AI for LHC experiments, MEG, and ultra-fast inference

A personal perspective with some biases…

Francesco Armando Di Bello

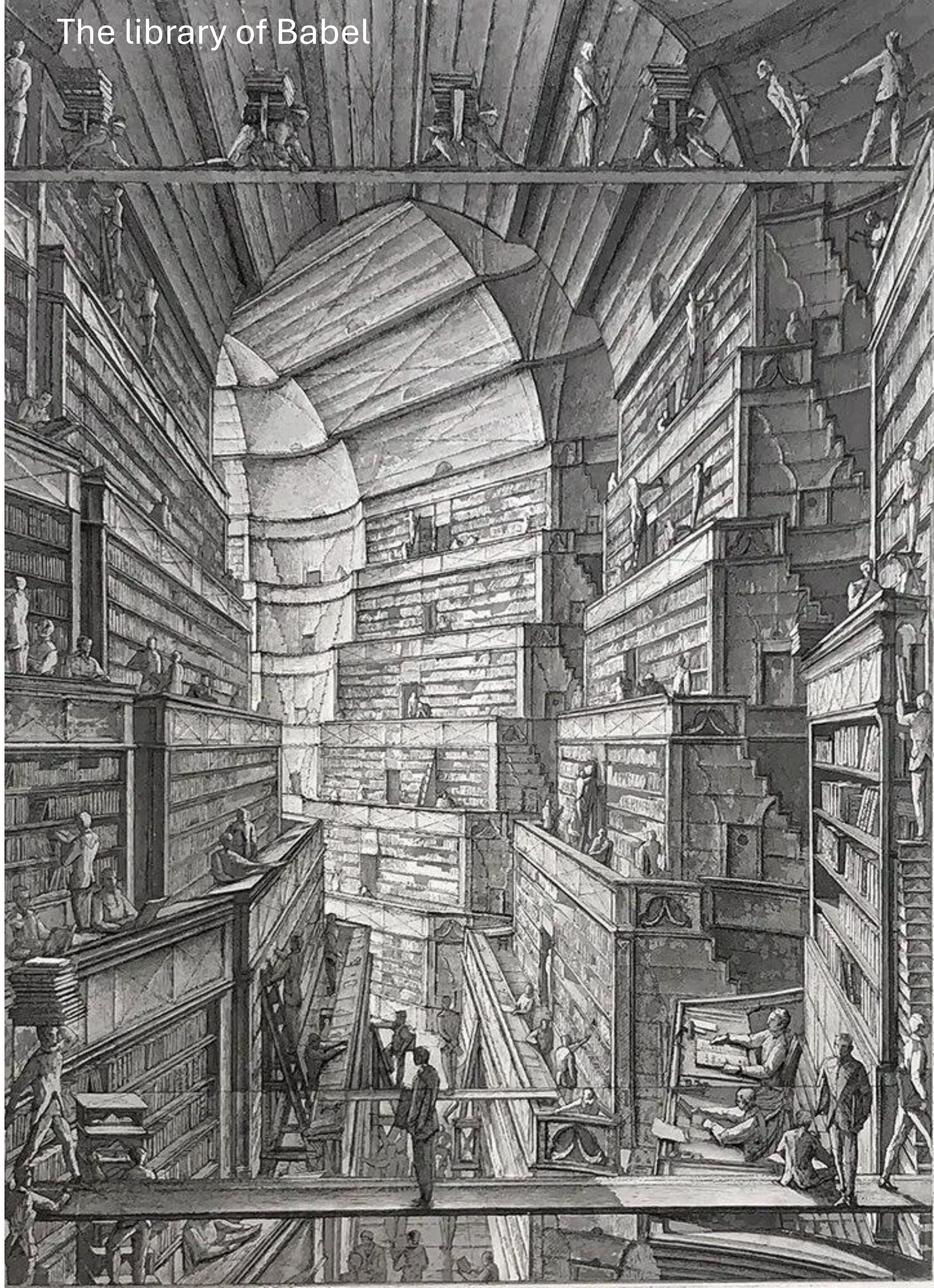
Università di Pisa. 27/05/26

With many thanks to:

A. Venturini, F. Cei, L. Dispoto, E. Graverini, F. Lazzari, G. Punzi, A. Rizzi, F. Cattafesta, F. Vaselli, A. Annovi, F. Castiglioni

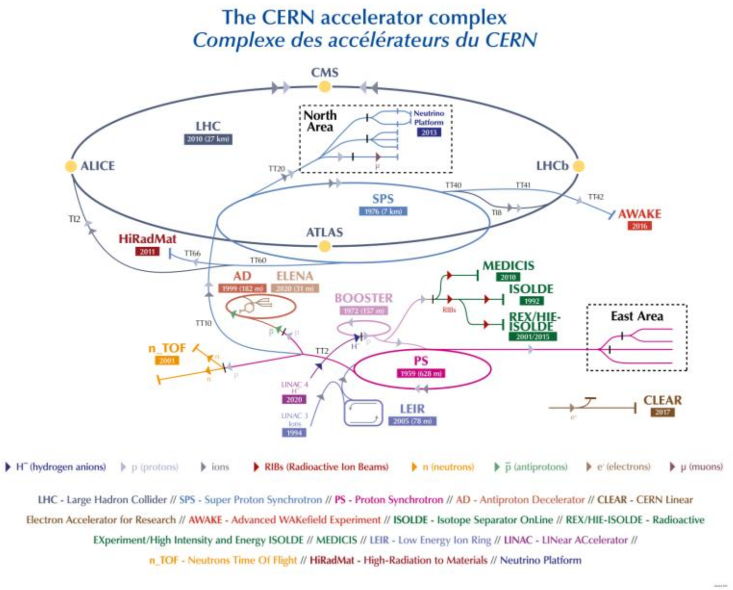
Dipartimento di Fisica "E. Fermi"
Università di Pisa

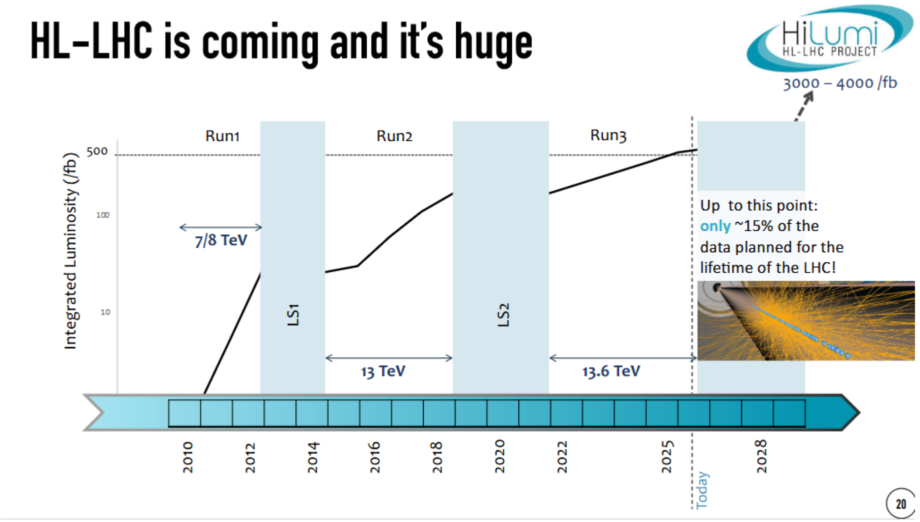
Università di Pisa



Outline of the talk

- I will try to merge contributions received by various groups (ATLAS, CMS, LHCb, MEG)
- Ultra-fast inference and machine learning techniques are extremely active research field within hep-ex area in the dep, not limited to what I will talk about today
- I will try to provide general context, then go deeper into specific examples and state-of-the art techniques, I will then give some *personal* outlook with some personal biases



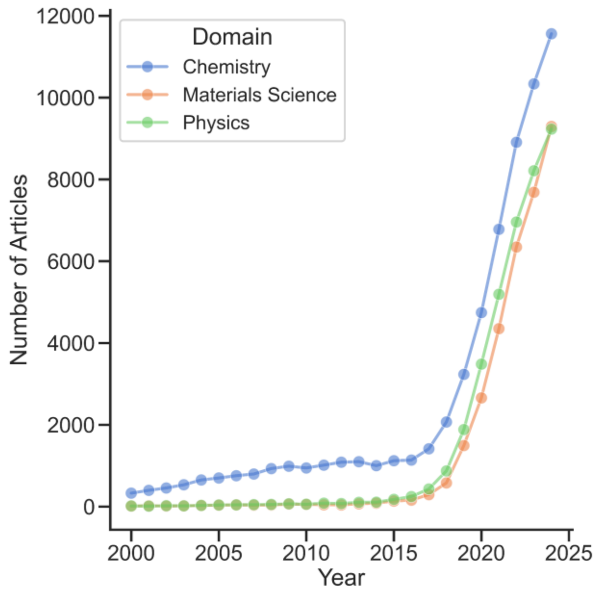


[end of RUN3](#)

[A. Sfyrla, LHCP 2026](#)

Part one:

Machine learning for *hep-ex* status and prospects



Number of papers are exploding with great majority originates from proof-of-principles

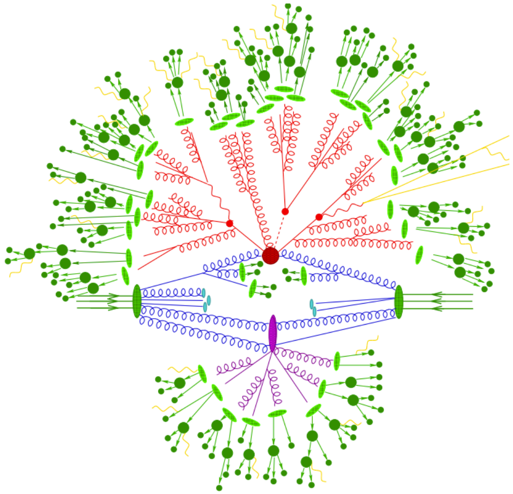
ml_publication_charts

Introduction

- Particle physics experiments are based on hierarchical, factorized pipelines whose goal is to allow unbiased parameters inference.

