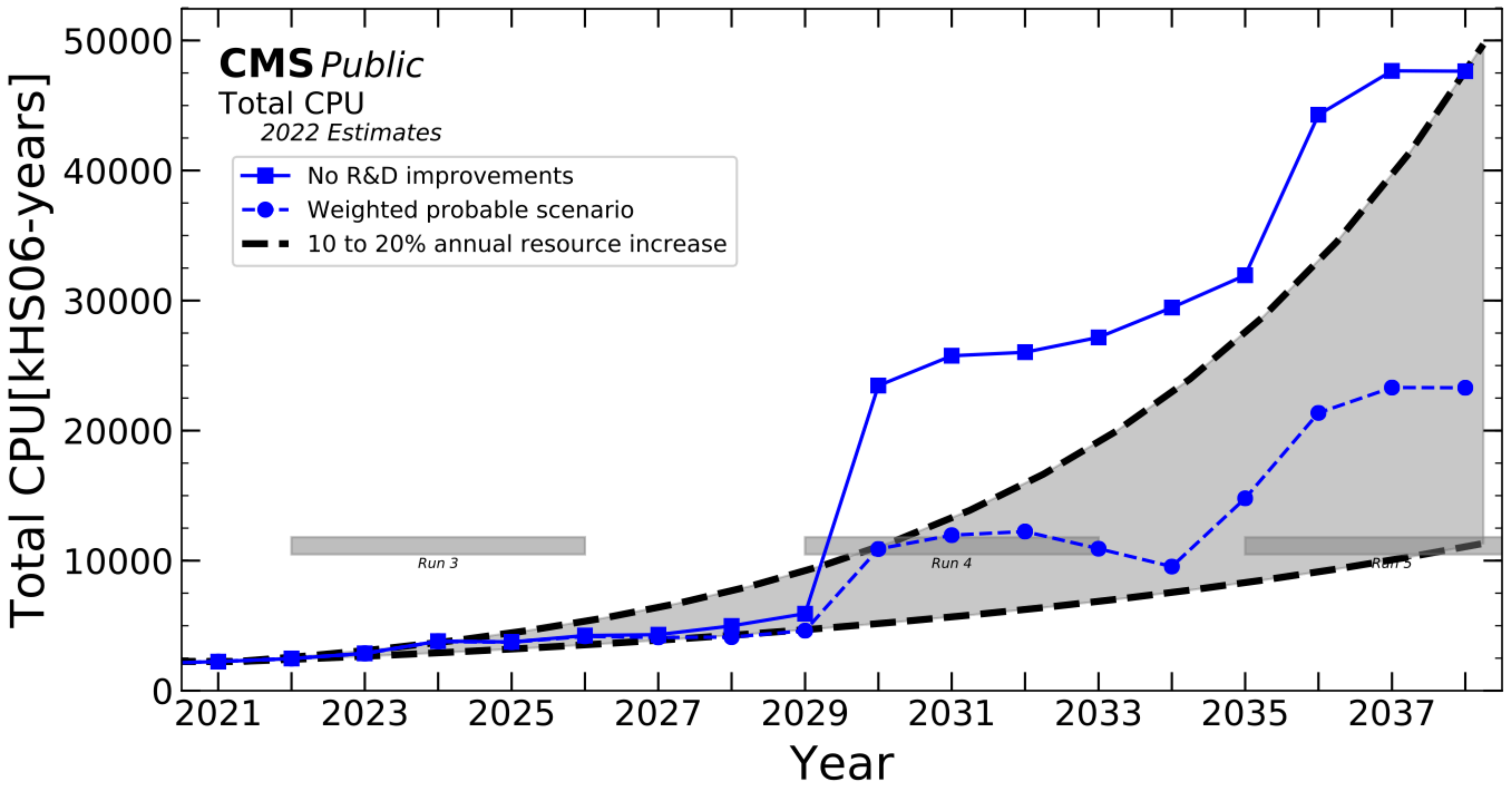
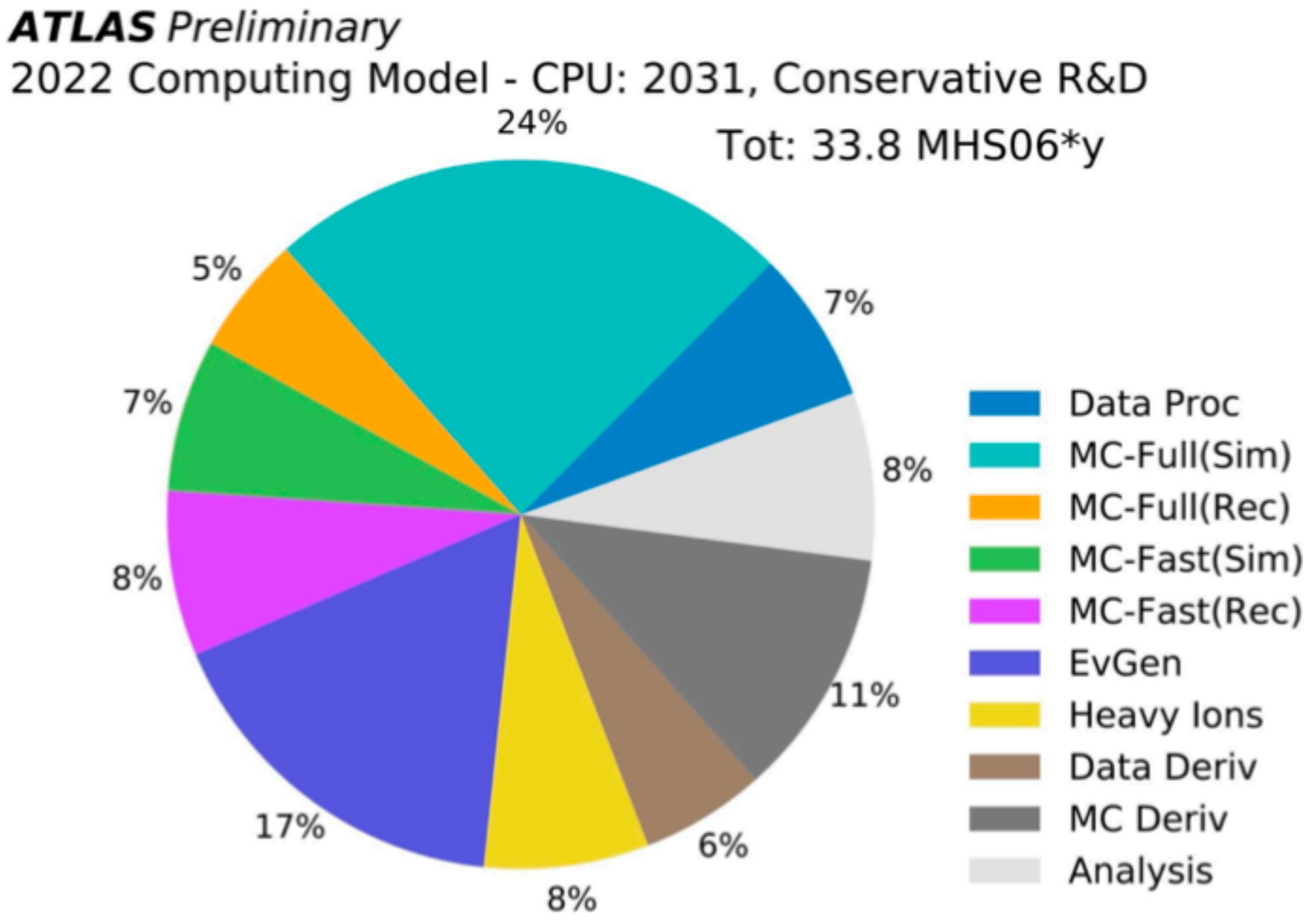
Lecture 2:
ML for Surrogate Modeling

Surrogate Modeling for Calorimeter Showers

https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults





CERN-LHCC-2022-005

Detector simulation with GEANT4 is the single largest computational bottleneck at LHC and other experiments

Surrogate Modeling for Calorimeter Showers

GEANT4 → 10^{10} events **SLOW but ACCURATE**

Surrogate Modeling for Calorimeter Showers

GEANT4 → 10^{10} events **SLOW but ACCURATE**

GEANT4 → 10^{5} events

Surrogate Modeling for Calorimeter Showers

GEANT4 → 10^{10} events **SLOW but ACCURATE**

GEANT4 → 10^{5} events → Surrogate model

(GAN, VAE, Normalizing Flow, …)
Learn underlying distribution of GEANT4 events

Surrogate Modeling for Calorimeter Showers

GEANT4 → 10^{10} events **SLOW but ACCURATE**

GEANT4 → 10^{5} events → Surrogate model → 10^{10} events

(GAN, VAE, Normalizing Flow, …)
Learn underlying distribution of GEANT4 events

Surrogate Modeling for Calorimeter Showers

GEANT4 → 10^{10} events **SLOW but ACCURATE**

GEANT4 → 10^{5} events → Surrogate model → 10^{10} events

(GAN, VAE, Normalizing Flow, …)
Learn underlying distribution of GEANT4 events

FAST *and* ACCURATE?

