New avenues for ML in HEP

or

"Places where ML could have a big impact, but where it has not been widely used traditionally"

Snowmass Community Summer Study 2022

Computational Frontier Colloquium on AI/ML

David Shih
July 18, 2022

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Modern Machine Learning

Modern Machine Learning

Modern machine learning is not physics.

Modern Machine Learning

Modern machine learning is not physics.

Rather, it is a powerful new tool which will enable us to do new kinds of physics that we couldn't do before.

Modern Machine Learning

Modern machine learning is not physics.

Rather, it is a powerful new tool which will enable us to do new kinds of physics that we couldn't do before.

- Calculus => Classical mechanics

Modern Machine Learning

Modern machine learning is not physics.

Rather, it is a powerful new tool which will enable us to do new kinds of physics that we couldn’t do before.

- Calculus => Classical mechanics
- Linear algebra => Quantum mechanics

Modern Machine Learning

Modern machine learning is not physics.

Rather, it is a powerful new tool which will enable us to do new kinds of physics that we couldn't do before.

- Calculus => Classical mechanics
- Linear algebra => Quantum mechanics
- Group theory => Standard Model

Modern Machine Learning

Modern machine learning is not physics.

Rather, it is a powerful new tool which will enable us to do new kinds of physics that we couldn't do before.

- Calculus => Classical mechanics
- Linear algebra => Quantum mechanics
- Group theory => Standard Model
- Statistics => Experimental Design

Modern Machine Learning

Modern machine learning is not physics.

Rather, it is a powerful new tool which will enable us to do new kinds of physics that we couldn't do before.

- Calculus => Classical mechanics
- Linear algebra => Quantum mechanics
- Group theory => Standard Model
- Statistics => Experimental Design
- Modern Machine Learning => ?

Modern Machine Learning

Modern machine learning is not physics.

Rather, it is a powerful new tool which will enable us to do new kinds of physics that we couldn't do before.

- Calculus => Classical mechanics
- Linear algebra => Quantum mechanics
- Group theory => Standard Model
- Statistics => Experimental Design
- Modern Machine Learning => ?

What new fields and discoveries await?

Modern Machine Learning

How will modern ML enable new kinds of physics?

With modern ML, we can extract more information from data than ever before.

Data: "events" $x_i \in R^d$ drawn iid from some distribution $p(x)$

Modern Machine Learning

How will modern ML enable new kinds of physics?

With modern ML, we can extract more information from data than ever before.

Data: "events" $x_i \in R^d$ drawn iid from some distribution $\boxed{p(x)}$

- **All the information contained in the data is contained in $p(x)$.**

Modern Machine Learning

How will modern ML enable new kinds of physics?

With modern ML, we can extract more information from data than ever before.

Data: "events" $x_i \in R^d$ drawn iid from some distribution $\boxed{p(x)}$

- **All the information contained in the data is contained in $p(x)$.**
- Generally, the underlying $p(x)$ of the data is unknown.

Modern Machine Learning

How will modern ML enable new kinds of physics?

With modern ML, we can extract more information from data than ever before.

Data: "events" $x_i \in R^d$ drawn iid from some distribution $\boxed{p(x)}$

- **All the information contained in the data is contained in $p(x)$.**
- Generally, the underlying $p(x)$ of the data is unknown.
- Modern ML can access $p(x)$ (explicitly or implicitly) from data, even for very high dimensional *x!*

Modern Machine Learning

In what ways can modern ML access the full likelihood of the data?

- $p(x)$ **itself** [density estimation, eg Normalizing Flows]
- **conditional** densities $p(x \mid y)$ [conditional density estimation, also NFs]
- **sampling** from $p(x)$ [generative modeling, eg GANs, VAEs, NFs]
- **ratios** of densities $p_1(x)/p_2(x)$ [classification, eg CNNs, RNNs, transformers, GNNs, …]
- ….

Modern Machine Learning

Modern Machine Learning will enable us to extract much more **physics** from data than ever before

Data → Modern Machine Learning → **Physics**

- Opens up entirely new frontiers in data analysis
- Qualitatively new kinds of physics analyses that weren't possible before
- A Golden Era of **method development, proofs-of-concept** and **new results**

Modern Machine Learning

Modern Machine Learning will enable us to extract much more **physics** from data than ever before

Data → Modern Machine Learning →

New physics searches

Triggering

Fast simulation

Instrumentation

Measurement

Theory

...

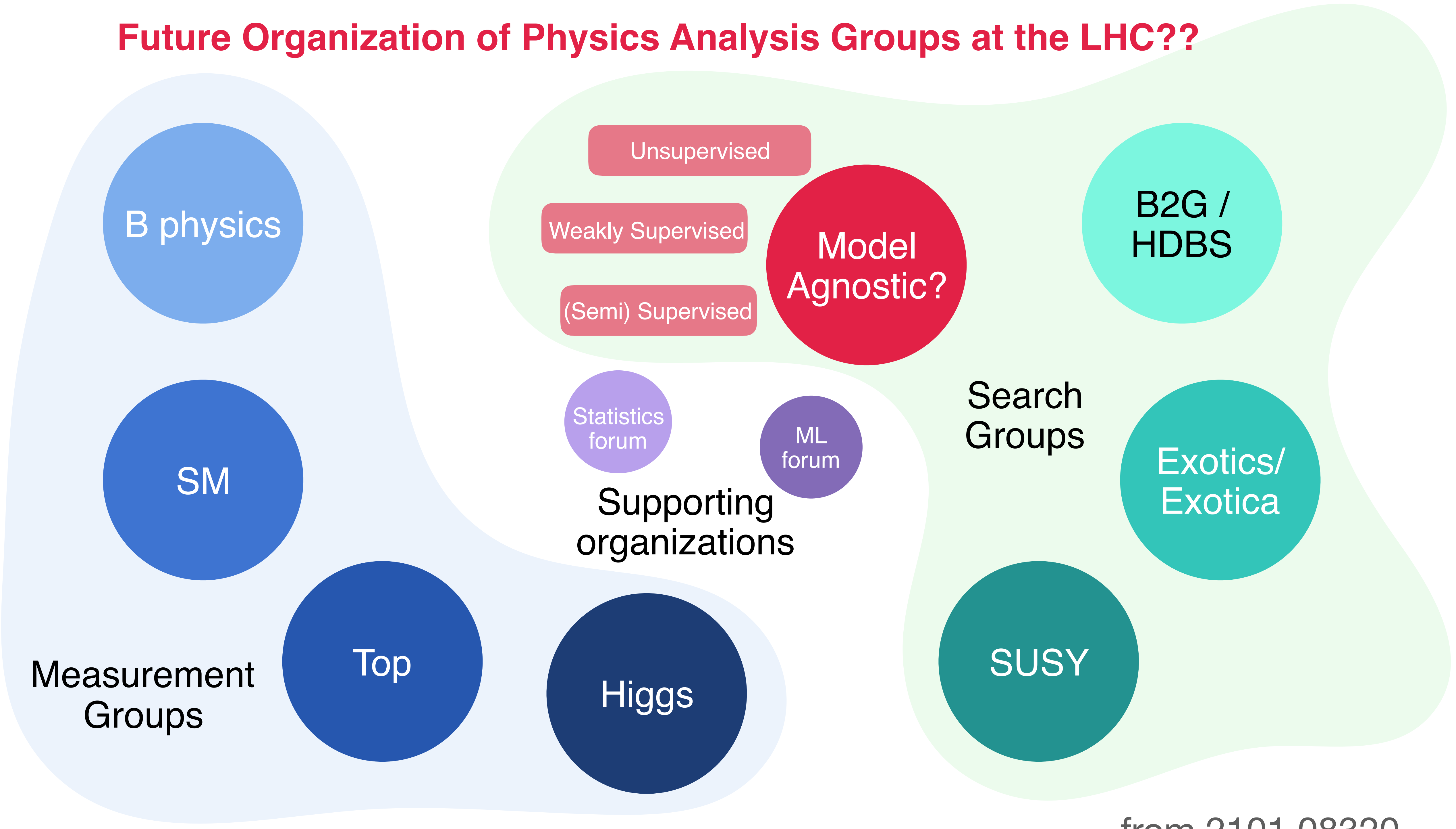
Apologies in advance if I can't cover everything in this talk!!

ML for New Physics Searches

ML for New Physics Searches

The vast majority of LHC searches for new physics are very model specific

ML for New Physics Searches



from 2101.08320

The vast majority of LHC searches for new physics are very model specific

Why aren't there more model-agnostic new physics searches?

ML for New Physics Searches

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

Gregor Kasieczka (ed),[1] Benjamin Nachman (ed),[2,3] David Shih (ed),[4] Oz Amram,[5] Anders Andreassen,[6] Kees Benkendorfer,[2,7] Blaz Bortolato,[8] Gustaaf Brooijmans,[9] Florencia Canelli,[10] Jack H. Collins,[11] Biwei Dai,[12] Felipe F. De Freitas,[13] Barry M. Dillon,[8,14] Ioan-Mihail Dinu,[5] Zhongtian Dong,[15] Julien Donini,[16] Javier Duarte,[17] D. A. Faroughy[10] Julia Gonski,[9] Philip Harris,[18] Alan Kahn,[9] Jernej F. Kamenik,[8,19] Charanjit K. Khosa,[20,30] Patrick Komiske,[21] Luc Le Pottier,[2,22] Pablo Martín-Ramiro,[2,23] Andrej Matevc,[8,19] Eric Metodiev,[21] Vinicius Mikuni,[10] Inês Ochoa,[24] Sang Eon Park,[18] Maurizio Pierini,[25] Dylan Rankin,[18] Veronica Sanz,[20,26] Nilai Sarda,[27] Uroš Seljak,[2,3,12] Aleks Smolkovic,[8] George Stein,[2,12] Cristina Mantilla Suarez,[5] Manuel Szewc,[28] Jesse Thaler,[21] Steven Tsan,[17] Silviu-Marian Udrescu,[18] Louis Vaslin,[16] Jean-Roch Vlimant,[29] Daniel Williams,[9] Mikaeel Yunus[18]

https://arxiv.org/abs/2101.08320

https://arxiv.org/abs/2105.14027

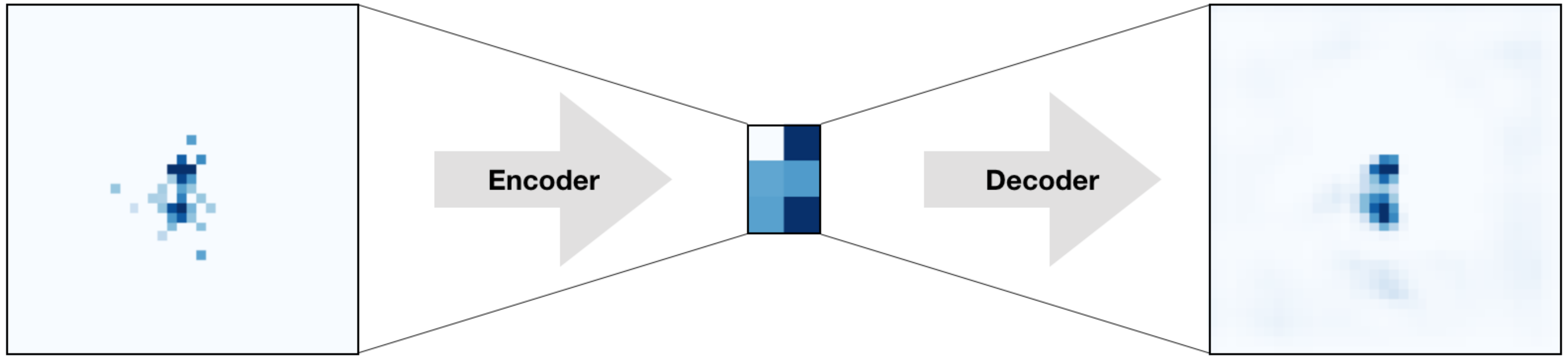
The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider

T. Aarrestad[a] M. van Beekveld[b] M. Bona[c] A. Boveia[e] S. Caron[d] J. Davies[c] A. De Simone[f,g] C. Doglioni[h] J. M. Duarte[i] A. Farbin[j] H. Gupta[k] L. Hendriks[d] L. Heinrich[a] J. Howarth[l] P. Jawahar[m,a] A. Jueid[n] J. Lastow[h] A. Leinweber[o] J. Mamuzic[p] E. Merényi[q] A. Morandini[r] P. Moskvitina[d] C. Nellist[d] J. Ngadiuba[s,t] B. Ostdiek[u,v] M. Pierini[a] B. Ravina[l] R. Ruiz de Austri[p] S. Sekmen[w] M. Touranakou[x,a] M. Vaškevičiūte[l] R. Vilalta[y] J.-R. Vlimant[t] R. Verheyen[z] M. White[o] E. Wulff[h] E. Wallin[h] K.A. Wozniak[α,a] Z. Zhang[d]

