# Modeling Experimental Effects and Generative Methods

Abdullah Farhat



June 28, 2024


# Acknowledgements

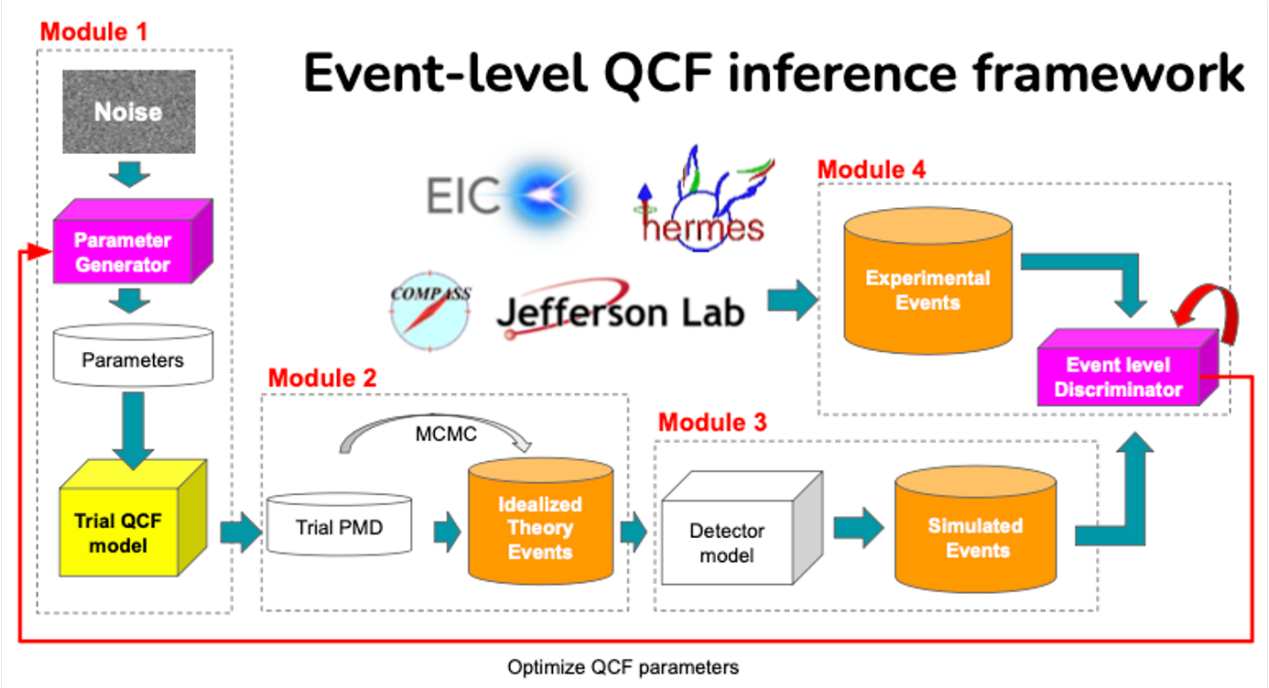
Thanks to Markus Diefenthaler and Daniel Lersch.

Thanks to Nobuo Sato.

Thanks to INT.

# Event Level Workflow

Practical Goal: Develop an advanced framework that can infer "pictures" of nucleons and nuclei from event-level data to reveal their 3D quark and gluon structure.




- Workflow constructed for a joint analysis of theory and experiment.
- Optimal precision with the consideration of full event details.
- Support for real-time analysis.

# Modeling Detector Effects

Some experimental effects considered:

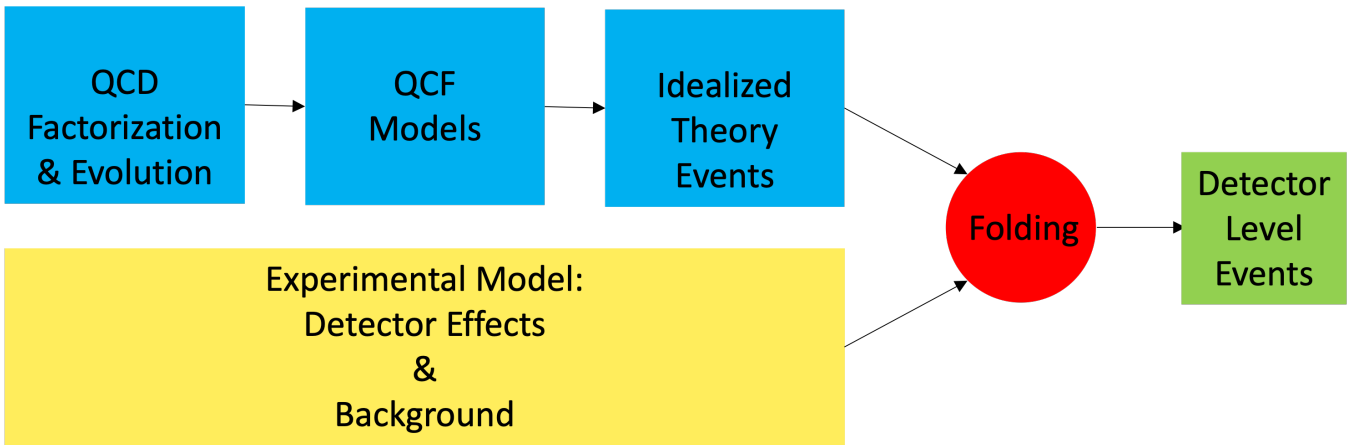
- Radiative effects
- Resolution effects
- Misalignment and calibration of detectors
- Detector inefficiency
- Acceptance, for example, obstructions to measurements due to support structures in detectors or kinematic constraints in analysis that limit the phase space

To be considered later:

- misidentification of particle type
- accelerator background
- background events, like photoproduction for DIS or $\pi^0$-production for DVCS

# Folding

Recent advances in theory allow to compare experiment and theory at the detector level.



