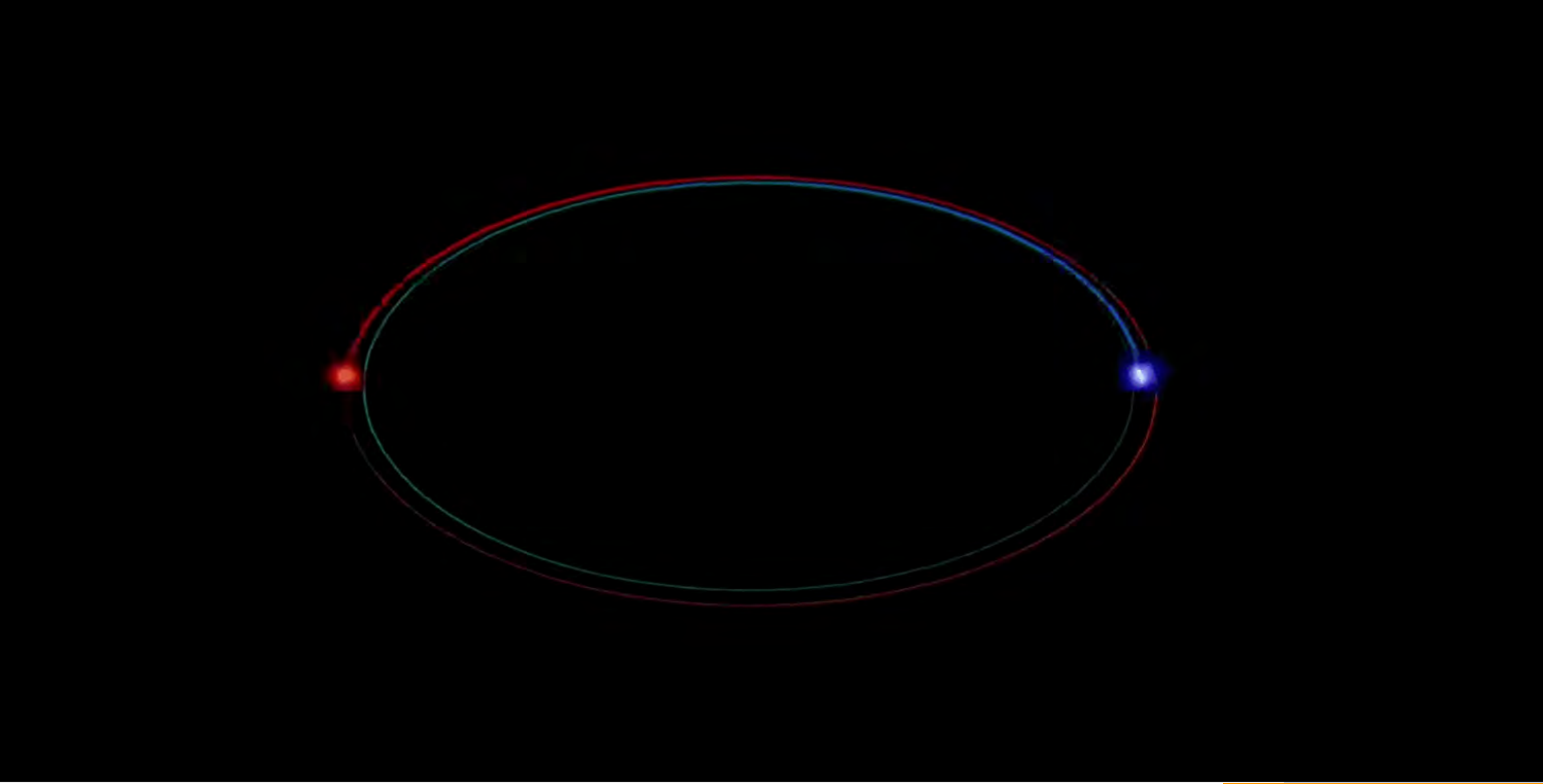
# Collider Physics Innovations Powered by Machine Learning

vmikuni@lbl.gov

vinicius-mikuni

Vinicius M. Mikuni

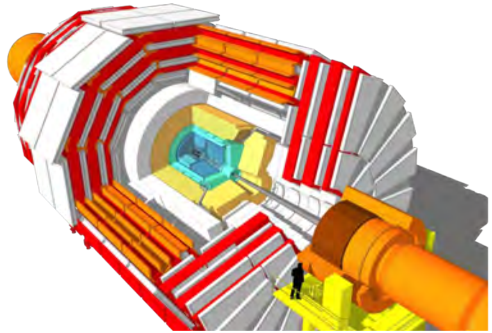


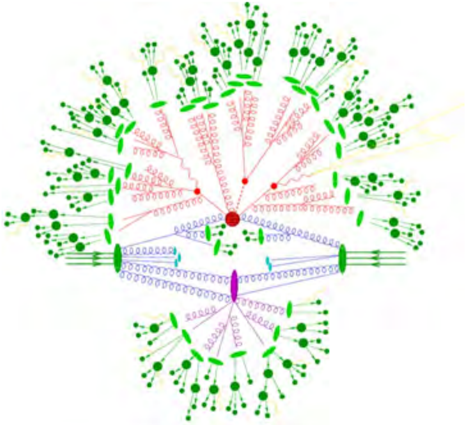
# Particle colliders

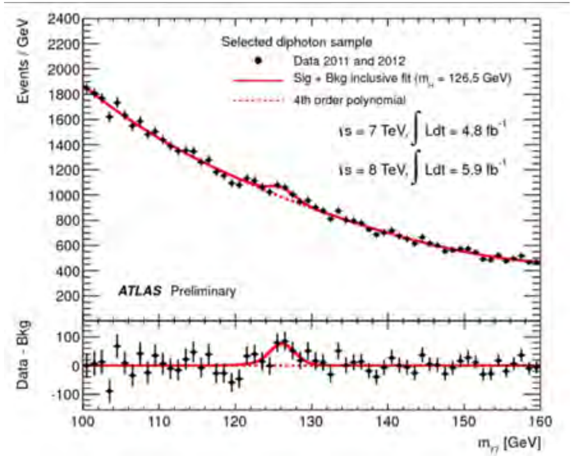
Animation from business insider

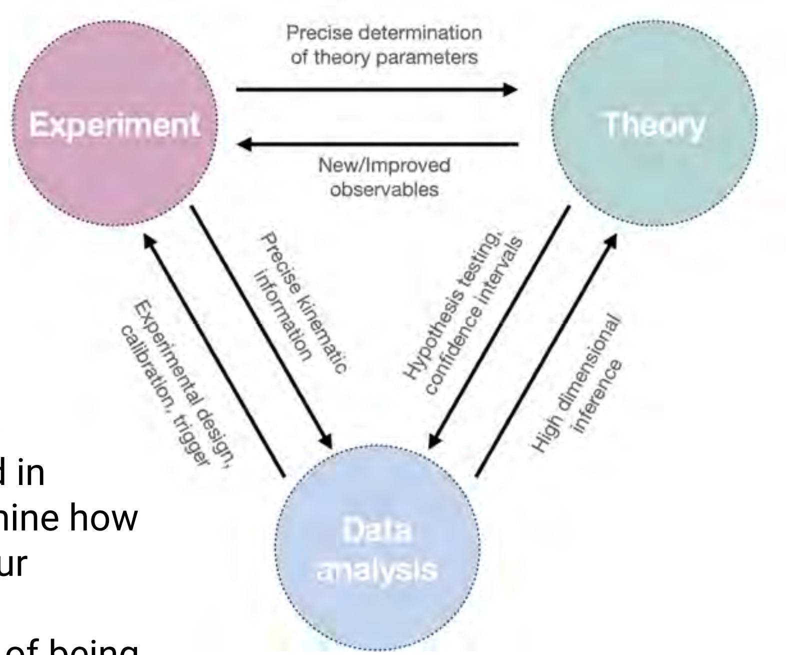
# Introduction

Picture from [arXiv:1411.4085](arXiv:1411.4085)







