

 **ATLAS**  
EXPERIMENT  
Candidate Event:  
 $pp \rightarrow H(\rightarrow bb) + W(\rightarrow \mu\nu)$   
Run: 338712 Event: 335908183  
2017-10-19 23:31:18 CEST

# Generative Models in HEP: Examples from the experiments



Sofia Vallecorsa

9th December 2024

# Table of Content

---

## **Introduction**

**A few words on Generative Models**

**Running in real time, challenges and constraints**

Anomaly detection

## **Running Reconstruction**

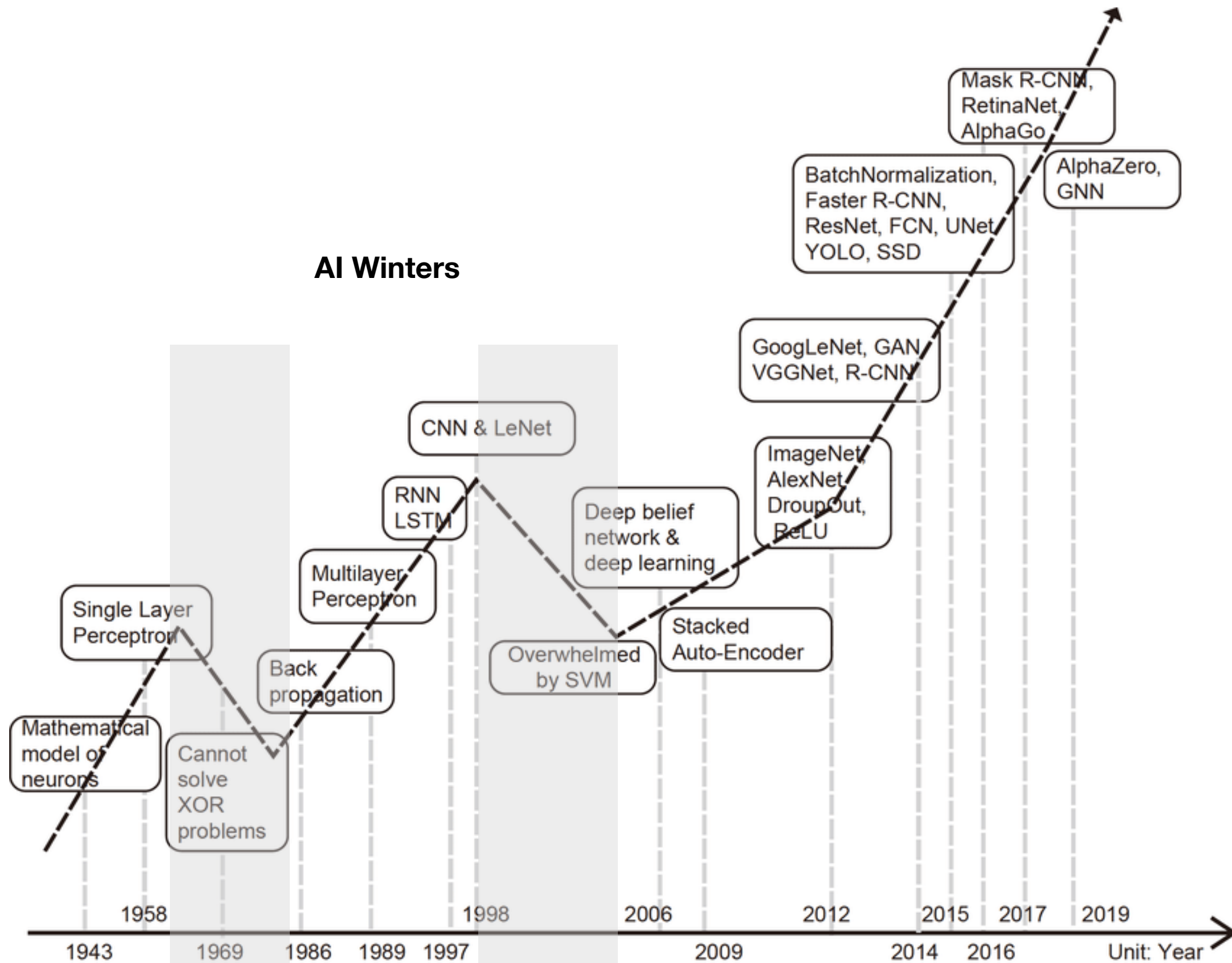
Jets

## **Simulating LHC events**

Event Generation & Detector Simulation

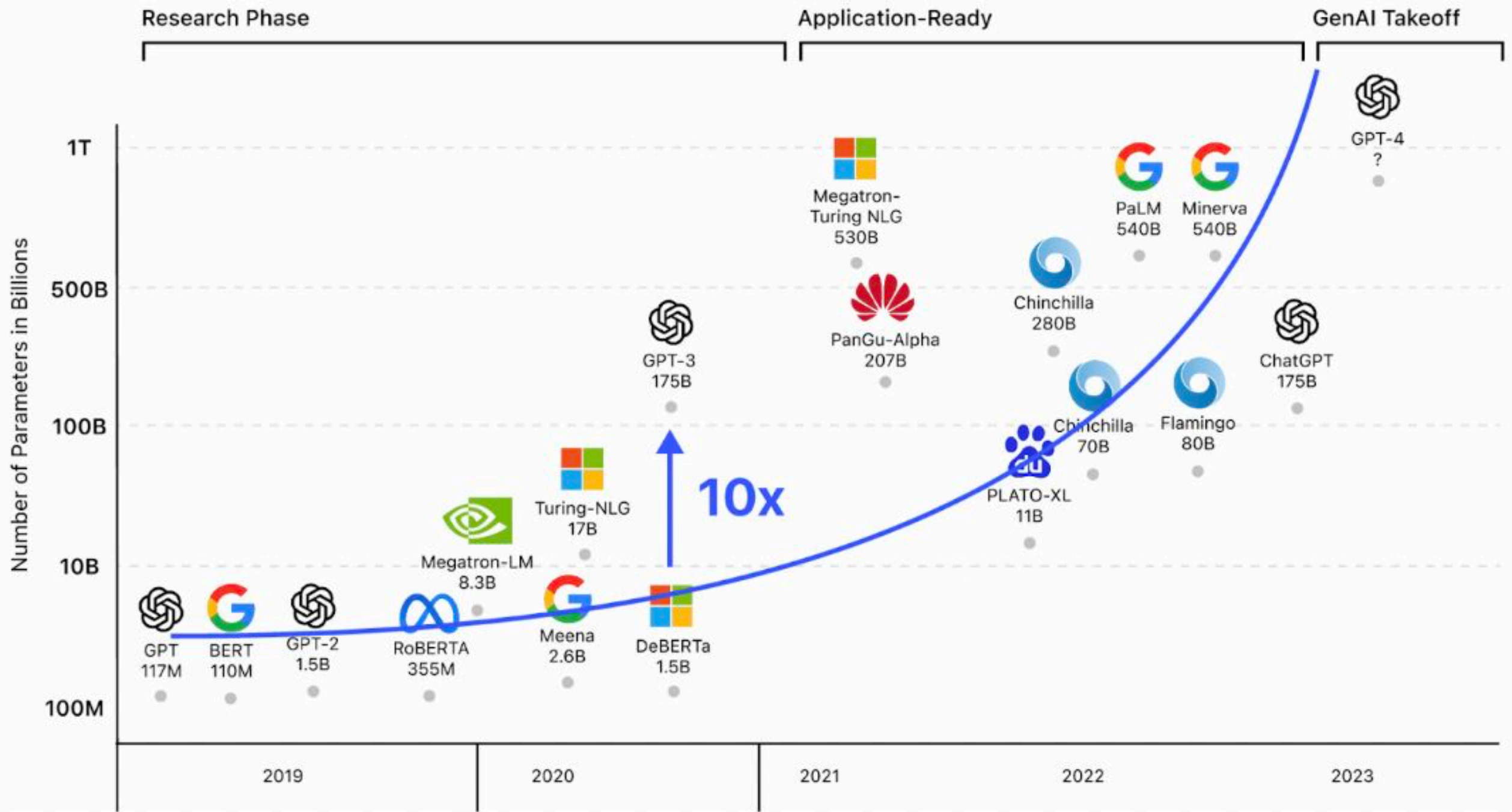
## **Summary**

# A bit of history





# Then.. AI TakeOff....



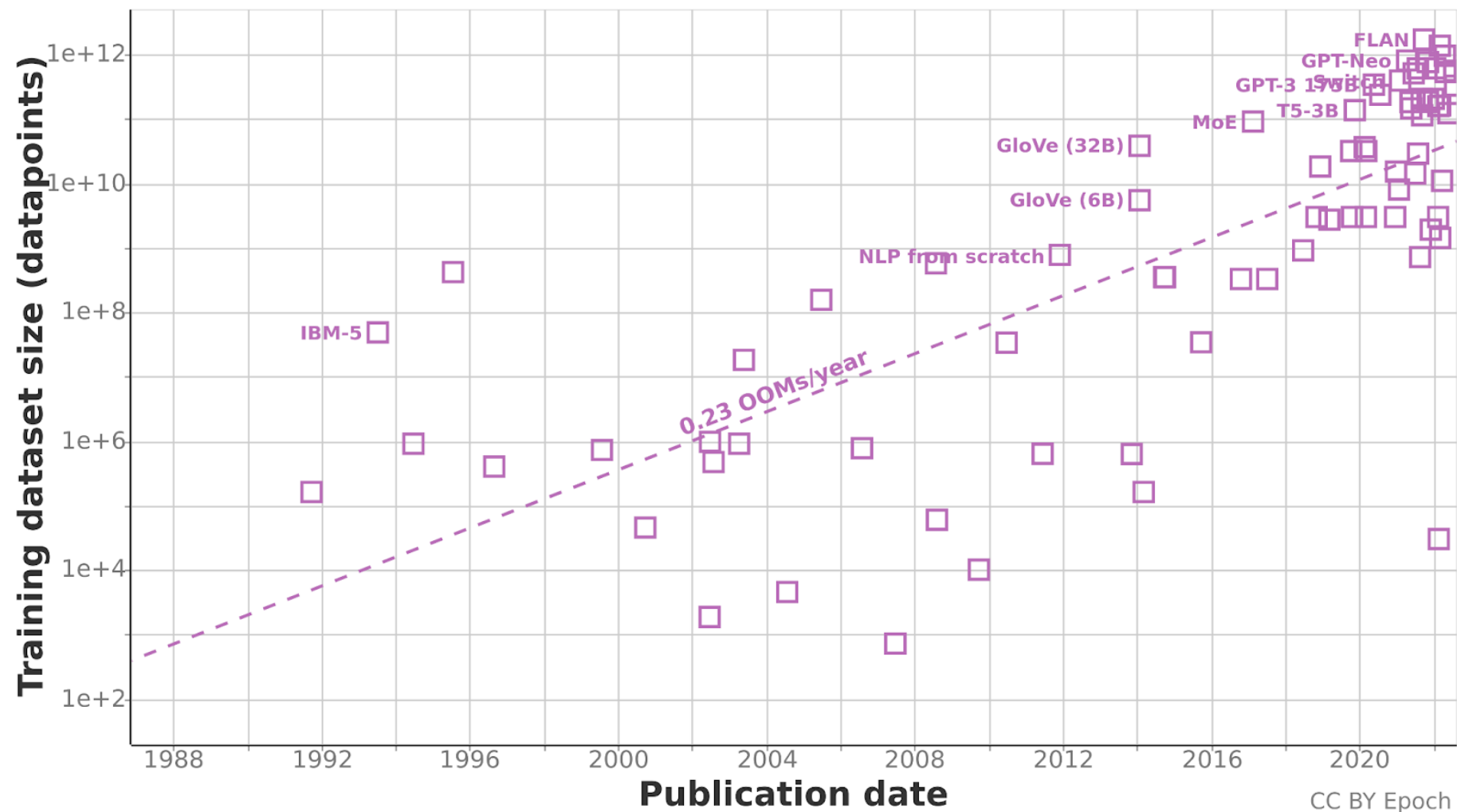


# Machine learning at scale, for science

**Machine learning has been proven a very good tool to:**

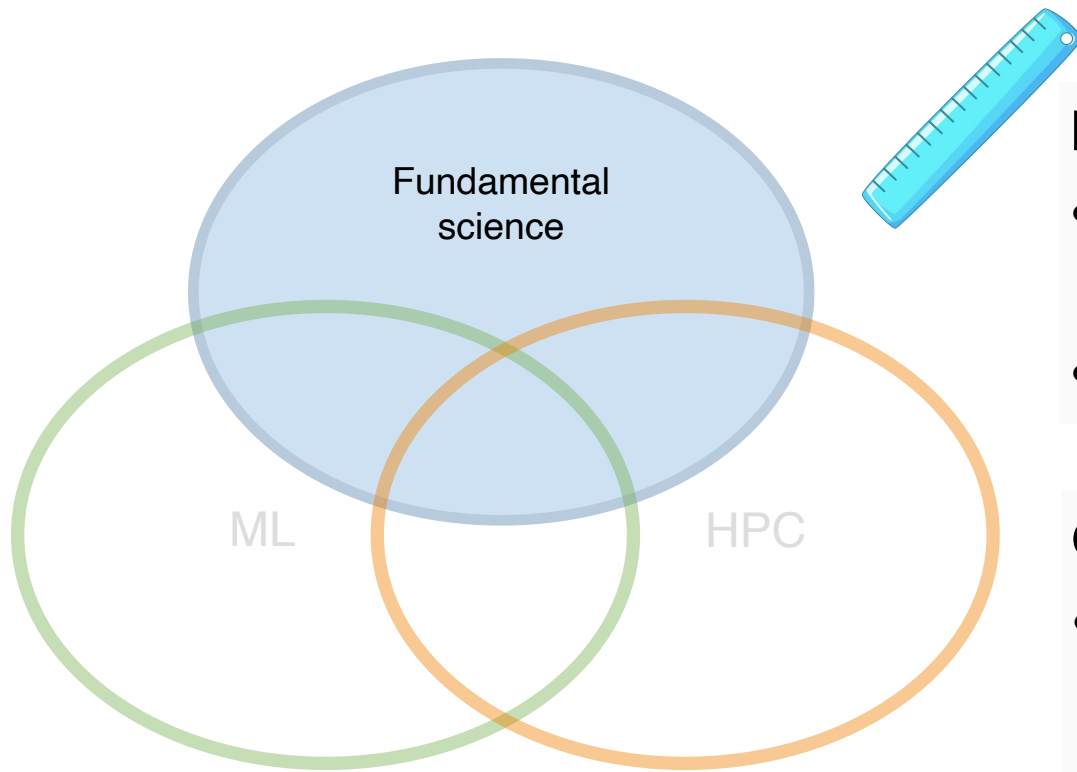
- Extract information from (very large) datasets
- Efficiently analyse very large amounts of data
- Easily handle data from different sources
- Scalability to HPC environments

**Observation based datasets in physics are comparable or larger than these!**



**Can we use these tools for fully data-driven science?**

# Scientific opportunities

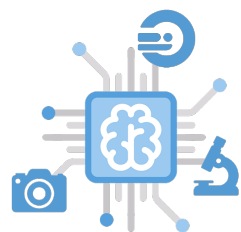


## Multi-scale dependencies:

- **Model complex higher-order, statistical relationships between observations, fields, ...**
- improve current simulations

## Compact representations:

- **Condense dataset information in a compact representation**
- eg. condense the info in a few GB rather than TB



## Multi-source models:

- **Enable multimodal and multi-source learning**
- eg. build models based on scientific data, GDP, birth rate etc..



## New discoveries:

- **Explore the potential of unsupervised learning to extract new information directly from data**
- Learn unknown correlation patterns

# Deep Learning in HEP

## Re-cast physics problems as “DL problems”

Interpret detector output as **images** and apply techniques borrowed from **computer vision**

Interpret physics events as **sentences** and apply **NLP techniques**

Better performances if applied directly to “**raw**” data

## Adapt DL to HEP requirements

In terms of model **interpretability**

Results **validation** against classical methods

Detailed **systematics**

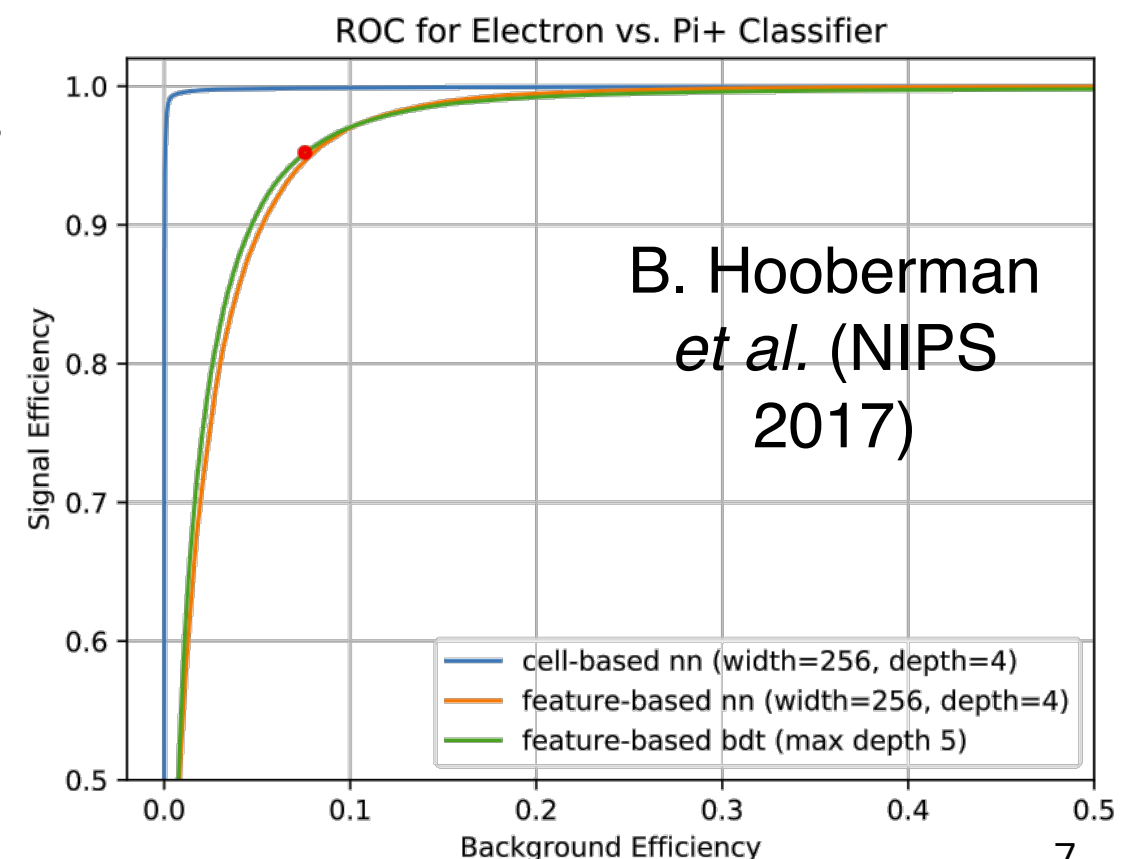
Adopting “new” computing models

**Accelerators** and dedicated hardware

**HPC** integration

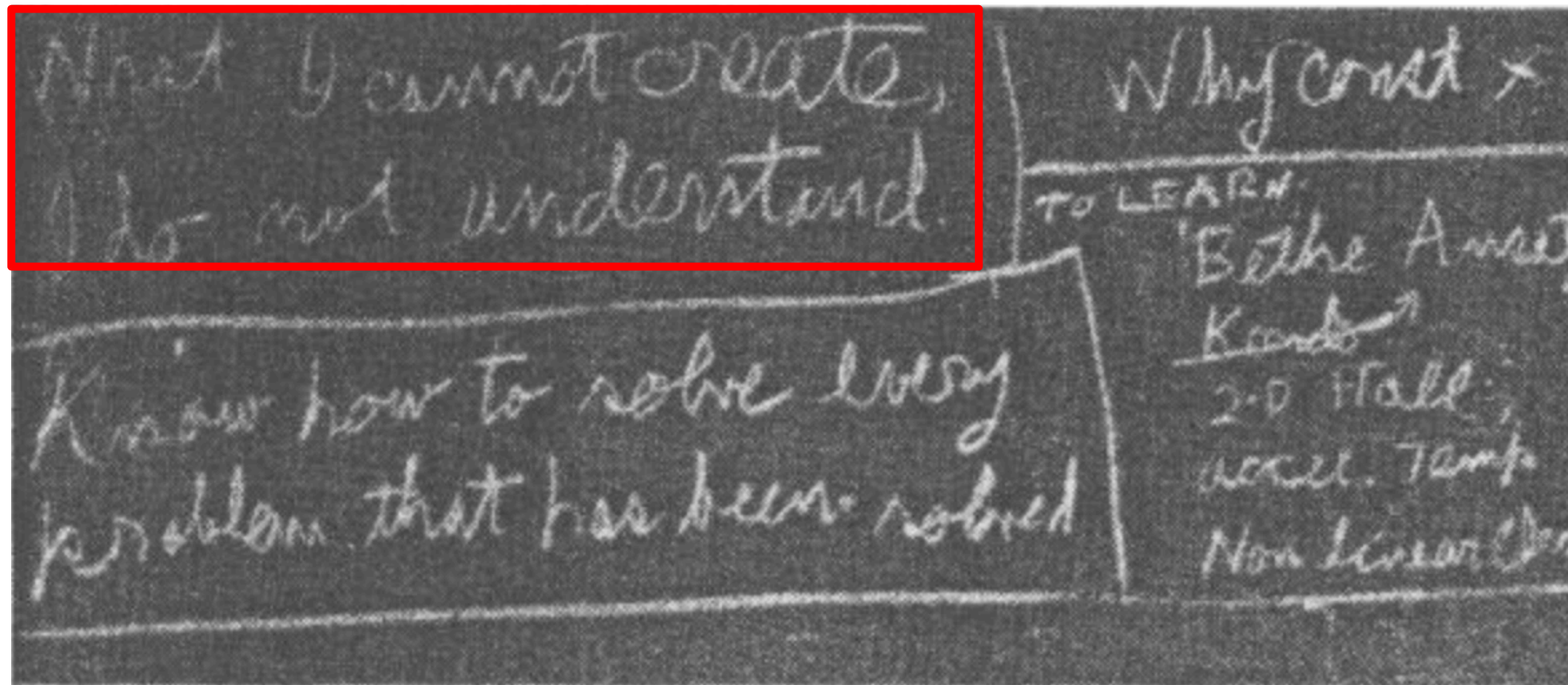
**Cloud** resources

**Big Data** platforms





# Generative Models



# Generative models

---

## The problem:

Assume data sample follows  $p_{\text{data}}$  distribution

Can we draw samples from distribution  $p_{\text{model}}$  such that  $p_{\text{model}} \approx p_{\text{data}}$ ?

