CERN

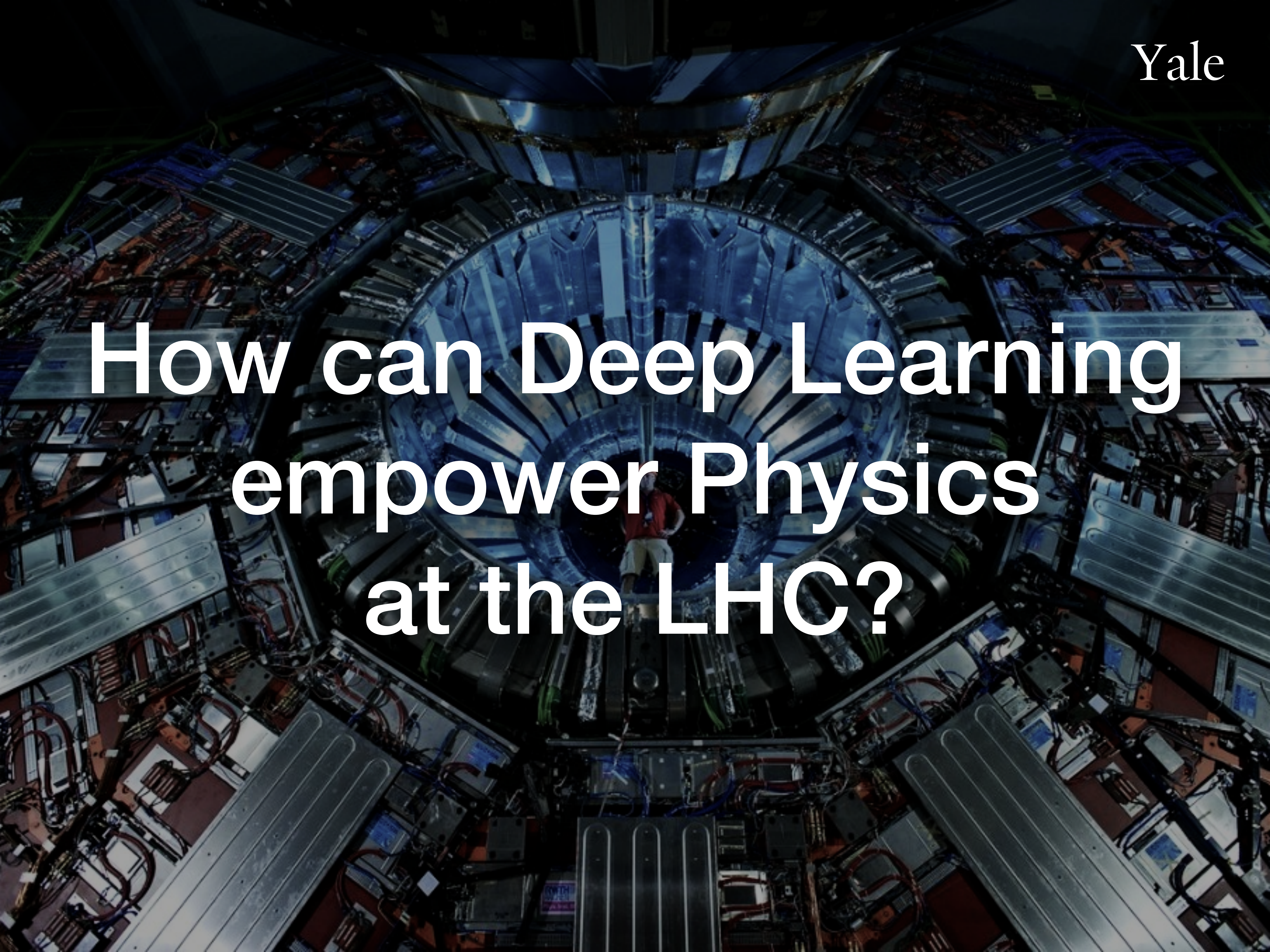
Accelerating Science

with Deep Learning

Michela Paganini
Yale
mickypaganini

Yale

How can Deep Learning empower Physics at the LHC?



So far…

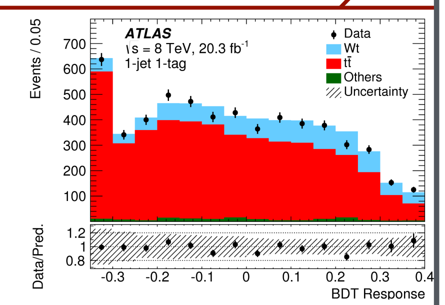
Yale

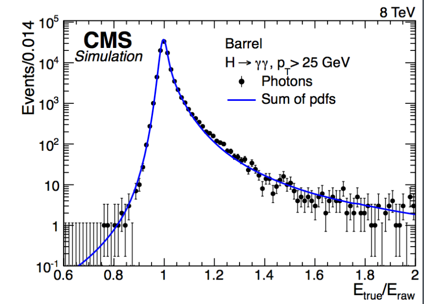
- Lots of convenient, but minor improvements across all aspects of the experiments

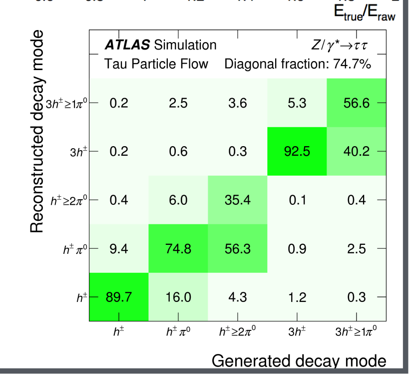
How Does Machine Learning Fit In?

15

- **In analysis**:
  - Classifying signal from background, especially in complex final states
  - Reconstructing heavy particles and improving the energy / mass resolution
- **In reconstruction**:
  - Improving detector level inputs to reconstruction
  - Particle identification tasks
  - Energy / direction calibration
- **In the trigger**:
  - Quickly identifying complex final states
- **In computing**:
  - Estimating dataset popularity, and determining how number and location of dataset replicas







from M. Kagan

- But nothing has really stood out as a major game-changer

until…

CaloGAN:
Simulating 3D High Energy Particle Showers in Multi-Layer EM Calorimeters with Generative Adversarial Networks

Michela Paganini
with Luke de Oliveira, Ben Nachman

BERKELEY LAB

Simulation

Yale

THEORY

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \not{D} \psi + h.c. + \bar{\psi}_i y_{ij} \psi_j \phi + h.c. + |D_\mu \phi|^2 - V(\phi)$$

HARD INTERACTIONS (ME CALCULATIONS)

MADGRAPH 5

PARTON SHOWERING & HADRONIZATION

DETECTOR SIM. & MATERIAL INTERACTIONS

Geant 4

DIGITIZATION

…

Simulation

Yale

THEORY

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\bar{\Psi} \not{D} \psi + h.c. + \bar{\Psi}_i y_{ij} \psi_j \phi + h.c. + |D_\mu \phi|^2 - V(\phi)$$

HARD INTERACTIONS (ME CALCULATIONS)

PARTON SHOWERING & HADRONIZATION

DETECTOR SIM. & MATERIAL INTERACTIONS

CaloGAN

Geant 4

DIGITIZATION

...

Motivation and Challenges

Yale

Full Simulation is slow

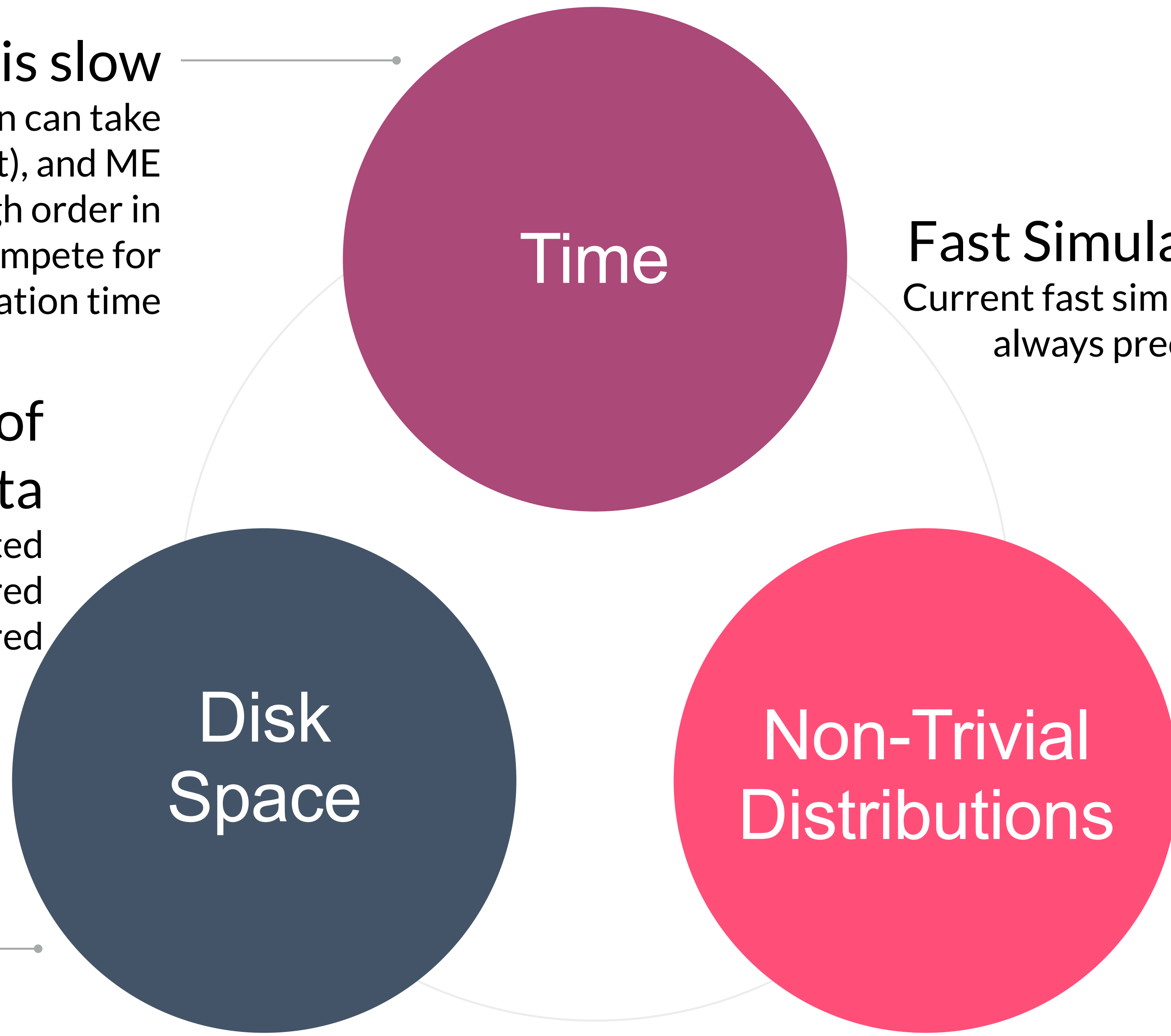
Detector simulation can take *O*(min/event), and ME calculations to high order in perturbation can compete for total generation time

