Scaling the SciDAC QuantOm Workflow

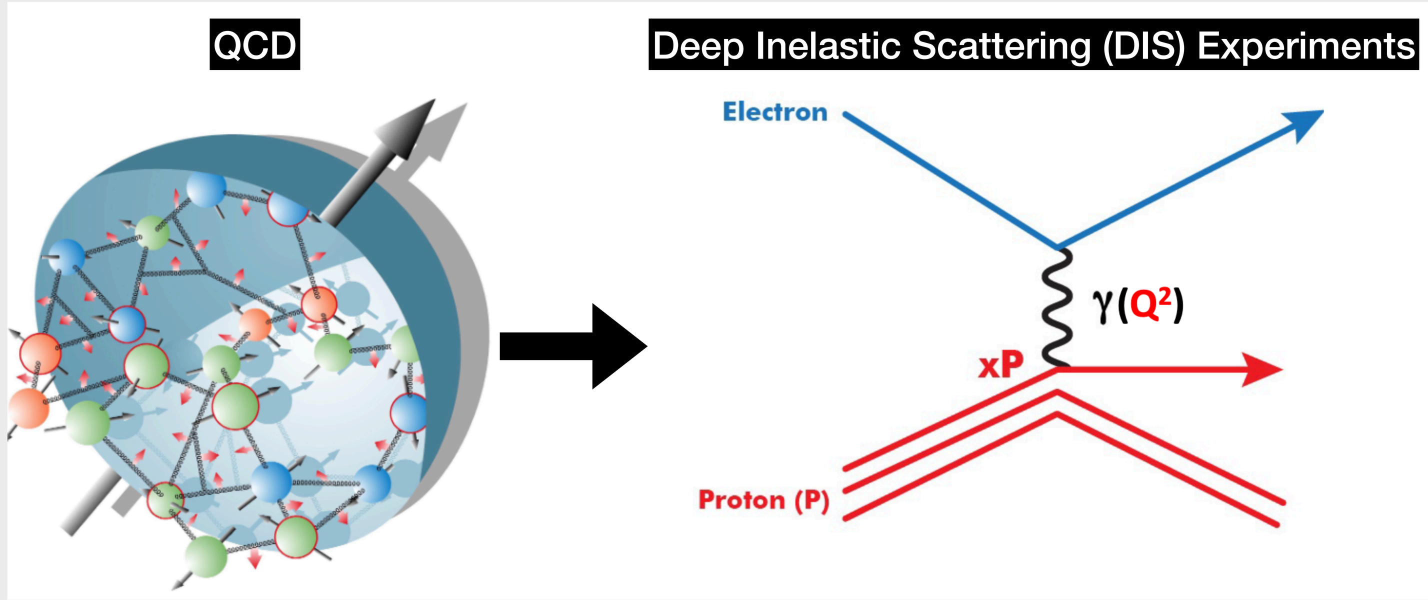
Daniel Lersch[1], Malachi Schram[1], Kishan Rajput[1], and Zhenyu Dai[1]

[1]Data Science Department, Jefferson Lab

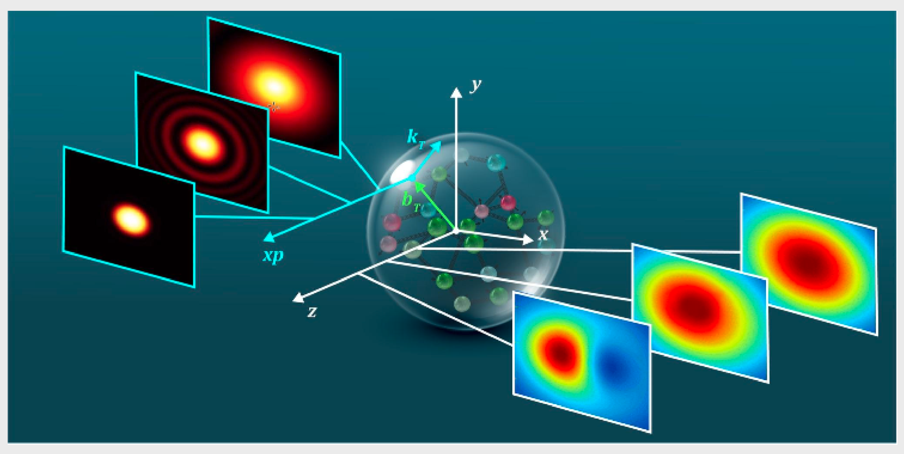
Jefferson Lab Exploring the Nature of Matter