# Generative models

---

## The problem:

Assume data sample follows  $p_{\text{data}}$  distribution

Can we draw samples from distribution  $p_{\text{model}}$  such that  $p_{\text{model}} \approx p_{\text{data}}$ ?

### Maximum Likelihood Estimator:

- Assume some form for  $p_{\text{model}}$  (prior knowledge, parameterized by  $\theta$ )
- draw samples from  $p_{\theta^*}$

$$\theta^* = \arg \max_{\theta} \sum_{\mathbf{x} \in \mathcal{D}} \log(p_{\text{model}}(\mathbf{x}; \theta))$$

Generative models don't look for mathematical expression of  $p_{\text{model}}$

Train NN as a generator  $g: \mathbb{R}^m \rightarrow \mathbb{R}^n$

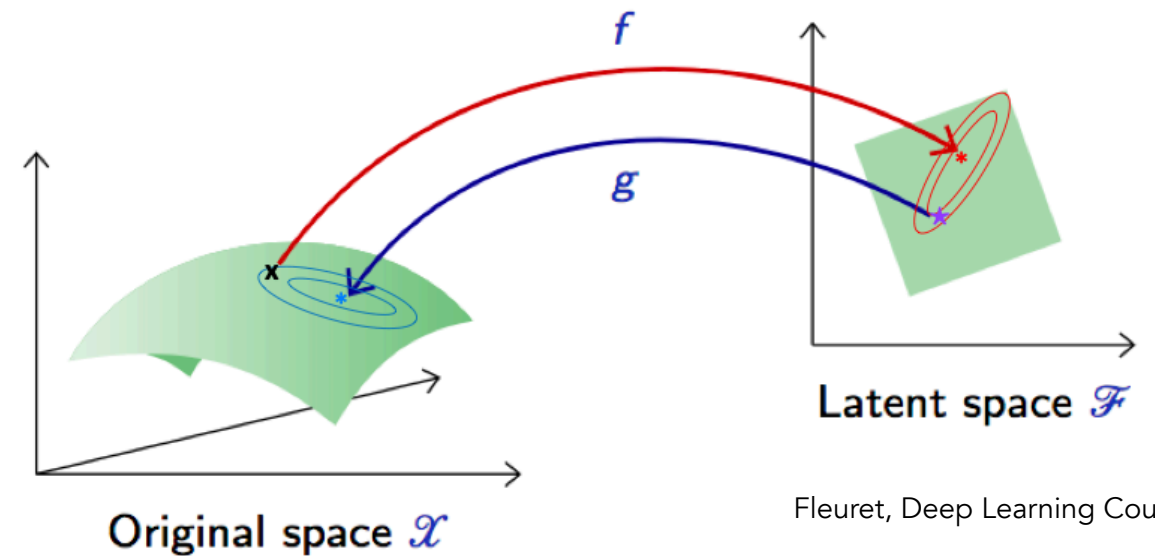
that maps samples from a tractable distribution supported in  $\mathbb{R}^m$  to points in  $\mathbb{R}^n$

31



# Latent Representation

- Information content is preserved within a **hidden manifold with lower dimension**
- Can manipulate **latent space** (style specification, hypothesis testing directly in data, ...)
- Can optimise latent representation according to a specific task (**guided compression**)
- Can help with **multi-modality**

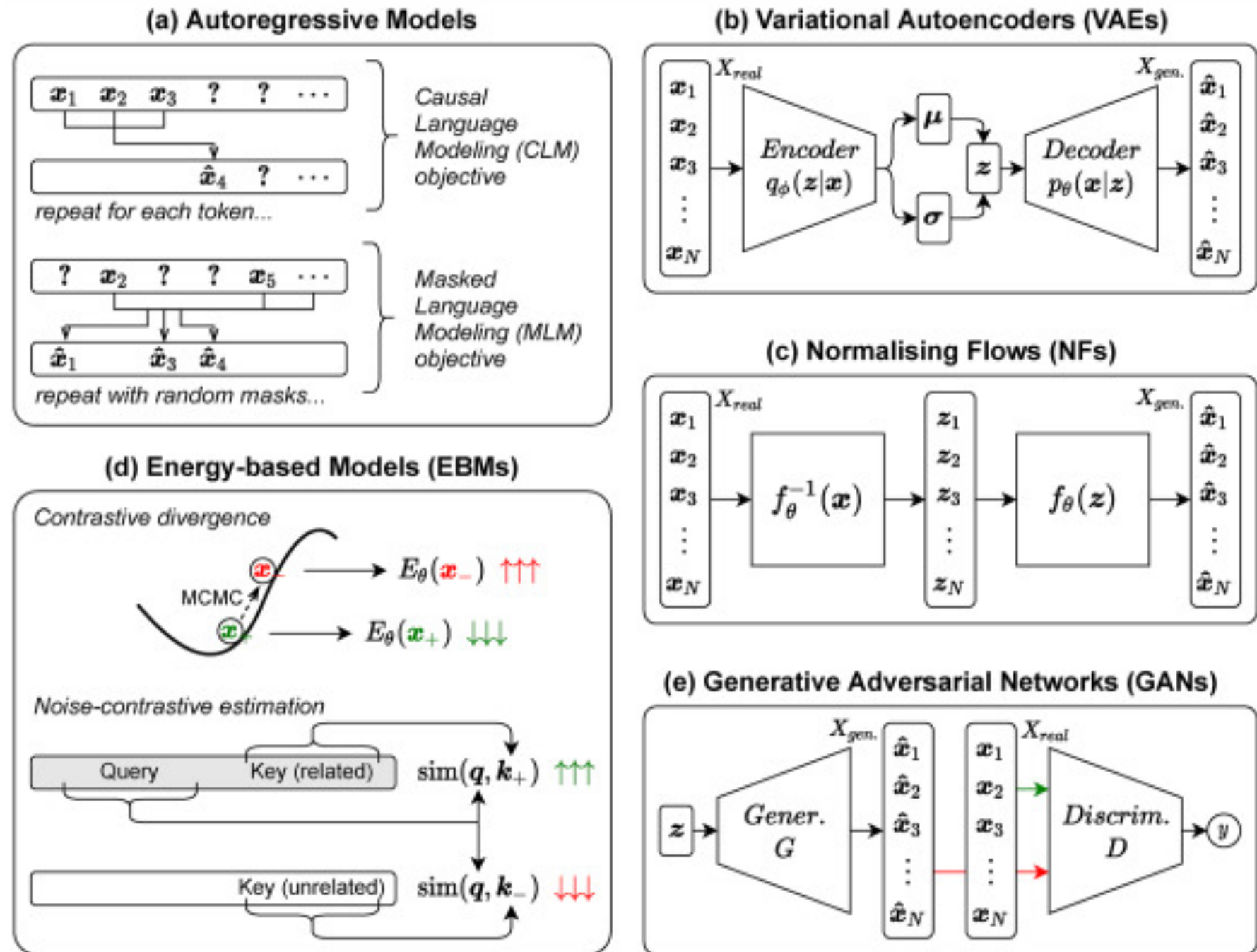


Fleuret, Deep Learning Course

**NB: Problems exhibiting complex symmetries may benefit from latent space representations connected to the specific underlying symmetry group!**

# Deep Generative Models

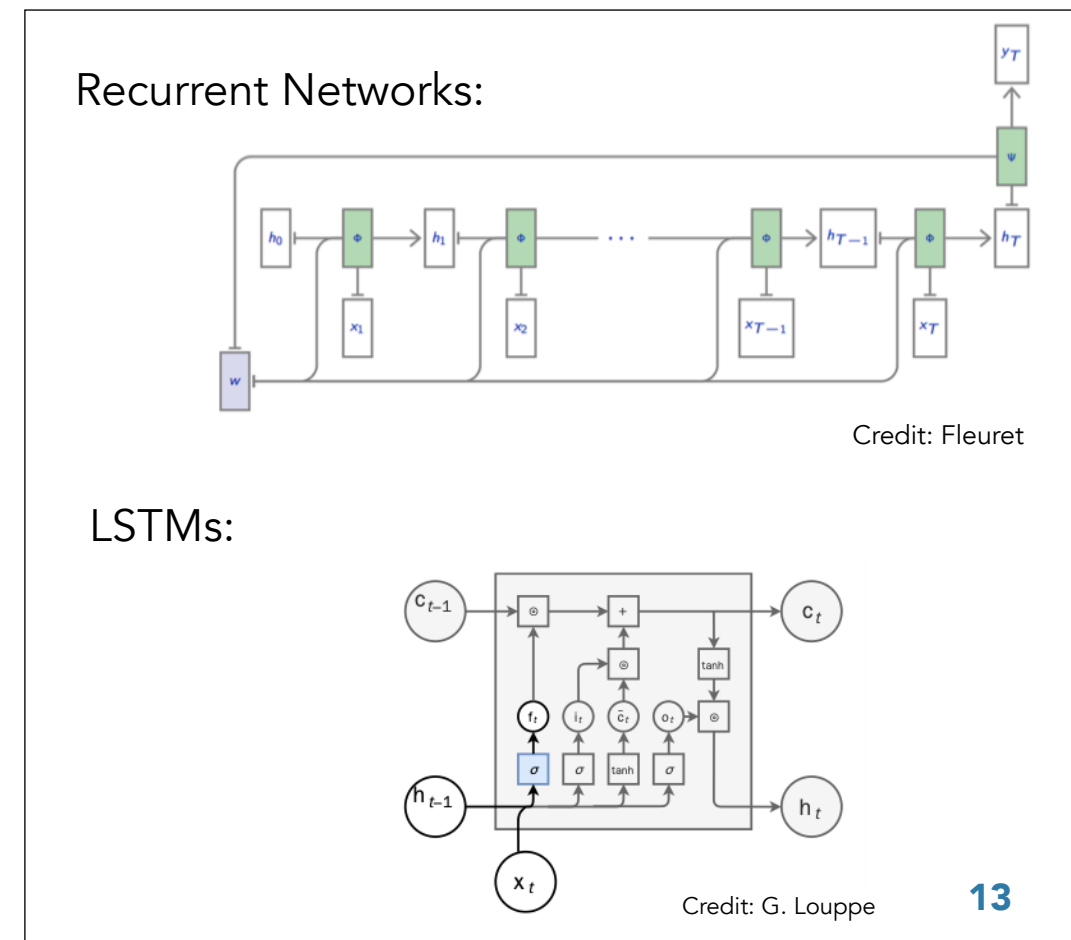
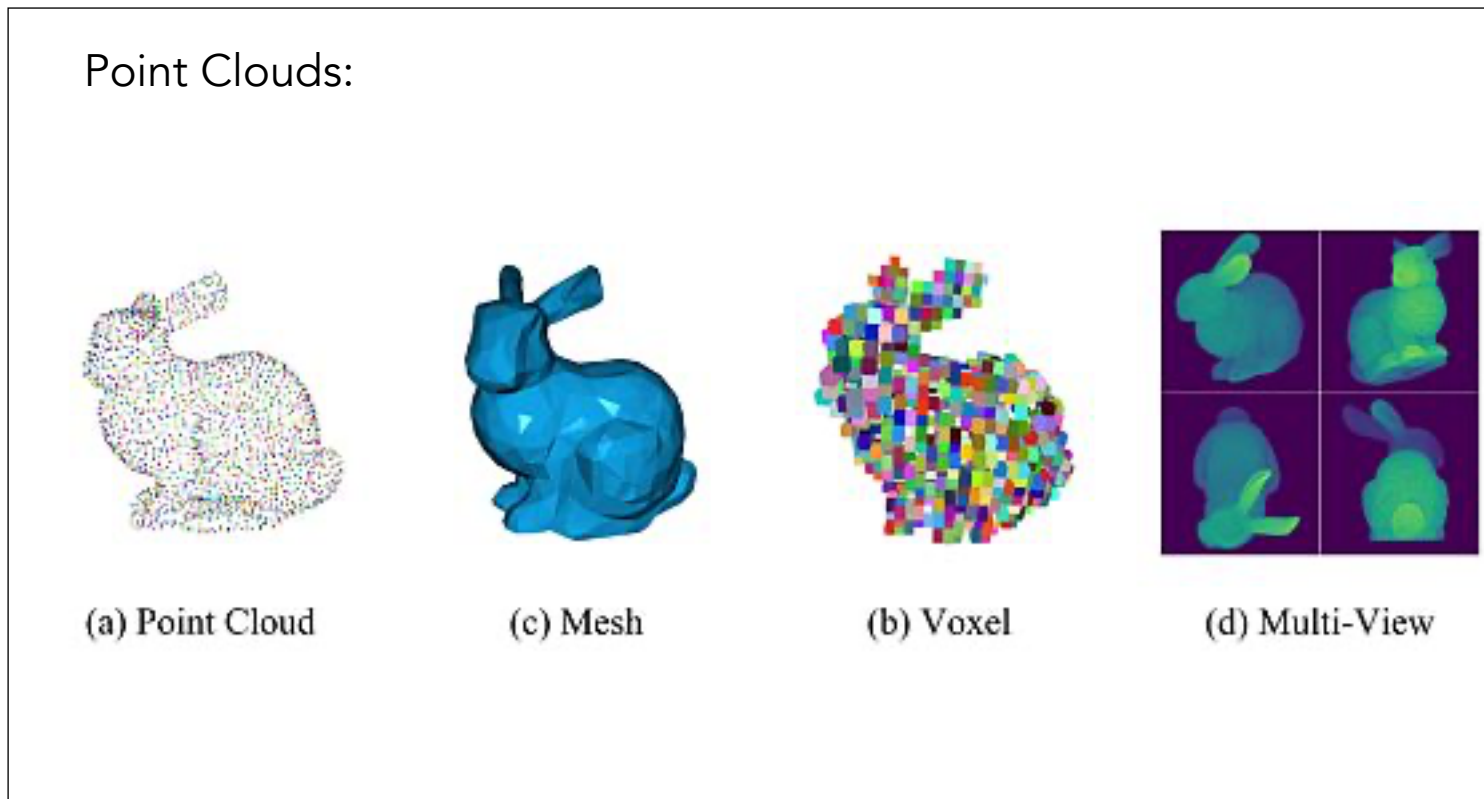
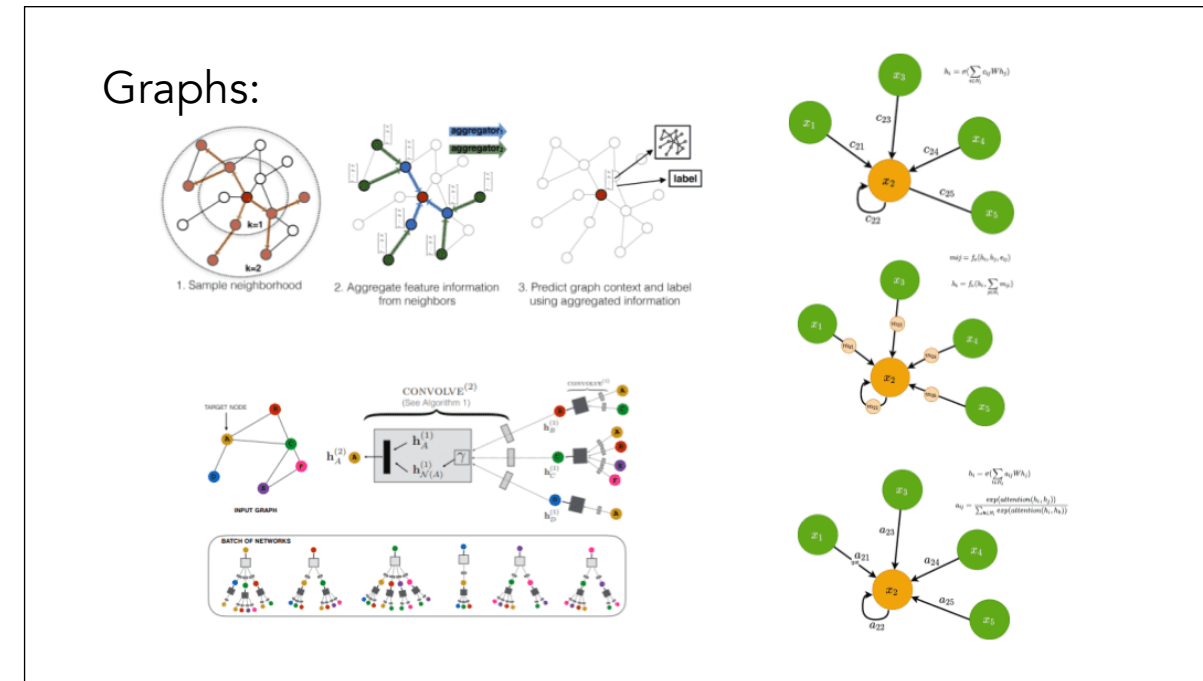
Deep models allow higher levels of abstraction and improve generalization wrt to shallow models



See Danilo Rezende tutorial on Deep Generative Models

# Different primitives for different data representations

- Perceptrons and MLP
- Convolutions
- Graphs
- Recurrent Units (and LSTMs)
- Point Clouds
- ...

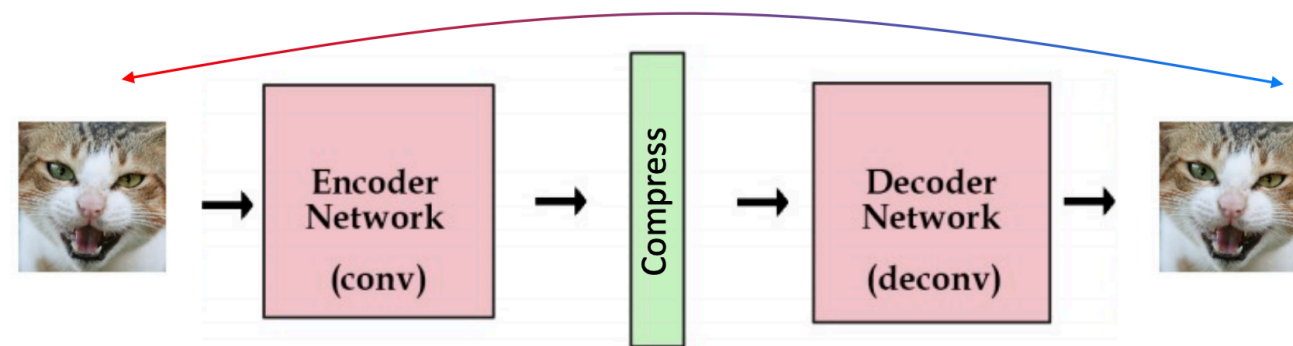




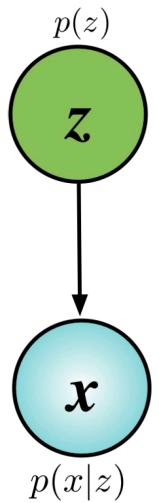
# Auto-Encoders

Example of latent variables models (and implicit...)