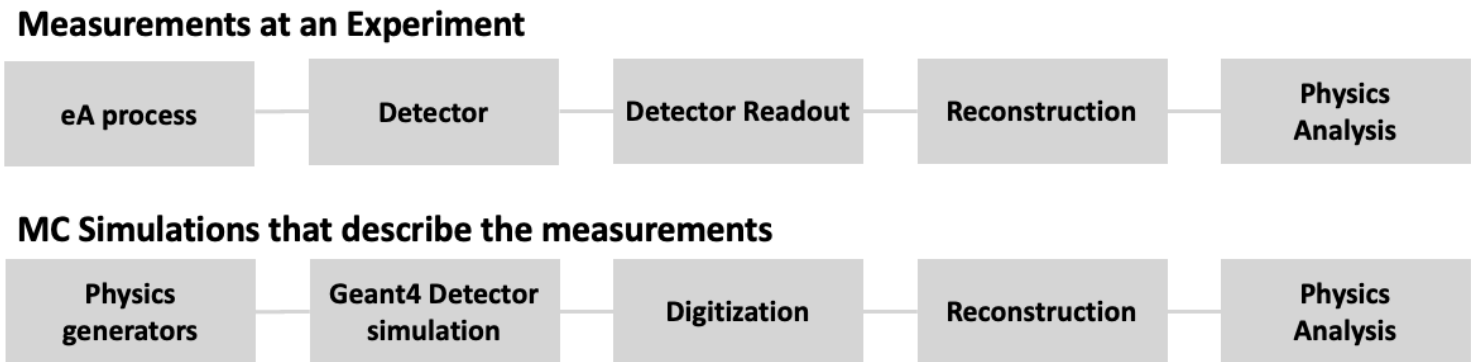

The folding approach enables us to align assumptions and cuts in theory in an unprecedented manner, reducing mismatches between the phase space covered by theory and experiment.

Additionally, folding is robust against changes in both theory and experimental data. Variations in the theory can be rigorously studied, and additional experimental data can be incorporated without necessitating any modifications.

# Experimental Modeling

**Goal:** Fold *idealized events* from theory with experimental effects to generate *detector-level events*.

Developed an event-level approach to model experimental effects from detailed simulations of the experiment.



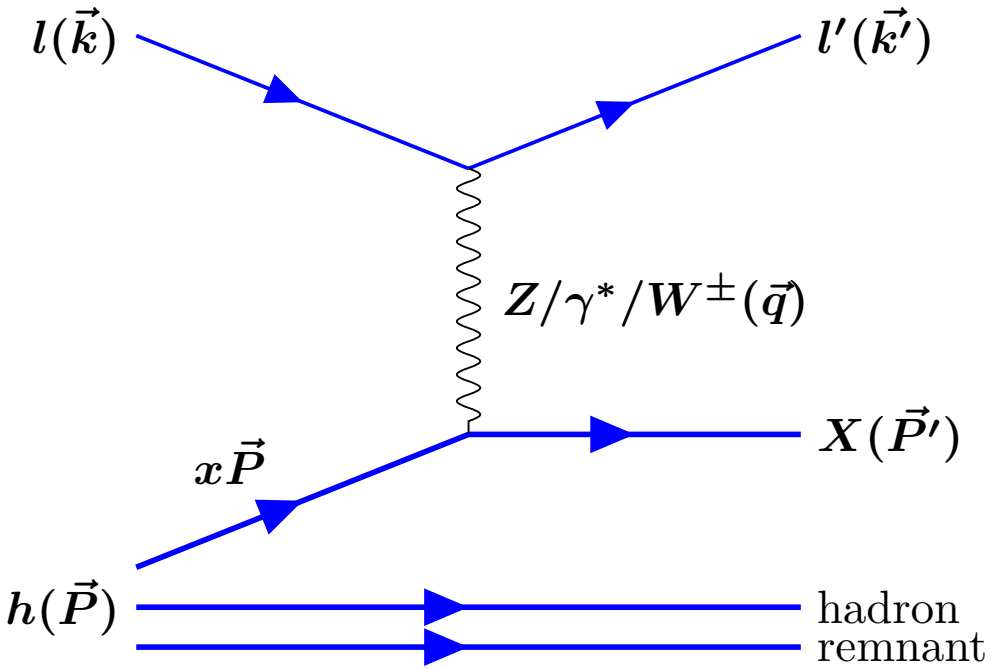

Two methods studied for the detector effects:

- *Deterministic:* DNN
- *Probabilistic:* Generative NN (Variational Auto-Encoder)

The DNN and VAE results are comparable. VAE demonstrates better performance for the examples studied.

# Physics Case

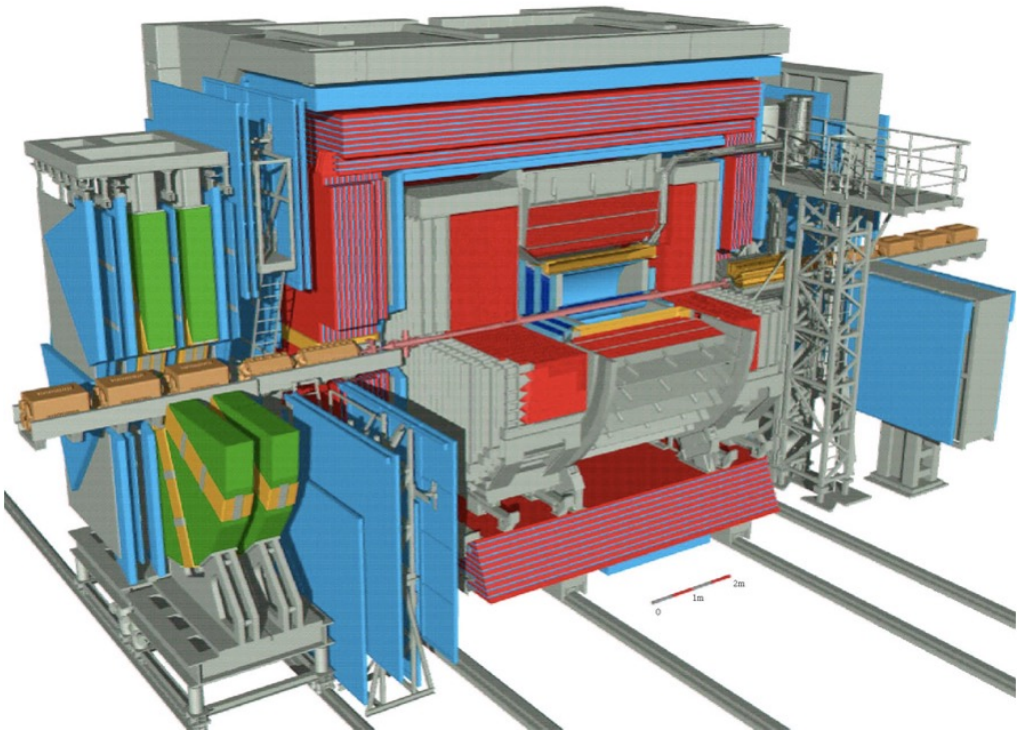
The studies of deep inelastic scattering (DIS) in the lepton-nucleon collisions give insight into the structure of nucleons.




$$l(\vec{k}) + h(\vec{P}) \rightarrow l'(\vec{k'}) + X(\vec{P'}).$$

# Experimental Setup: ZEUS

Using ZEUS data was motivated by a previous analysis...