Fast Simulation is inaccurate

Current fast simulation techniques are not always precise enough to describe all fluctuations correctly

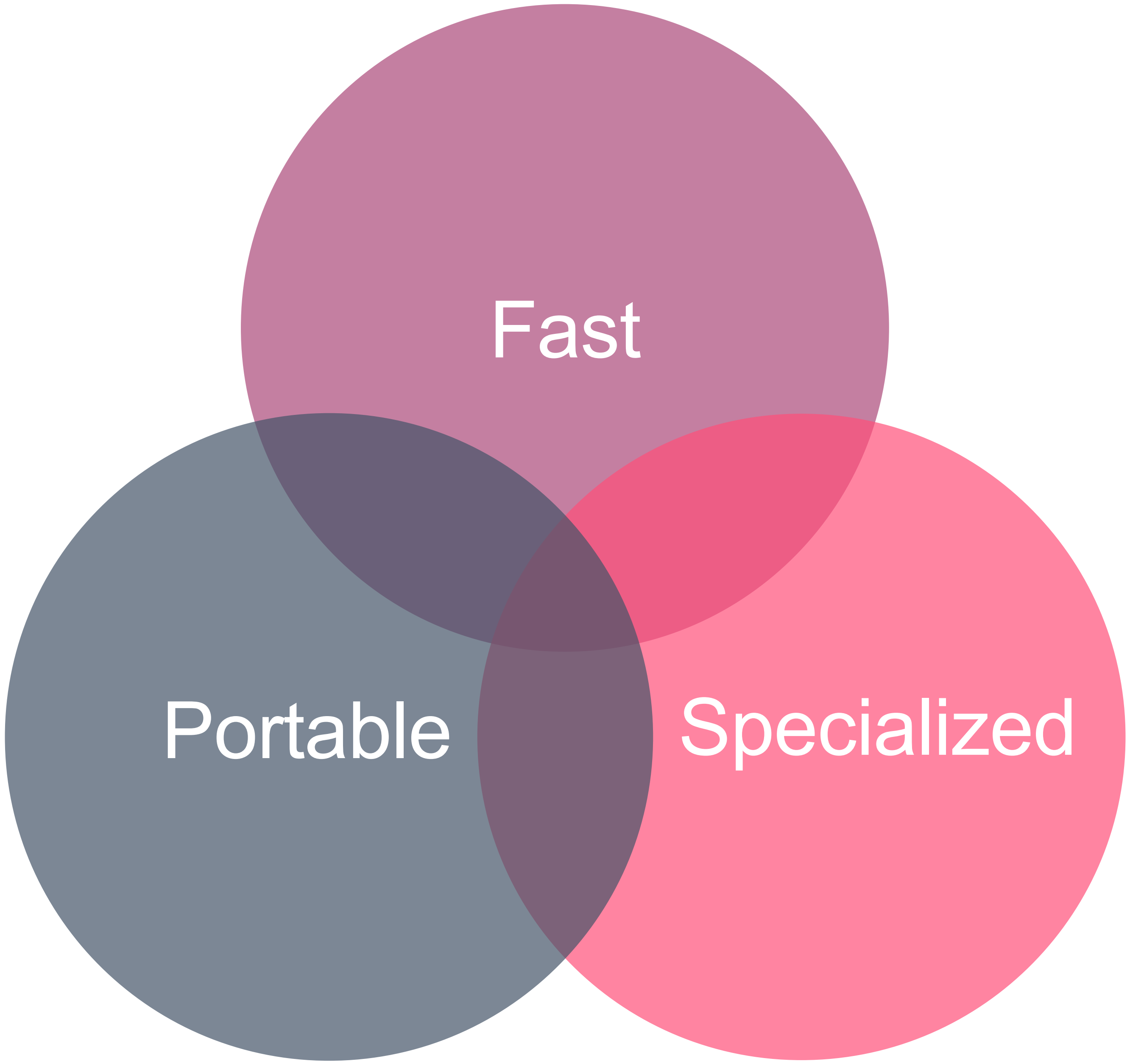
Petabytes of Simulated Data

Large amounts of simulated data needs to be stored and transferred



Looking for a Solution

Yale



Fast Simulation

Yale

1. "**Fast Simulation**":
   - Parametrized showers, frozen showers, …
   - Not accurate or fast enough

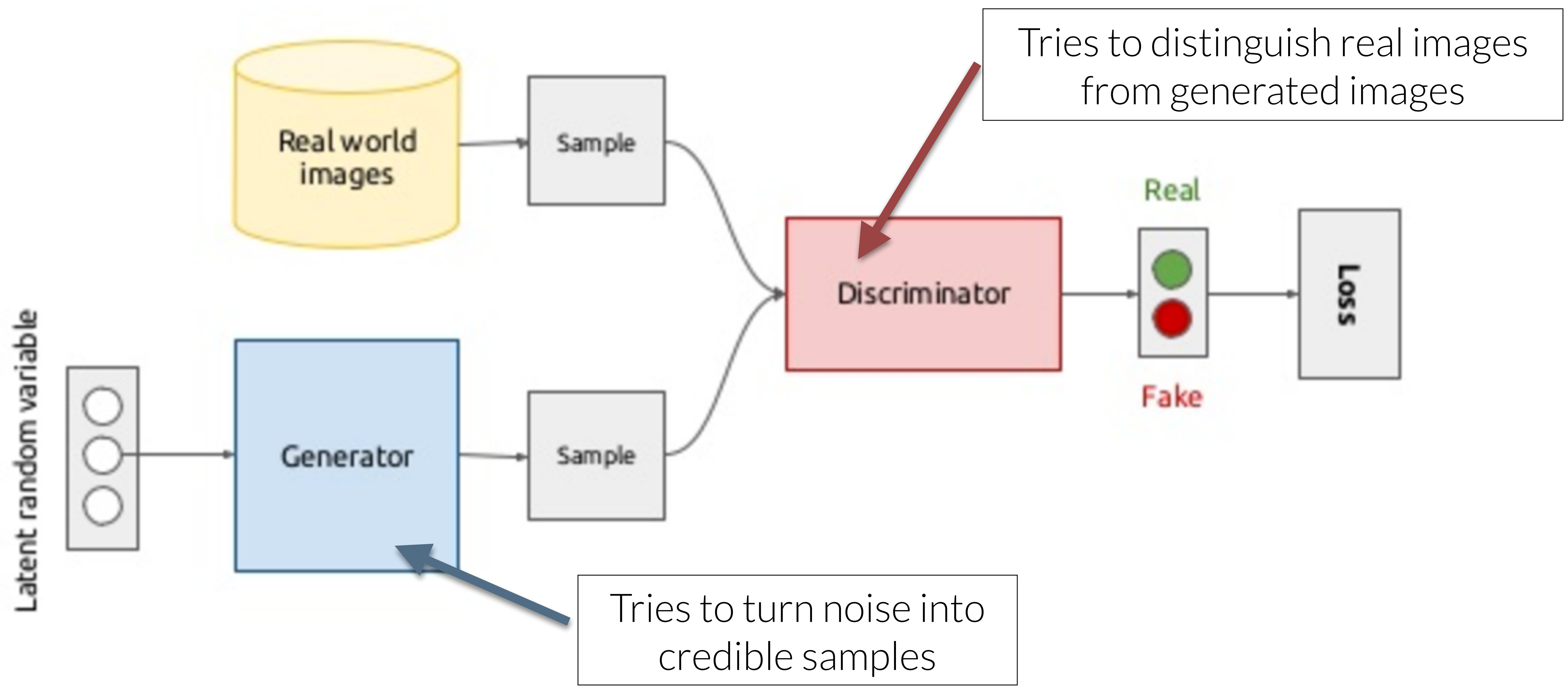
2. **Deep Learning** approaches:
   - Variational Auto-Encoders, Autoregressive Models, Generative Adversarial Networks, …