Surrogate Modeling for Calorimeter Showers

GEANT4 → 10^{10} events **SLOW but ACCURATE**

GEANT4 → 10^{5} events → Surrogate model → 10^{10} events

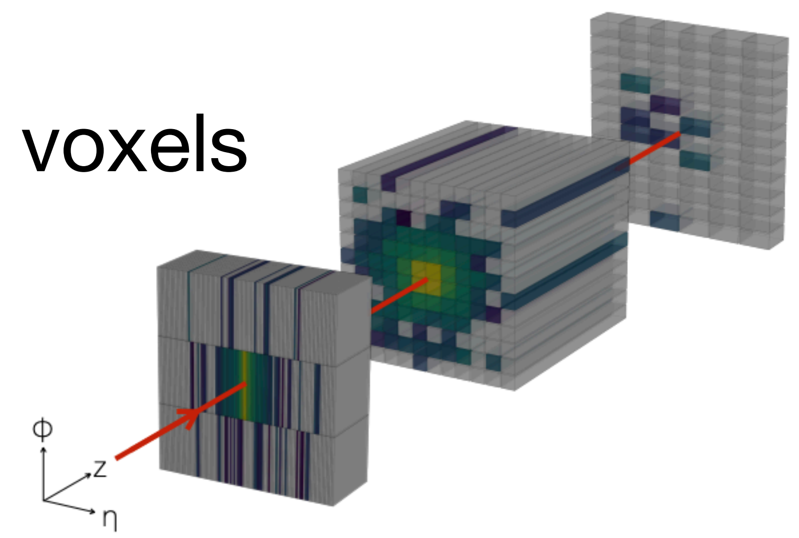
(GAN, VAE, Normalizing Flow, …)
Learn underlying distribution of GEANT4 events

FAST *and* ACCURATE?

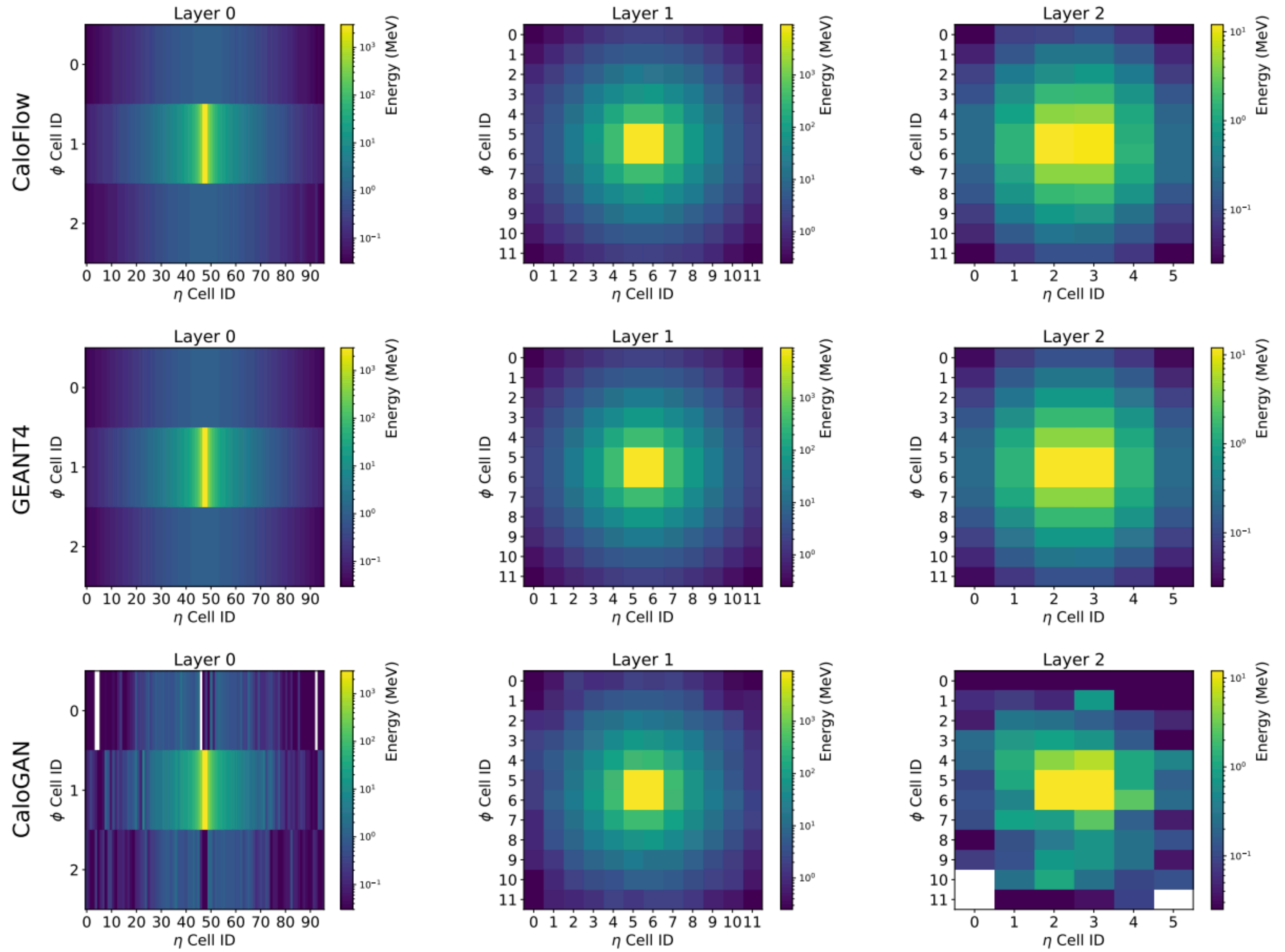
ML methods can provide fast ***and*** accurate “surrogate models” for GEANT4 etc

- Snowmass WP — detector sim — 2203.08806
- Snowmass WP — event generation — 2203.07460