M. Diefenthaler, A. Farhat, A. Verbytskyi, and Y. Xu, "Deeply learning deep inelastic scattering kinematics," Eur. Phys. J. C, vol. 82, no. 11, p. 1064, 2022.

# Experimental Setup: ZEUS

We concentrate on the neutral current DIS events, i.e. those with an electron in the final state and utilize the **simulated** data of ZEUS experiment and reconstruct the four-momentum transferred to the hadronic system, $Q^2 = -\vec{q} \cdot \vec{q} = -(\vec{k} - \vec{k'})^2$ and the Bjorken scaling variable $x = \frac{Q^2}{2\vec{P} \cdot \vec{q}}$.

---

- Simulated and real data is available for analyses.
- Data preservation efforts led to convenient and accessible data samples.
- The documentation was appropriate to start an analysis of the ZEUS data.

# ZEUS MC

- The Ariadne and Lepto programs were used to simulate the inclusive DIS process using parts of the Pythia6 framework for the simulation of hadronization processes and decays of particles.
- The following models were trained on the events generated from Ariadne, but tested on events generated from Lepto.
- HERACLES was used for implementations of the higher order QED and QCD corrections.
- The simulation of the particle transport through the detector material and the simulation of detector response from the simulation of the DIS collision event in the detector was performed in Geant.

# Event selection (most important cuts)

- **Detector status:** It was required that for all the events the detector was functional.
- **Electron energy:** At least one electron candidate with energy greater than $10GeV$
- **Electron identification probability:** The SINISTRA probability of lepton candidate being the DIS lepton was required to be greater than 90%.
- **Electron isolation:** The fraction of the energy not associated to the lepton was required to be less than 10% over the total energy deposited within a cone around the lepton candidate. The cone is defined with a radius of $0.7$ units in the pseudorapidity-azimuth plane around the lepton momentum direction.
- **Electron track matching:** The tracking system covers the region of polar angles restricted to $0.3 < \theta < 2.85$. If the lepton candidate was within the tracking system acceptance region, there must be a matched track. This track must have a distance of closest approach between the track extrapolation point at the front surface of the CAL and the cluster center-of-gravity-position of less than 10 cm. The track energy must be greater than 3GeV.
- **Electron position:** To remove regions poorly described by Monte Carlo simulations, additional requirements on the position of the electromagnetic shower were imposed. The events in which the lepton was found in the following regions were rejected: RCAL where the depth was reduced due to the cooling pipe for the solenoid, regions in-between calorimeter sections, regions close the the beam pipe.
- **Primary vertex position:** It was required that the reconstructed primary was close to the central part of the detector, implying $-28.5 < Z_{\mathrm{vtx}} < 26.7cm$.
- **Energy-longitudinal momentum balance:** To suppress photoproduction and beam-gas interaction background events and poor Monte Carlo simulations, restrictions are put on the energy-longitudinal momentum balance. This quantity is defined as: $\delta = \delta_e + \delta_{\mathcal{H}} = (E_{e'} - P_{z,e'}) + (E_{\mathcal{H}} - P_{z,\mathcal{H}}) = \sum_i (E_i - P_{z,i})$ where the final summation index runs over all energy deposits in the detector. In this analysis we've applied a condition $38 < \delta < 65$ GeV.
- **Missing transverse energy :** To remove the beam-related background and the cosmic-ray events an cut on the missing energy was imposed. $P_{T,miss}/\sqrt{E_T} < 2.5^{1/2}$, where $P_{T,miss}$ is the missing transverse momentum as measured with the CAL and $E_T$ is the total transverse energy in the CAL.

# Detector Surrogate

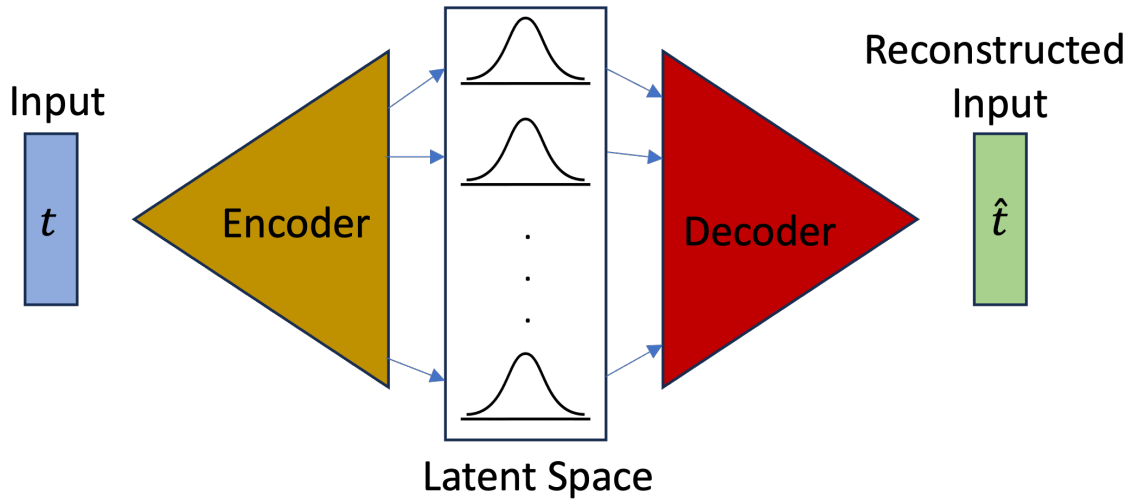
Goal: construct a model that takes *idealized theory events* and folds in detector effects to produce *simulated events*.

$$(x, Q^2) \rightarrow (x', Q^{2\prime})$$

Simulated events can be meaningfully compared to measured experimental events

For training, the simulated events $(x', Q^{2\prime})$ were reconstructed via the Electron Method.