Ex. Auto-Encoder



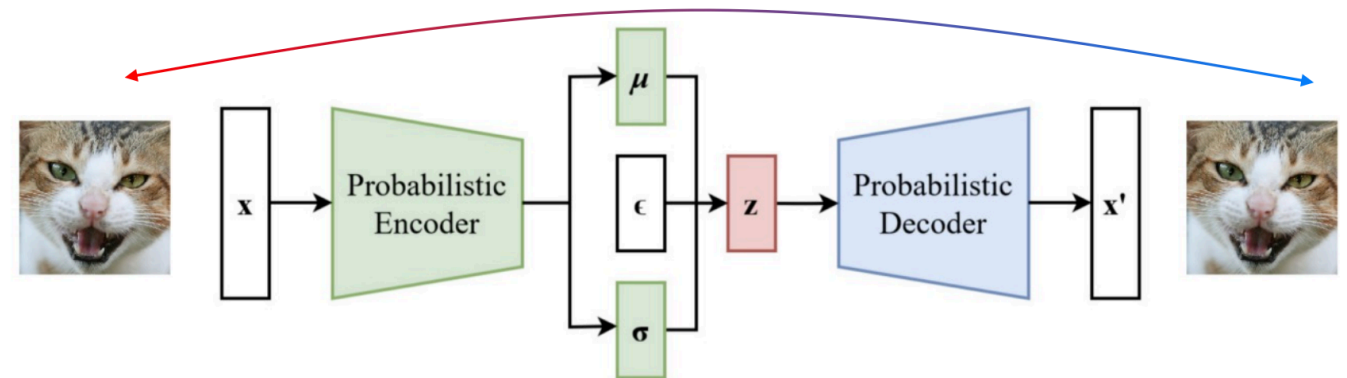
$$x \in \mathbb{R}^{d_x} \quad z \in \mathbb{R}^{d_z} \quad \theta \in \mathbb{R}^{d_\theta}$$
$$\mathcal{D} = \{x_i\} \quad i \in \{1, \dots, N\}$$



Ex. Variational Auto-Encoder

**Explicit constraints** on encoded representations (learn the **latent variable distribution**)

Two components in the loss function (**reconstruction loss** and **KL divergence** to constrain latent to prior)



Multiple AE variants and flavours have been developed in the past few years

# Diffusion models

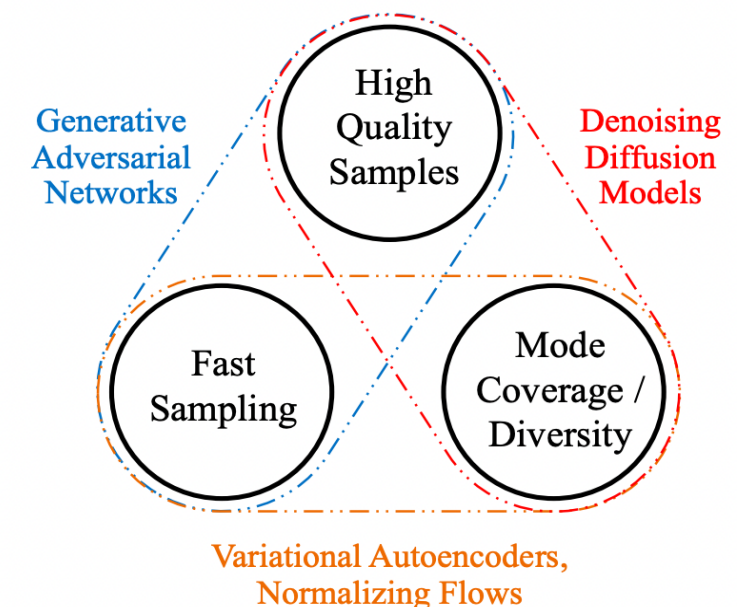
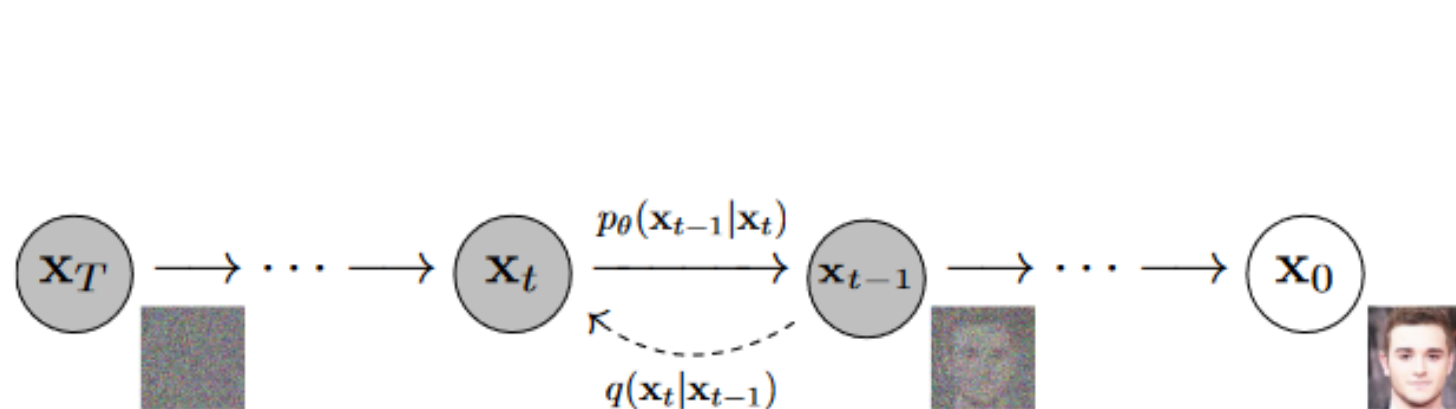
**Parametrized Markov Chains** trained using variational inference to produce samples matching the data after finite time.

Chain transitions are **reverse diffusions** (gradually adding noise to the data)

Ex. Diffusion Denoising Probabilistic Models (DDPM, [arxiv:2006.11239](https://arxiv.org/abs/2006.11239)) based on U-Net:

Iteratively add Gaussian noise to input image, eventually reaching pure noise

Generation process **inverts the diffusion**: start from pure noise sample, then iteratively de-noise it.



# Attention and Transformers

---



# Seq2seq models and the information bottleneck

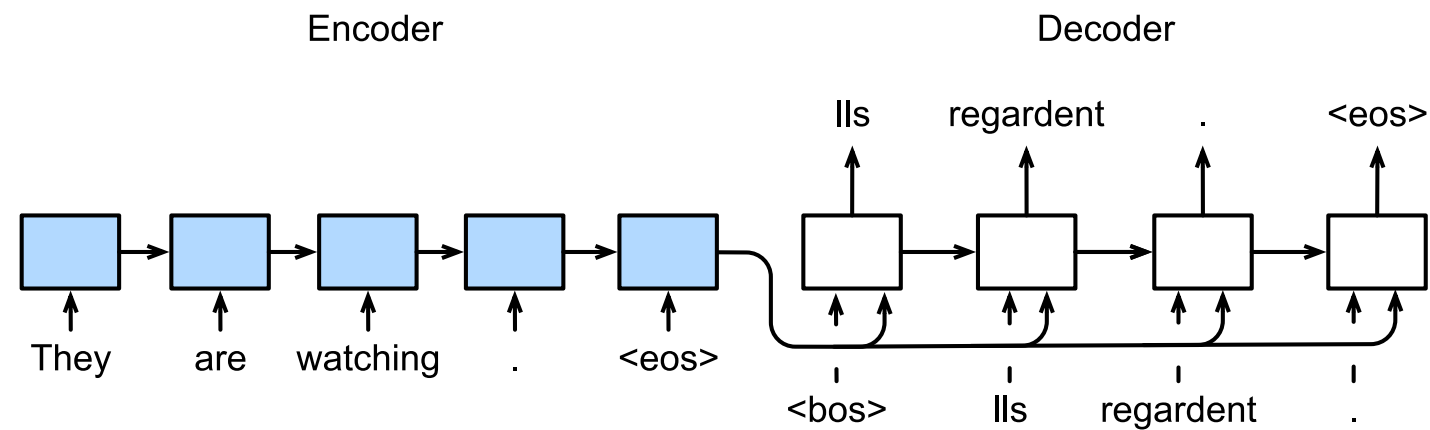


Image Credit: d2l.ai

## Seq2seq models analyse sequences

Predict probability distributions of the next token given previous context

Encoder compresses the sequence in a fixed size vector

## Compression in fixed size latent vector is a bottleneck

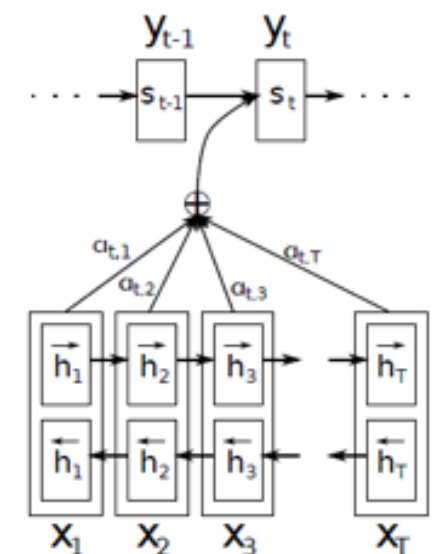
Need a mechanism to **focus on most relevant** input tokens at each prediction step

## Introduce (Self-) Attention Maps

Use **softmax to calculate probability** (maintain differentiable architecture)

Output is **independent of the order** of input examples (set instead of sequences)

Highlight **relationships between input elements**



Attention mechanism as originally formulated in a bi-directional LSTM Auto-Encoder (arxiv:1409.0473)

# Attention - Transformers

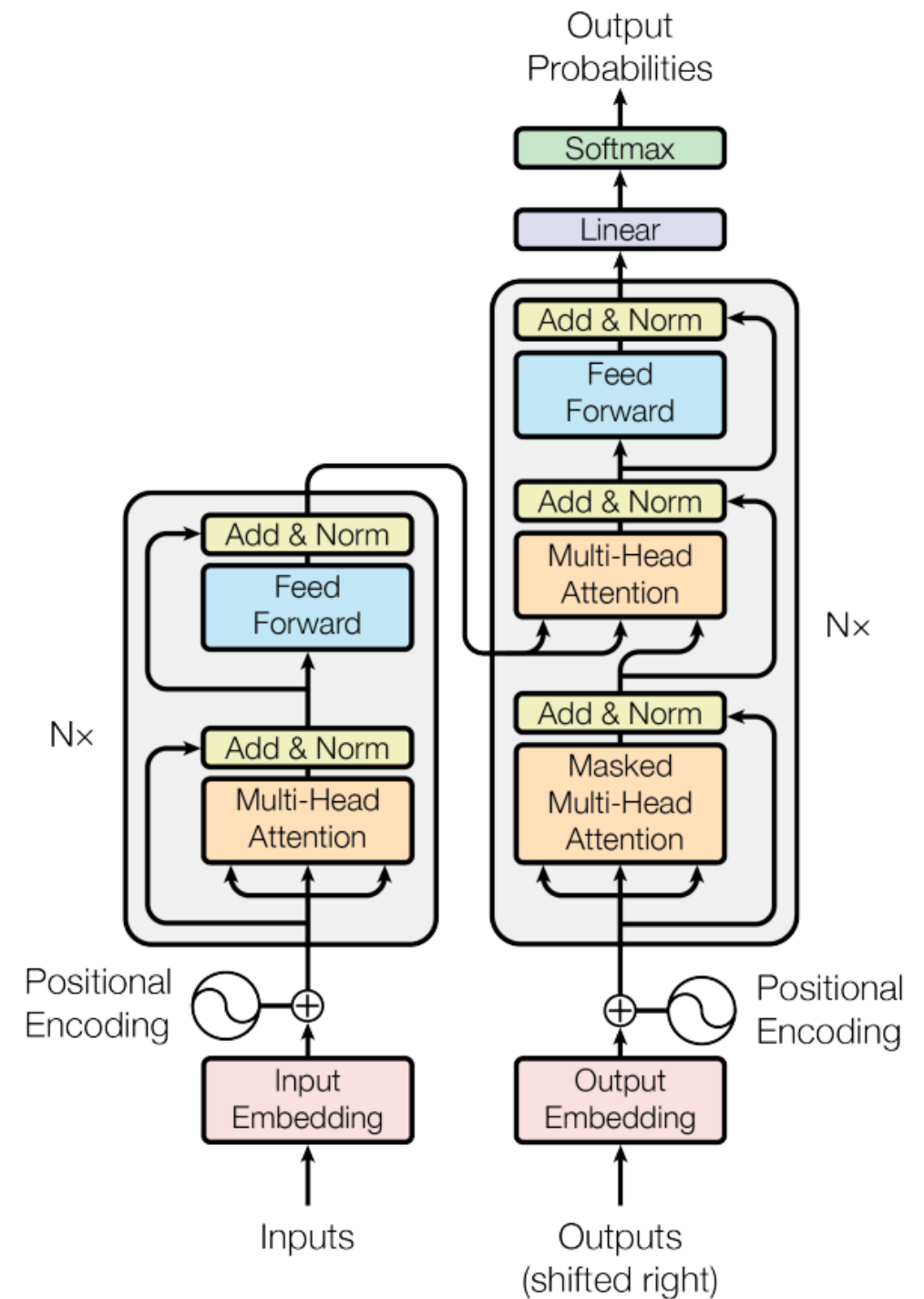
Transformer components include:

Multi Head **Attention**

**Normalisation** layers

Position Independent **Feed Forward Layers**

**Skip Connections**



See tutorial G.. Weiss tutorial at IML workshop :  
<https://indico.cern.ch/event/1297159/>

Vaswani et al., *Advances in Neural Information Processing Systems*, 2017, 5998–6008

# Example applications

---

# Online Machine/Deep Learning

## LHC Run3 Fact sheet:

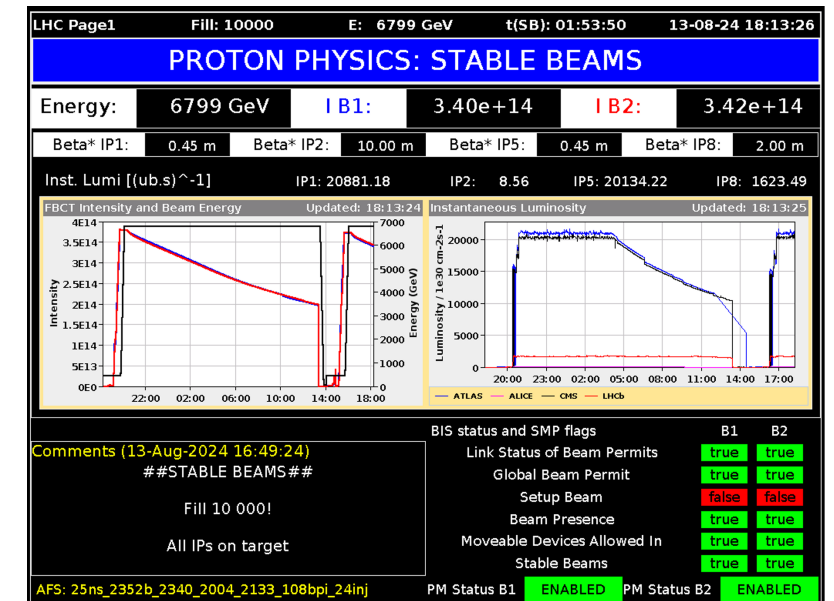
Since 2022, collisions at 6.8 TeV

25 ns bunch crossing

Peak collision rate at 30 MHz (2017-2018)

Peak instantaneous luminosity of  $2 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$  (2023)

About 50 pileup collisions



<https://home.cern/news/news/accelerators/accelerator-report-10-000-lhc-fills>

## Many ML/DL applications for real time detectors operation:

Data Quality Monitoring , Adaptive Data Acquisition Systems , Triggers

## Constraints on Latency:

Accelerate inference through dedicated ASICs, FPGAs

## Constraints on Model Complexity:

Reduce model size through quantisation, compression, distillation, ...

## Constraints on the quality of data available:

Input features are known with limited resolution (or limited detector information)

—> is this a limitation for ML/DL ?

# Anomaly detection for model independent searches

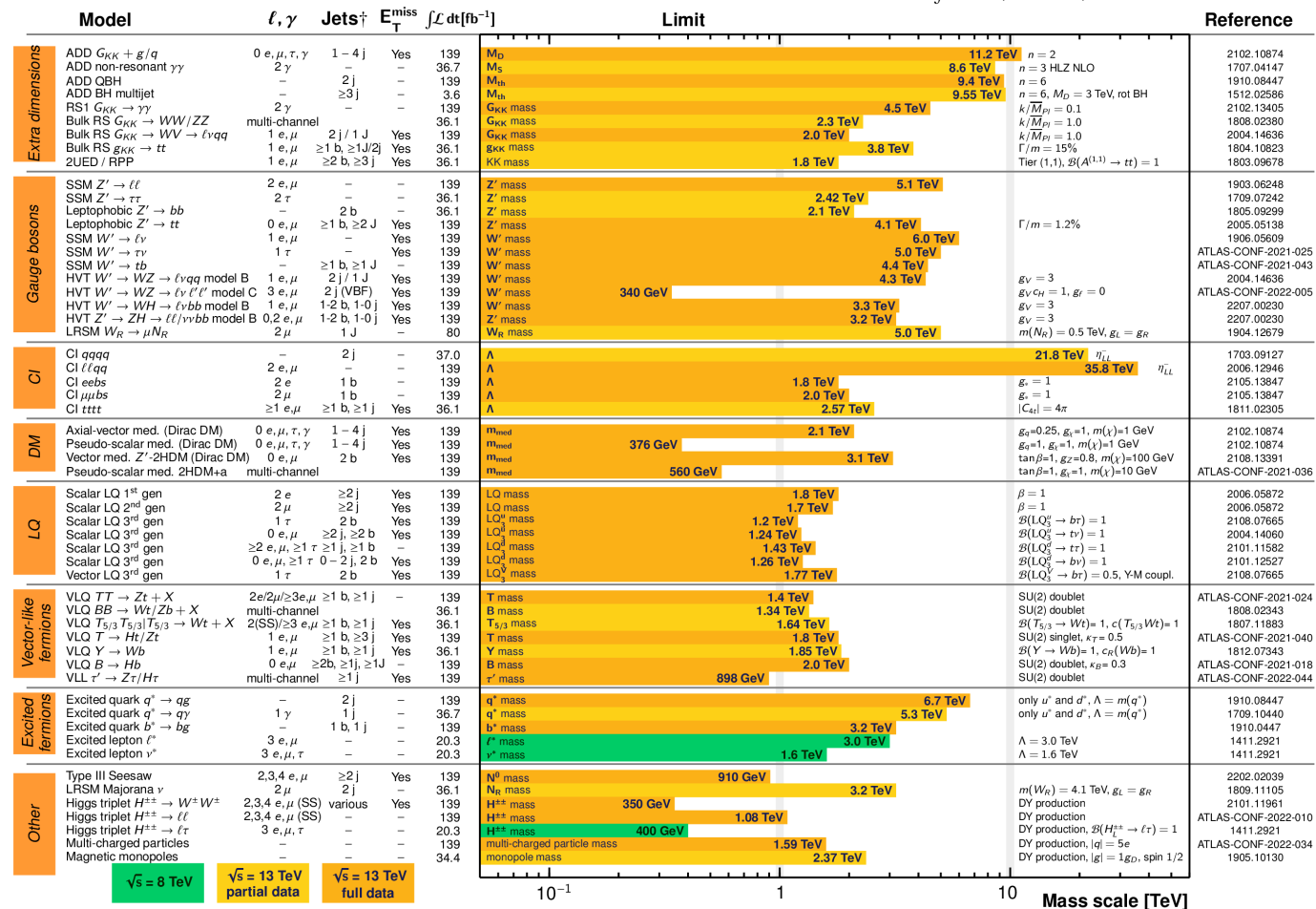
## ATLAS Heavy Particle Searches\* - 95% CL Upper Exclusion Limits

Status: July 2022

ATLAS Preliminary

$$\int \mathcal{L} dt = (3.6 - 139) \text{ fb}^{-1}$$

$$\sqrt{s} = 8, 13 \text{ TeV}$$



\*Only a selection of the available mass limits on new states or phenomena is shown.

