs.odu.edu

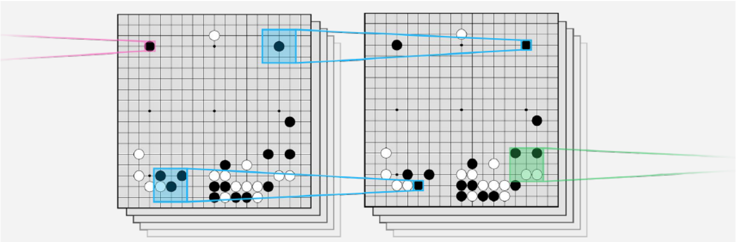
5/26/2022 @ ODU Physics Department

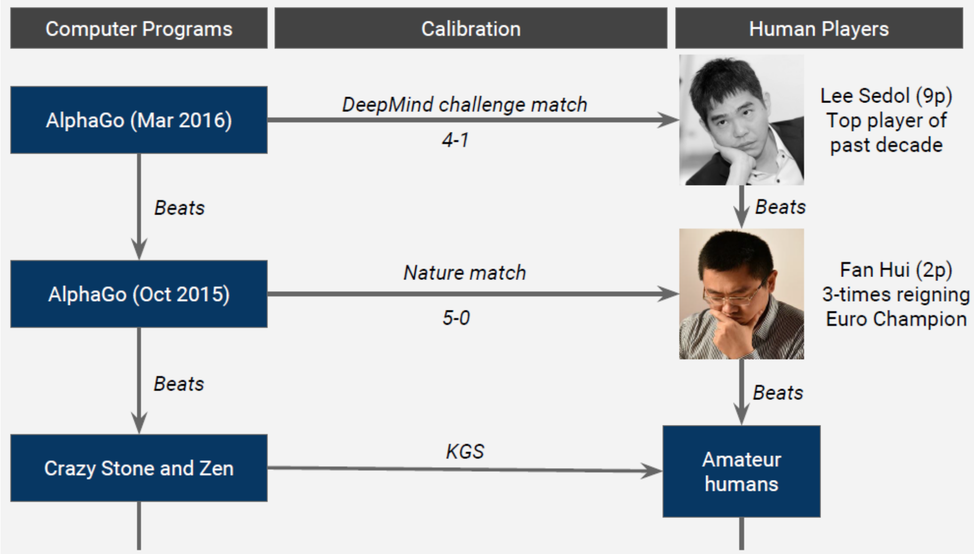
Agenda

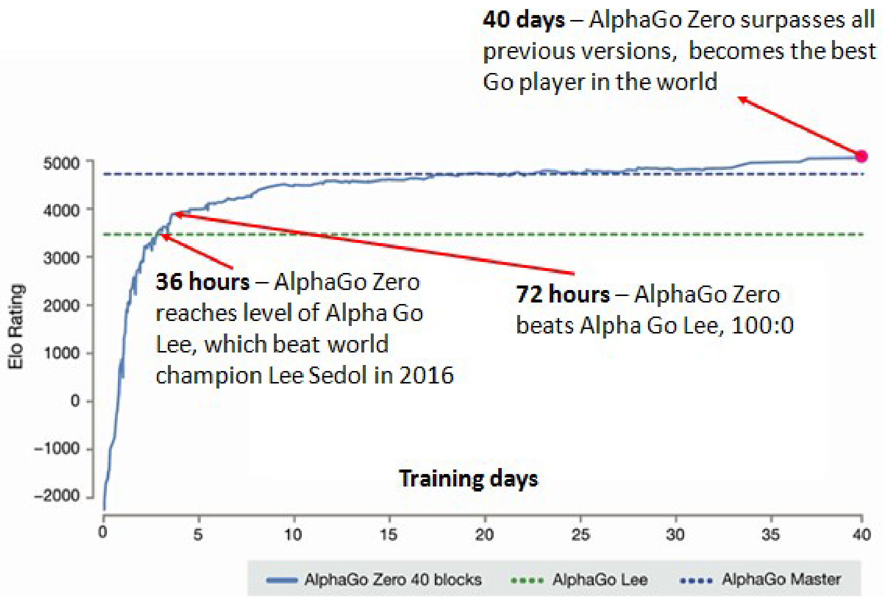
- **Introduction to Adversarial Learning and GAN**
- **Why can GAN work?**
- **Training a GAN-based Monte Carlo Event Generator**
  - Challenges
  - Electron-Proton Scattering
  - Fitting HERA Data
  - Conditional GAN
  - Pion photoproduction on the proton
- **Open Questions**

Adversarial Learning

- **AlphaGo**
  - Convolutional Neural Network (CNN)
- **AlphaGo Zero**
  - Adversarial Network



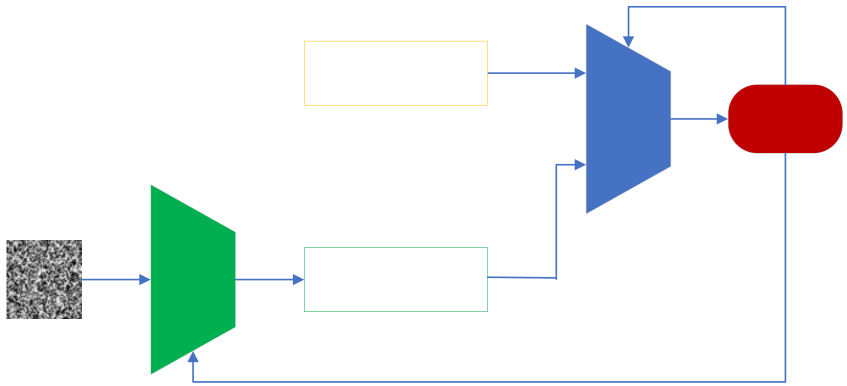




Source: DeepMind

Generative Adversarial Network (GAN)

- **Generative Adversarial Networks**
  - Introduced by Ian Goodfellow et al. in 2014
  - Deep neural network architectures comprised of two nets
    - A Generator
    - A Discriminator

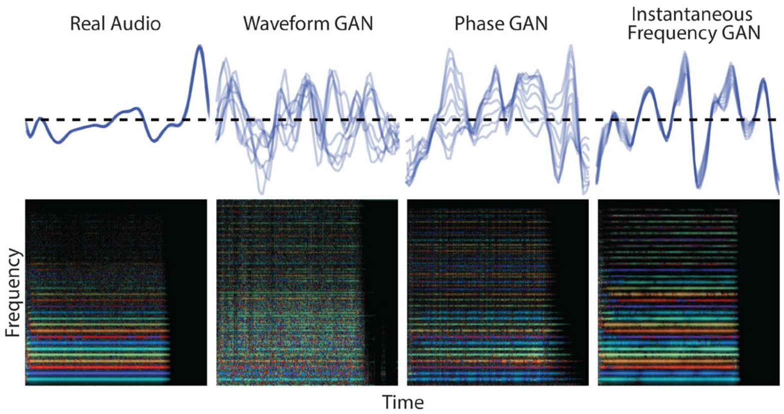


  - Both nets are trying to optimize a different and opposing loss function in a zero-sum game
- **Potential of GAN**
  - Can be trained to mimic any distribution of data
  - Create worlds eerily similar to our own in any domain

"The most interesting idea in the last 10 years in machine learning" – Yann LeCun

The Power of GAN

- **Can be trained to mimic any distribution of data**
- **Applications**
  - Artificial Arts
  - Virtual Reality
  - New Characters
  - Artificial Music



Source: gansynth

Agenda

- Introduction to Adversarial Learning and GAN
- **Why can GAN work?**
- Training a GAN-based Monte Carlo Event Generator
  - Challenges
  - Electron-Proton Scattering
  - Fitting HERA Data
  - Conditional GAN
  - Pion photoproduction on the proton
- Open Questions

Fundamentals of GAN

- **Generator G**
  - A Function: Input z, Output x
  - Given a prior distribution $P_{prior}(z)$, a probability distribution $P_G(x)$ is defined by function G
- **Discriminator D**
  - A Function: Input x, Output a scalar
  - Evaluate the difference between $P_G(x)$ and $P_{data}(x)$

Kullback–Leibler Divergence

- **Kullback–Leibler divergence (Relative Entropy)**
  - measures how one probability distribution is different from a reference probability distribution
  - Given probability distributions *P* and *Q*
    - Discrete version

$$D_{\mathrm{KL}}(P||Q) = -\sum_{x} P(x)\log\left(\frac{Q(x)}{P(x)}\right)$$

    - Continuous version

$$D_{\mathrm{KL}}(P||Q) = -\int P(x)\log\left(\frac{Q(x)}{P(x)}\right)dx$$

Properties of Kullback–Leibler Divergence

- **Explanation of KL divergence**

Cross Entropy of *P* and *Q*

Entropy of *P*

$$D_{\mathrm{KL}}(P||Q) = -\sum_x P(x)\log\left(\frac{Q(x)}{P(x)}\right)$$

$$= -\sum_x P(x)\log Q(x) - \left(-\sum_x P(x)\log P(x)\right)$$

- **Properties of KL divergence**
  - Non-symmetric
  - Non-negative

Jensen-Shannon Divergence

- **Jensen-Shannon Divergence**
  - Measures the similarity between two probability distributions
  - A symmetrized and smoothed version of the Kullback–Leibler divergence
  - Definition

$$JSD(P||Q) = \frac{1}{2} D_{\mathrm{KL}}(P||M) + \frac{1}{2} D_{\mathrm{KL}}(M||Q)$$

where

$$M = \frac{1}{2}(P + Q)$$

  - Bounds

$$0 \leq JSD(P||Q) \leq \log(2)$$

GAN Cost Function

- **An optimization problem**
  - Find an optimal generator G^* such that

$$G^{*}=\text{arg min}_{G}\text{max}_{D}V(G,D)$$

  - A MiniMax algorithm
- **Cost Function of Binary Classifier**
  - $V = E_{x\sim P\_data} [\log D(x)] + E_{x\sim P\_G}[\log(1-D(x))]$
    - Minimizing Cross-Entropy
      - x is real, minimize $-\log D(x)$
      - x is fake, minimize $-\log(1-D(x))$

$\max_D V(G, D)$

- $\max_D V(G, D)$
  - Given a generator G
  - *$max_D V(G,D)$* evaluates the “difference” between P_G and P_{data}
- **What is the optimal D^* that maximize $V(G, D)$?**

$$V = E_{x \sim P_{data}}[\log D(x)] + E_{x \sim P_G}[\log(1 - D(x))]$$
$$= \sum_x P_{data}(x) \log D(x) + \sum_x P_G(x) \log(1 - D(x))$$

Then

$$D^* = P_{data}(x) / (P_{data}(x) + P_G(x))$$

$\min_G \max_D V(G, D)$

$\max_D V(G, D)$

$= V(G, D^*)$ **where** $D^* = P_{data}(x)/(P_{data}(x) + P_G(x))$

$= E_{x \sim P_{data}}[\log D^*(x)] + E_{x \sim P_G}[\log(1 - D^*(x))]$

$= \sum_x P_{data}(x) \log D^*(x) + \sum_x P_G(x) \log(1 - D^*(x))$

$= -2\log 2 + 2JSD(P_{data} || P_G)$

What is G^* **with** $\min_G \max_D V(G, D)$**?**

$$JSD(P_{data} || P_G) = 0$$

i.e., $P_{data} = P_G$

Agenda

- Introduction to Adversarial Learning and GAN
- Why can GAN work?
- **Training a GAN-based Monte Carlo Event Generator**
  - Challenges
  - Electron-Proton Scattering
  - Fitting HERA Data
  - Conditional GAN
  - Pion photoproduction on the proton
- Open Questions

Challenges in GAN Training

- **Training a GAN is notoriously difficult**
  - Perfect Discriminator
  - Mode Collapse
  - Non-convergence
  - Imbalance Generator and Discriminator Training
  - Model parameter oscillation
  - Destabilization
  - Vanishing gradient

Additional challenges in training an event generation GAN

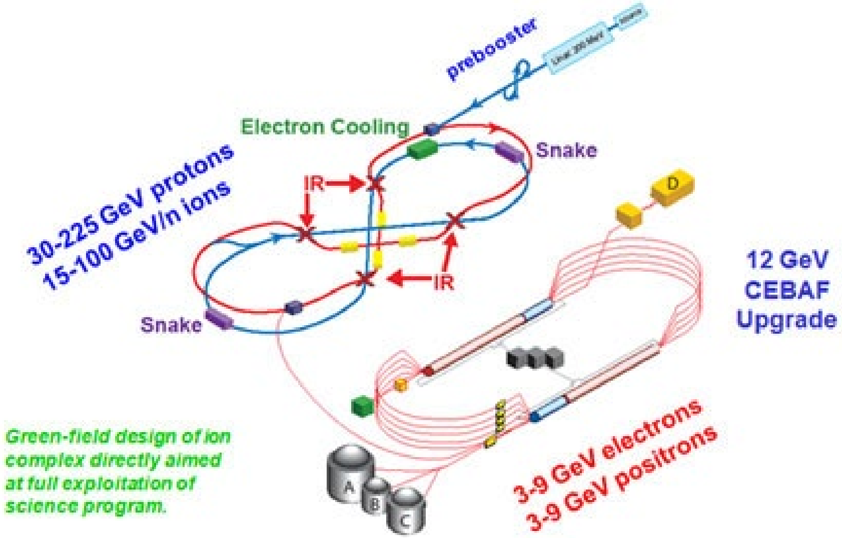
- **Precise Event Feature Distributions**
  - Replicate the nature of particle reactions faithfully
- **Obeying the Fundamental Physics Laws**
  - Energy Conservation
  - Momentum Conservation
- **Handling Detector Effects**
  - Smearing
  - Acceptance
  - Detector Inefficiency

Agenda

- Introduction to Adversarial Learning and GAN
- Why can GAN work?
- **Training a GAN-based Monte Carlo Event Generator**
  - Challenges
  - Electron-Proton Scattering
  - Fitting HERA Data
  - Conditional GAN
  - Pion photoproduction on the proton
- Open Questions

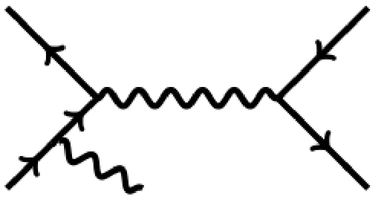
Electron-Proton Scattering

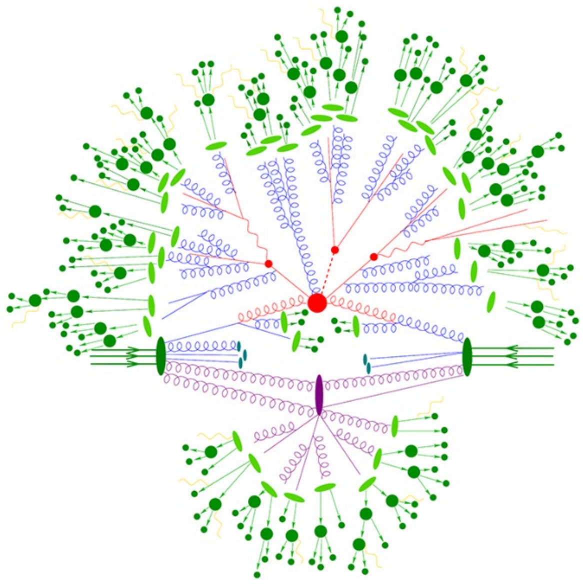
- **Pythia Events**
  - Center-of-mass energy of 100 GeV
- **Inclusive Simulation**
  - GAN is only trained on the momenta of the final state electrons



Classic Monte Carlo Event Generator (MCEG)

- **Important tools for studies of high energy scattering reactions**
  - Understanding detector effects
  - Building expectations on how experimental data should look like under different theoretical assumptions
  - Justifying the validity of the quantum field theory in the underlying models
- **Popular MCEGs**
  - Pythia
  - Herwig
  - Sherpa





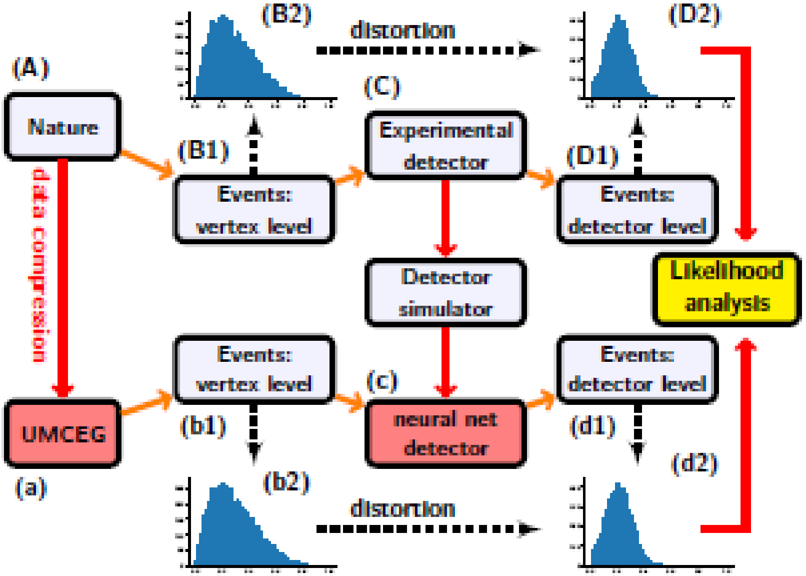
https://theory.slac.stanford.edu/our-research/simulations