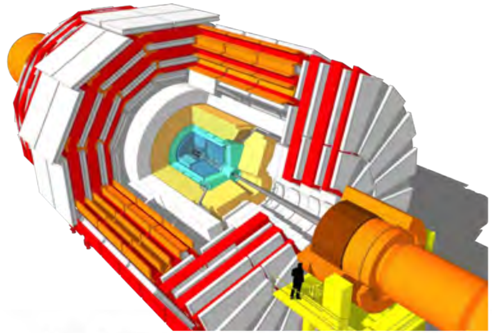



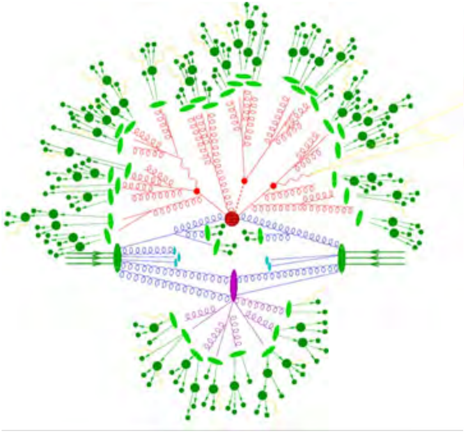

**Significance**: metric used in collider physics to determine how confident we are about our claims.

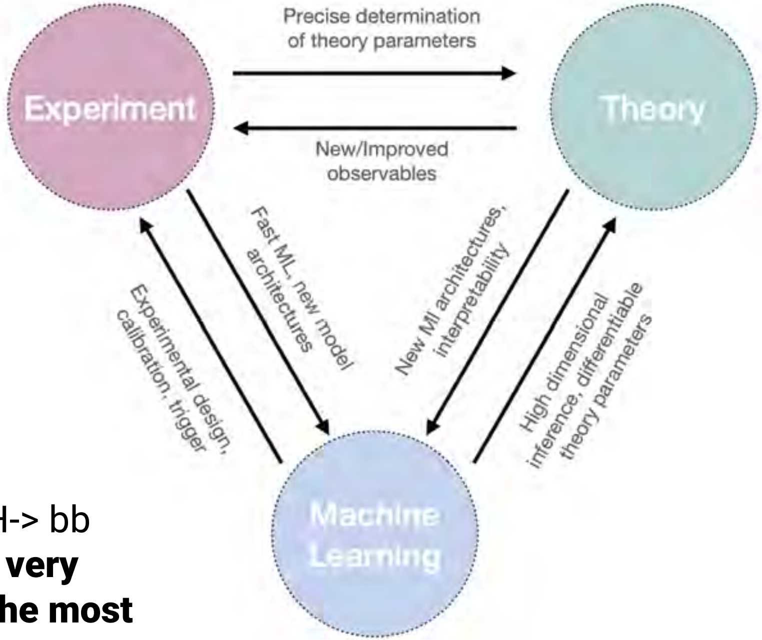
- **4σ:** 1 in 1M chance of being a spurious observation
- **5σ:** 1 in 3.5M chance of being a spurious

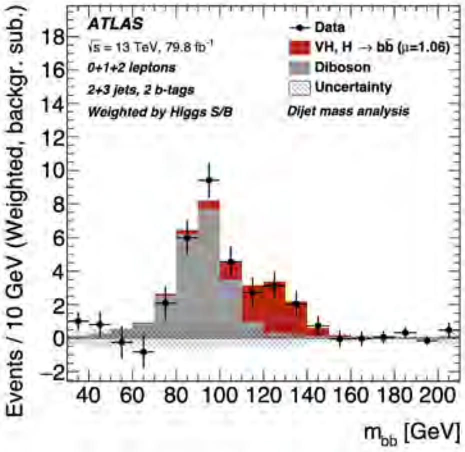
# Introduction

Picture from [arXiv:1411.4085](https://arxiv.org/abs/1411.4085)







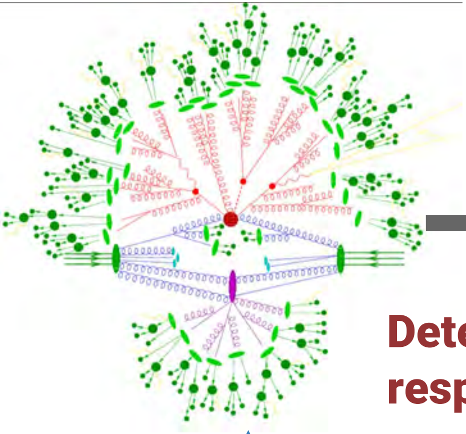




"The extraction of a signal from H-> bb decays in the WH channel **will be very difficult at the LHC, even under the most optimistic assumptions**" - SNOWMASS-2001-P111

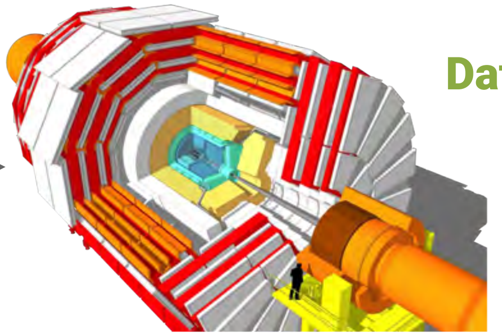
# What I'm talking about

Picture from [arXiv:1411.4085]

Surrogate



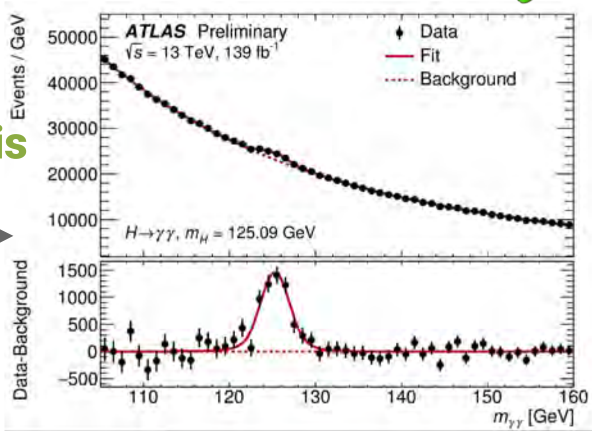
Detector response



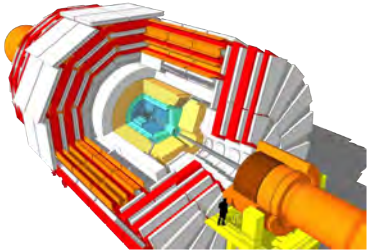
Unfolding

Data analysis

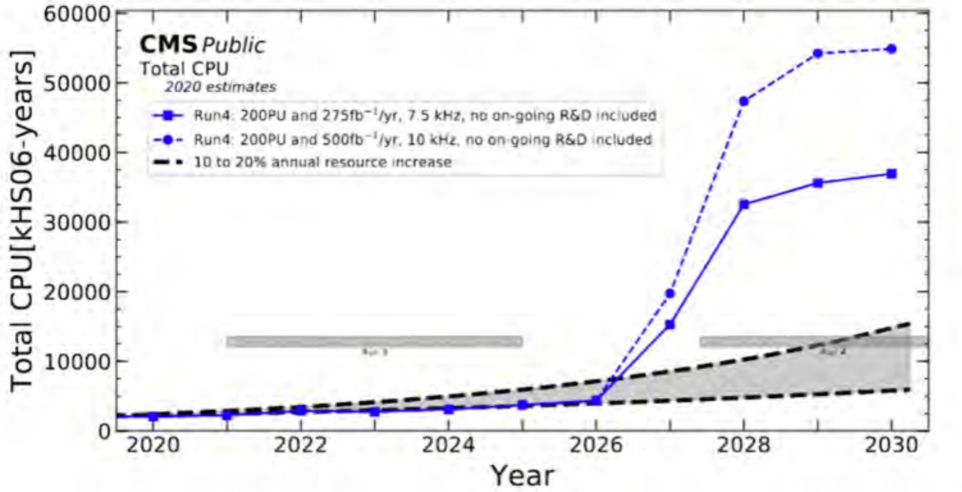
Anomaly detection



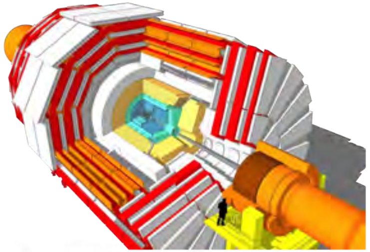
# Surrogate modeling for detector simulation



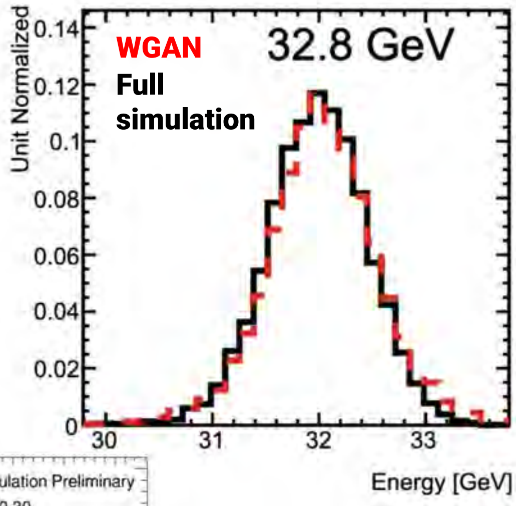
- Detector simulation is **computationally expensive**:
  - Full detector simulation of a particle can take up to **a minute** and we still need **many billions of particles simulated**
- For previous LHC runs, detector simulation used around **40% of all computing resources** and may go beyond the available budget for future runs