Toy ATLAS ECAL — 3 layers, 504 voxels

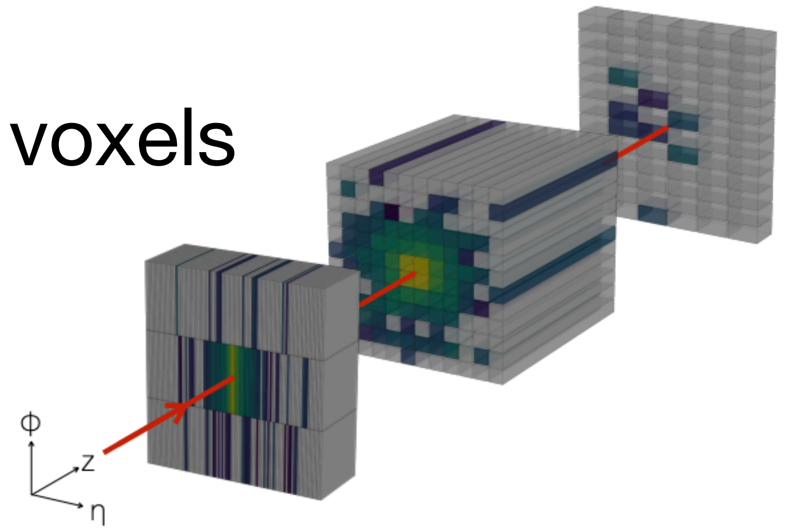


CaloGAN and CaloFlow



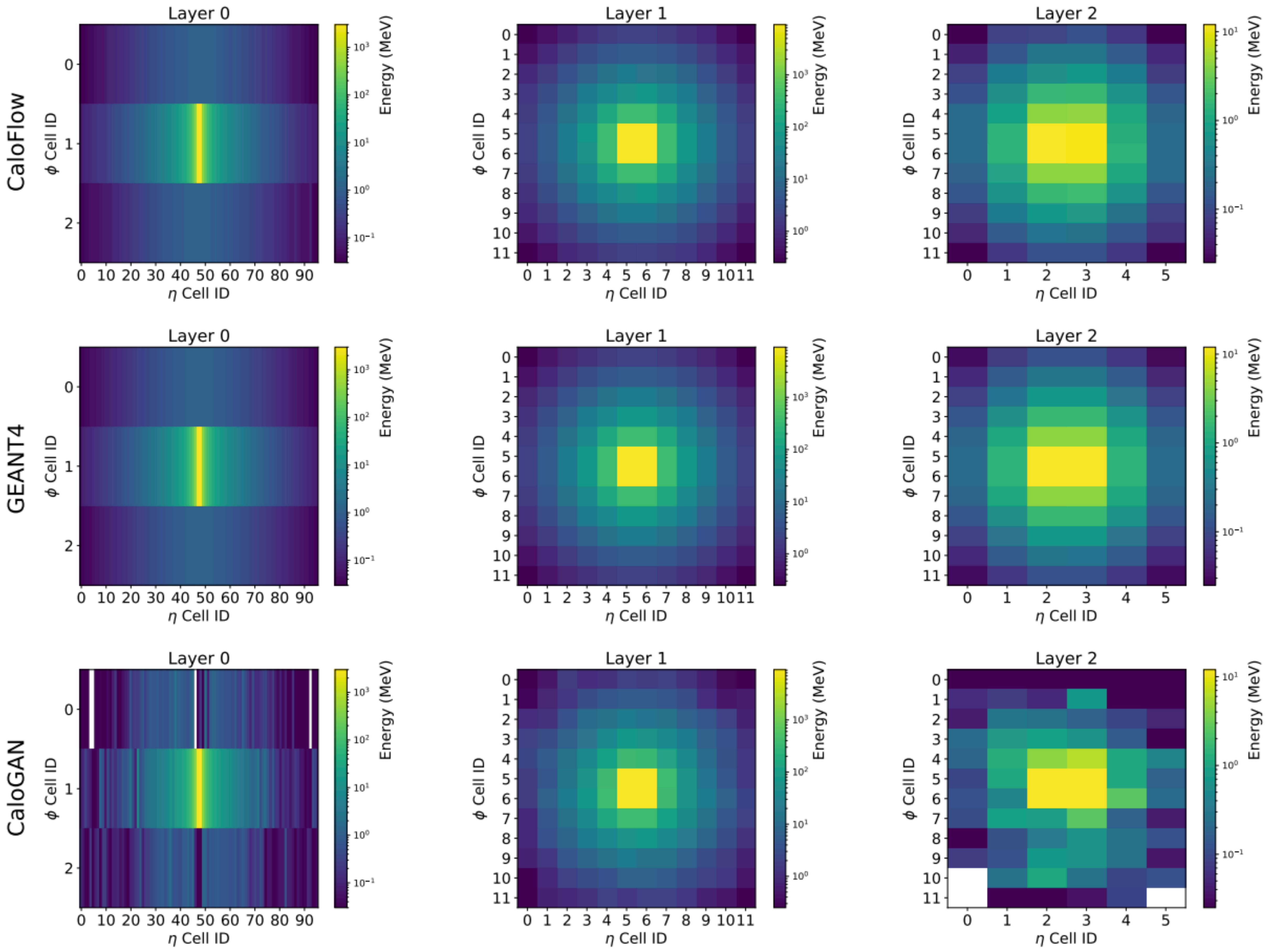
CaloGAN [Paganini, de Oliveira & Nachman (2017)]
First-ever deep surrogate model for fast calosim

Toy ATLAS ECAL — 3 layers, 504 voxels



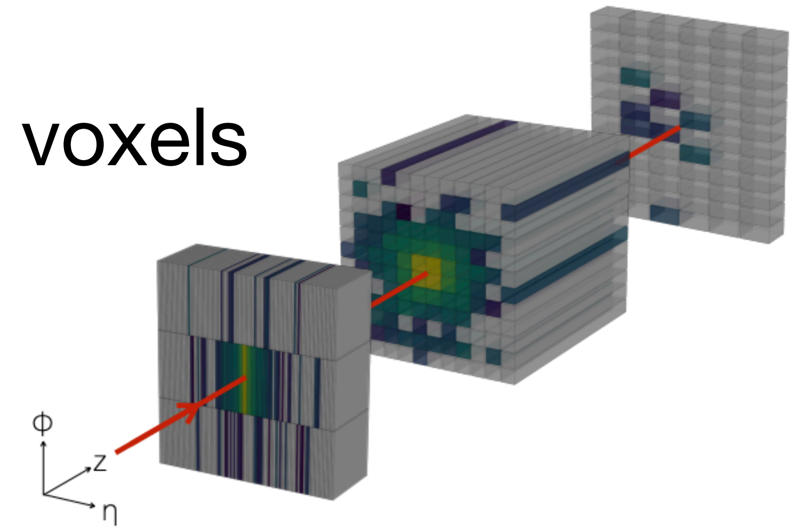
CaloGAN and CaloFlow

CaloFlow [Krause & DS (2021)]
First fast calosim using Normalizing Flows, major advance in quality



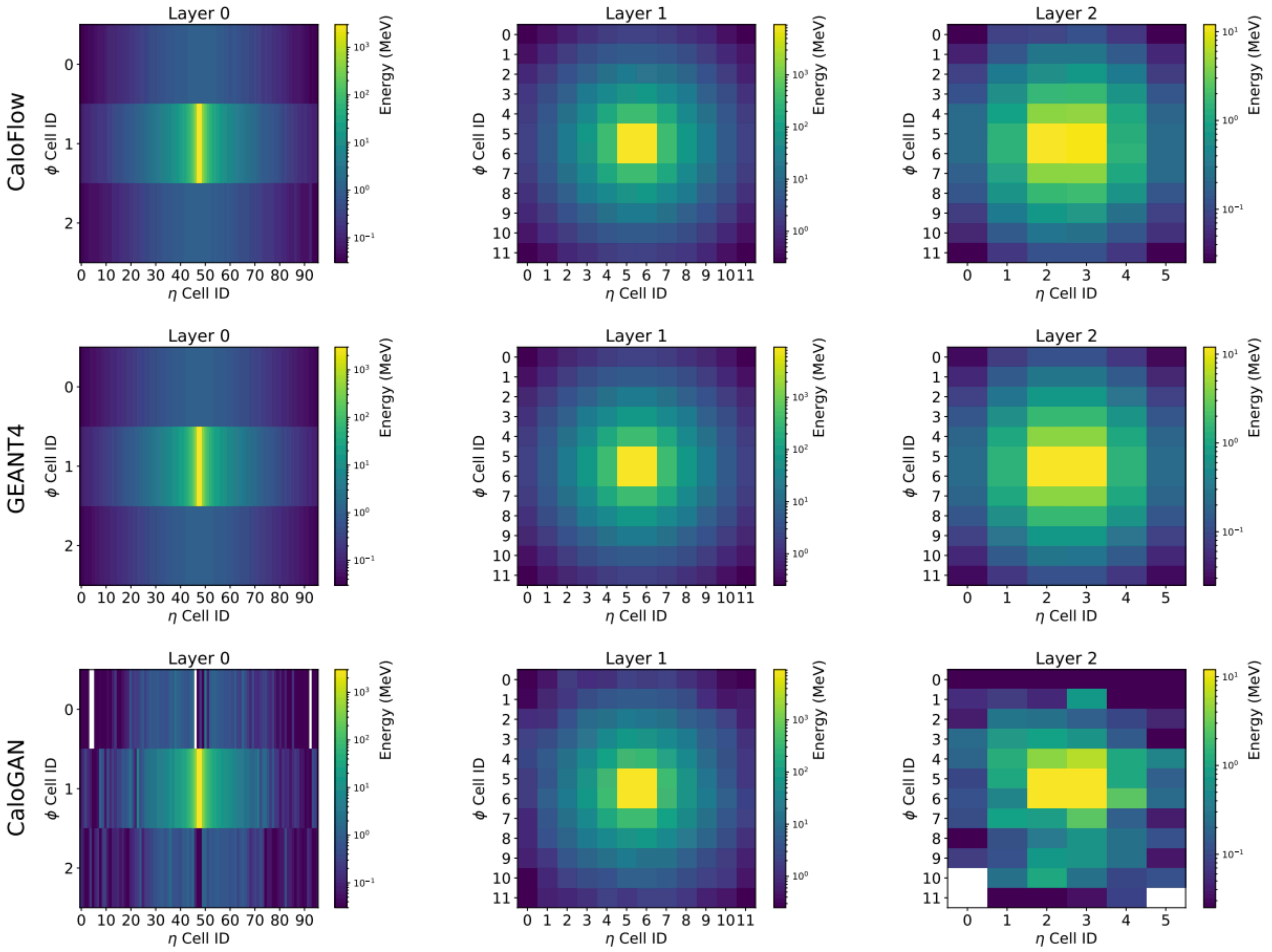
CaloGAN [Paganini, de Oliveira & Nachman (2017)]
First-ever deep surrogate model for fast calosim

Toy ATLAS ECAL — 3 layers, 504 voxels



CaloGAN and CaloFlow

CaloFlow [Krause & DS (2021)]
First fast calosim using Normalizing Flows, major advance in quality

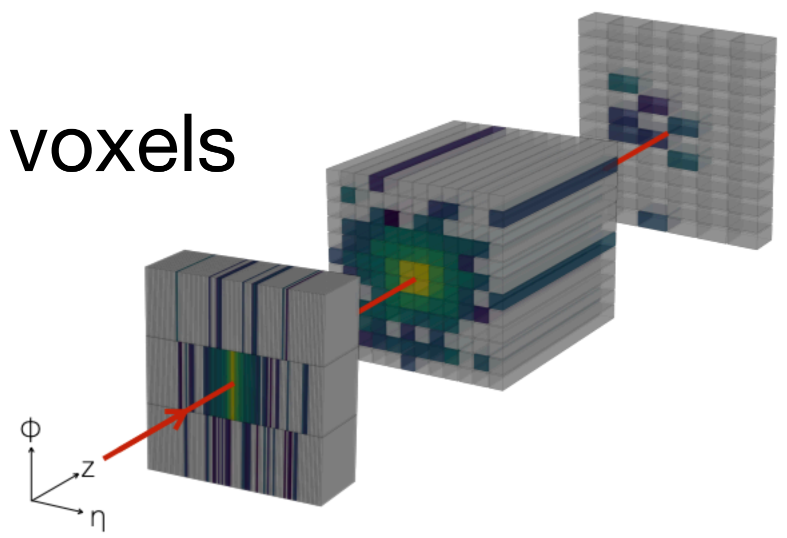


Also, first application of the classifier test to HEP generative modeling

AUC	GEANT4 vs. CaloGAN	GEANT4 vs. CaloFlow
e^+	1.000(0)	0.847(8)
γ	1.000(0)	0.660(6)
π^+	1.000(0)	0.632(2)