Limitations of MCEG

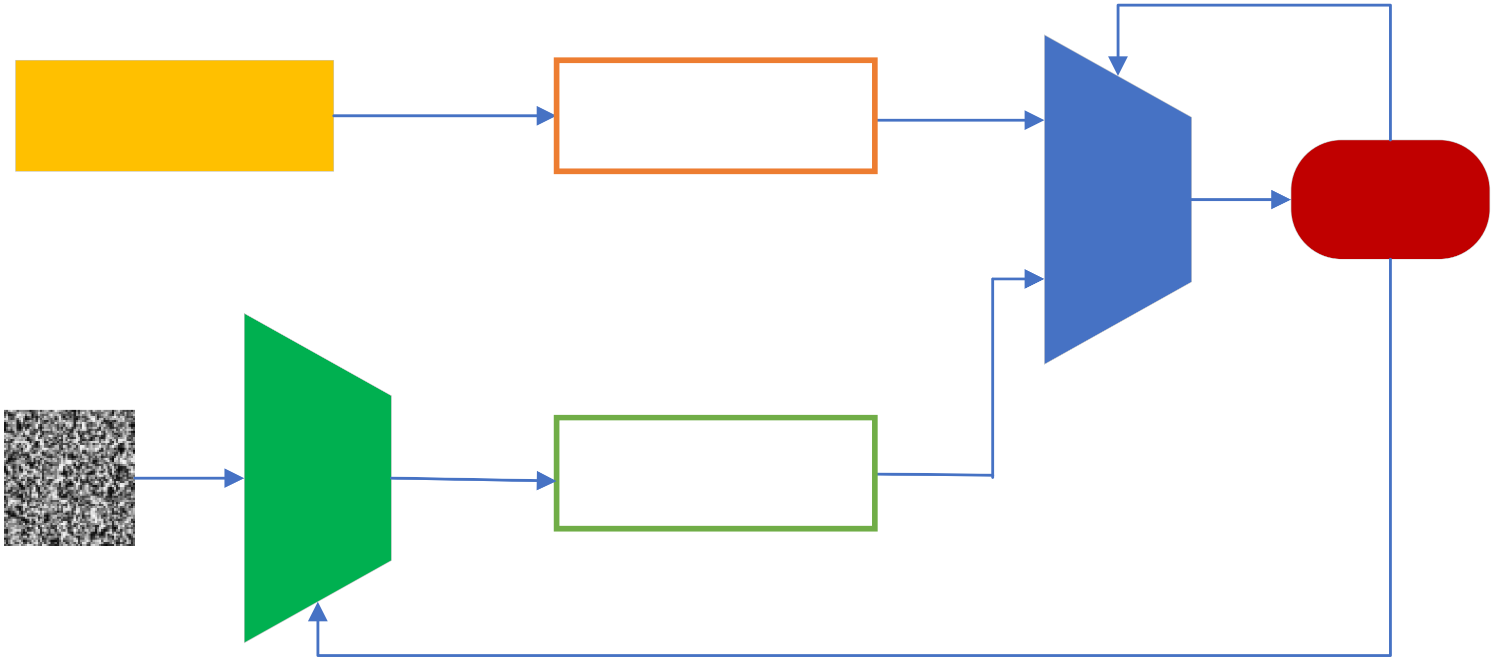
- **Assumptions of Monte Carlo event generators**
  - The underlying physics theories that govern the production of particles in a given reaction
  - Femto-scale physics
- **Computation**
  - Efficacy of QCD (quantum chromodynamics) factorization
  - Approximation
    - partonic dynamics
    - nonperturbative amplitudes
    - probability distributions
- **Limited capability to capture the full range of possible correlations between the particles' momenta and spins**

GAN-based Event Generators

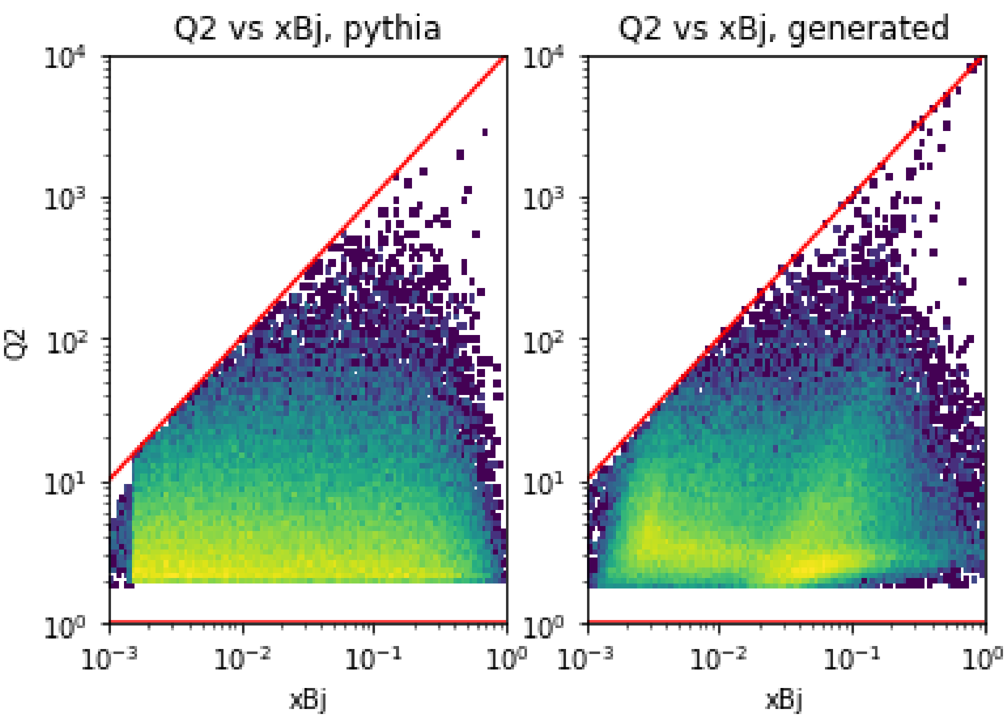
- **Learn from real electron-proton scattering data**
  - Capture rich underlying distributions over data
    - Difficult to model using explicit parameters
- **Faithfully reproducing particle reaction events**
  - No assumptions on femtometer-scale physics theory
- **Overcome the limitations of MCEGs**
- **Proof-of-concept on inclusive electrons**

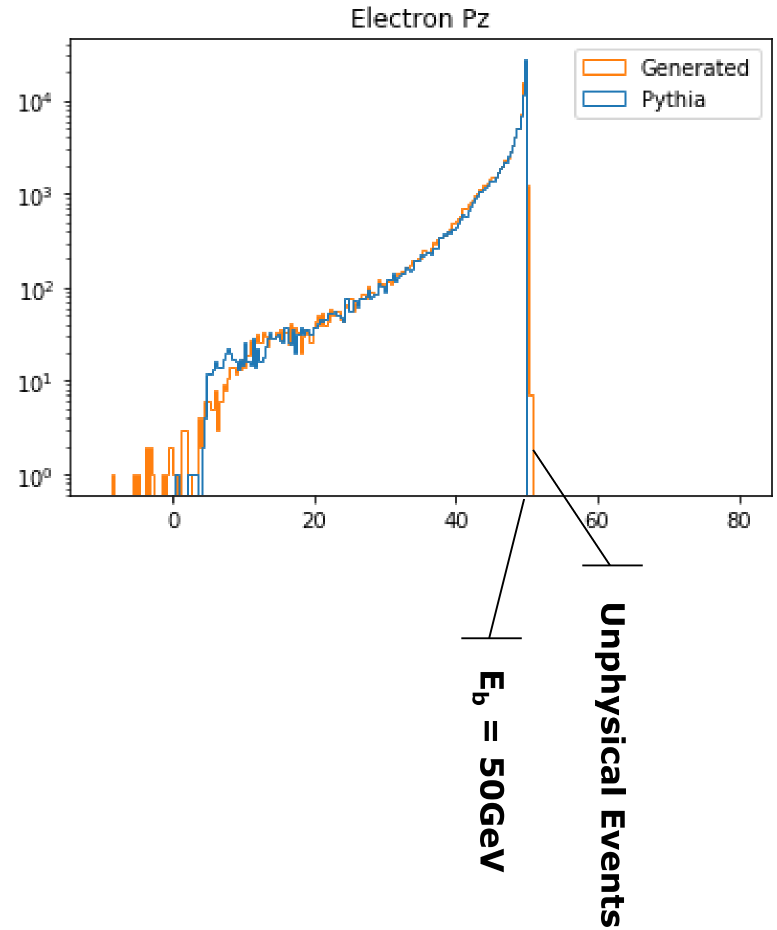


Initial Attempt: Direct Simulation GAN



Results of Director Simulation



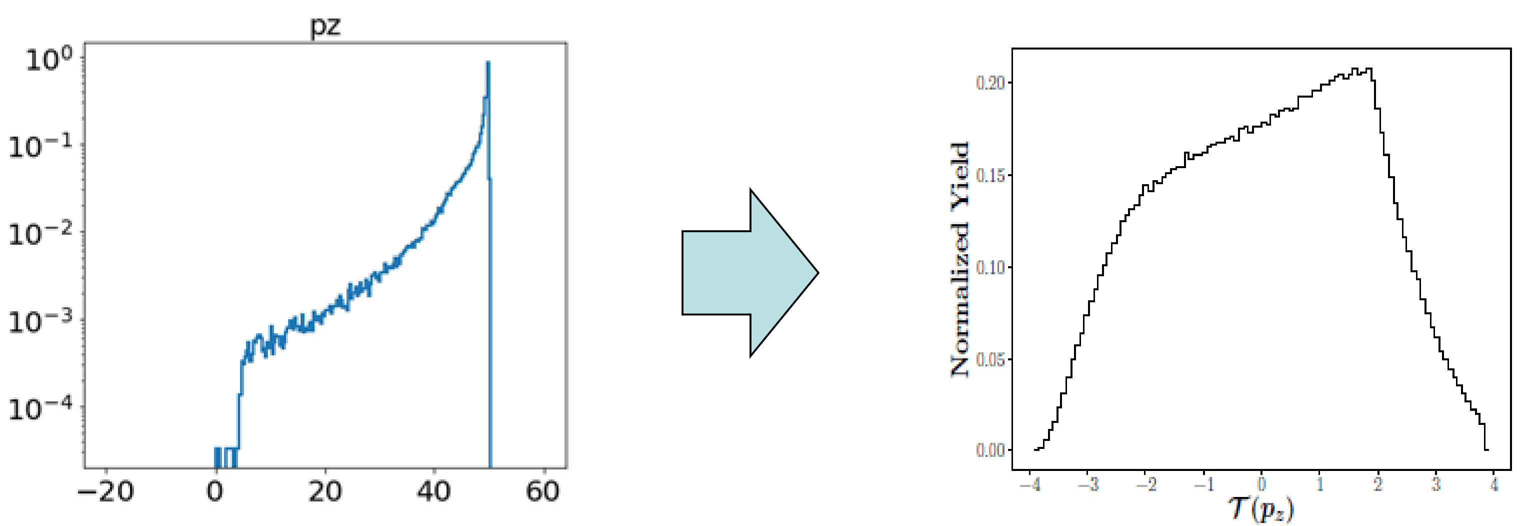


$\chi^2 = 118.29$

Features Transformation

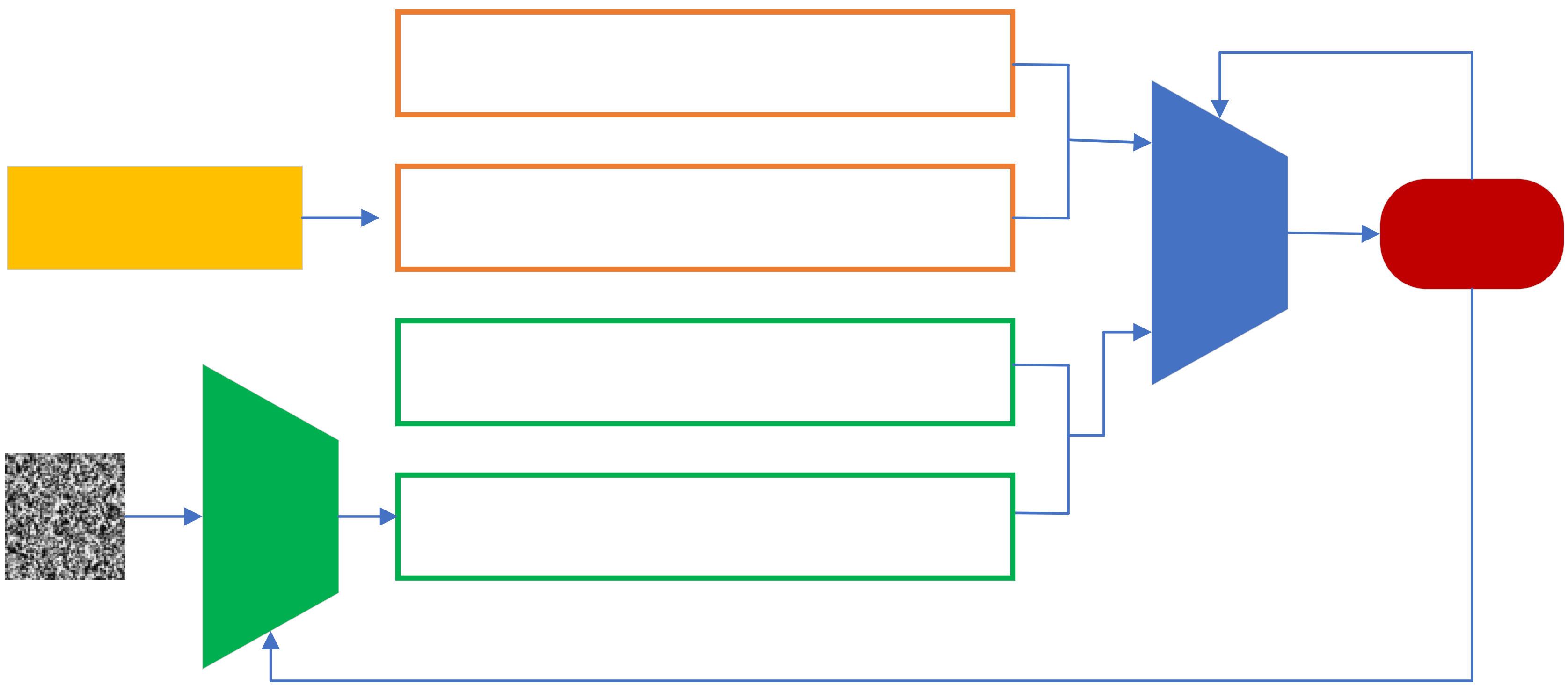
$$\mathcal{T}(p_z) = \log(E_{\mathrm{b}} - p_z)$$

- **Conversion to eliminate sharp edges**
- **Guarantee no generation of non-physical electrons**

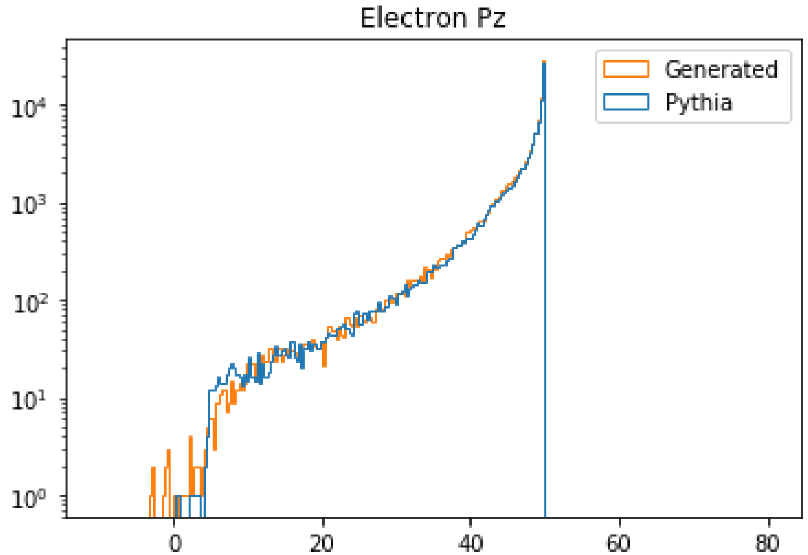


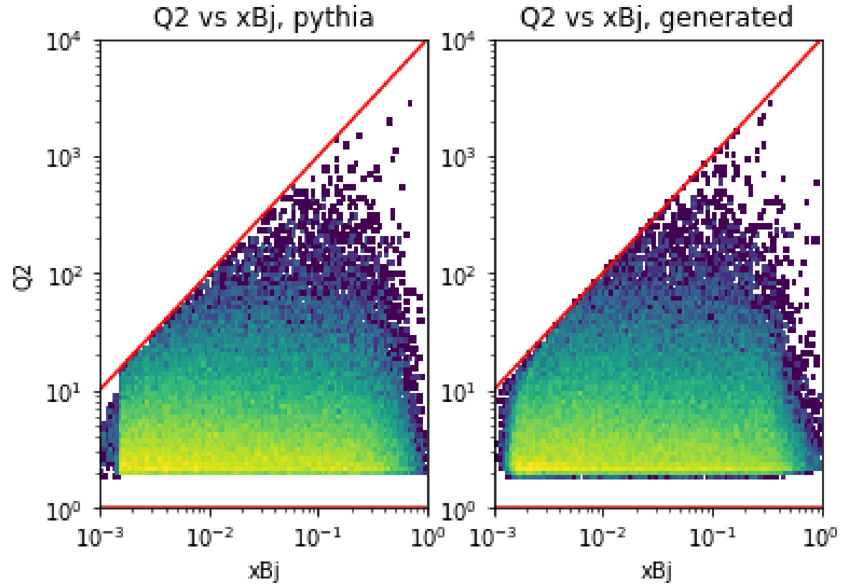
Features Augmentation and Transformation GAN (FAT-GAN)