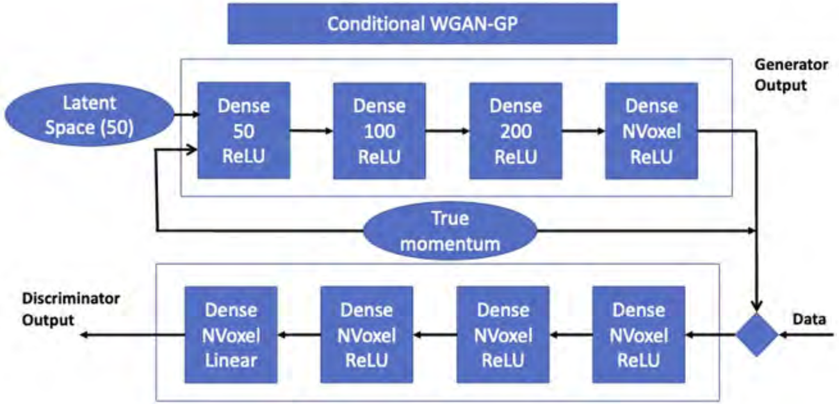


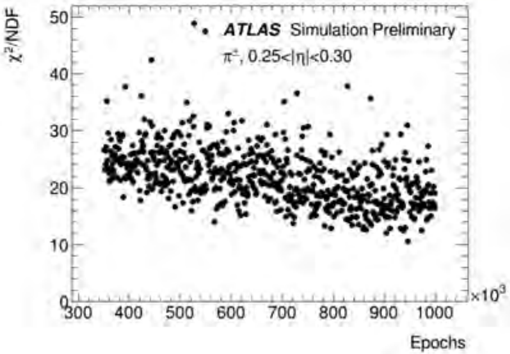
# Surrogate modeling for detector simulation



- At the **Large Hadron Collider (LHC)**, experiments already have a WGAN-GP planned to replace part the full simulation routine
- **Fully-connected** architecture that leads to orders of magnitude faster generation compared to full simulation
- 1000s of times faster than the current full simulation

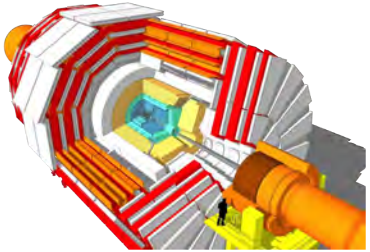


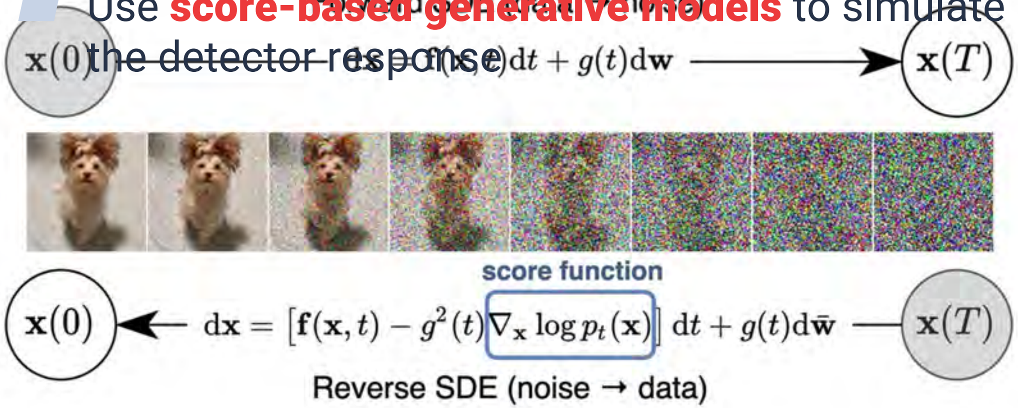




**See also:** M. Paganini, L. de Oliveira, and B. Nachman, Phys. Rev. D 97, 014021

# Surrogate modeling for detector simulation

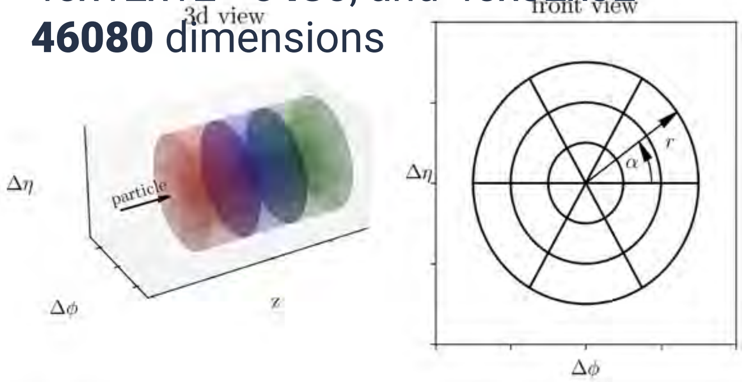


- Bring modern generative models to improve the model fidelity and tackle **high dimensional datasets**
- Explore the detector geometry using **3D convolutions**
- Use **score-based generative models** to simulate the detector response

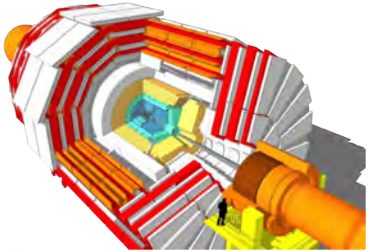



**Song, Yang, et al.** *arXiv preprint arXiv:2011.13456* (2020).

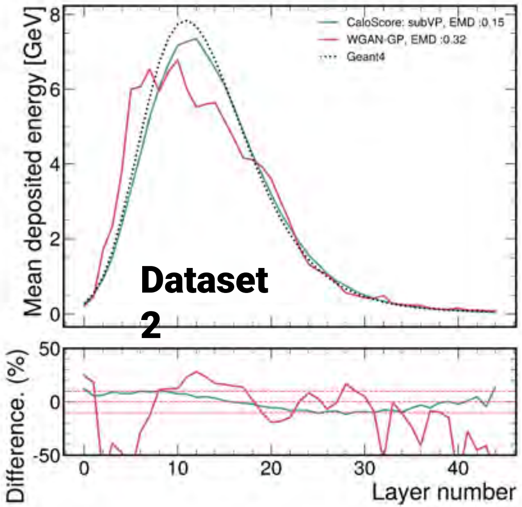
- Proof of concept using the [Fast Calorimeter Simulation Challenge 2022]
- 3 datasets available with **368**, 45x12x12= **6480**, and 45x32x32= **46080** dimensions