† Small-radius (large-radius) jets are denoted by the letter j (J).

How to insure we do not miss potential discoveries?

Model agnostic searches represent an alternative

Multiple strategies exist

Deep Learning provides particularly powerful tools

Suitable for online deployment (trigger)



# Anomaly Detection with VAE

First demonstrations as early as 2018 !

Variational Auto Encoders as **model-independent** (unsupervised) BSM search tools

## Train on known physics

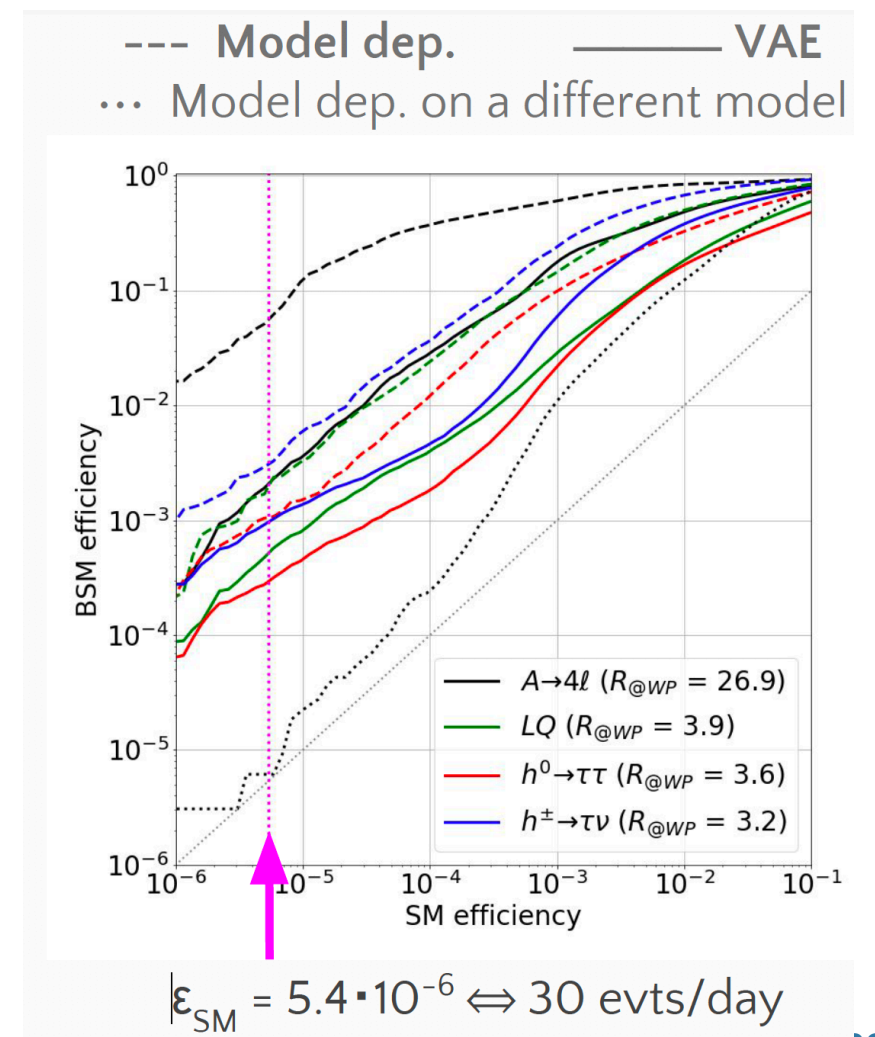
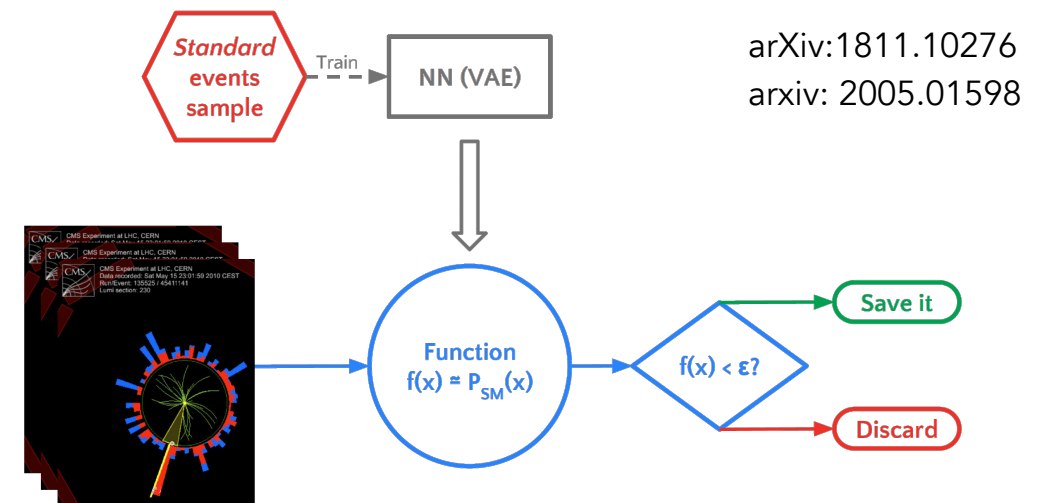
Monte Carlo

Real detector data

Minimise input-output difference  $\mathcal{L} = \|x - x'\|^2$

Anomalies will exhibit large error

Build an anomaly score



# Run3 running examples @ CMS



## How do we train it ?

- Learning *typicality* : By training on Zero Bias dataset



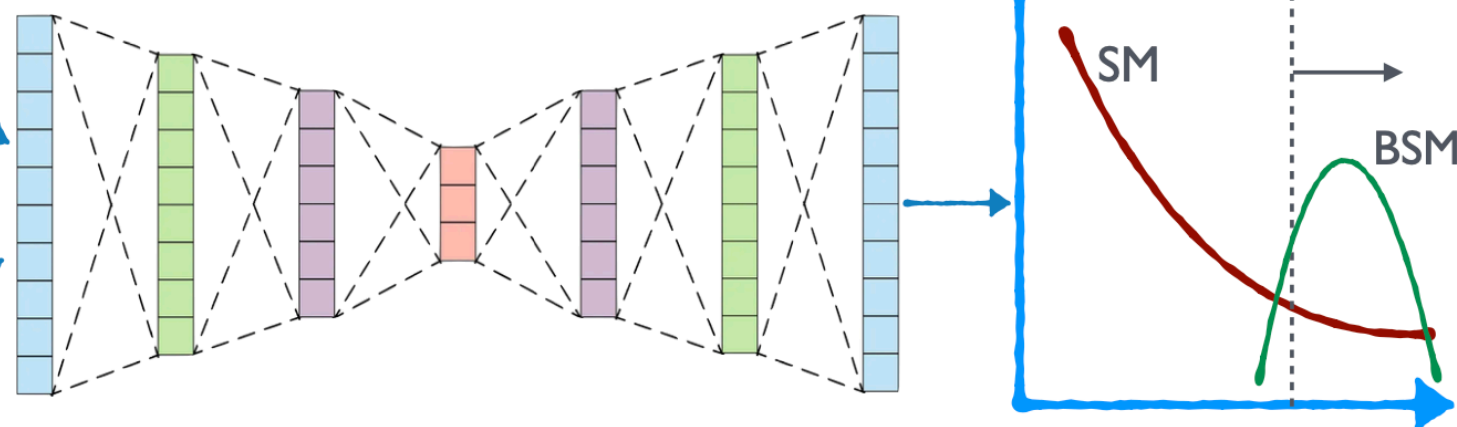
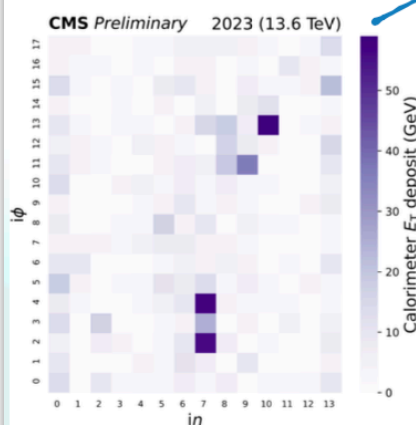
7 million

	$p_T$	$\eta$	$\phi$
MET		N/A	
4 $e/\gamma$			
4 $\mu$			
10 jets			

From calorimeter and muon trigger systems:

Objects : Jets(x10) ,  $e/\gamma$  (x4),  $\mu$  (x4), MET

Attributes:  $P_T, \eta, \phi$  in raw integer value



Low level input :

Calorimeter towers, grouped as calo regions

Basically the energy pixels



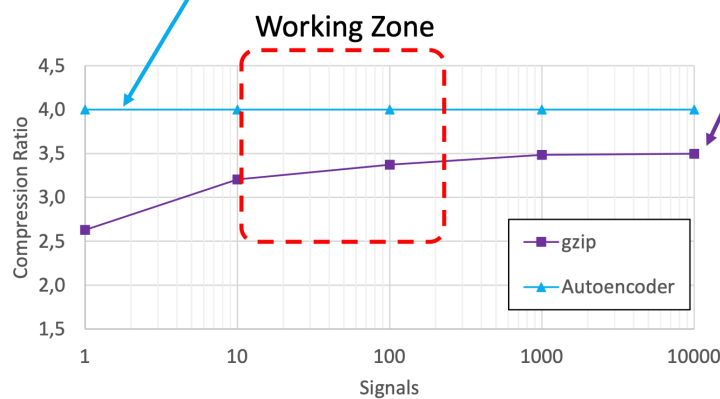
# Other Online Applications (Ex. from CHEP2024)

**Compressed data streaming at BDX:** replace trigger based data acquisition with compressed data stream via AutoEncoder



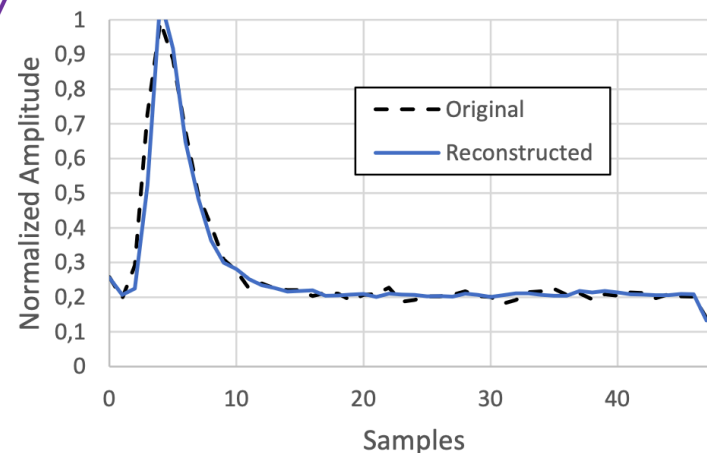
Autoencoder compression ratio is a Parameter

Gzip compression ratio depends on signals number

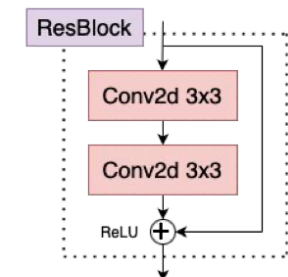
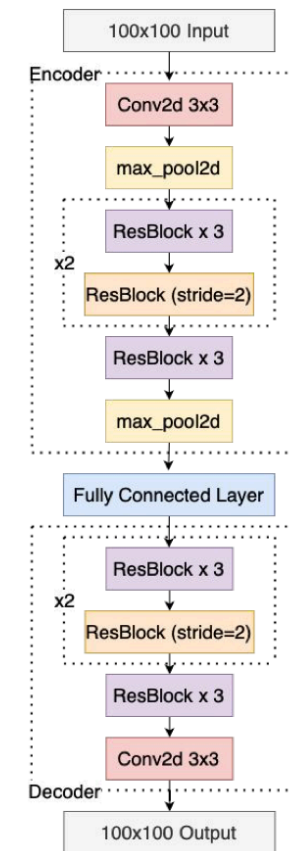


**Better compression ratio**

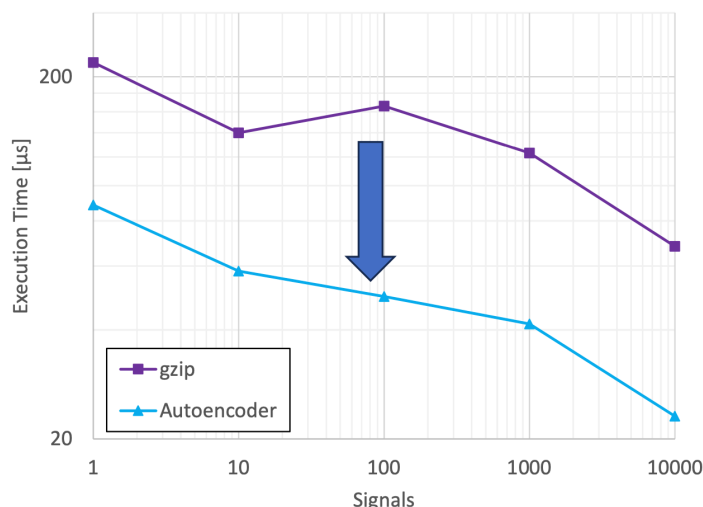
Real-time implementation of Artificial Intelligence compression algorithm for High-Speed Streaming Readout signals, CHEP2024



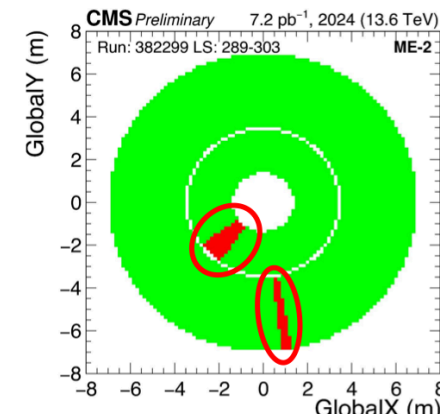
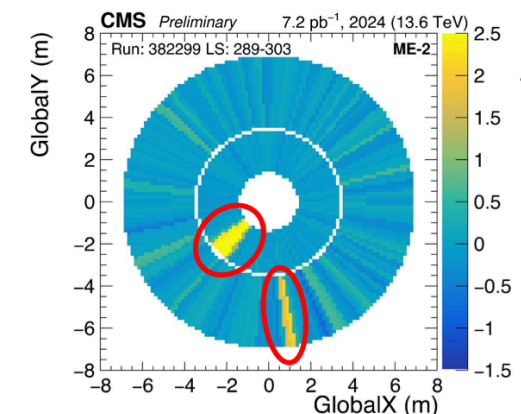
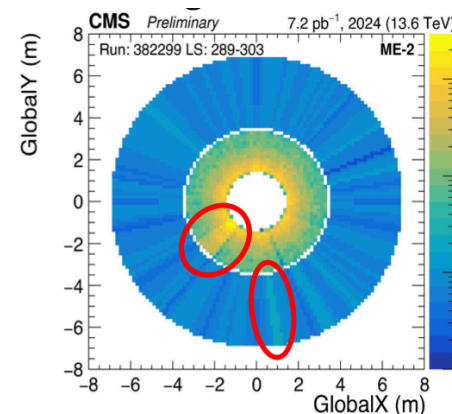
**Data Quality Monitoring in CMS:** ResNet AutoEncoder