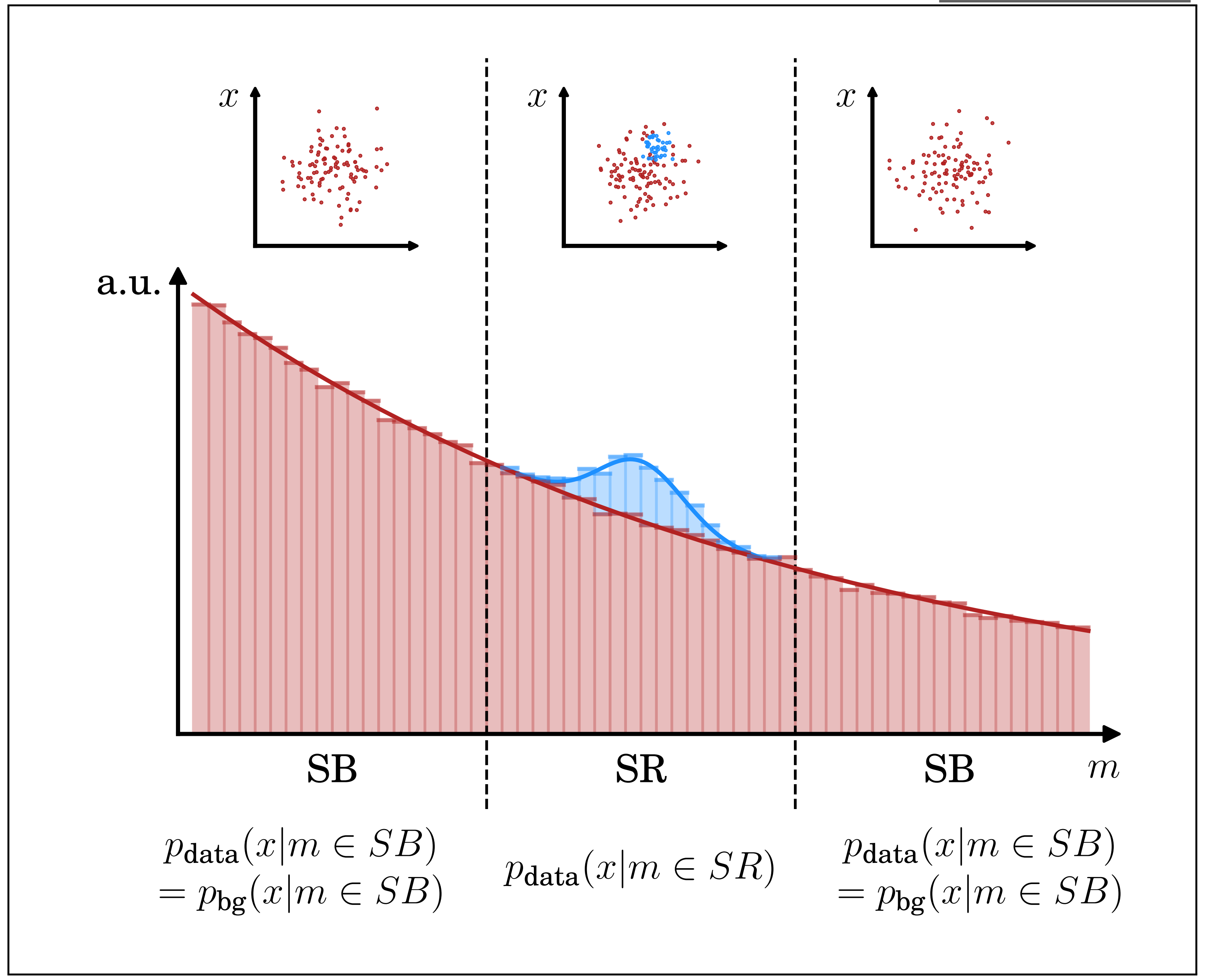
A lot of community interest in model-agnostic NP searches!

Both theorists and experimentalists are proposing many new approaches using modern ML

ML for New Physics Searches



from 1808.08992



from 2109.00546

Autoencoders

Fully unsupervised

Sensitive to outliers (low $p(x)$)

Farina, Nakai & **DS** 1808.08992

Heimel et al 1808.08979

and many more!!

Enhanced bump hunts

Weakly supervised

Sensitive to overdensities (high $p_{data}(x)/p_{bg}(x)$)

CWoLa Hunting [Collins, Howe & Nachman 1805.02664, 1902.02634]

ANODE [Nachman & **DS** 2001.04990]

CATHODE [Hallin et al 2109.00546]

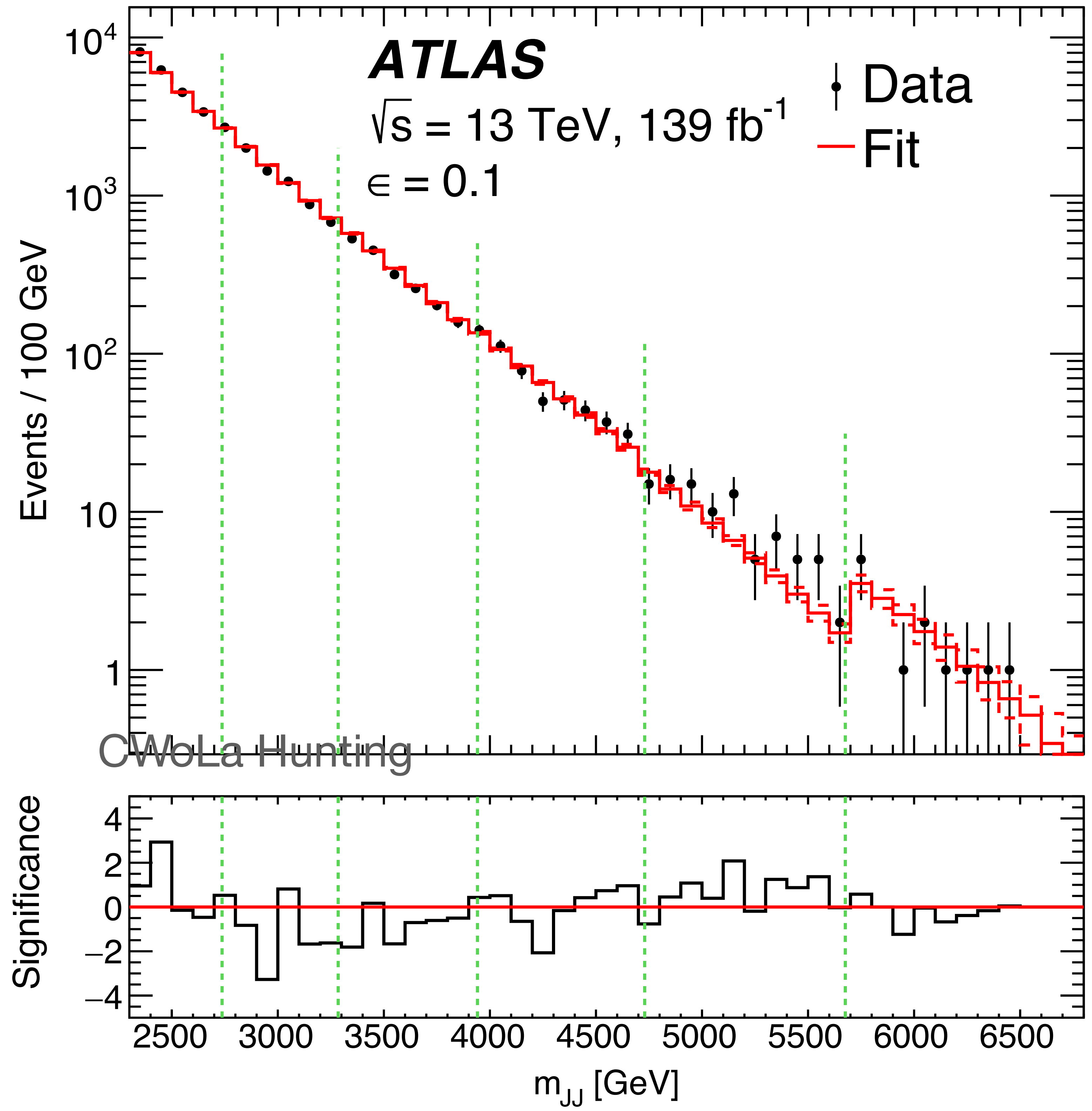
CURTAINS [Raine et al 2203.09470]

and more…

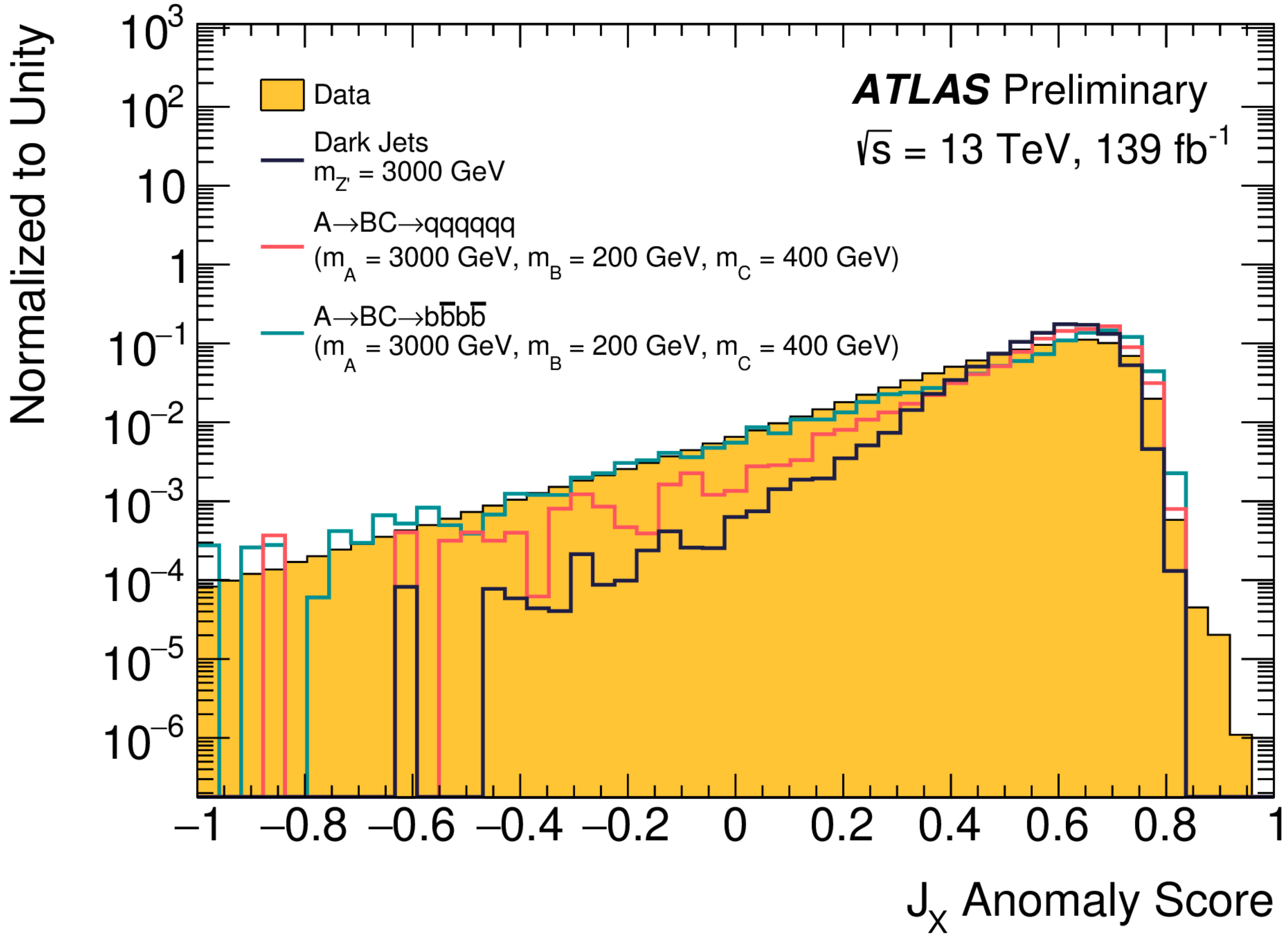
ML for New Physics Searches

Proofs-of-concept are becoming actual LHC searches!