GANs

Yale

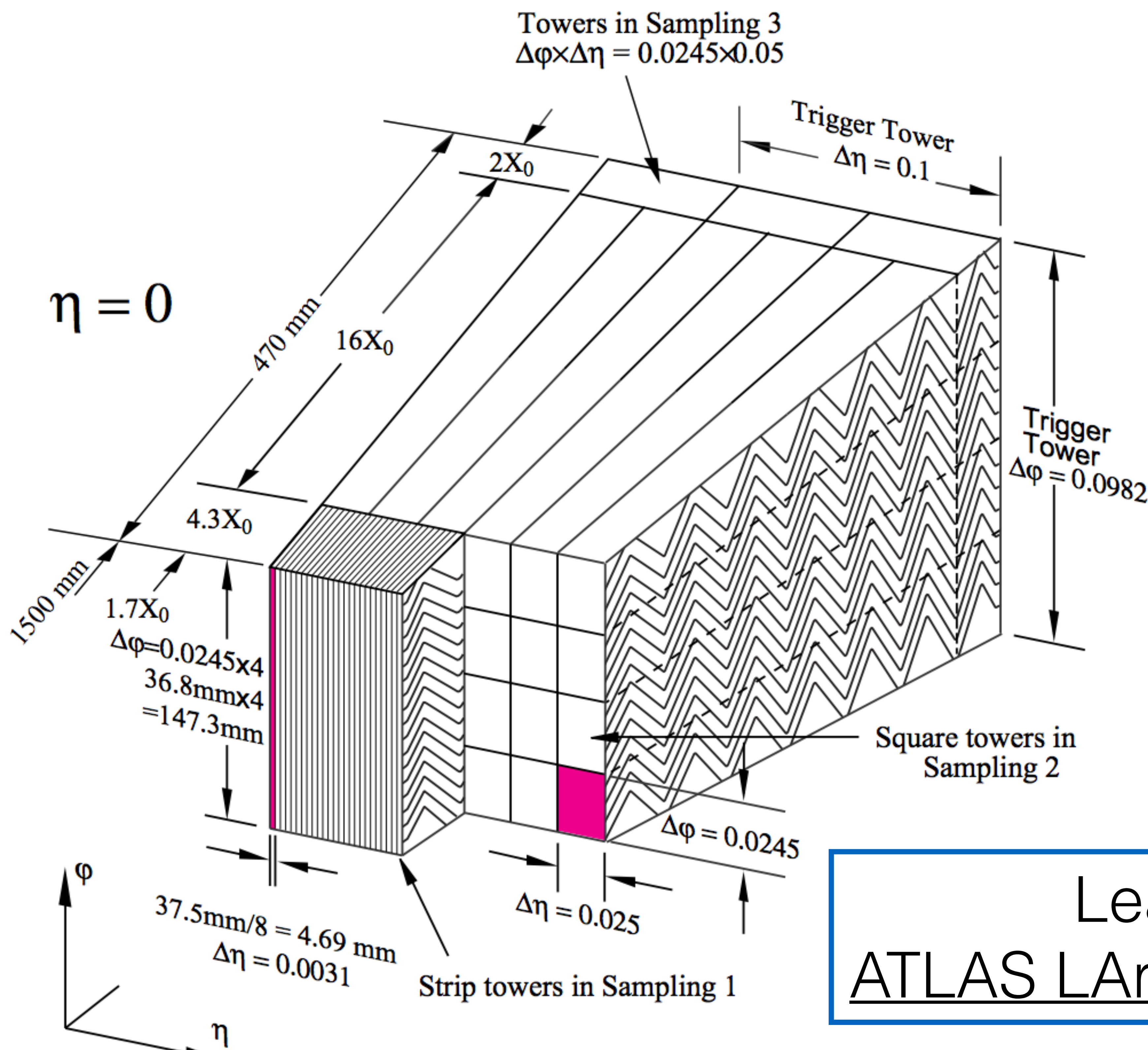
Generative Adversarial Networks (**Goodfellow et al., 2014**):

- Two player non-cooperative game between two deep neural networks, the Generator and the Discriminator



ATLAS EM Calorimeter

Yale

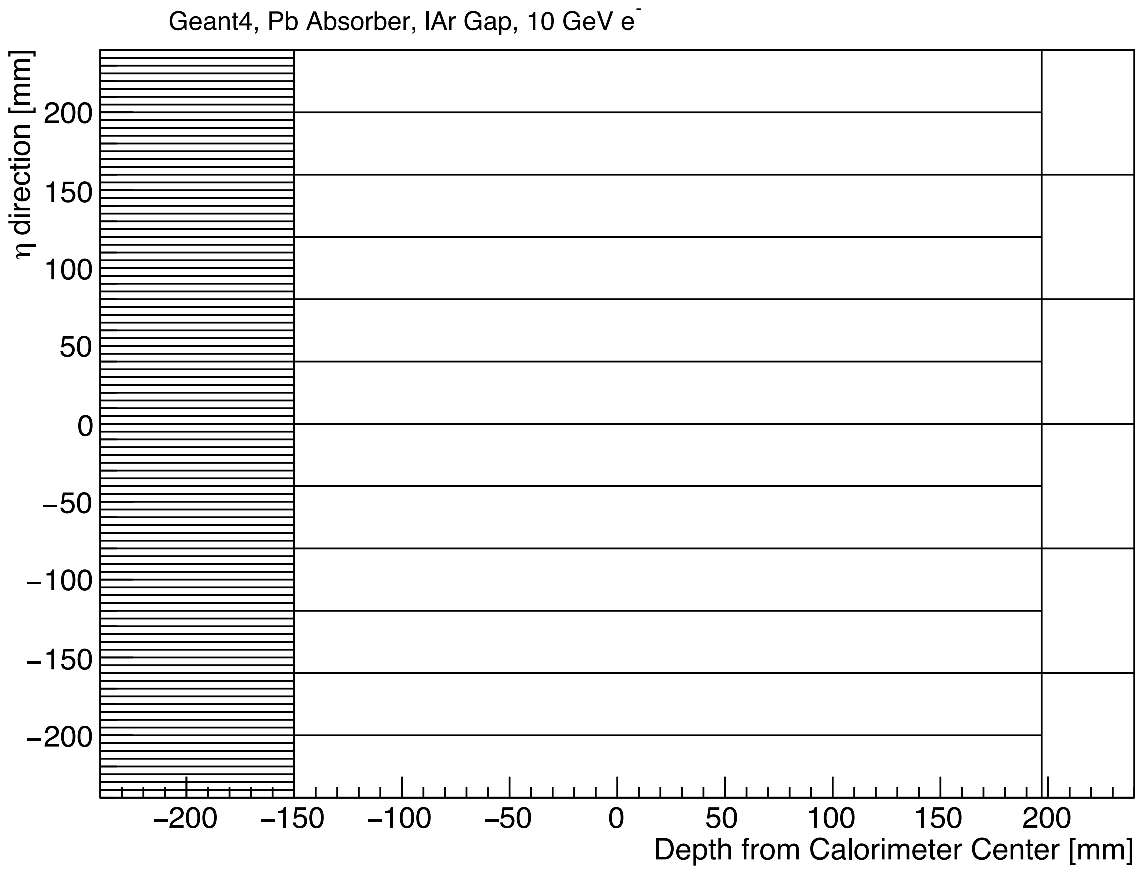


Learn more:
ATLAS LAr Calorimeter TDR

Shower Images

Yale

- EM calorimeter drawing inspiration from the ATLAS geometry.
- Built with GEANT4.



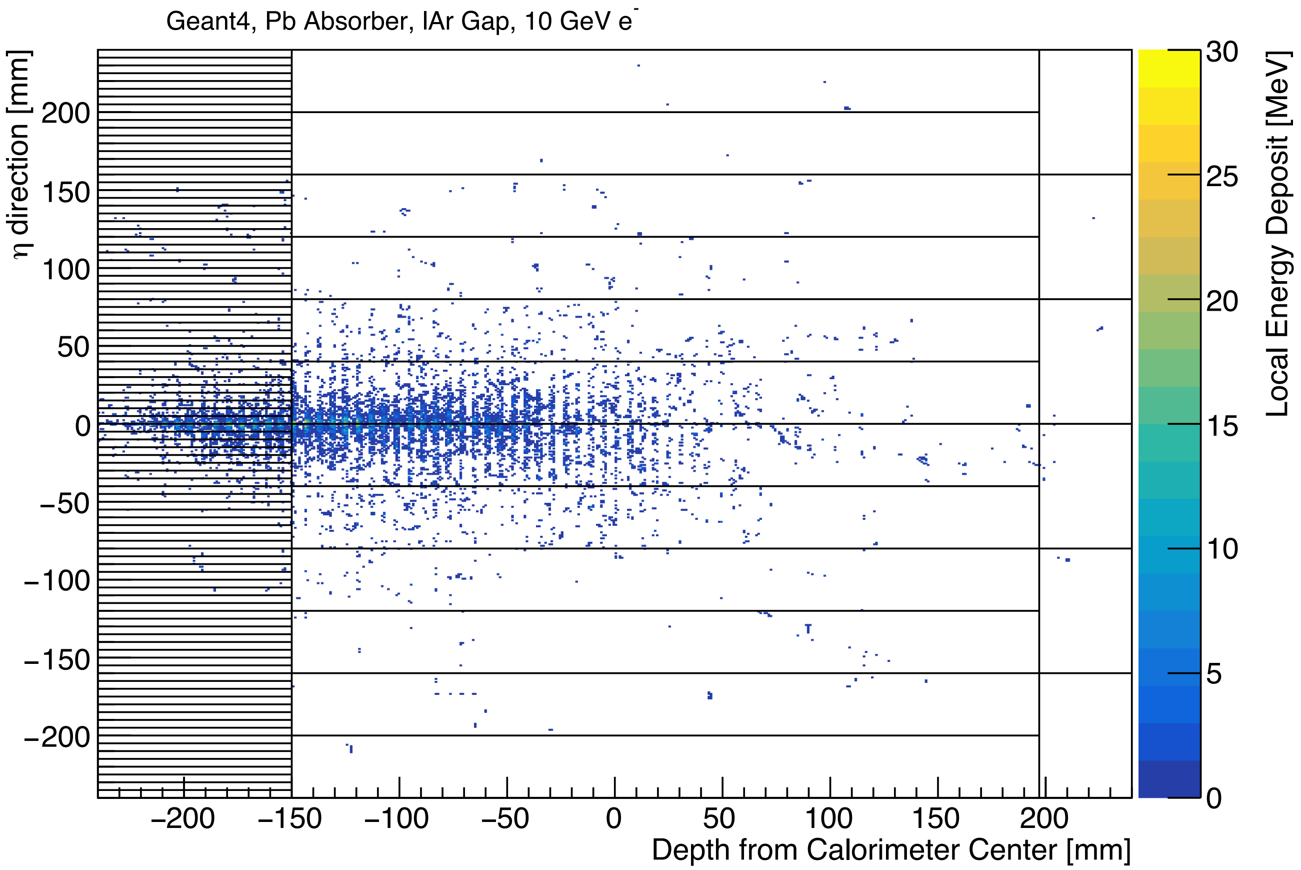
- Heterogeneous longitudinal segmentation into **3 layers**.
- **Irregular granularity** in eta and phi.
- Sequence of alternating lead and liquid argon sublayers.