- Features Transformation
- Features Augmentation



Results of FAT-GAN

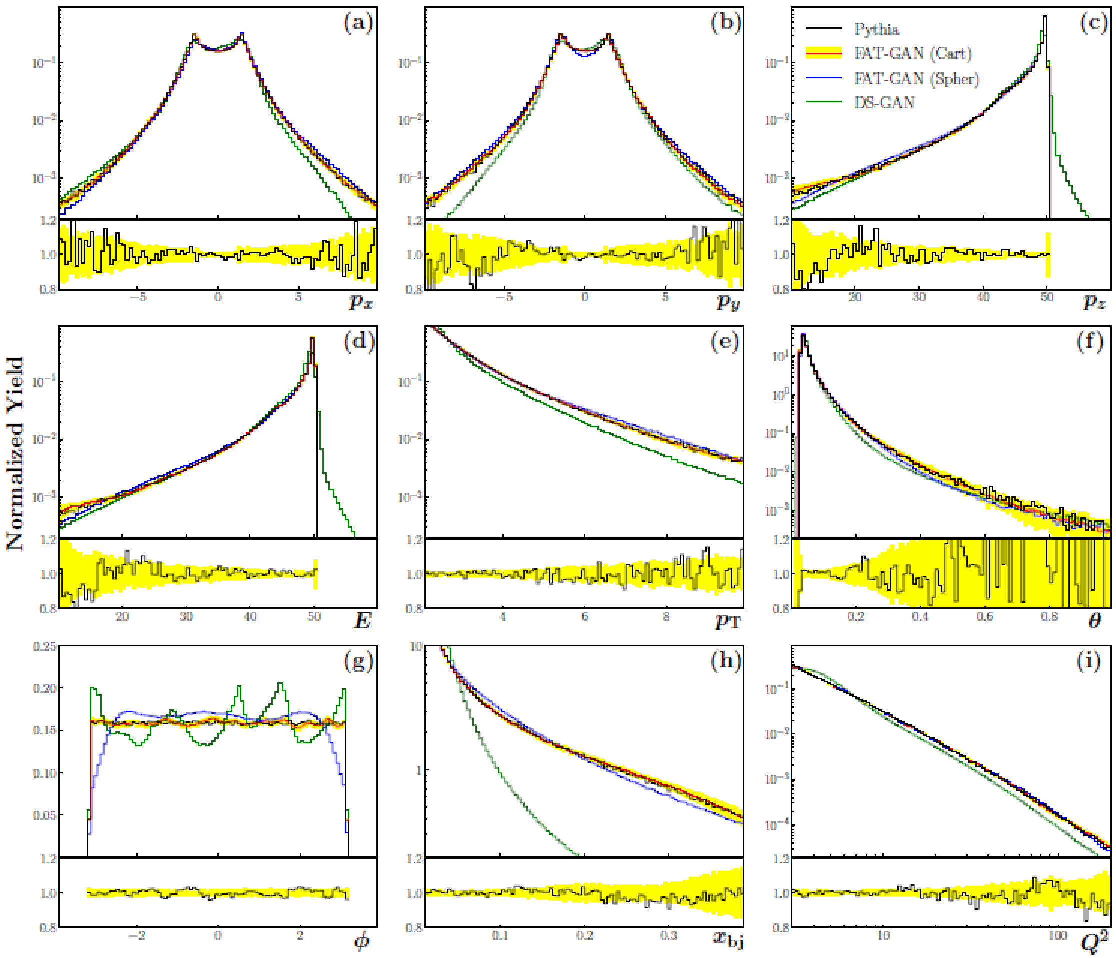




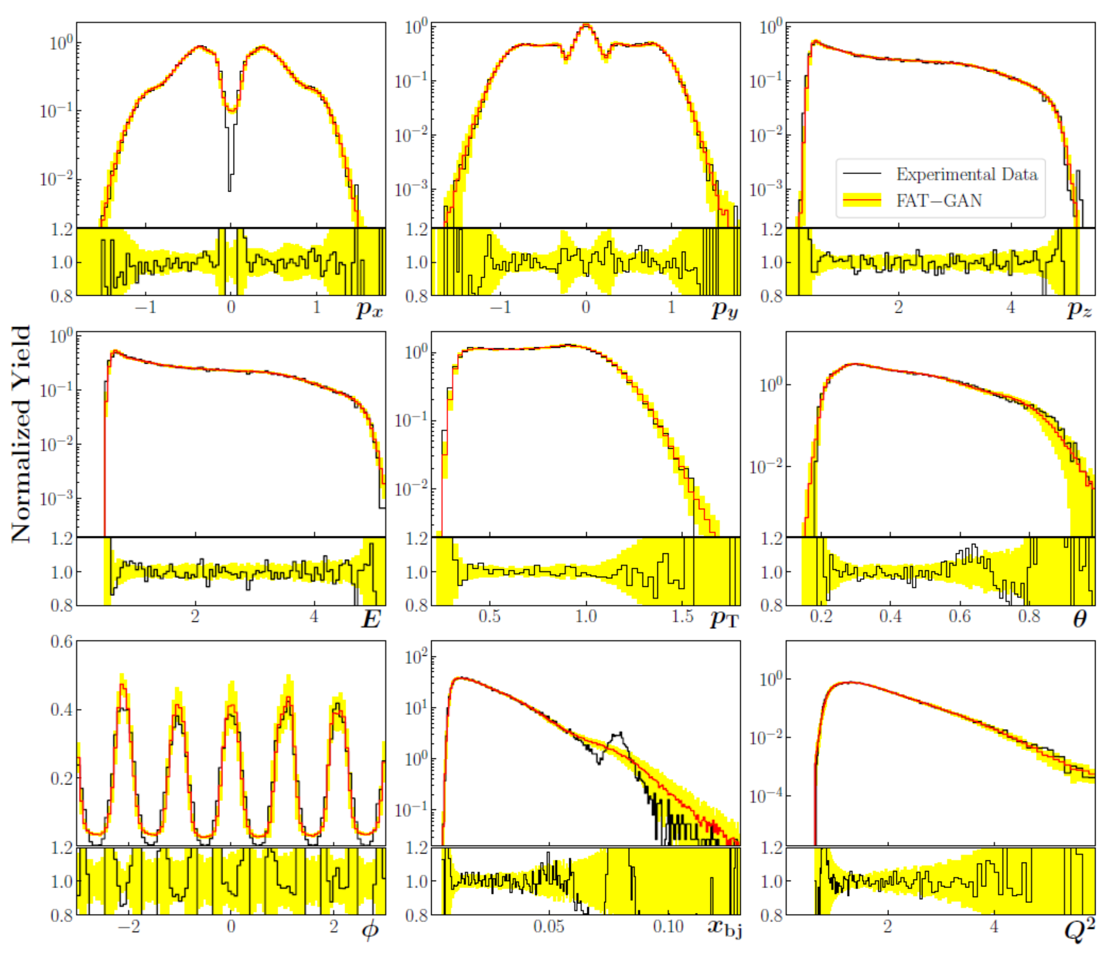
No more non-physical events with p_z > 50 GeV

Good approximation of Q^2 and x_{Bj} correlation with $\chi^2 = 1.52$

Distributions of Generated Physical Properties



FAT-GAN on experimental electron-proton scattering data

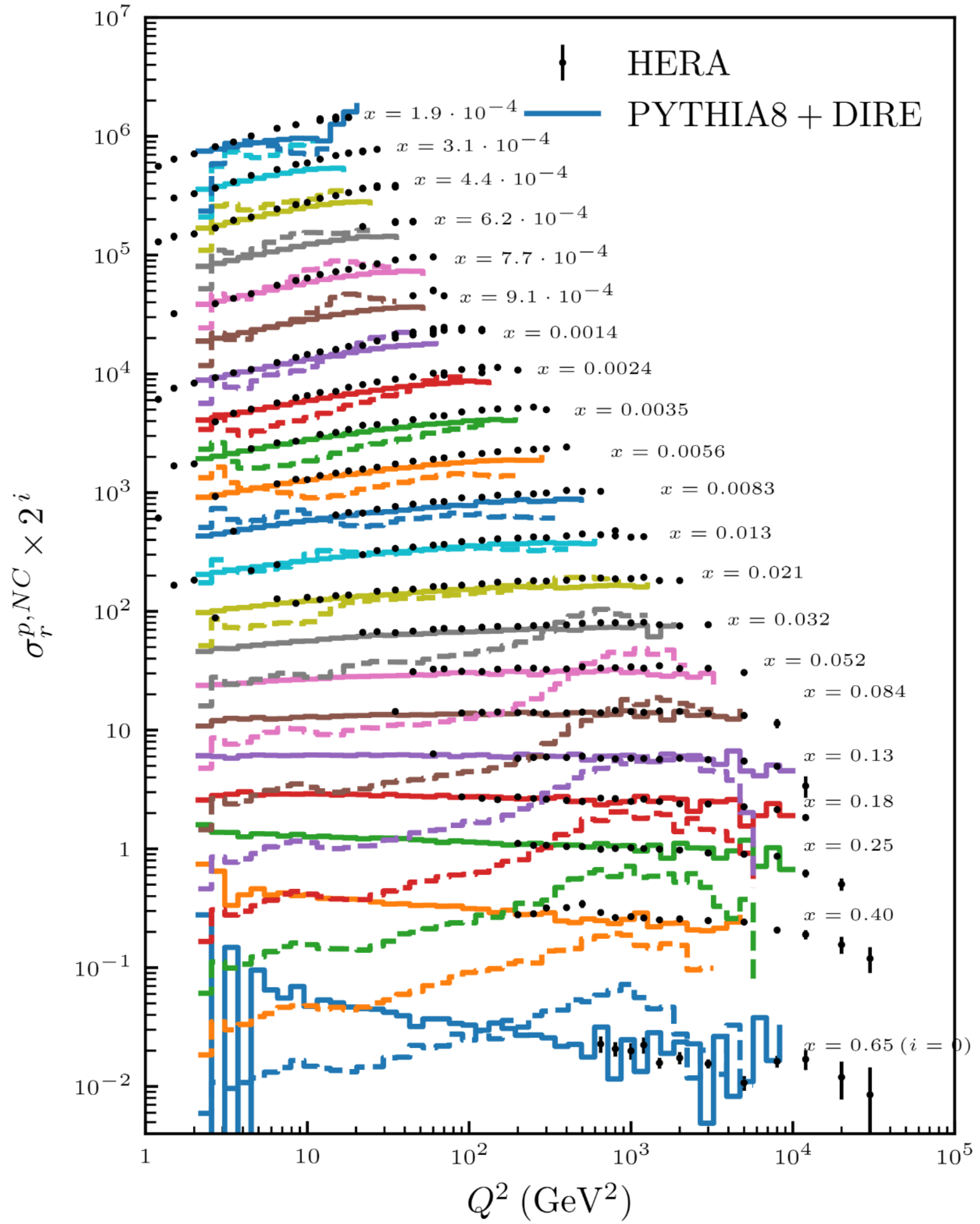


Agenda

- Introduction to Adversarial Learning and GAN
- Why can GAN work?
- **Training a GAN-based Monte Carlo Event Generator**
  - Challenges
  - Electron-Proton Scattering
  - Fitting HERA Data
  - Conditional GAN
  - Pion photoproduction on the proton
- Open Questions

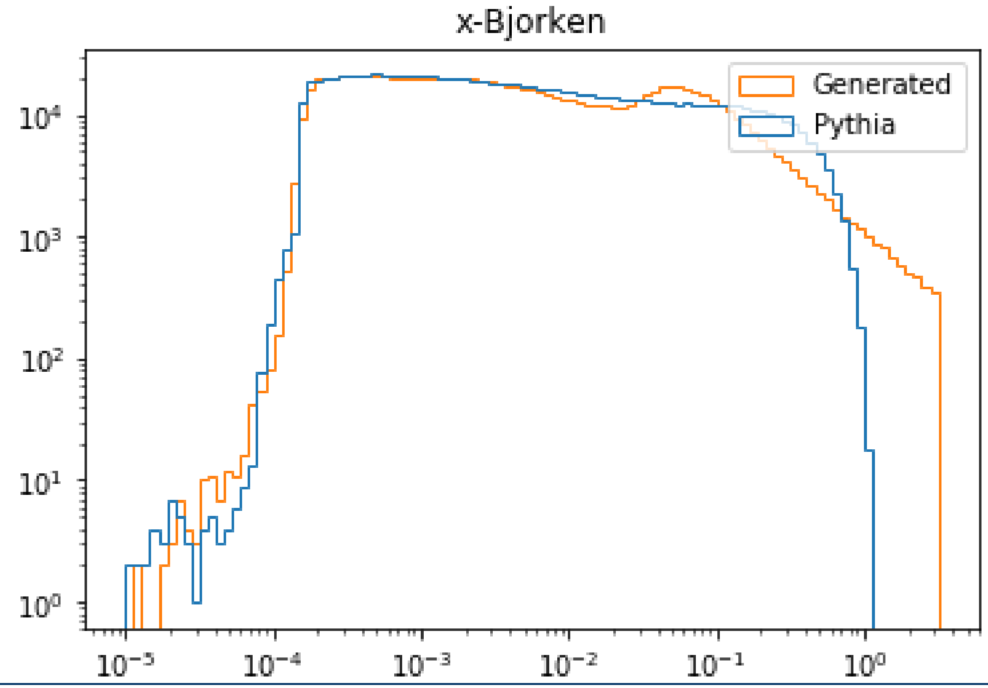
A New Problem

- **FAT-GAN has difficulty to reproduce HERA data**
  Electron Beam Energy: 27.5 GeV
  Proton Beam Energy: 920 GeV



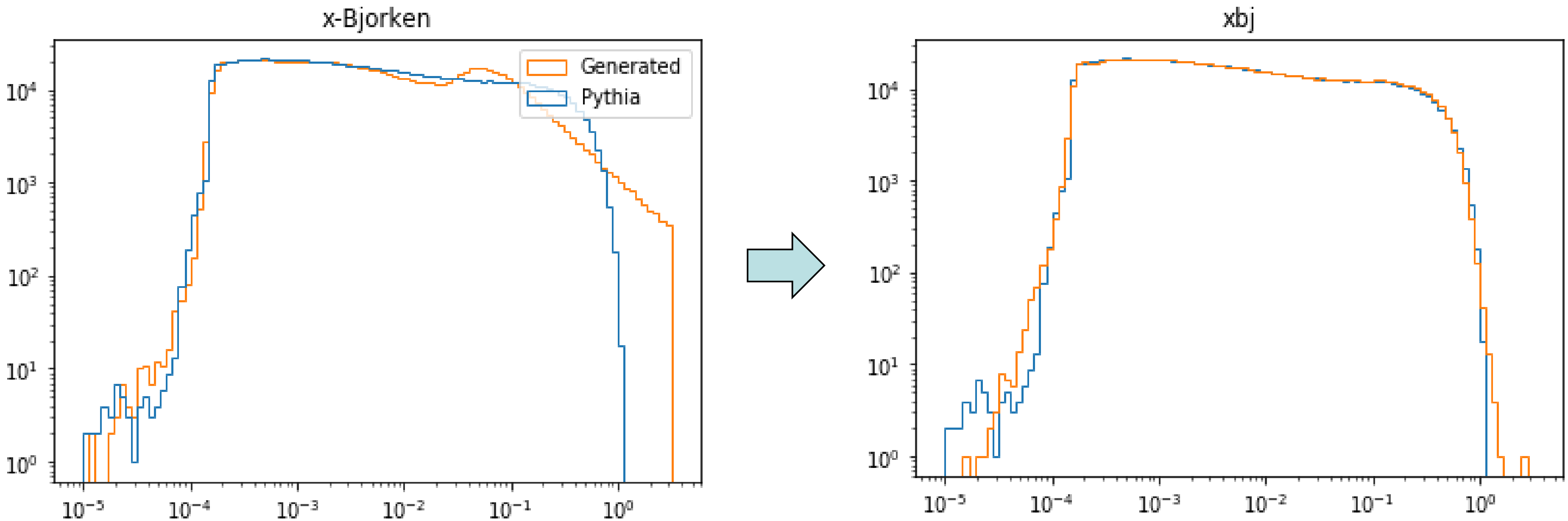
The Problem

- **$\log(E_b - p_z)$ is not enough**
  - Need to be aware of the other conditions for physical feasible events
  - For example
    - $X_{Bj} < 1.0$ (energy conservation)
    - $2E_b - E - p_z > 0$