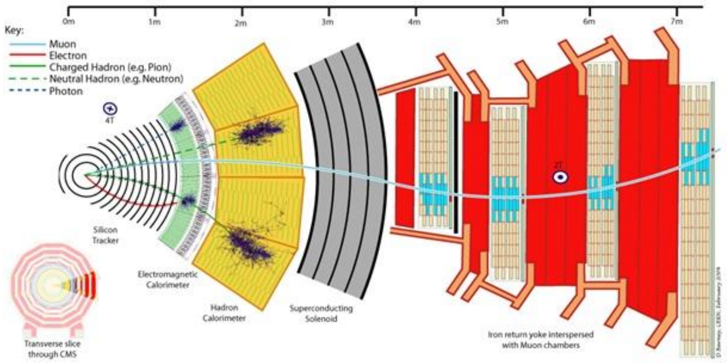
$$p(x \mid \theta) = \int dR'\, dR\, dH\, dz\; \delta\big(x - g(R')\big)\; \delta\big(R(H) - R\big)\; p_{\mathrm{sim}}(H \mid z)\, p_{\mathrm{gen}}(z \mid \theta)$$

$p_{\mathrm{gen}}(z \mid \theta)$



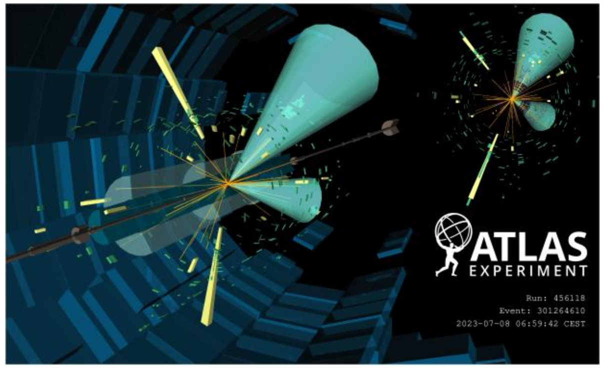
Event generation

$p_{\mathrm{sim}}(H \mid z)\, p_{\mathrm{gen}}(z \mid \theta)$



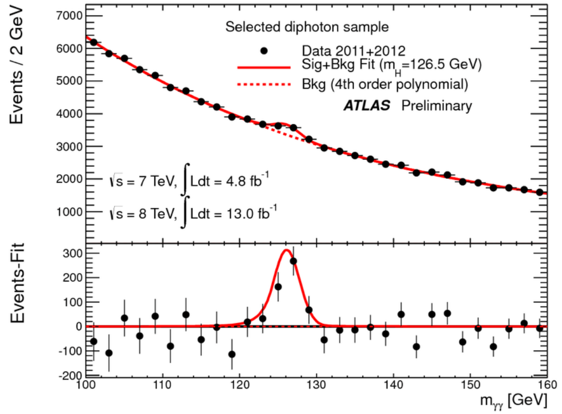
Detector interaction

$\delta\big(R(H) - R\big)\, p_{\mathrm{sim}}(H \mid z)\, p_{\mathrm{gen}}(z \mid \theta)$



Reconstruction

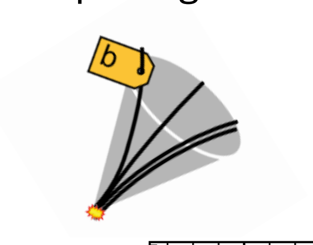
$p(x \mid \theta)$



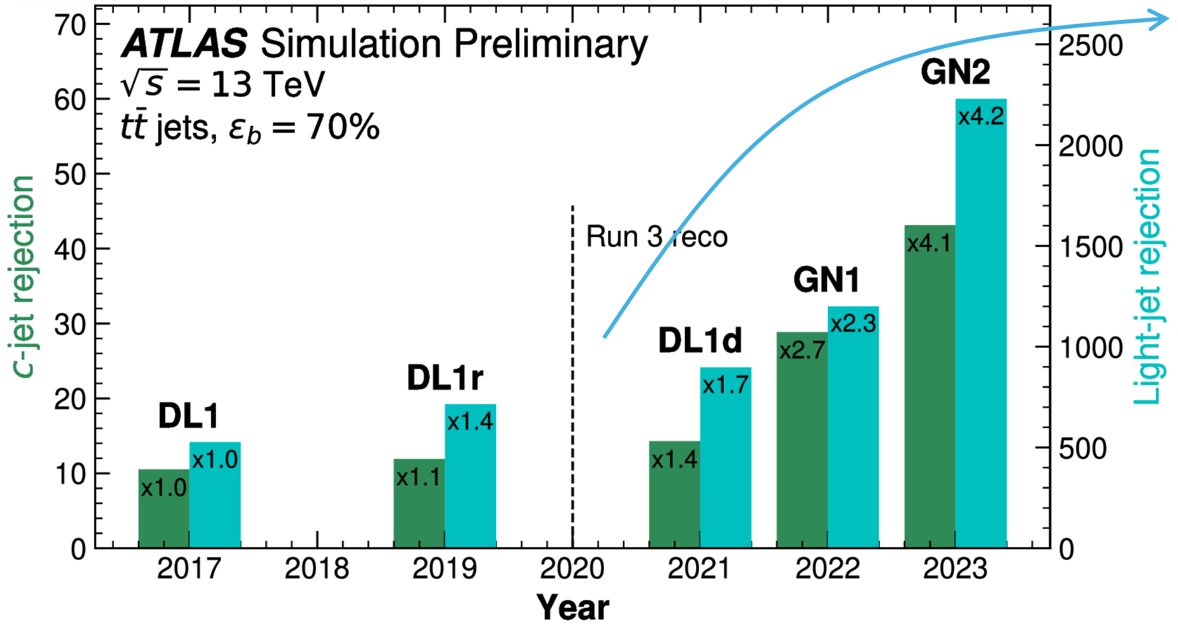
Inference

Jet identification

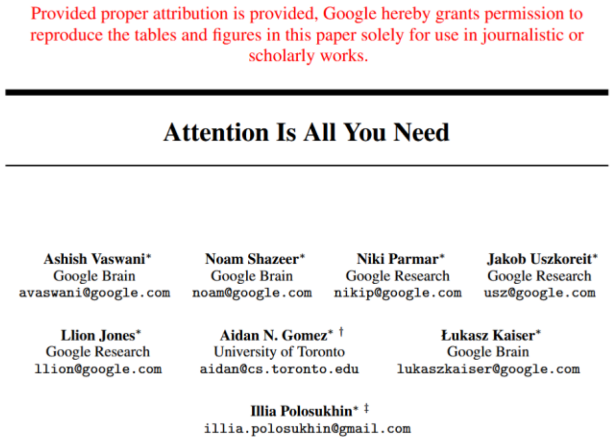
How is it ML impacting hep-ex? In multiple ways, peraphs one of the clearest example comes from jet tagging
Opening new frontiers, e.g. can we measure Higgs-charm coupling at the (HL-)LHC?



Are we saturating?



[Nature Comm. Tranforming Jet Flavour Tagging at ATLAS]


Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

Attention Is All You Need

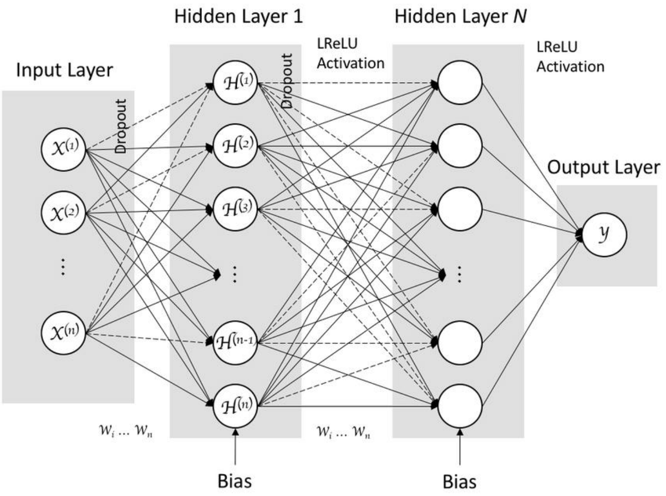
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Building block of a Transformer

Review of architectures

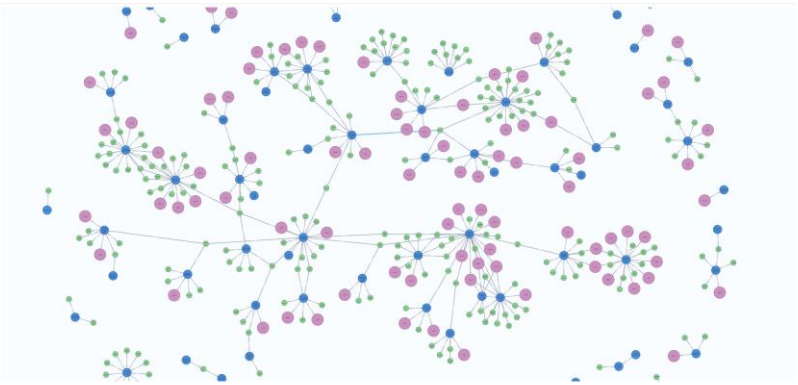
- Taking a step-back on the algorithms used in exp. HEP to improve data analysis
- A very large portion of problems are related to the more complex **Set-to-Set** category