**Anomaly detection for data quality monitoring of the Muon system at CMS, CHEP2024**

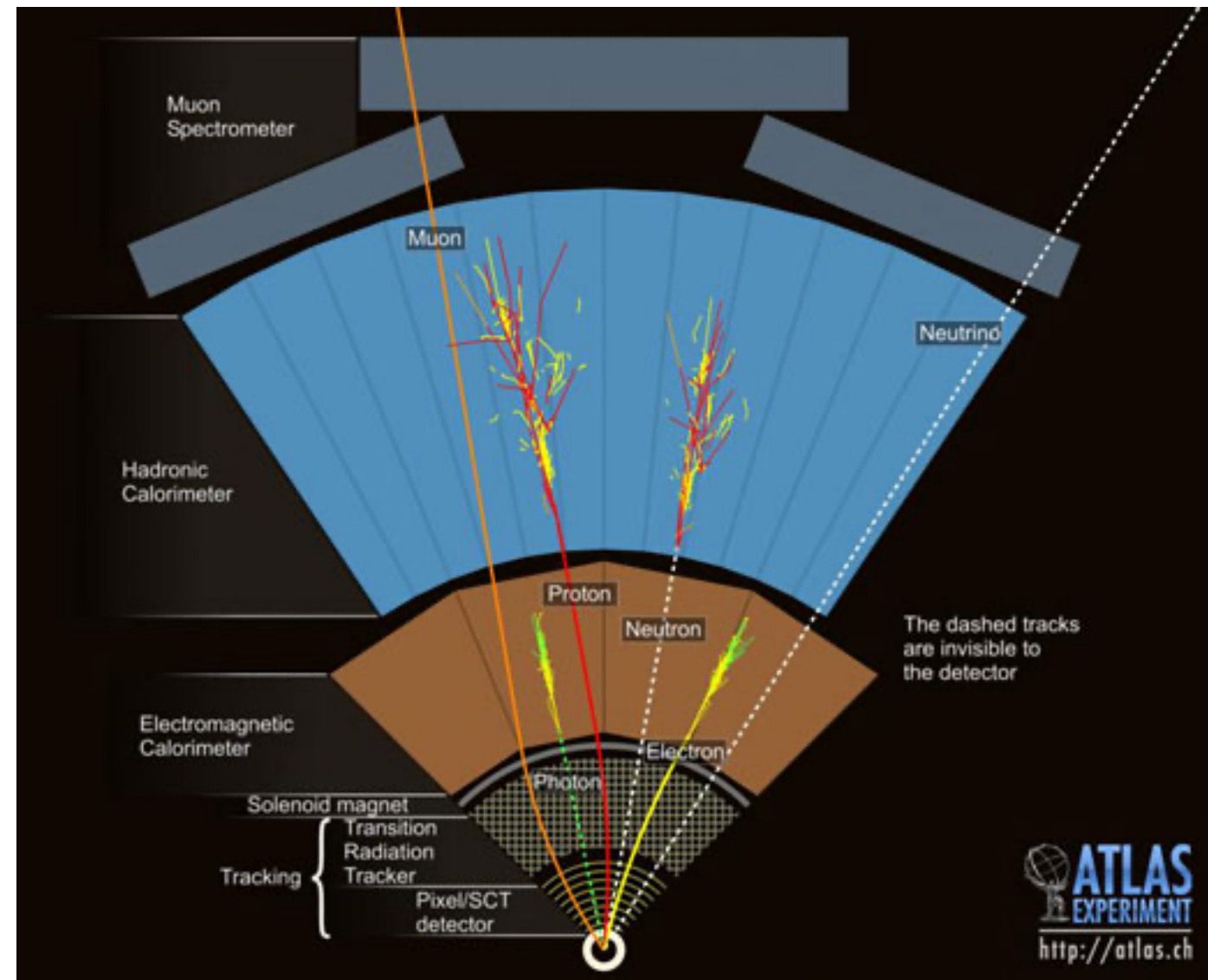
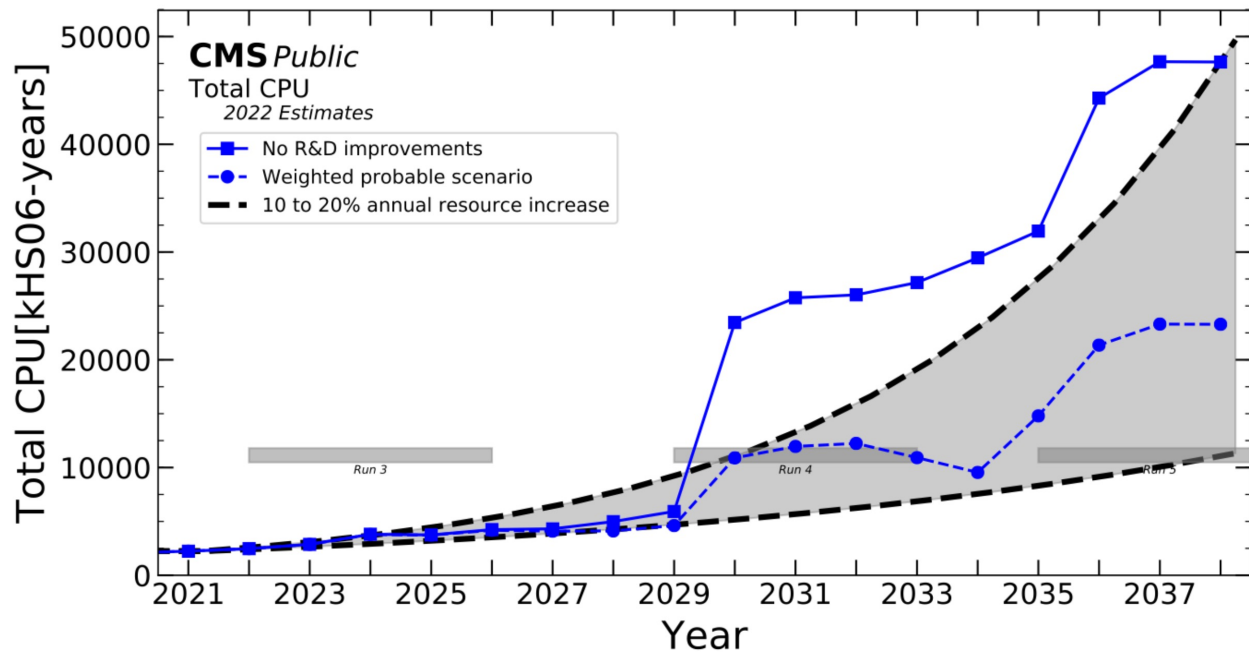
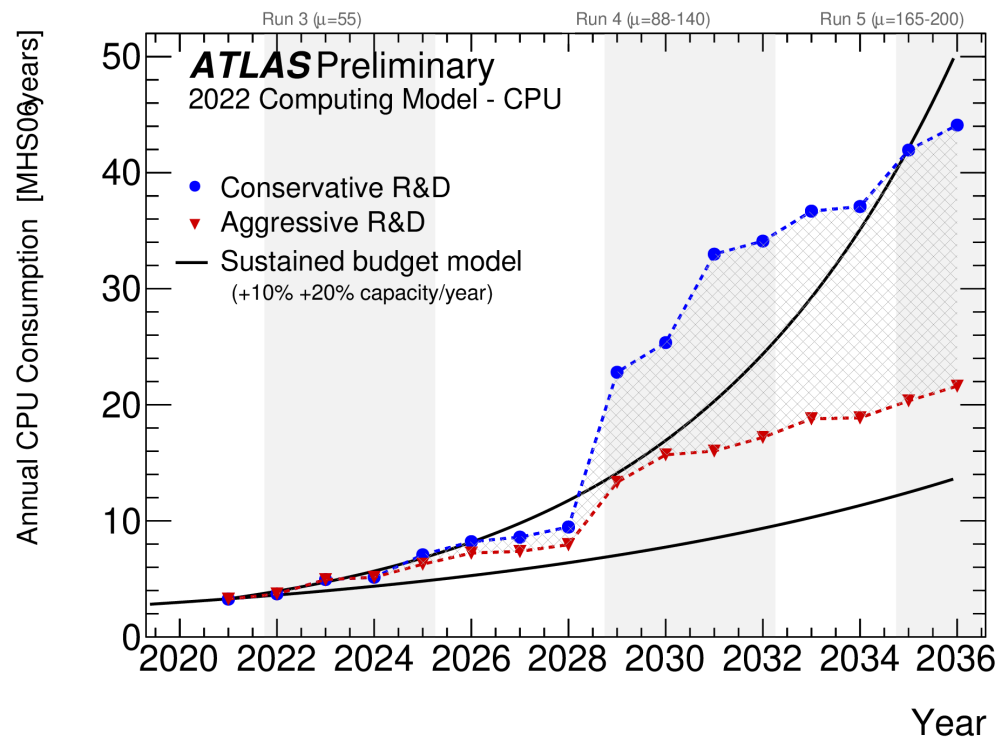


**Better also on execution time**





# Offline processing challenges





# Jets

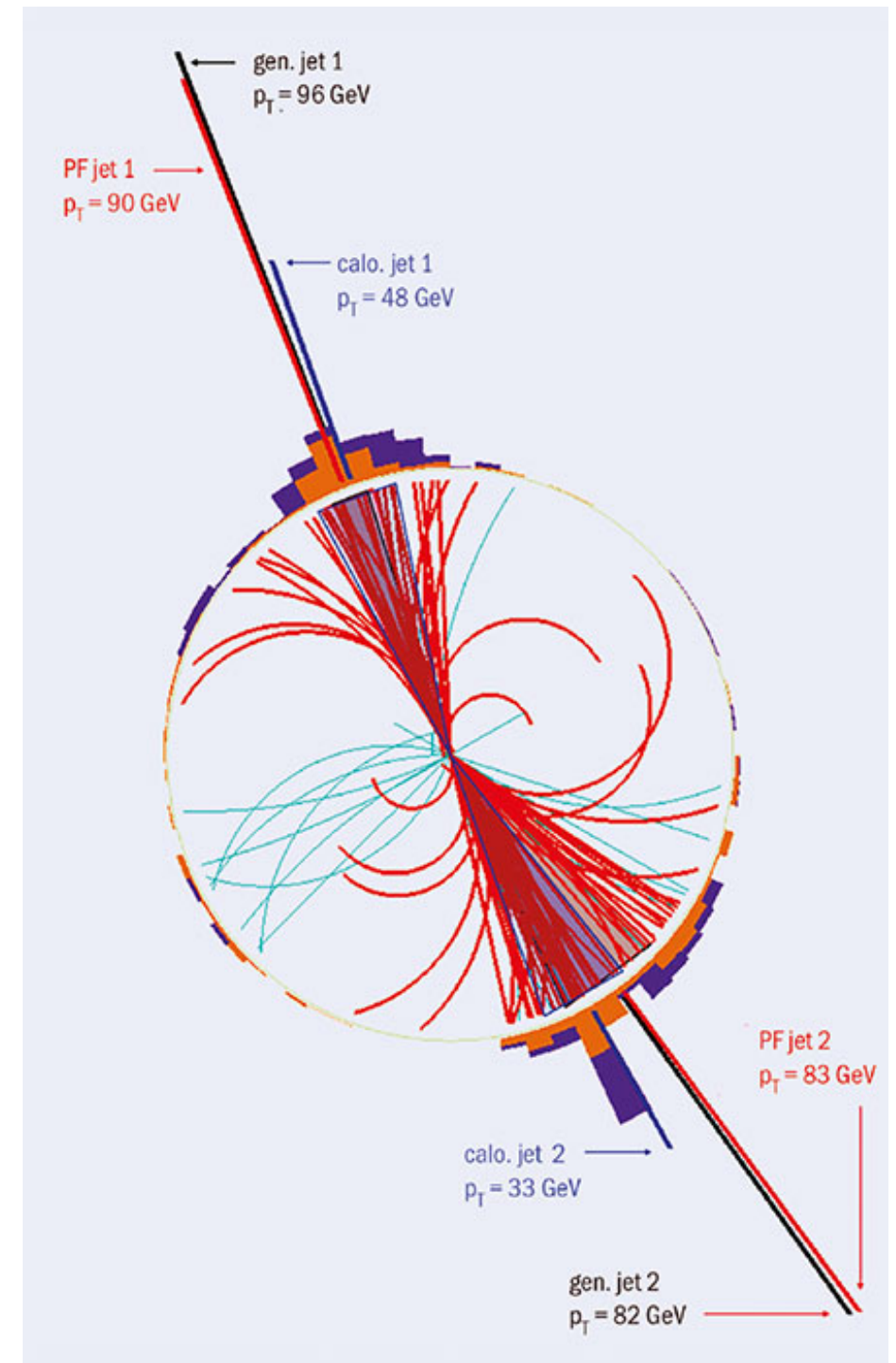
See ML4Jets <https://indico.cern.ch/event/1253794/overview>

## Jets represent a major area of applications for ML

- **Truth Jets:** stable particles defined by MC generators
- **Track Jets:** Use charged-particle tracks. Particularly useful for pile-up mitigation or jet tagging.
- **Topo Jets:** Calorimeter energy deposits. Requires cells clustering and calibration.
- **"Particle Flow" Jets:** Combine tracks and energy deposits.
- **A few notes:**

**Tracks info is limited to charged-particles**, while topo-clusters are built from both charged and neutral particles

**Angular resolution** of the trackers is "still" better than calorimeters. **Calorimeter extend pseudo rapidity coverage.**



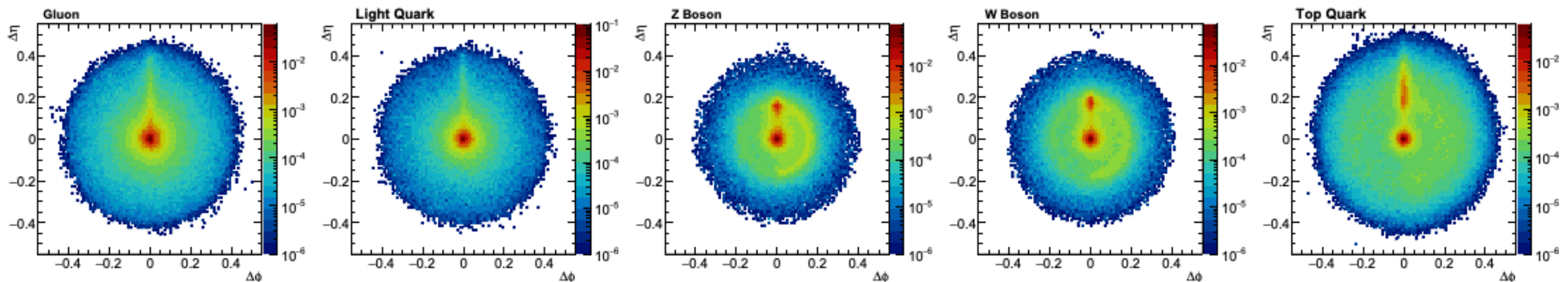
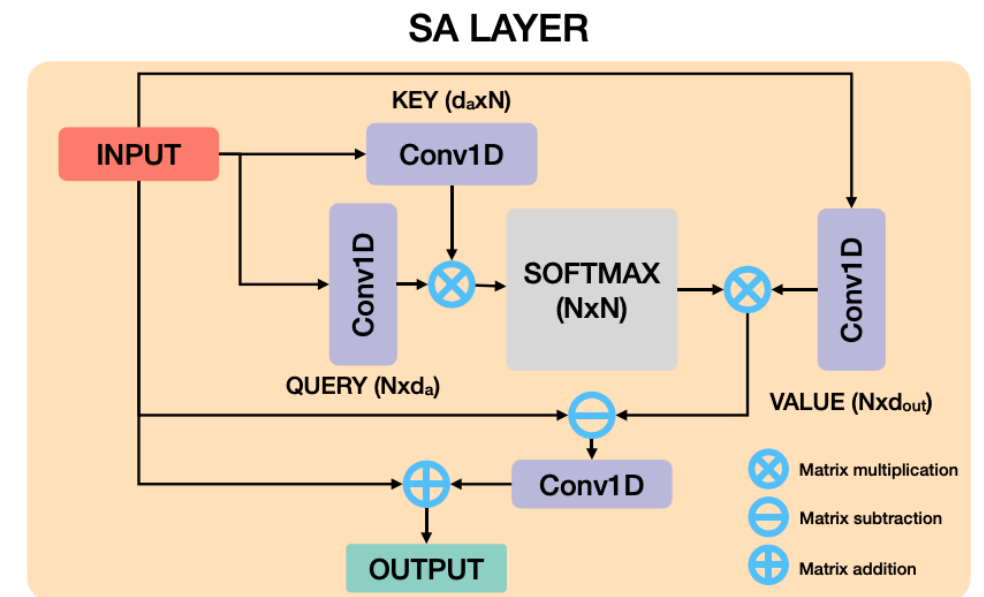
**Both jet reconstruction and Jet tagging (classification) are major applications for ML/DL**

27

# Point Cloud Transformers

Use Self Attention on point-cloud particle data to learn "semantics"

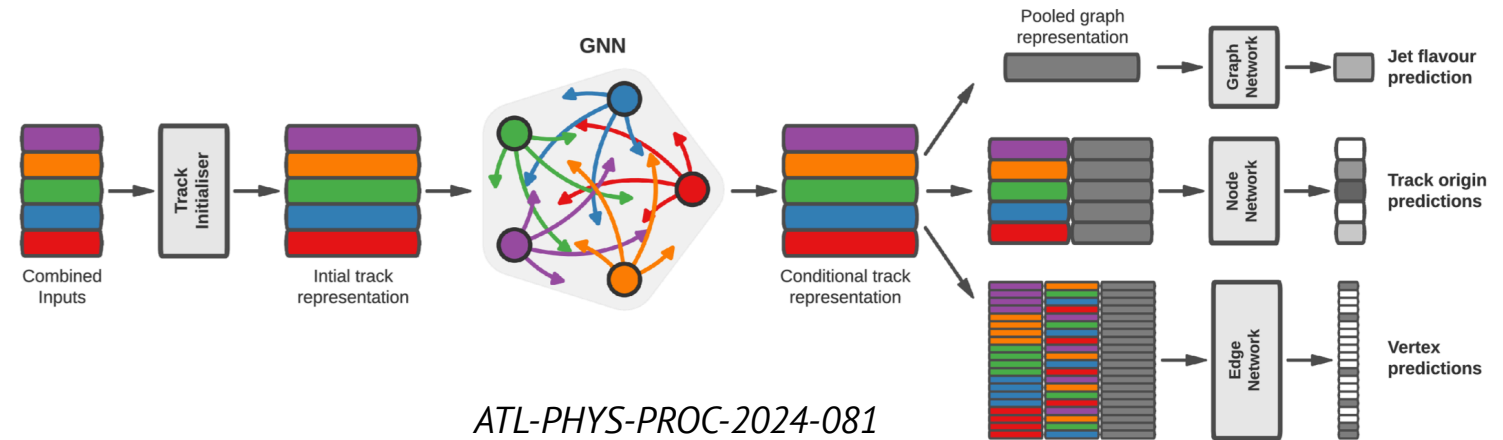
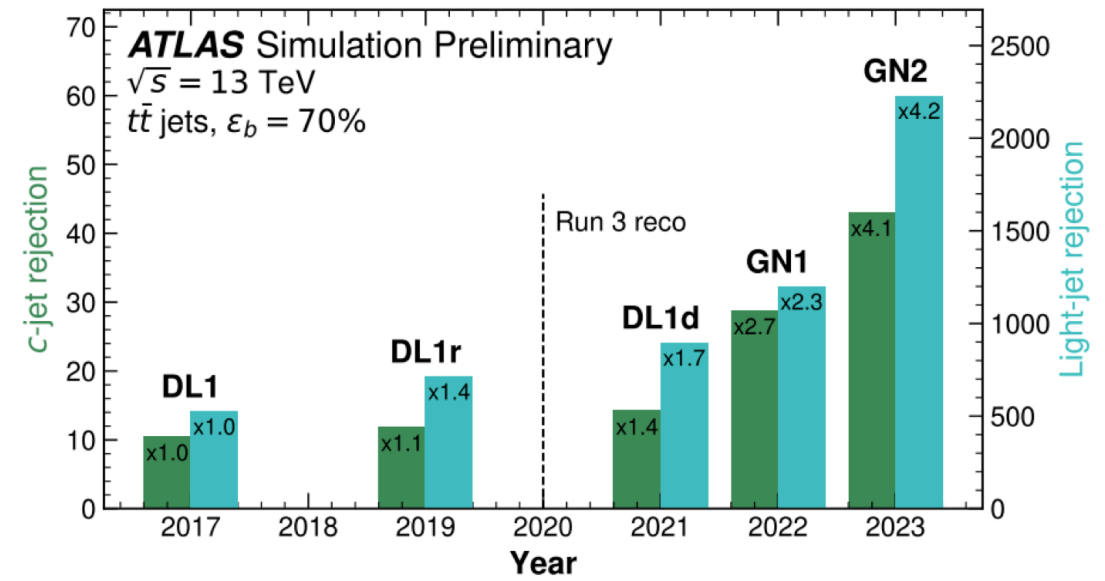
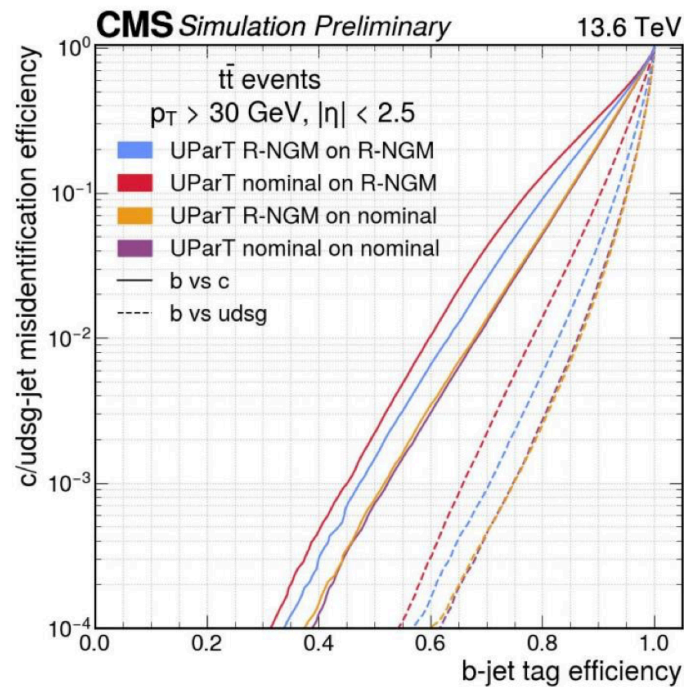
- SA layers extract **different information** for each jet (jet sub-structure)
- **Increased relevance to harder sub-jets** in the case of Z boson, W boson, and top quark initiated jets.
- Light quark and gluon jets have **homogeneous radiation pattern**



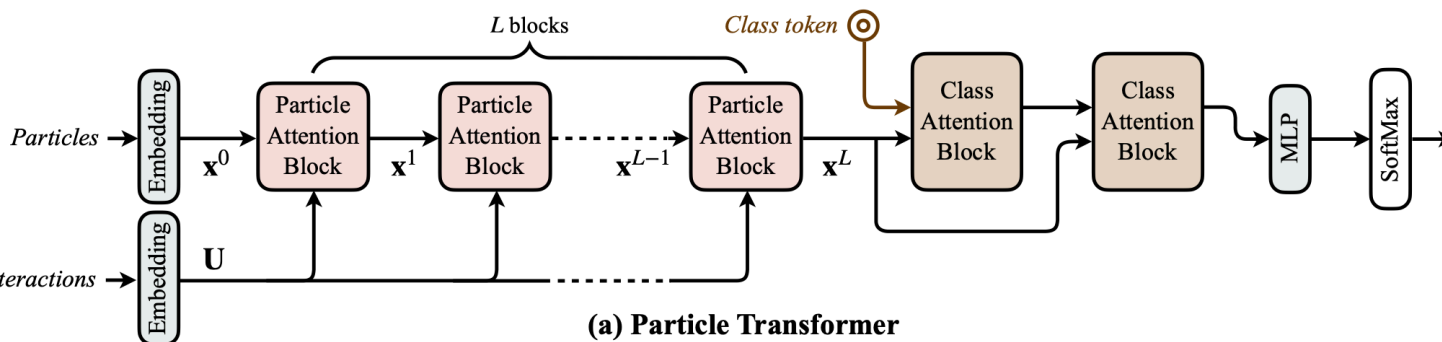
# Jet Tagging: highlights from ML4Jets 2024

<https://indico.cern.ch/event/1386125/overview>

**CMS Jet Tagging: ParticleTransformer** trained to classify b, c, tau, and s and regress on energy and resolution quantiles (no positional encoding since jets are permutation invariant)