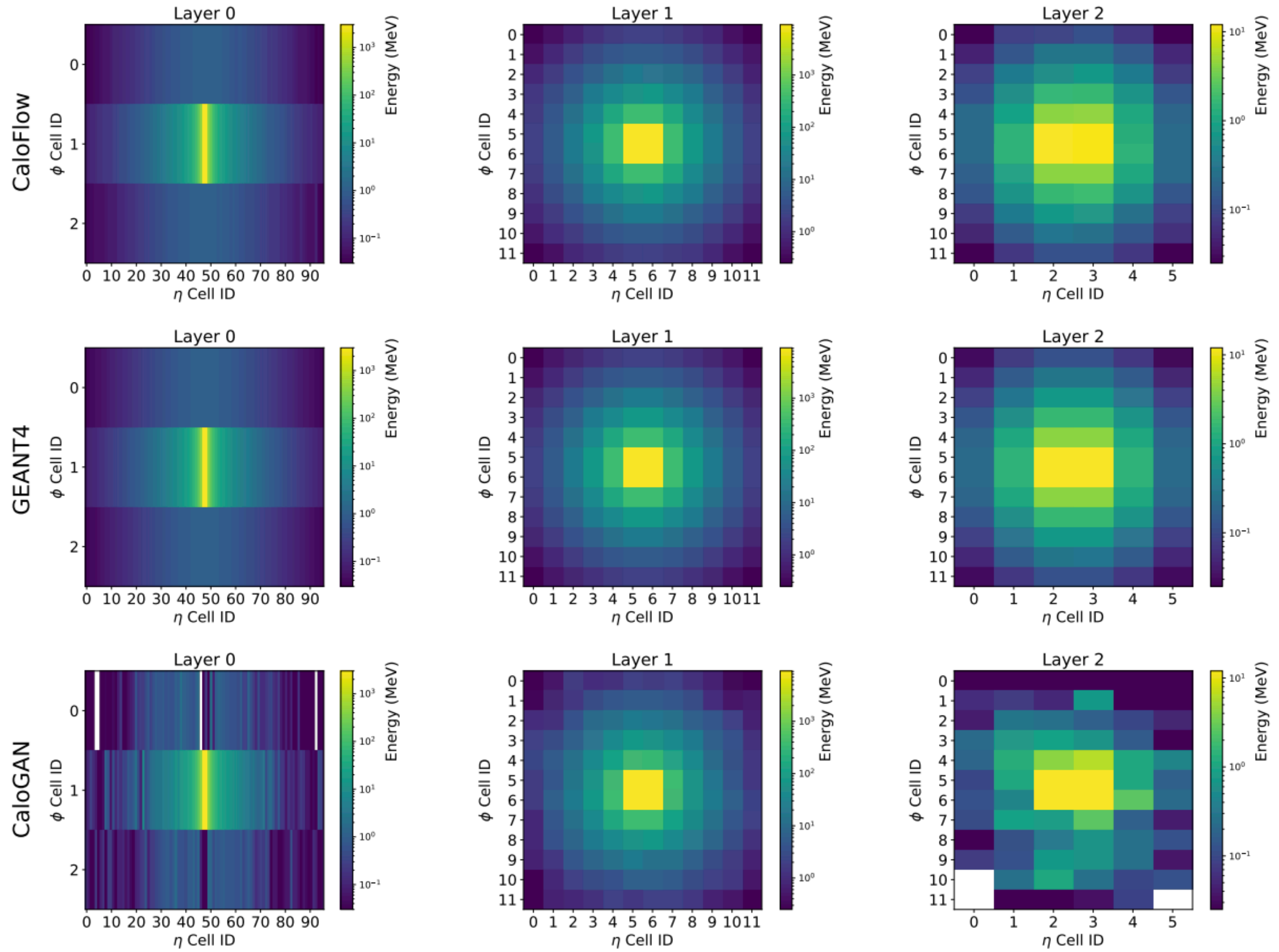
CaloGAN [Paganini, de Oliveira & Nachman (2017)]
First-ever deep surrogate model for fast calosim

Toy ATLAS ECAL — 3 layers, 504 voxels



CaloGAN and CaloFlow

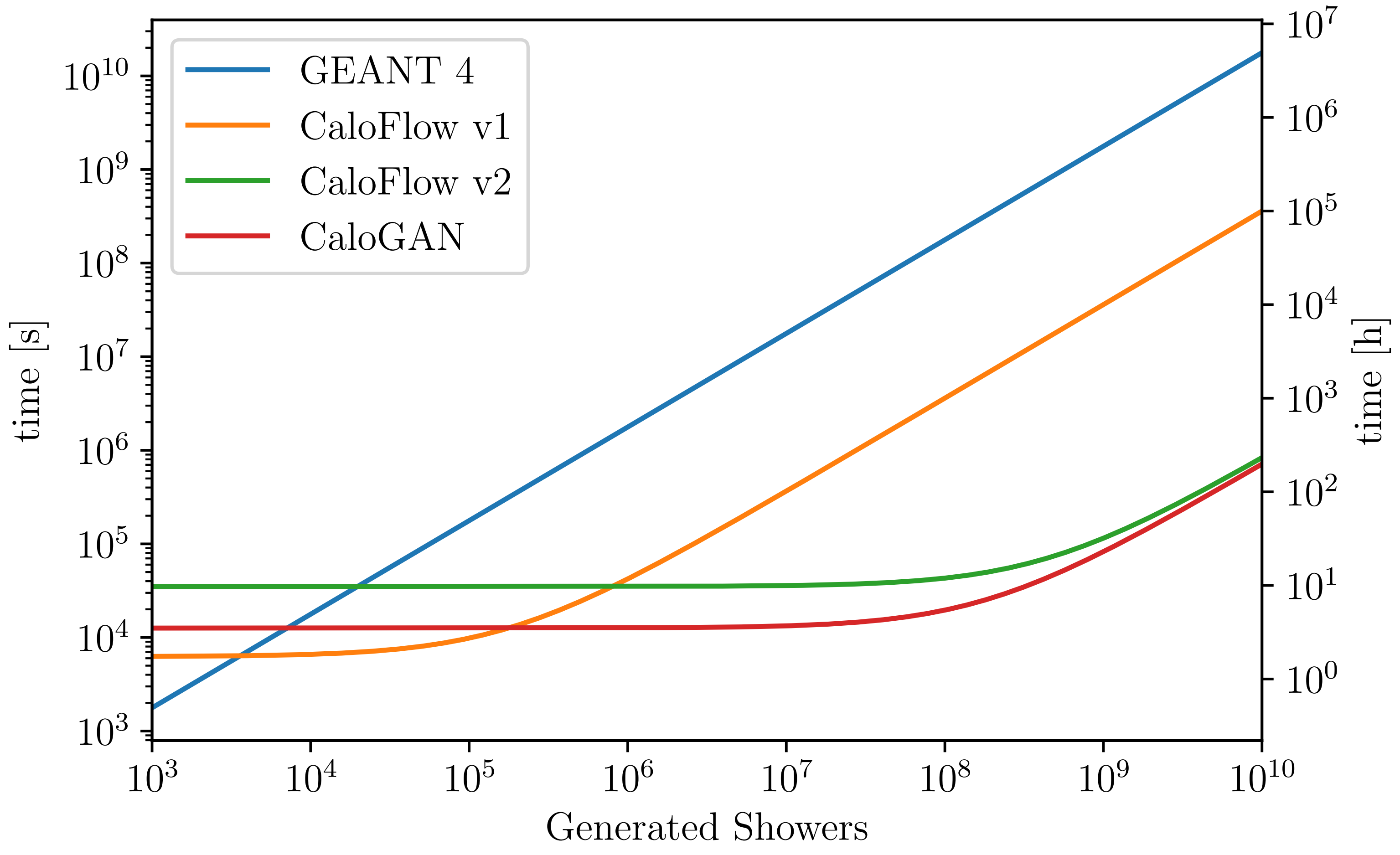
CaloFlow [Krause & DS (2021)]
First fast calosim using Normalizing Flows, major advance in quality



CaloGAN [Paganini, de Oliveira & Nachman (2017)]
First-ever deep surrogate model for fast calosim

Also, first application of the classifier test to HEP generative modeling

AUC	GEANT4 vs. CaloGAN	GEANT4 vs. CaloFlow
e^+	1.000(0)	0.847(8)
γ	1.000(0)	0.660(6)
π^+	1.000(0)	0.632(2)



Can speed up GEANT4 by up to ~10^5

CaloChallenge

Fast Calorimeter Simulation Challenge 2022

View on GitHub

Welcome to the home of the first-ever Fast Calorimeter Simulation Challenge!