# Variational Auto-Encoder (VAE)

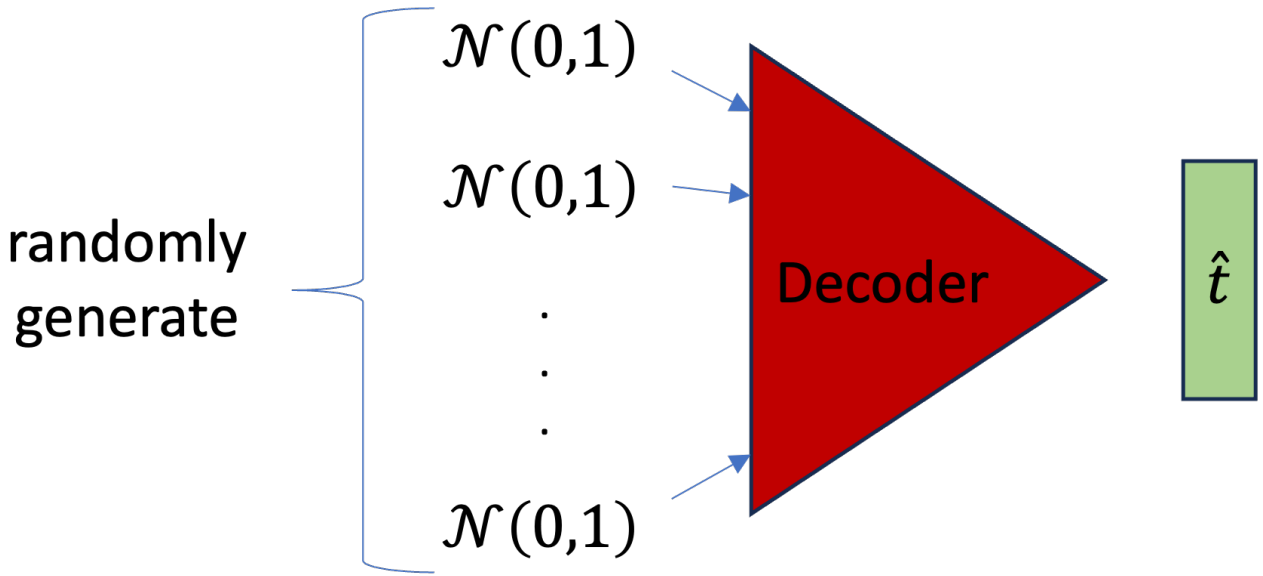



Train model by minimizing the loss over a sample data set:

$$Loss = (MSE) + (KL)$$

- MSE = mean square error, measures difference between input and reconstructed input
- KL Divergence = Kullback-Leibler Divergence, measures difference between latent space distribution and the standard normal distribution

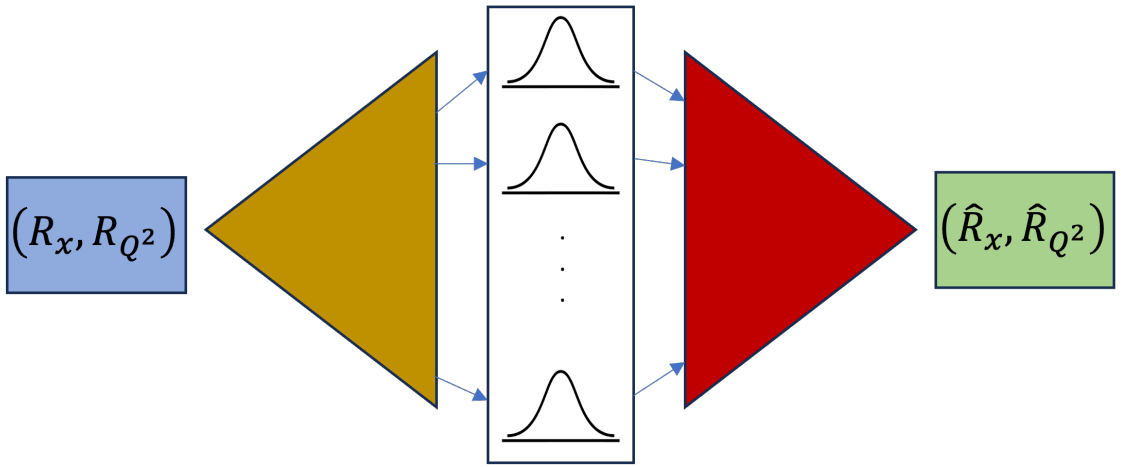
# VAE Inference




VAE is a generative model.

# Detector Surrogate: Variational Auto-Encoder (VAE)

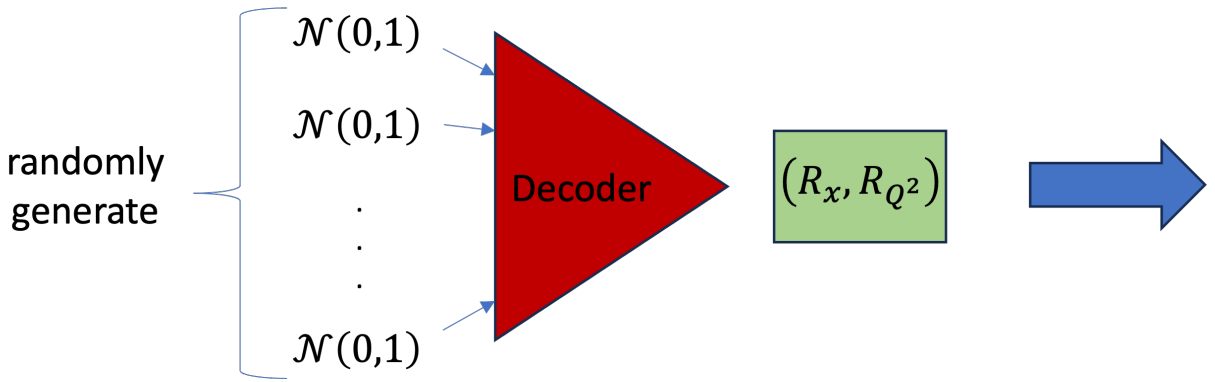
**Method**: Learning normalized residuals.




$$R_x = \frac{x - x'}{x}$$

$$R_{Q^2} = \frac{Q^2 - Q^{2\prime}}{Q^2}$$

**Inference:**




$$x' = (1 - R_x) \cdot x$$

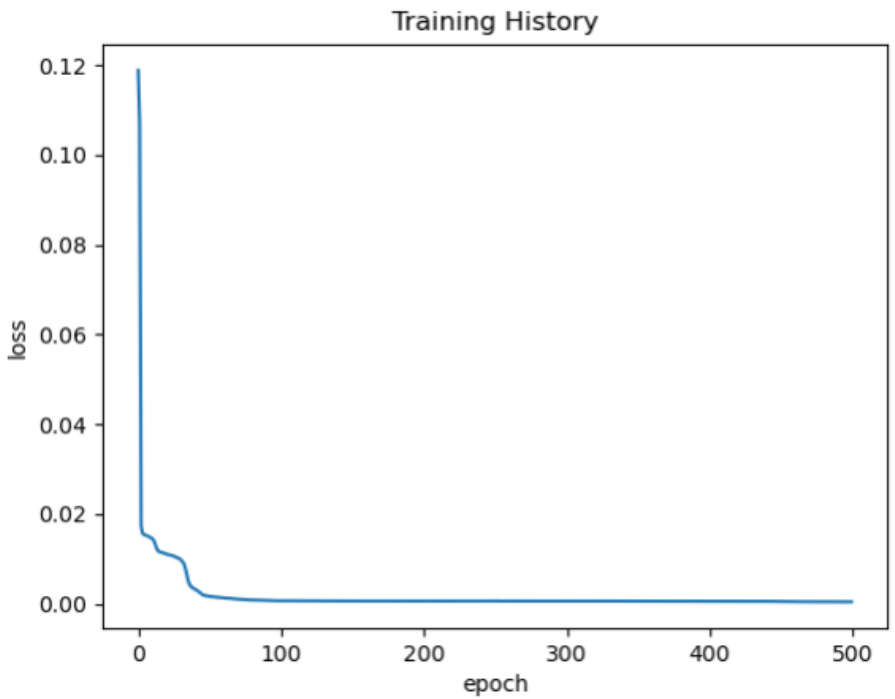
$$Q^{2\prime} = \left(1 - R_{Q^2}\right) \cdot Q^2$$