QuantOm Collaboration

Quantum Nuclear Tomography - QuantOm



- Part of the **Sci**entific **D**iscovery through **A**dvanced **C**omputing (**SciDAC**) program
- **Goal:** Understand quark-gluon system by analyzing data from deep inelastic scattering experiments
  - Extract Quantum Correlation Functions (QCFs) from experimental data
  - 3D imaging of the proton

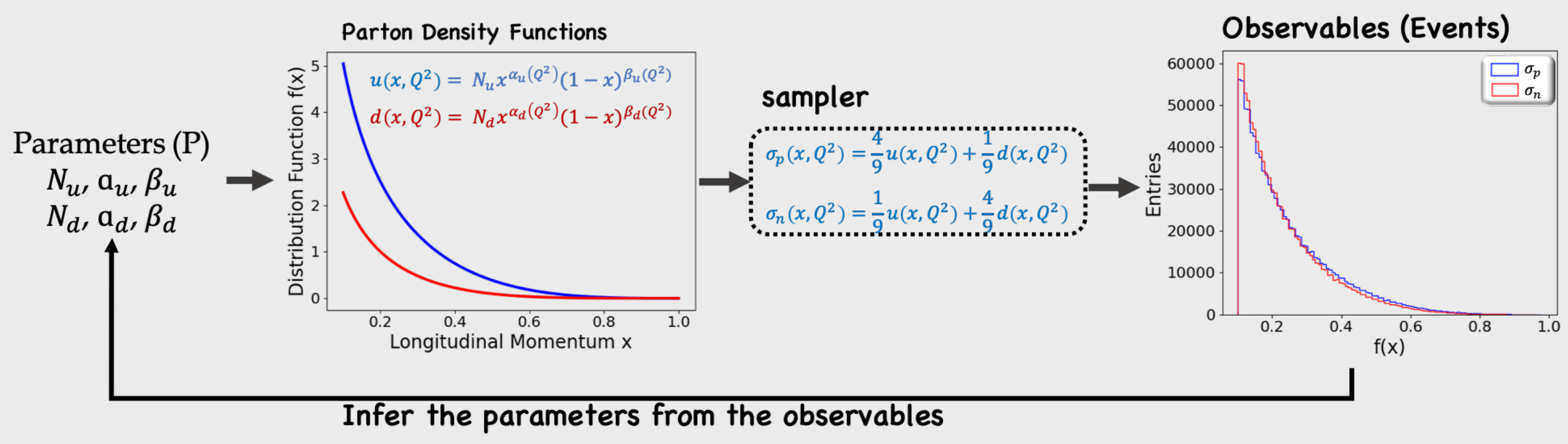


- **Problem:** QCFs are not directly accessible in experiment
- **Approach:** Solve inverse problem on an event level