CWoLa Hunting
ATLAS, PRL **125** 131801 (2020)



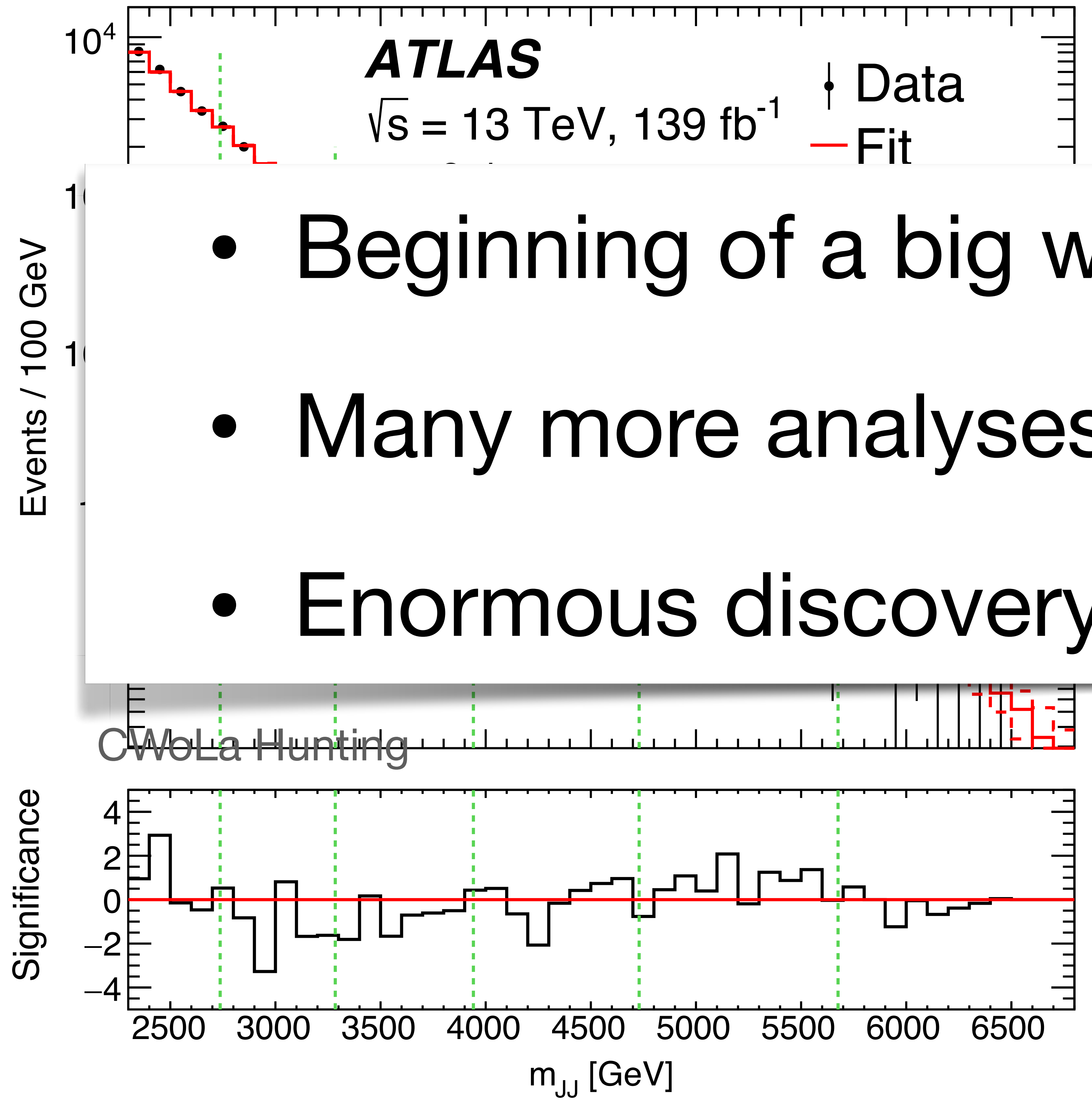
RNN VAE
ATLAS-CONF-2022-045



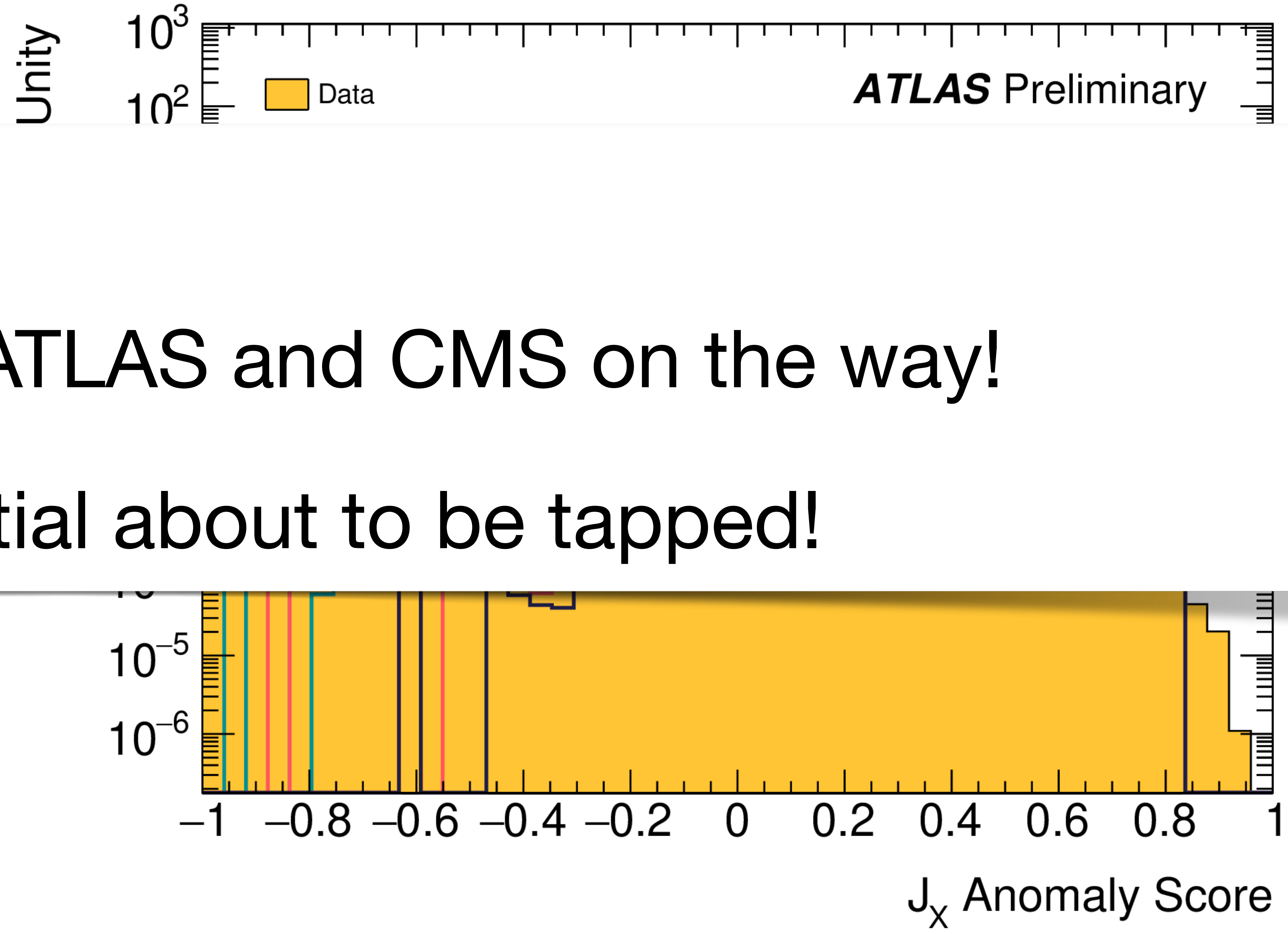
ML for New Physics Searches

Proofs-of-concept are becoming actual LHC searches!

CWoLa Hunting
ATLAS, PRL **125** 131801 (2020)



RNN VAE
ATLAS-CONF-2022-045

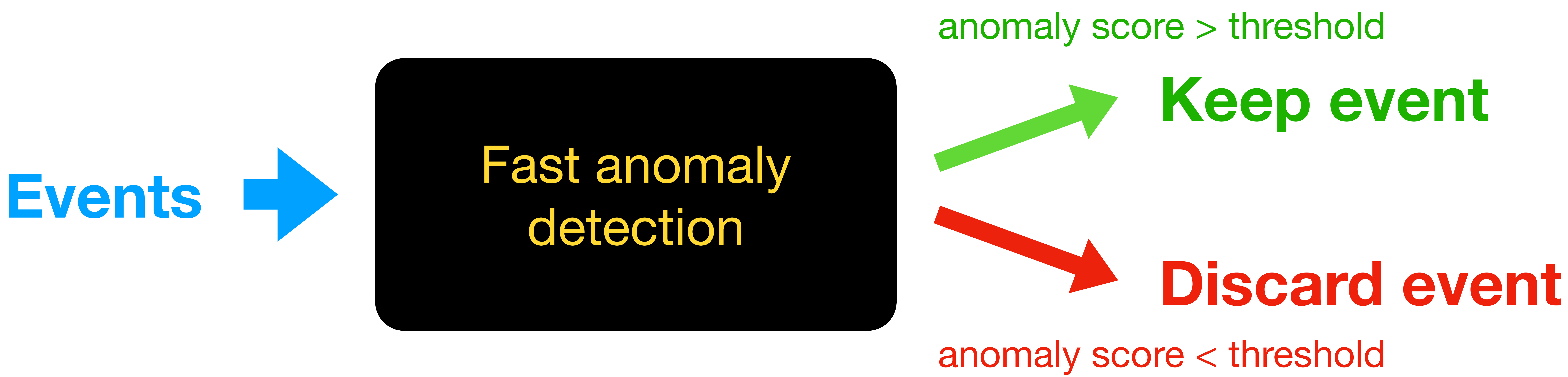


- Beginning of a big wave?
- Many more analyses from ATLAS and CMS on the way!
- Enormous discovery potential about to be tapped!

Fast ML for
Online Anomaly Detection

Fast ML for Online Anomaly Detection

- ML for triggers and DAQ used since the 90s (CDF, H1); widely used at LHC at both L1 and HLT
- New avenue with modern ML: **anomaly detection at trigger level**



Fast ML for Online Anomaly Detection

- ML for triggers and DAQ used since the 90s (CDF, H1); widely used at LHC at both L1 and HLT
- New avenue with modern ML: **anomaly detection at trigger level**
    - Autoencoders for anomaly detection at trigger level [Cerri et al 1811.10276, Knapp et al 2005.01598, Dillon et al 2206.14225, …]
    - Autoencoders on FPGAs for L1T [Govorkova et al. 2108.03986]
    - **Double Decorrelated Autoencoders** for anomaly detection **and** background estimation at trigger level [Mikuni, Nachman & **DS** 2111.06417]

Fast ML for Online Anomaly Detection

Unsupervised New Physics detection at 40 MHz

In this challenge, you will develop algorithms for detecting New Physics by reformulating the problem as an out-of-distribution detection task. Armed with four-vectors of the highest-momentum jets, electrons, and muons produced in a LHC collision event, together with the missing transverse energy (missing E_T), the goal is to find a-priori unknown and rare New Physics hidden in a data sample dominated by ordinary Standard Model processes, using anomaly detection approaches.

Real-time event filtering

The algorithms are intended to be deployed in the first stage of the real-time event filter processing system of LHC experiments (Level 1 or L1 trigger), where the available bandwidth, latency and resources are strictly limited. Such limitations constrain the design of the algorithm. To emulate the constraints in terms of bandwith only the leading 10 jets, 4 muons, 4 electrons and the missing E_T will be provided to be used as input to the algorithm. Furthermore, only a maximum number of bits is available for the representation of the η, ϕ, and the transverse momentum p_T of each physics object. The effect of such *quantization* of the inputs can be studied for instance with QKeras (see below).

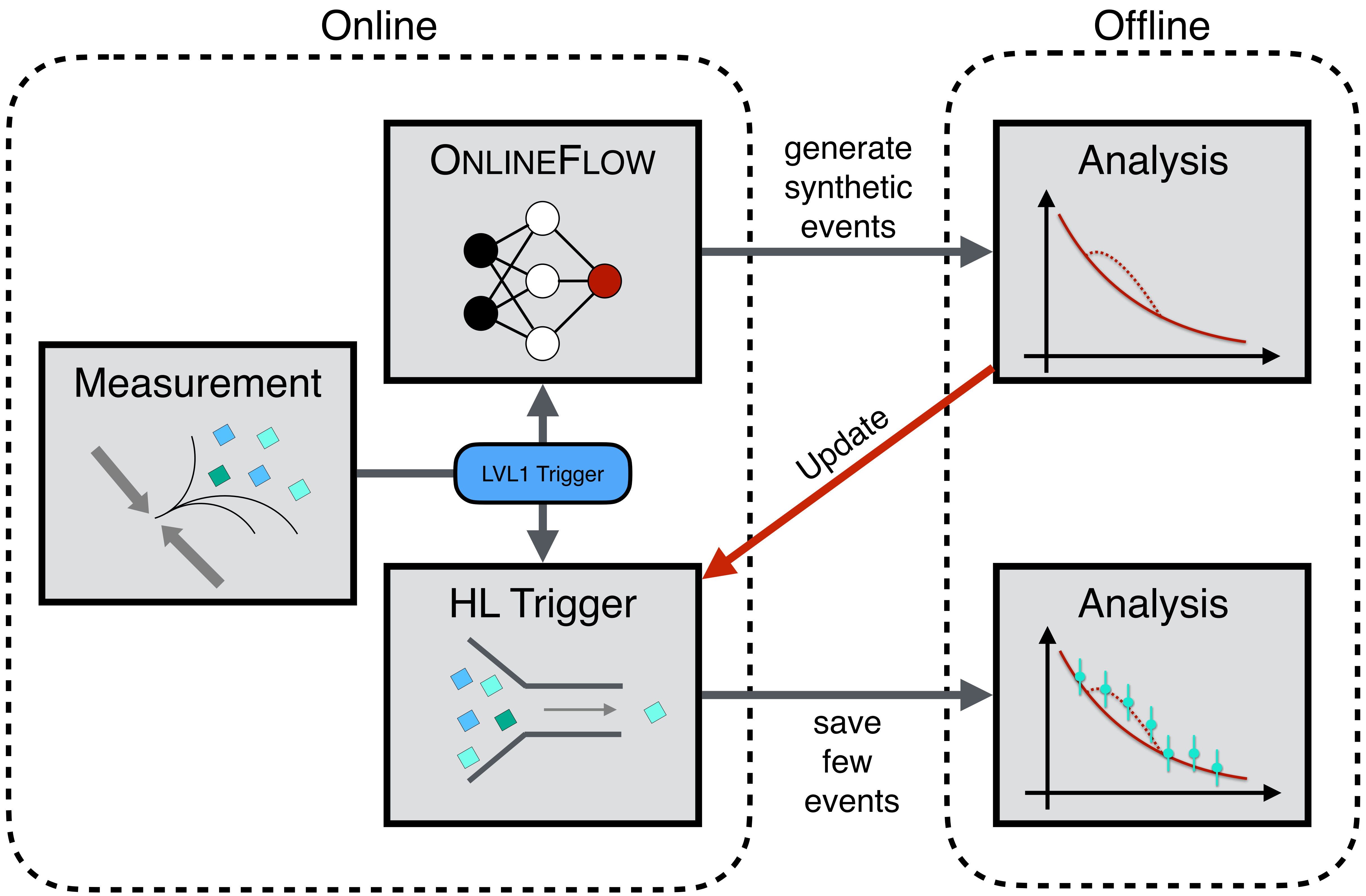
Ongoing data challenge for Fast Anomaly Detection [2107.02157]

Organizers: Govorkova, Puljak, Ngadiuba, Pierini, Aarrestad

Deadline: ML4Jets2022@Rutgers in November

Fast ML for Online Anomaly Detection

Crazy idea: what if we could replace LHC with a generative model?