**ATLAS Jet Tagging: GNN -based transformer encoder** Also multi-task training (tagging, tracks origin and vertex)



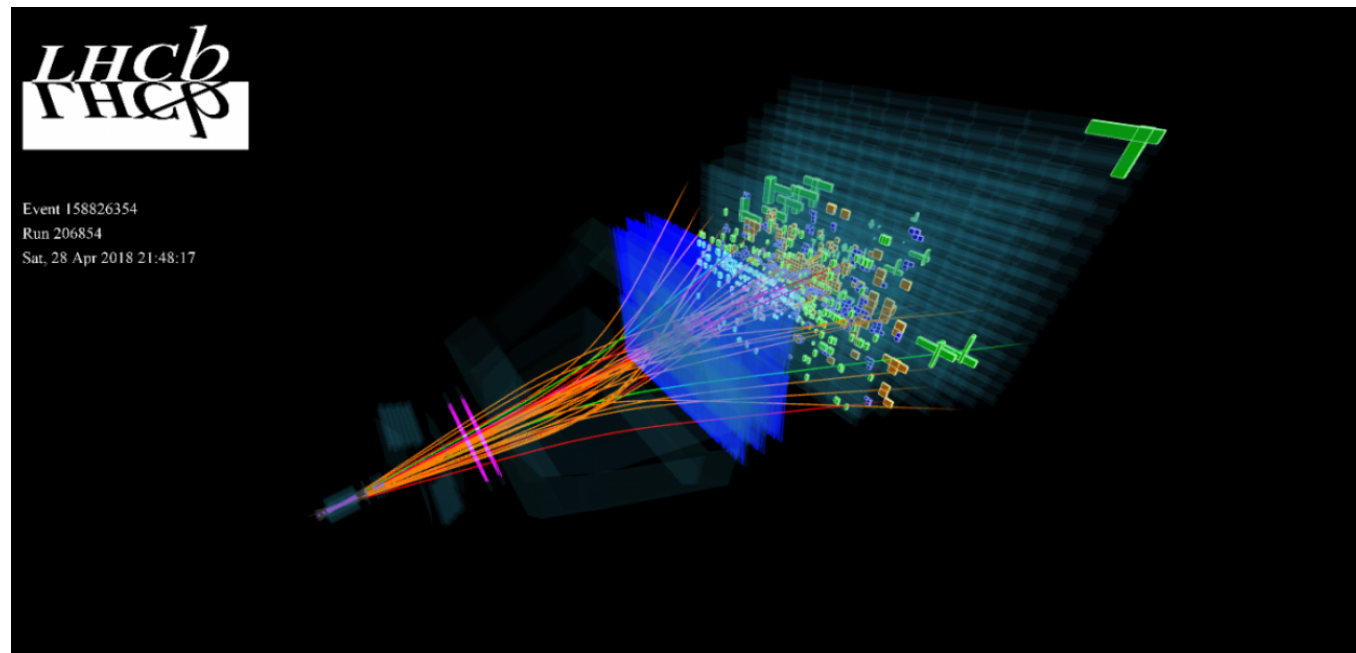
<https://cds.cern.ch/record/2904702>  
[arXiv:2202.03772 \[hep-ph\]](https://arxiv.org/abs/2202.03772)

# Monte Carlo Simulation

**Monte Carlo and simulation** related tasks account for largest computational costs within offline data processing

**Calorimeters** are particularly expensive

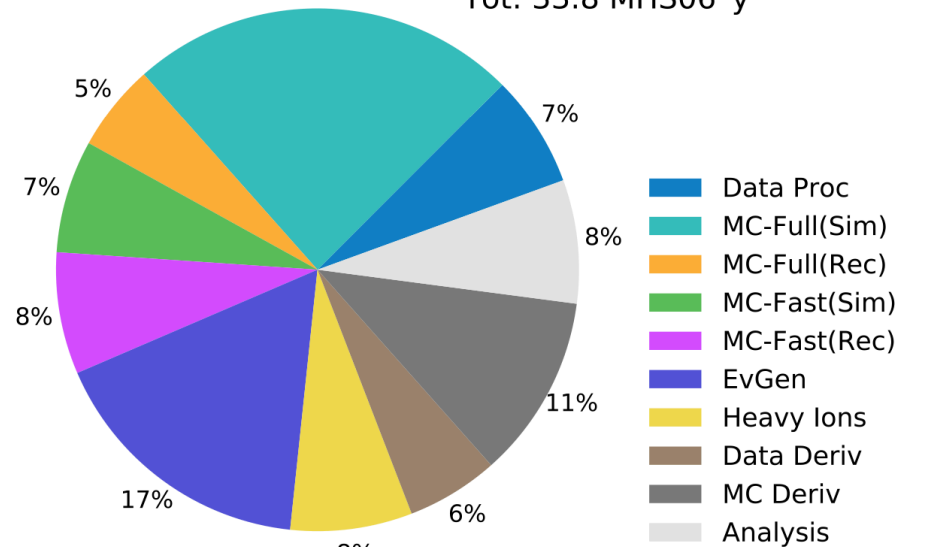
Multiple **fast simulations** techniques exist



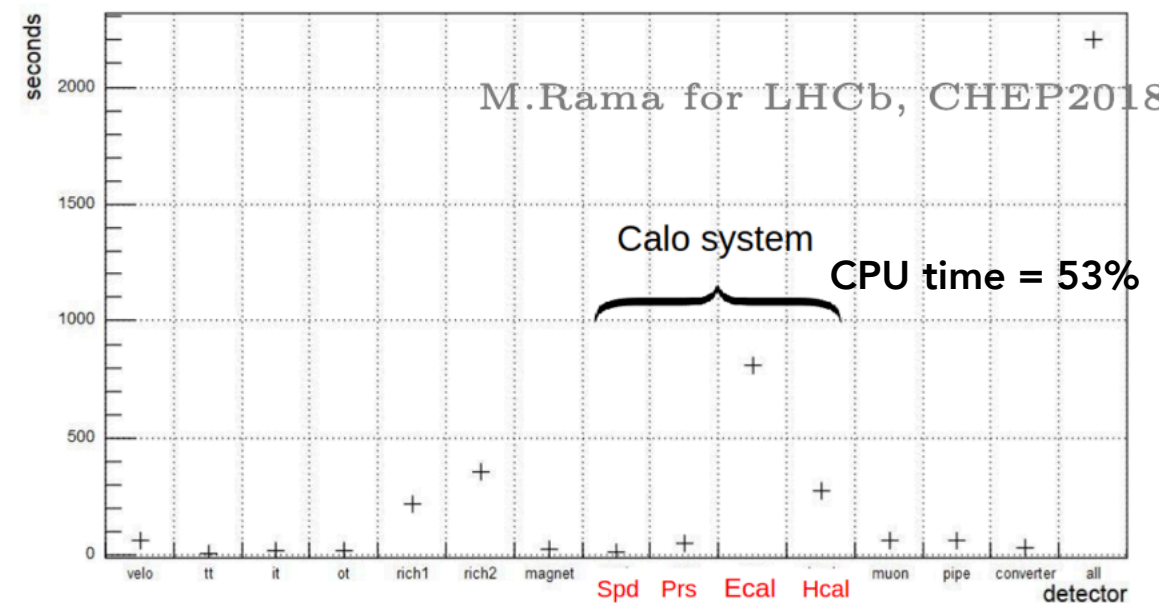
**Ideal task for state-of-the-art generative AI**

Used for fast simulation in HEP as early as 2017

**ATLAS Preliminary** ATLAS CERN-LHCC-2022-005  
2022 Computing Model - CPU: 2031, Conservative R&D  
Tot: 33.8 MHS06\*y



Total time spent in Gauss in different detector volumes





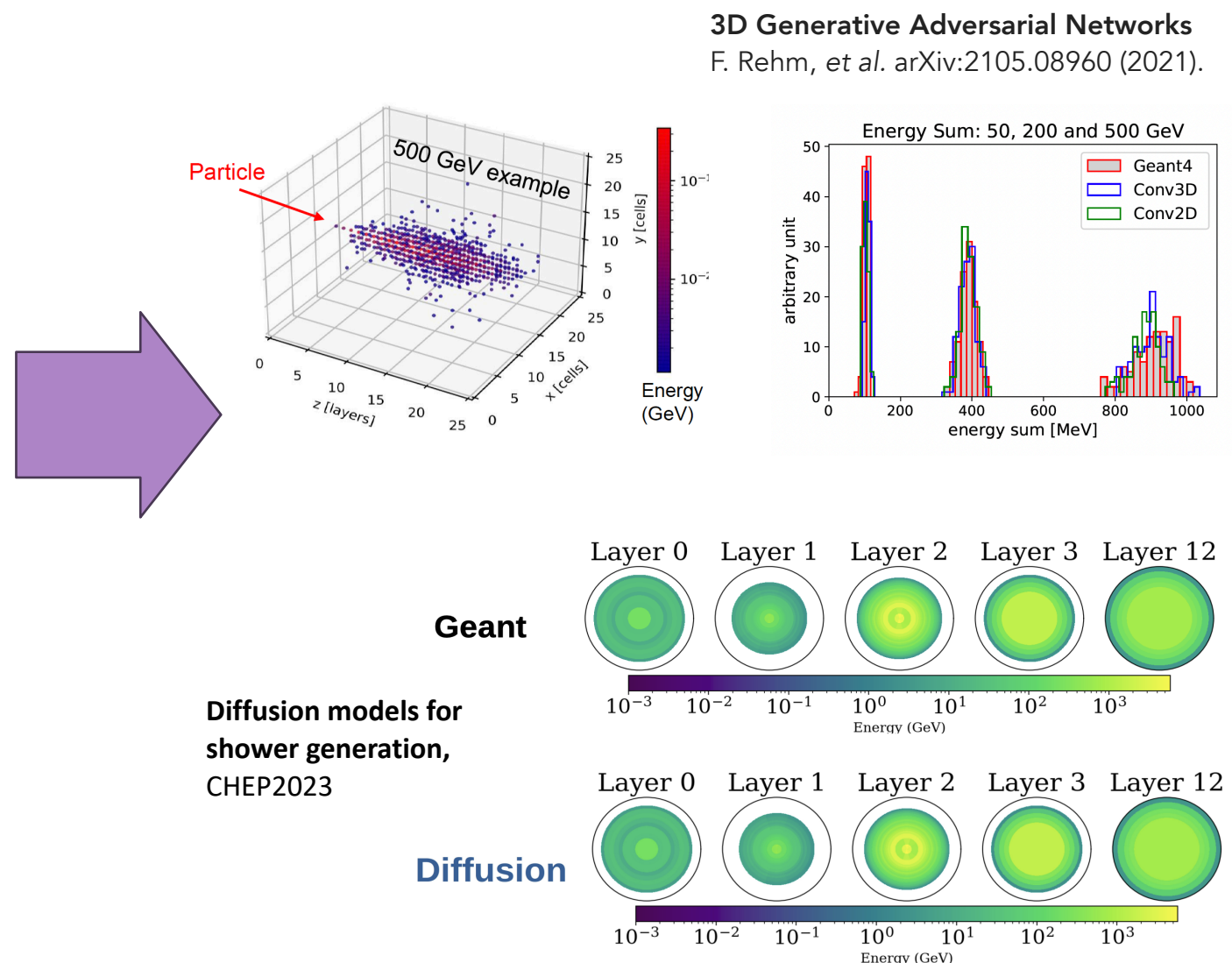
# Synthetic data generation through DL

**Initially use computer vision** approaches and interpret data as 3D grids to simulate energy deposition patterns in calorimeters

**Gradually increase model complexity** and extend fast simulation "concept" (ultra-fast sim)



J. M. Allen, *Space Opera Theatre*, MidJourney (2022)

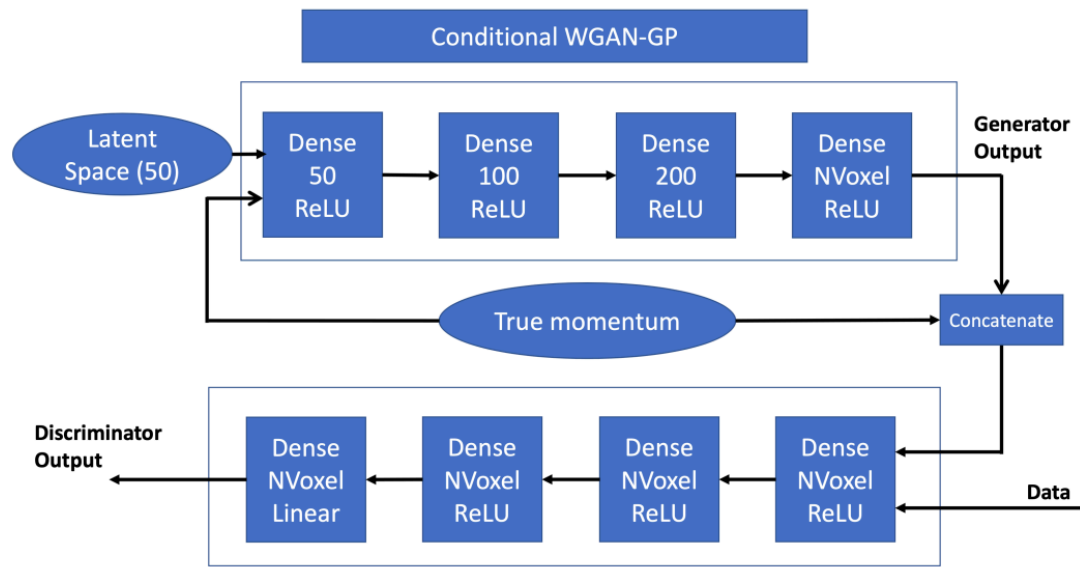




# GAN for calorimeters

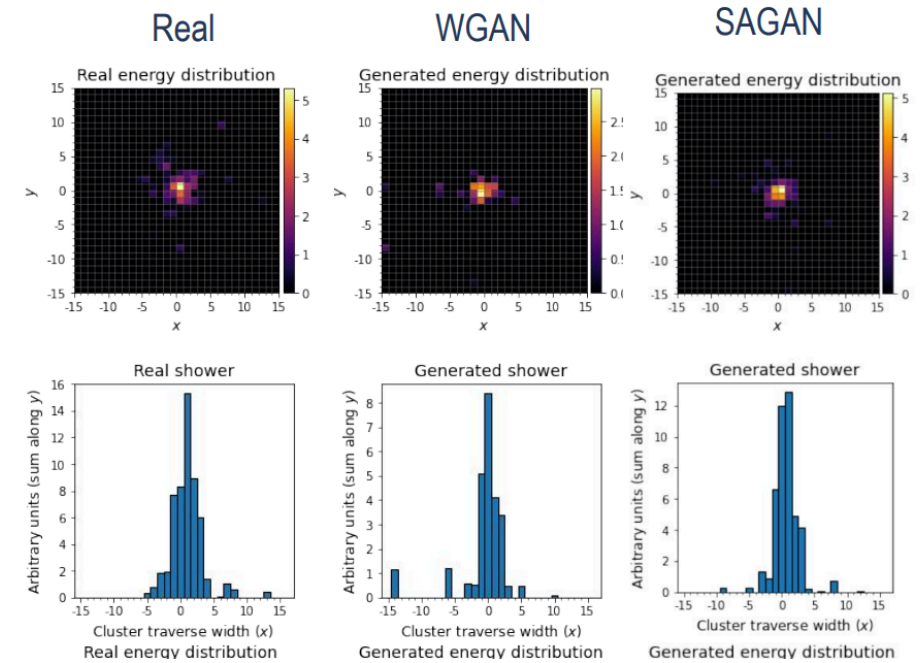
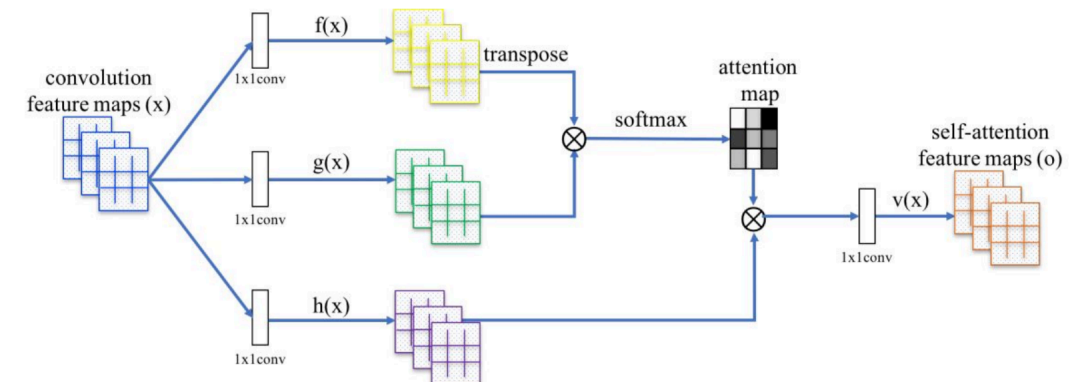
## FastCaloGAN: 300 GANs

M. Faucci CHEP 2023



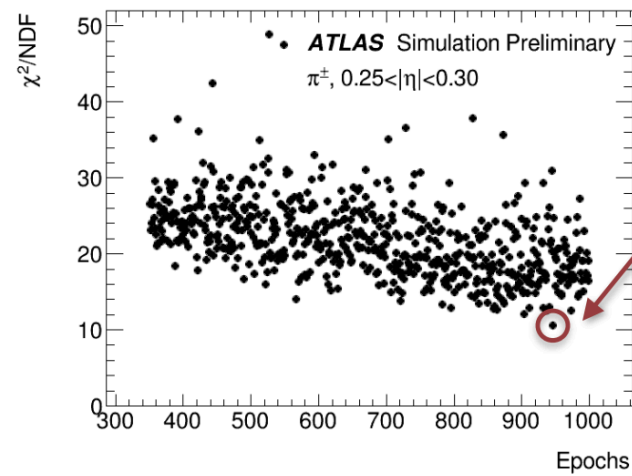
## Self-Attention GANs

F. Ratnikov, A. Rogachev, CHEP2021

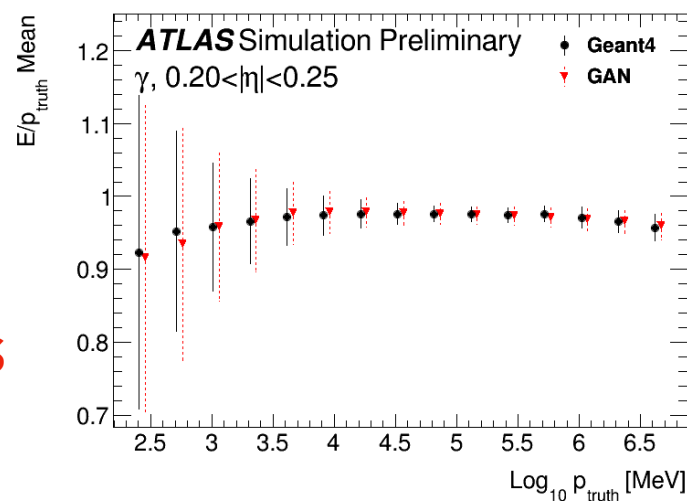


Model	Physics PRD-AUC	Raw Images PRD-AUC
WGAN	0.936	0.971
SAGAN+SN D	0.895	0.901
SAGAN+SN G and D	0.948	0.975

Zhang H. et al. Self-attention generative adversarial networks. – PMLR, 2019 C. 7354-7363.