Vector-to-vector



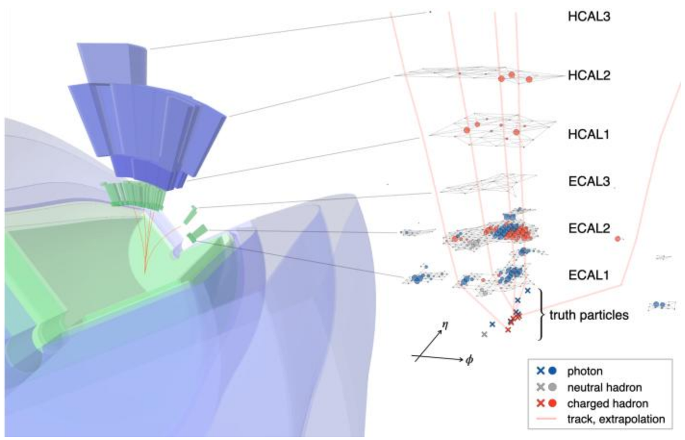
Known and used since many years

Set-to-Vector



Cardinality of the inputs is not fixed but output is

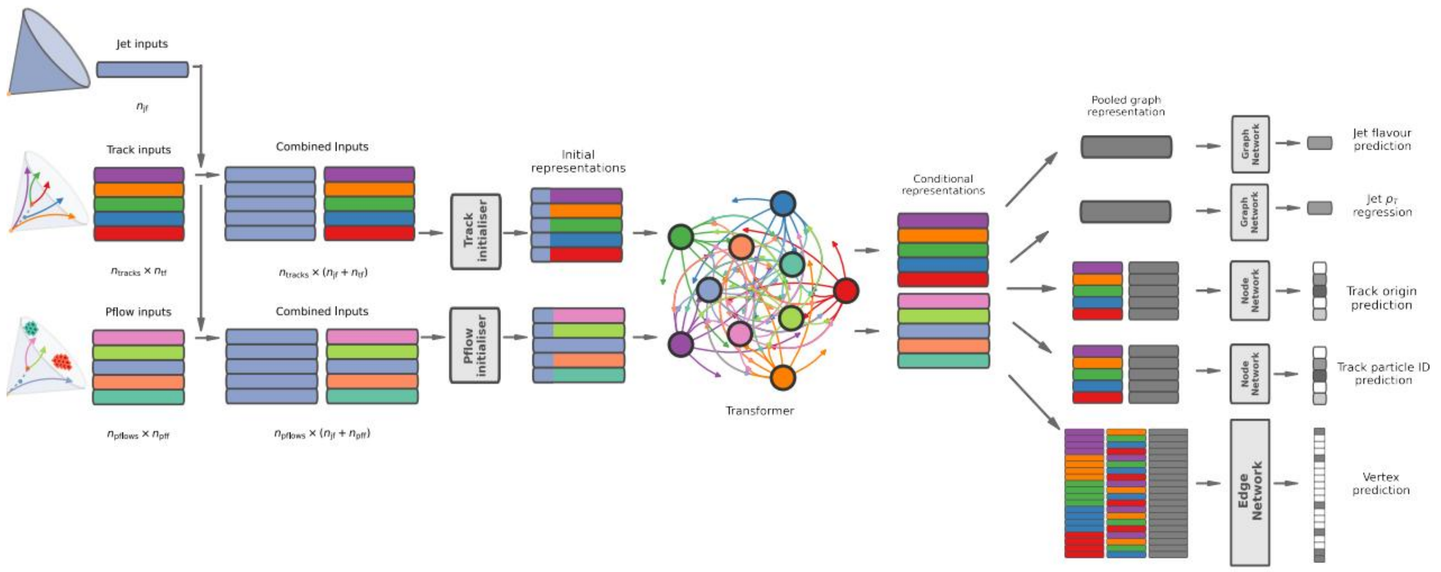
Set-to-Set



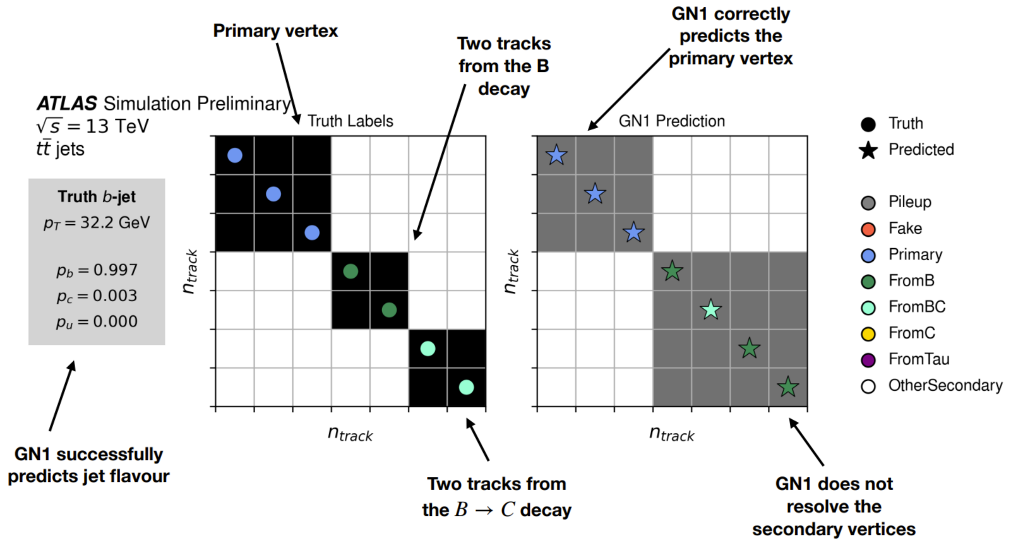
Cardinality of both the input and the output are not fixed

The state-of-the art now (meaning is old already)

- A jet is a set of inputs, the point is to represent a jet as a fixed size vector in high dimension efficiently
- State of the art flavour tagging algorithm. This start looking like a foundation model (yet a selected one)



This is a general workflow: heterogeneous inputs



Can also predict more granular information: decay chain topology

This is all trained using MC simulations, can we have control the performance on real collision data?

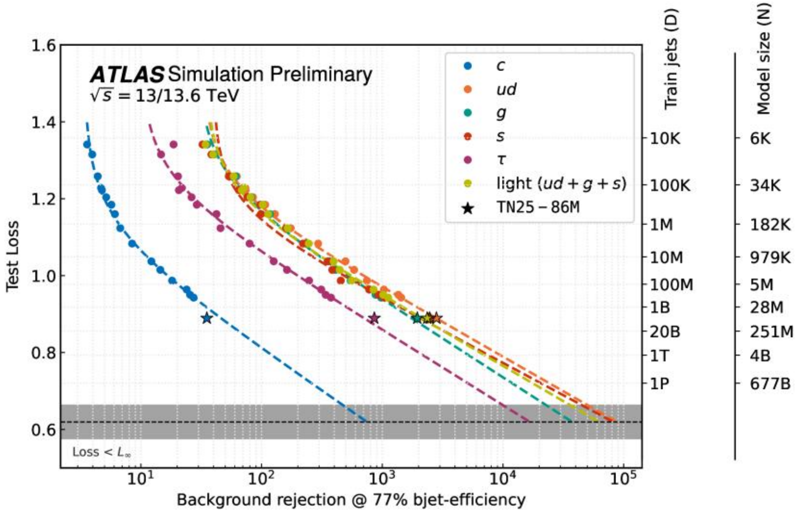
GN3: Multi-task, Multi-modal Transformers for Jet Flavour Tagging in ATLAS

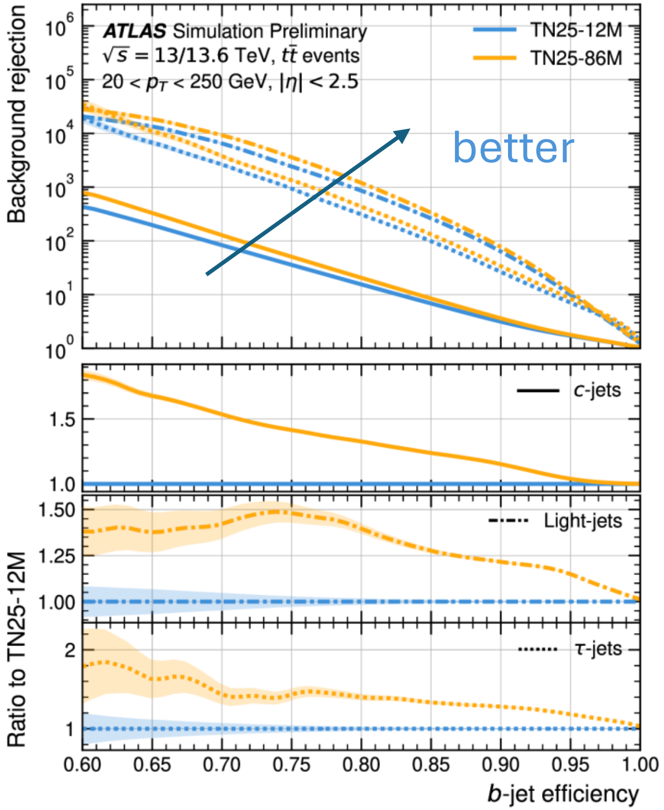
Question of scale?