Train generative model (eg Normalizing Flow) on every event (or every event after L1T).

If generative model is perfect, we have successfully encoded SM (plus any NP in the data)!

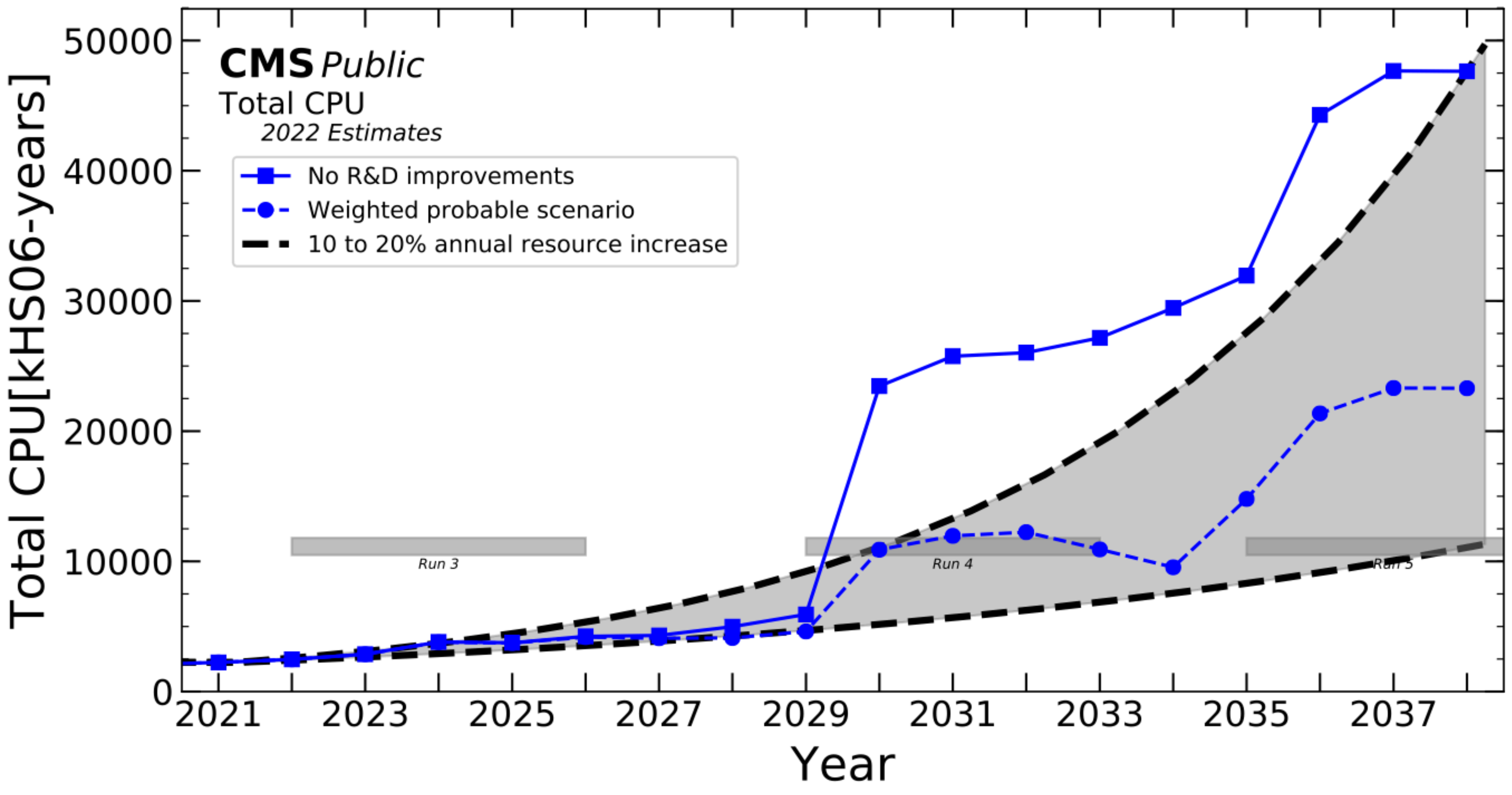
Can potentially discard LHC (after all the data is taken) and just perform offline analysis on events from generative model?!

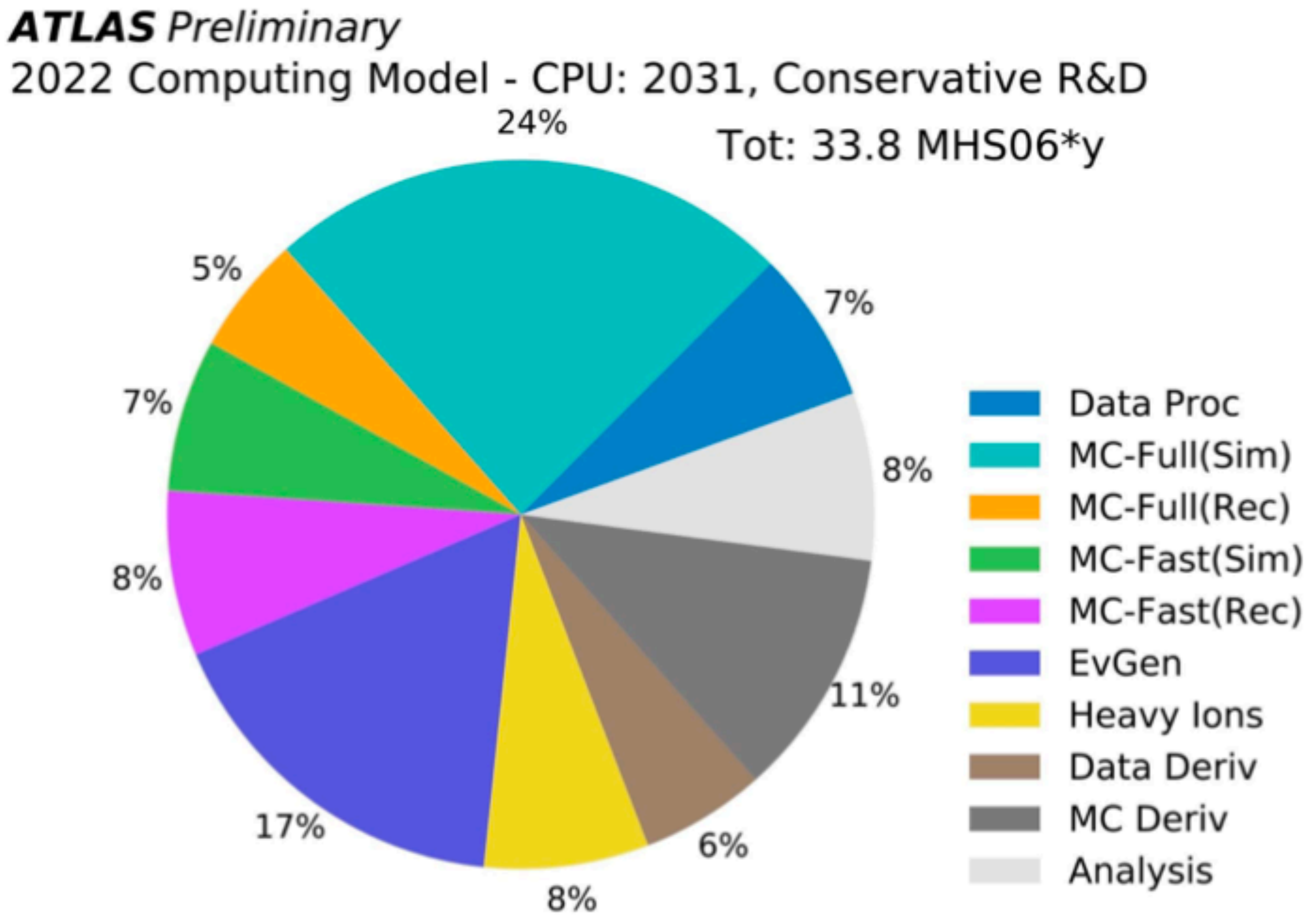
Butter, Diefenbacher, Kasieczka, Nachman, Plehn, **DS** & Winterhalder 2202.09375

Fast ML for Surrogate Modeling

Fast ML for Surrogate Modeling

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





CERN-LHCC-2022-005

Detector simulation (GEANT4) and event generation (MG5, Pythia, Herwig, …) are major — and growing — bottlenecks at LHC and other experiments

Fast ML for Surrogate Modeling

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

Fast ML for Surrogate Modeling

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

GEANT4 → 10^{5} events

Fast ML for Surrogate Modeling

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

GEANT4 → 10^{5} events → Surrogate model

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

Fast ML for Surrogate Modeling

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 ML for Surrogate Modeling

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?

Fast ML for Surrogate Modeling

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

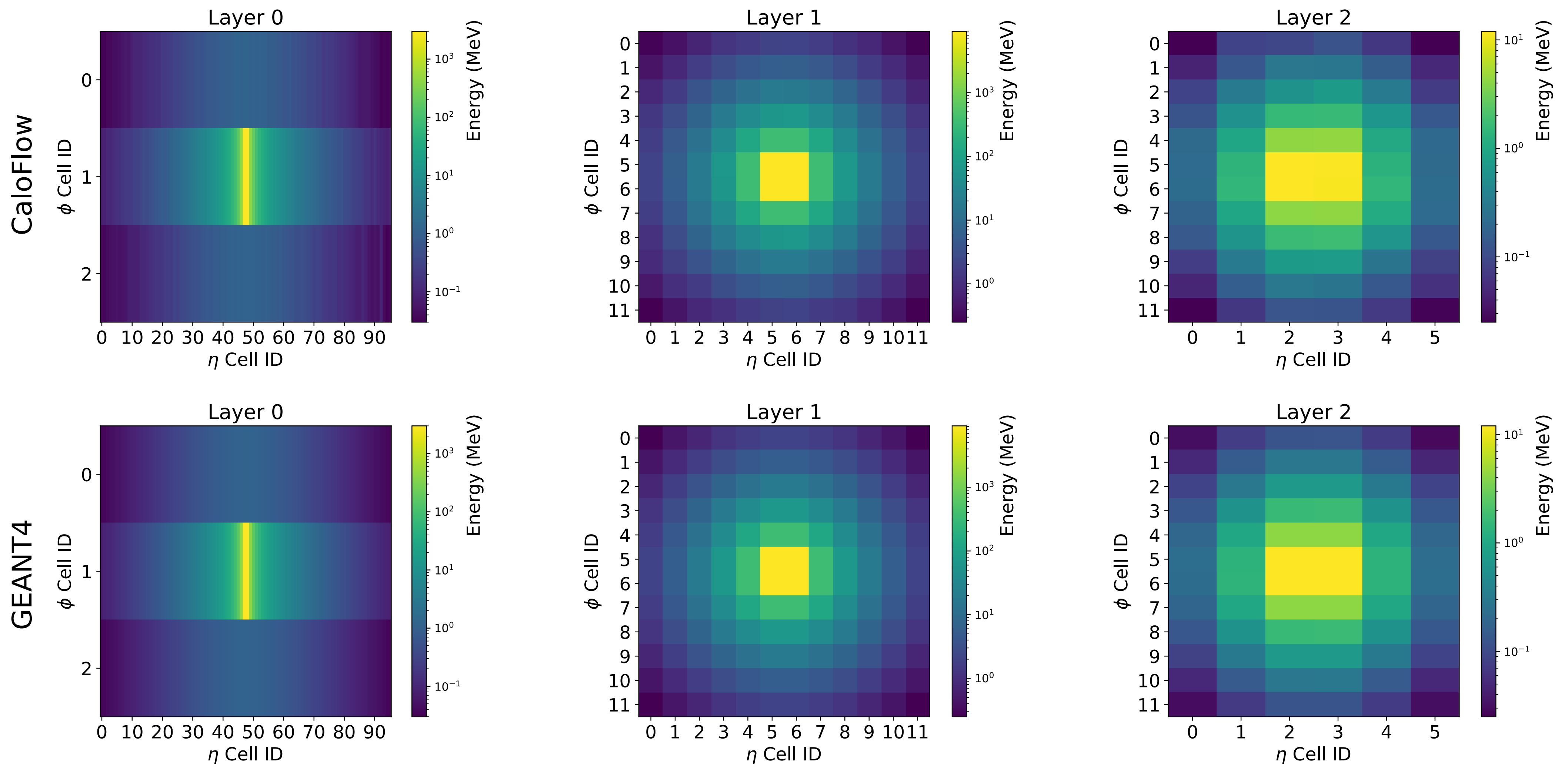
Fast ML for Surrogate Modeling

ML methods are achieving impressive performance on high-dimensional surrogate modeling tasks