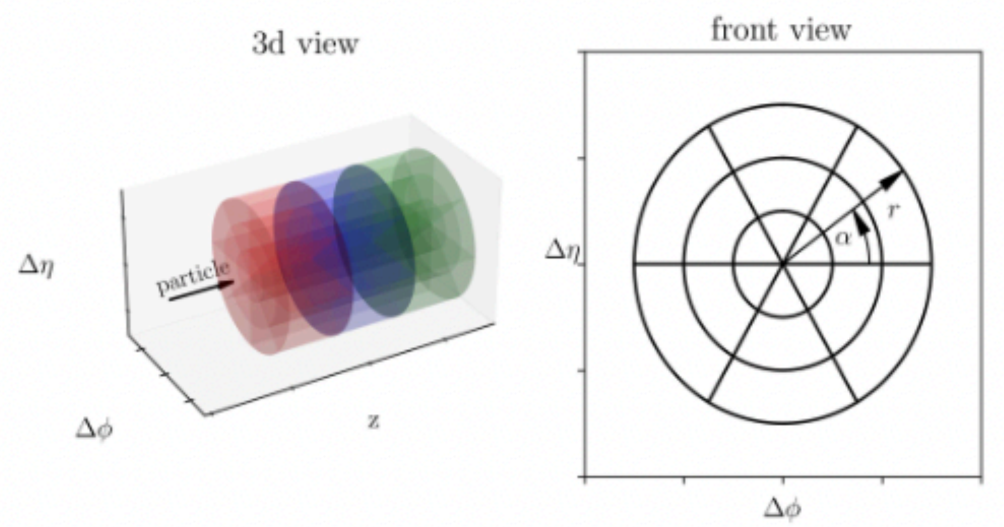
The purpose of this challenge is to spur the development and benchmarking of fast and high-fidelity calorimeter shower generation using deep learning methods. Currently, generating calorimeter showers of interacting particles (electrons, photons, pions, ...) using GEANT4 is a major computational bottleneck at the LHC, and it is forecast to overwhelm the computing budget of the LHC experiments in the near future. Therefore there is an urgent need to develop GEANT4 emulators that are both fast (computationally lightweight) and accurate. The LHC collaborations have been developing fast simulation methods for some time, and the hope of this challenge is to directly compare new deep learning approaches on common benchmarks. It is expected that participants will make use of cutting-edge techniques in generative modeling with deep learning, e.g. GANs, VAEs and normalizing flows.

This challenge is modeled after two previous, highly successful data challenges in HEP – the top tagging community challenge and the LHC Olympics 2020 anomaly detection challenge.

Datasets

The challenge offers three datasets, ranging in difficulty from “easy” to “medium” to “hard”. The difficulty is set by the dimensionality of the calorimeter showers (the number layers and the number of voxels in each layer).

Each dataset has the same general format. The detector geometry consists of concentric cylinders with particles propagating along the z-axis. The detector is segmented along the z-axis into discrete layers. Each layer has bins along the radial direction and some of them have bins in the angle α. The number of layers and the number of bins in r and α is stored in the binning .xml files and will be read out by the HighLevelFeatures class of helper functions. The coordinates Δφ and Δη correspond to the x- and y axis of the cylindrical coordinates. The image below shows a 3d view of a geometry with 3 layers, with each layer having 3 bins in radial and 6 bins in angular direction. The right image shows the front view of the geometry, as seen along the z axis.



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

arXiv > physics > arXiv:2410.21611

Search...
Help | Adva

Physics > Instrumentation and Detectors

[Submitted on 28 Oct 2024]

CaloChallenge 2022: A Community Challenge for Fast Calorimeter Simulation

Claudius Krause, Michele Faucci Giannelli, Gregor Kasieczka, Benjamin Nachman, Dalila Salamani, David Shih, Anna Zaborowska, Oz Amram, Kerstin Borras, Matthew R. Buckley, Erik Buhmann, Thorsten Buss, Renato Paulo Da Costa Cardoso, Anthony L. Caterini, Nadezda Chernyavskaya, Federico A.G. Corchia, Jesse C. Cresswell, Sascha Diefenbacher, Etienne Dreyer, Vijay Ekambaram, Engin Eren, Florian Ernst, Luigi Favaro, Matteo Franchini, Frank Gaede, Eilam Gross, Shih-Chieh Hsu, Kristina Jaruskova, Benno Käch, Jayant Kalagnanam, Raghav Kansal, Taewoo Kim, Dmitrii Kobylianskii, Anatolii Korol, William Korcari, Dirk Krücker, Katja Krüger, Marco Letizia, Shu Li, Qibin Liu, Xiulong Liu, Gabriel Loaiza-Ganem, Thandikire Madula, Peter McKeown, Isabell-A. Melzer-Pellmann, Vinicius Mikuni, Nam Nguyen, Ayodele Ore, Sofia Palacios Schweitzer, Ian Pang, Kevin Pedro, Tilman Plehn, Witold Pokorski, Huilin Qu, Piyush Raikwar, John A. Raine, Humberto Reyes-Gonzalez, Lorenzo Rinaldi, Brendan Leigh Ross, Moritz A.W. Scham, Simon Schnake, Chase Shimmin, Eli Shlizerman, Nathalie Soybelman, Mudhakar Srivatsa, Kalliopi Tsolaki, Sofia Vallecorsa, Kyongmin Yeo, Rui Zhang

We present the results of the "Fast Calorimeter Simulation Challenge 2022" - the CaloChallenge. We study state-of-the-art generative models on four calorimeter shower datasets of increasing dimensionality, ranging from a few hundred voxels to a few tens of thousand voxels. The 31 individual submissions span a wide range of current popular generative architectures, including Variational AutoEncoders (VAEs), Generative Adversarial Networks (GANs), Normalizing Flows, Diffusion models, and models based on Conditional Flow Matching. We compare all submissions in terms of quality of generated calorimeter showers, as well as shower generation time and model size. To assess the quality we use a broad range of different metrics including differences in 1-dimensional histograms of observables, KPD/FPD scores, AUCs of binary classifiers, and the log-posterior of a multiclass classifier. The results of the CaloChallenge provide the most complete and comprehensive survey of cutting-edge approaches to calorimeter fast simulation to date. In addition, our work provides a uniquely detailed perspective on the important problem of how to evaluate generative models. As such, the results presented here should be applicable for other domains that use generative AI and require fast and faithful generation of samples in a large phase space.

Comments: 204 pages, 100+ figures, 30+ tables

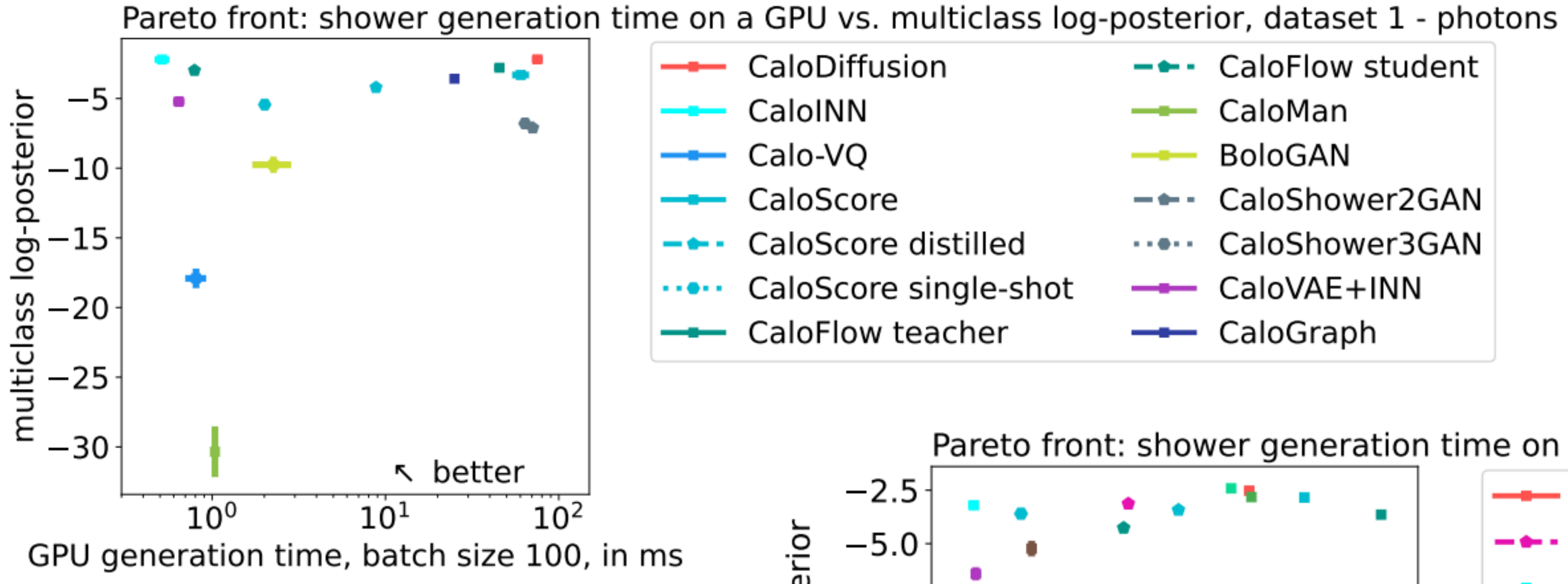
Completed data challenge for fast calorimeter simulation