Shower Images

Yale

- EM calorimeter drawing inspiration from the ATLAS geometry.
- Built with GEANT4.

- Heterogeneous longitudinal segmentation into
- **Irregular granularity**
- Sequence of alternating lead and liquid argo

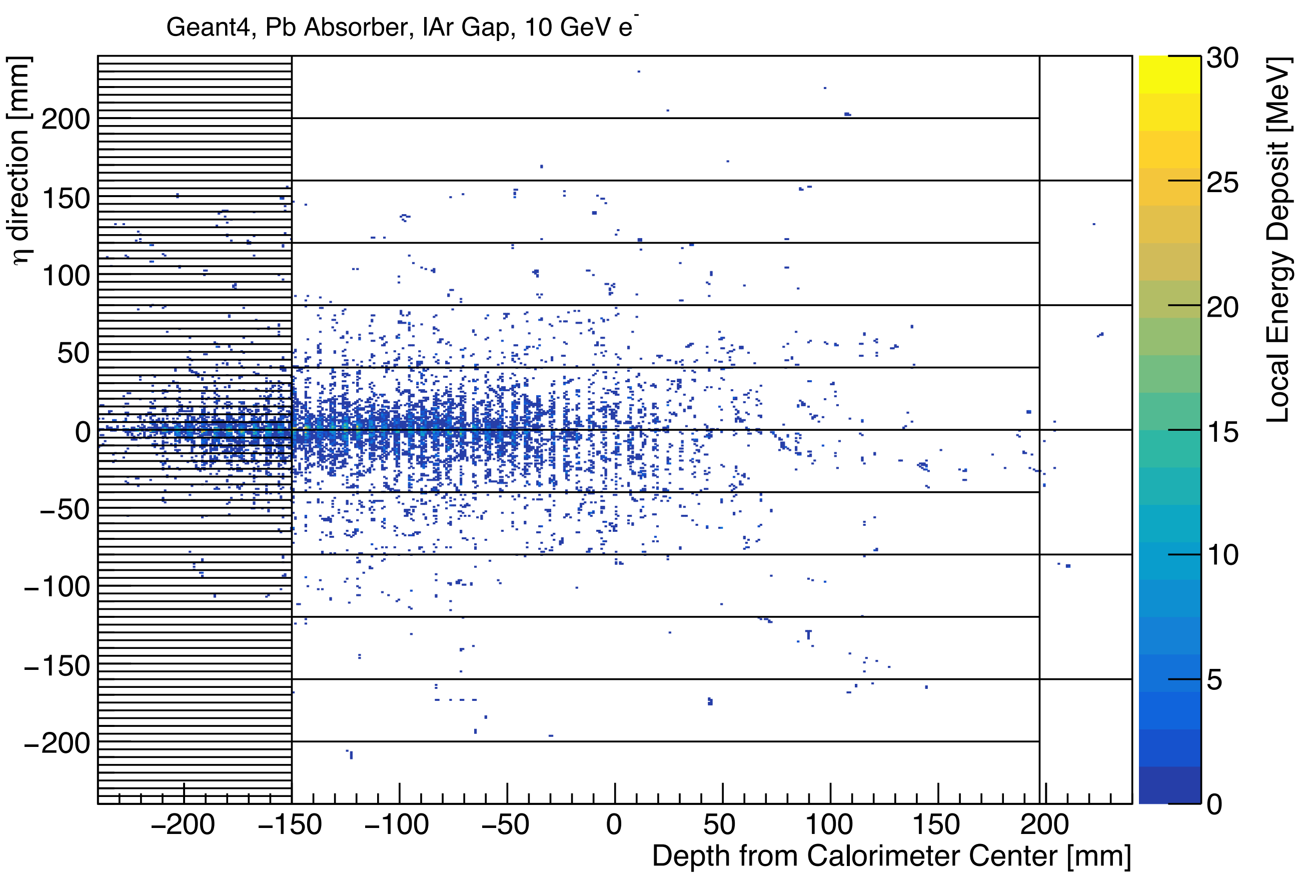


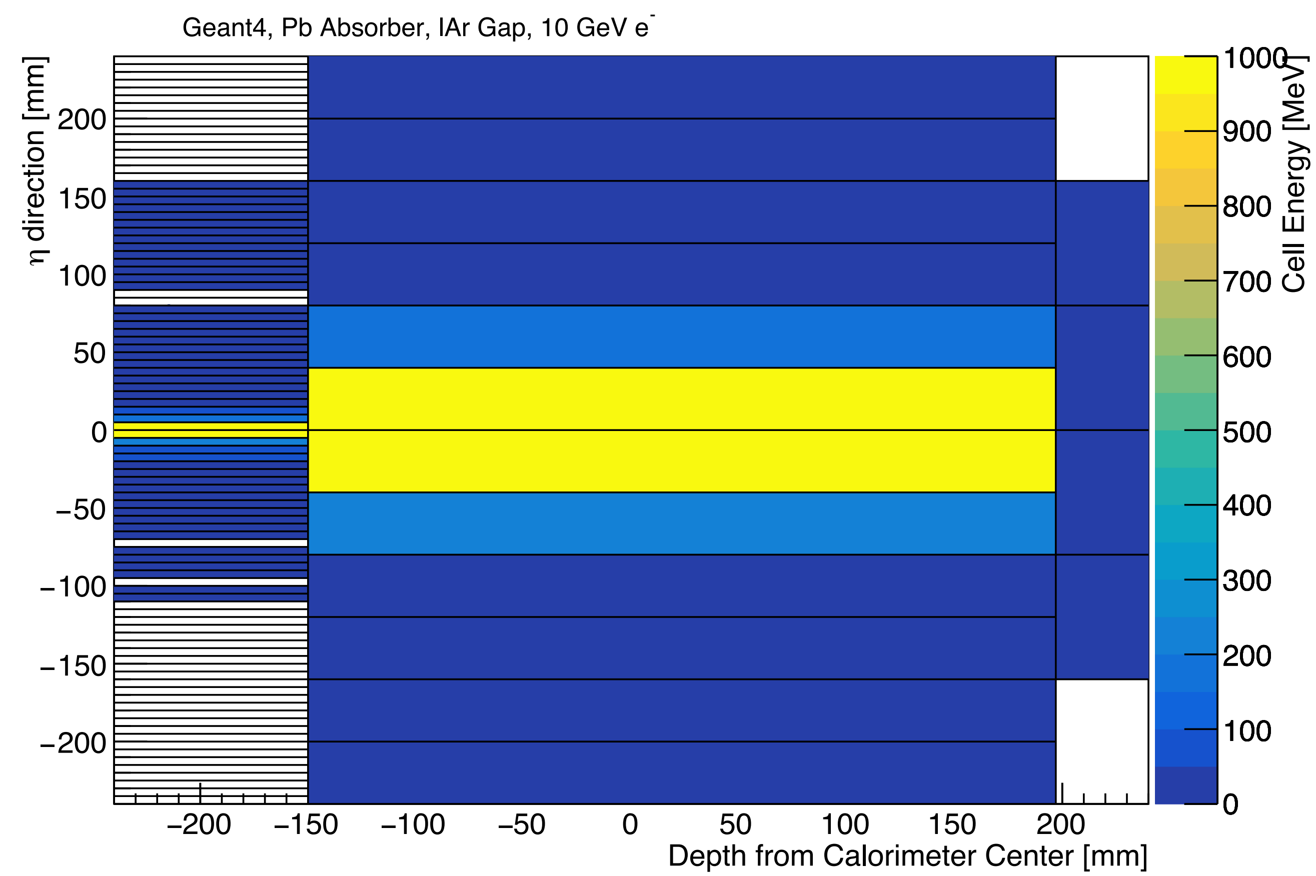
Simulated showers of e^+, pi^+ and y incident ⊥ to the center of the detector with uniform energy in [1, 100] GeV

Shower Images

- EM calorimeter drawing inspiration from the ATLAS geometry.
- Built with GEANT4.