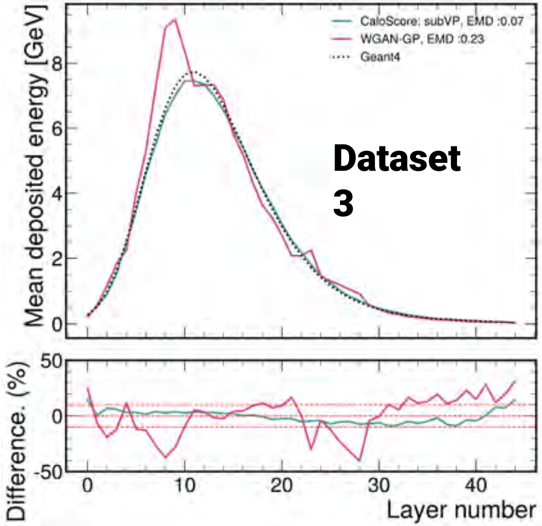


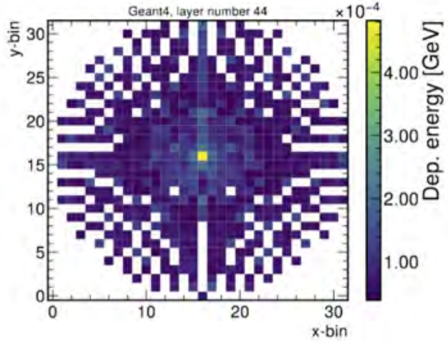

**For a NF approach see also:**
Krause, C. and Shih, D.,*arXiv preprint*, arXiv:2106.05285 (2021)



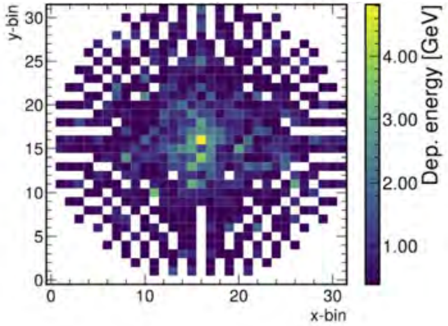
# CaloScore




**Full simulation**




**Generated by the surrogate**

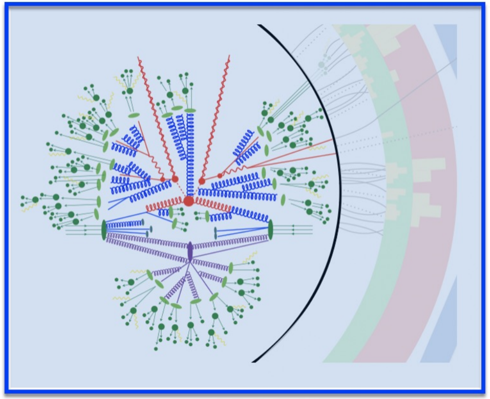
| Dataset | N. of voxels | N. of weights | Time to 100 showers [s] | | |
|---|---|---|---|---|---|
| | | | CaloScore | WGAN-GP | Geant |
| dataset 1 | 384 | 32M | 4.0 | 1.3 | $\mathcal{O}(10^2 - 10^3)$ |
| dataset 2 | 6480 | 1.4M | 5.8 | 1.33 | $\mathcal{O}(10^4)$ |
| dataset 3 | 46080 | 1.7M | 33.4 | 2.06 | $\mathcal{O}(10^4)$ |

**Mikuni, Vinicius**, and Nachman, Benjamin. *Phys. Rev. D 106 (2022), 092009.*

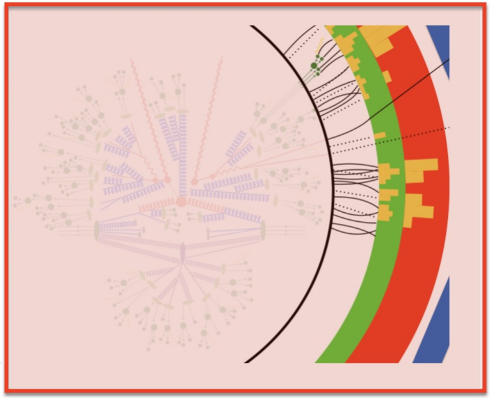
# Detector unfolding

- The **opposite problem** is how to report physics measurements that are **corrected for detector effects:**
- Also referred to **Unfolding** or **solving an inverse problem** or **deconvolution**
- Hard task to accomplish in high dimensional spaces:
  - stograms

**Want this**



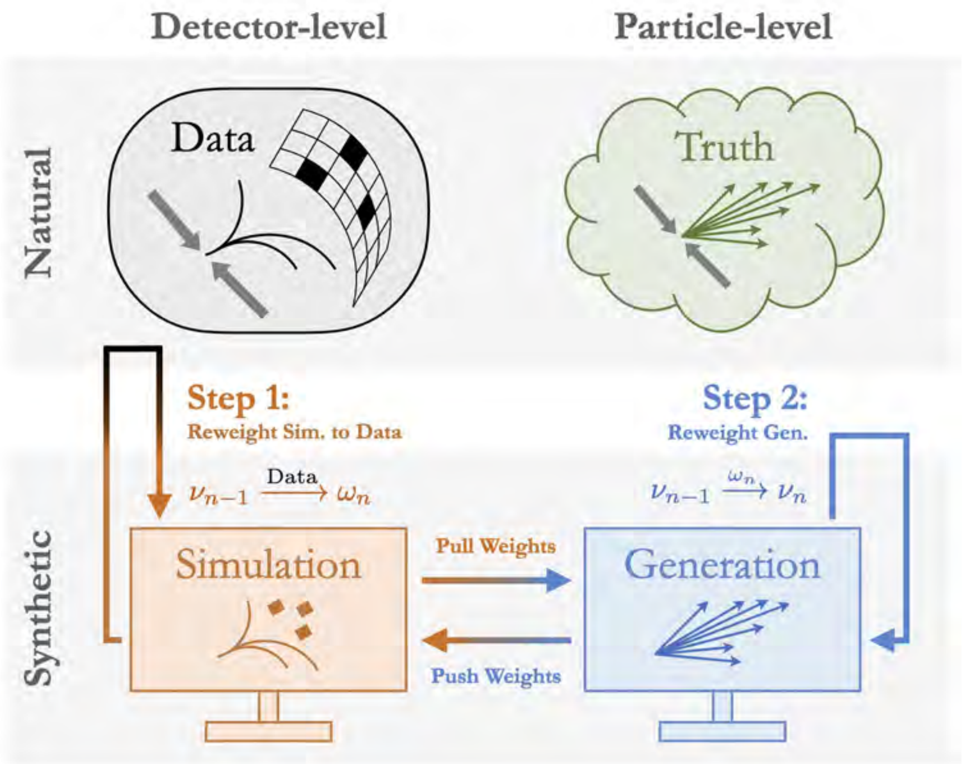
**Measure this**



**Benefits**

- **Precision**: Multiple dimensions can be unfolded simultaneously
- **Reusability**: People using the same data can also use the unfolding information
- **Data Preservation**: Recast or combination of measurements becomes trivial for future experiments

# ML-based unfolding*




Machine learning is used to overcome these limitations in an Expectation-Maximization style

**2 step** iterative approach

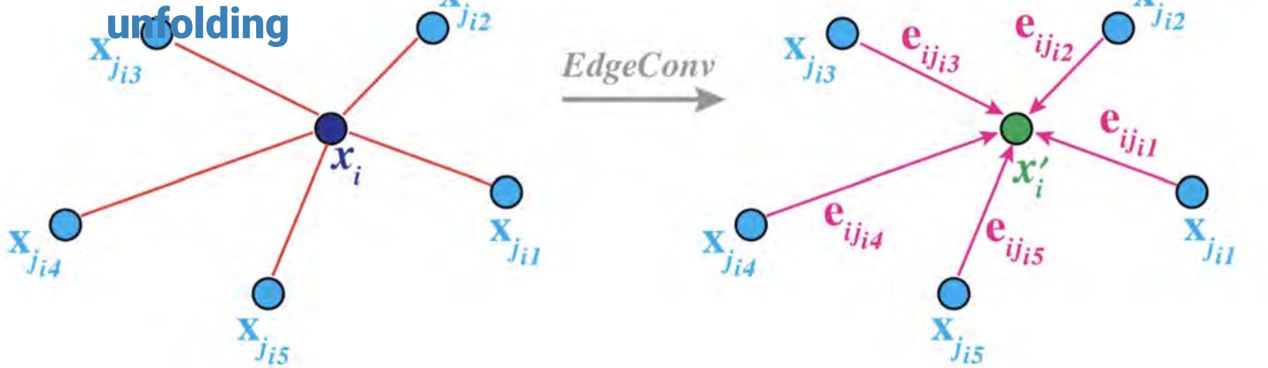
- Simulated events after detector interaction are reweighted to match the data
- Create a "new simulation" by transforming weights to a proper function of the generated events

Machine learning is used to derive the reweighting functions

* Andreassen et al. PRL 124, 182001 (2020)

# Extracting particle information

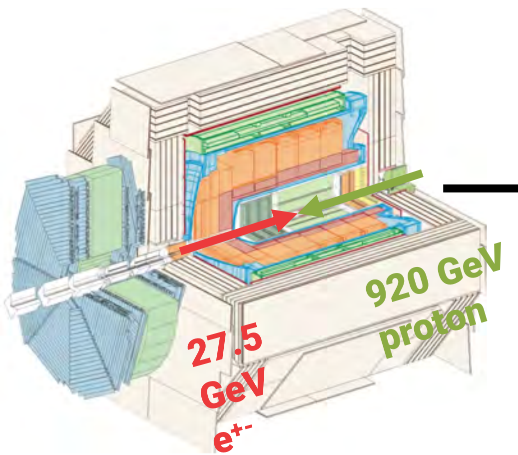
- **Particle collisions** are described by **graphs** where **particles** are **nodes**
- **Graph structure** naturally incorporate concepts such as varying number of particles and non-intrinsic ordering due to quantum mechanics
- Nearby particles carry the information on how they **decay** and **radiate**, encoded through **edges**
- Use a **Point Cloud Transformer*** model to learn the differences between particles during **unfolding**



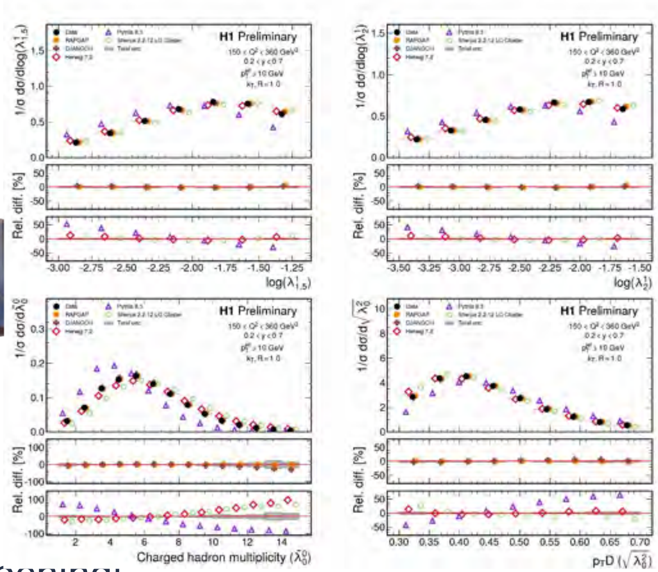

* **V. Mikuni** and F. Canelli 2021 *Mach. Learn.: Sci. Technol.* **2** 035027
Wang, Yue, et al. *Acm Transactions On Graphics (tog)* 38.5 (2019): 1-12.