# ZEUS Example

VAE detector surrogate specifications:

- Encoder hidden layers and units: [50,50,50,100,100]
- Decoder Hidden layers and units: [100,100,50,50,50]
- Latent Dimension 128, RELU activation function.

Training: 20k events, 80/20 train/test split, outliers removed.

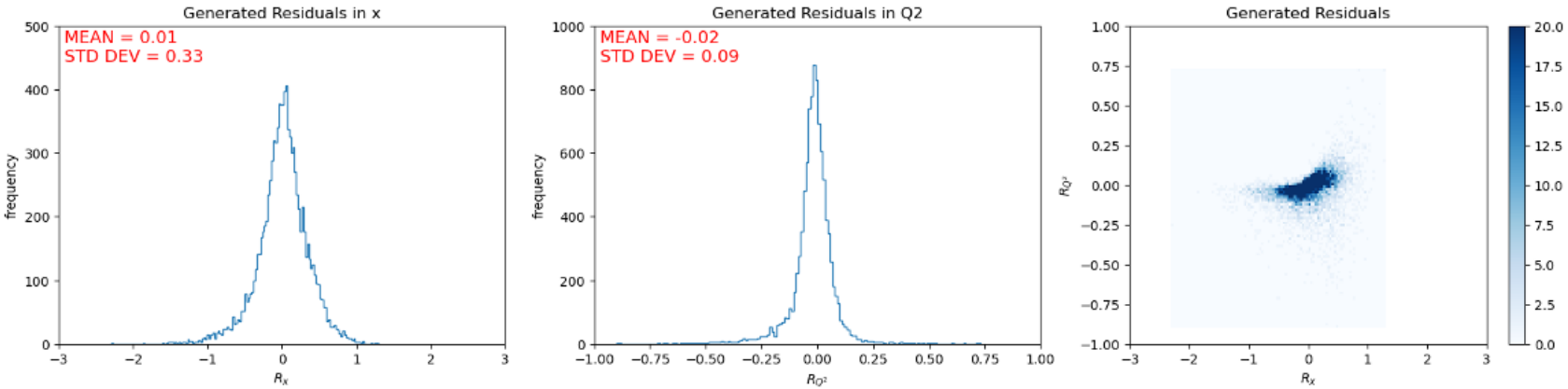



Regularization

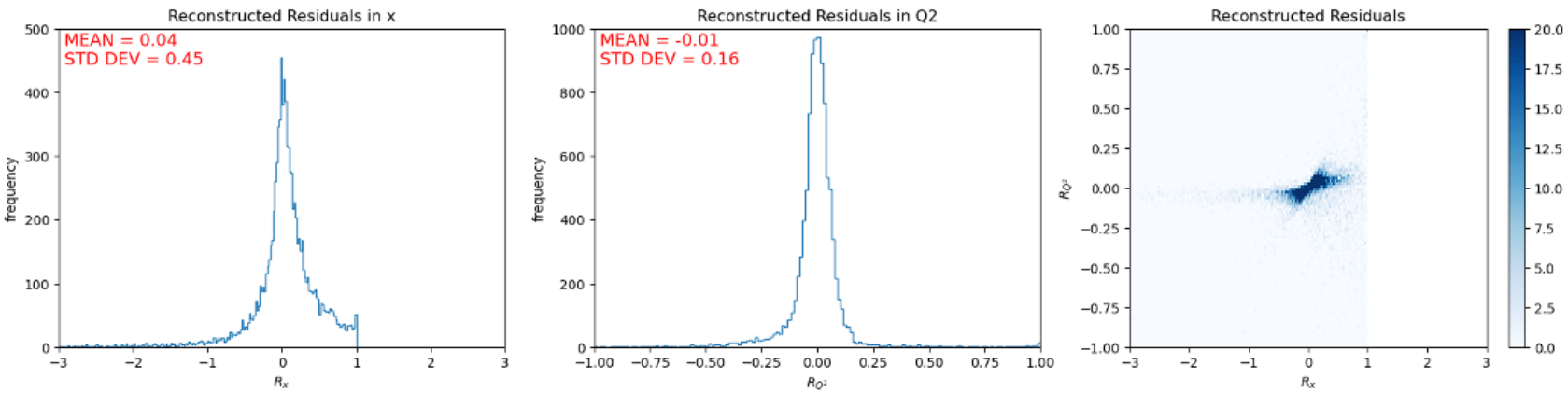
- L2: $10^{-4}$
- KL: $\frac{1}{150}$

# ZEUS Example: Residual Distributions

## VAE Detector Surrogate

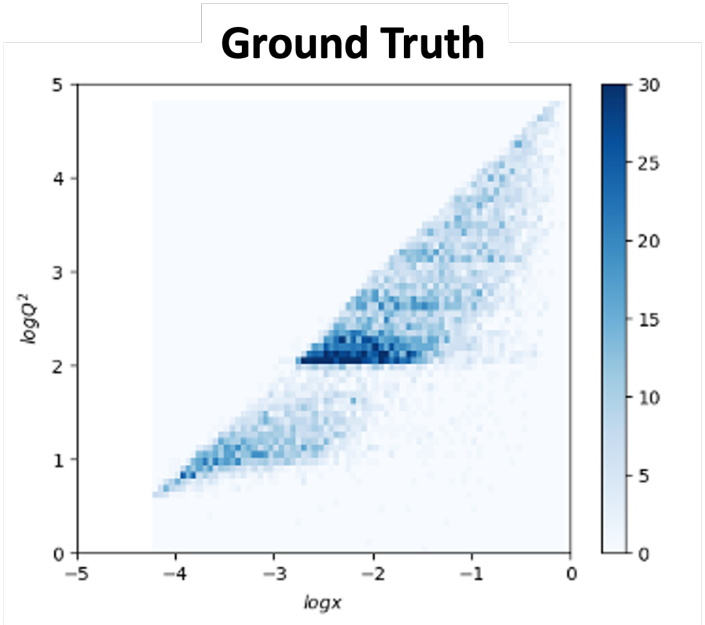


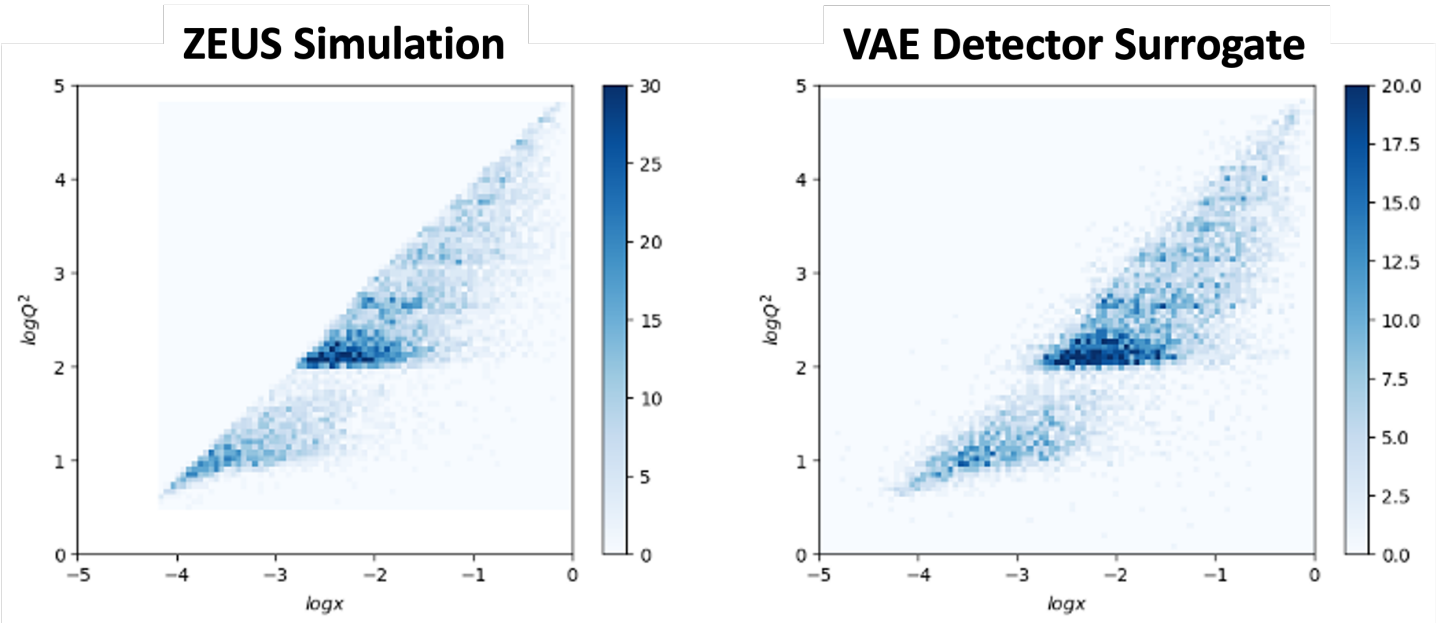