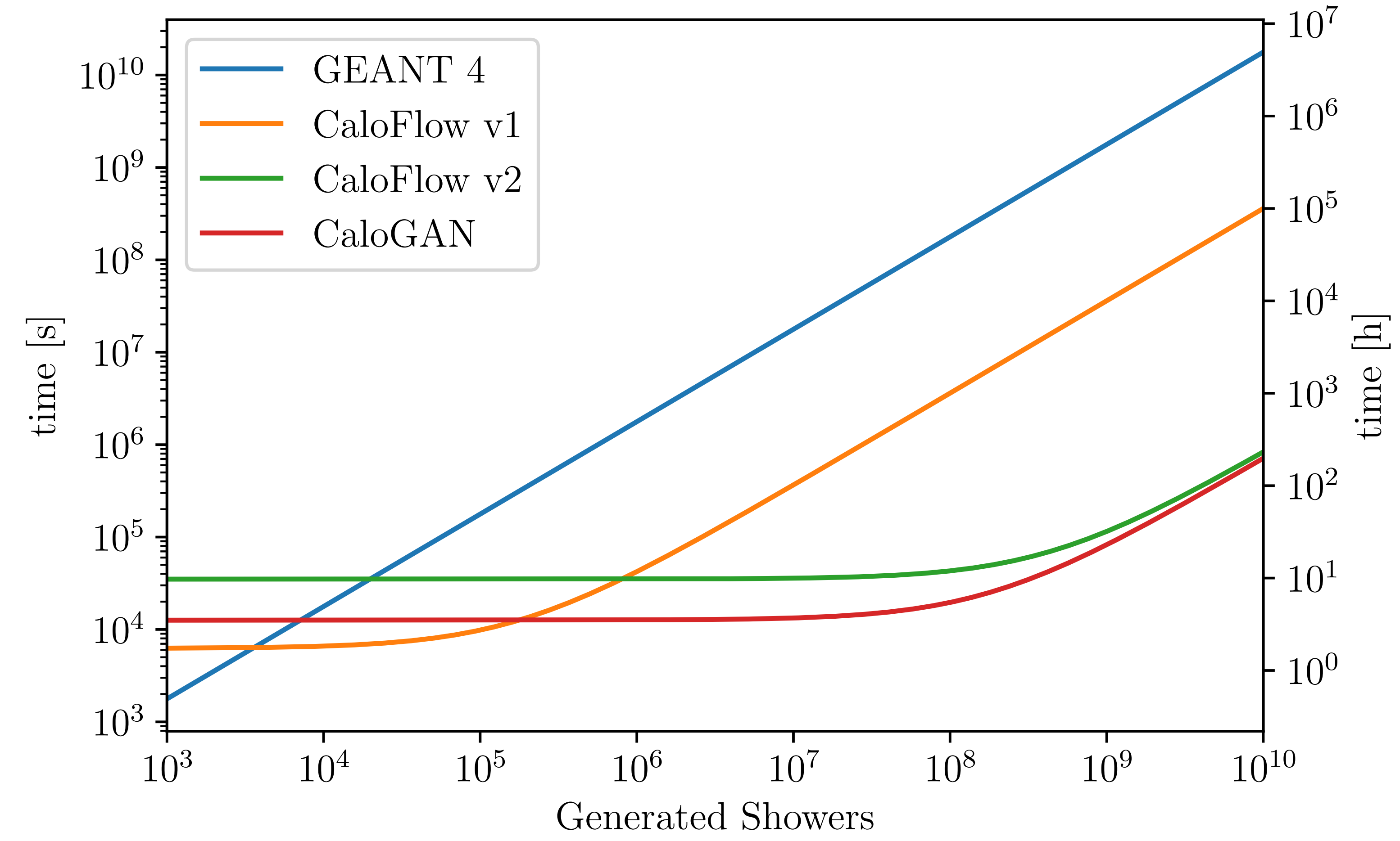
CaloFlow [Krause & DS, 2106.05285, 2110.11377] — first ever GEANT4 surrogate model based on normalizing flows

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)

First to ever pass the "ultimate classifier metric" test



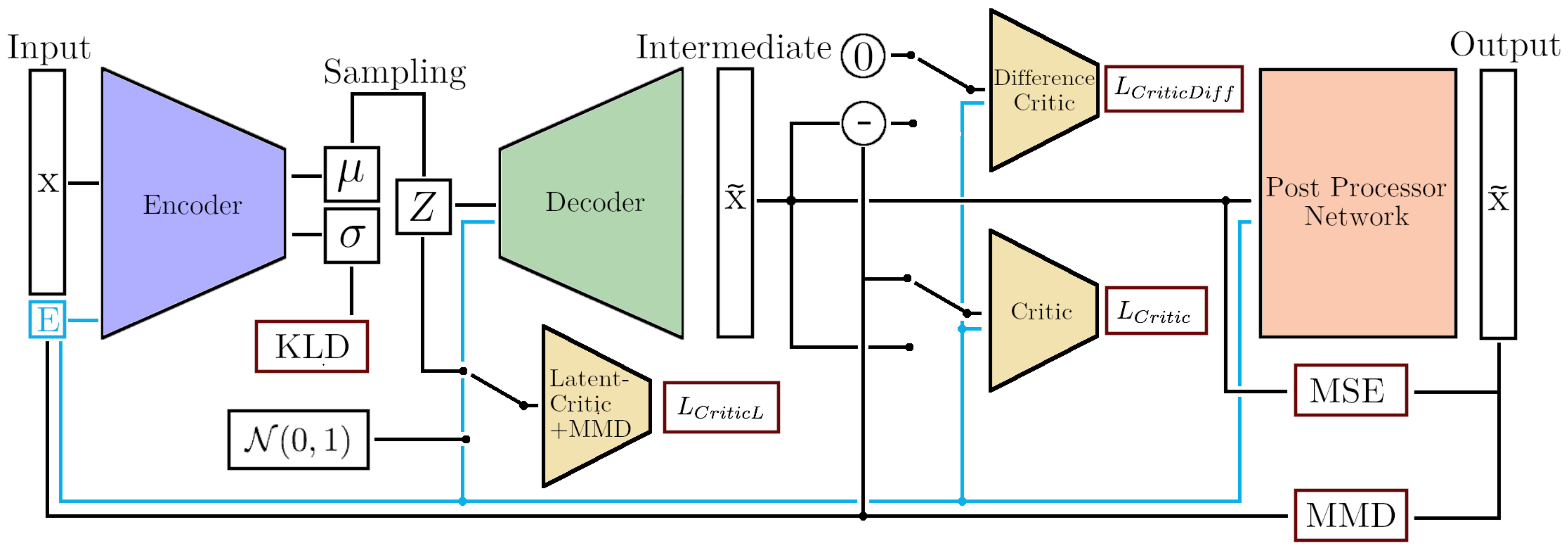
Toy ATLAS ECAL from CaloGAN [Paganini, de Oliveira & Nachman 1705.02355, 1712.10321] — 3 layers, 504 voxels

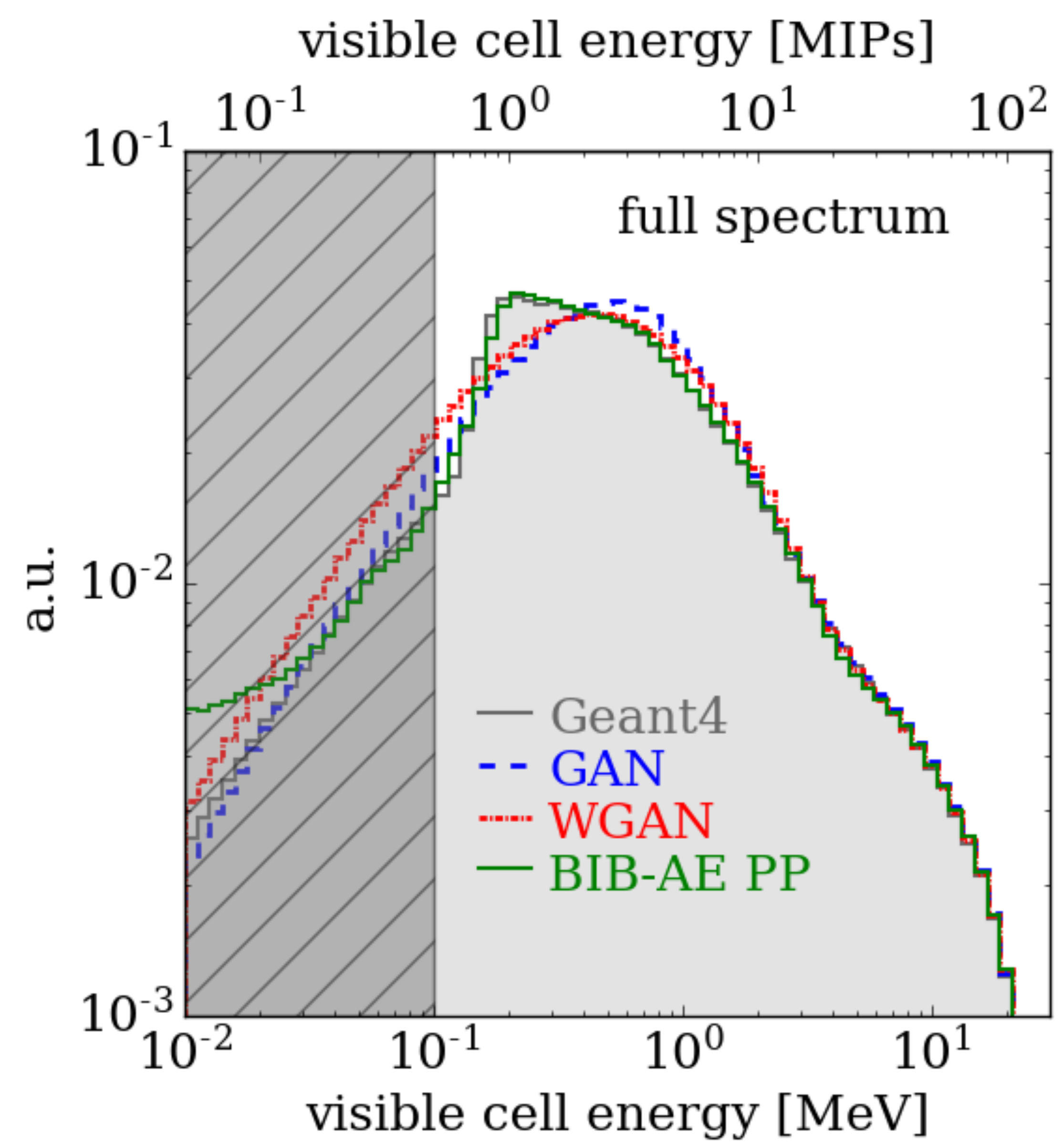


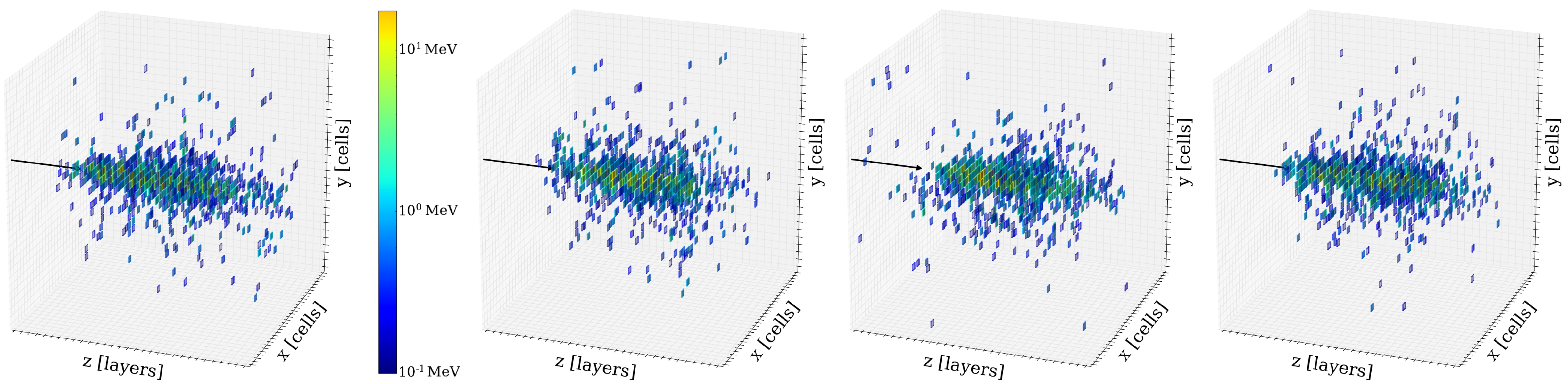
$10^4\times$ faster than GEANT4!