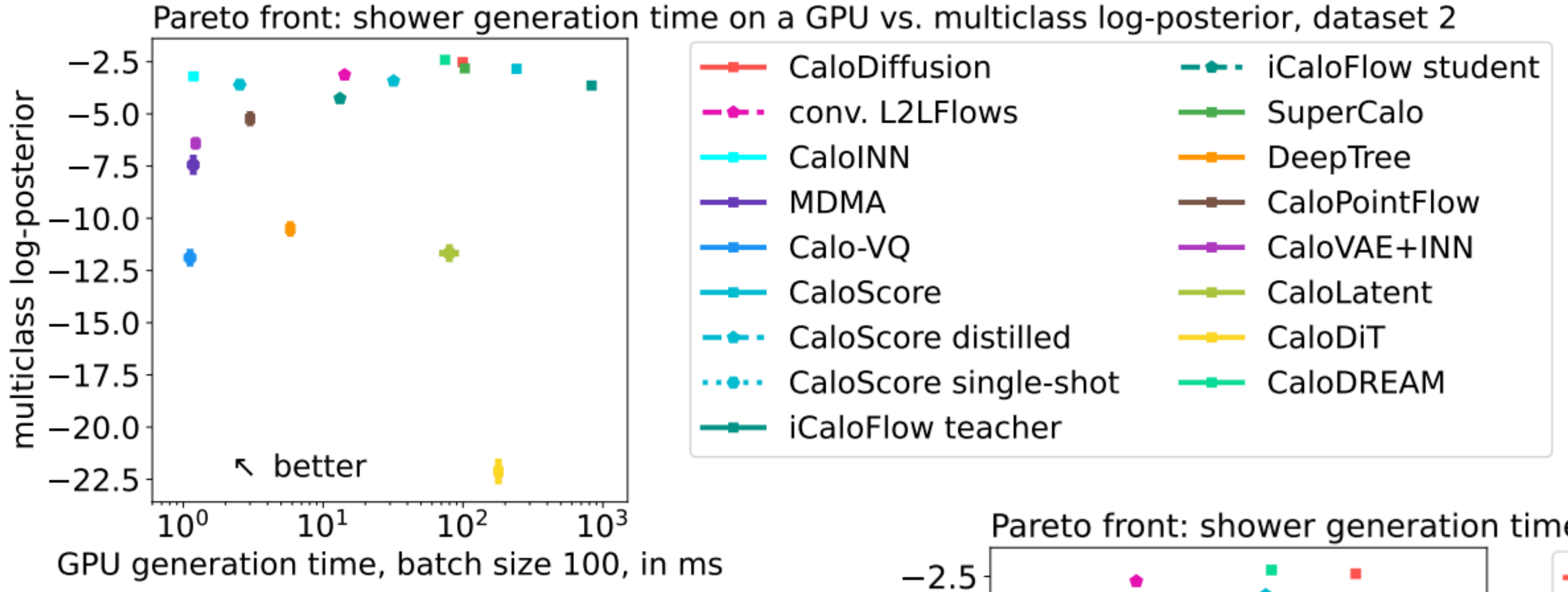
Organizers: Giannelli, Kasieczka, Krause, Nachman, Salamani, **DS**, Zaborowska

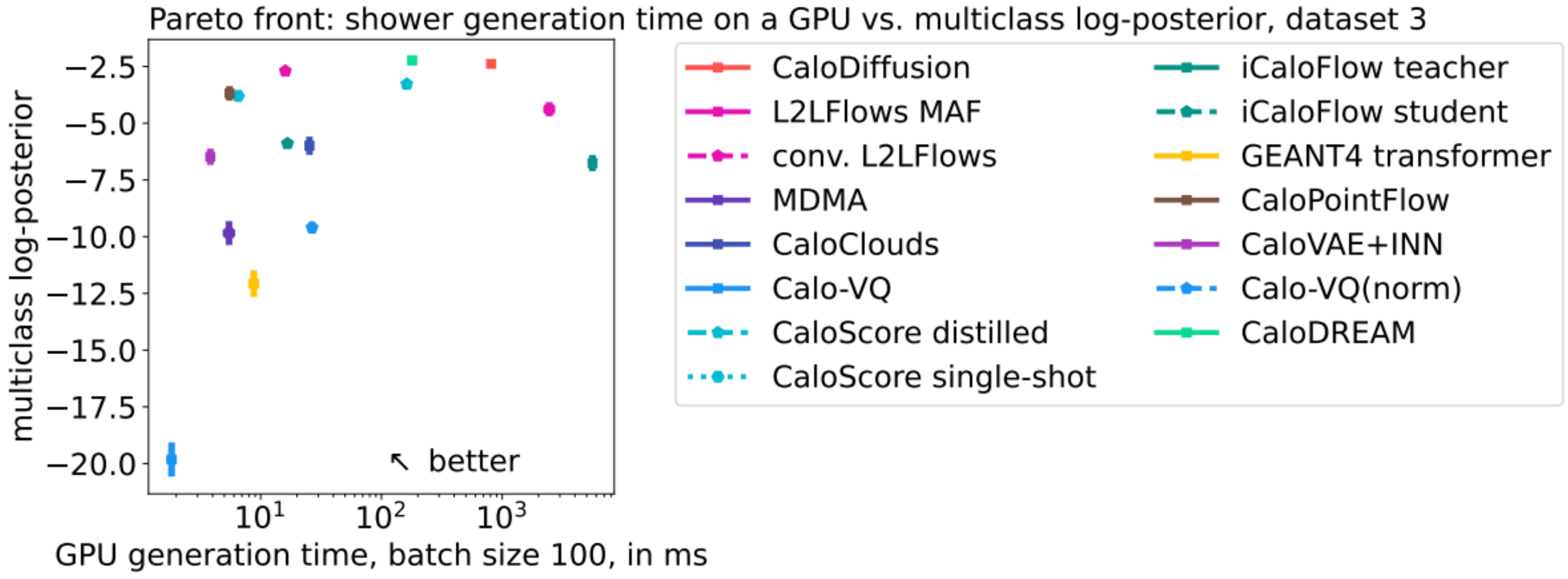
3 datasets:

- “easy” — official ATLAS CaloSim (~10^2 voxels)
- “medium” — GEANT4 example detector (~10^3 voxels)
- “hard” — GEANT4 example detector (~10^4 voxels)