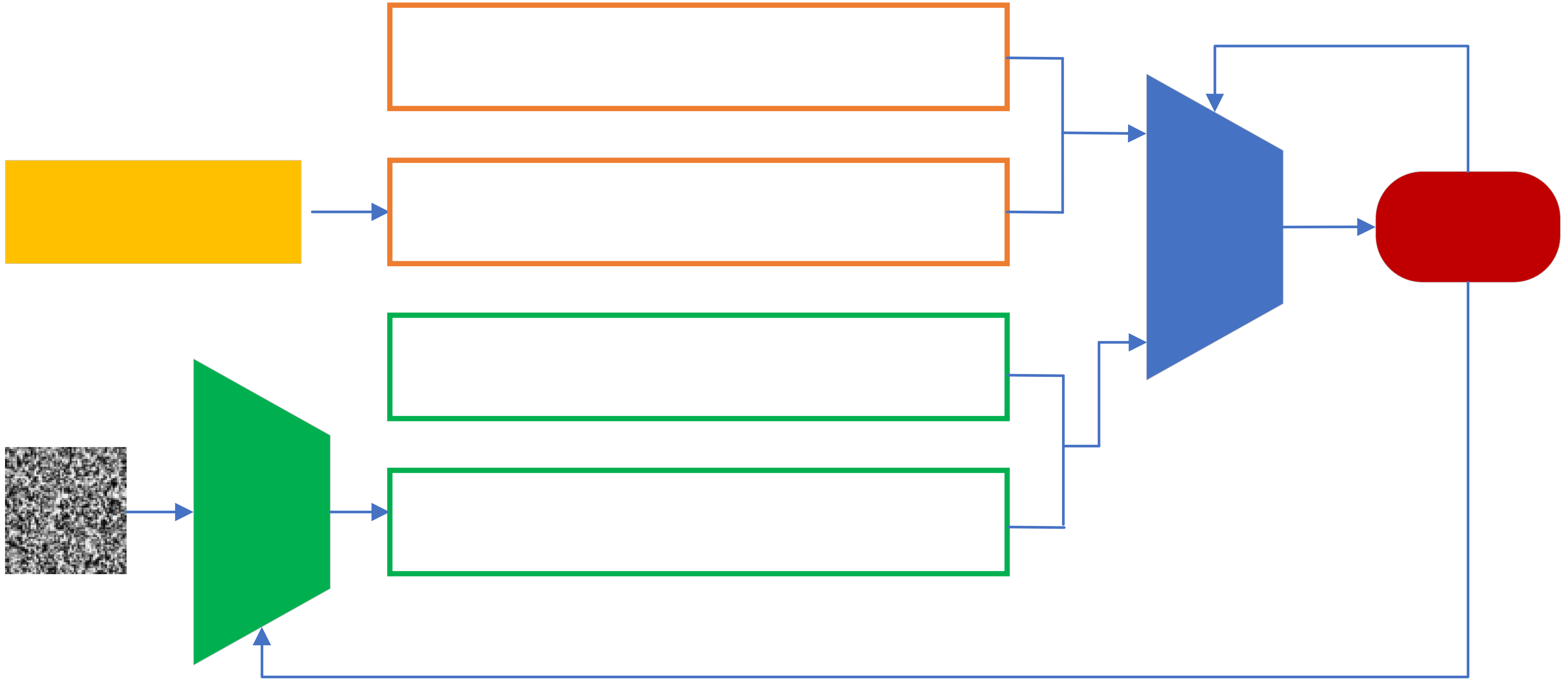


The solution

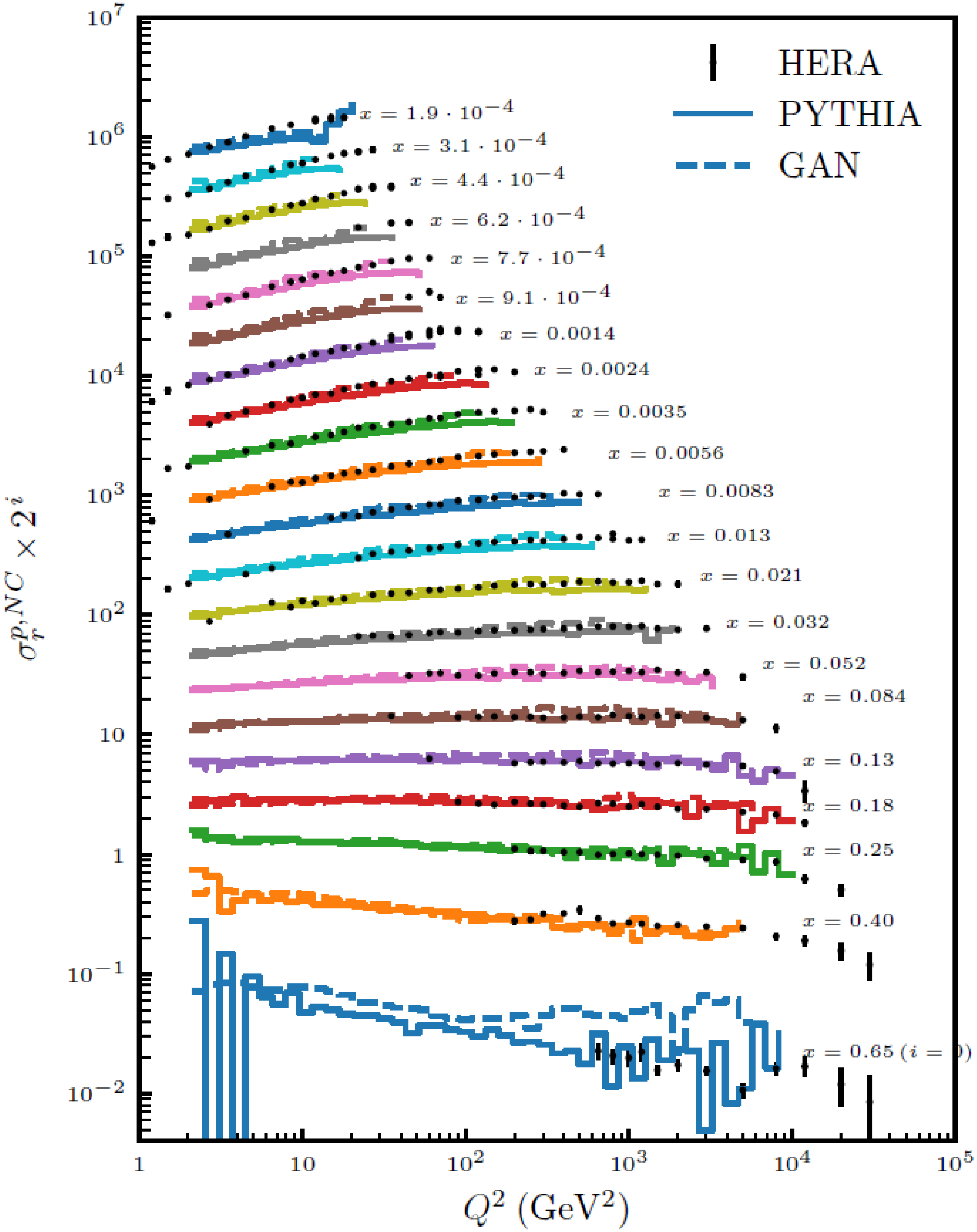
- **New Generated Features**
  - Log($E - p_z$)
  - Log($2E_b - E - p_z$)
  - Φ
- **Recalculate (E, p_x, p_y, p_z) from the generated features**



A New FAT-GAN

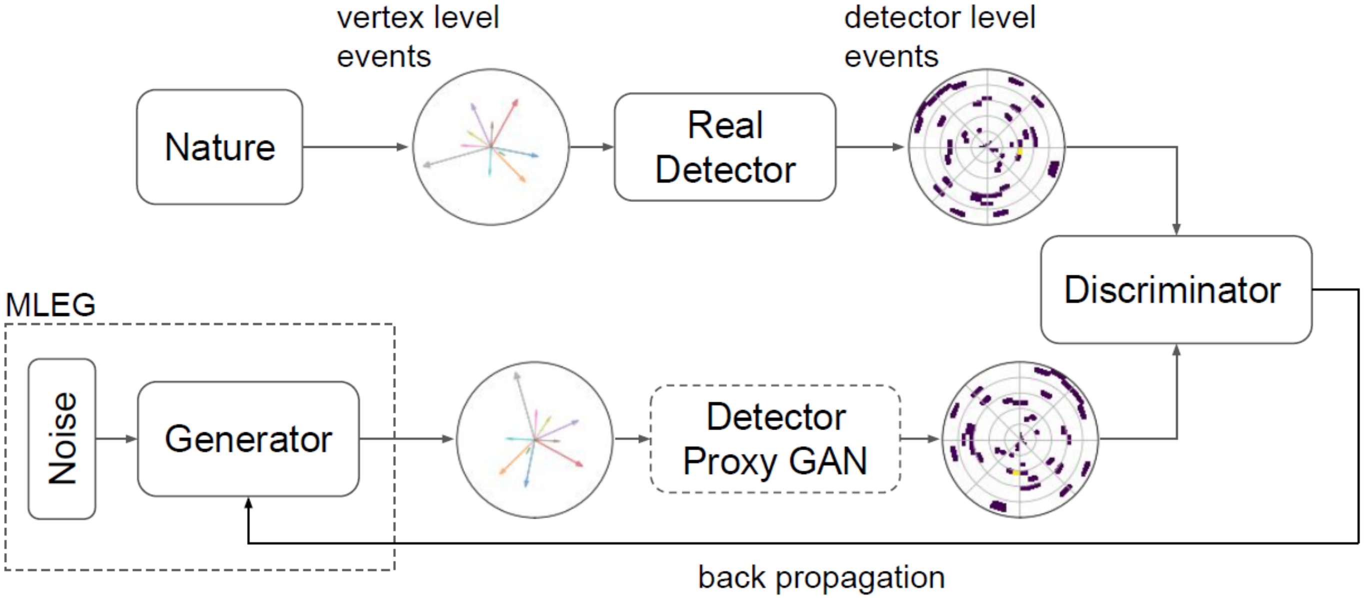


New Results for HERA



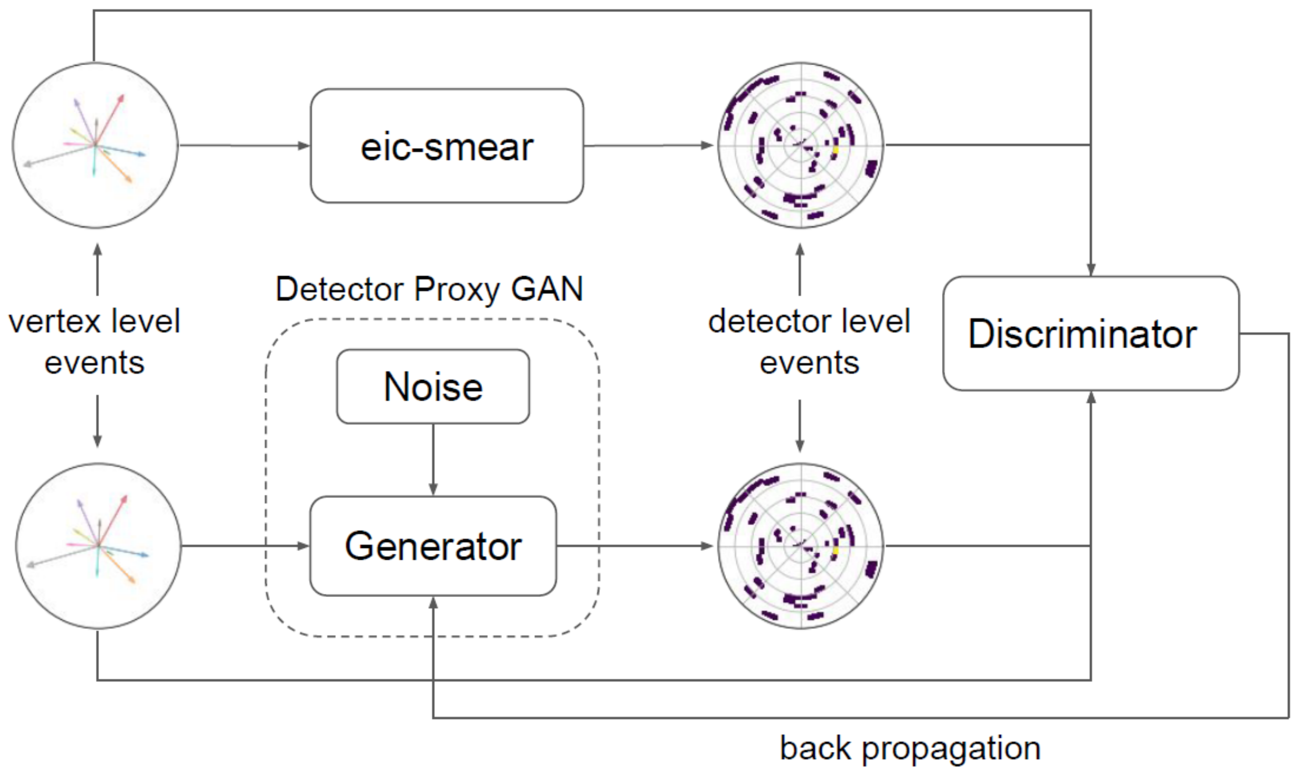
Unfolding Vertex-level Events from Detector-level Events

- **MLEG**
  - Transform noise into vertex-level simulated events
- **Detector Proxy GAN**
  - Detected simulator
  - Mimic synthetic detector-level events
- **Discriminator**
  - Differentiate detector-level events