Fast ML for Surrogate Modeling

Bib-AE Buhmann et al [2005.05334, 2112.09709]
Combination of VAE and GAN

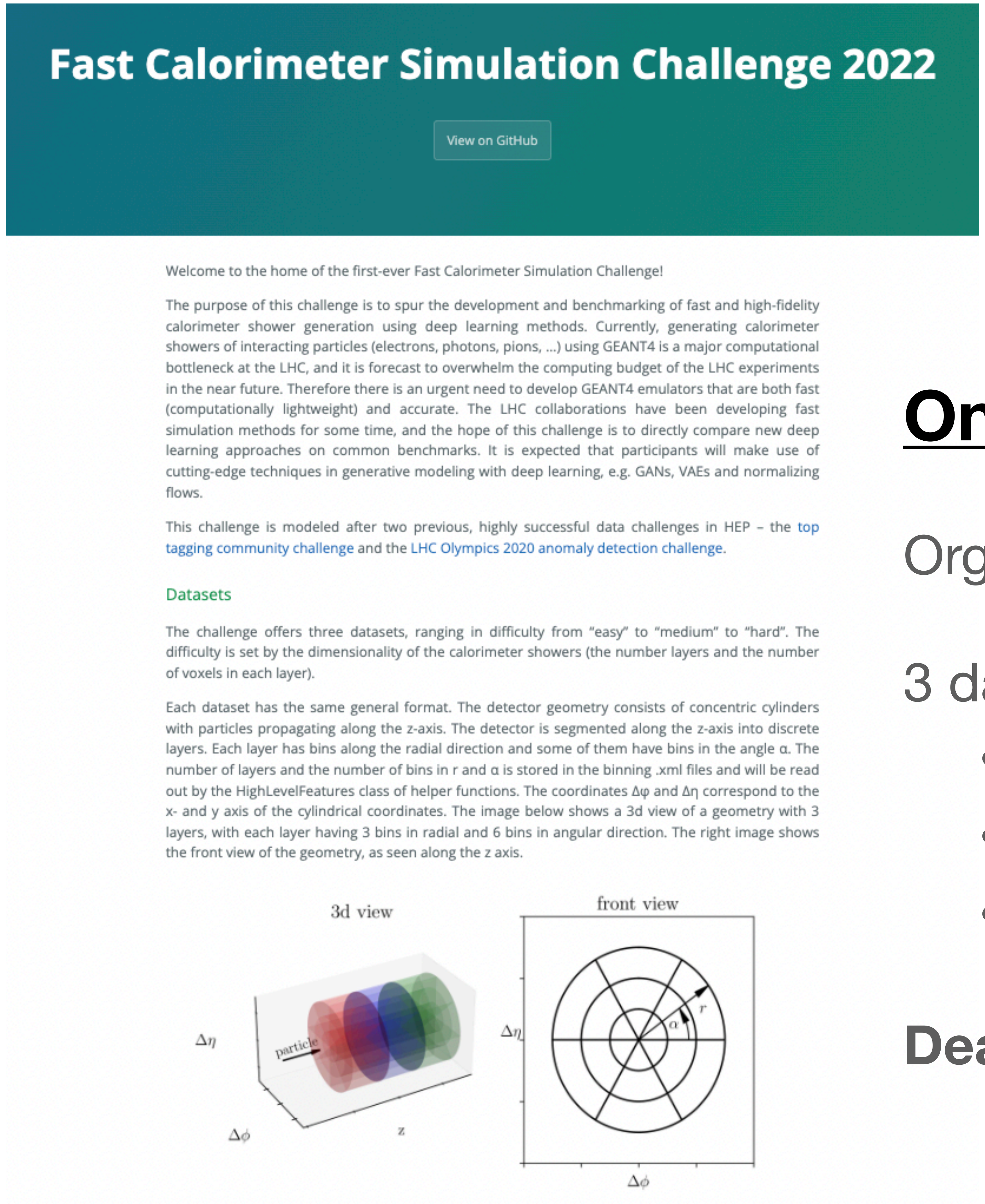






30x30x30 = 27,000 voxels ILD prototype (similar scale to CMS HGCAL) — current frontier in dimensionality

Fast ML for Surrogate Modeling


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.

<u>Ongoing data challenge for fast calorimeter simulation</u>

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)

Deadline: ML4Jets2022@Rutgers in November

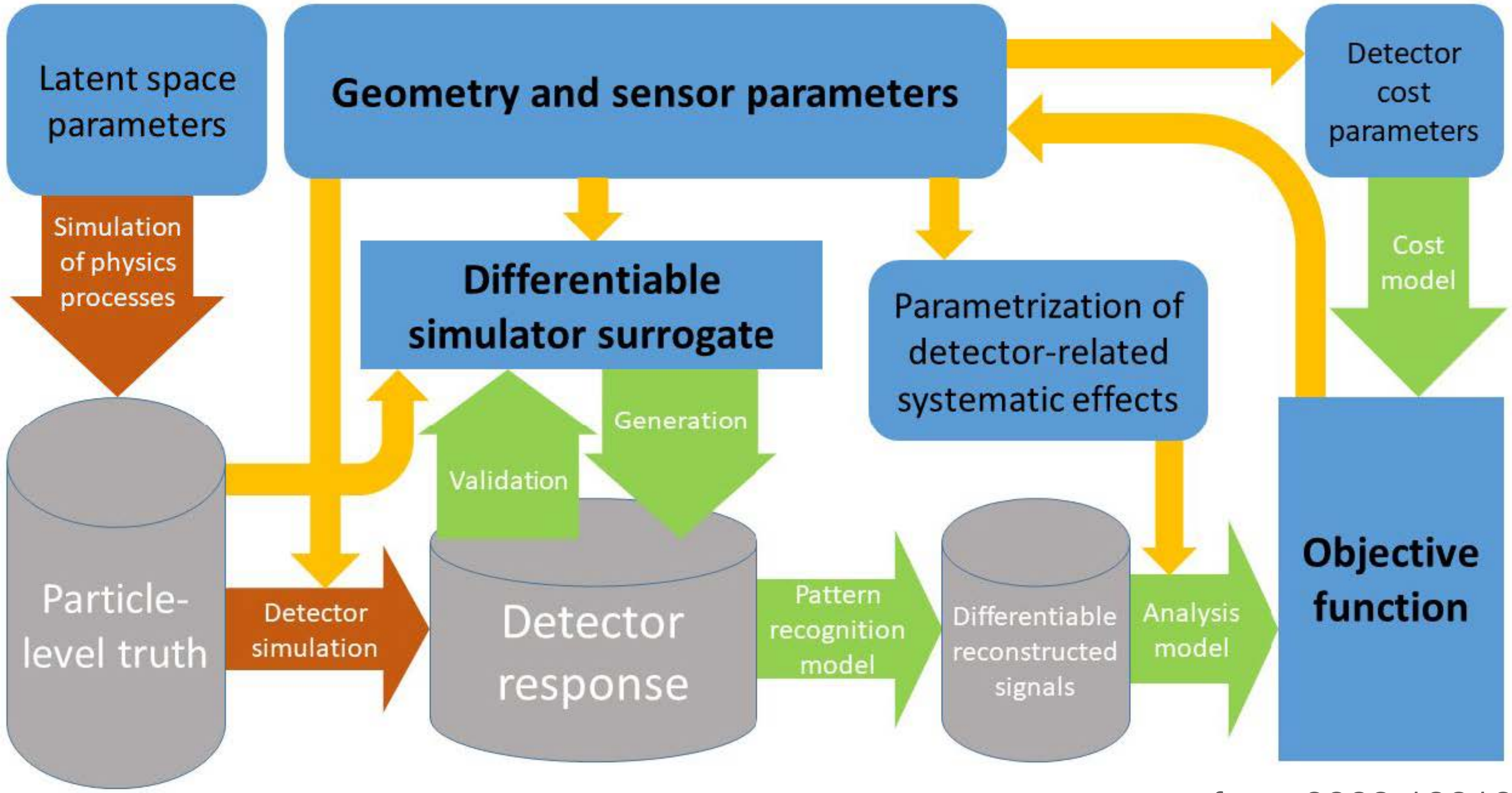
https://calochallenge.github.io/homepage/

Other new avenues for ML in HEP

ML for Instrumentation

Optimizing detector design

Fully differentiable surrogate model => could be very useful in designing experiments

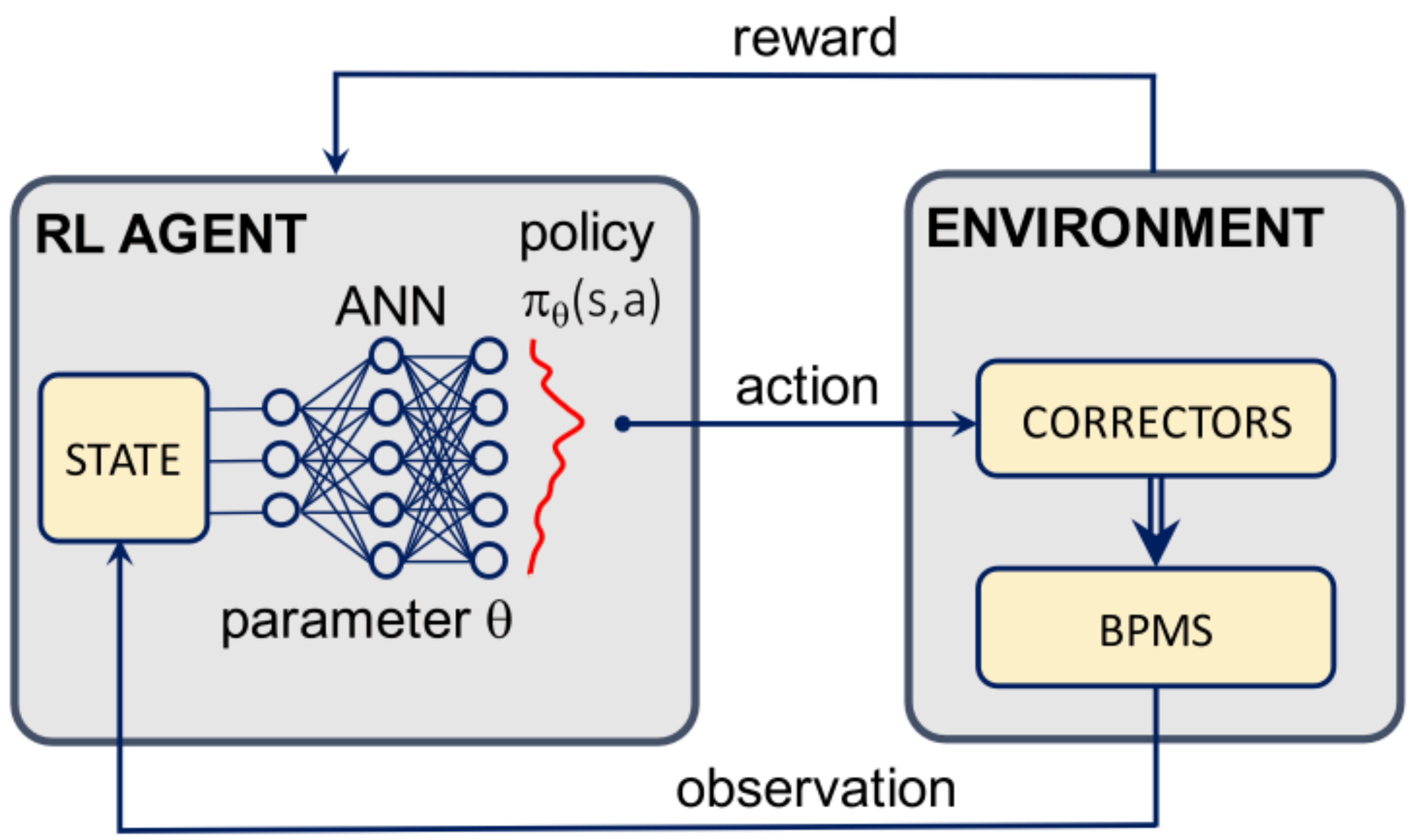


from 2203.13818

- MODE collaboration WP “End-to-End Optimization of Particle Physics Instruments with Differentiable Programming” 2203.13818
- See also AI-assisted design of EIC detector [Fanelli et al 2205.09185]

ML for Instrumentation

Accelerator/detector operations



from Kain et al

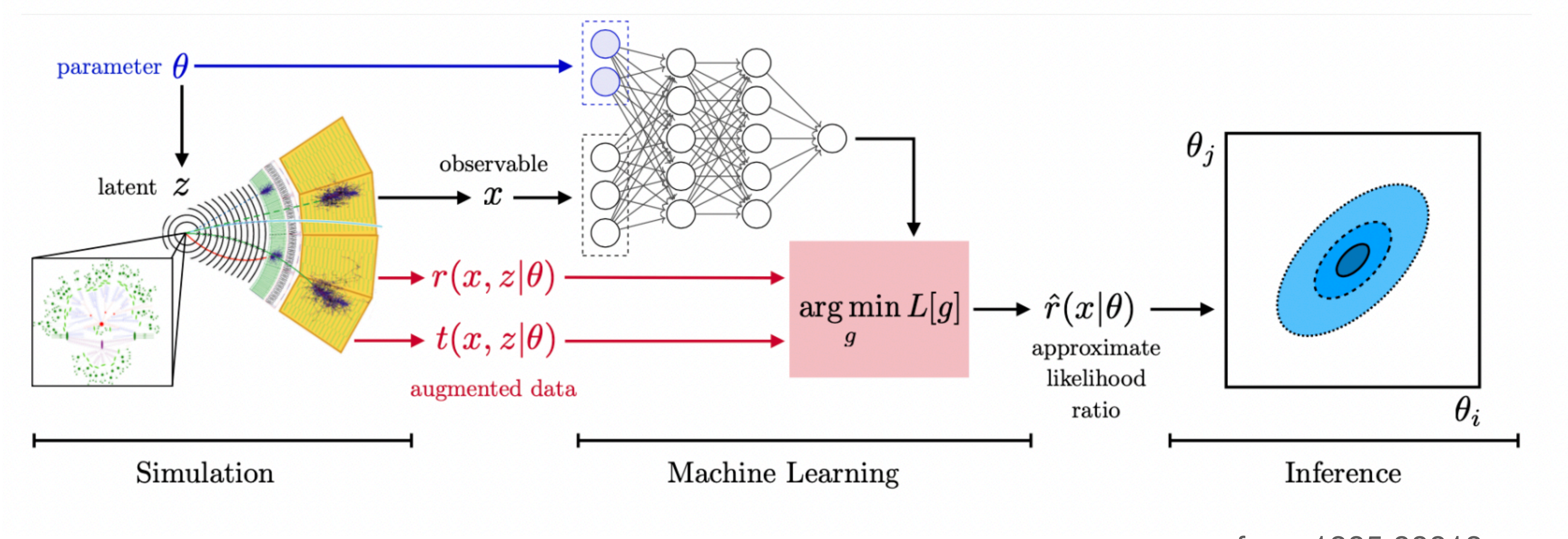
- Many promising applications of **Reinforcement Learning** to real-time accelerator operations
  - Pang et al "Autonomous Control of a Particle Accelerator using Deep Reinforcement Learning" 2010.08141
  - St. John et al "Real-time Artificial Intelligence for Accelerator Control: A Study at the Fermilab Booster" 2011.07371
  - Kain et al "Sample-efficient reinforcement learning for CERN accelerator control" Phys.Rev.Accel.Beams 23 (2020) 12, 124801
  - Scheinker et al "Advanced Control Methods for Particle Accelerators (ACM4PA) 2019 Workshop Report" 2001.05461
- "self-driving triggers"
  - Bartoldus et al Snowmass WP 2203.07620
  - Y. Chen et al., "Self-driving data trigger, filtering, and acquisition", Snowmass LOI (2020)
- "self-driving telescopes"
  - Nord et al, "Cycle and symbiosis: AI and Cosmology intersect to produce new knowledge and tools", Snowmass LOI (2020)