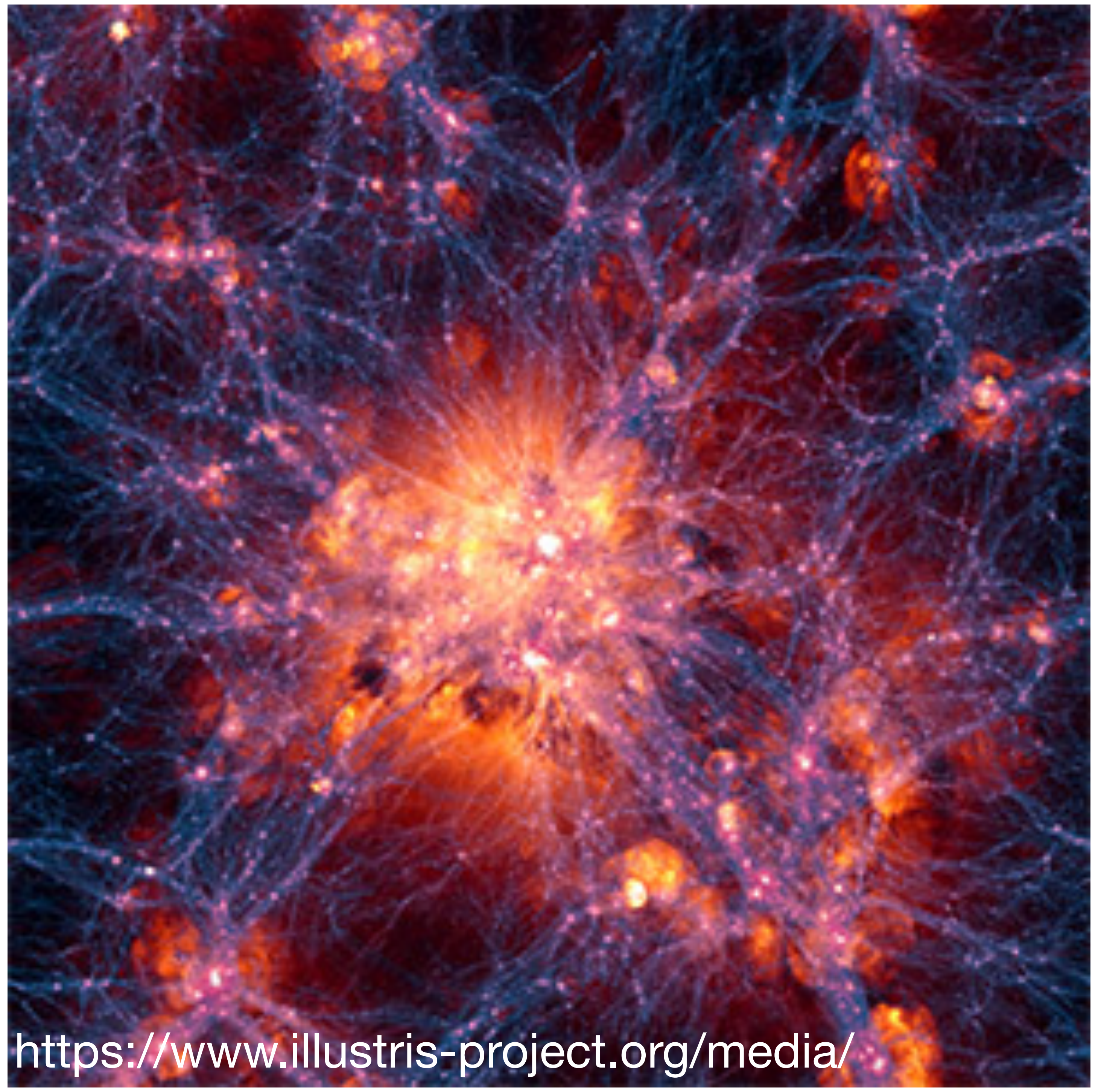
CaloChallenge — Pareto Fronts

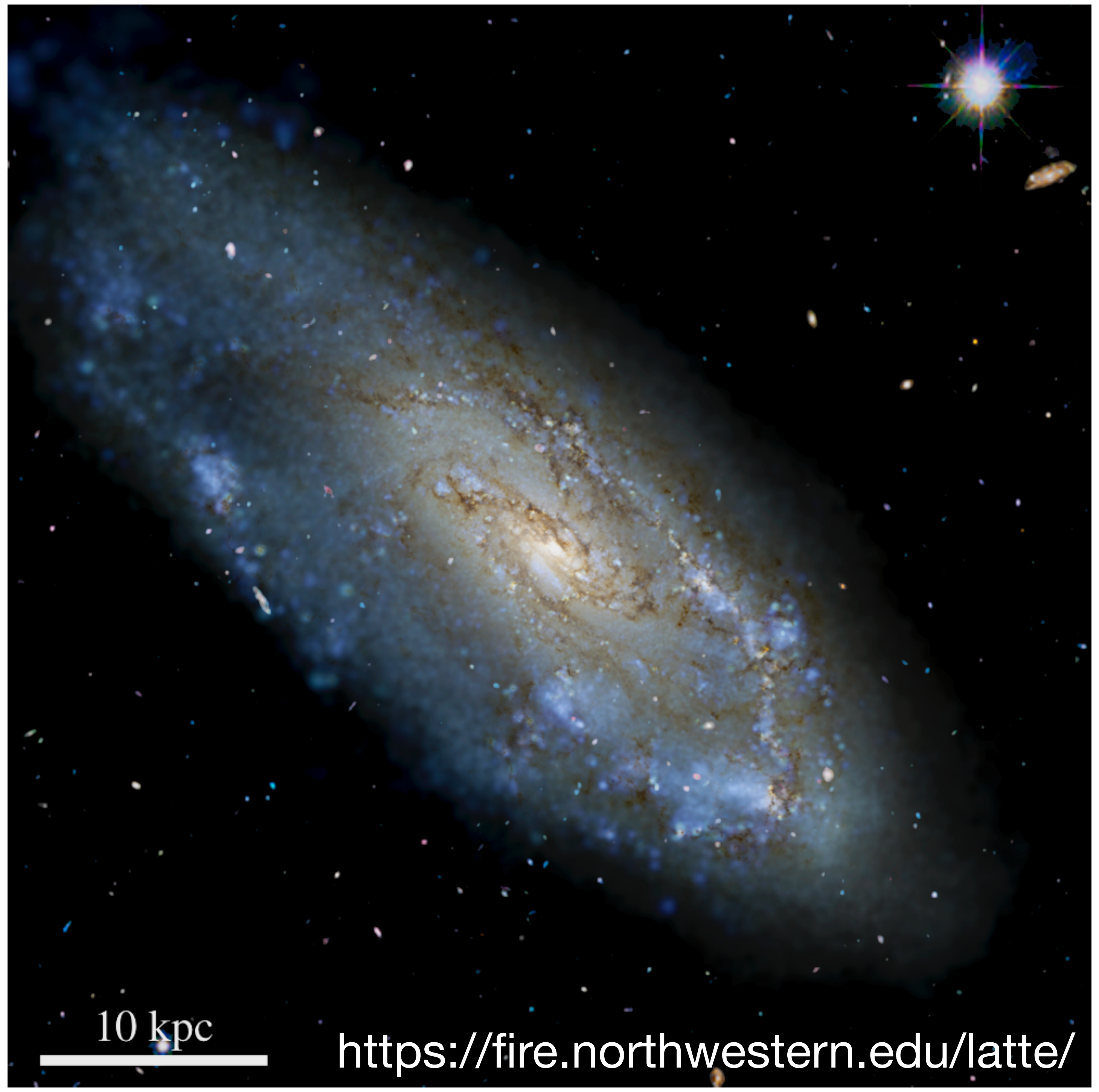






Upsampling Cosmological Simulations for Mock Star Catalogs





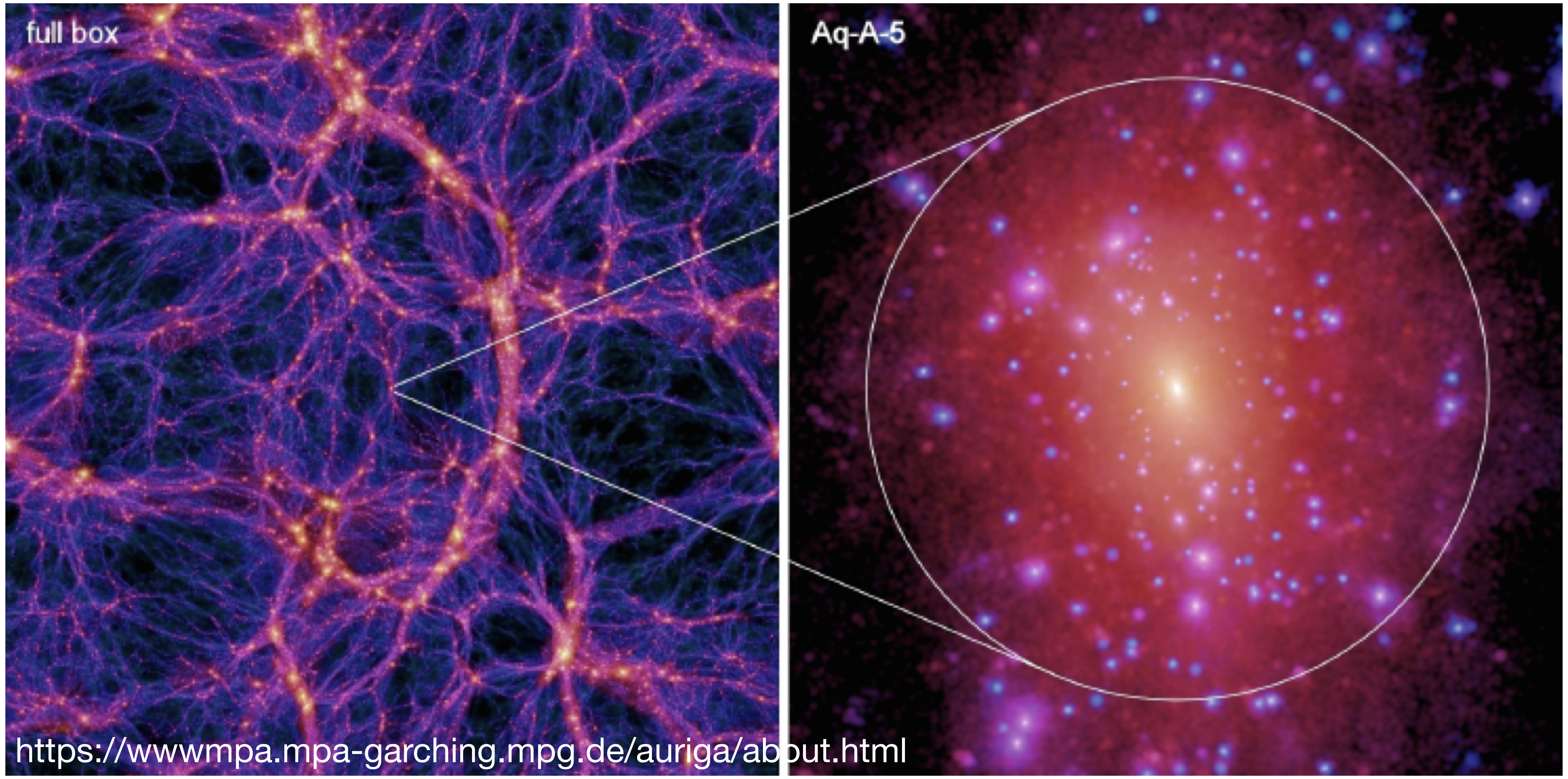
- Cosmological simulations of DM, stars and gas are increasingly sophisticated and realistic