- Industry applications driven by large language model (LLM) studied in detail how these models ML scale
- Can predict expected improvements from data and model size. Compared to industry we are small, but is there room for improvements? How much?

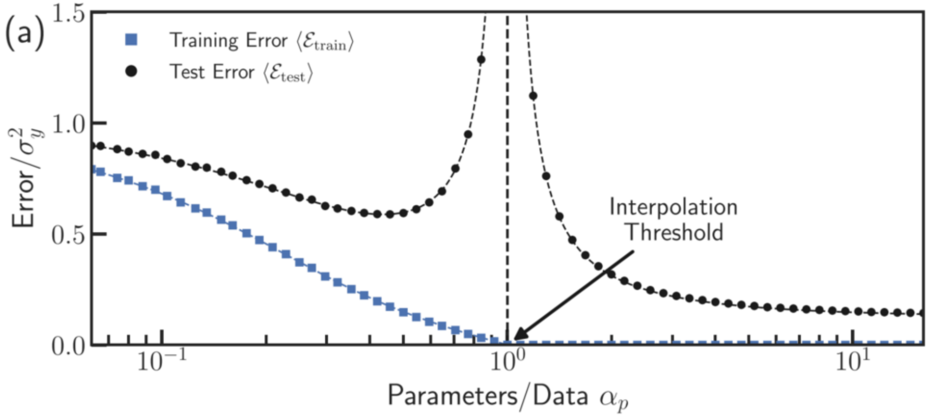
Scaling laws, DeepMind 2022

$$L(N, D) = L_\infty + \frac{A}{N^\alpha} + \frac{B}{D^\beta},$$





Double descent known from CS – where are we?



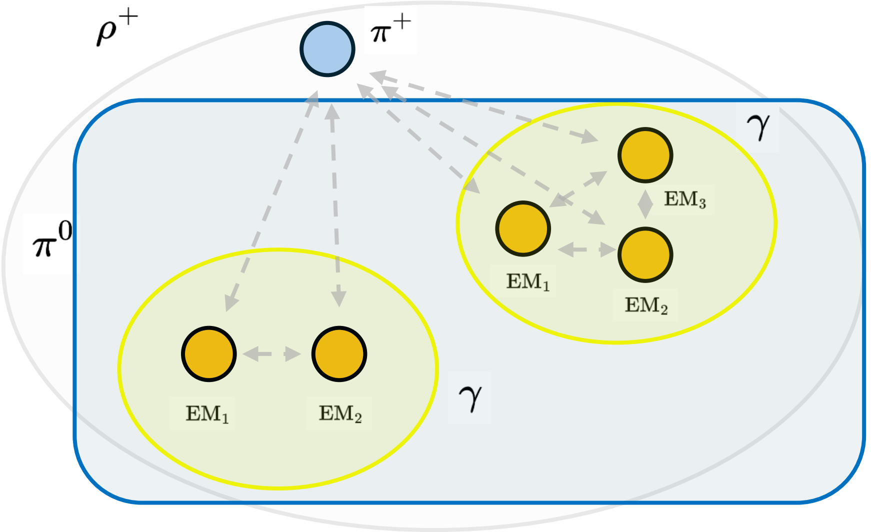
Wikipedia

Carpe datum [ATLAS PUB]

Towards set-to-set problems

Tracking and global particle flow algorithms are major example of set-to-set problems , decay chain reconstruction is also an application

Hypergraph reconstruction



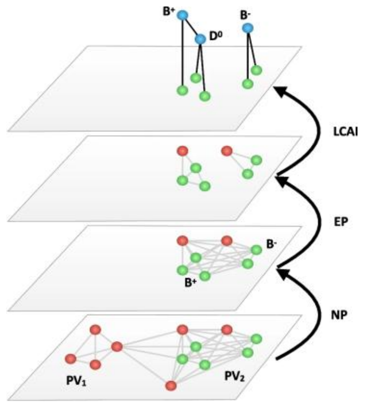
$$\tau^+ \to \rho^+ + \nu \to \pi^+ + \pi^0 + \nu \to \pi^+ + \gamma + \gamma + \nu$$

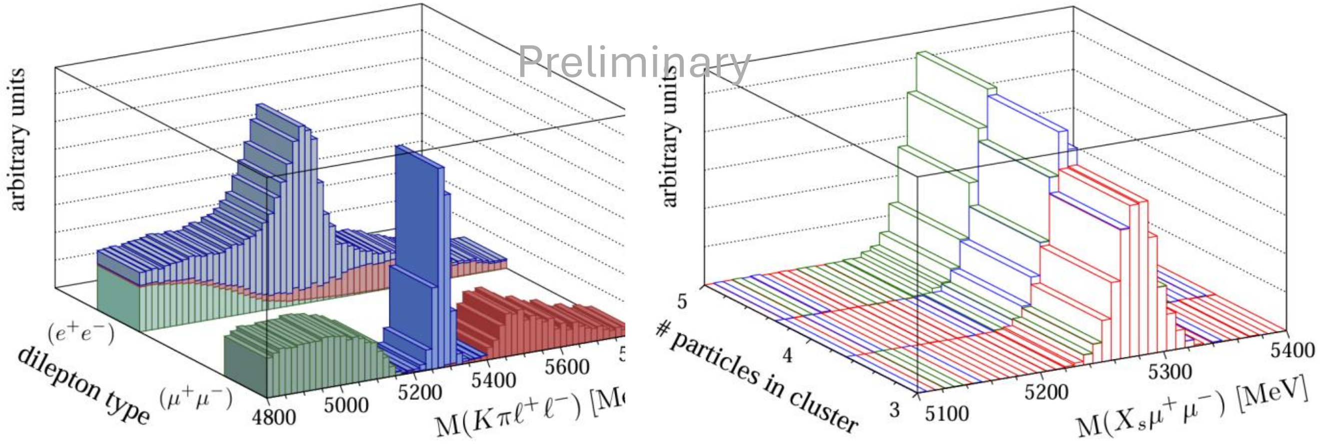
Global particle flow with hypergraphs

Topology-agnostic, hierarchical graph network for event reconstruction

Hierarchical decay tree reconstruction

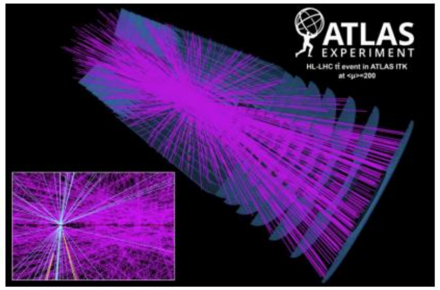
$B^+ \to K^+\pi^-\pi^+\ell^+\ell^-$
$B^0 \to K^+\pi^-\ell^+\ell^-$
$B^+ \to K^+\ell^+\ell^-$



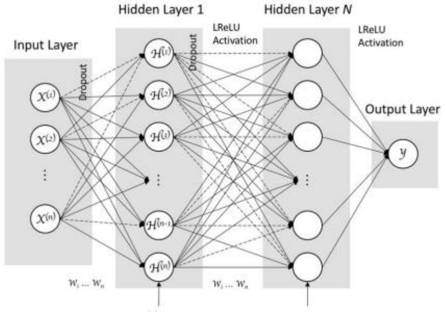


DFEI_lhcb_pisa_2025

End-to-end algorithms?



0101011101010101
0011010101010011
1111000101010101
1010100100001111
1000101010101010
1010101001010100
1100000011111100
1010101010010101



Is this a Higgs boson?

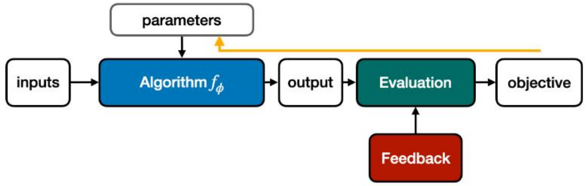
Will we ever reach this stage?

1. Dimensionality: need to probe from a space of around O(10^7) dimensions.

1. Data inherently have structures (locality) that are essential reduce the dimension of the problem as well as for generalization.

