AI-driven detector design for the EIC

PI: Miguel Arratia

Co PIs: Aaron Angerami, Ron Soltz, Benjamin Nachman, Kenneth Barish

UC RIVERSIDE BERKELEY LAB Lawrence Livermore National Laboratory

High-level description of what we want to do:

Detector Model

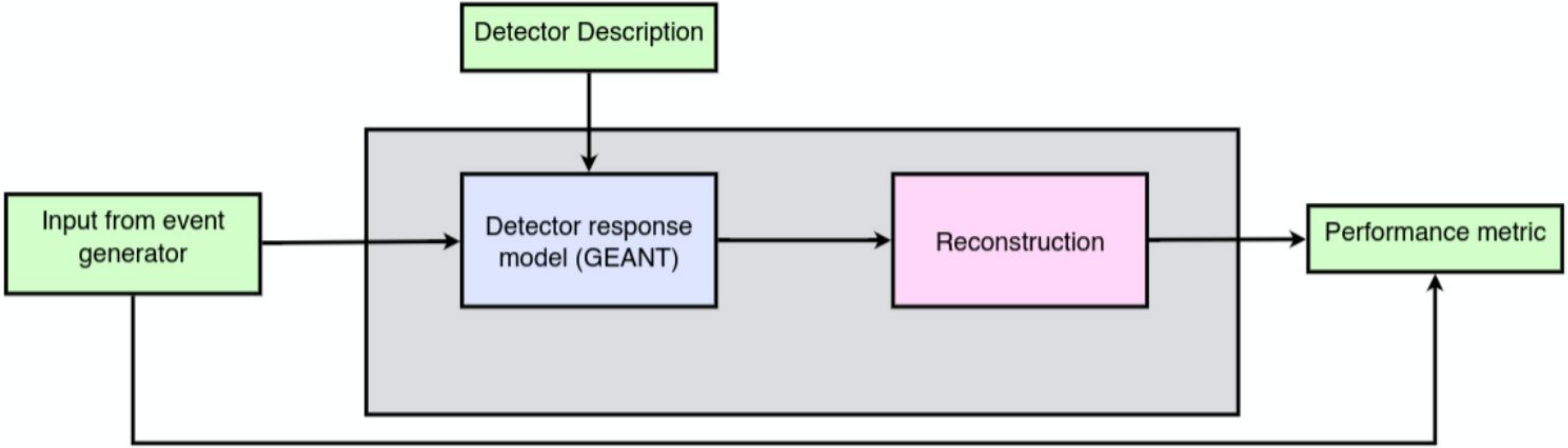
parameters of interest θ

Goal: find best θ given a metric(s).

Challenge: detector output is high-dimensional and θ may be high-dimensional.

Today: can we use ML to (1) interpolate in the high-dimensional space, (2) define optimal metrics, and (3) find the best values of θ.

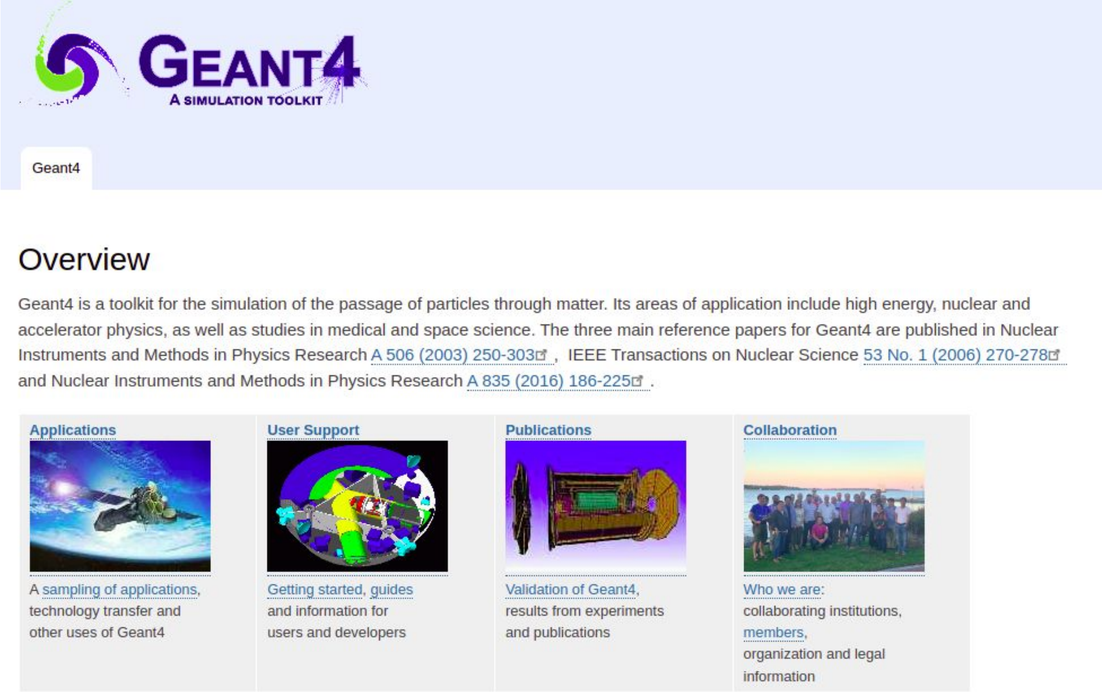
Traditional approach to evaluate detector designs



- Small number of detector setups are considered, and the reconstruction algorithm is not optimized for a given detector setup.
- **Not amenable to gradient-based optimization**

The backbone of detector design

“Geant4 is a toolkit for the simulation of the passage of particles through matter”


GEANT4
A SIMULATION TOOLKIT

Geant4

Overview

Geant4 is a toolkit for the simulation of the passage of particles through matter. Its areas of application include high energy, nuclear and accelerator physics, as well as studies in medical and space science. The three main reference papers for Geant4 are published in Nuclear Instruments and Methods in Physics Research A 506 (2003) 250-303 , IEEE Transactions on Nuclear Science 53 No. 1 (2006) 270-278 and Nuclear Instruments and Methods in Physics Research A 835 (2016) 186-225 .

Applications
A sampling of applications, technology transfer and other uses of Geant4

User Support
Getting started, guides and information for users and developers

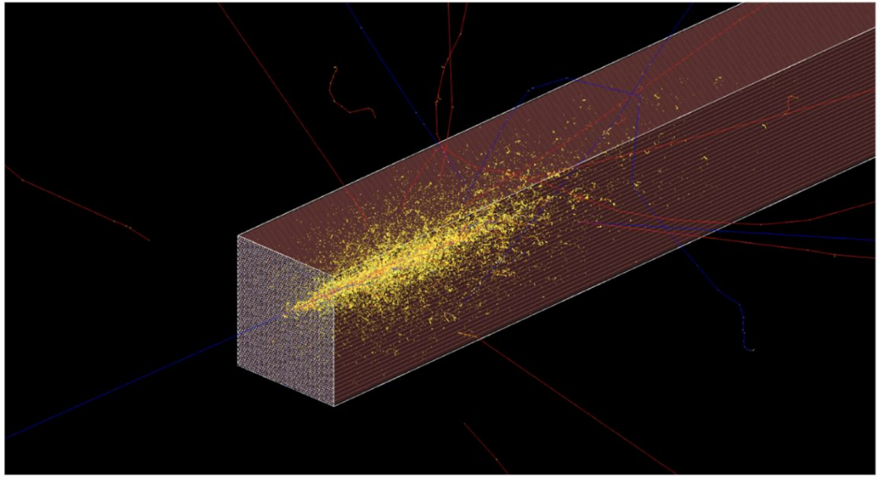
Publications
Validation of Geant4, results from experiments and publications

Collaboration
Who we are: collaborating institutions, members, organization and legal information

Used in many fields: particle and nuclear, space, medical physics, etc.