ML for Measurements

Potential for performing measurements using full unbinned phase space

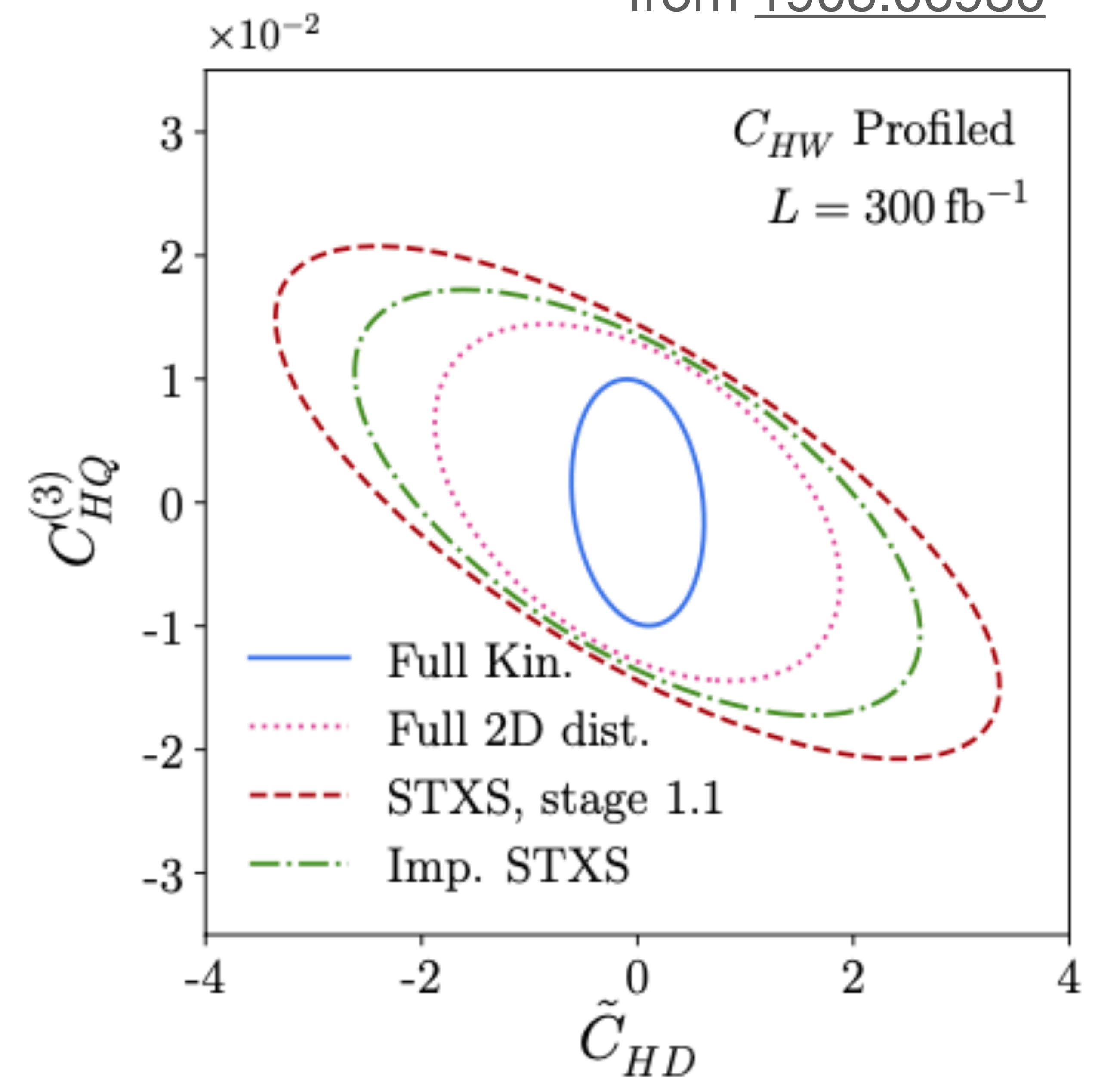
<u>"Simulation based inference"</u>

Cranmer, Brehmer, Louppe 1911.01429
Brehmer & Cranmer 2010.06439



from 1805.00013

from 1908.06980

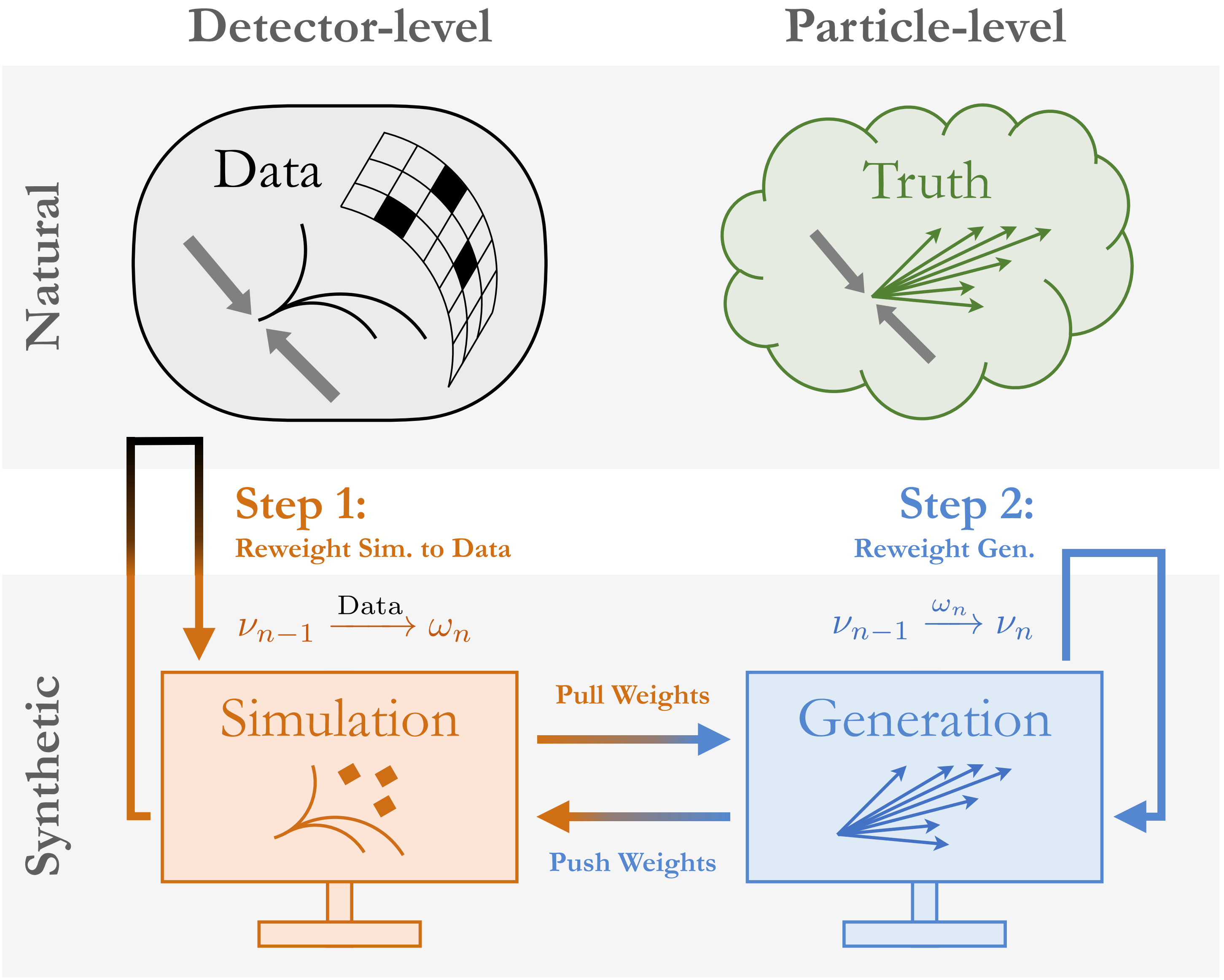


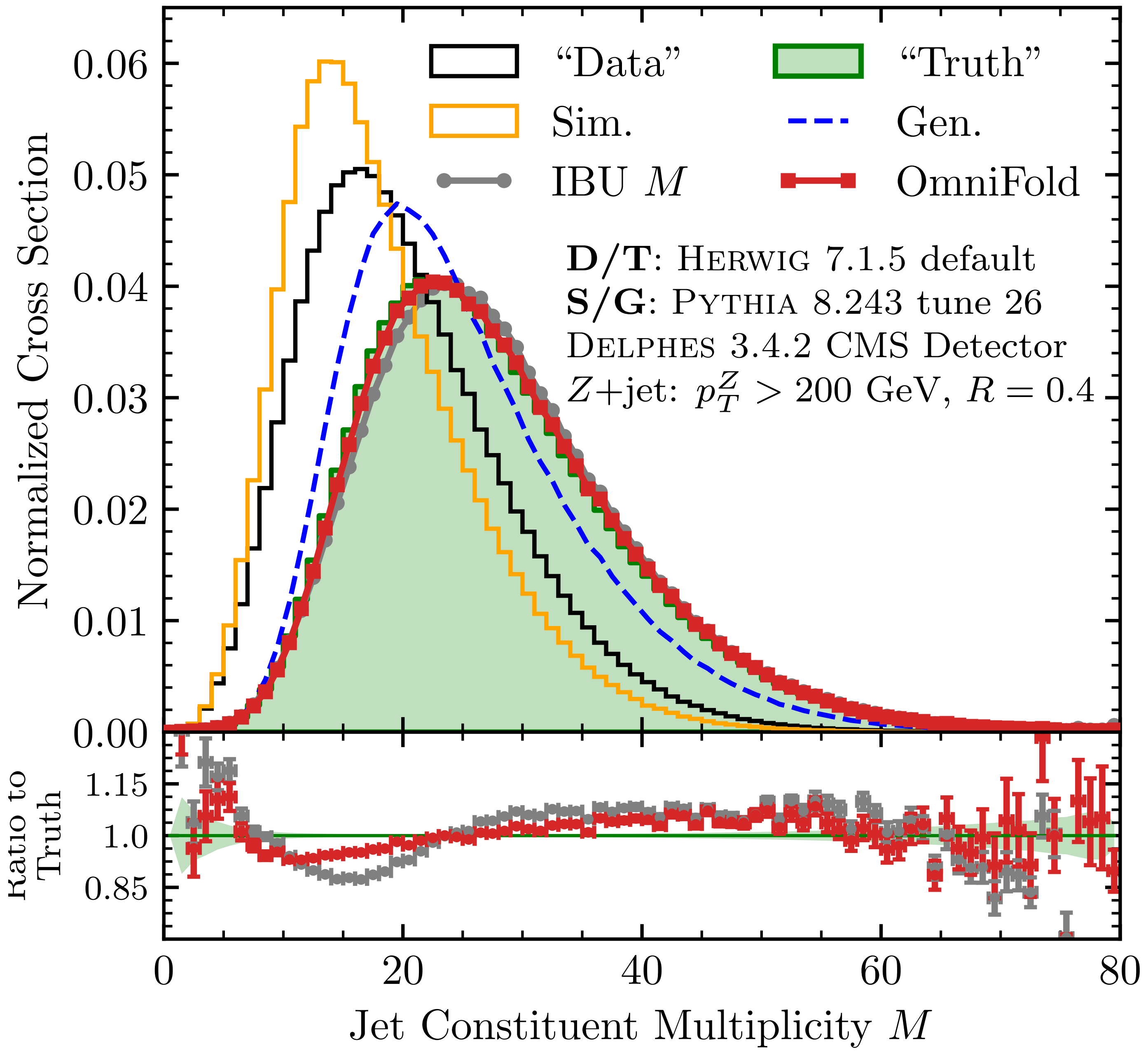
ML for Measurements

Potential for performing measurements using full unbinned phase space

"Omnifold" Andreassen et al 1911.09107

Full phase space unfolding detector->particle level



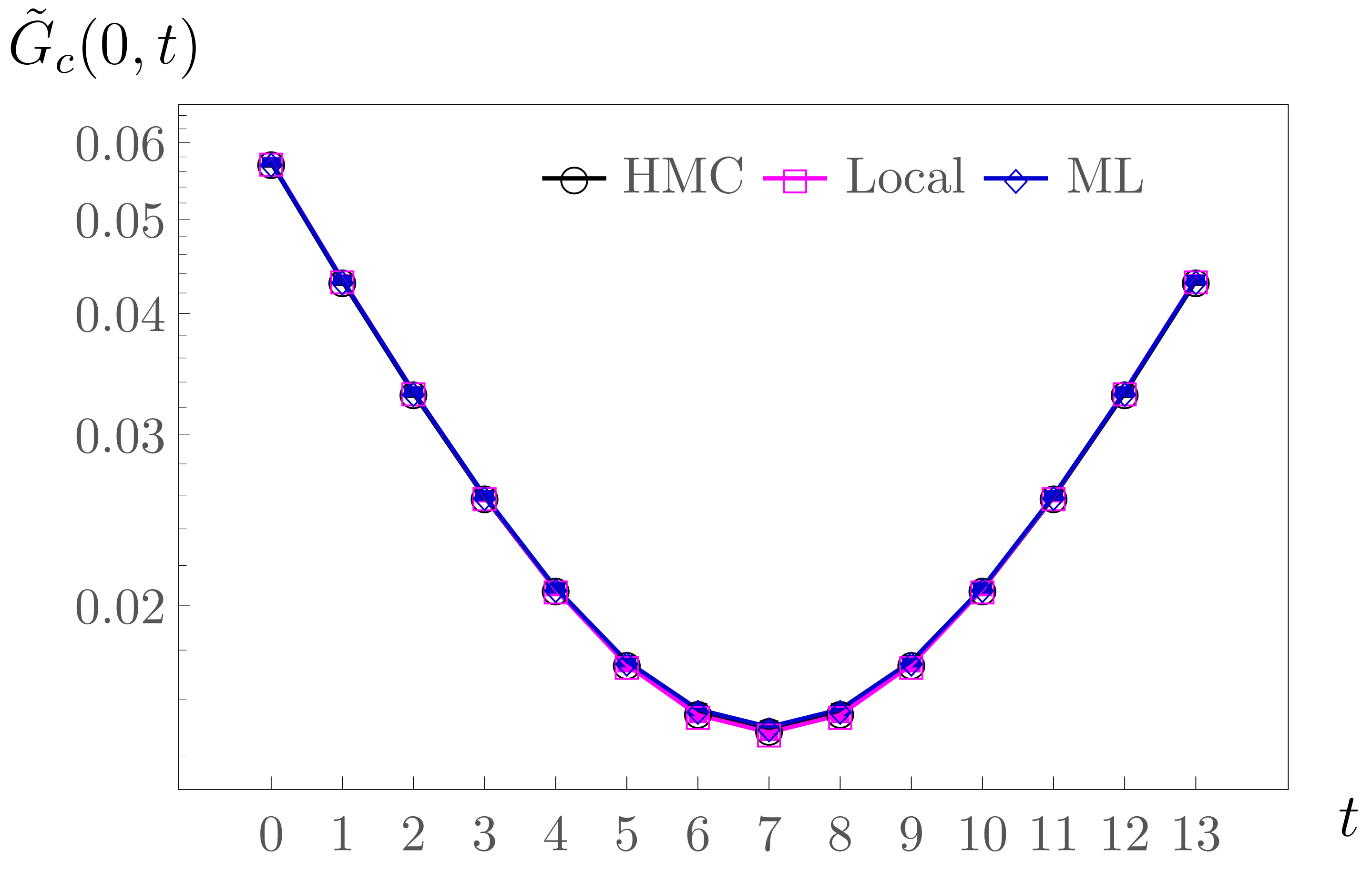


ML for Theory

Modern ML is also making inroads into Theory

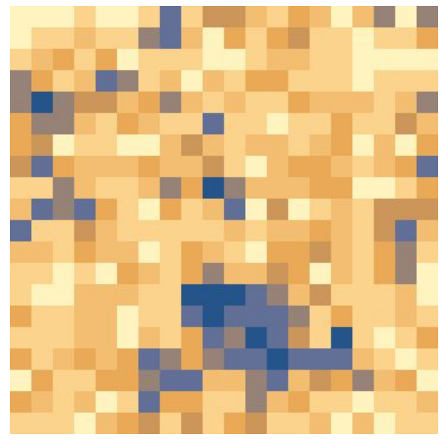
- ML4Lattice Snowmass WP [2202.05838]

Eg using NFs to sample efficiently from lattice configurations

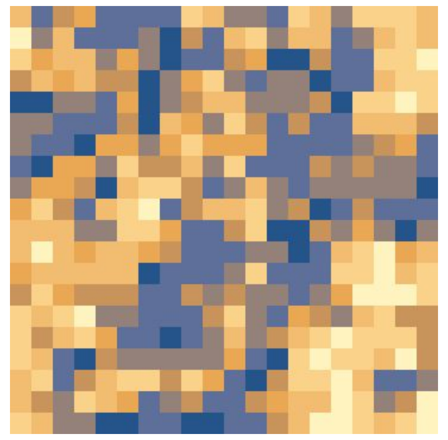


from 1904.12072

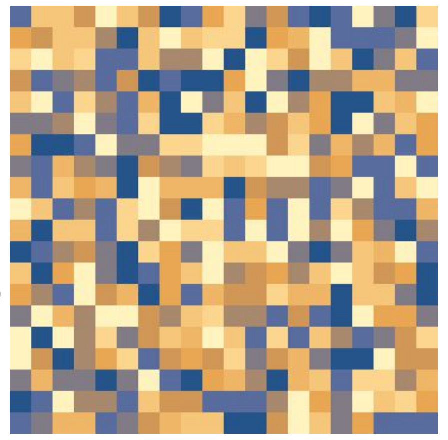
likely
(log prob = 22)



likely