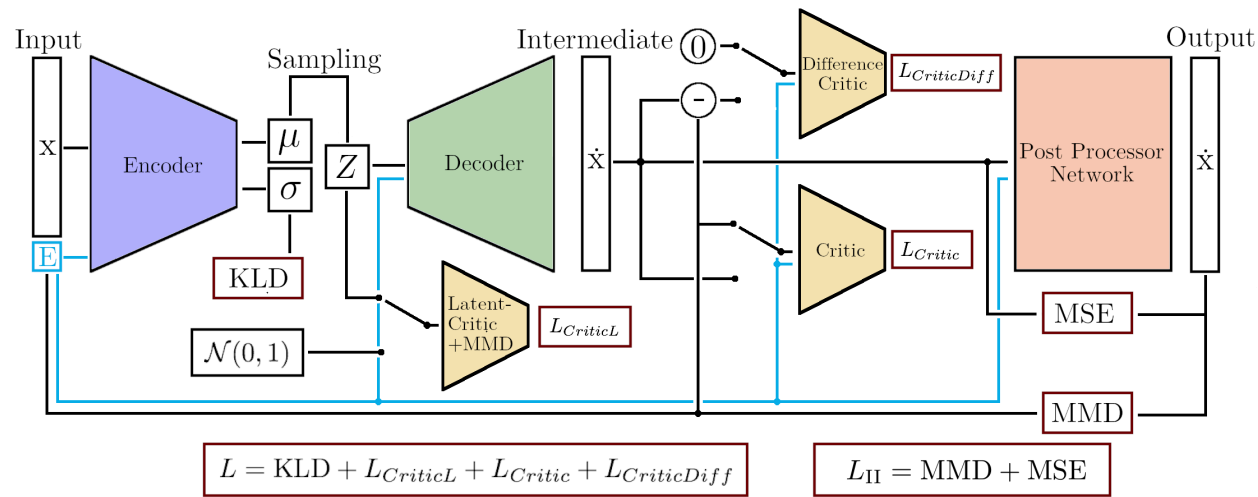
Selected epoch



IN PRODUCTION IN ATLAS

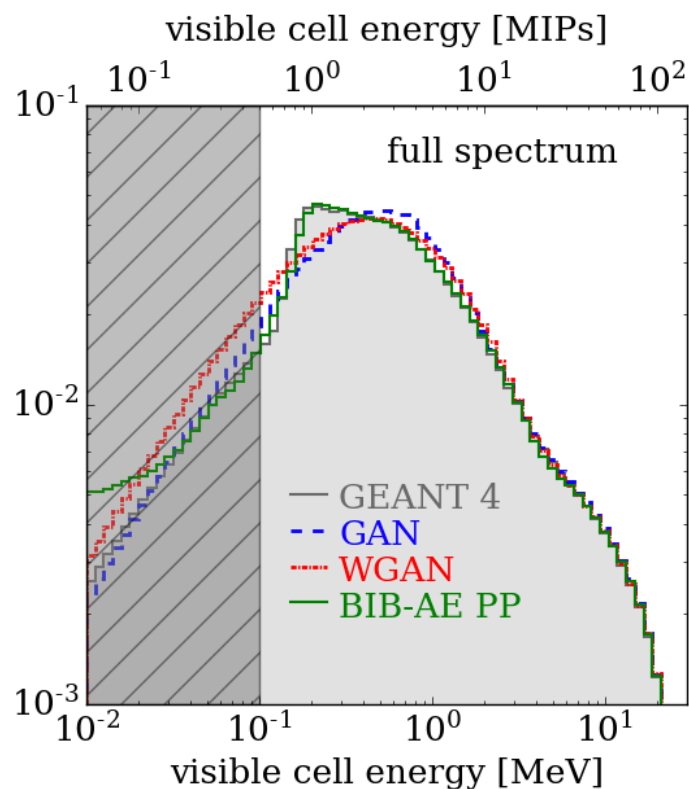
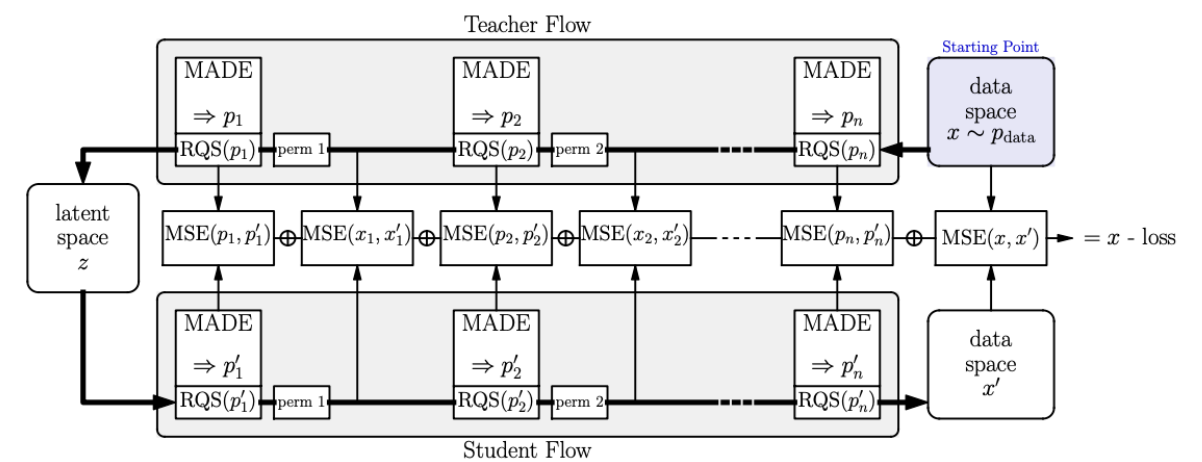
# Increasing complexity

## GAN – AutoEncoder hybrid

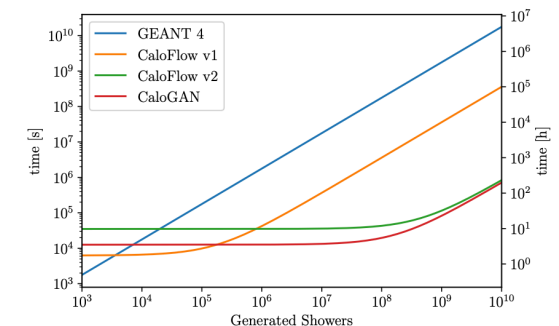
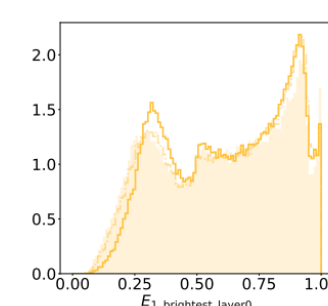
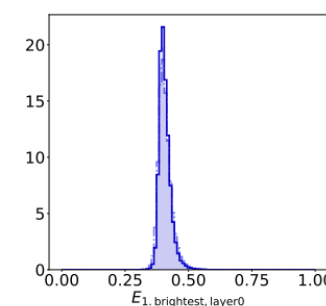
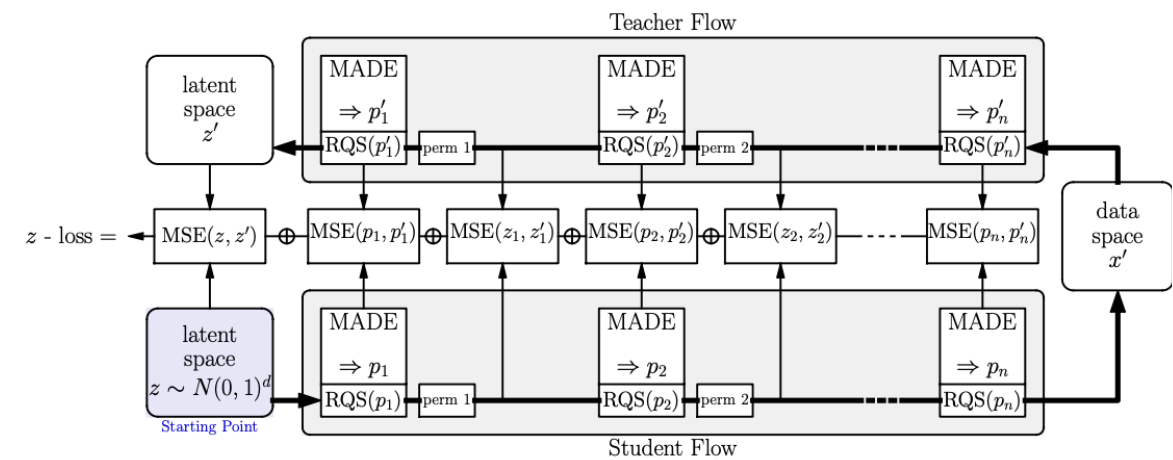


## Normalizing Flows

Krause, Claudius, and David Shih. "CaloFlow II: Even Faster and Still Accurate Generation of Calorimeter Showers with Normalizing Flows." *arXiv:2110.11377*



Buhmann, Erik, et al. "Getting high: high fidelity simulation of high granularity calorimeters with high speed." *Computing and Software for Big Science* 5.1 (2021): 1-17.

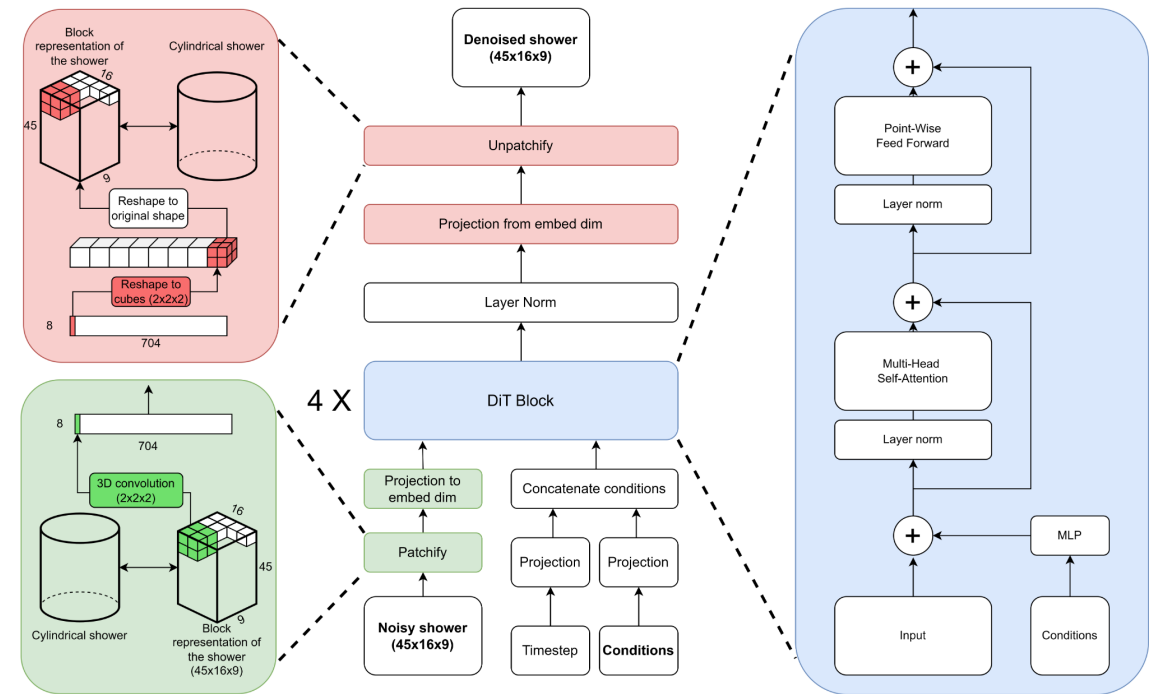


# Conditional Diffusion based Transformer

Architecture based on **visual transformers**

**Input condition** on Energy, Particle Trajectory, Geometry

**Heavy data preprocessing** necessary to map calorimeter geometry to image tiles



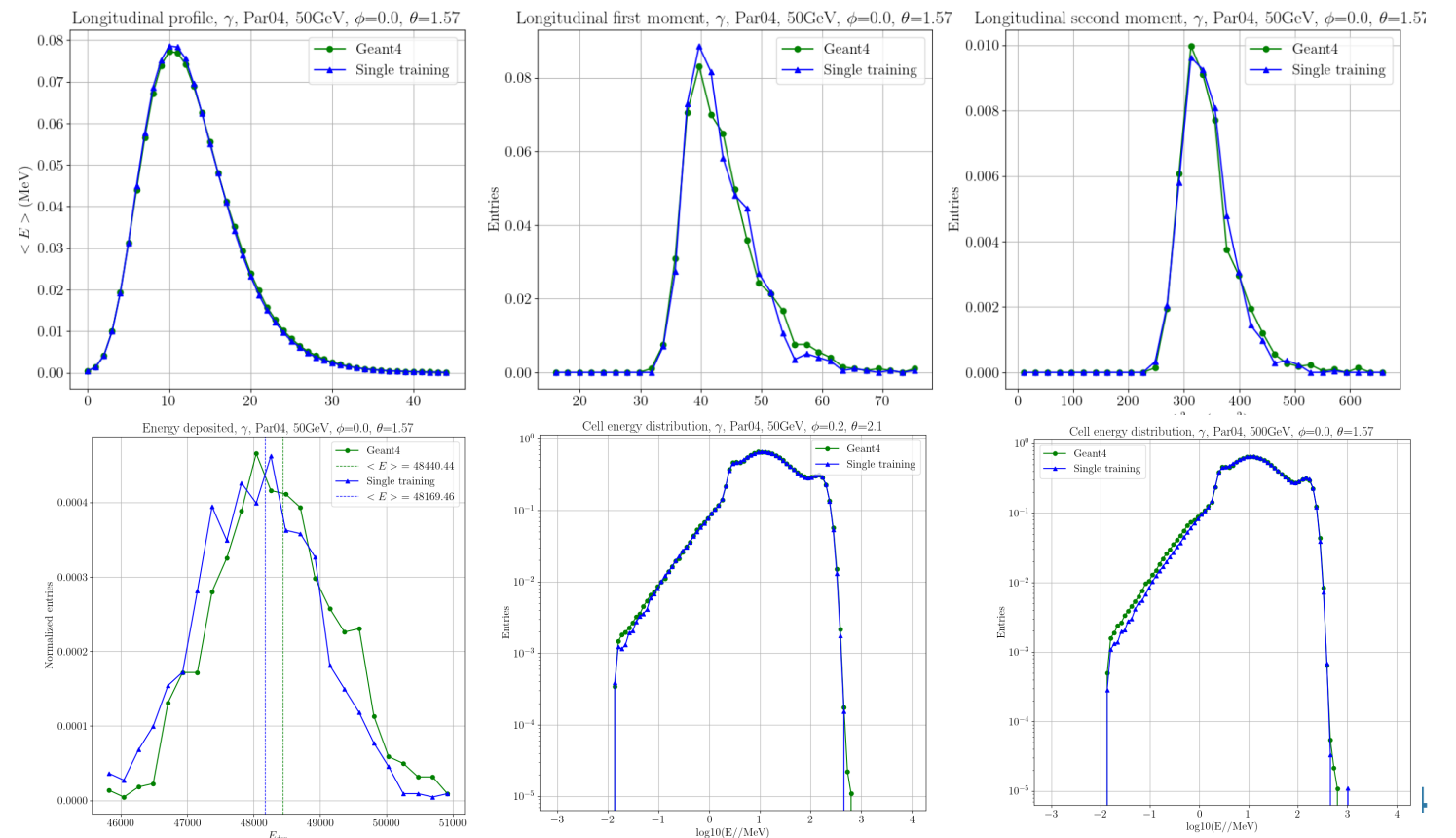
R. Cardoso, CHEP 2023

Maybe different data representation could be more convenient?

**Results:**

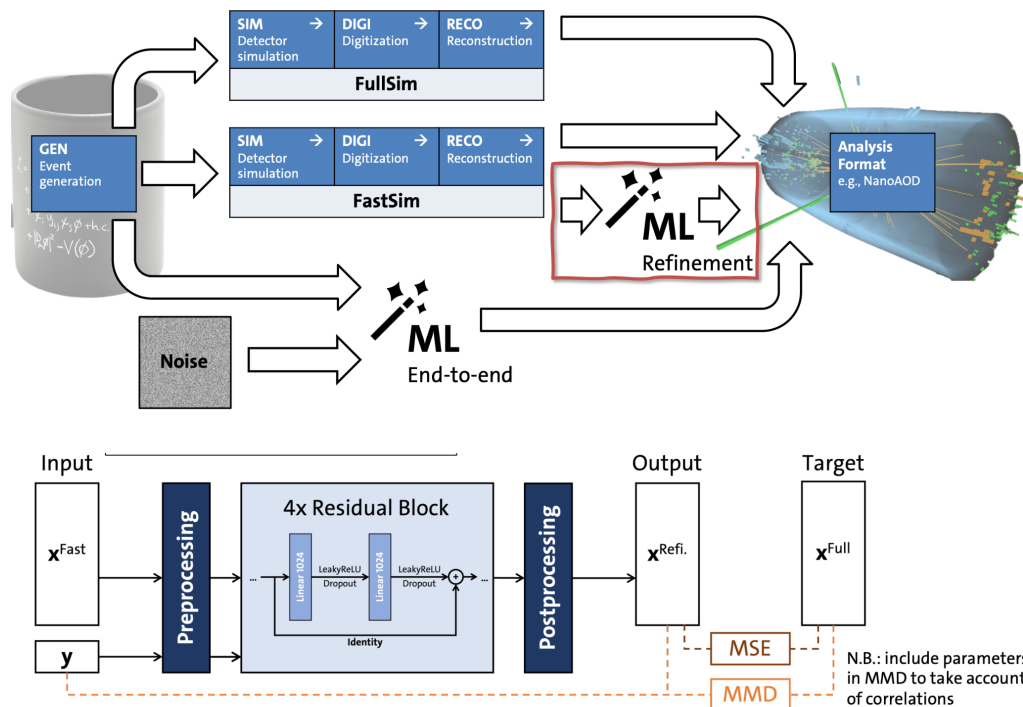
**Good accuracy** throughout all profiles

**Cell energy shows particular good results** compared to other generative models

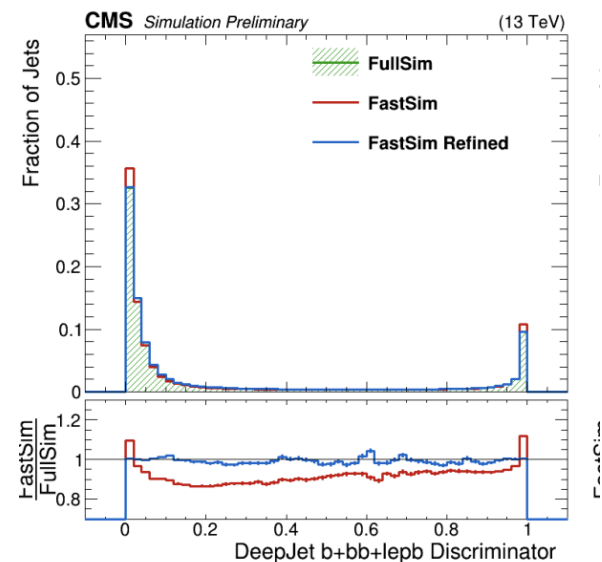


# More Simulation

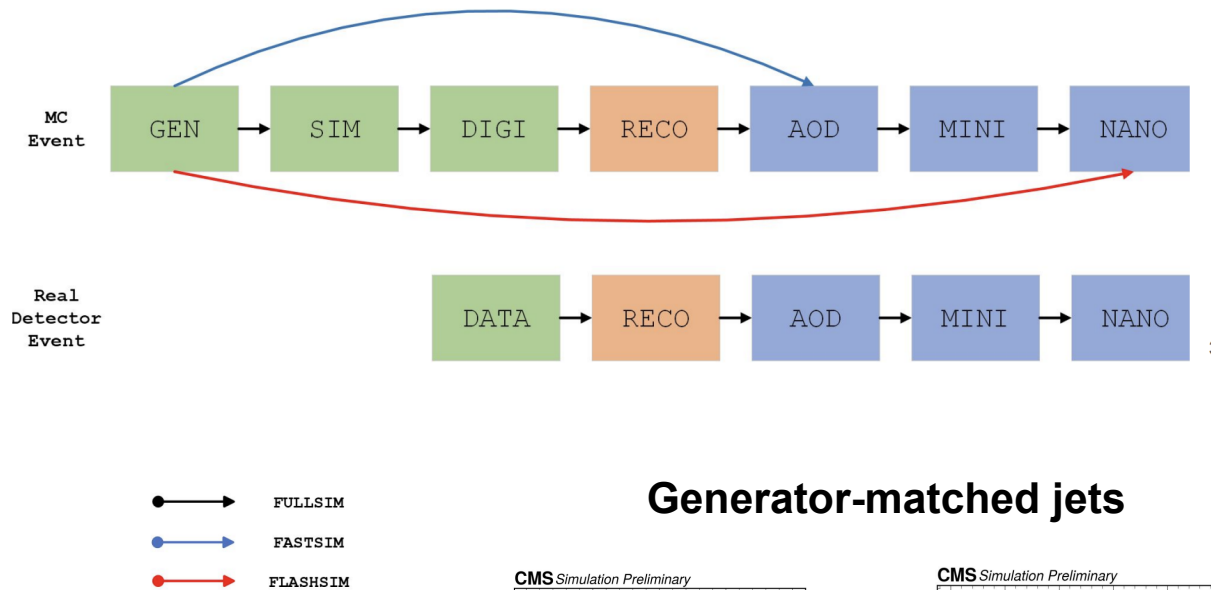
Deep learning to **match fast-sim to fullsim** at analysis level Increases fidelity of fastsim



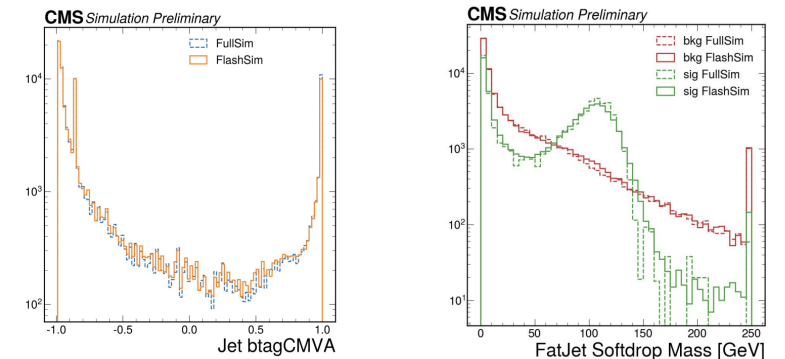
Refining fast simulation using machine learning, CHEP2023



A normalizing flow - based **end-to-end super-fast-sim**, transforming Monte Carlo events directly into high-level analysis objects.