# Experimental results

H1 Collaboration. H1prelim-22-034




To account for all uncertainties,
**2800 neural networks** were trained!

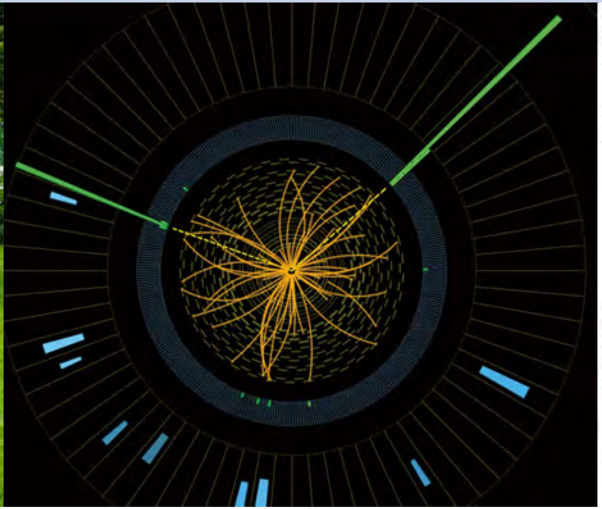



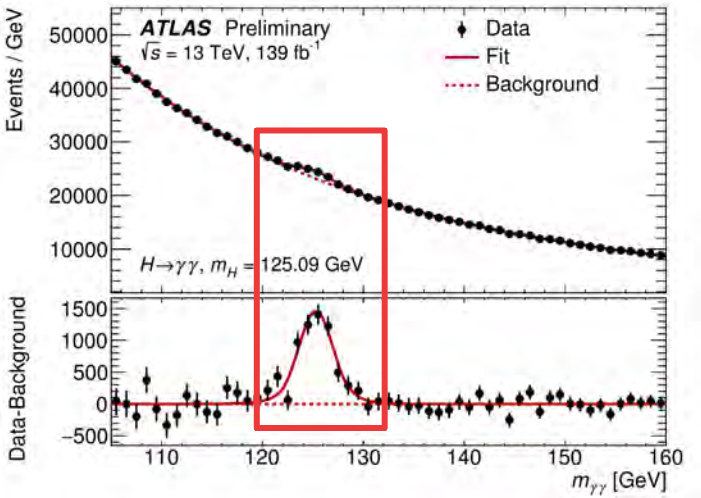

ML-based unfolding results using experimental data are already happening!

- Full unfolding procedure carried out with the Perlmutter Supercomputer
- **Multiple observables** are measured and **unfolded simultaneously** with **high precision**!
- **Multiple energy regimes** are investigated to highlight different physics

Picture from: https://www.symmetrymagazine.org/article/december-2013/four-things-you-might-not-know-about-dark-matter

# Anomaly detection

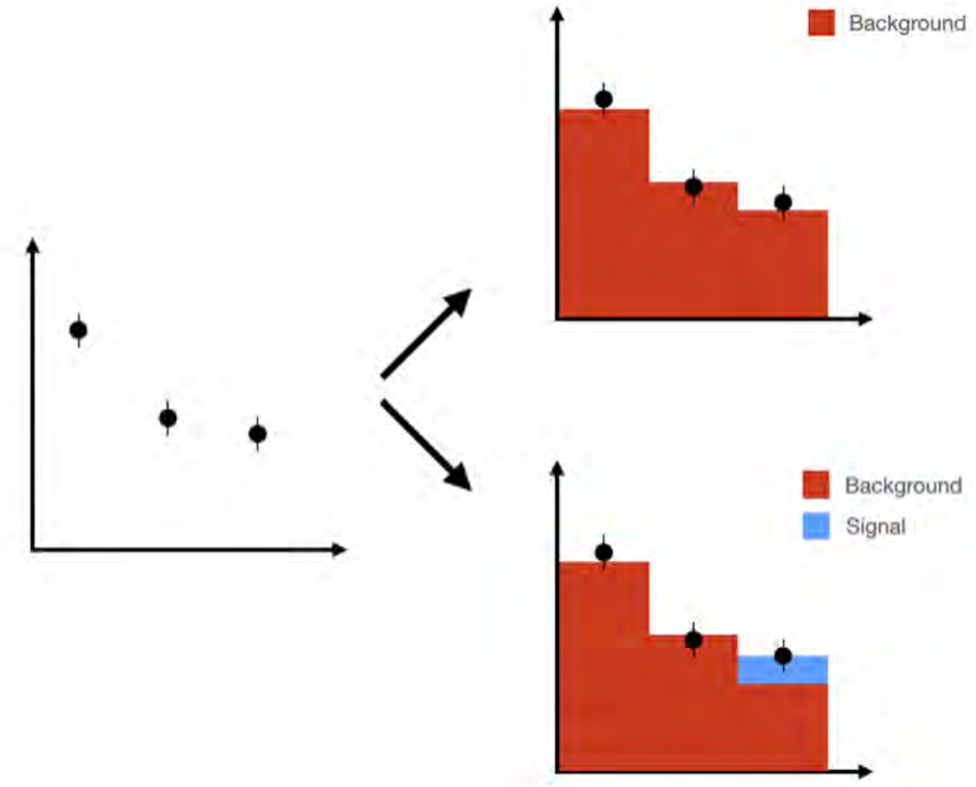





**Anomaly!** ? **Anomaly!**

- **Anomaly detection** is often associated to **outlier detection**
- For new physics, a single observation is not enough: an **ensemble** of observations is necessary to provide **context**

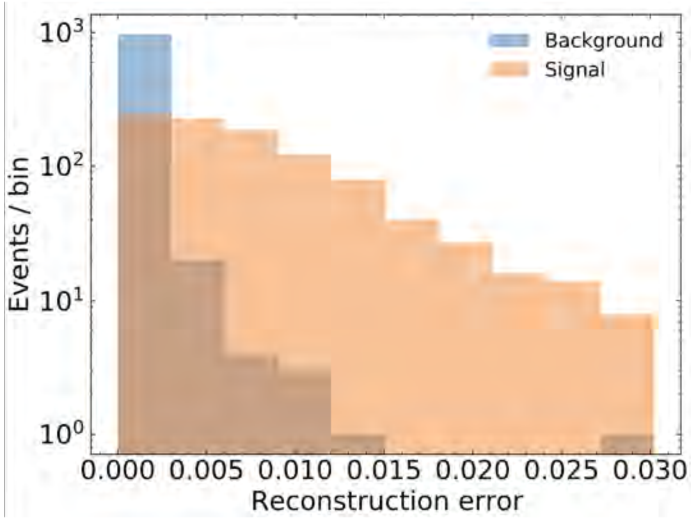
# Anomaly detection



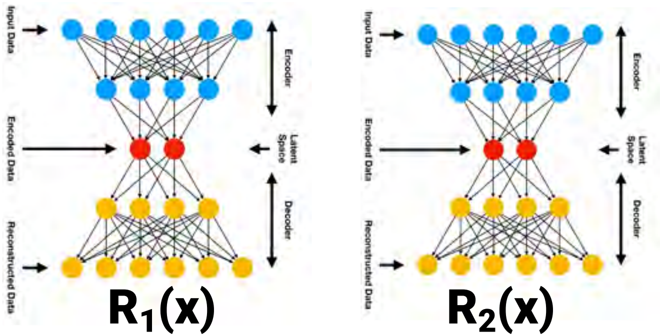

- A good **anomaly detection method** should be able to identify anomalies as well as provide context for false positives or **background events** events misidentified as anomalies: **False positives**