- They are extremely computationally expensive, taking weeks-months and enormous supercomputing resources to run

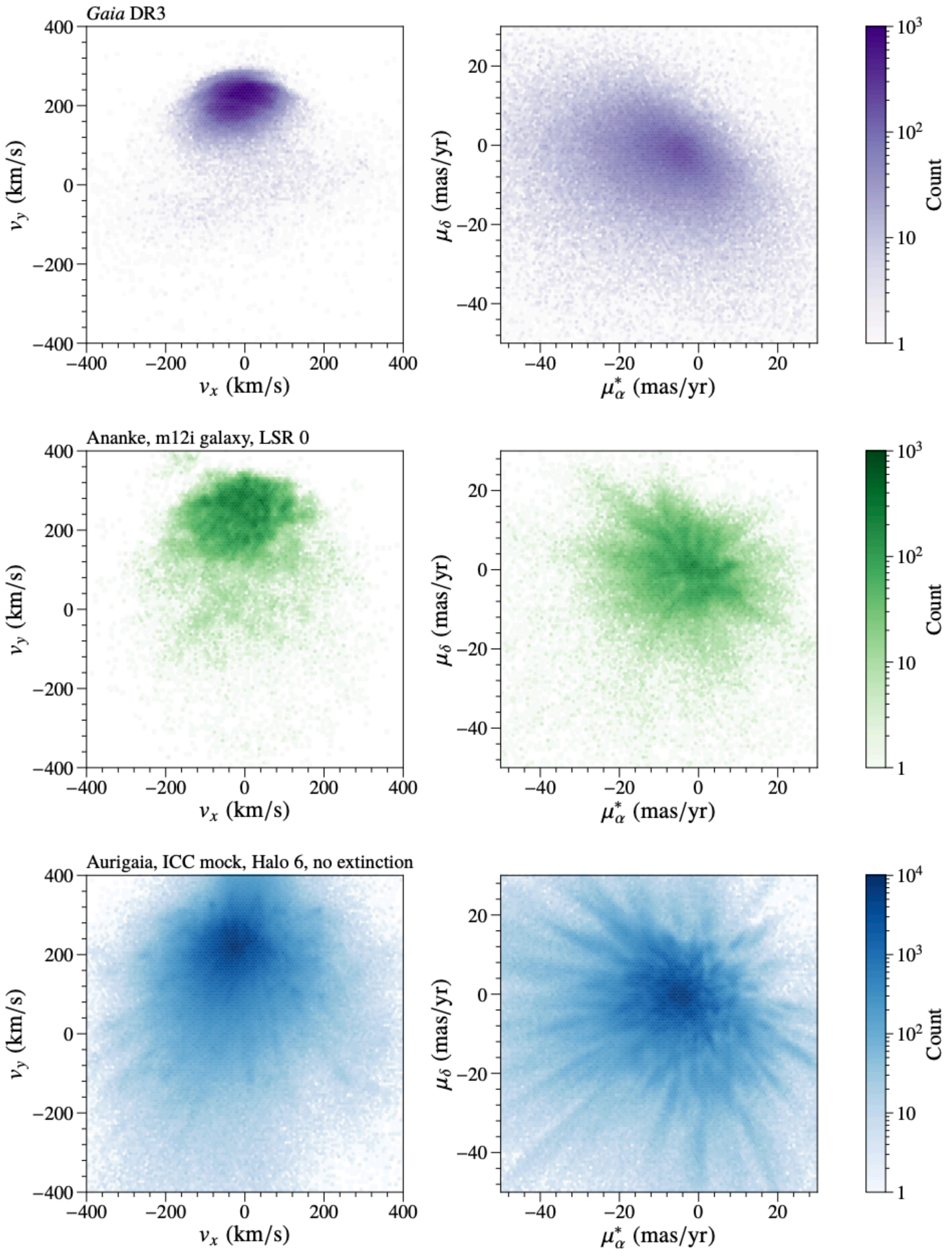
Upsampling Cosmological Simulations for Mock Star Catalogs

- However, turning these cosmological simulations into mock star catalogs (eg for Gaia) is not automatic: the simulations produce "star particles" that each represent 10^4 stars

- Need to "upsample" (superresolve) cosmological simulations to turn star particles into stars



Upsampling Cosmological Simulations for Mock Star Catalogs



- Previous attempts at upsampling used a crude form of density estimation (EnBiD) + Gaussian kernel sampling
- Led to a “blotchy” star catalog!
- Perfect opportunity for deep generative models

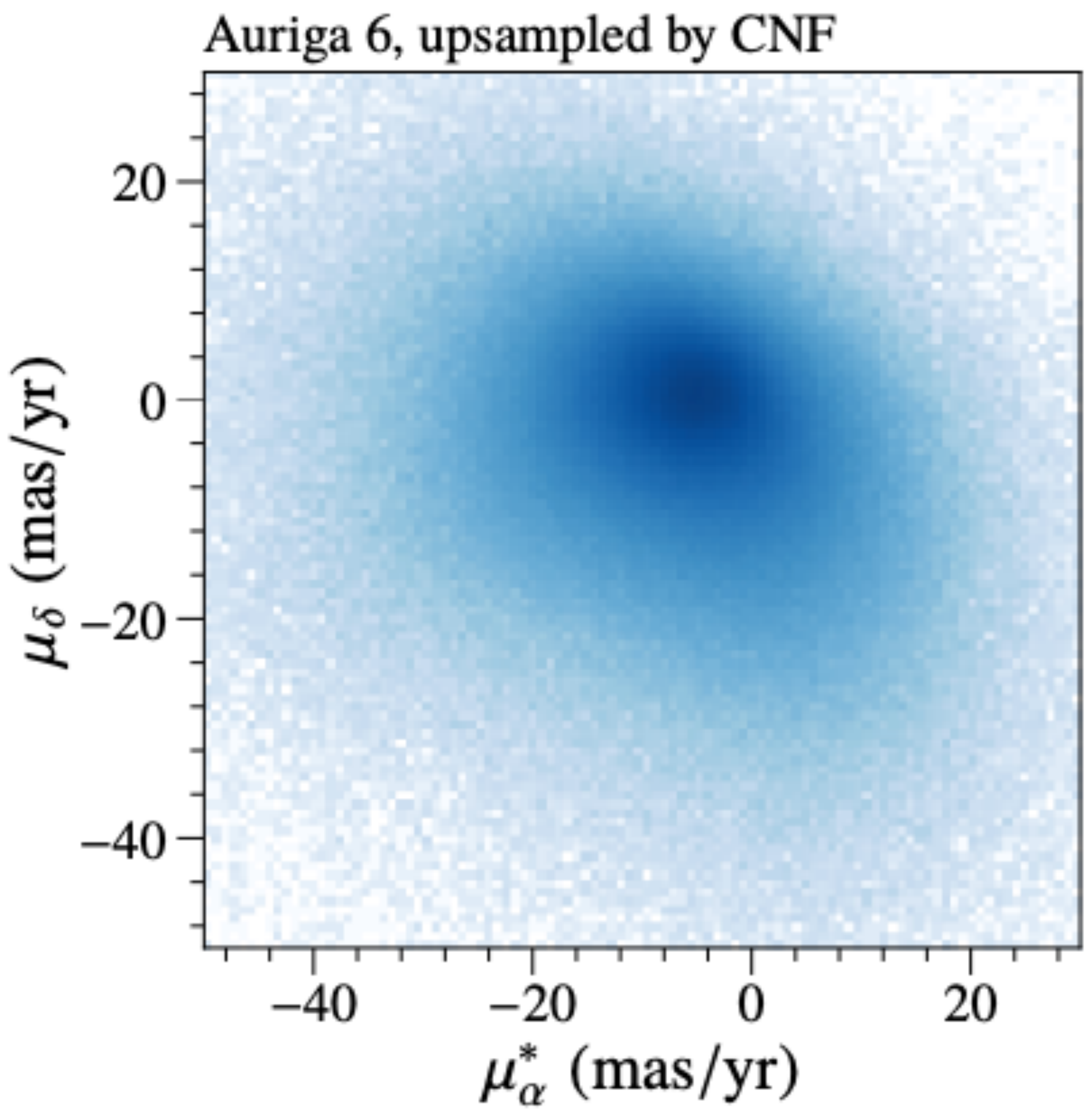
$$x \sim p_{gen}(x) \approx p_{sim}(x)$$

Should be much smoother than Gaussian kernel sampling!

GalaxyFlow

Lim, Raman, Buckley & DS 2211.11765

Trained a continuous normalizing flow on a hydrodynamical cosmological simulation from the Auriga collaboration.



Classifier:	EnBiD vs. CNF
Upsampler	LP(upsampler)
EnBiD	-0.724 ± 0.003
CNF	$\mathbf{-0.683 \pm 0.003}$

The result is much smoother and agrees better with the reference star particle distribution!