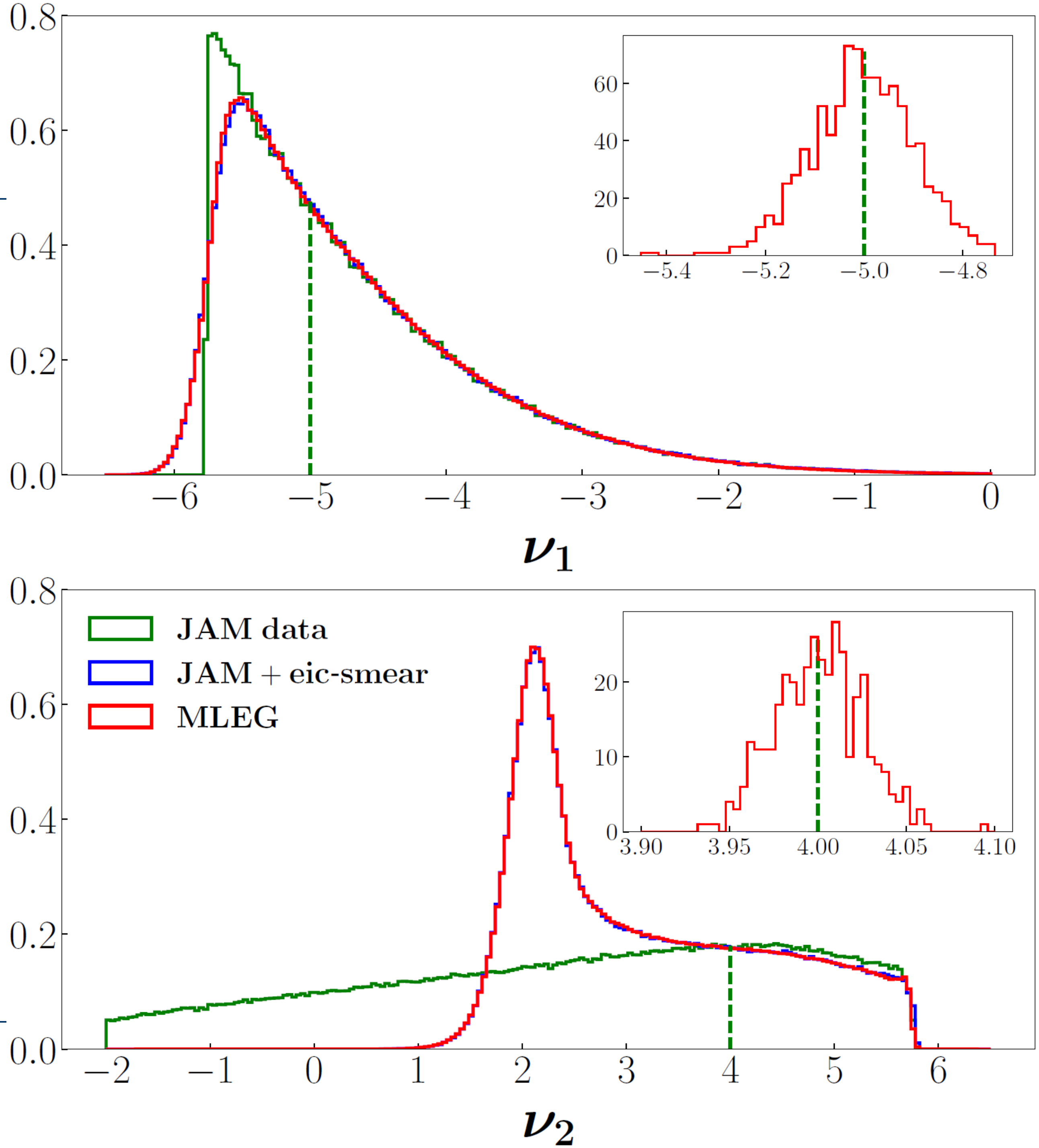


Detector Surrogate

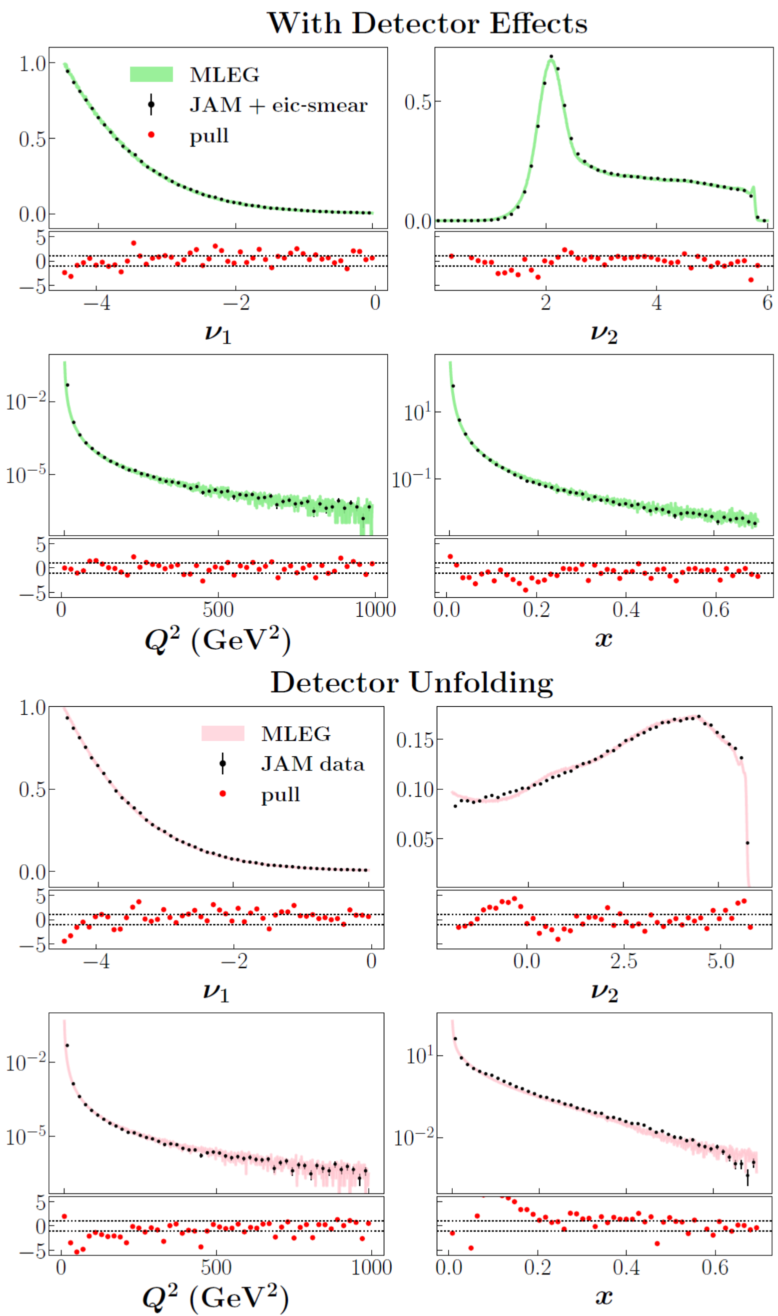
- **Detector Proxy GAN**
  - Conditional GAN
  - Training samples
    - From guess vertex-level samples and corresponding detector-level samples using a detector simulator

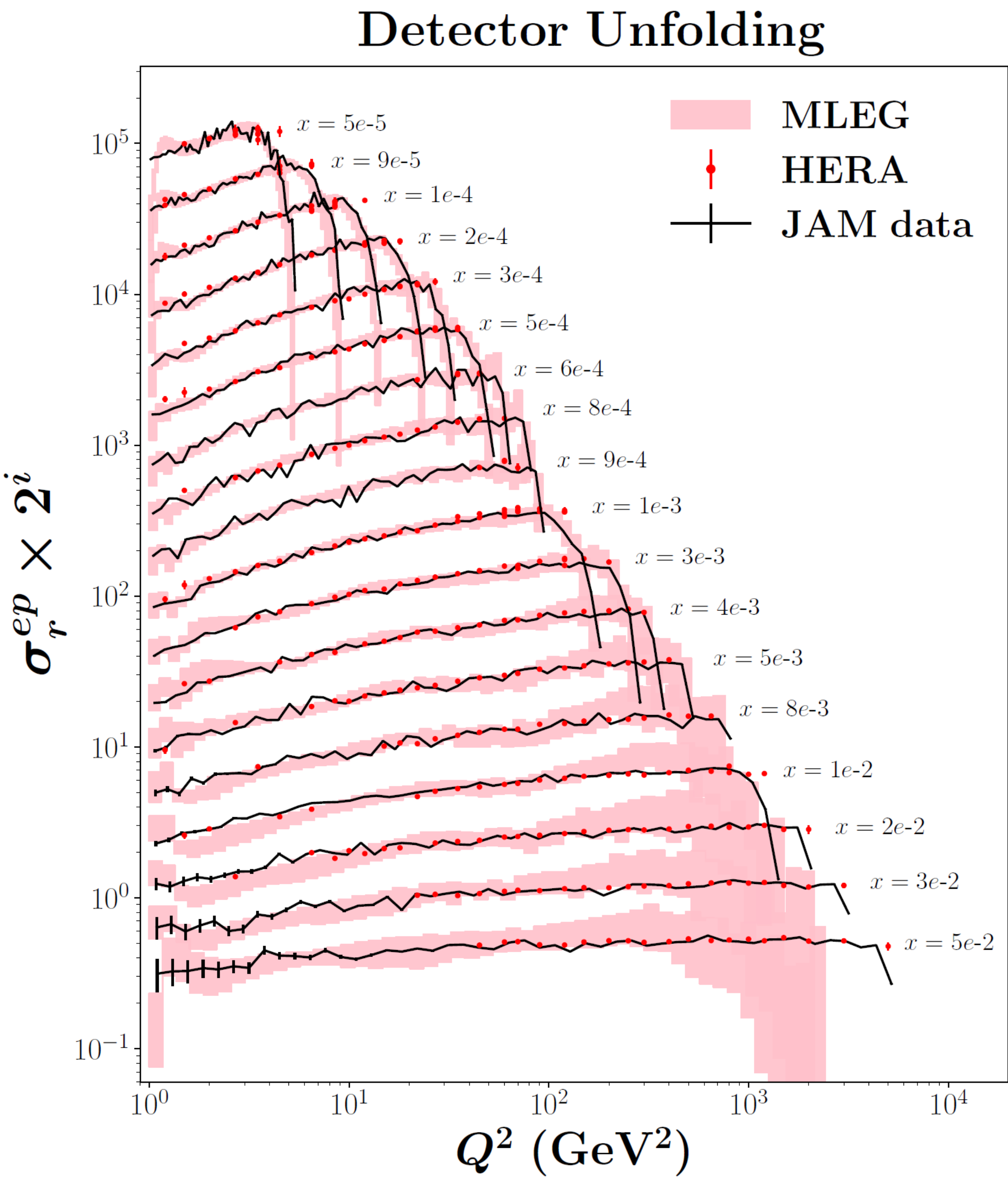


EIC Smearing



Unfolding Results



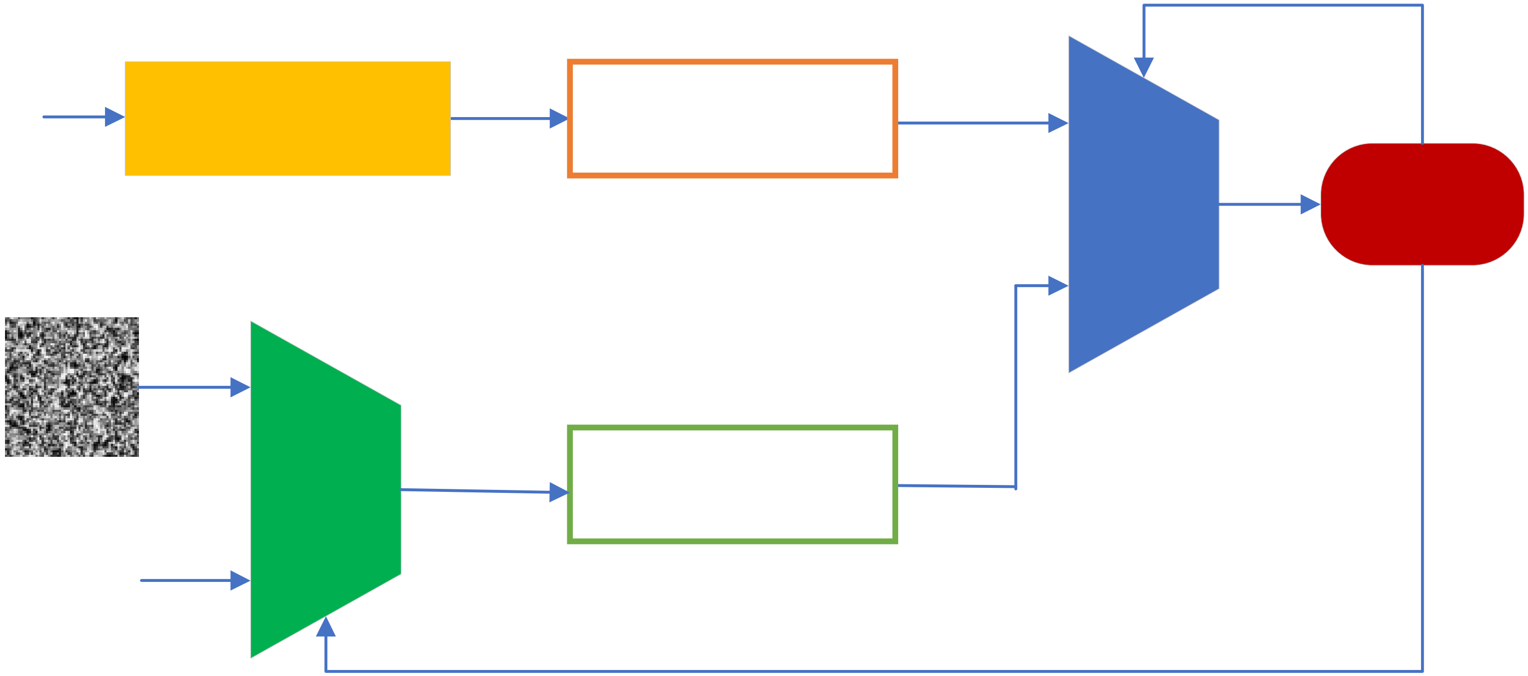


Agenda

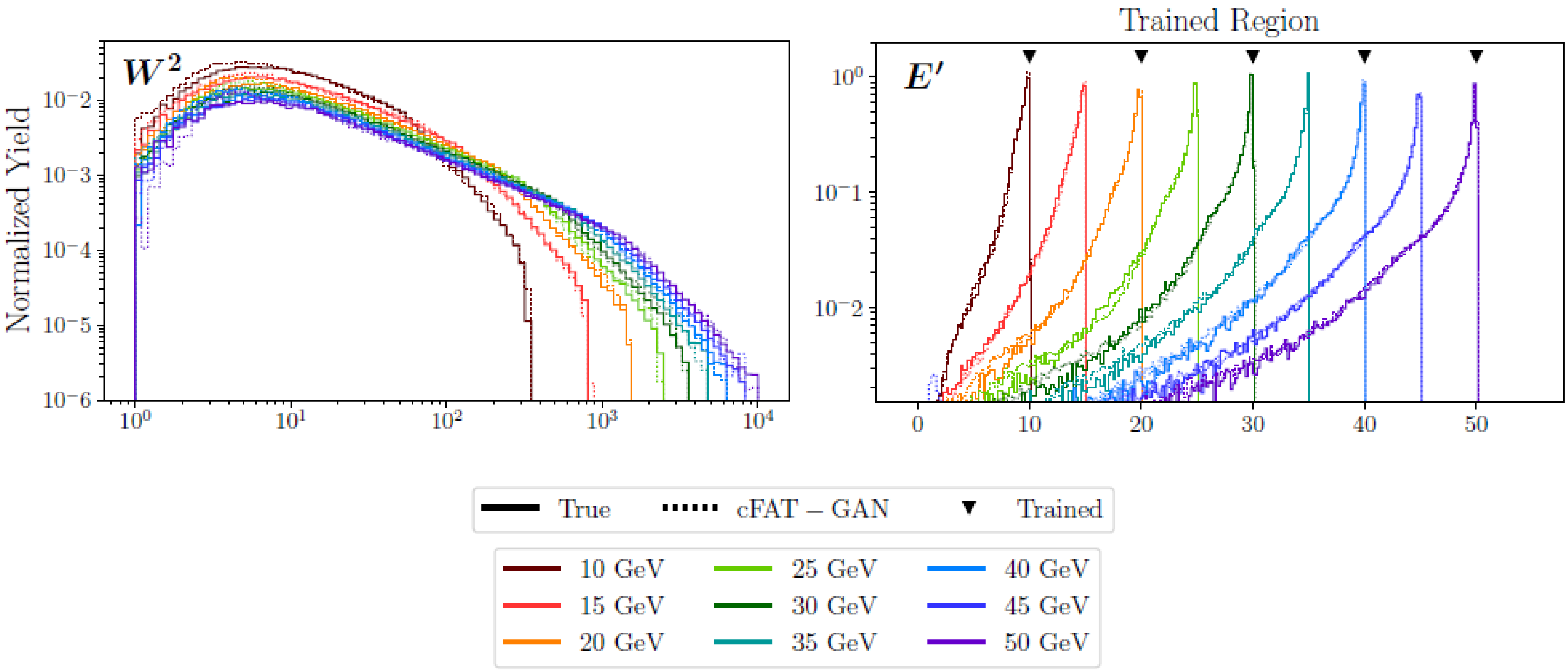
- Introduction to Adversarial Learning and GAN
- Why can GAN work?
- **Training a GAN-based Monte Carlo Event Generator**
  - Challenges
  - Electron-Proton Scattering
  - Fitting HERA Data
  - Conditional GAN
  - Pion photoproduction on the proton
- Open Questions

Conditional GAN

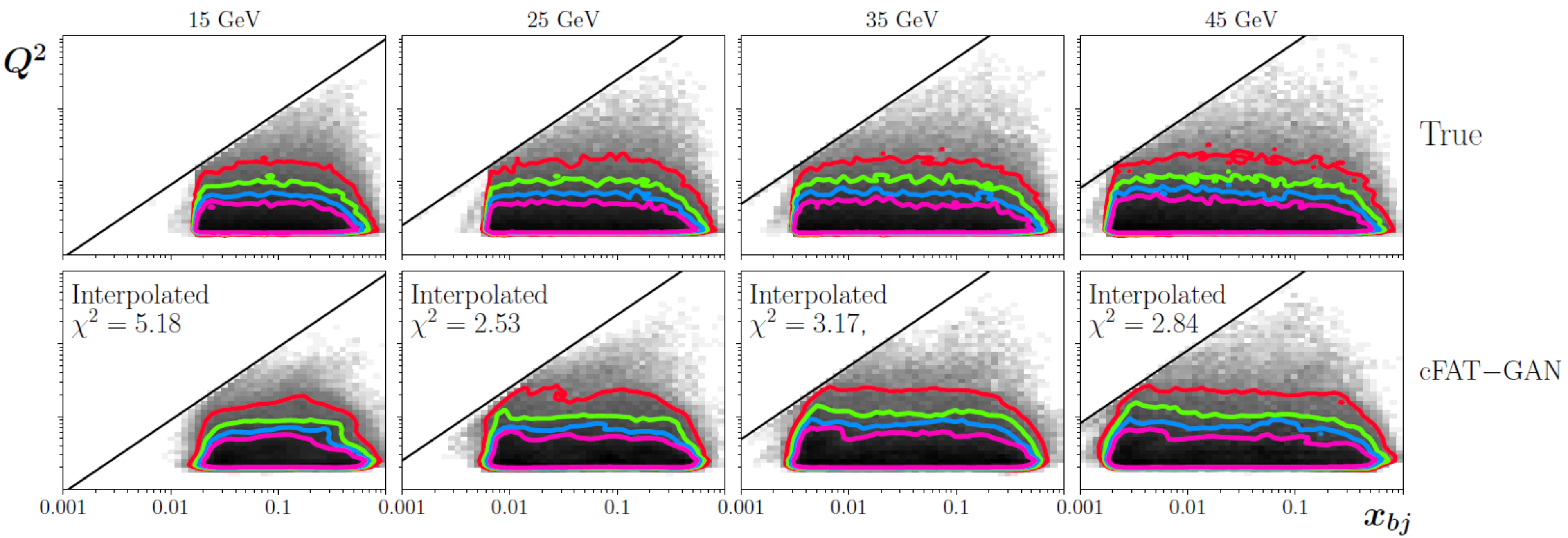
- **A GAN-based Event Generator w.r.t. Beam Energy Input**