# Decorrelated autoencoders

- **Autoencoders** learns to **compress** and **decompress** data using background events
- **Anomalies** are often poorly reconstructed, yielding a **high reconstruction error**




- Train multiple **autoencoders** such that their reconstruction error is **independent** for background events



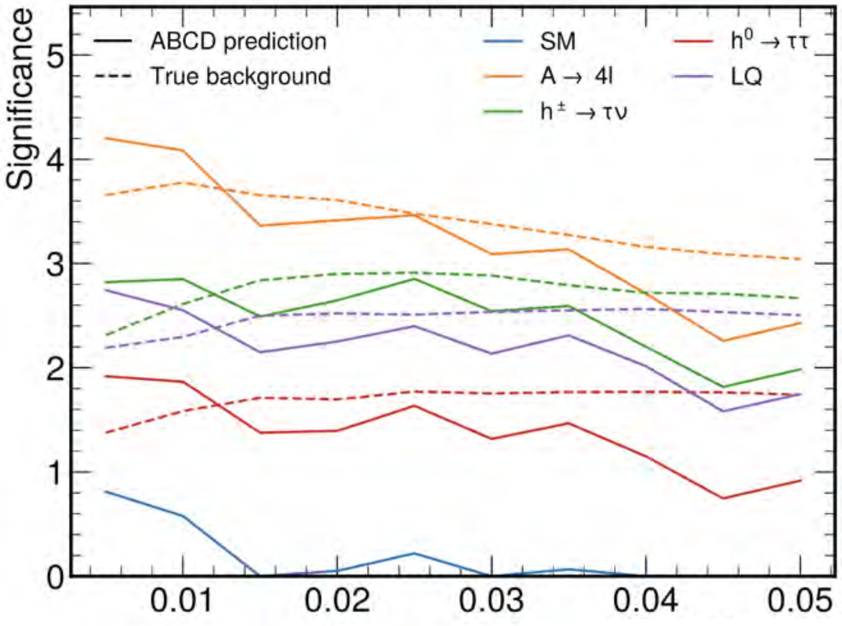

$$L[f_1, f_2, g_1, g_2] = \sum_i R_1(x_i)^2 + \sum_i R_2(x_i)^2 + \lambda\,\mathrm{DisCo}^2[R_1(X), R_2(X)]$$

**See also:** Kasieczka, G., Mastandrea, R., **Mikuni, V.**, Nachman, B., Pettee, M., & Shih, D. (2022). *arXiv preprint arXiv:2209.06225.*

# Anomaly detection performance

**No anomalies**

**Other colors**: datasets with **0.1%** anomalies and **99.9%** standard physics processes




False positive rate

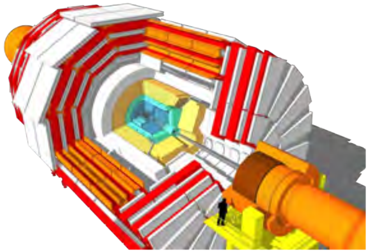
- **Number of false positives** determined using the independent reconstruction error
- In the **absence of new physics**, the algorithm reports the **correct number of observations**
- Anomalies identified as an **excess on the number of observations** translated as a **Significance** or **signal-to-noise ratio**

**Mikuni, Vinicius,** Benjamin Nachman, and David Shih. *Physical Review D* 105.5 (2022): 055006.

# Conclusions

See more **HEP-related developments** at
https://iml-wg.github.io/HEPML-LivingReview/

- Modern **data analysis** methods and **machine learning** are a **fundamental part of collider physics**
- In this talk I covered only a small part of a large number of exciting projects and ideas



- **Full detector simulation** is expensive and not easily scalable
  - **Surrogate models using ML** are necessary to keep up with the large amount of data collected by experiments
  - Use **score-based generative models** for the first time in particle physics to enhance simulation fidelity
- **Machine learning unfolding** to solve **inverse problems**:
  - Able to measure **precisely many observables simultane**
  - **First results using experimental data** are out!

- **Anomaly detection** is a new way to look for **new physics processes**
- Understanding the strengths and weaknesses of the algorithms is an important step to interpret results