## **ZEUS Simulation** (electron method only)




Developed a customizable detector surrogate and training procedure to model various eA experiments.

# ZEUS Example: $(x, Q^2)$ Distributions



# DNN Model

Construct a DNN that directly modifies observables:

$$(\log x, \log Q^2) \to (\log x', \log Q^{2\prime})$$

# DNN Model

DNN surrogate specifications:

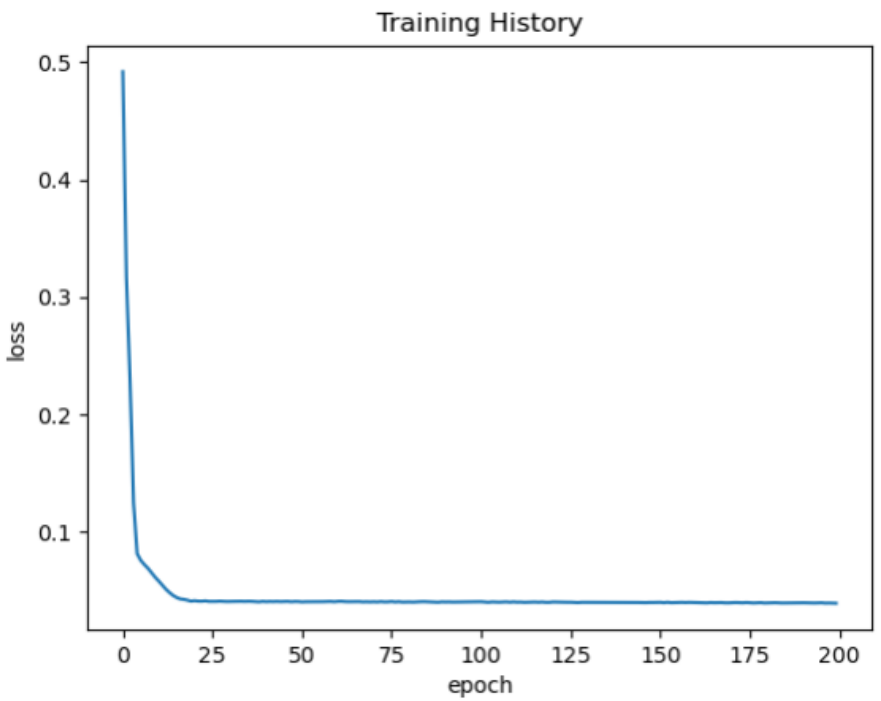
- Hidden layers and units: [100,100,100,50,50,50,50,50,50,50]
- Sigmoid Activation Function

Training: 20k events, 80/20 train/test split.