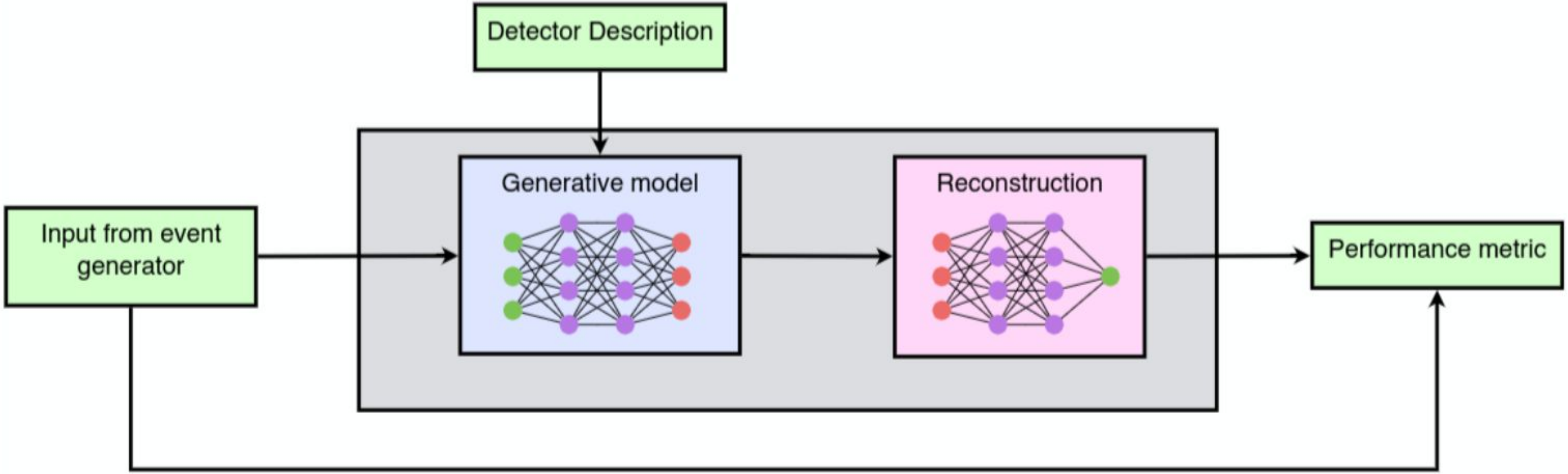
Issues with Geant4 simulations that create bottlenecks for automated detector design

- **Computationally expensive.** (bottleneck for collider experiments simulations, especially that of particle calorimeters)
- **Not amenable to automated optimization** (uses a stochastic model of Detector response)



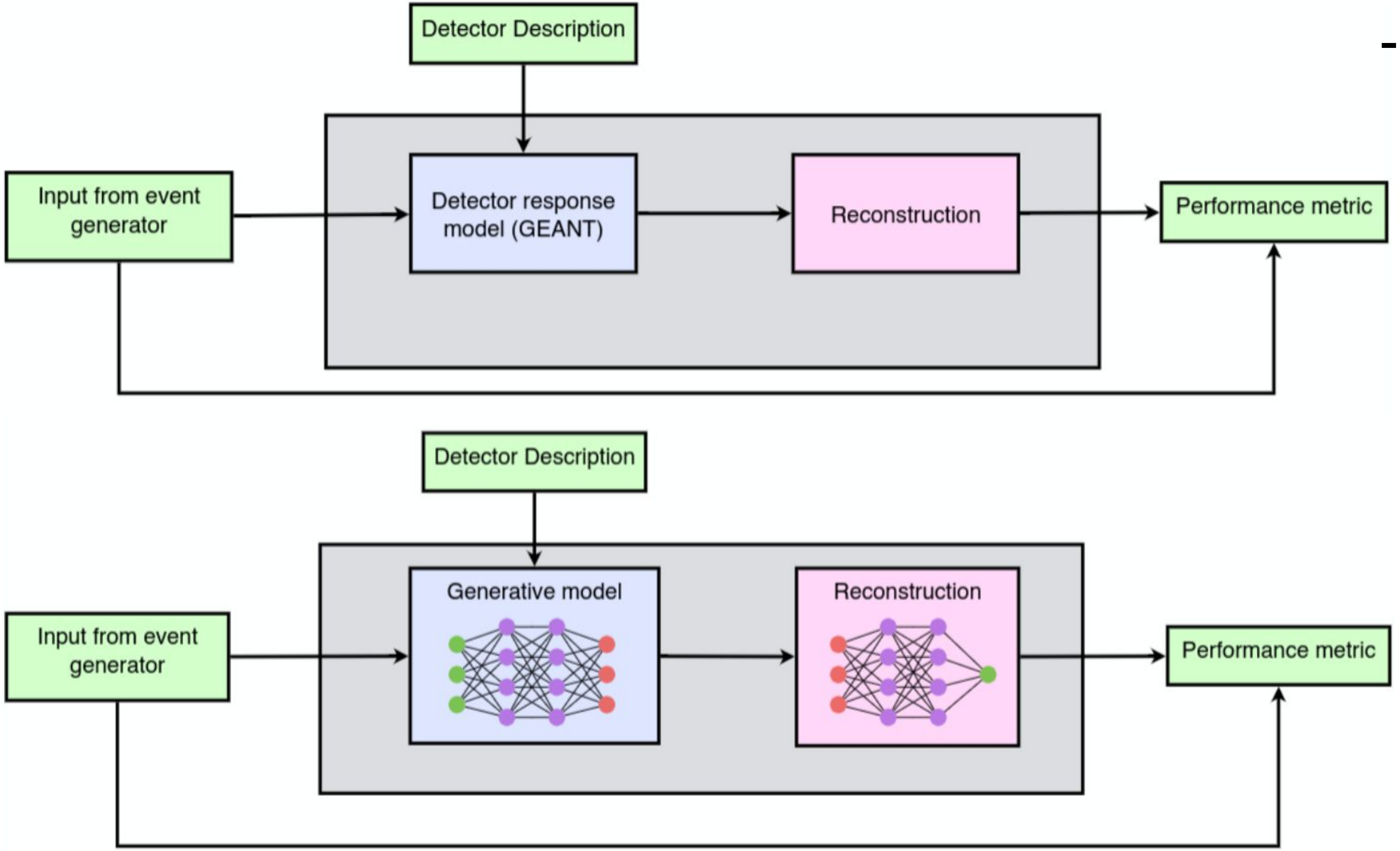
Example of simulation of a particle shower developing in a calorimeter. Source: Geant4