- Heterogeneous longitudinal segmentation into
- **Irregular granularity**
- Sequence of alternating lead and liquid argo



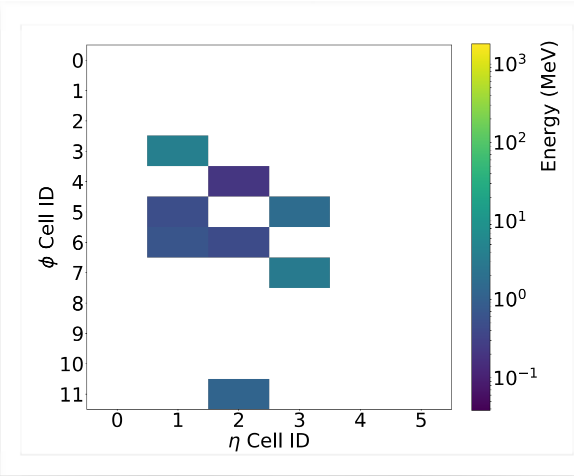


Simulated showers of e^+, pi^+ and y incident ⊥ to the center of the detector with a uniform energy in [1, 100] GeV

Shower Images

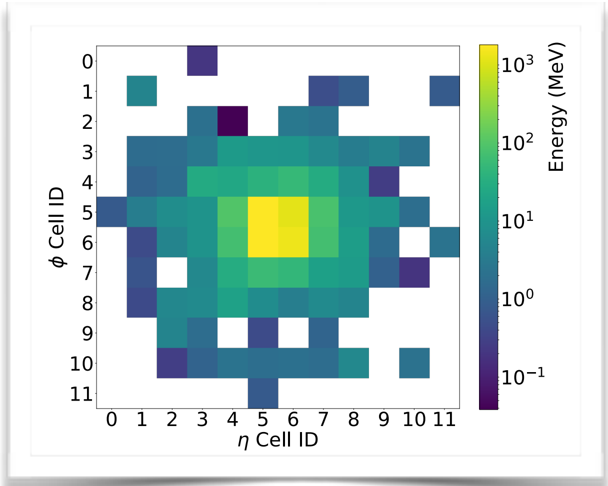
Yale

- Energy depositions in each layer as a **2D image**

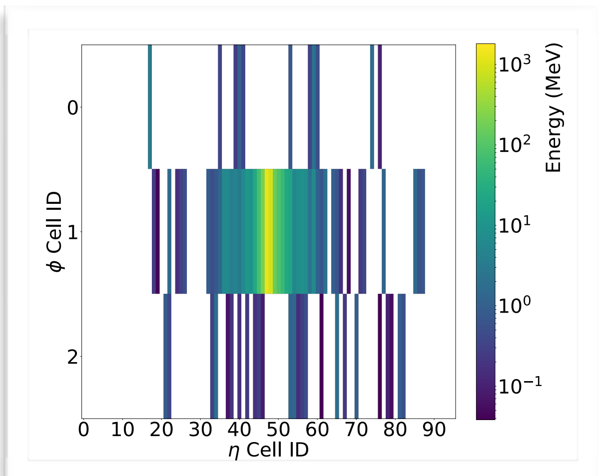


12x6

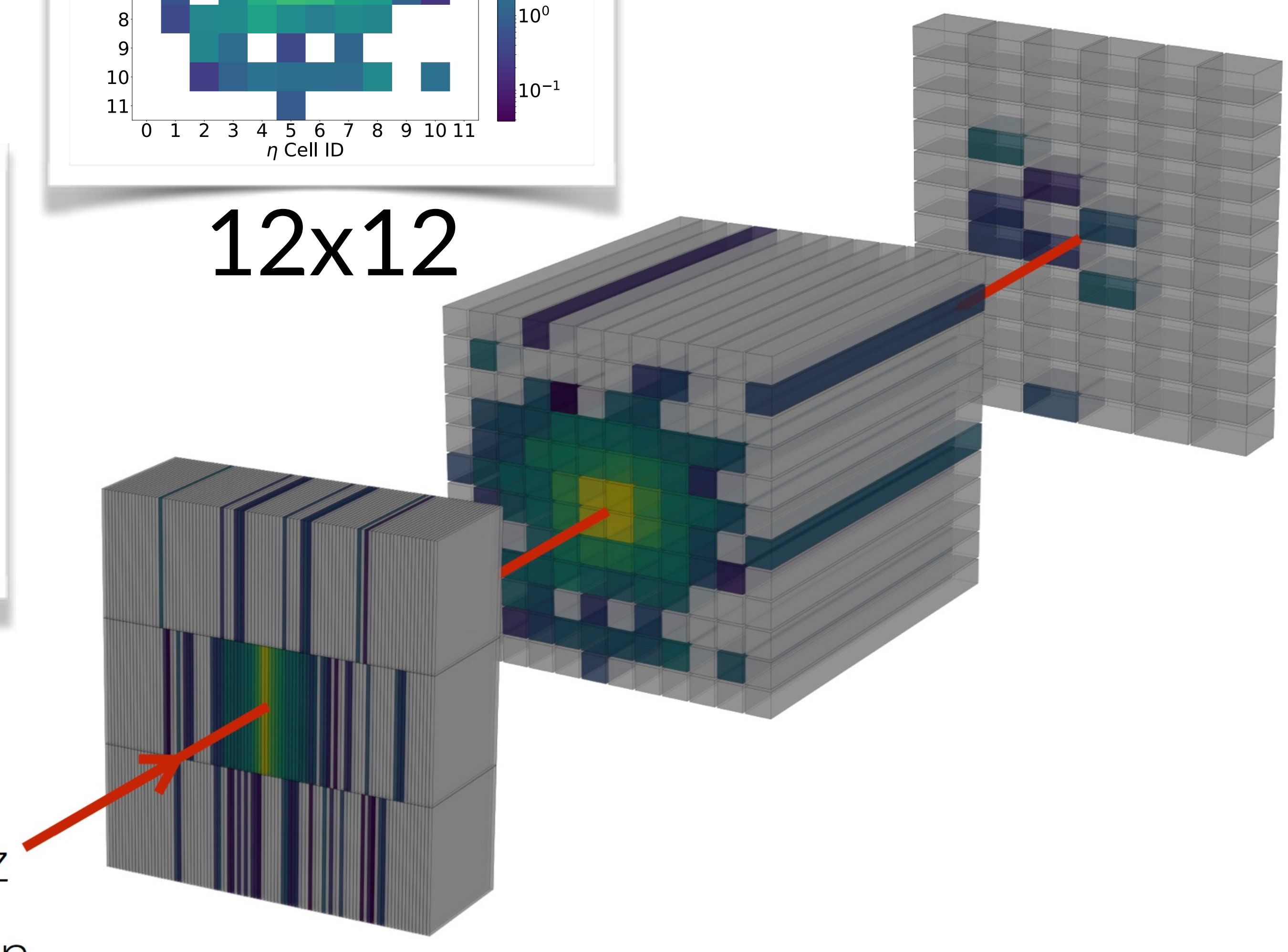
Layer	z segmentation [mm]	η segmentation [mm]	ϕ segmentation [mm]
0	90	5	160
1	347	40	40
2	43	80	40

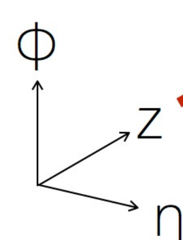


12x12



3x96



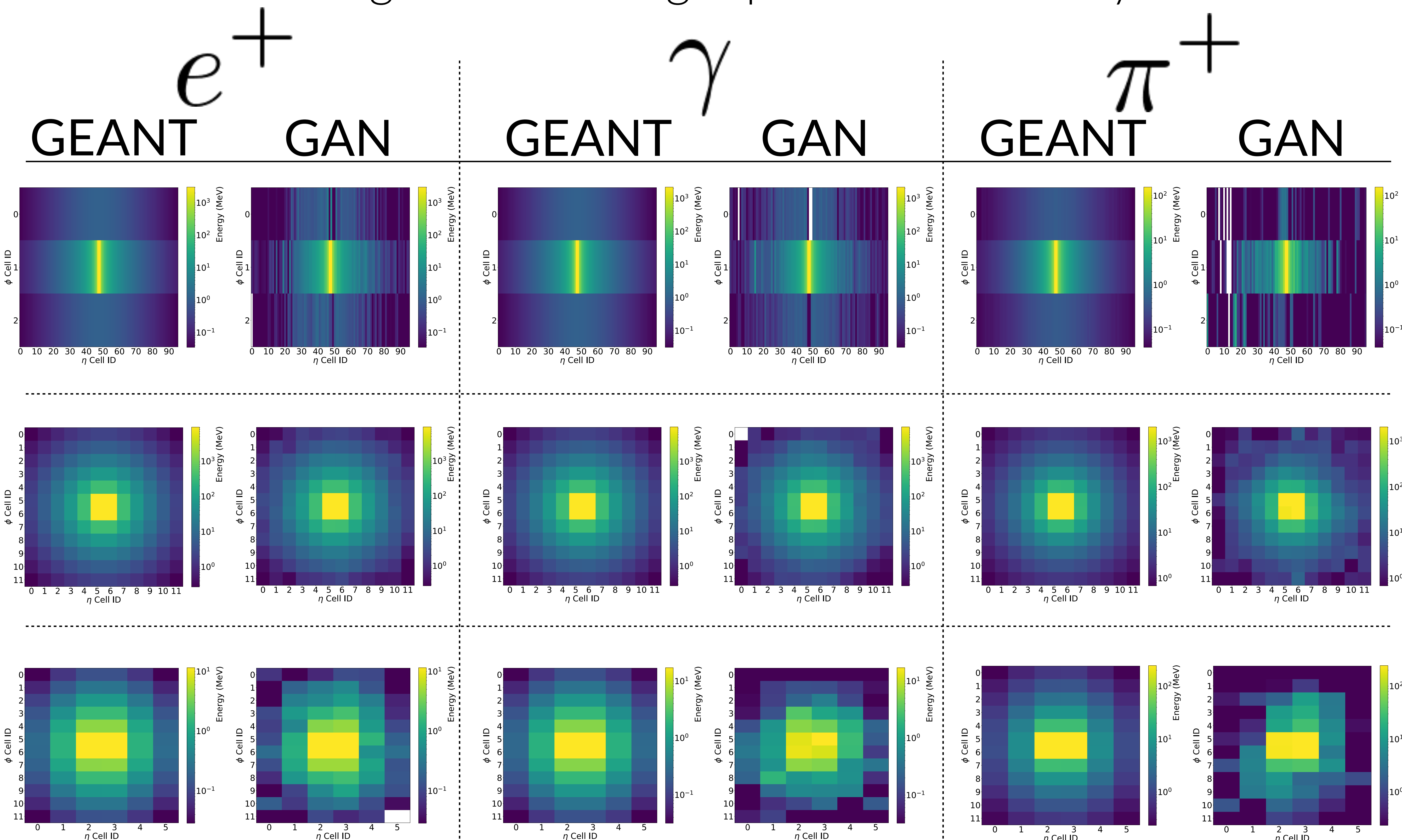


- <u>Goal</u>:
  generate this
  fixed representation

Qualitative Performance (1)