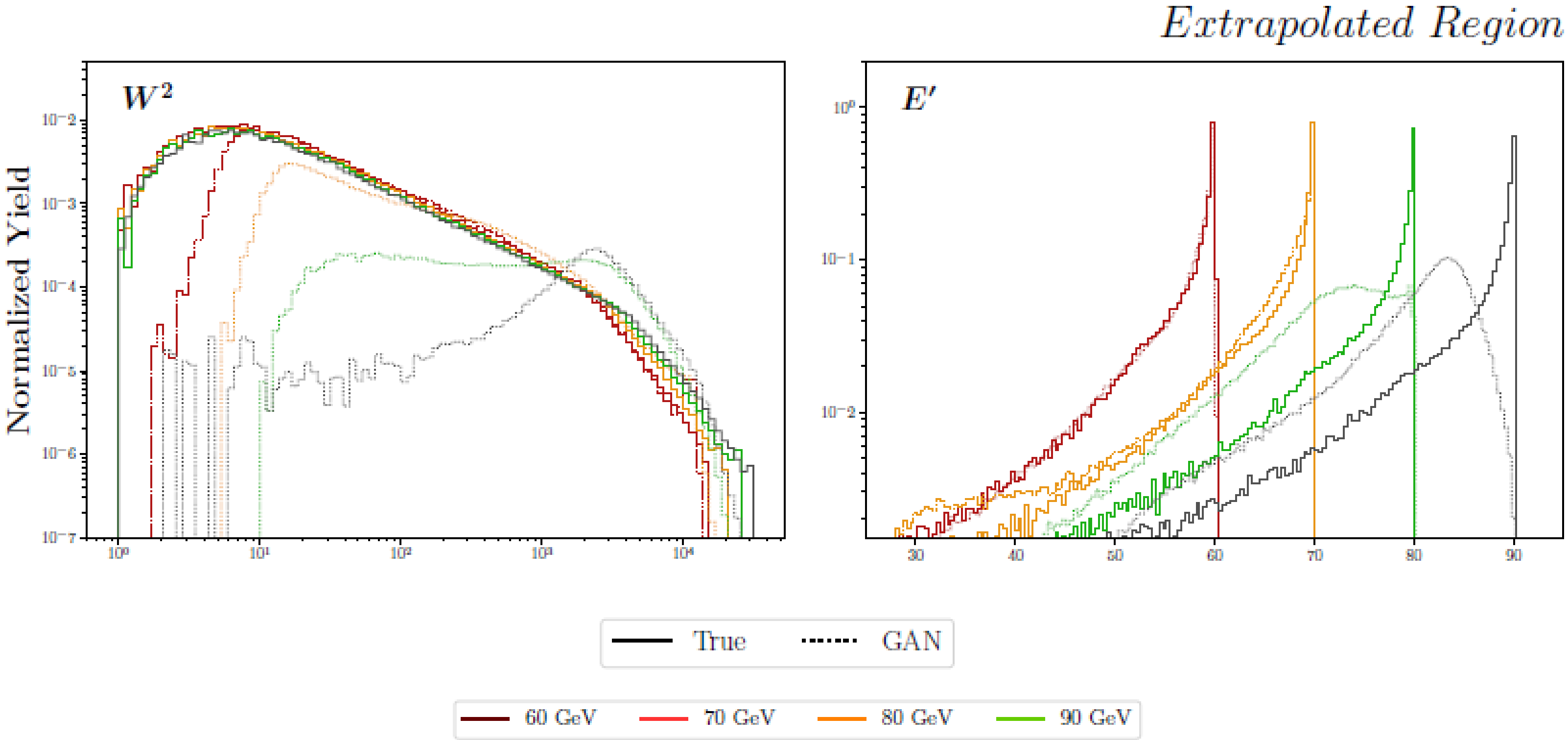
Interpolation



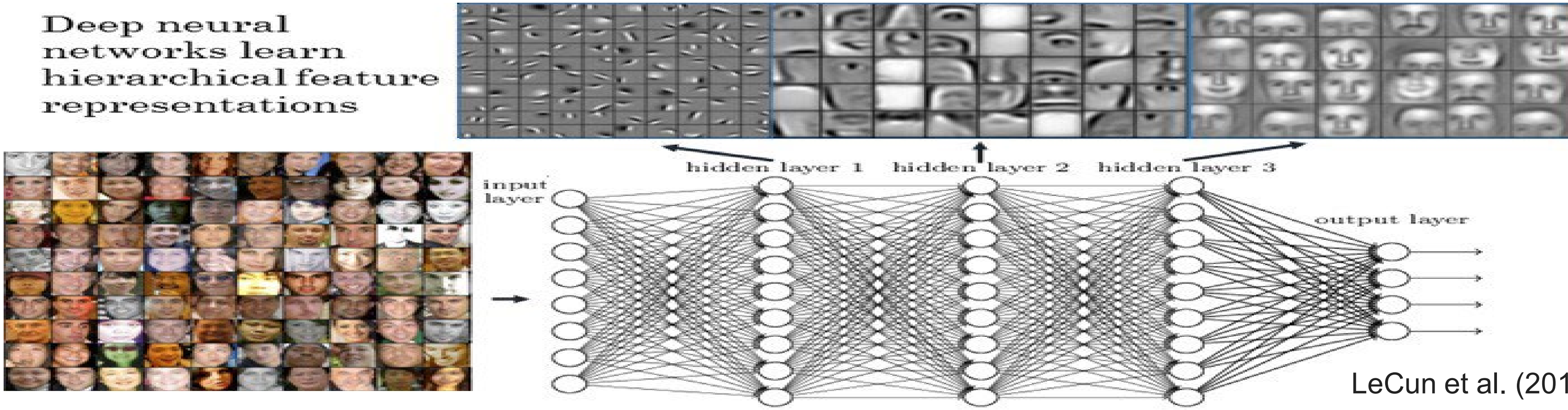
Correlations in Interpolated Beam Energy Levels



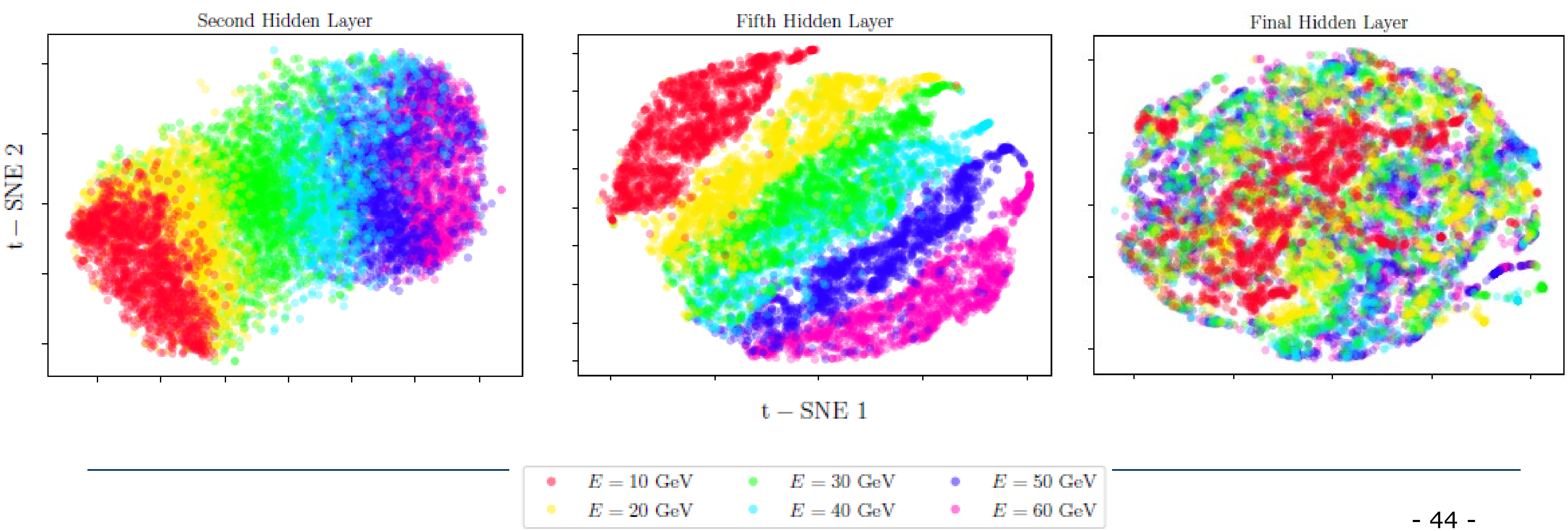
Extrapolation



Patterns in Hidden Layers



LeCun et al. (2015)

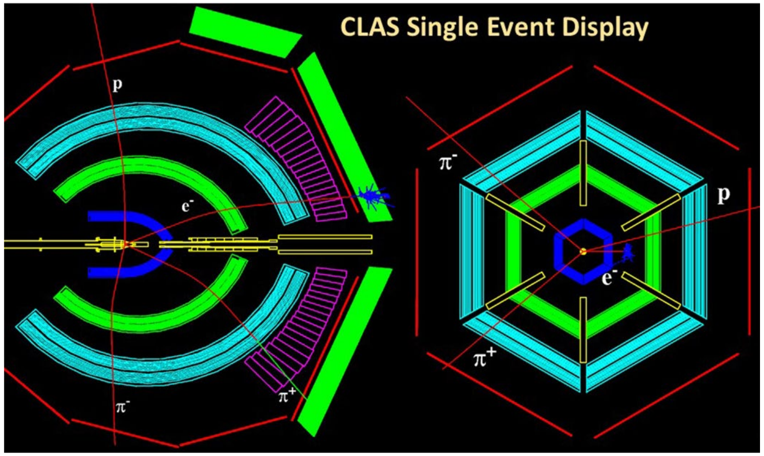


Agenda

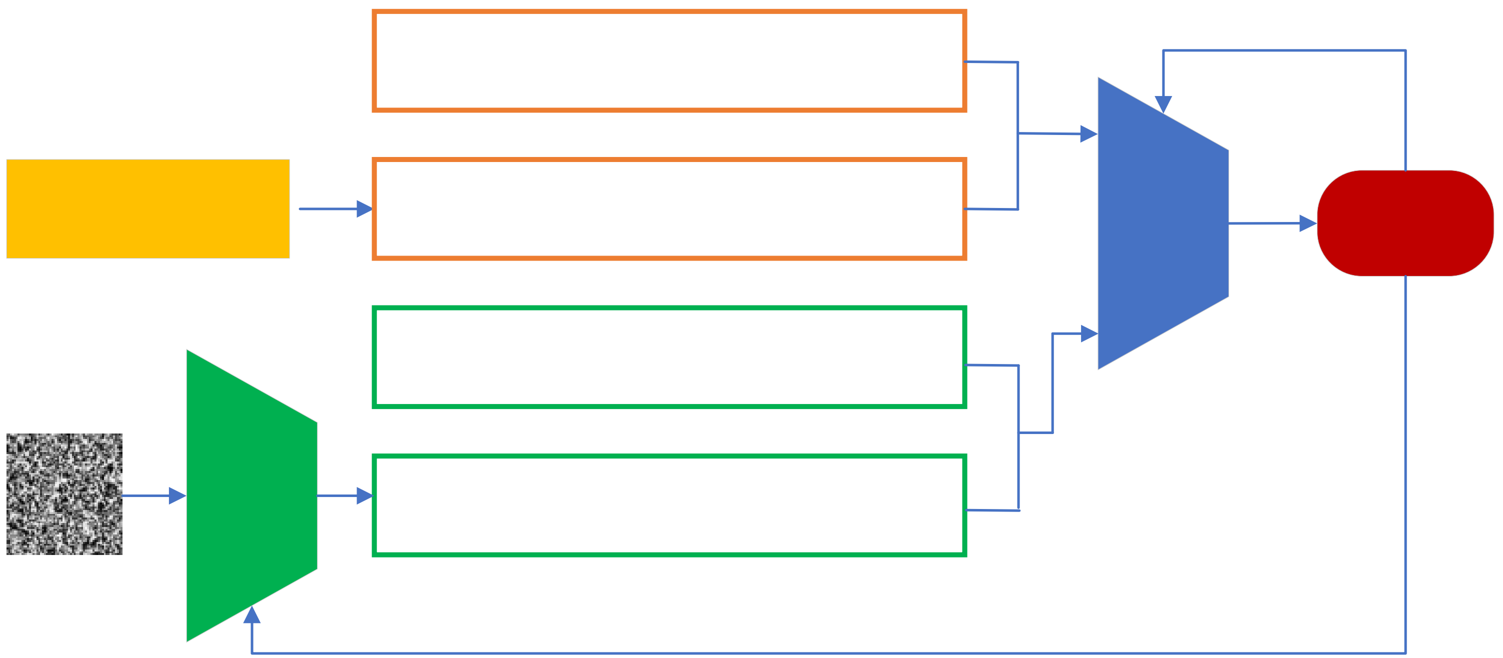
- **Introduction to Adversarial Learning and GAN**
- **Why can GAN work?**
- **Training a GAN-based Monte Carlo Event Generator**
  - Challenges
  - Electron-Proton Scattering
  - Fitting HERA Data
  - Conditional GAN
  - Pion photoproduction on the proton
- **Open Questions**

Pion Photoproduction on the Proton

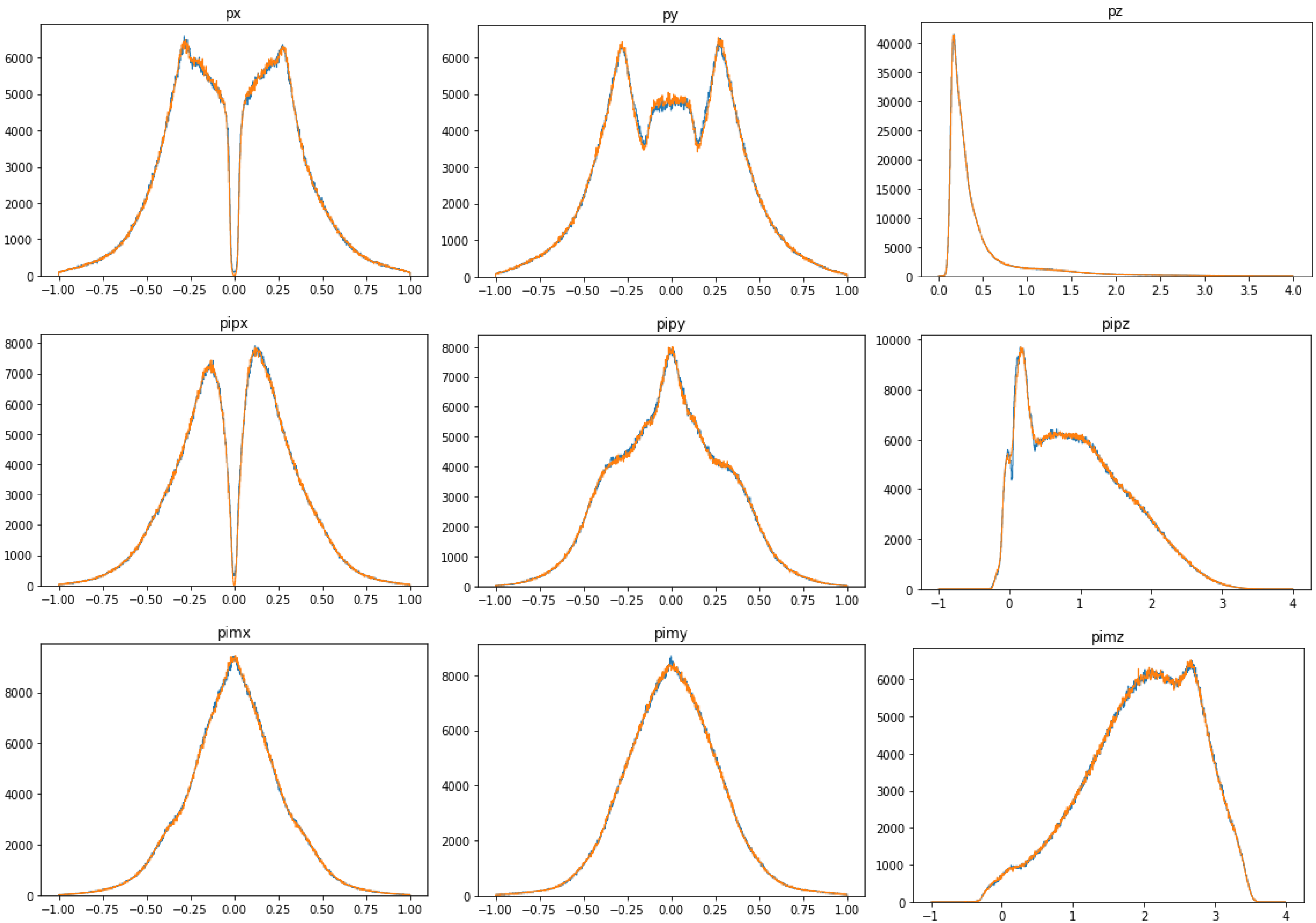
- $\boldsymbol{\gamma p \to p\pi^+\pi^-}$
  - $\boldsymbol{p}$ and $\boldsymbol{\pi^+}$ generated
  - $\boldsymbol{\pi^-}$ reconstructed
  - $\boldsymbol{\gamma} \in [3, 3.75]\mathrm{GeV}$
- **Detector Effect**