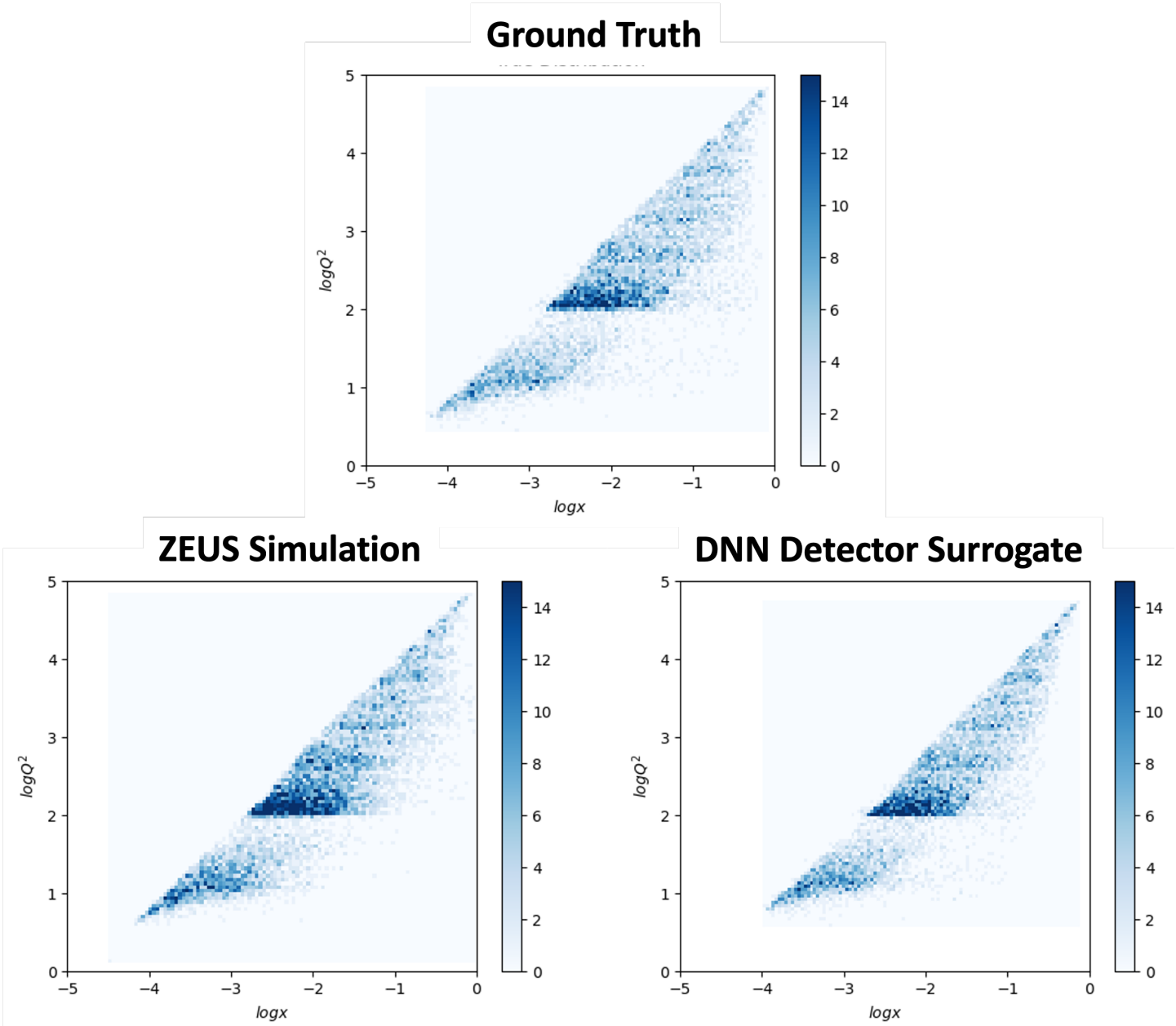
ADAM Optimizer

- Learning rate: 10-4
- Batch size: 16

Regularization: L2: $10^{-7}$



# ZEUS Example: $(x, Q^2)$ Distributions



# Discussion

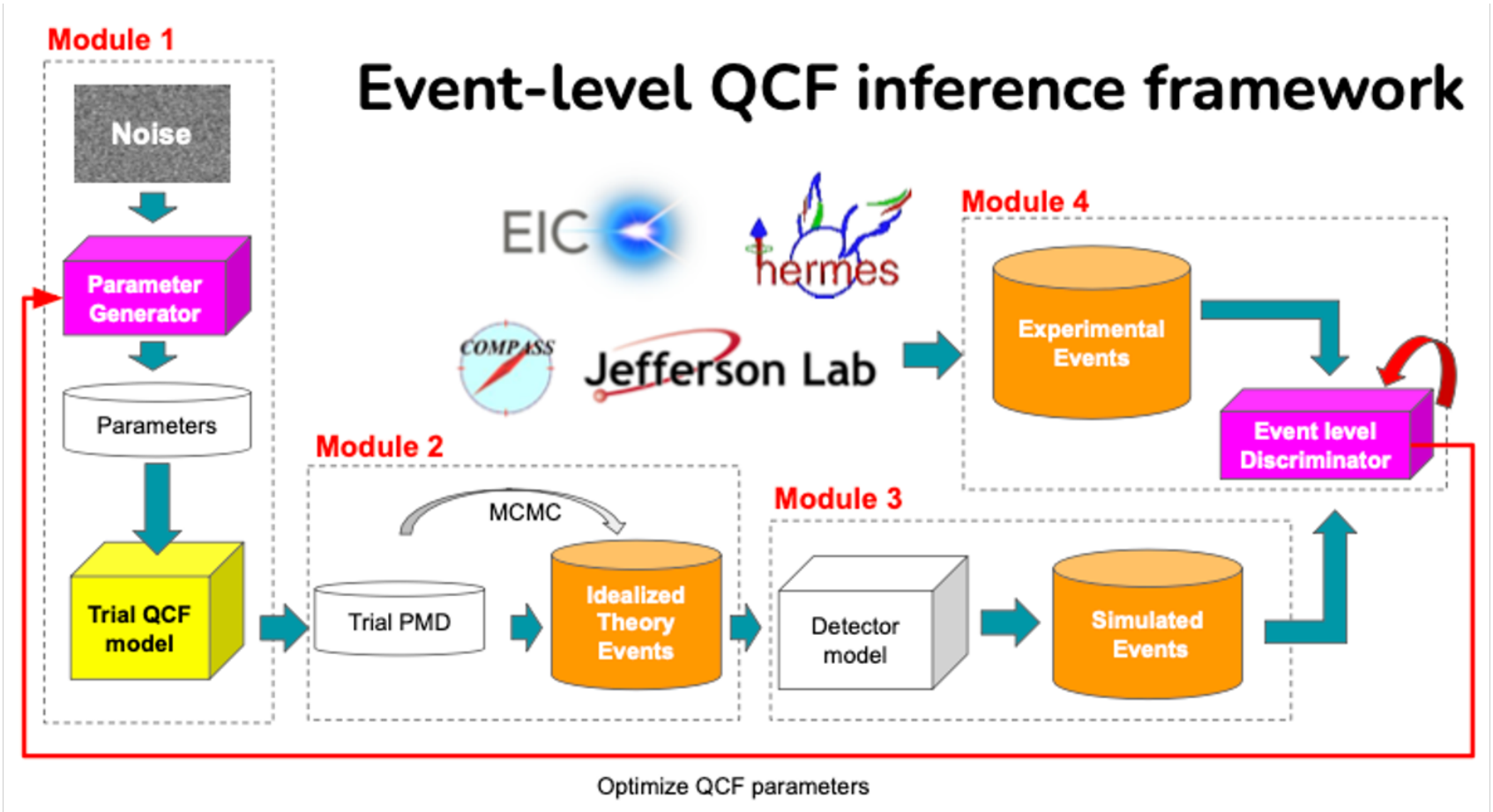
Both DNN and VAE models:

- Maintain a reasonable distribution of events in the $(x, Q^2)$-space.
- Control for outliers where ordinary reconstruction methods are limited or fail.