3. There is a problem of domain shift: we need a way to control mis-modelling of the simulations we use to train our algorithms.

$$\frac{\partial L}{\partial \phi} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial \phi}$$

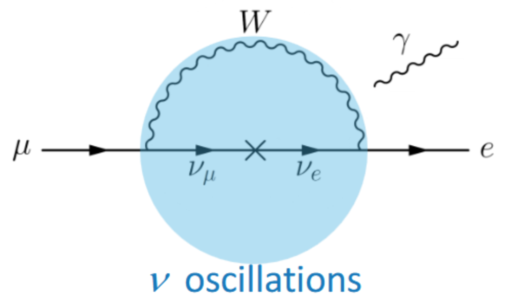


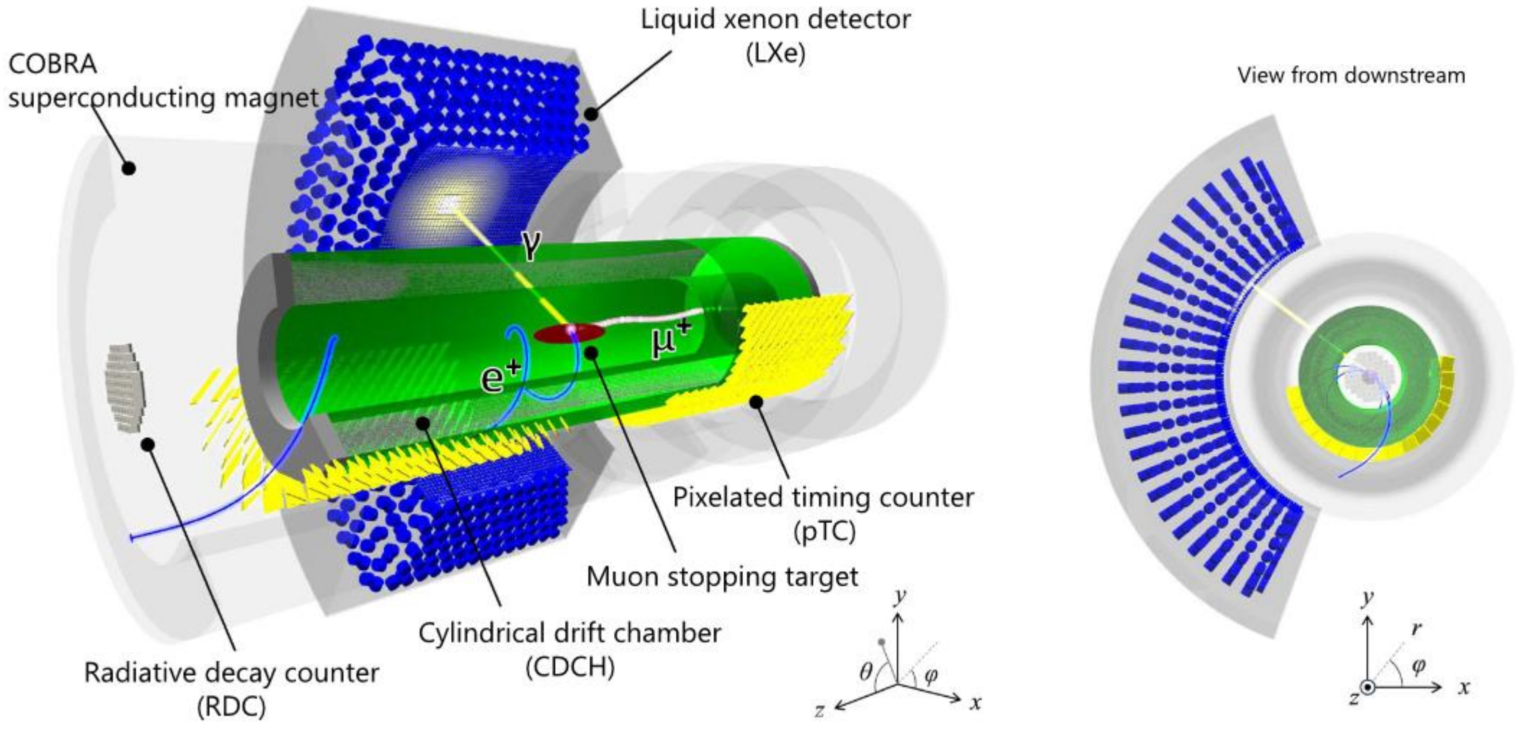
→ **requires algorithms and evaluation to be differentiable**

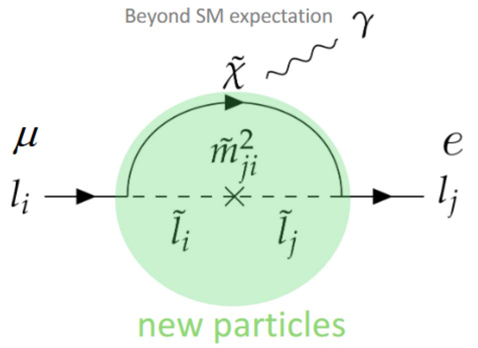
Henirich et al., Kagan et al, Rambelli et al

MEG II – application of a transformer for pattern recognition

- Intensity frontier – searching for lepton flavour violation
- Contineous stream of particle.. Tracking struggles at high pile-up
- Pecularity here is there is no reliable simulations that can be used for target definitions



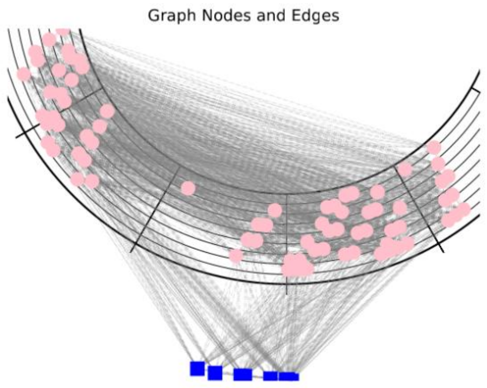




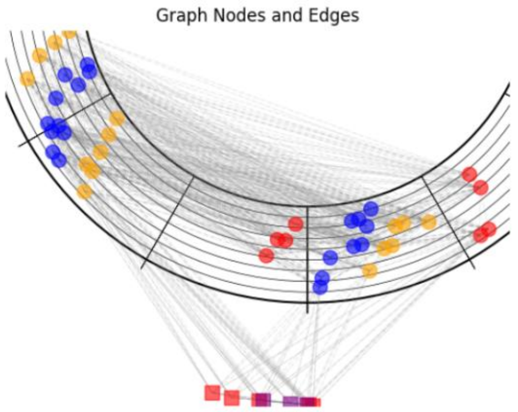
Pattern Reco. improvements

[2512.19482] Transformer-Based Approach to Enhance Positron Tracking Performance in MEG II

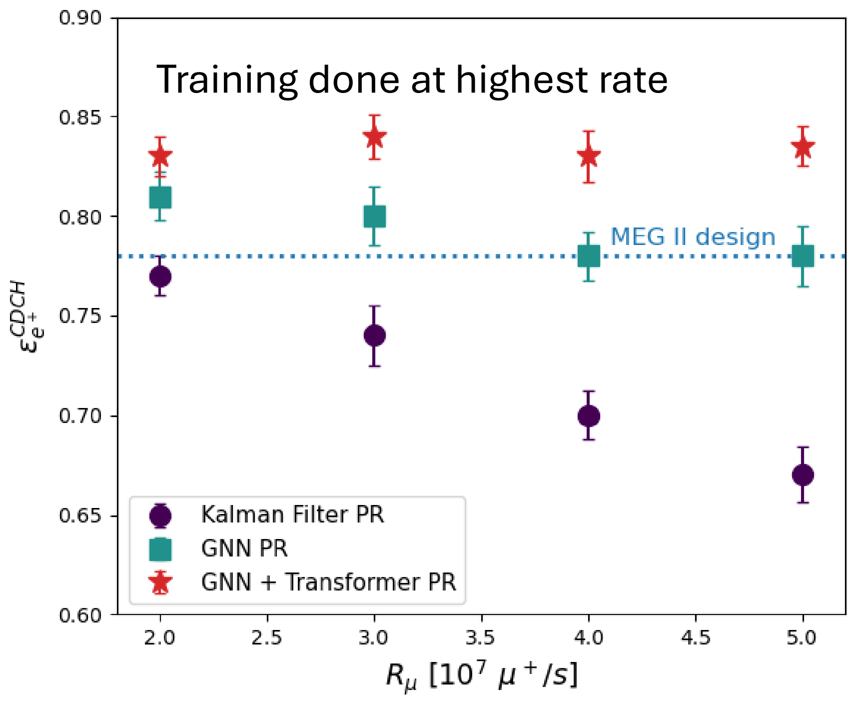
Constrained by full reco on data



O(1k) per event

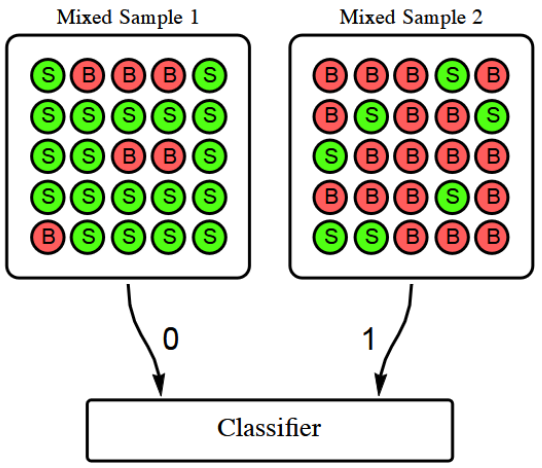


Two approaches investigated based on a GNN and a transformer using different inputs



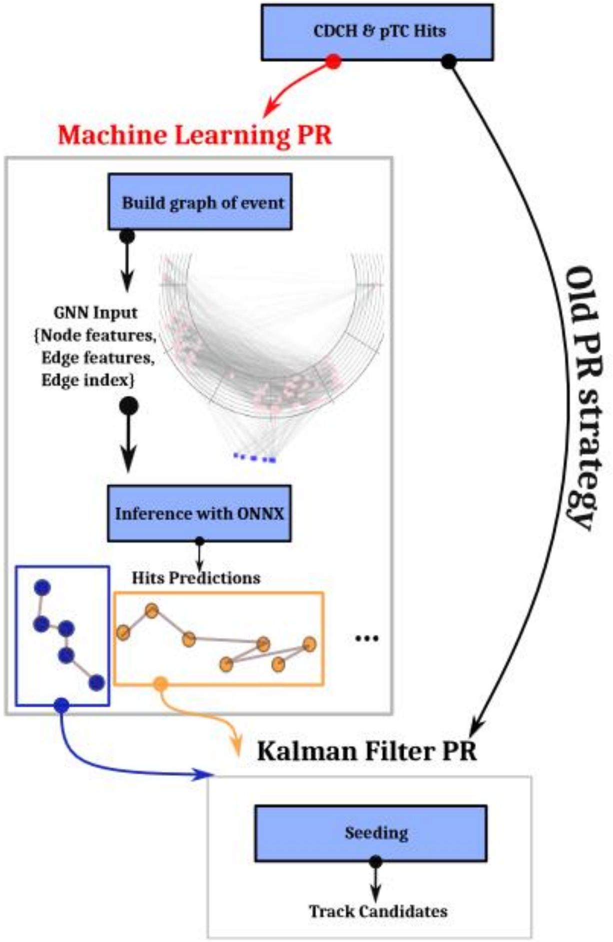
Combination provides best performance, **flat vs rate!**