Our DNN-based approach (Co-design)



Amenable to automated optimization
(e.g gradient-based optimization)

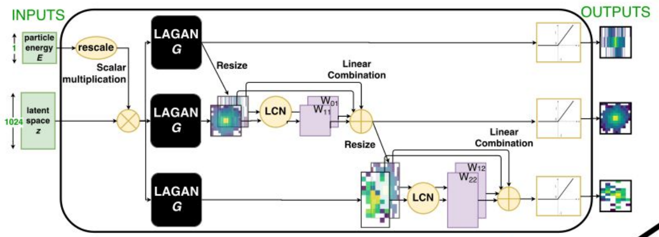
Traditional vs DNN-based approach



- We need to swap Geant4 for DNN-surrogate, and reconstruction algorithm must be DNN-based, so the entire chain is **differentiable with respect to detector parameters**

Our proposed solution:

*"to use deep neural networks (DNNs) to transform Geant4 simulations into **differentiable models**, which will make **gradient-based optimization** possible"*



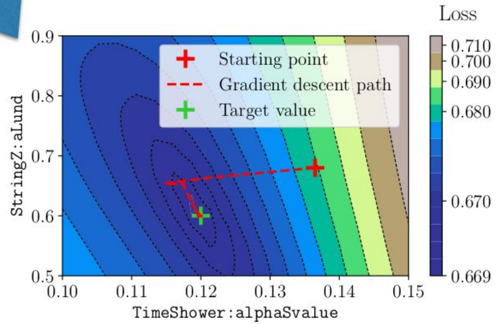
Detector Modeling

Surrogate Model
Differentiable Simulation

For free: GPU-enabled fast sim.

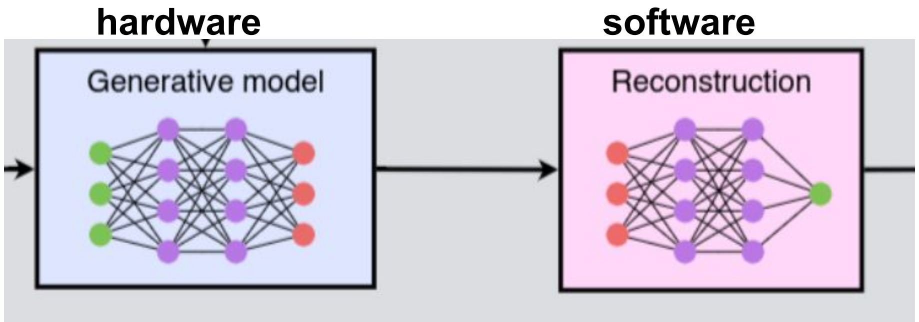
Gradient-based
Gradient-free

ML-based Optimization



Not just optimal solution but also model gradient!

“Co-design”



This block can be thought of as a single network

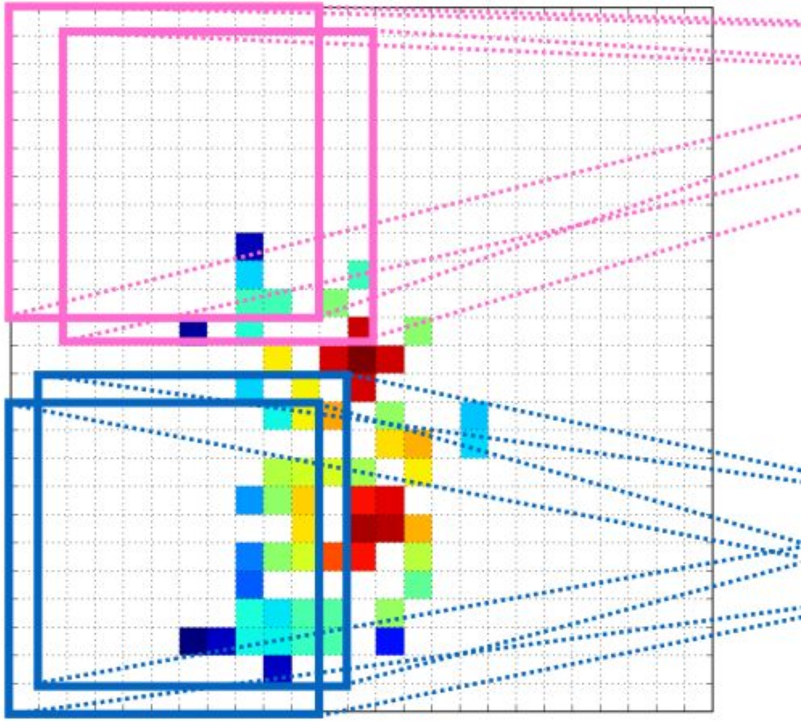
- Overall optimization will achieve “codesign” in the sense that the reconstruction algorithm will be **optimal for a given detector**

- Codesign is a common concept in many fields, but was never used in particle or nuclear physics before.

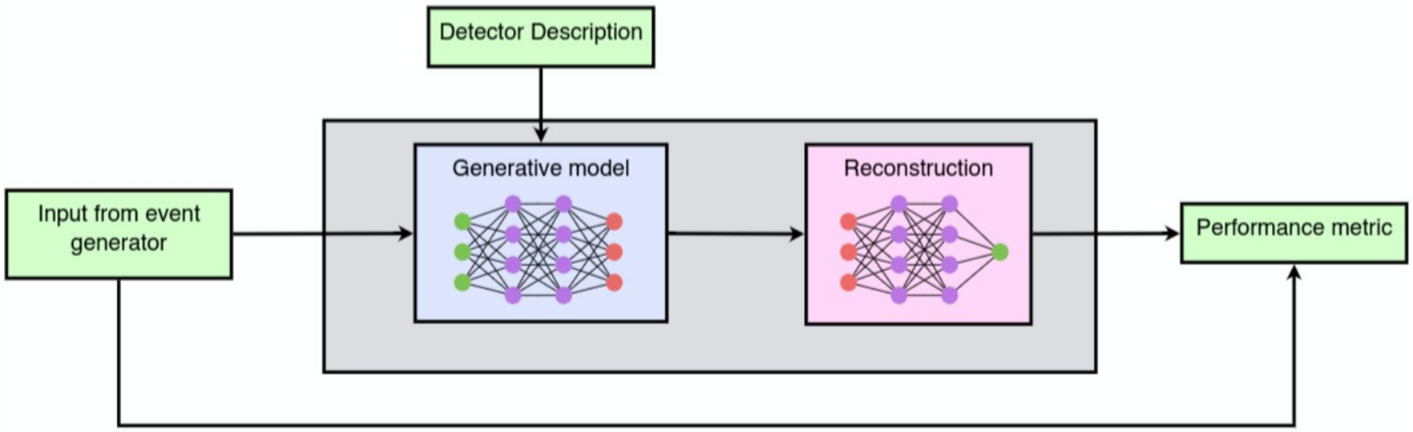
The advent of EIC detectors is a perfect opportunity to pursue this.