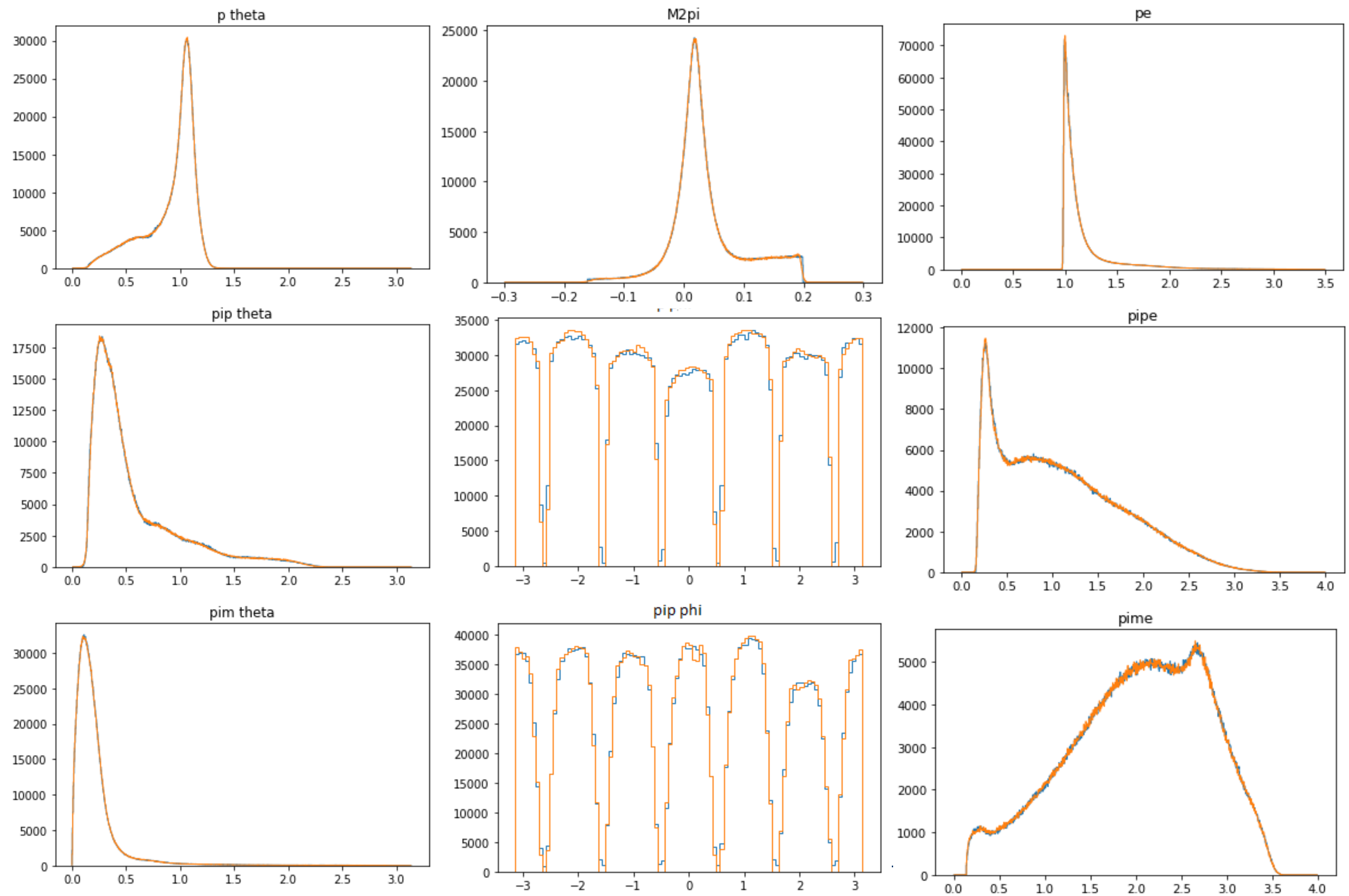
GAN Architecture



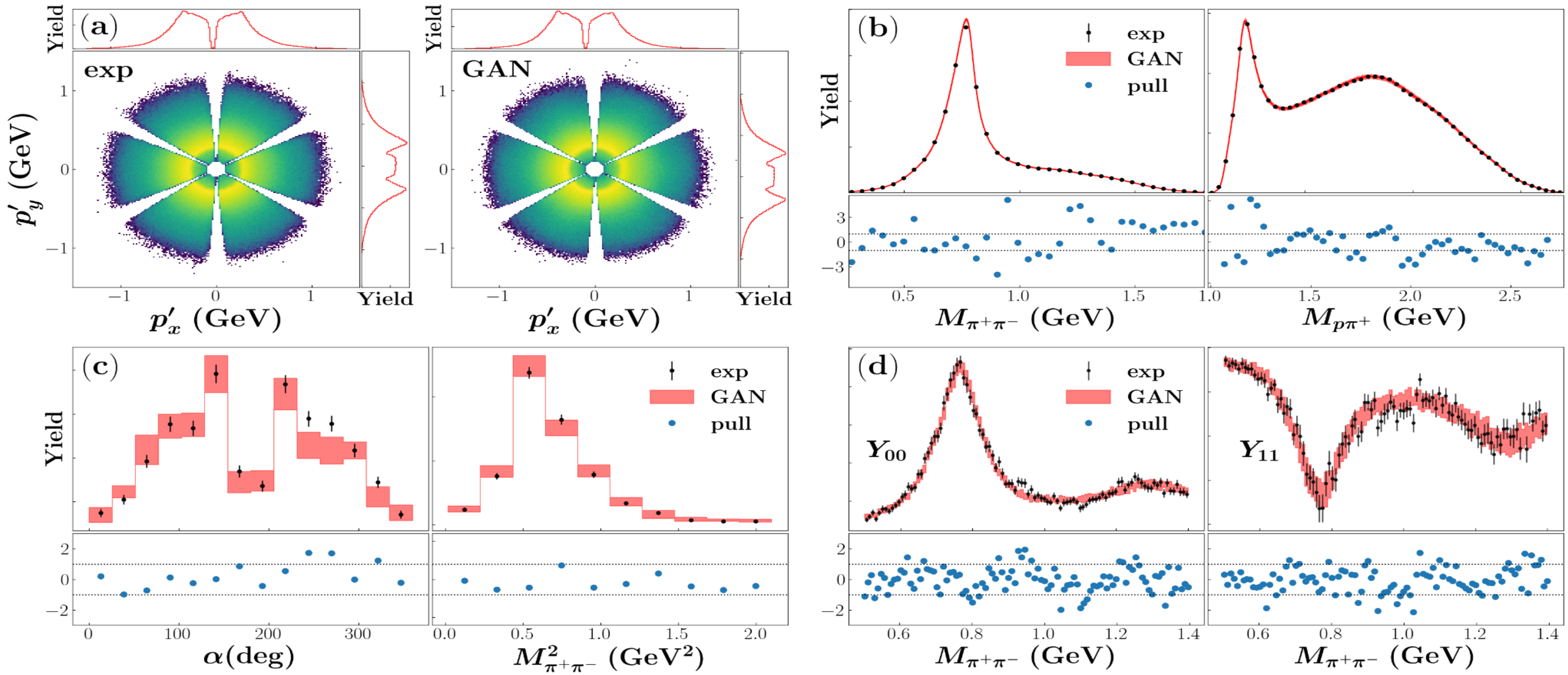
1D Distributions



1D Distributions (cont.)

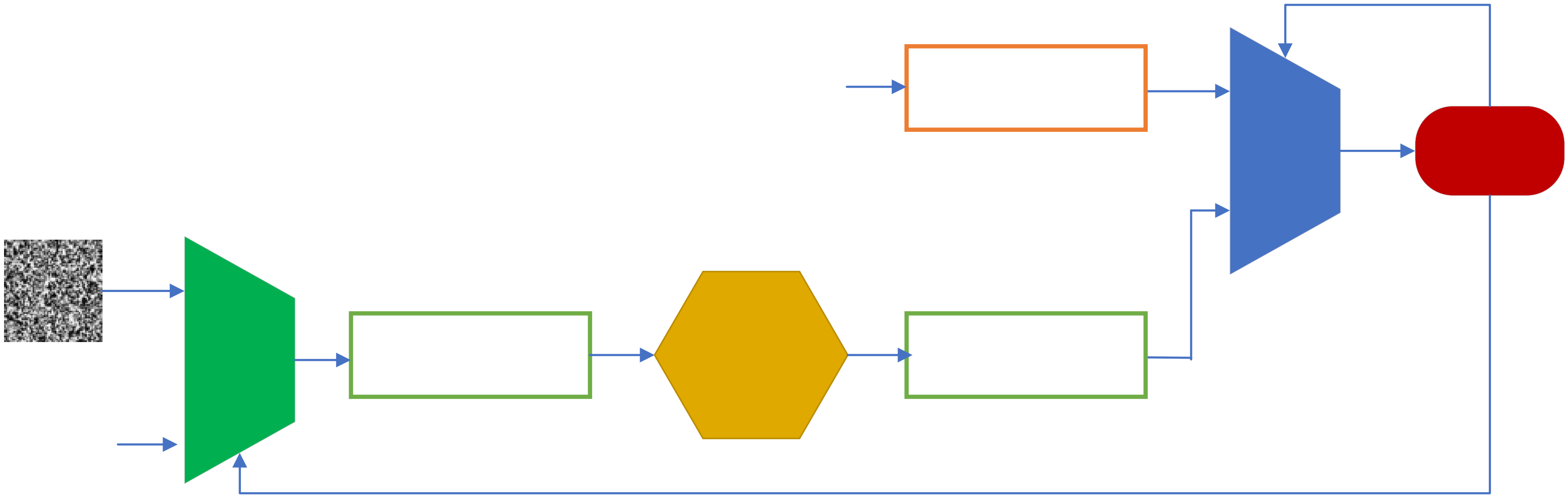


High-ordered Correlations



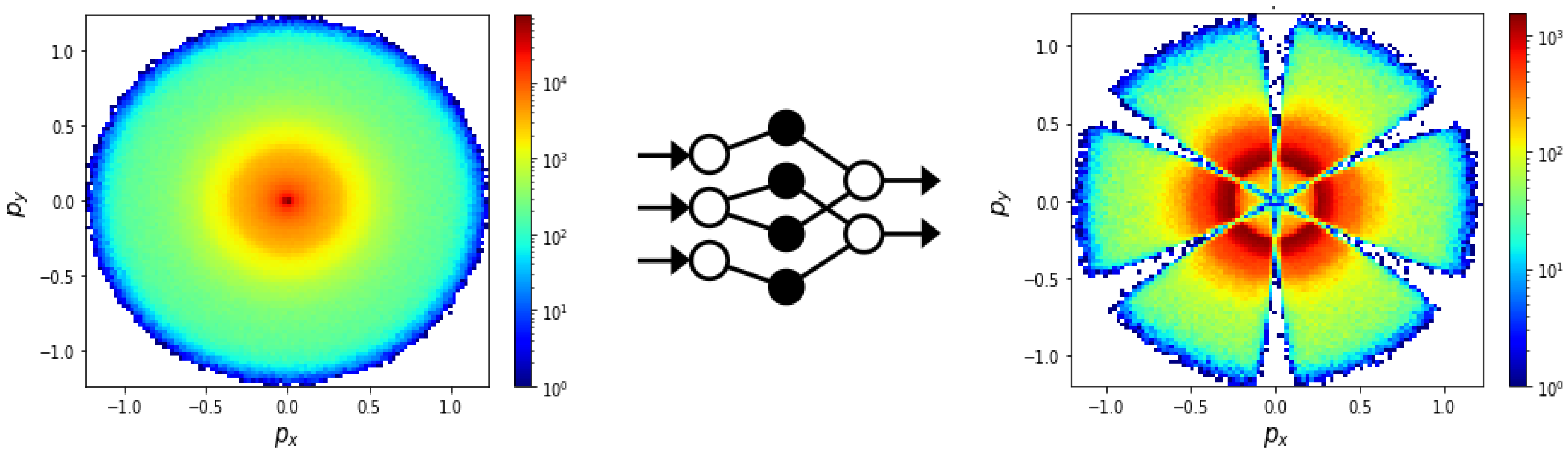
Comparison of selected observables derived from experimental CLAS data and GAN-generated synthetic data: (a) yield for the p'_x and p'_y components of the scattered proton momentum in the lab frame, (b) invariant mass distributions for the pion-pion and proton-pion systems, (c) yields for the angle α (for the bin with 2.55<W<2.60 GeV, 0.74<$M_{\pi\pi}$<0.87 GeV, 1.21<$M_{p\pi}$<1.35 GeV, 0.8<cos(θ_π)<0.9), and invariant mass $M^2_{\pi^+\pi^-}$ (for the bin with 2.70<W<2.75 GeV, 1.37<$M_{p\pi}$<1.51 GeV, 0.9<cos(θ_π)<1.0, 0<α<60°), (d) moments of the angular distributions Y_{00} and Y_{11} versus the ππ invariant mass. For panels (b), (c) and (d), the experimental data (solid black points with error bars) are compared with the GAN-generated results (red bands), with the uncertainty quantification shown in the form of pull distributions given by $(\mu_C-\mu_G)/\sqrt{\sigma_C^2+\sigma_G^2}$ (blue circles at the bottom of the panels).