# THANKS!

**Any questions?**

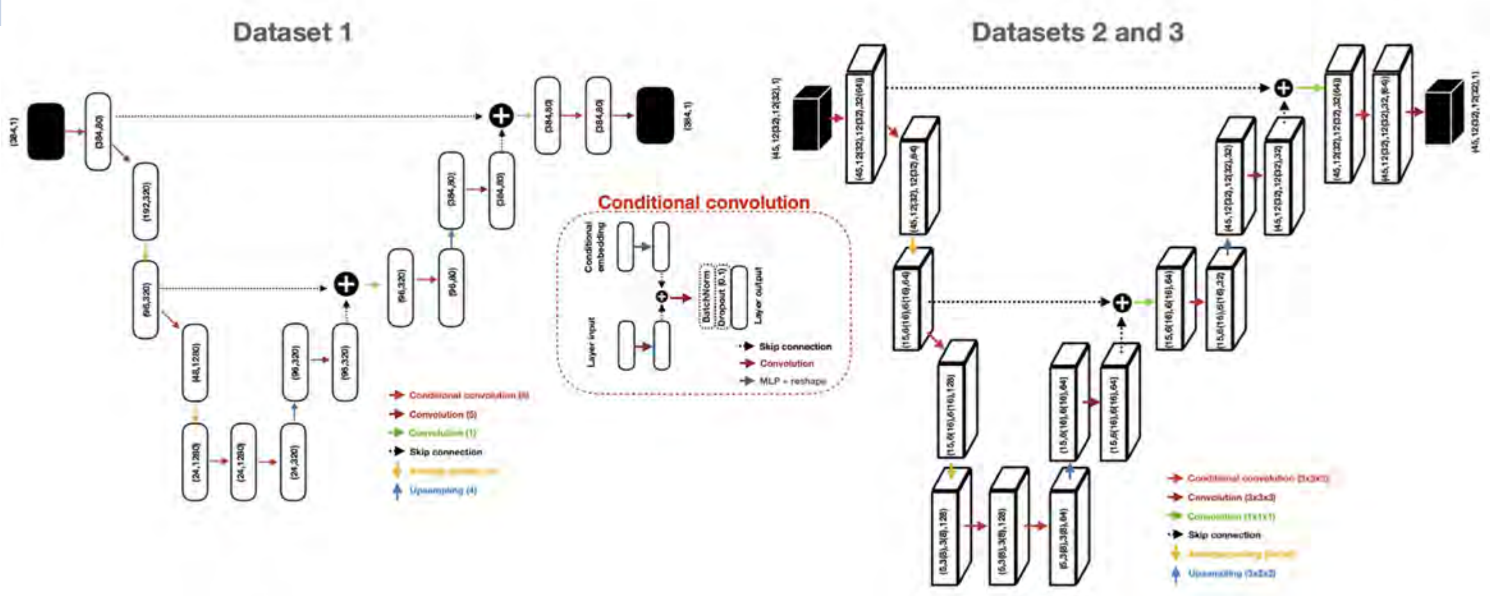
vmikuni@lbl.gov

vinicius-mikuni

# BACKUP

# Surrogate model for detector simulation

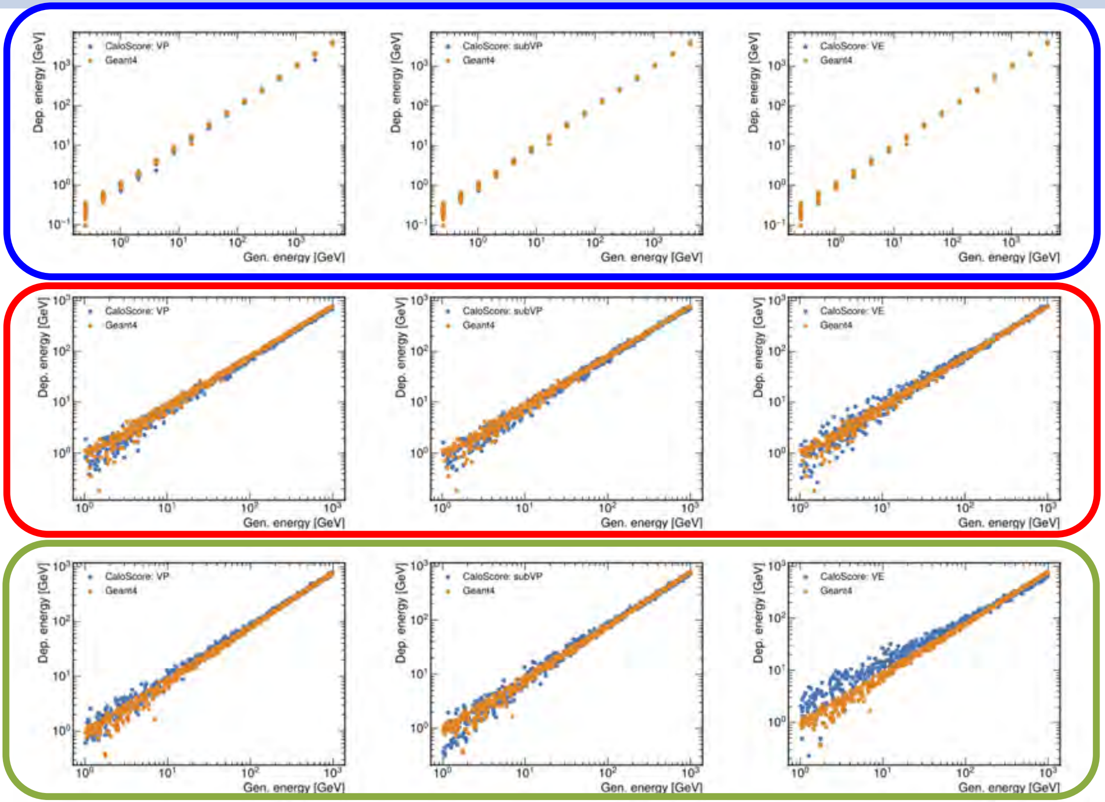
# Calorimeter shower generation




Very simple **U-NET** model used to build the score function

- Lots of new developments over the years, adding attention between layers, additional skip connections, but kept it simple for this application

# Results



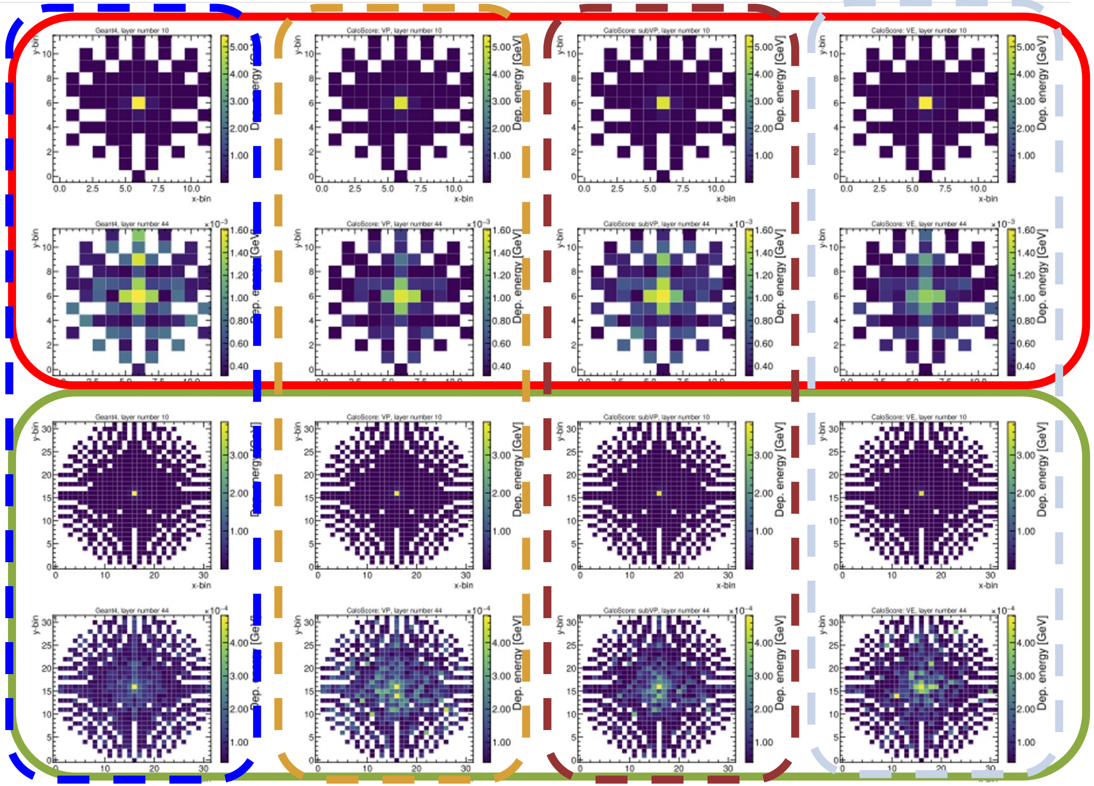

- **Deposited energy** (sum of voxels) vs. the **conditional energy**
- Good agreement between **full simulation** and **different diffusion models**
- **VE** shows the same shift observed for dataset 3

**Dataset 1**

**Dataset 2**

**Dataset 3**

# Results

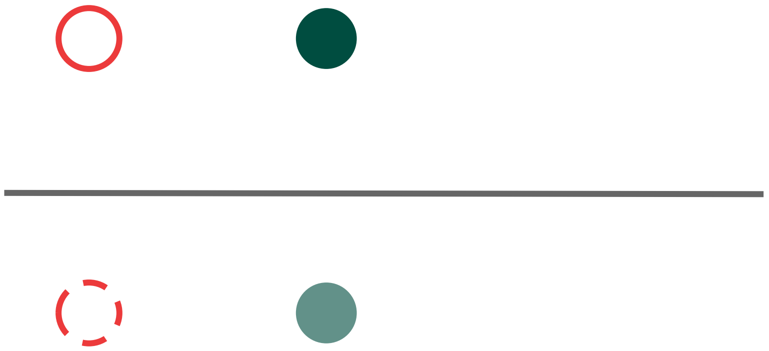



Weird shapes are a result of the coordinate transformation

# Unfolding

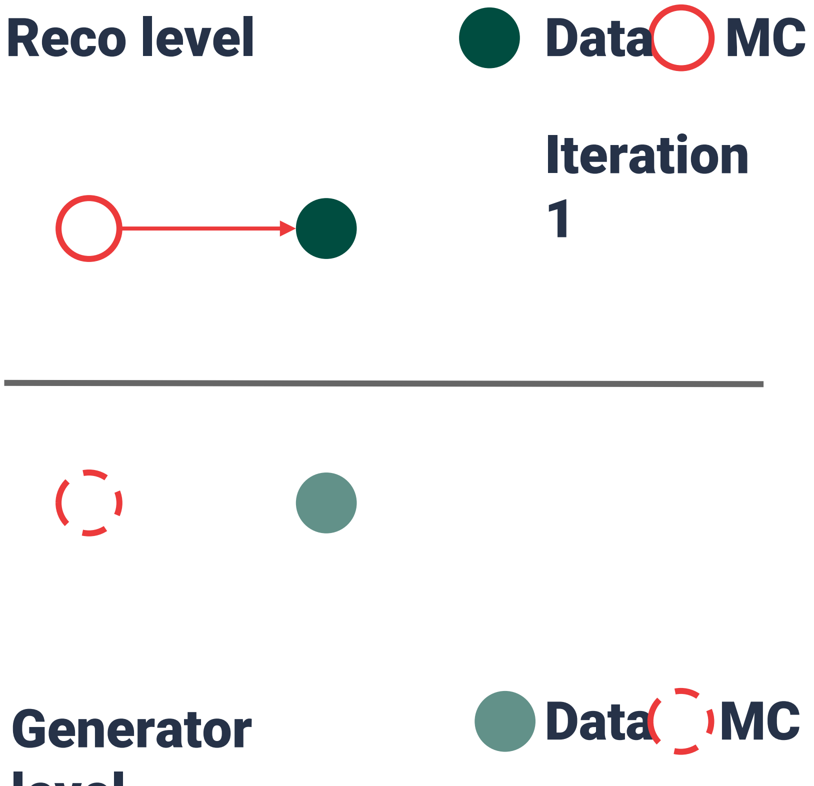
# Omnifold

**Reco level** Data MC



**Generator level** Data MC

# Omnifold




**Step 1:**

- Train a classifier to separate **data** from **MC** events
- Reweight **reco level MC** with weights:

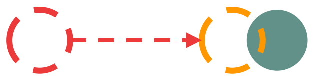
$$W(\mathrm{reco}) = p_{\mathrm{Data}}(\mathrm{reco})/p_{\mathrm{MC}}(\mathrm{reco})$$