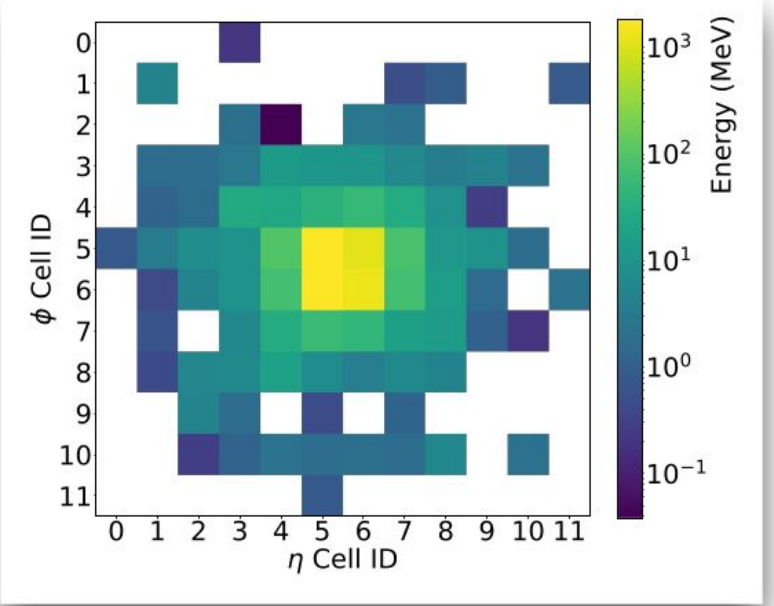
AI/ML can do more than improve data analysis!

We can use these tools to optimize our detectors - a qualitatively new application of ML!



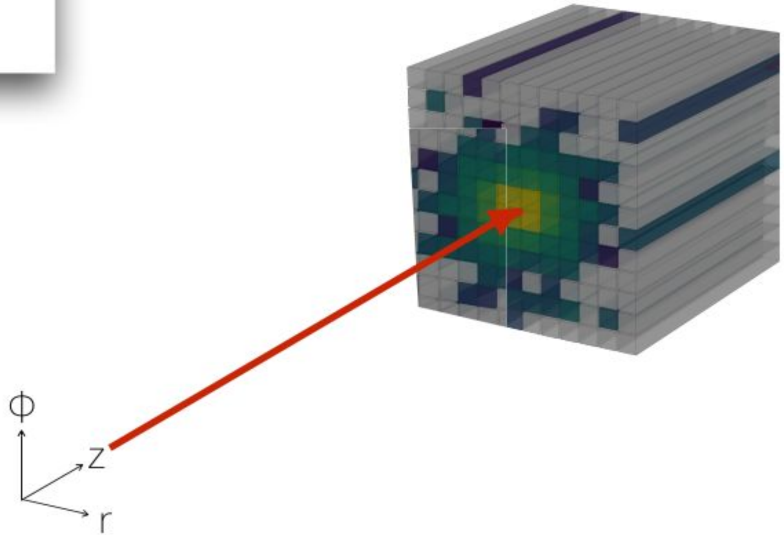
How do we achieve this?

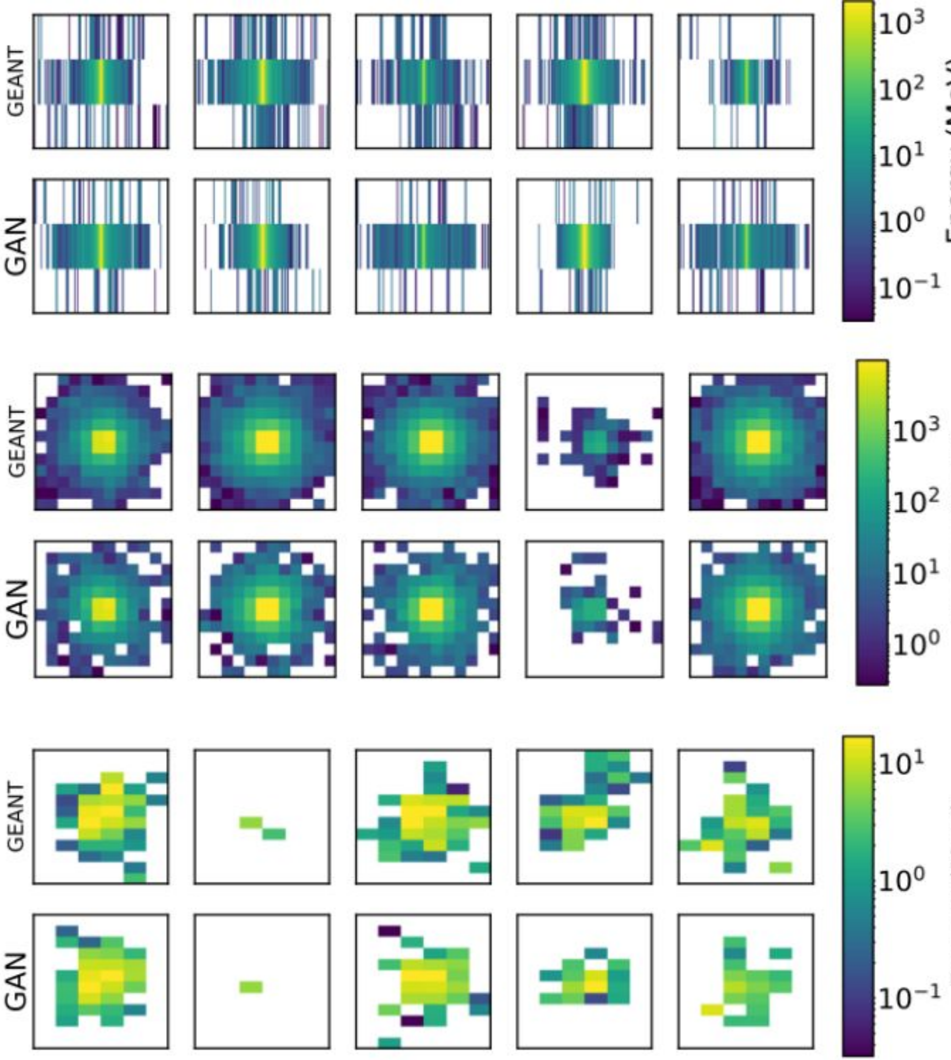




Can we train a neural network to emulate the detector simulation?