Detector Effect Simulator

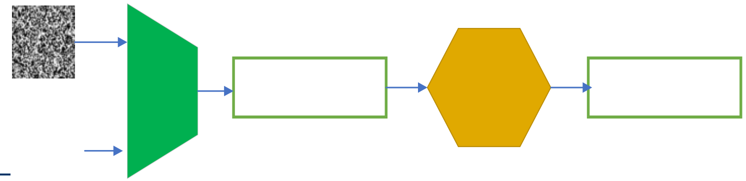


Detector Efficiency Simulator

- **A Neural Network to Map the Detector Efficiency**

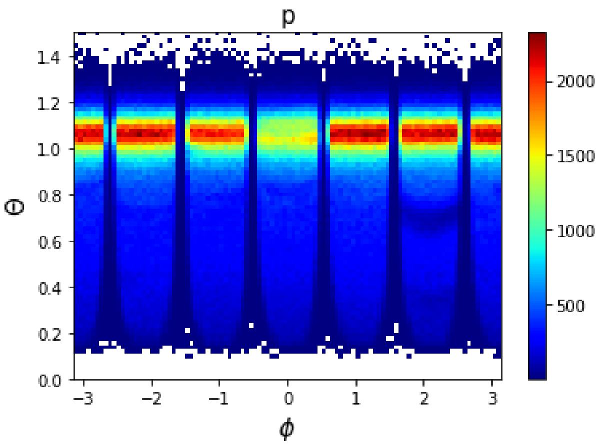


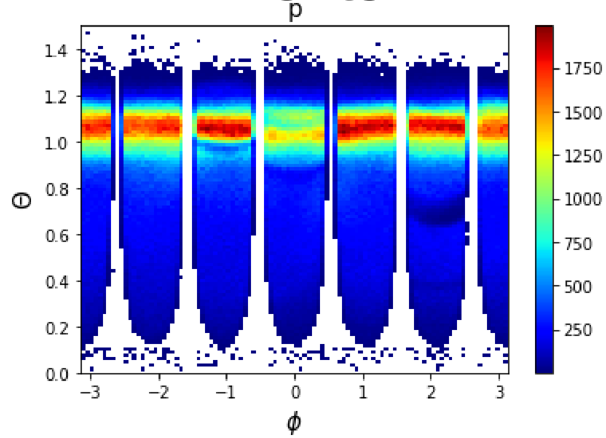
Unfolding Preliminary Results

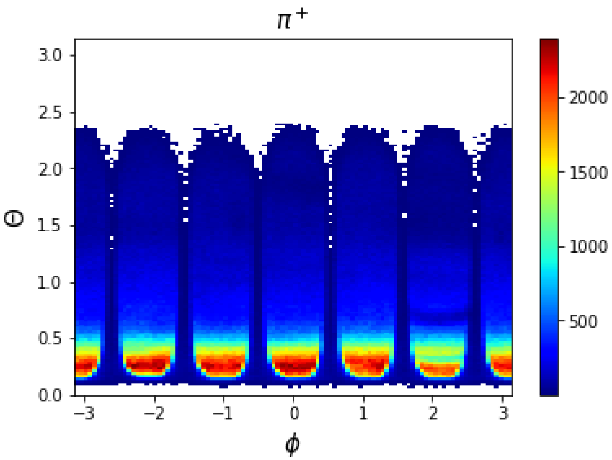


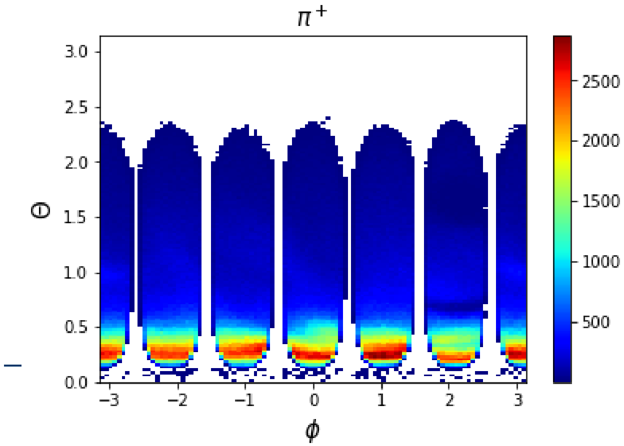
CLAS Events

GAN Detector-Level Events

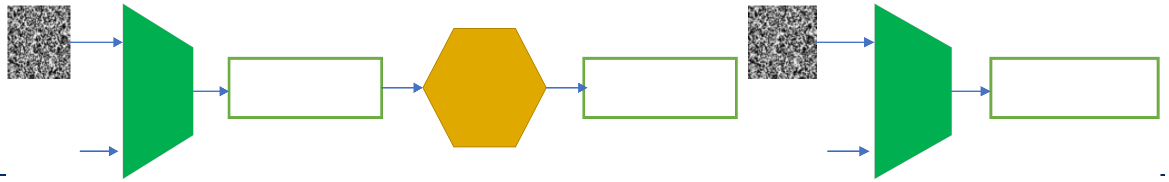




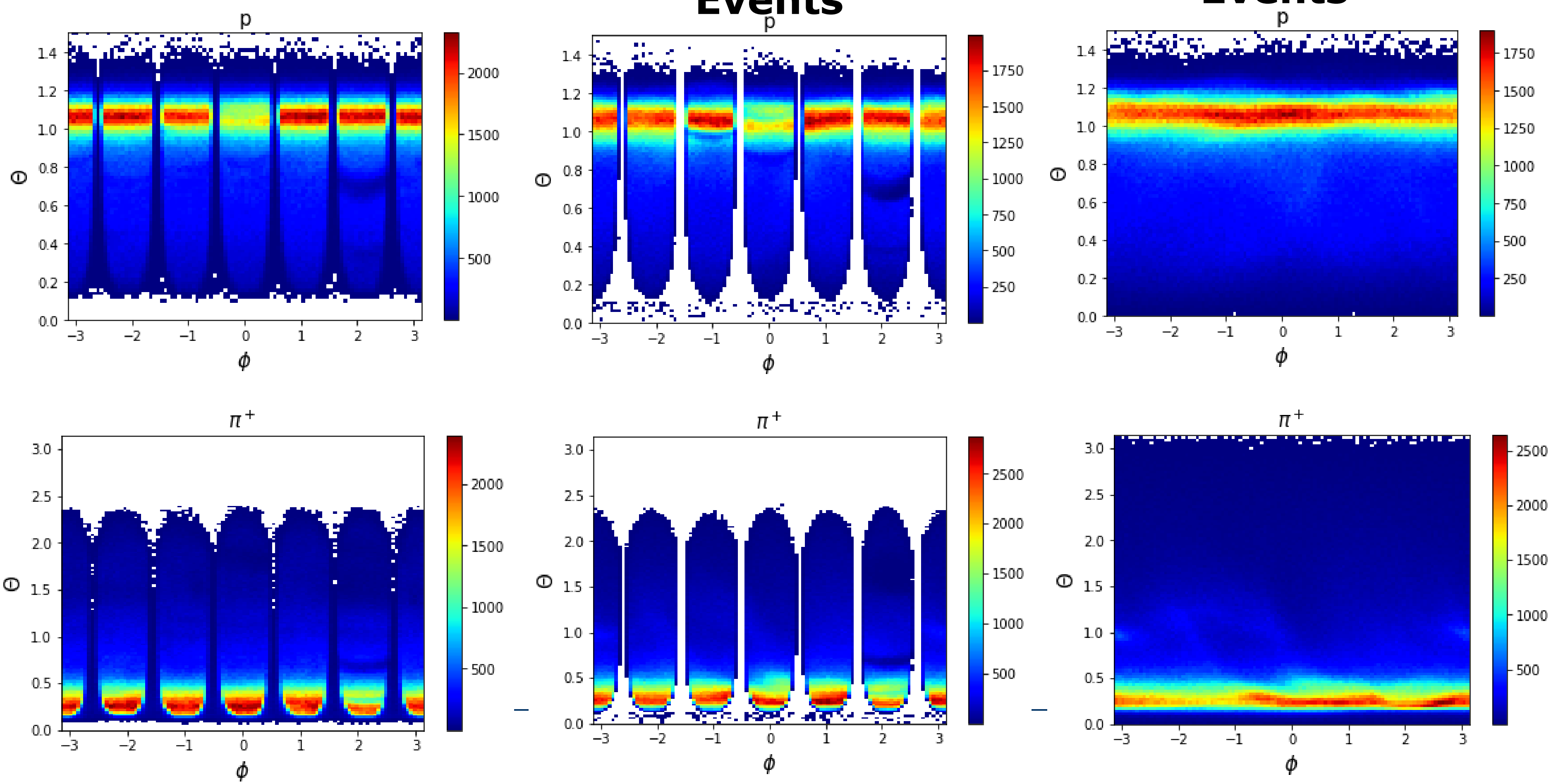




Unfolding Preliminary Results



CLAS Events | **GAN Detector-Level Events** | **GAN Vertex-Level Events**



Agenda

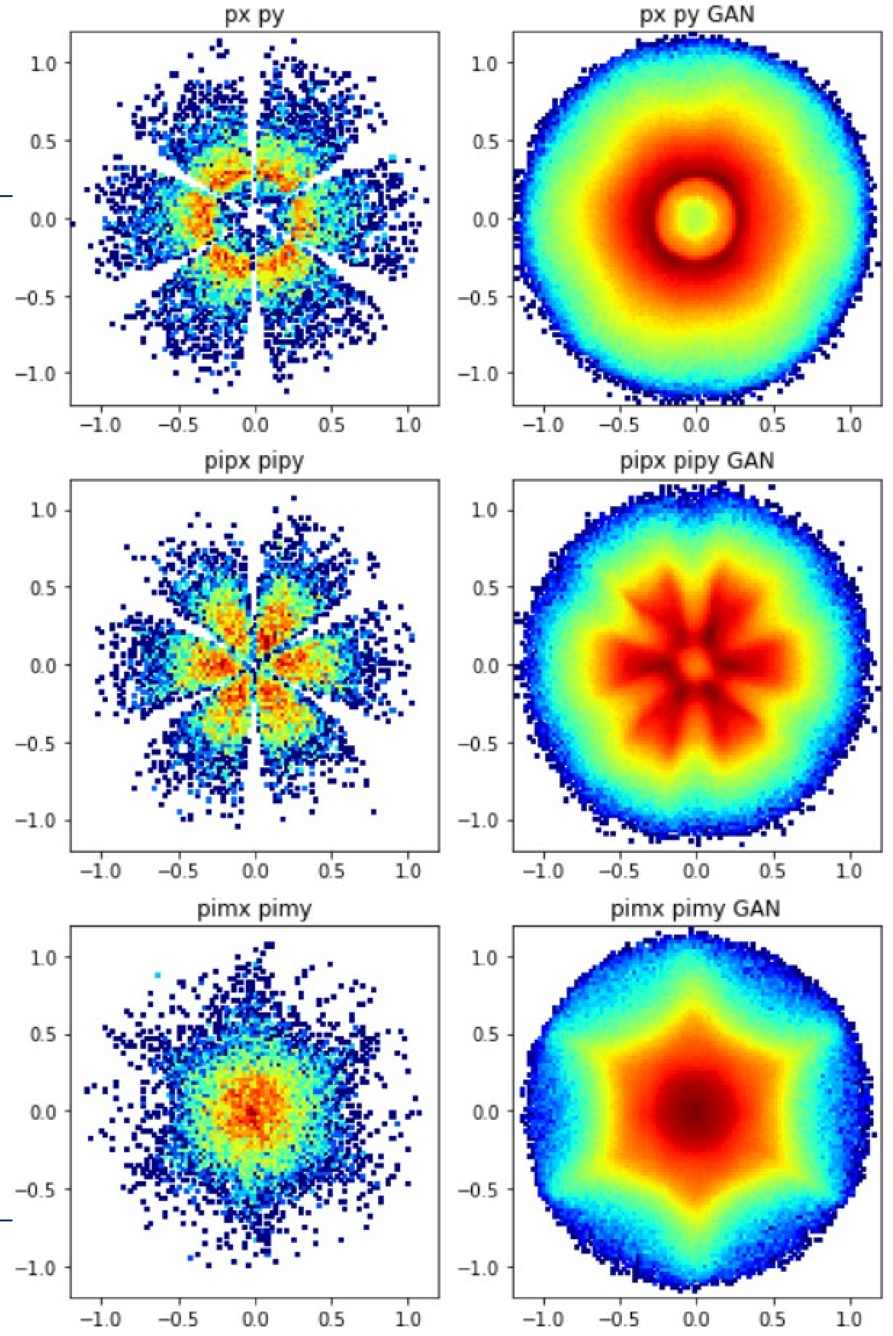
- Introduction to Adversarial Learning and GAN
- Why can GAN work?
- Training a GAN-based Monte Carlo Event Generator
  - Challenges
  - Electron-Proton Scattering
  - Fitting HERA Data
  - Conditional GAN
  - Pion photoproduction on the proton
- **Open Questions**

Open Questions 1) Can GAN-based MCEGs Display Super-Resolution?

- **Can GAN-based MCEGs go beyond the statistical precision of the training event samples?**
- **Only as much statistical precision as the training data can be achieved [Matchev and Shyamsundar, 2020]**
  - An MLEG does not add any physics knowledge
- **Events can be amplified before reaching the limitation of the statistics of the training data [Butter et al., 2020]**
  - MLEGs are powerful interpolation tools
  - Can add to discrete event data sets by enabling denser binning -> higher resolution

Image Source: SRGAN

Super-resolution in CLAS Data



Open Questions 2) Can GAN-based MCEGs Faithfully Reproduce Physics?

- **Can GAN-based MCEGs fully represent the underlying physics of a reaction?**
  - Critical to many MLEG applications in particle physics
  - If not fully, to what extent?
- **Currently, lack of comprehensive evaluation framework to thoroughly evaluate the quality of GAN-based MCEG events**
  - Uncertainty Quantification
  - Quantifying the correlation among event features with physics meaning
  - Measuring the quality of rare events

Open Questions 3) Can GAN-based MCEGs Provide New Physics Insights?

- **Can a GAN-based MCEG go beyond the manifold of its training event data and bring physical insight into regions without any data?**
  - Can GAN-based MCEGs be used for extrapolation?
- **Extrapolation Capability of Neural Network**
  - Output of a neural network is NOT reliable outside of the range of training samples
  - GANs, VAEs, and NFs are fundamentally neural networks
  - GAN-based MCEG yields good agreement for interpolating events, but not in extrapolating events in electron-proton scattering [Velasco et al., 2020]
- **Potential Ways for GAN-based MCEGs to Generate Correct Events in Unknown Regions**
  - Regularizations
    - Physics laws in regularization
  - Use artificial data samples in the unknown region by physics theory or simulation to correct the behavior of GAN-based MCEGs

Physics-informed Machine Learning

- **Pure ML Models**
  - Promising in Physics Applications
    - Computational costly/infeasible
    - Not fully understood process
  - Limitations
    - Large amount of (experimental) data requirement
    - Generalization to lack of sample scenarios
    - Physically inconsistent results
- **Physics-informed ML**
  - Integrate physics and ML in a synergistic way
  - Tackle more complex problems
    - Better generalization
    - Less demand on data
    - Physically consistent
  - ML can reveal unknown physics

Summary

- **Development of GAN-based MCEGs is still in its infant stage**
  - Many Challenges
  - Incorporating physics into Machine Learning models is the KEY
- **GAN-based MCEGs are not likely to replace classic MCEGs**
  - MCEGs are used to verify the underlying theory
  - Alternative approach of MCEGs to generate physics events
    - Much faster event generation
    - Agnostic of theoretical assumptions
  - Important Applications if the open questions can be justified:
    - Super-resolution
      - Remedy the statistical weakness of MCEGs
    - Extrapolation
      - New Physics Insights
    - Faithful reproduction
      - Compactified data storage utility

Related Publications

1. Y. Alanazi, P. Ambrozewicz, M. Battaglieri, G. Costantini, A. Hiller-Blin, E. Isupov, T. Jeske, Y. Li, L. Marsicano, W. Melnitchouk, V. Mokeev, N. Sato, A. Szczepaniak, T. Viducic, “Artificial Intelligence based data reduction and interpretation for subatomic particle scattering,” to be submitted, Nature Machine Intelligence, 2022.
2. Y. Alanazi, N. Sato, T. Liu, W. Melnitchouk, M. P. Kuchera, E. Pritchard, M. Robertson, R. Strauss, L. Velasco, Y. Li, “Simulation of electron-proton scattering events by a Feature-Augmented and Transformed Generative Adversarial Network (FAT-GAN),” Proceedings of 30th International Joint Conference on Artificial Intelligence (IJCAI-21), 2021.
3. Y. Alanazi, N. Sato, P. Ambrozewicz, A. N. Hiller-Blin, W. Melnitchouk, M. Battaglieri, T. Liu, Y. Li, “A Survey of Machine Learning based Physics Event Generation,” Proceedings of 30th International Joint Conference on Artificial Intelligence (IJCAI-21), 2021.
4. M. Almaeen, Y. Alanazi, N. Sato, W. Melnitchouk, M. Kuchera, Y. Li, “Variational Autoencoder Inverse Mapper: An End-to-End Deep Learning Framework for Inverse Problems,” Proceedings of International Joint Conference on Neural Networks (IJCNN2021), 2021.
5. Y. Alanazi, P. Ambrozewicz, M. P. Kuchera, Y. Li, T. Liu, R. E. McClellan, W. Melnitchouk, E. Pritchard, M. Robertson, N. Sato, R. Strauss, L. Velasco, “AI-based Monte Carlo event generator for electron-proton scattering,” arXiv:2008.03151, 2020.
6. L. Velasco, Y. Alanazi, E. McClellan, P. Ambrozewicz, N. Sato, T. Liu, W. Melnitchouk, M. P. Kuchera, Y. Li, “cFAT-GAN: Conditional Simulation of Electron-Proton Scattering Events with Variate Beam Energies by a Feature Augmented and Transformed Generative Adversarial Network,” Proceedings of 19th IEEE International Conference on Machine Learning and Applications (ICMLA2020), 2020.

Codes: https://github.com/JeffersonLab/FAT-GAN

Collaborator Acknowledgements

Nobuo Sato
JLab

Wally Melnitchouk
JLab

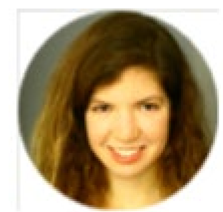
Tianbo Liu
JLab/Shandong Univ.

Marco Battaglieri
INFN/JLab

Pawel Ambrozewicz
JLab

Victor Mokeev
JLab

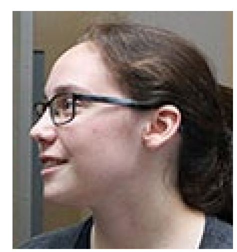
Evgeny Isupov
INFN



Astrid Hiller Blin
JLab

Michelle P. Kuchera
Davidson

Yasir Alanazi
ODU



Luisa Velasco
University of Dallas

Florian Hauenstein
MIT

Acknowledgements

- We thank Jianwei Qiu for helpful discussions.
- This work was supported by the LDRD project No. LDRD19-13 and No. LDRD20-18.

MLEG without Detector Effects as Baseline

MLEG

- Trained MLEG using DIS pseudodata
- Without detector effects