## Generator-matched jets



Flashsim: a ML based simulation for analysis datatiers, CHEP2023

More interesting developments in constructing **ML models for event generation (hadronization)** or to have fundamental **data-driven ML representation** for hadronic physics models in Geant4

MLHad: Simulating Hadronization with Machine Learning, CHEP2023

Simulation of Hadronic Interactions with Deep Generative Models, CHEP2023

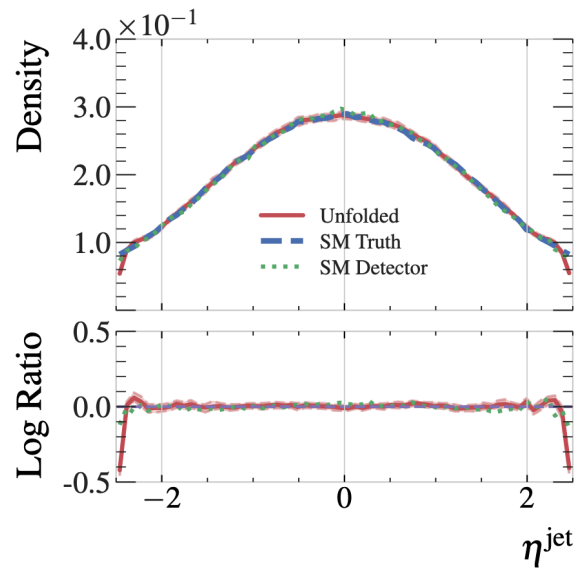
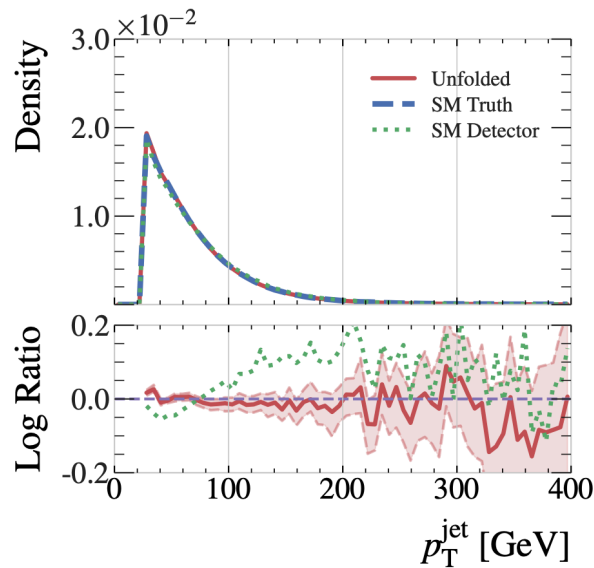
# Comparing experimental data to theory

## Generative Unfolding

### Latent Variational Diffusion:

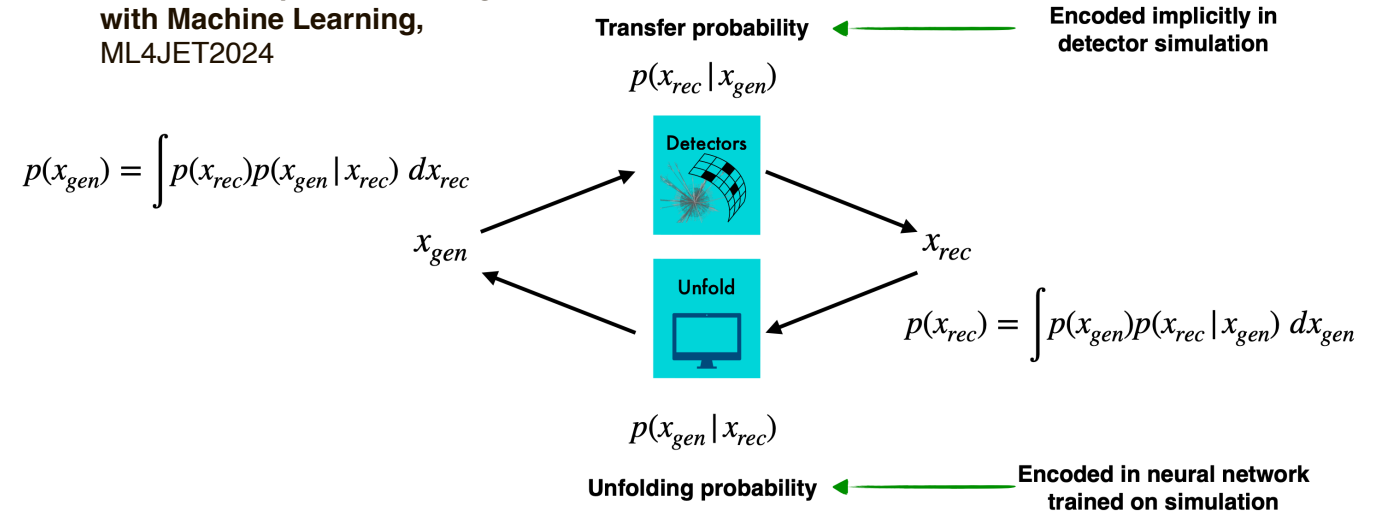
Perform the diffusion process in the latent space of a pre-trained VAE (2112.10752)

Variational diffusion model (2107.00630): interpretation of the diffusion model as an infinitely deep chain of VAE



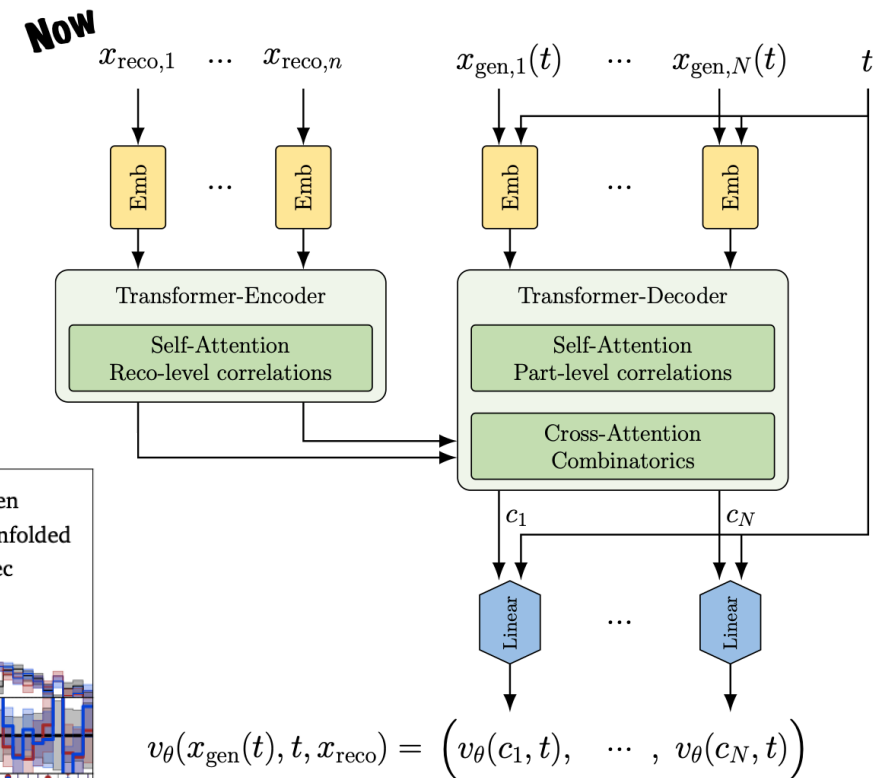
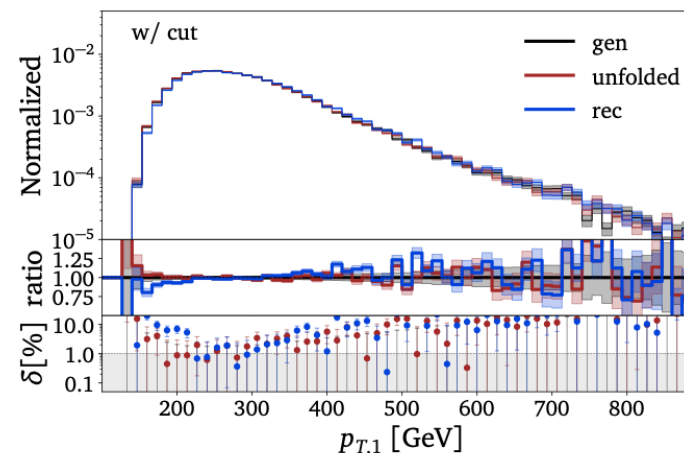
Full Event Particle-Level Unfolding with Variable-Length Latent Variational Diffusion, ML4JET2024

The Landscape of Unfolding with Machine Learning, ML4JET2024



## Transformer based top unfolding

How to unfold Top decays, ML4JET2024



Adapted from [arXiv:2310.07752](https://arxiv.org/abs/2310.07752)



# Event Generators: a Lorentz Equivariant Transformer

Transformer components are modified to learn data in a geometric algebra over space-time, **equivariant under Lorentz transformations**.

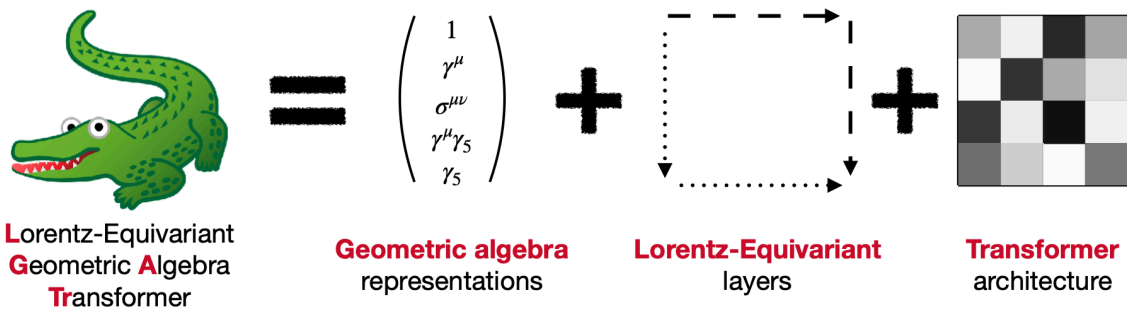
Test on **Amplitude Regression, Jet Tagging and Event Generation**

## Event Generation:

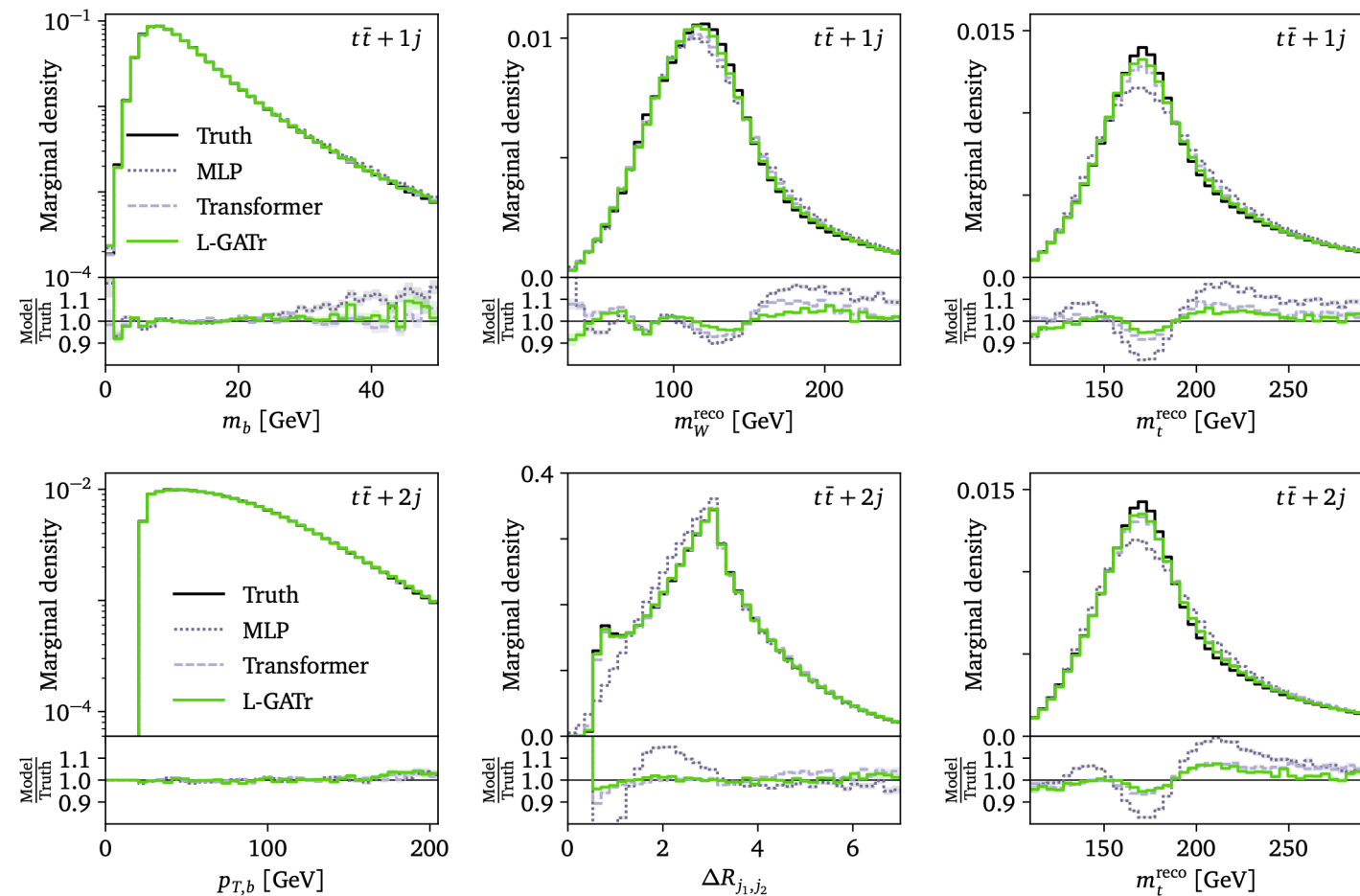
Use L-GATr blocks in a normalising flow  
Focus on hadronic top decay

$$pp \rightarrow t_h \bar{t}_h + n j, \quad n = 0 \dots 4$$

$$n = 0 \dots 4$$



Layer type	Transformer	L-GATr
Linear( $x$ )	$v \cdot x + w$	$\sum_{k=0}^4 v_k \langle x \rangle_k + \sum_{k=0}^4 w_k \gamma^5 \langle x \rangle_k$
Attention( $q, k, v$ ) $_{ic}$	$\sum_{j=1}^{n_t} \text{Softmax}_j \left( \sum_{c'=1}^{n_c} \frac{q_{ic'} k_{jc'}}{\sqrt{n_c}} \right) v_{jc}$	$\sum_{j=1}^{n_t} \text{Softmax}_j \left( \sum_{c'=1}^{n_c} \frac{\langle q_{ic'}, k_{jc'} \rangle}{\sqrt{16n_c}} \right) v_{jc}$
LayerNorm( $x$ )	$x \left[ \frac{1}{n_c} \sum_{c=1}^{n_c} x_c^2 + \epsilon \right]^{-1/2}$	$x \left[ \frac{1}{n_c} \sum_{c=1}^{n_c} \sum_{k=0}^4 \left  \langle x_c \rangle_k, \langle x_c \rangle_k \right  + \epsilon \right]^{-1/2}$
Activation( $x$ )	GELU( $x$ )	GELU( $\langle x \rangle_0$ ) $x$
GP( $x, y$ )	—	$xy$



# Summary and Conclusion

The number of Generative Models applications in experimental HEP continues to increase

In many cases, these tools are already in production for Run 3

Interest also on Large Language Models and AI-based assistants (information retrieval, code assistants, etc..) (I did not talk about this!)

(see for example: <https://indico.desy.de/event/38849/>)

Generative Models based research in the theory domain seems increasing

See HEP ML living review : <https://iml-wg.github.io/HEPML-LivingReview/>

Thanks!  
Question?

<https://indico.cern.ch/event/1386125>

## Contribution List

8 / 105

Event Generator

### 75. Exploring phase space with Flow Matching

Timo Janssen

05/11/2024, 10:50

Event generation

Generative models can speed up parton-level Monte Carlo event generation. Normalizing F due to their exact likelihood evaluation. Compared to discrete, layer-based flows, continuous flows have shown to offer better performance. This presentation focuses on the application of flow matching to the generation of parton-level Monte Carlo events.

### 35. Differentiable MadNIS-Lite

Theo Heimel (Heidelberg University)

05/11/2024, 11:10

Event generation

Differentiable programming opens exciting new avenues in particle physics, also affecting future event generators. These new techniques boost the performance of current and planned MadGraph implementations. Combining phase-space reweighting with a set of new, small, learnable flow generators, MadNIS-Lite, we are generating the same event distributions with the same

### 59. Event Generation with Lorentz-Equivariant Geometric Algebra Transformers

Jonas Simon Spinner

05/11/2024, 11:30

Event generation

Extracting scientific understanding from particle-physics experiments requires solving diverse learning problems with high precision and good data efficiency. We propose the Lorentz Geometric Algebra Transformer (L-GATr), a new multi-branch architecture for high-precision simulation of particle-physics events. It is designed to be a general-purpose framework for

### 9. Classifying importance regions in Monte Carlo simulations with machine learning

Raymundo Ramos (Korea Institute for Advanced Study)

05/11/2024, 11:50

Event generation

We attempt to extend the typical stratification of parameter space used during Monte Carlo simulations by considering regions of arbitrary shape. Such regions are defined by directly using their importance for the simulation, for example, a

### 23. Data-driven hadronization models

Manuel Szewc

05/11/2024, 13:50

Event generation

I'll discuss recent and ongoing developments related to the tuning and construction of machine-learning-based models of hadronization. Specifically, I will discuss efforts related to the extraction of microscopic hadronization dynamics from

### 84. Fast simulation of backgrounds at LHCb - a generalised tool

Alex Marshall (University of Bristol (GB))

05/11/2024, 14:10

Event generation

Background estimation is already a bottleneck in several analyses at LHCb, and with the upcoming larger datasets, the demand for efficient background simulation will continue to grow. While there are existing tools that can provide quick, rough estimates of background distributions, a more detailed, more general tool is needed for the efficient

### 102. Generating particle-clouds with discrete features using Markov jump processes

Dr Darius Faroughy (Rutgers University)

05/11/2024, 14:30

Event generation

In many real-world scenarios, data is hybrid — i.e. described by both continuous and discrete features. At high-energy accelerators like the LHC, jet constituents exhibit discrete properties such as electric charge or particle-id. In this talk, we

### 87. (R) Application of generative models for full-detector, whole-event simulated event generation and jet background subtraction

Yeonju Go (Brookhaven National Laboratory (US))

05/11/2024, 14:50

Event generation

AI generative models, such as generative adversarial networks (GANs), have been widely used and studied as efficient alternatives to traditional scientific simulations like Geant4. Diffusion models, which have demonstrated great capability