Grayscale images: Pixel intensity = energy deposited



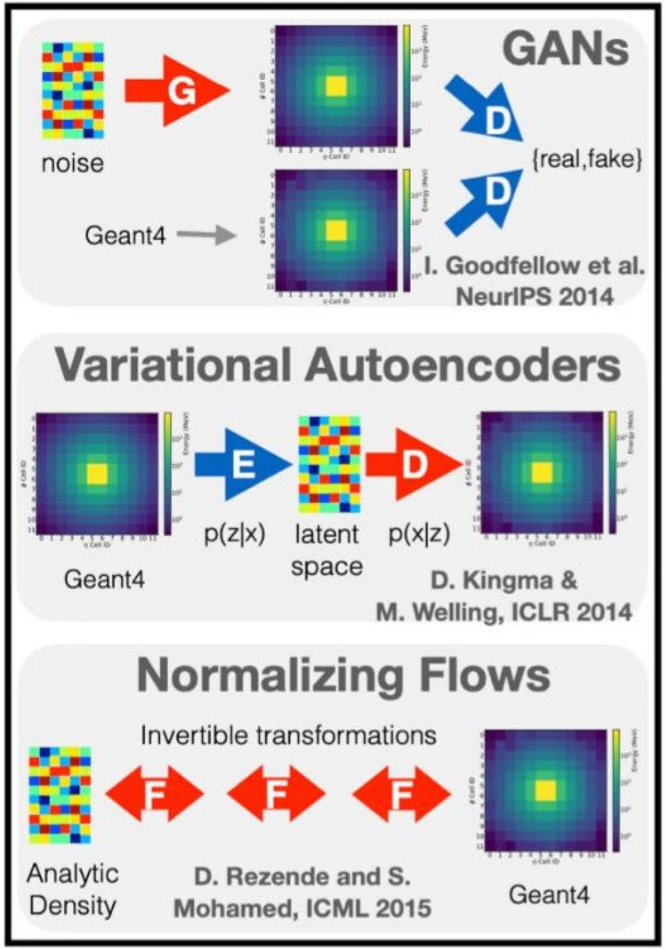


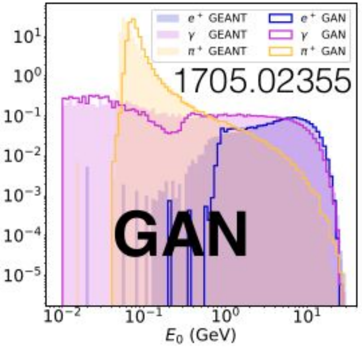
Five randomly selected photon showers in two calorimeter
layers from Geant4 (top rows) and their five nearest neighbors from a set of CaloGAN candidates.

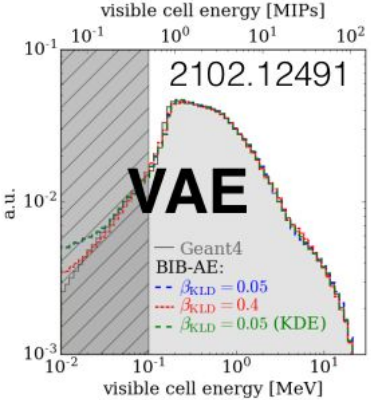
Figure source: M. Paganini, L. Oliveira, B.Nachman, Phys. Rev. Lett. 120 (2018) 4, 042003.

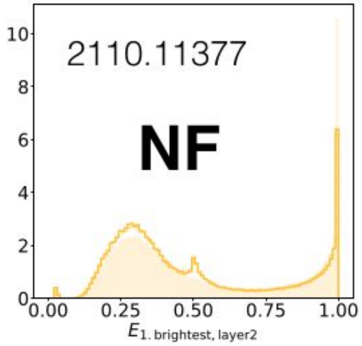
Although CaloGAN is not the only approach possible, nowadays there are various competing approaches:

Learn a differentiable **surrogate**

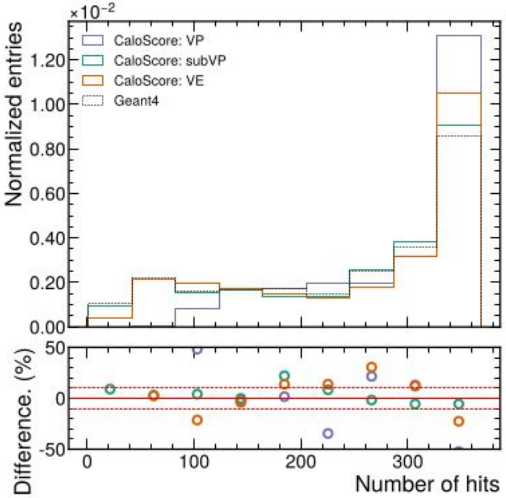


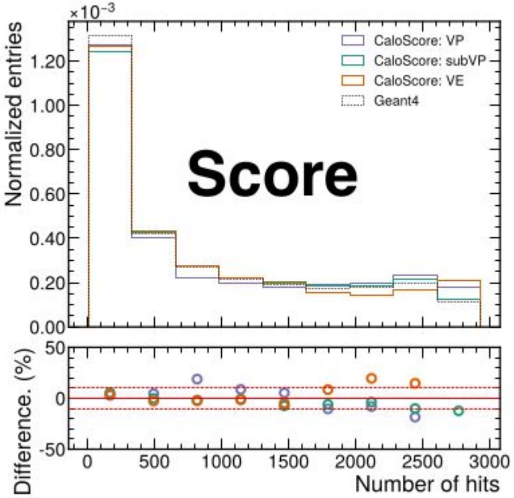


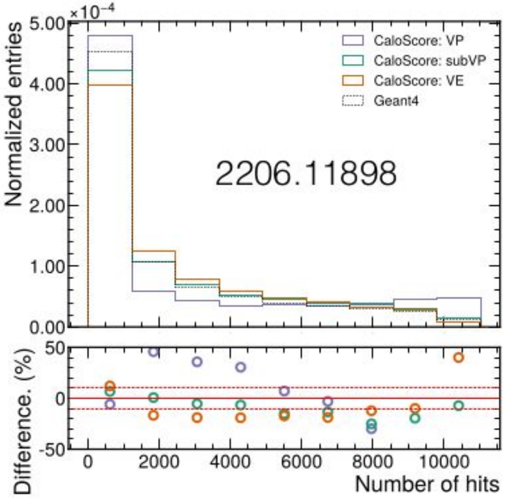




Many papers on this subject - see the <u>living review</u> for all

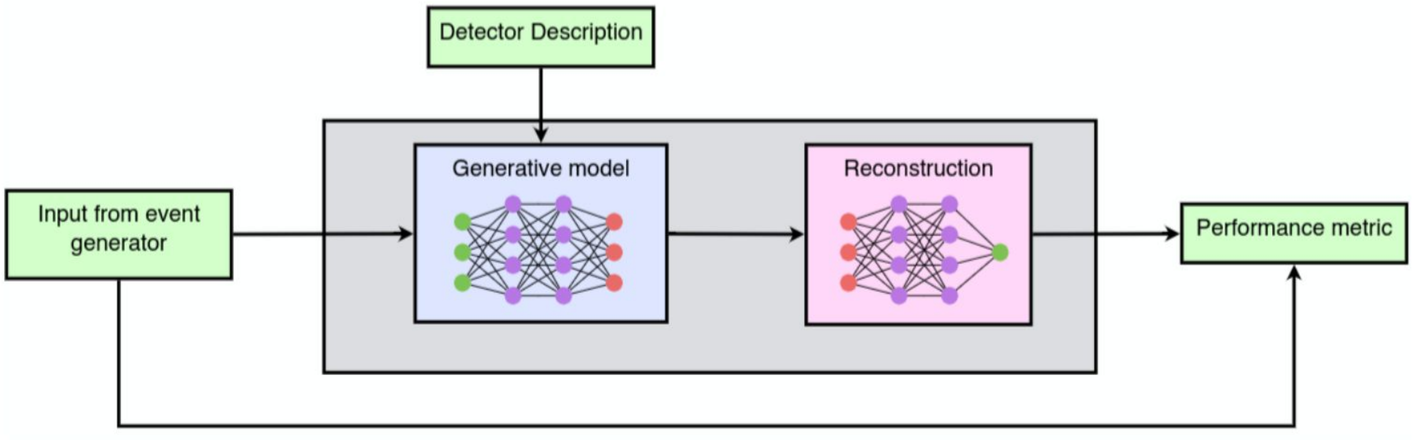






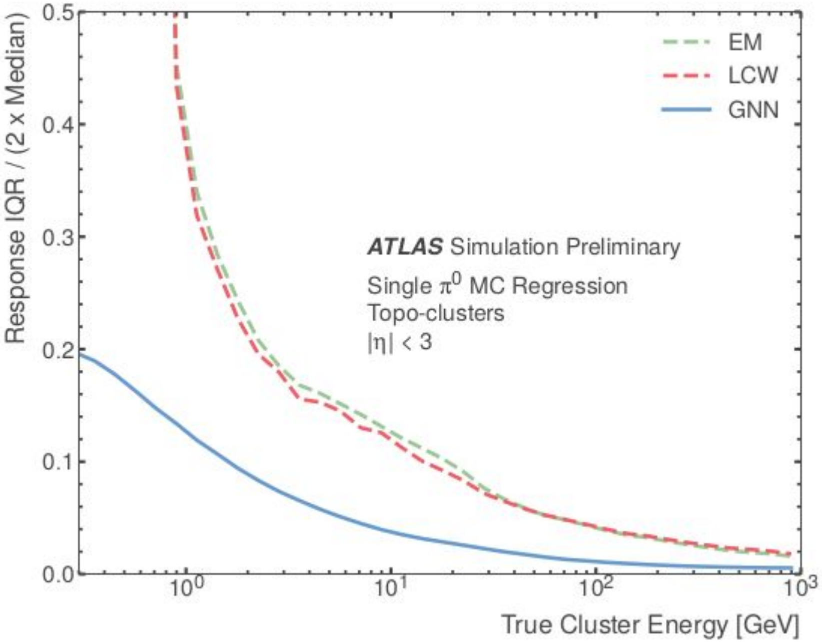
See also https://calochallenge.github.io/homepage/

How do we achieve this?



Reconstruction based on DNN

If doing gradient-based optimization, the target also needs to be differentiable. For example, target could be resolution of some reconstructed object. This could itself be a neural network!

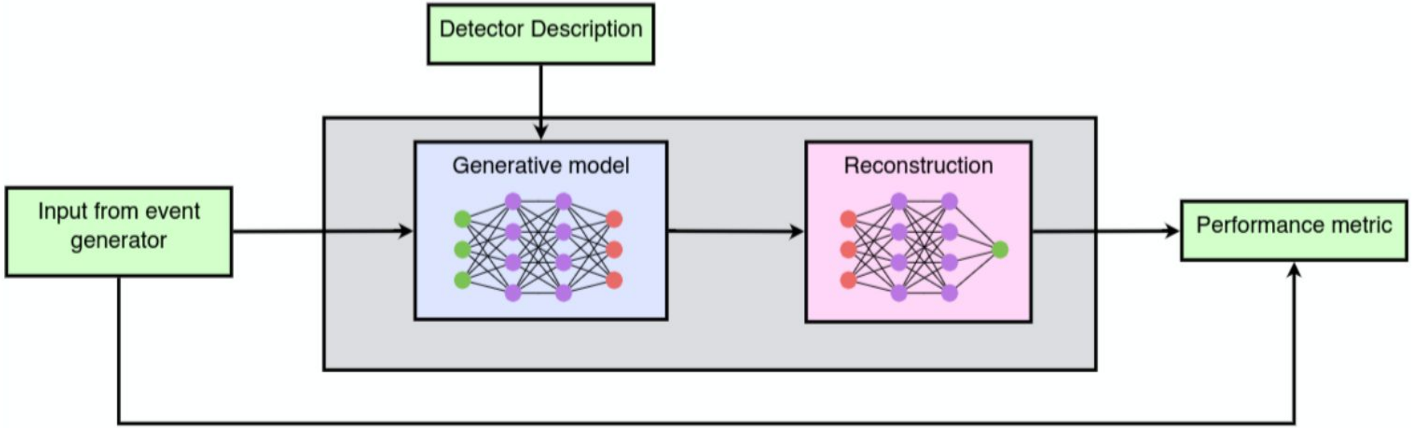


State of the art work, led by members of our team (Angerami, Nachman)

ATLAS PUB Note

ATL-PHYS-PUB-2022-040

Innnovation



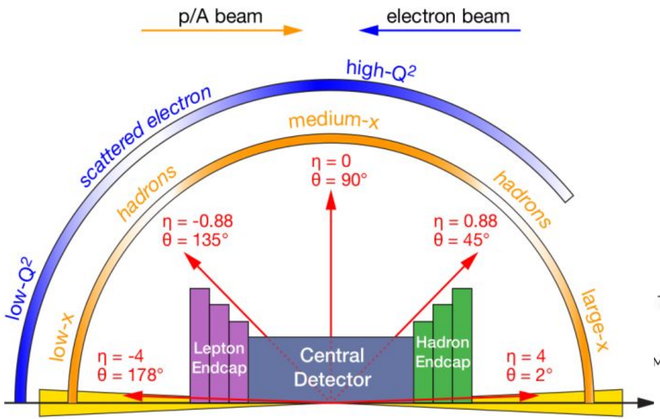
- Generative models through surrogate is well established (although not conditional on detector parameters)
- Reconstruction with DNN is well established

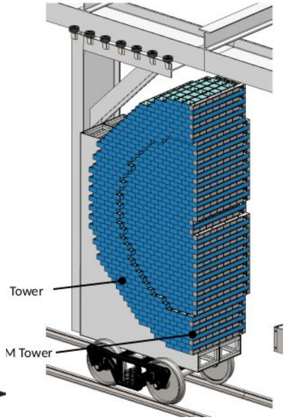
But nobody had merged the two and used it in an optimization setting

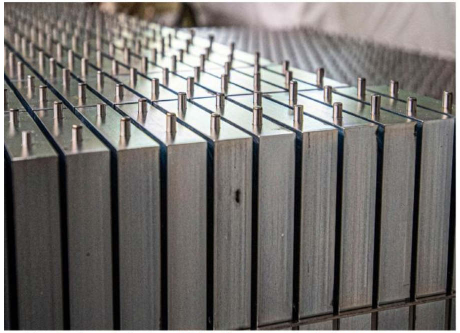
Case study: Calorimeters for Electron-Ion Collider