Weak supervision: 'classification without labels'

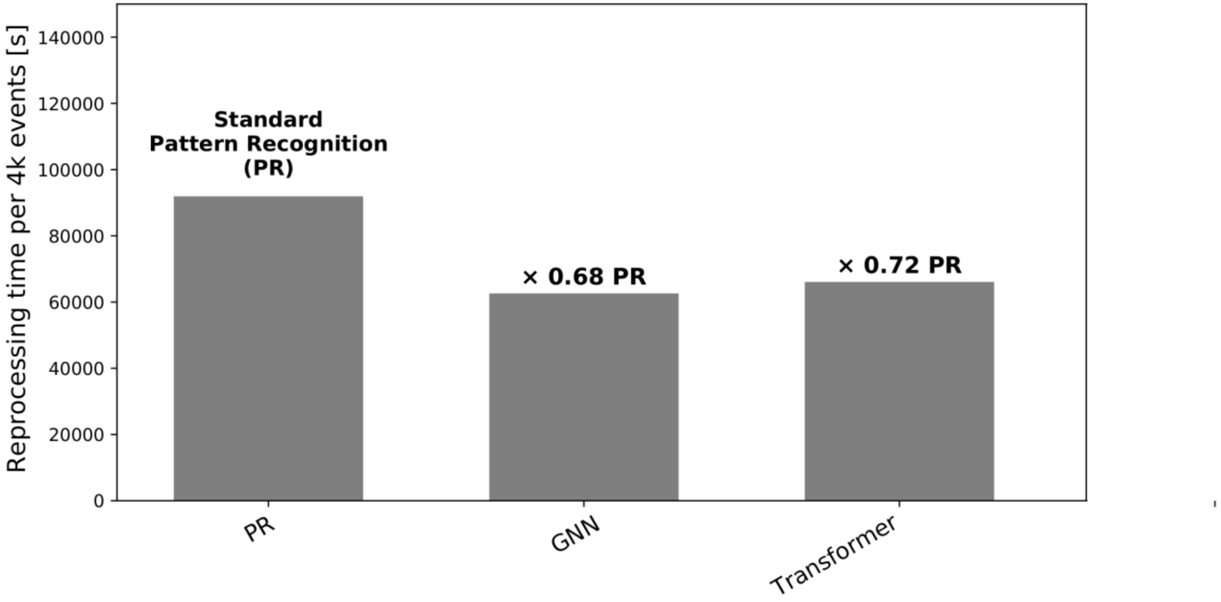


weak supervision

Deployment



Fully deployed and running!

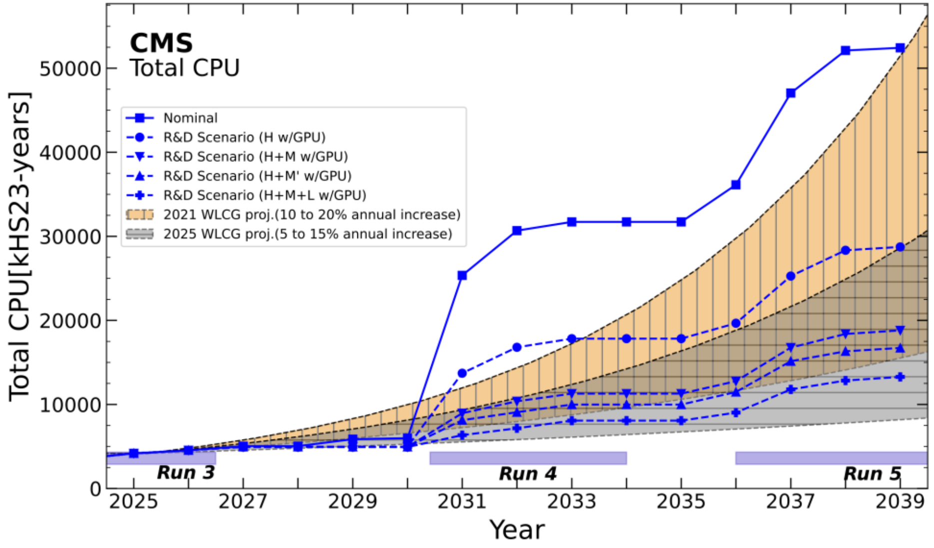


Next step end-to-end reconstruction: End-to-End Multi-track Reconstruction Using Graph Neural Networks at Belle II

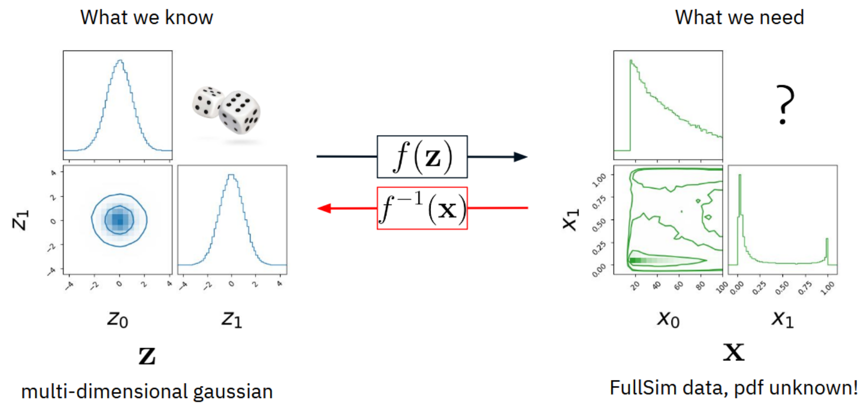
Flash sim

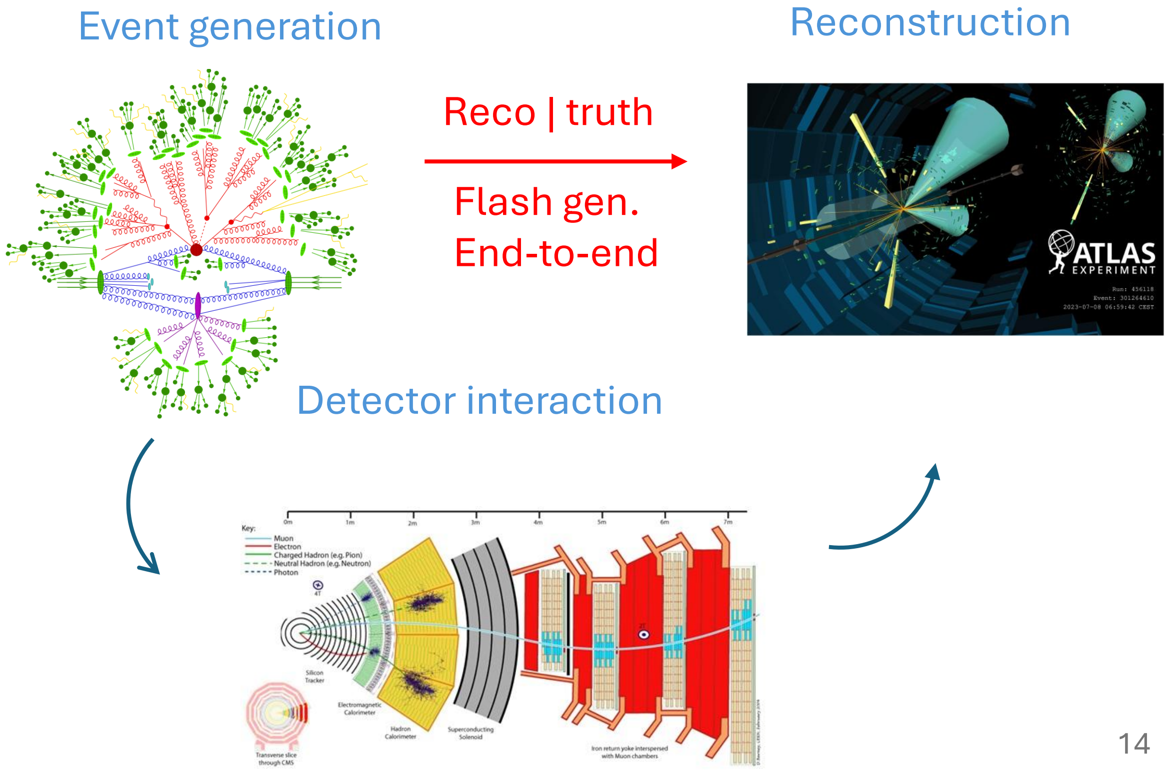
A. Rizzi CHEP

- The increasing volume of data demands a massive scale-up in Monte Carlo statistics
- The event full simulation accounts for **~50%** of the total CPU budget
- Fast simulation is a big deal at LHC experiments
- Based on normalizing flow – or its contineous version (more recent) flow matching
- FlashSim can produce 100M events in 1 day on a single commercial GPU. 2–4 orders of magnitude faster than FullSim