# Omnifold

Reco level

Data MC

Iteration 1



**Step 2:**

- **Pull weights** from **step 1** to generator level events
- Train a classifier to separate **initial MC at gen level** from **reweighted MC** events
- Define a **new simulation** with weights that are a **proper function of gen level kinematics**

$$W(gen) = p_{weighted\ MC}(gen)/p_{MC}(gen)$$

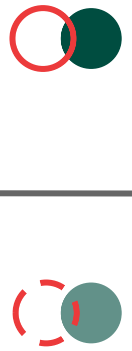
Generator level

Data MC MC reweighted

## Omnifold

**Reco level** **Data** **MC**

**Iteration 1**

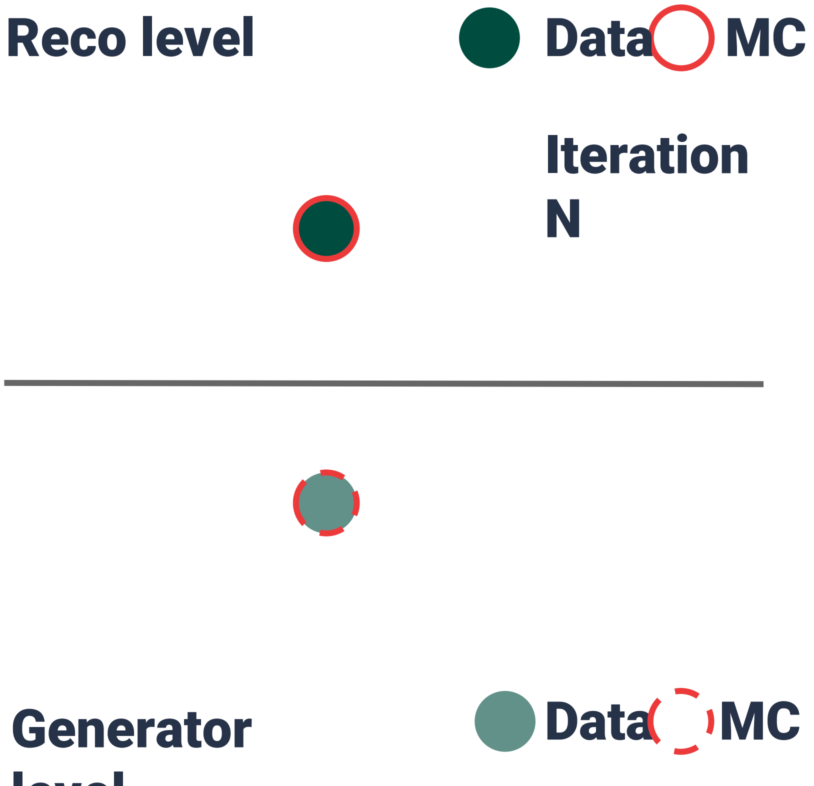


**Generator level** **Data** **MC**

Start again from **step 1** using the **new simulation** after **pushing** the weights from **step 2**

- Guaranteed convergence  to the maximum likelihood estimate of the generator-level distribution when number of iterations go to infinite
- In practice, less than 10 iterations are enough to achieve convergence

# Omnifold

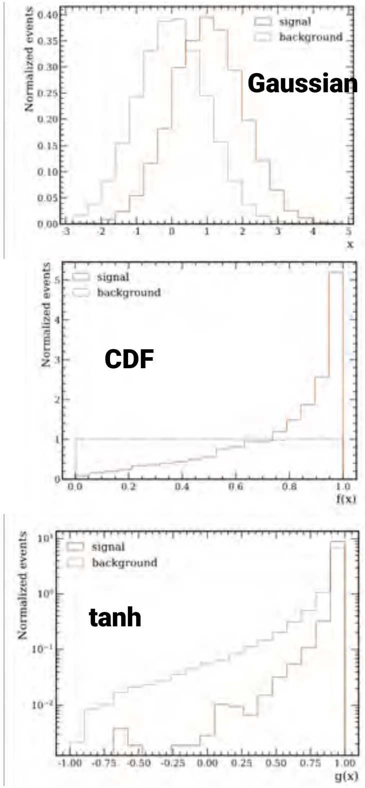


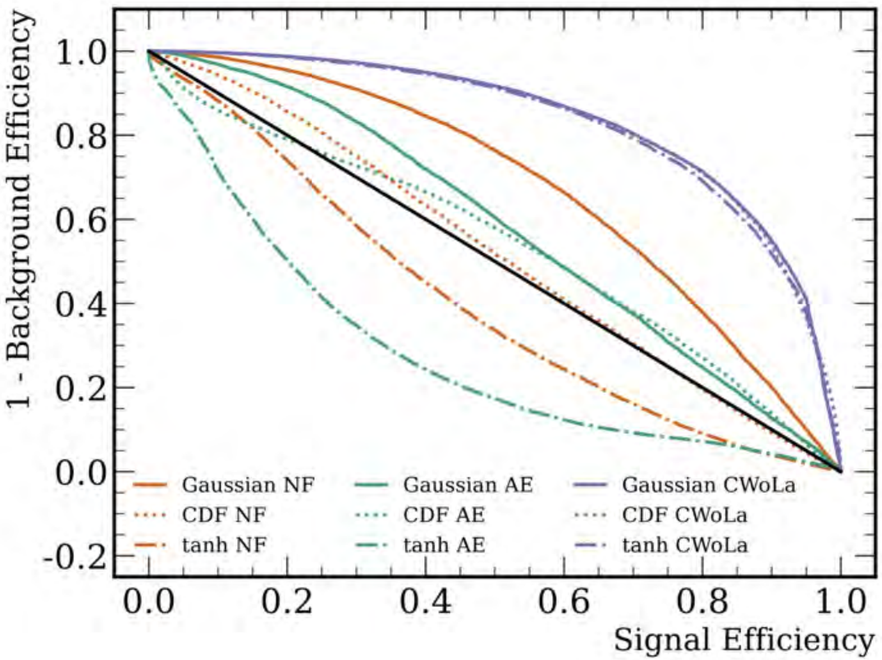

Start again from **step 1** using the **new simulation** after **pushing** the weights from **step 2**

- **Guaranteed convergence** to the maximum likelihood estimate of the generator-level distribution when number of iterations goes to infinite
- In practice, **less than 10 iterations** are enough to achieve convergence

# Anomaly detection

# Anomaly detection



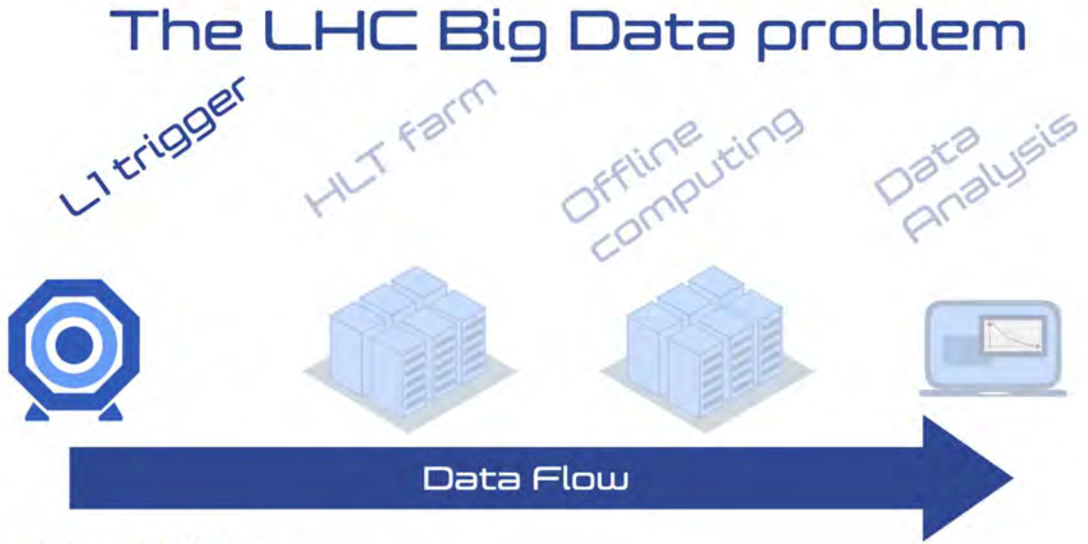




Kasieczka, G., Mastandrea, R., **Mikuni, V.**, Nachman, B., Pettee, M., & Shih, D. (2022). *arXiv preprint arXiv:2209.06225*.

- The **set of features** used to search for anomalies can also have a big impact on the algorithm performance, as statements regarding $\mathbf{p_s(x)}$ and $\mathbf{p_b(x)}$ are not invariant under **change of coordinates**

# Online compatibility

Slides from [Maurizio Pierini]




- 40 MHz in / 100 KHz out
- ~ 500 KB / event
- Processing time: ~10 µs
- Based on coarse local reconstructions
- FPGAs / Hardware implemented

- More than **99%** of events are rejected due to **bandwidth restrictions**
- Given the algorithm’s simplicity, it can also be deployed directly using modern hardware implementations such as **FPGAs**
- Possibility to identify **anomalous events** and store the information for further analysis