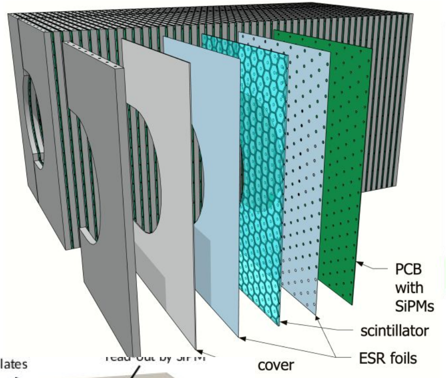
Chosen because :

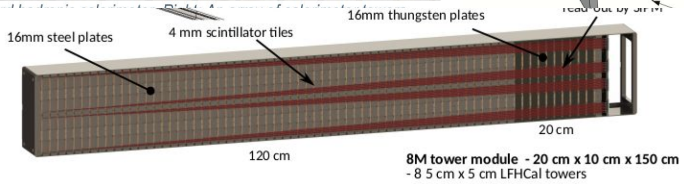
- **Maximal impact** (calorimeter is the bottleneck for Geant4 simulations)
- **Clear need for optimization in high-dimensional space** (granularity to be defined),
- **Domain knowledge within team** (active involvement of UCR in EIC R&D, previous work by LLNL & LBNL team in using AI techniques for ATLAS heavy ions)







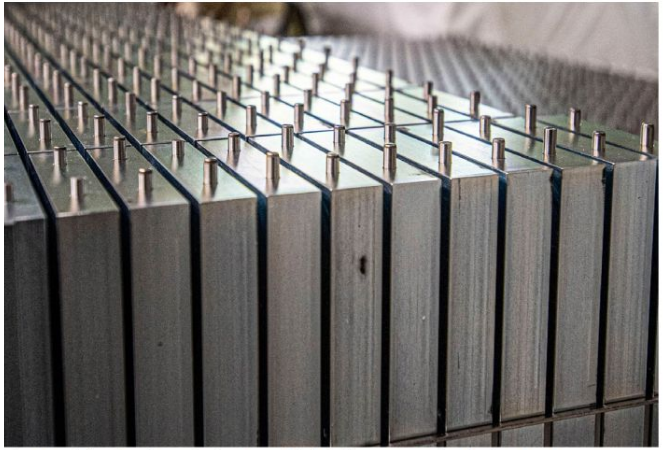




arXiv:2207.09437

Some of the key questions that our AI-driven optimization approach could answer are

- Given a certain budget, what is the optimal performance one can get and how it depends on number of readout layers?
- For which angles would a high-segmentation have the largest impact?
- Where should be the longitudinal layers be placed?



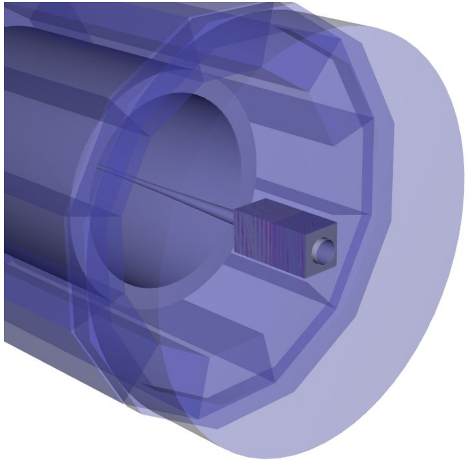
In addition to an optimal detector configuration, **our approach also provides the model gradient**, which will gives us quantitative insights for decision making

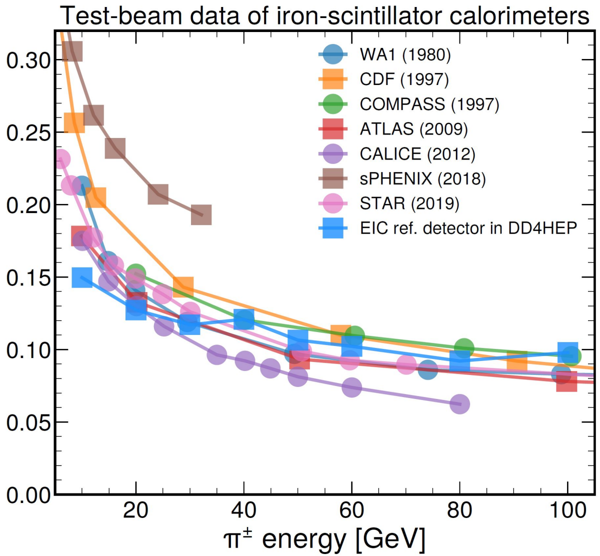
Timetable of Activities

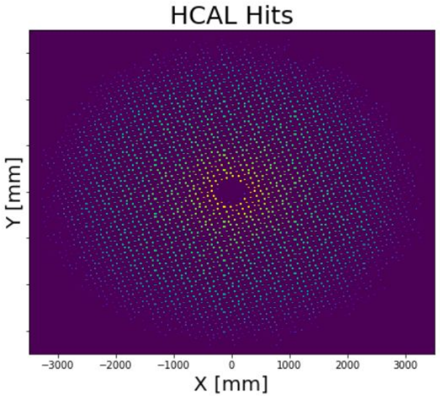
Tasks and Deliverables	FY22				FY23			
	1	2	3	4	1	2	3	4
T1: Implement G4 model of realistic calorimeter and develop training framework								
T2: Setup full learning pipeline using simplified setup								
D1: Write methods paper								
T3: Apply pipeline to EIC G4 simulation								
T4: Complete full detector/reconstruction codesign								
D2: Write optimization paper								
D3: Deliver DNN fast sim and event reconstruction tools to EIC community								

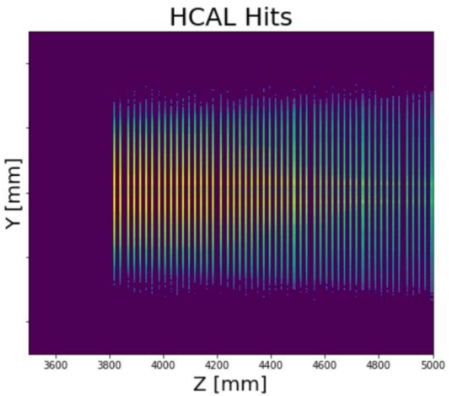
Pipeline building

- Geant4 models for simulations implemented with full detail.
- Data generation @ LLNL QA code, etc.
- Simulations validated against real data using “strawman” (non-AI) methods to avoid GIGO in our AI approaches.