Yale

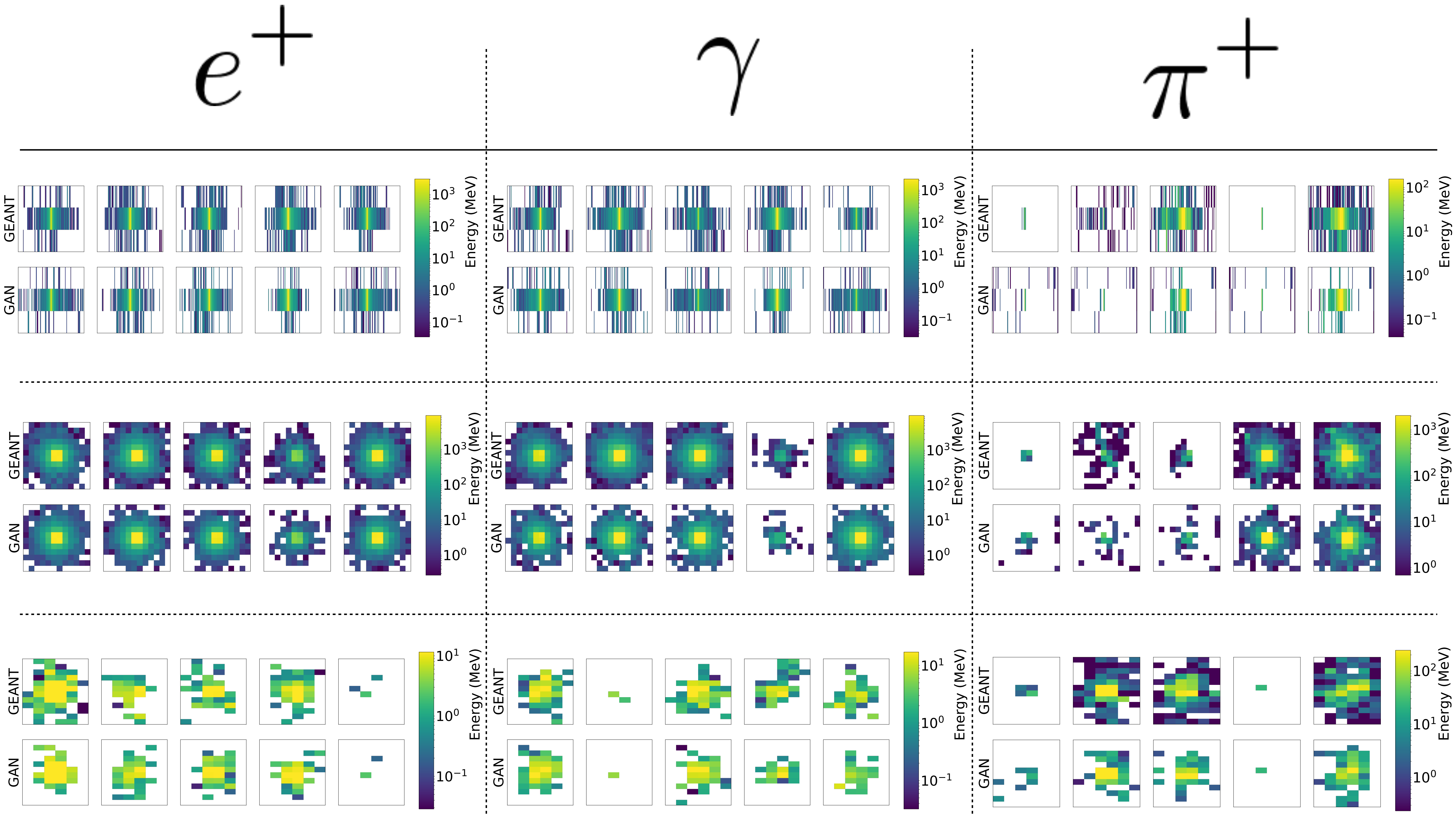
- Average shower images per calorimeter layer



Qualitative Performance (2)

Yale

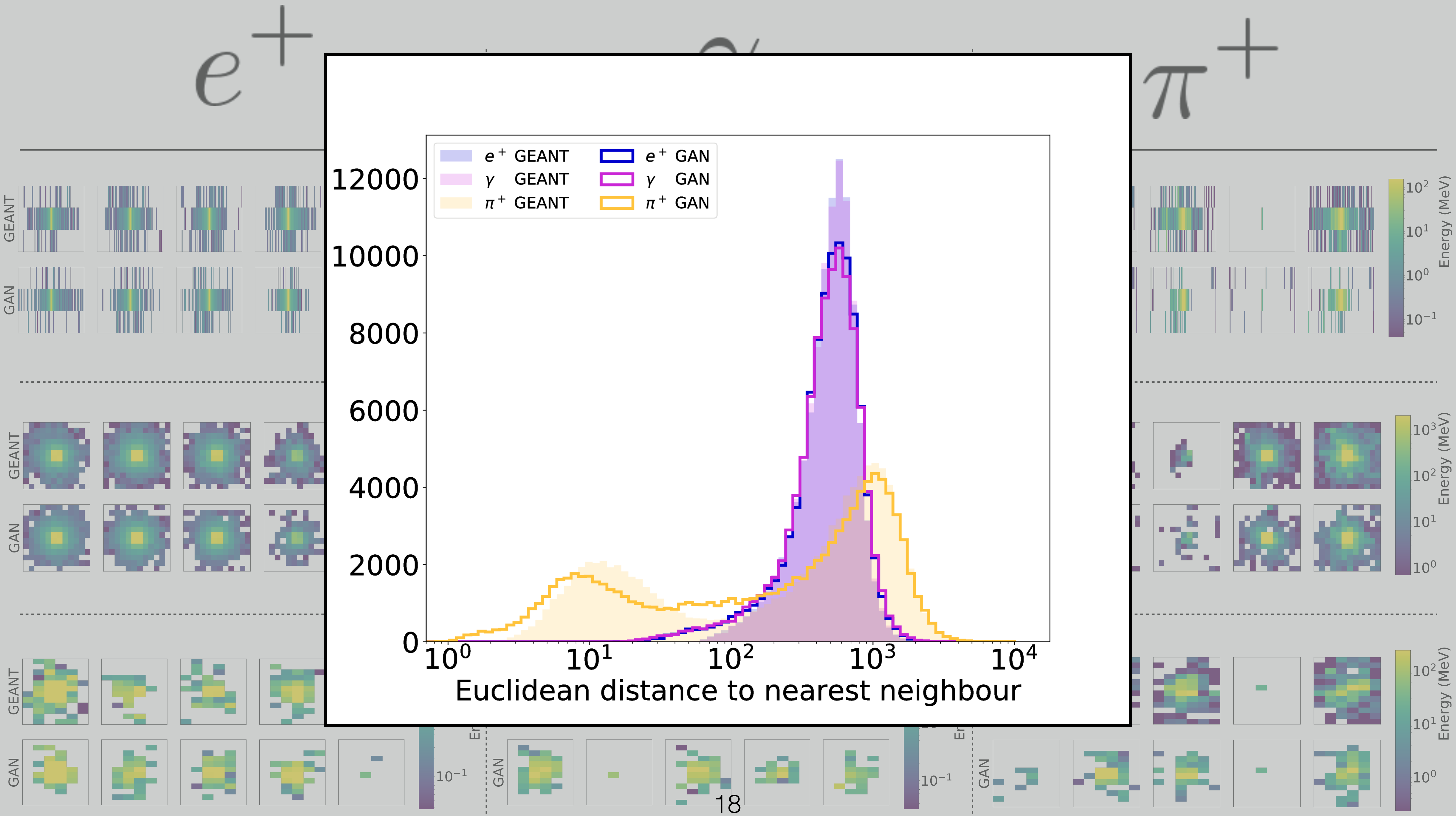
- 5 random shower images and their nearest GAN-generated neighbor



Qualitative Performance (2)