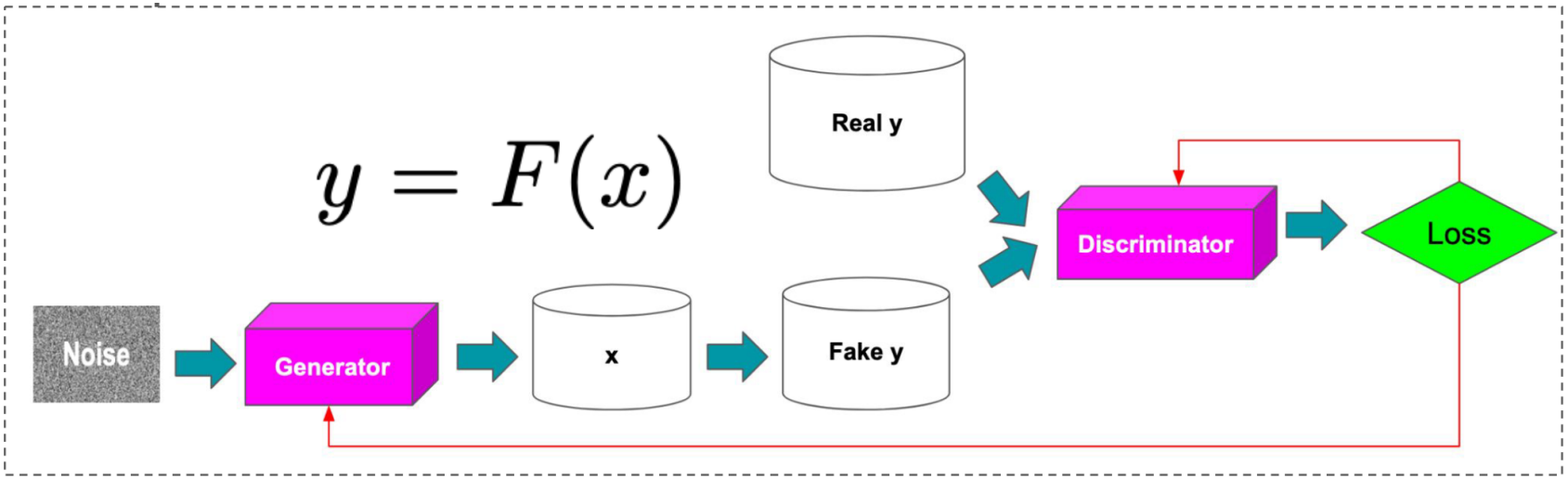
The selection of the ML model controls the inductive bias.

- Determine and quantify necessary requirements for an ML detector model.
- Consider using GANs and other generative architectures

Background contributions need to be considered.


Module 1
Event-level QCF inference framework
Noise
Parameter Generator
Parameters
Trial QCF model
EIC
hermes
COMPASS
Jefferson Lab
Module 4
Experimental Events
Event level Discriminator
Module 2
MCMC
Trial PMD
Idealized Theory Events
Module 3
Detector model
Simulated Events
Optimize QCF parameters

# Concept GAN Inference




A measure-theoretic proof is used to provide a sufficient condition for a description of the optimal discriminator and generator models for the concept GAN inference method.

# Concept GAN Inference

Call the generator $G$ and the disciminator $D$.

The optimal generator and discriminator models are selected by solving the following optimization problem:

$$\min_G \max_D V(D, G) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_z[\log(1 - D(F(G(z))))],$$

where $z$ is noise.

# Lemma

Let $Z \sim P_Z$ be a random variable. If, for some measurable function $f$, $X = f(Z) \sim P_{X,Z}$, then for any measurable function $g$,

$$\mathbb{E}_X[g(X)] = \mathbb{E}_Z[g(f(Z))].$$

Let $(X, A)$ and $(Y, B)$ be two measurable spaces. A function $f : X \to Y$ is measurable if for any $E \in B$, then $f^{-1}(E) \in A$, where $f^{-1}(E) = \{x : f(x) \in E\}$ is the pre-image of $E$.

# Proof of Lemma

For any measurable set $A$,

$$P_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(f(Z) \in A) = \mathbb{P}(Z \in f^{-1}(A)) = P_Z(f^{-1}(A))$$

Let $\chi_A$ be the indicator function of $A$. Then

$$\begin{aligned}
\mathbb{E}_X[\chi_A(X)] &= P_X(A) \\
&= P_Z(f^{-1}(A)) \\
&= \mathbb{E}_Z[\chi_{f^{-1}(A)}(Z)] \\
&= \mathbb{E}_Z[\chi_A(f(Z))].
\end{aligned}$$

Since any measurable function $g$ can be expressed as the limiting value of weighted sum of indicator functions, it follows that

$$\mathbb{E}_X[g(X)] = \mathbb{E}_Z[g(f(Z))].$$

# Optimal Discriminator

For fixed generator $G$, the optimal discriminator is:

$$D^*(y) = \frac{P_{data}(y)}{P_{data}(y) + P_F(y)}.$$

Since, by the lemma, we can rewrite:

$$\begin{aligned} V(D,G) &= \int_y P_{data}(y) \log D(y) + \int_z P_z(z) \log(1 - D(F(G(z))))dz \\ &= \int_y P_{data}(y) \log D(y) + \int_x P_x(x) \log(1 - D(F(x)))dx \\ &= \int_y P_{data}(y) \log D(y) + P_F(y) \log(1 - D(y))dy \\ &= \mathbb{E}_{y \sim P_{data}}[\log D(y)] + \mathbb{E}_{y \sim P_F}[\log(1 - D(y))]. \end{aligned}$$

Following the same method in the Goodfellow paper, the result of the proposition follows identically.

# Optimal Generator

With the optimal discriminator $D^*$,
the optimal generator is achieved if and only if

$$P_F = P_{data}$$

.

Questions?