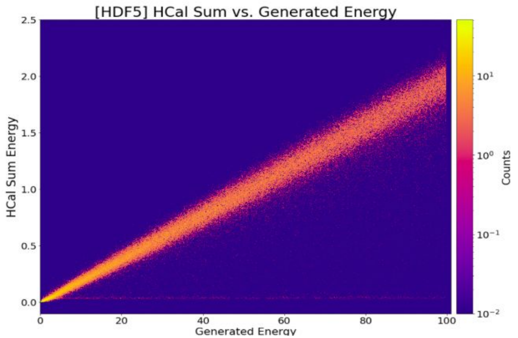


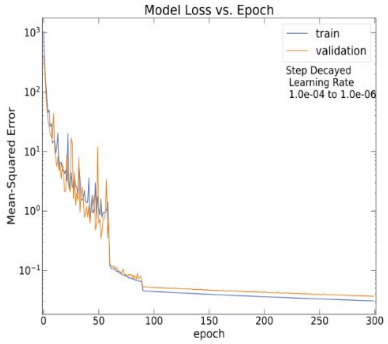


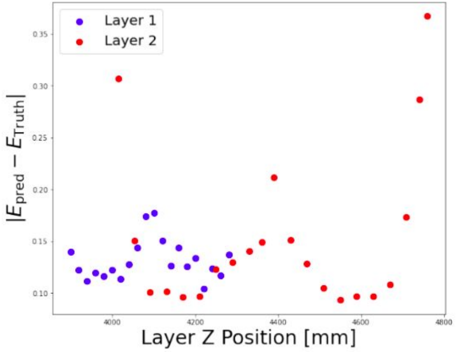
Pipeline building, continuation

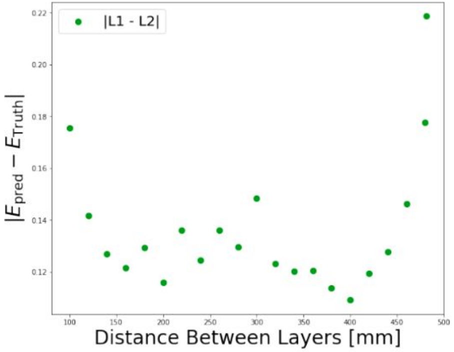
- Large-scale generation of Geant4 simulation for AI models trained achieved using HPC clusters at LLNL

- Training of AI models for reconstruction using HPC using distributing GPU is working.
  Have completed DNN-based workflow and are starting GNN one

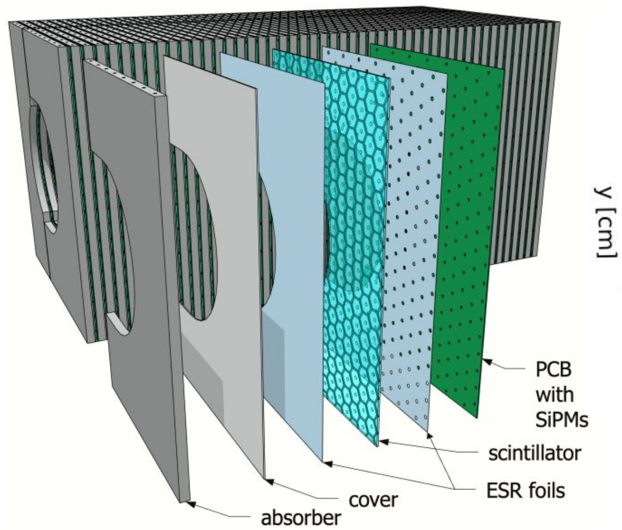




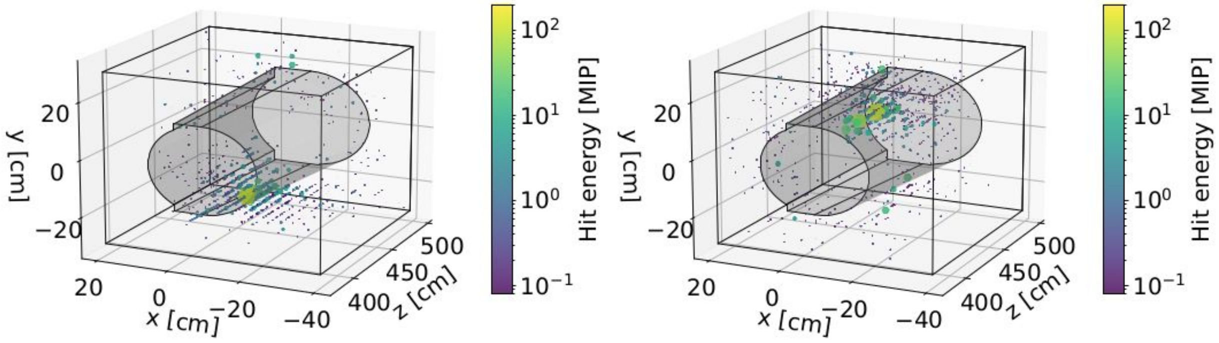




Our AI work motivated us to explore the potential of "imaging calorimetry" @EIC



Shower-shape examples. Highly-granular 5D hits



- Technology can yields 5D (energy, time, space) information → **Optimal for AI**
- Flexible technology, requires non-trivial optimization, and challenging reconstruction that **calls for AI-based methods and design.**
- Conceptual design submitted this year to NIMA, optimization paper with methods developed in this project to be published separately

Computer resources

We are using LLNL HPC resources such as *Borax* and *Lessen* at LLNL.

New collaboration between LLNL and UC Riverside.

A total of 5 students and 1 postdoc from UCR are involved and learnt how to use these for large-scale data generation as well as machine-learning model training.

Exchange between UC Riverside and National Labs

2 UCR graduate students traveled to Berkeley this summer.

Next summer, we plan the same for Livermore.

2 undergraduate students are being paid as research assistants (not just over summer)

Dual goal is also to expose students to broad applications of AI methods at LBNL and LLNL

Team & budget

Table 2. Key team members and their relevant expertise for this project.

Team Member	Expertise	Contribution to Project
Ben Nachman (LBNL)	ML/jets	0.10
Aaron Angerami (LLNL)	ML/jets	0.10
Miguel Arratia (UCR)	Jets/EIC	0.10
Ken Barish (UCR)	Cold QCD/Calorimetry/EIC	0.10
Computer Scientist (LBNL)	Data/Computer Science	0.15
Piyush Karande (LLNL)	Data/Computer Science	0.15
Postdoc (LBNL)	ML/jets	0.50
Postdoc (LLNL)	ML/jets	0.50
Postdoc (UCR)	ML/jets	0.50
Graduate Student (UCR)	Training for ML	0.50
Graduate Student (UCR)	Training for ML	0.50

	FY21 ($k)	FY22 ($k)	Totals ($k)
a) Funds allocated	490	490	980
b) Actual costs to date	325		

Postdocs: Fernando Torales-Acosta (LBNL), Dongwi Handiipondola Dongwi (LLNL), Vishnu Karki (UCR);
Grad students: Sebastian Moran (UCR) , Liam Blanchard (UCR)

Optimizing detector design with DNNs

Overarching goal: provide first-ever detector design optimized with DNNs

Key deliverables

- A framework applicable for any future experiment that rely on Geant4 simulations (the backbone for detector designs in many fields)
- High-fidelity DNN-based fast simulator for EIC
- DNN-based reconstruction software for EIC

Summary

- **Overarching goal: provide first-ever detector design optimized with DNNs**

  Building on recent advances in AI/ML on generative models and reconstruction
  which have been pioneered by members of our team.

- **New collaboration between**

  AI/ML experts and EIC detector experts
  Between LLNL, LBNL and UC Riverside

Project aims to influence EIC detector design and cement use of AI/ML methods on EIC at an early stage.