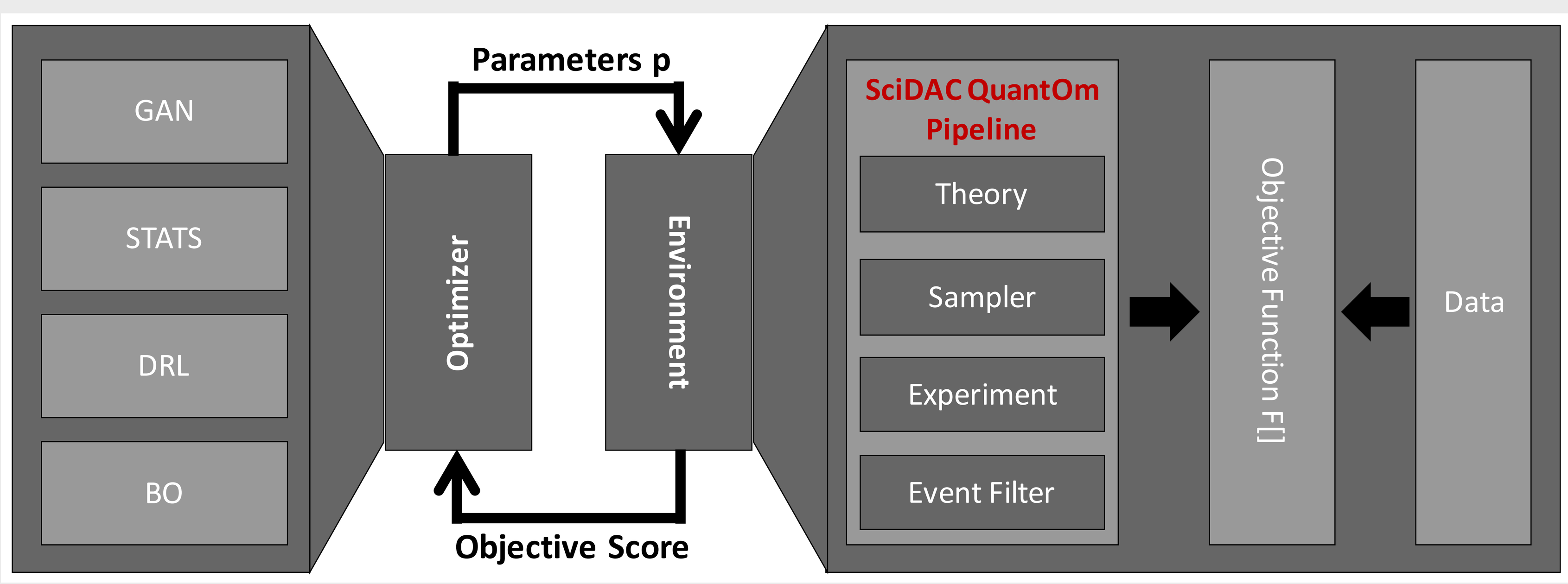
Inverse Problems and 1D Proxy App

- Well known forward function: $y = f(x)$
  - Observable y is measured in experiments
  - We are interested in $x \leftrightarrow$ QCFs
  - Parametric form provided by theory: $x \approx \hat{x}(p)$
- Solve inverse problem by minimizing: $F[y, f(\hat{x}(p))]$

1D Proxy App for workflow validation



The Generative Inverse Problem Solver - GIPS



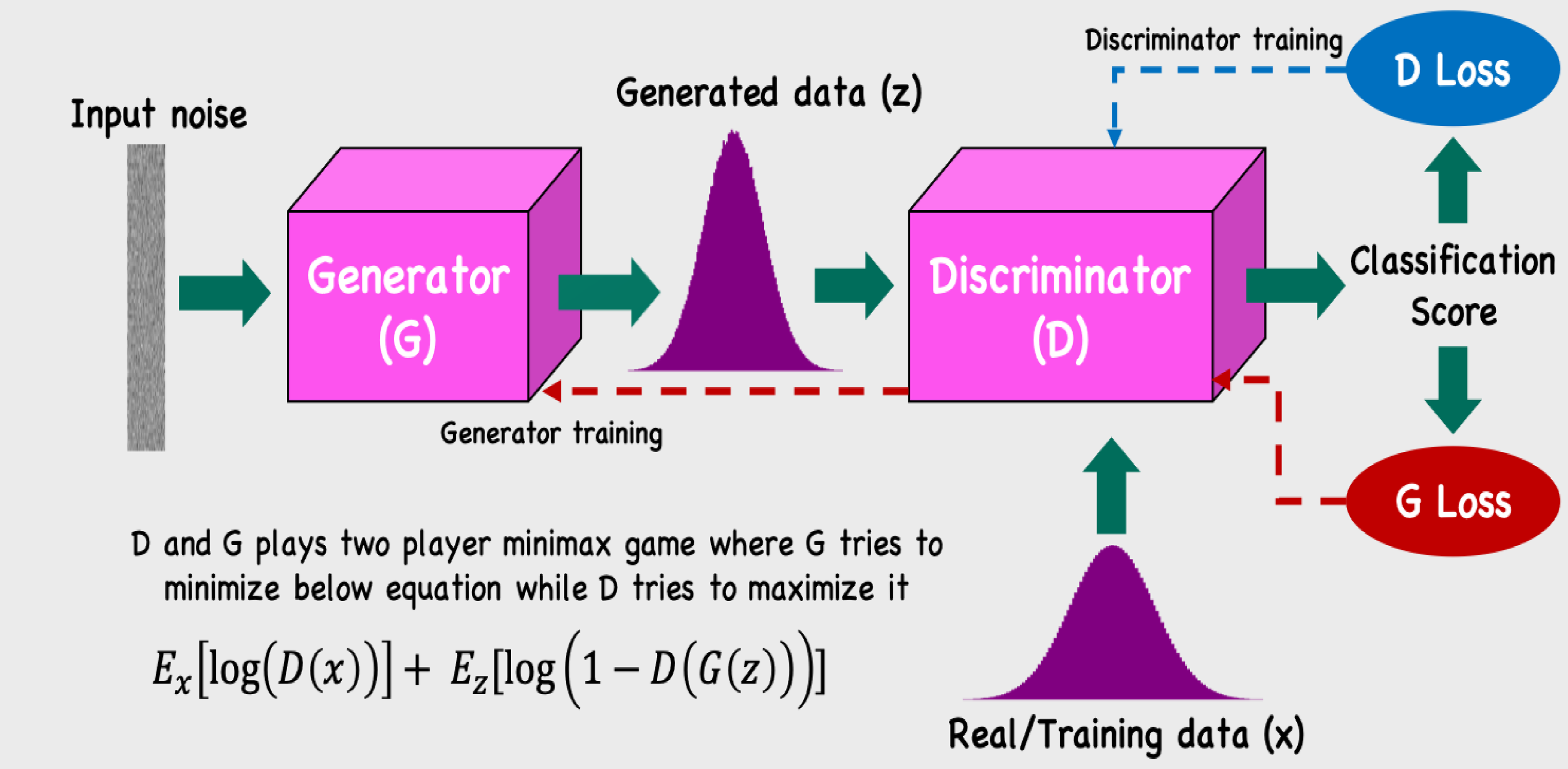
Modular workflow for event-level analysis