Yale

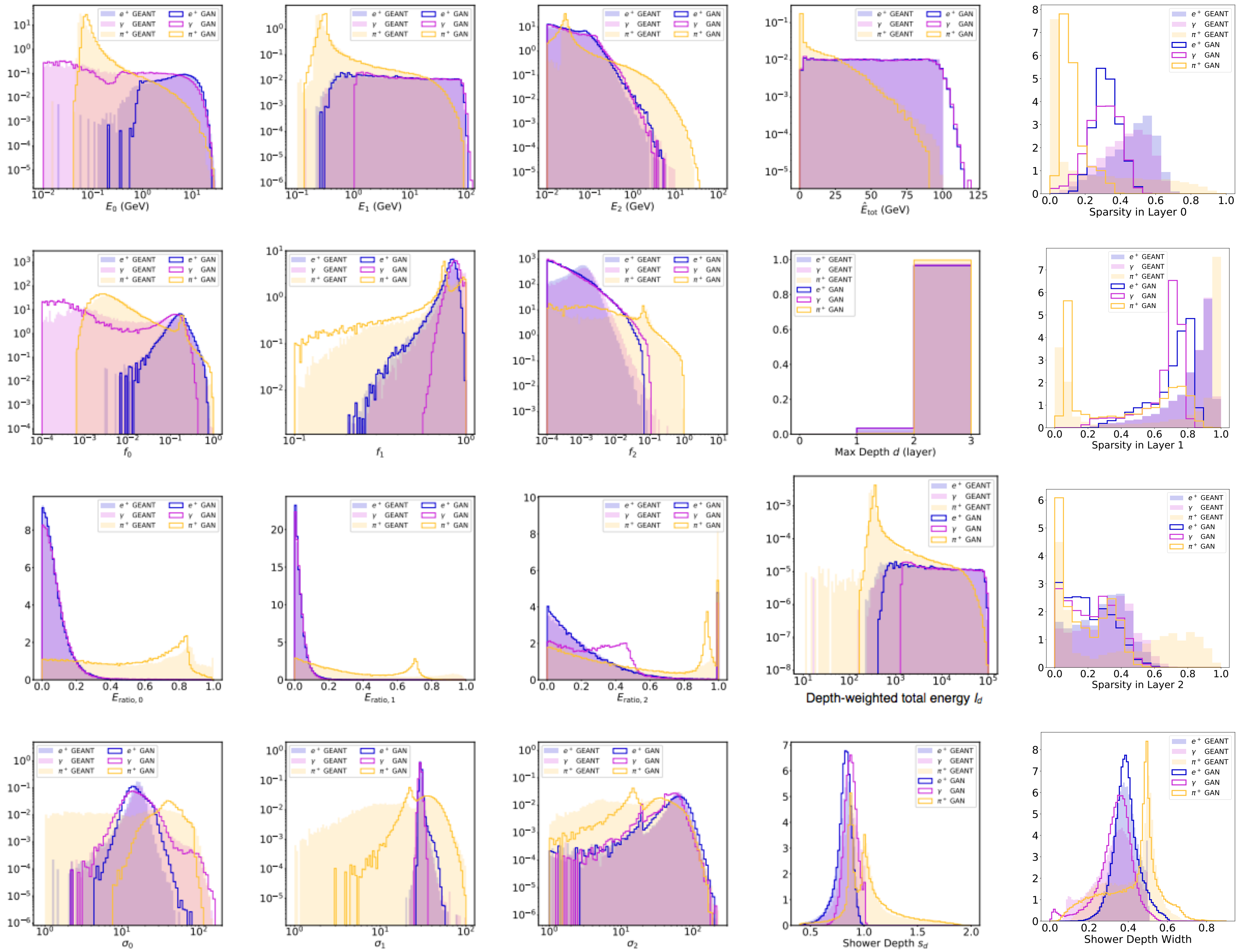
- 5 random shower images and their nearest GAN-generated neighbor



Shower Shapes

Yale

Check: does the GAN recover the true data distribution as projected onto a set of meaningful 1D manifolds?



Computing

Yale

Compared to single core sequential GEANT 4 simulation, can obtain an evaluation time speed **~100,000x** when evaluating GAN on GPU

Generation Method	Hardware	Batch Size	milliseconds/shower
GEANT4	CPU	N/A	1772
CALOGAN	CPU	1	13.1
		10	5.11
		128	2.19
		1024	2.03
	GPU	1	14.5
		4	3.68
		128	0.021
		512	0.014
		1024	0.012

GEANT 4 generation time grows with energy,
GAN generation time is constant

Summary

Yale

- Main point: **we have a working solution!** `arXiv:1705.02355`
- Problem we are trying to tackle: accurate, fast calo simulation
- Emphasis on **reproducibility**:

contribute!

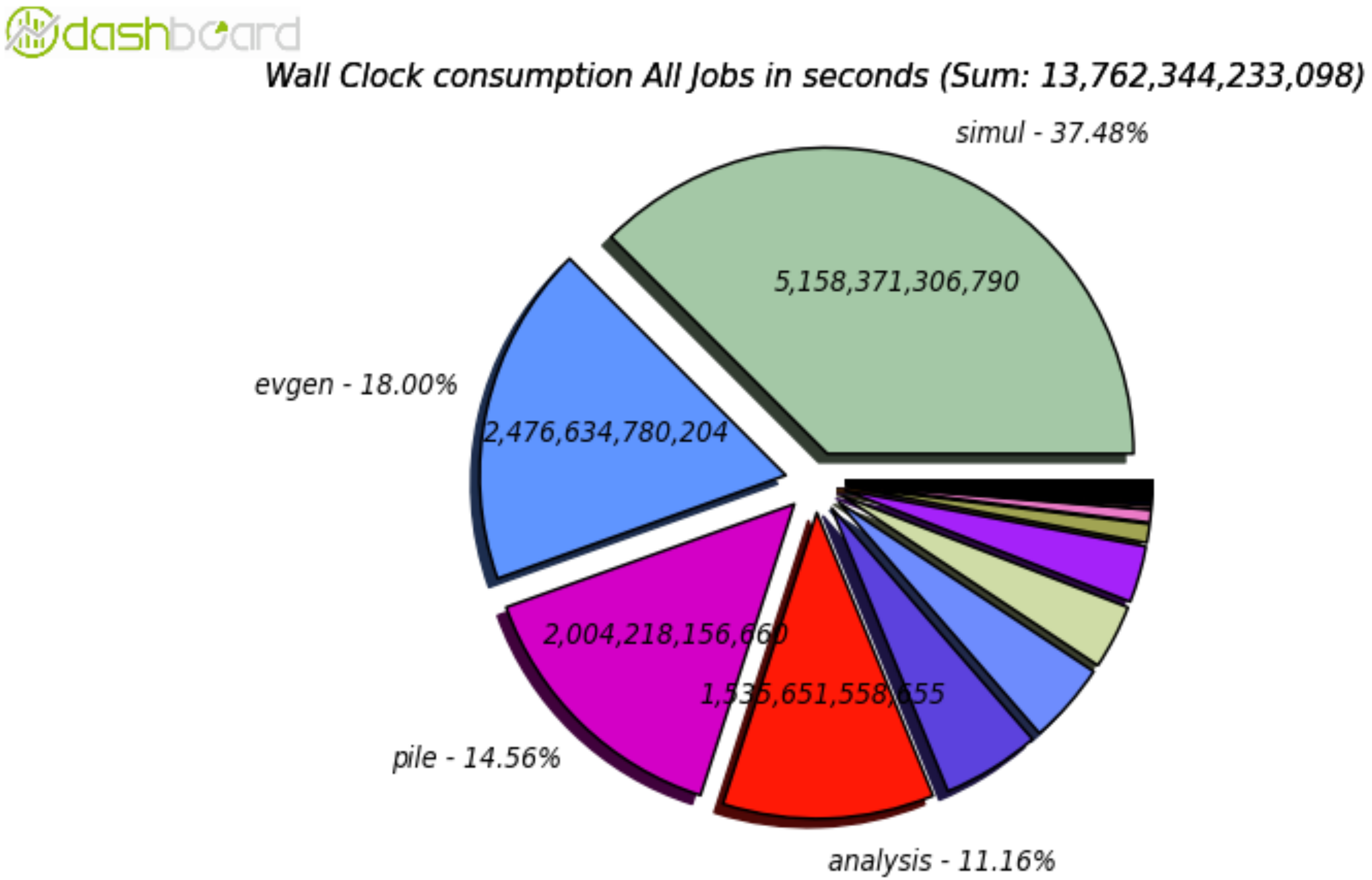
Data	Code
DOI 10.17632/pvn3xc3wy5.1	github.com/hep-lbdl/CaloGAN
zenodo MENDELEY DATA	jupyter