(log prob = 5)



unlikely
(log prob = -6107)



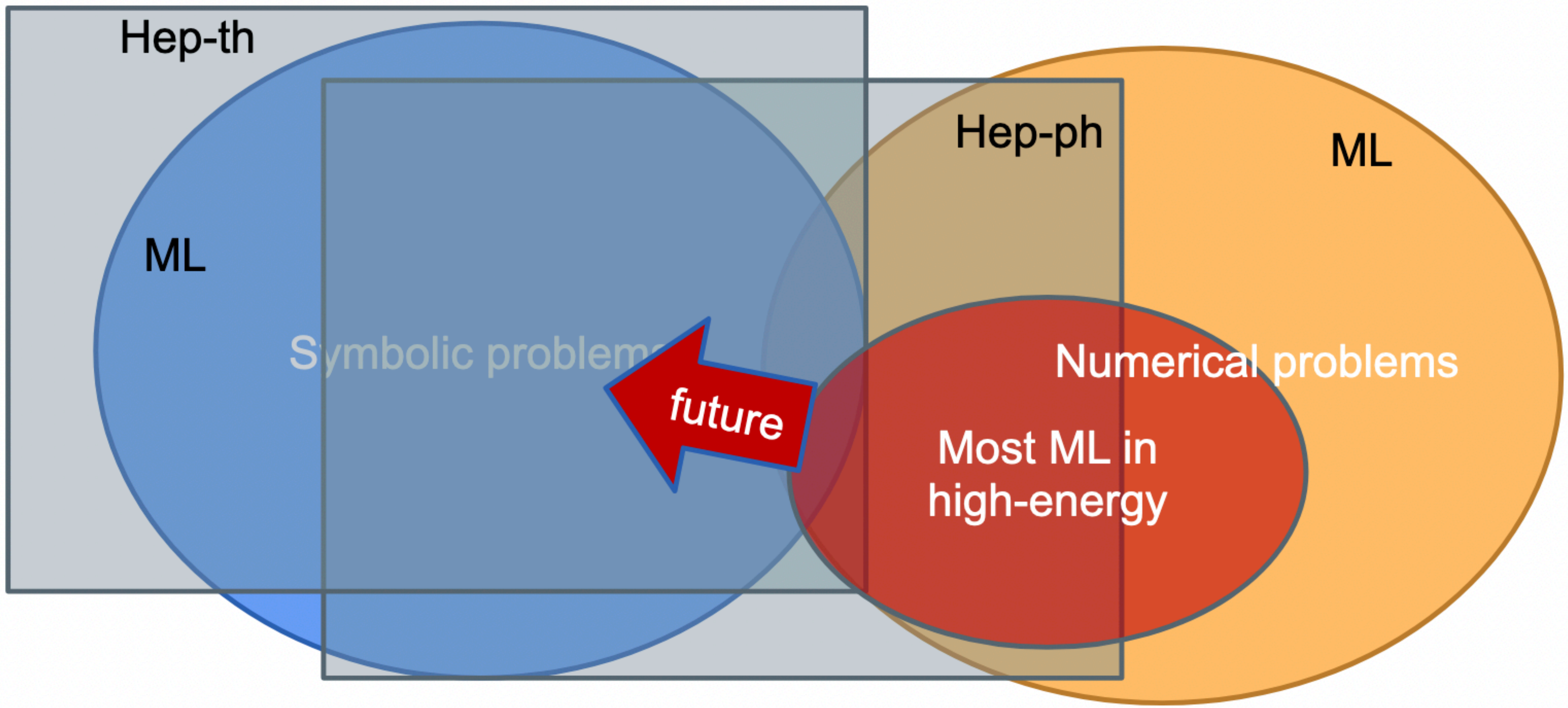
from Kanwar Lattice 2019 talk

- Symbolic tasks (regression, learning physical laws, simplification)

- …

ML for Theory

What the future holds?



from M. Schwartz Mainz DLEP22 Workshop talk

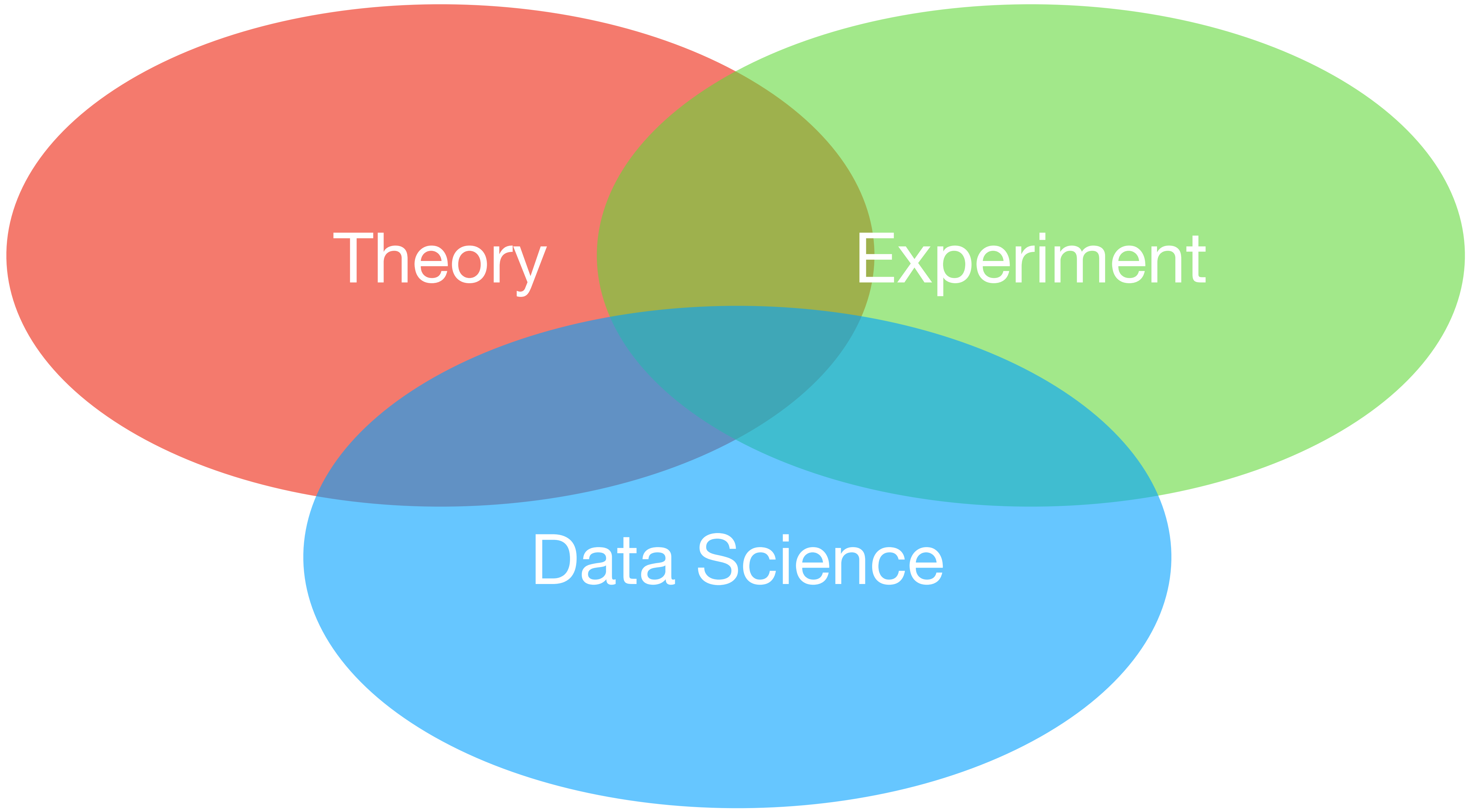
Summary

- Modern ML is a powerful new tool that enables qualitatively new kinds of physics analyses that weren't possible before.
- Modern ML holds enormous potential for new physics searches, triggering, fast simulation, instrumentation, theory and more.
- There has been an explosion of development of new methods and proofs-of-concept. Many of these are beginning to be ported over to real data.

Outlook

I believe we are witnessing the dawn of a new era of **data-driven physics…**

…and also the dawn of a new kind of physicist — **the "data physicist"**.



These are exciting times for ML and HEP!