- **Optimizer:** Predicts Parameters
- **Environment:** Returns objective score to update optimizers internal states
  - **Pipeline:** Translates parameters to physics events
  - **Objective Function:** Compare synthetic data to data from experiment
- Handle data volume ⇒ **Run workflow on HPC systems**

The GAN Optimizer

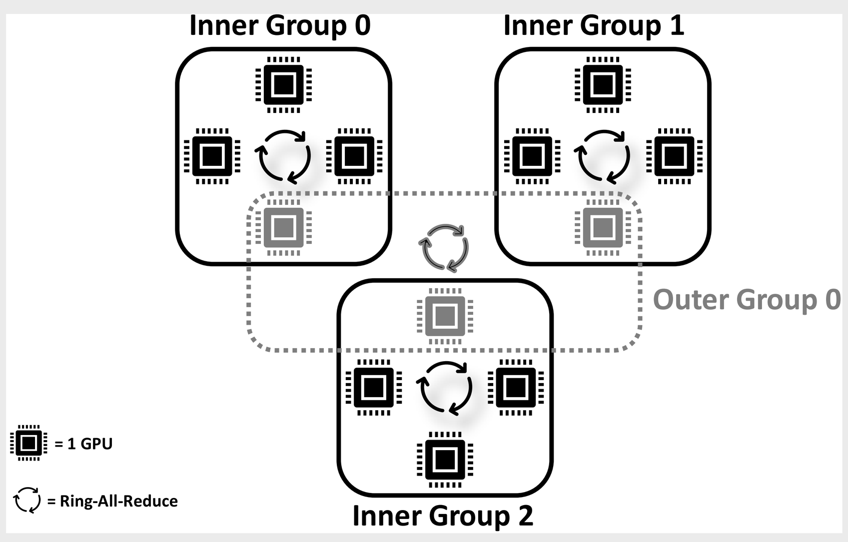
Generative Adversarial Networks (GAN)

- GANs are widely used generative models in computer vision
- GAN uses two ML models namely a Generator (G) and a Discriminator (D) that compete with each other
- Backpropagation is used to train the deep learning models (both G and D)