- Goal: getting this to work in ATLAS/GEANT in the not-so-distant future

Backup

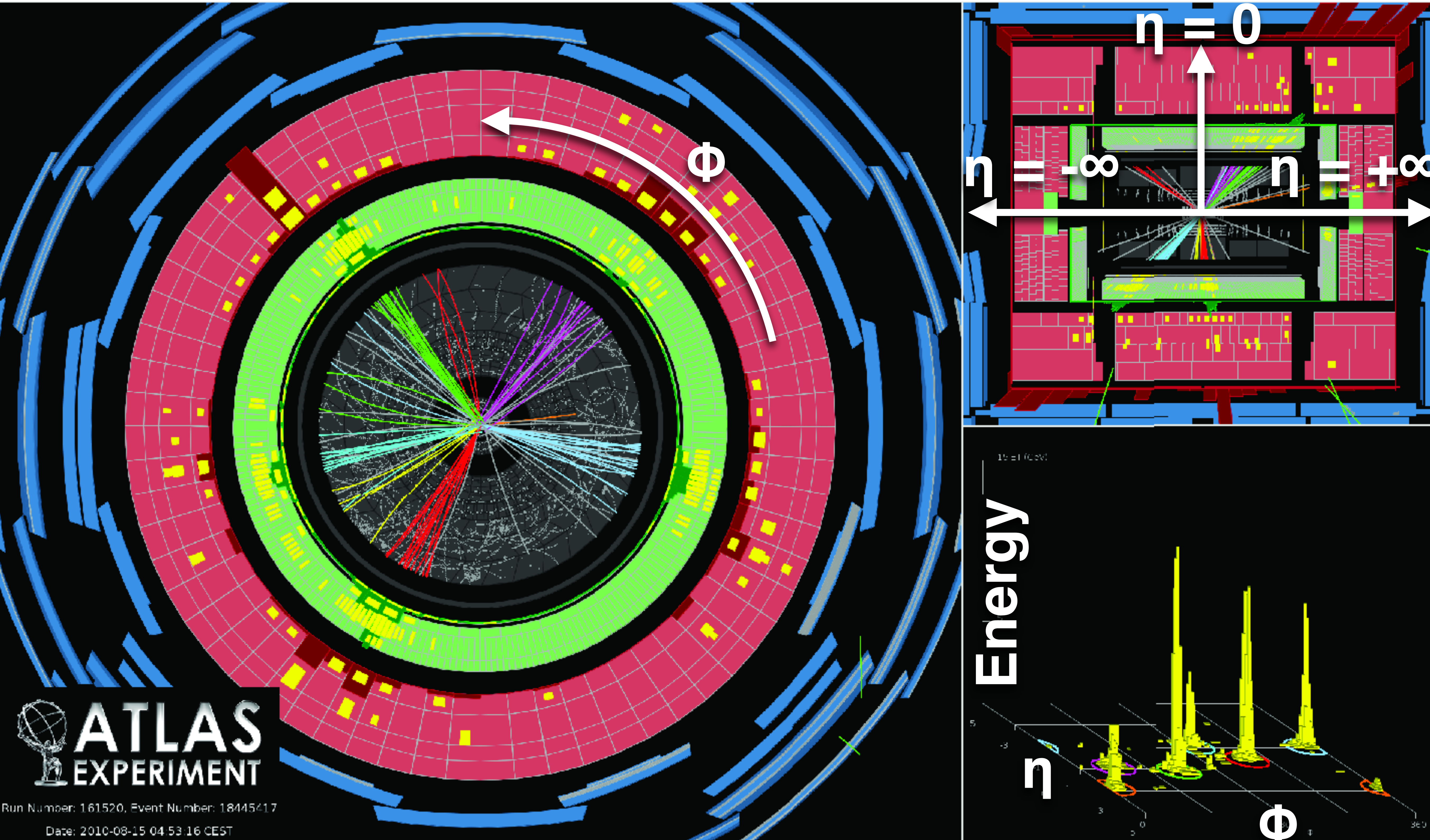
Grid Utilization

Yale



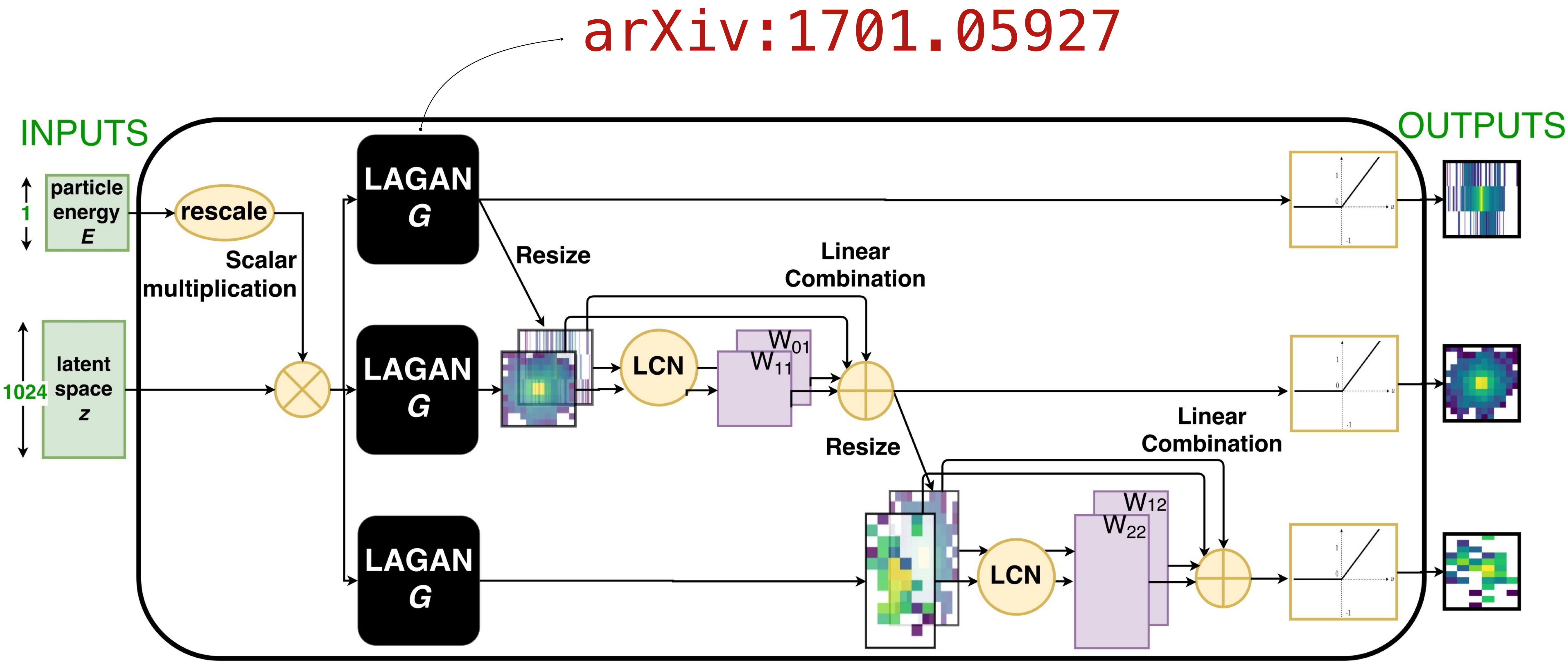
Events in ATLAS

Yale



CaloGAN Generator

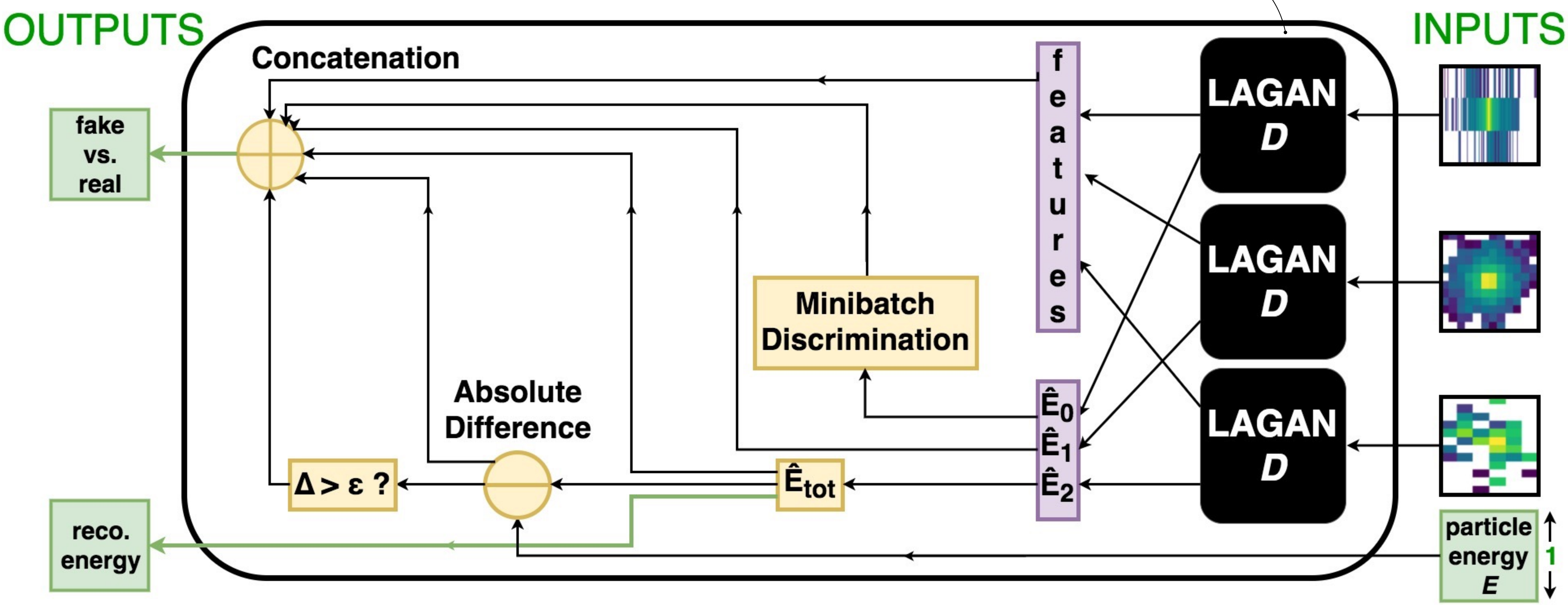
Yale



CaloGAN Discriminator

Yale

arXiv:1701.05927



Shower Shape Variables

Yale

Shower Shape Variable	Formula	*Notes*
E_i	$E_i = \sum_{\text{pixels}} \mathcal{I}_i$	Energy deposited in the i^{th} layer of calorimeter
E_{tot}	$E_{\text{tot}} = \sum_{i=0}^{2} E_i$	Total energy deposited in the electromagnetic calorimeter
f_i	$f_i = E_i / E_{\text{tot}}$	Fraction of measured energy deposited in the i^{th} layer of calorimeter
$E_{\text{ratio},i}$	$\dfrac{\mathcal{I}_{i,(1)} - \mathcal{I}_{i,(2)}}{\mathcal{I}_{i,(1)} + \mathcal{I}_{i,(2)}}$	Difference in energy between the highest and second highest energy deposit in the cells of the i^{th} layer, divided by the sum
d	$d = \max\{i : \max(\mathcal{I}_i) > 0\}$	Deepest calorimeter layer that registers non-zero energy
Depth-weighted total energy, l_d	$l_d = \sum_{i=0}^{2} i \cdot \mathcal{I}_i$	The sum of the energy per layer, weighted by layer number.
Shower Depth, s_d	$s_d = l_d / E_{\text{tot}}$	The energy-weighted depth in units of layer number.
Shower Depth Width, σ_{s_d}	$\sigma_{s_d} = \sqrt{\dfrac{\sum_{i=0}^{2} i^2 \cdot \mathcal{I}_i}{E_{\text{tot}}} - \left(\dfrac{\sum_{i=0}^{2} i \cdot \mathcal{I}_i}{E_{\text{tot}}} \right)^2}$	The standard deviation of s_d in units of layer number.
i^{th} Layer Lateral Width, σ_i	$\sigma_i = \sqrt{\frac{\mathcal{I}_i \odot H^2}{E_i} - \left(\frac{\mathcal{I}_i \odot H}{E_i} \right)^2}$	The standard deviation of the transverse energy profile per layer, in units of cell numbers.

Recent HEP-ML Conferences and Workshops

Yale

- NIPS Workshop 2014, 2015
- DS@HEP 2015, 2016, 2017
- Connecting the Dots 2016, 2017
- IML Workshop
- CERN OpenLab workshop on ML and Data Analytics