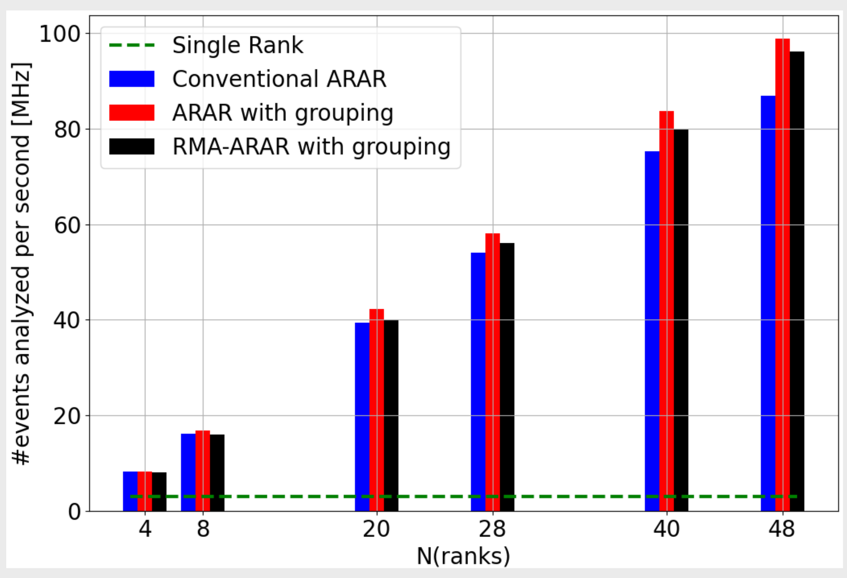


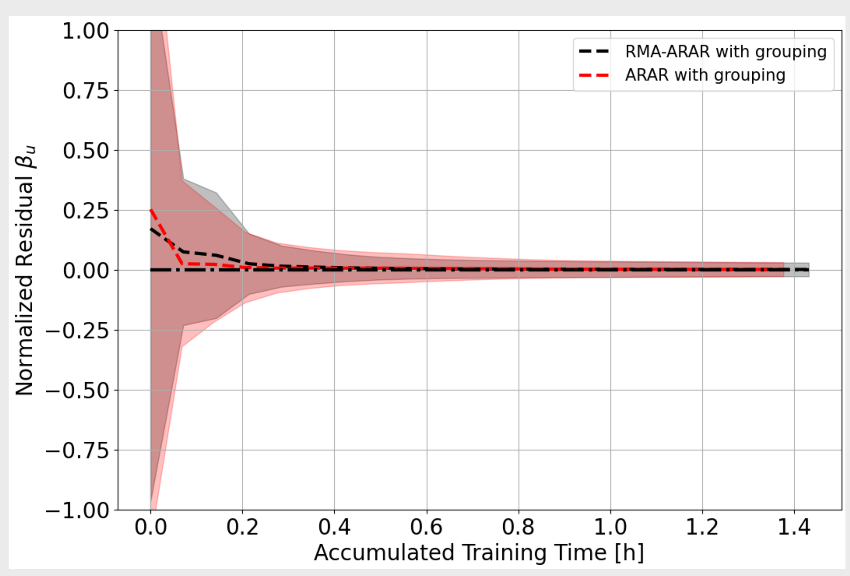
Asynchronous Data Parallel Training of GAN Workflow

- **Challenges**
  - Computation cost and stochastic nature of sampler module
  - GAN: Two networks need to be trained synchronously
- **Approach:**
  - Each rank analyzes portion of data (with overlap)
  - Accumulate gradients via **A**synchronous **R**ing-**A**ll-**R**educe (**ARAR**)
  - Utilize grouping to optimize communication among ranks
  - Test **R**emote **M**emory **A**ccess (**RMA**) for transferring gradients between ranks
  - Multiple modes for gradient transfer available



- **Scaling tests with 1D proxy app on Polaris @ Argonne**
  - 50% of toy data are analyzed by each rank
  - Run ensemble analysis with 20 GANs and 20 GPUs per GAN





Acknowledgements

This work is supported by the Scientific Discovery through Advanced Computing (SciDAC) program via the Office of Nuclear Physics and Office of Advanced Scientific Computing Research in the Office of Science at the U.S. Department of Energy under contracts DE-AC02-06CH11357, DE-AC05-06OR23177, and DE-SC0023472, in collaboration with Argonne National Laboratory, Jefferson Lab, National Renewable Energy Laboratory, Old Dominion University, Ohio State University, and Virginia Tech.