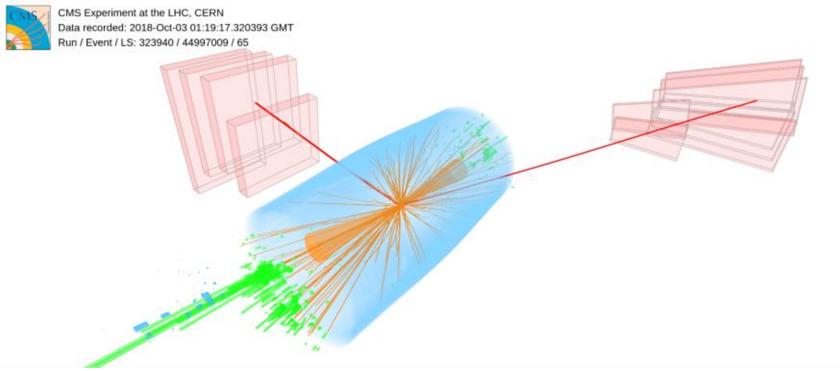
LHCC-G-186



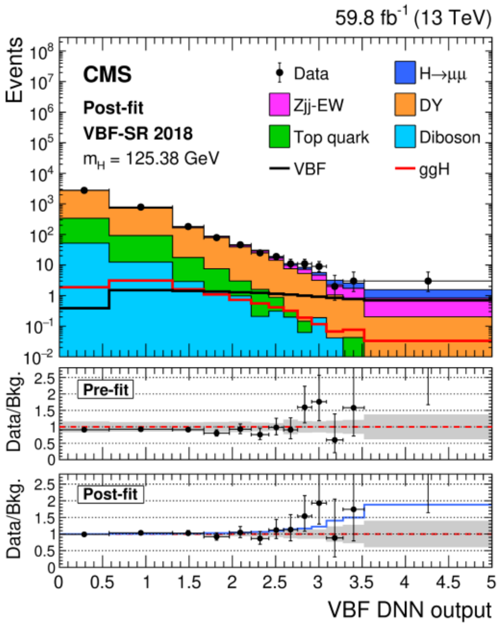


Flash sim

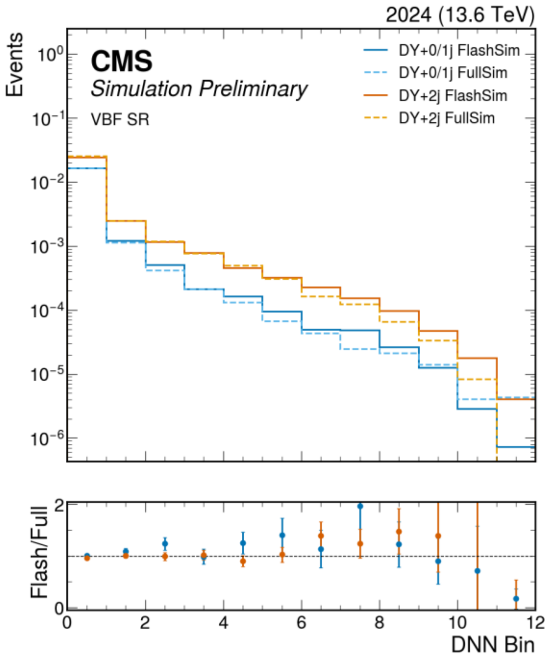
- Tested in recent years first on toy datasets, recently on real CMS simulation, and it works!
- Good performance in a realistic H⟶μμ analysis scenario and versatility in producing very large simulated sample. Show promising at trigger level as well!
- Large effort on inference infrastructures, long story short: speed improvement is more than 3 orders of magnitude!



Validation

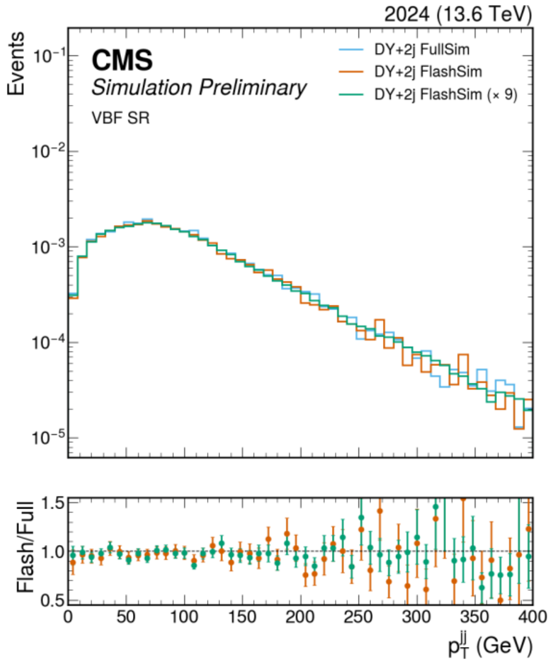


10.1007/JHEP01(2021)148



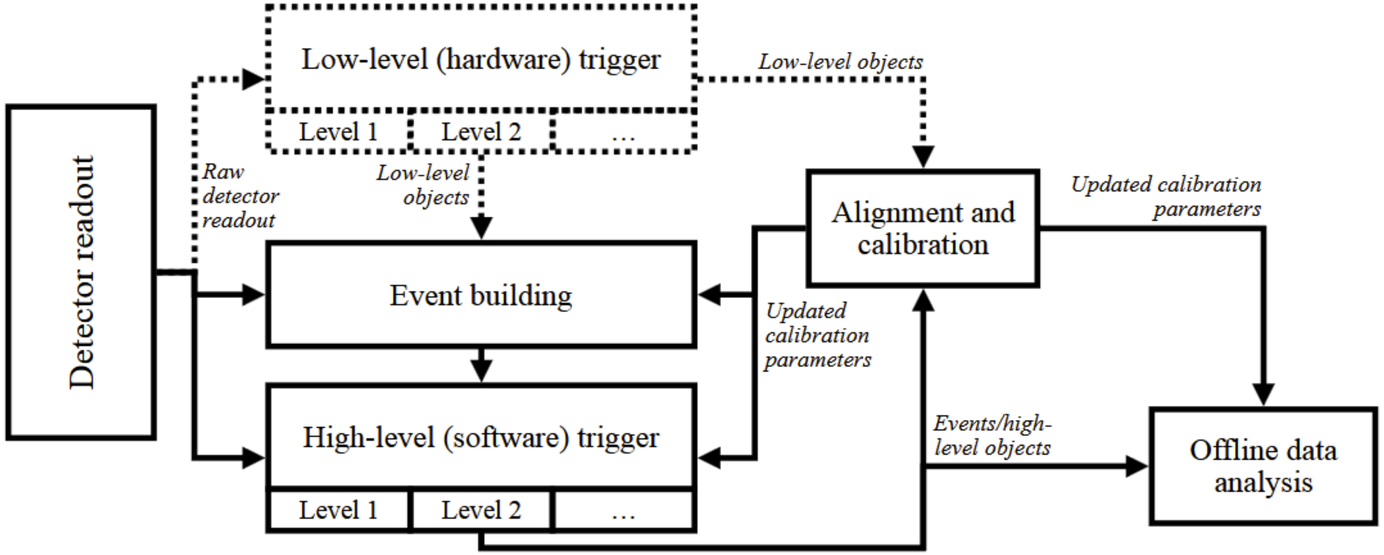
F. Cattafesta CHEP

Oversampling



Part two:

fast and ultra-past inference

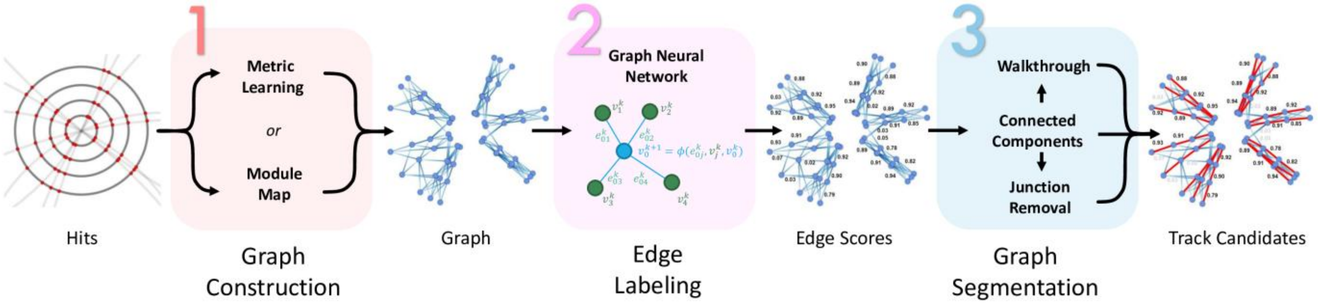


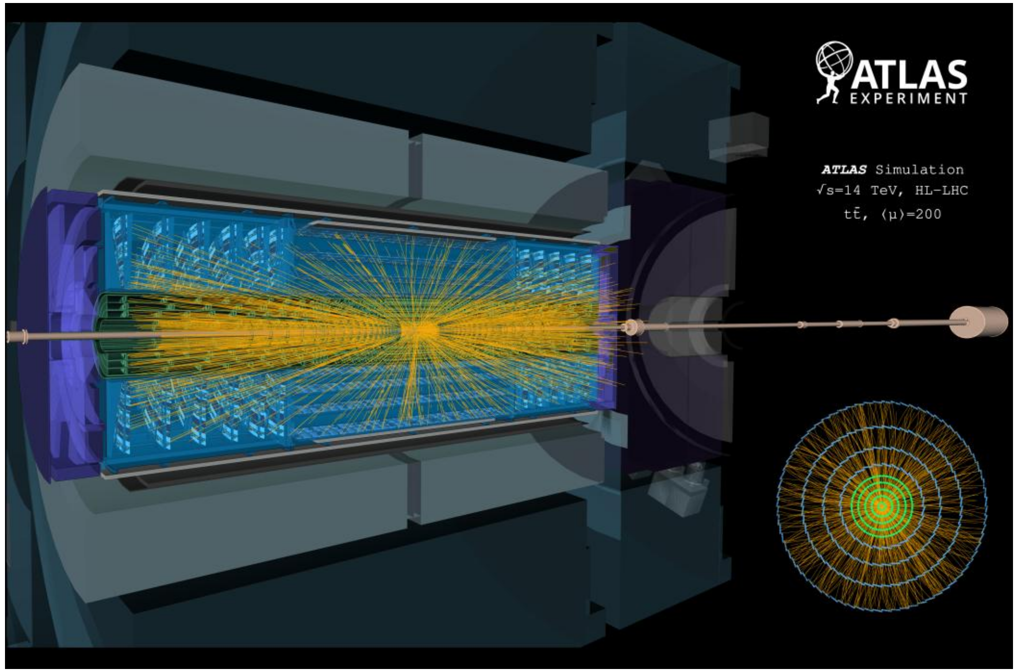
[Summary of the trigger systems of the Large Hadron Collider experiments ALICE, ATLAS, CMS and LHCb]

ATLAS tracking for software trigger at HL-LHC

ATLAS compared three strategies (tracking performance, cost, maintainability, risk assessment, …):

CPU only - CPU+GPU - CPU+FPGA



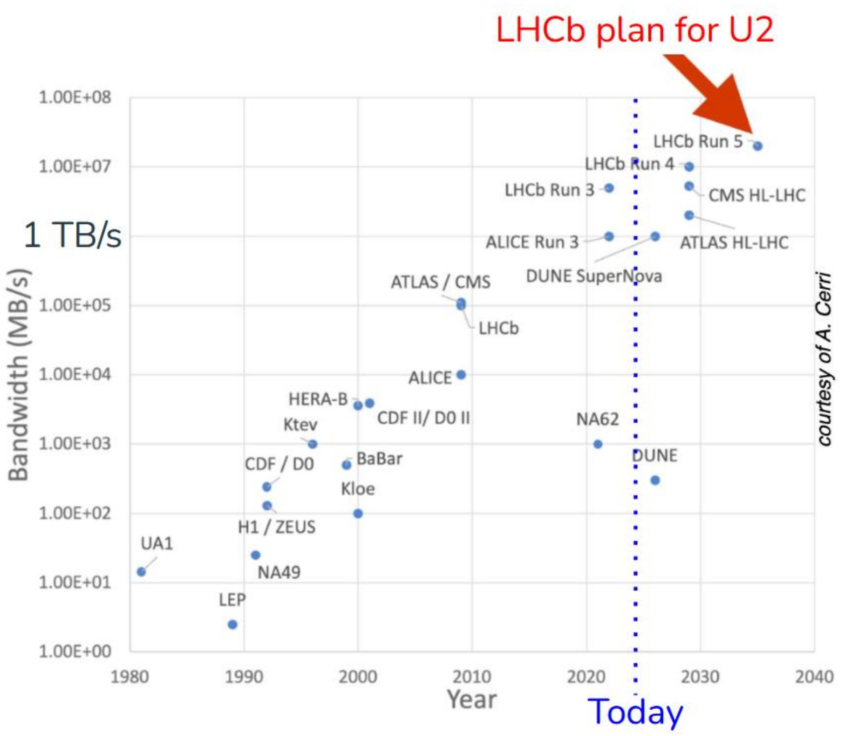


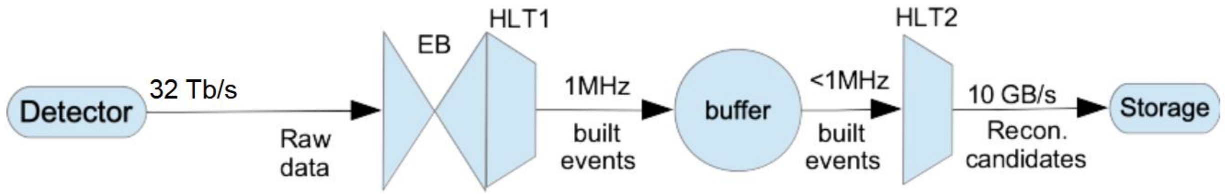
CPU+GPU is the selected choice

Code and name of pipeline (CPU / GPU / FPGA)	Clustering	Seeding	Track Finding	Ambiguity & fake resolution	Track fitting
C100 ACTS-based Fast Tracking	15%	40%	← 45% →		
C230 as C100 with Graph-Based Track Seeding		GBTS			
F100 as C100/C230 with clustering offloaded to FPGA		(GBTS)			
F150i as F100 with seeding run on the FPGA too					
G080 as C230 but clustering is offloaded to GPU		GBTS			
G130 as G080 with GBTS seeding also offloaded to GPU		GBTS			
G200 End-to-end GPU Traccc					

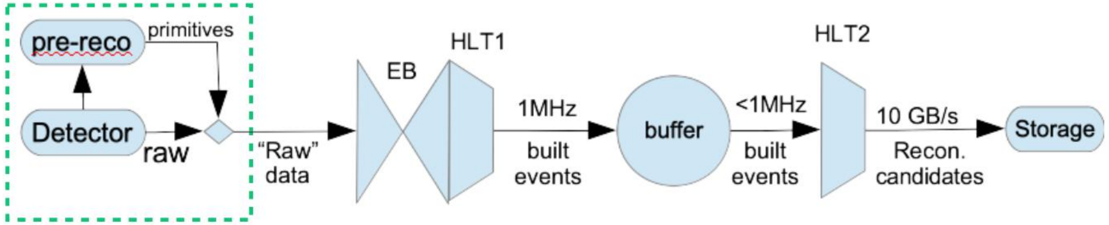
The Retina Project at LHCb

- Flavour physics at low energy needs full event reconstruction
- Very high cross-section of b-physics means many interesting events
- LHCb uses a triggerless readout of the whole detector + full event reconstruction before first trigger decision is made
- Two-level trigger: HLT1 on GPU, HLT2 on CPU - no hardware
- Is it possible to exploit hardware to accelerate reconstruction, maintaining a full-software trigger?





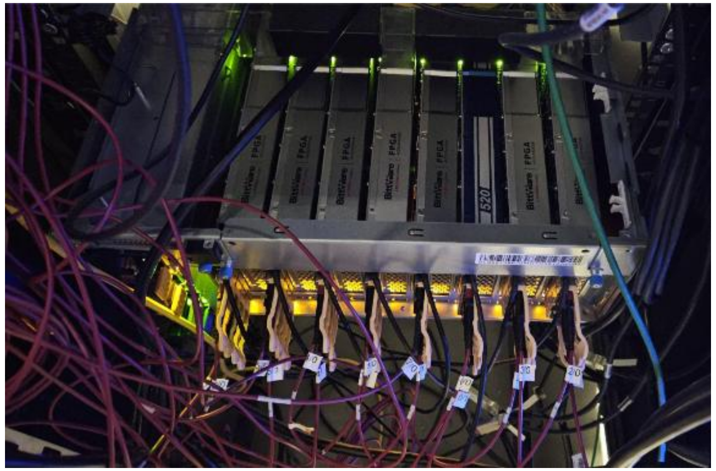
New with FPGA

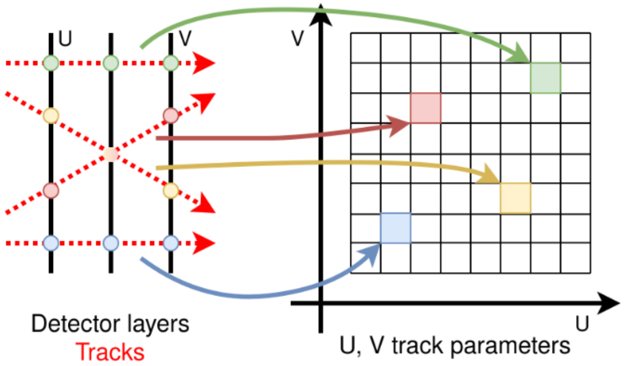


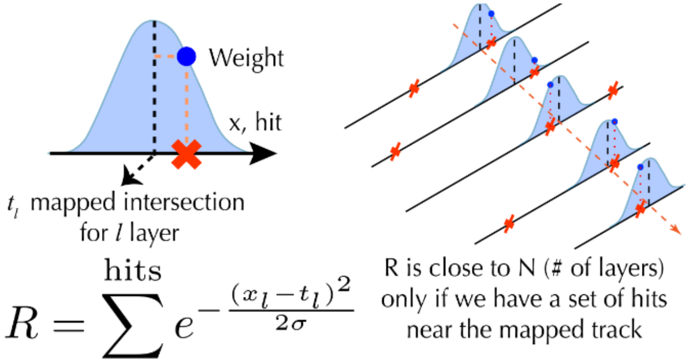
Retina core idea

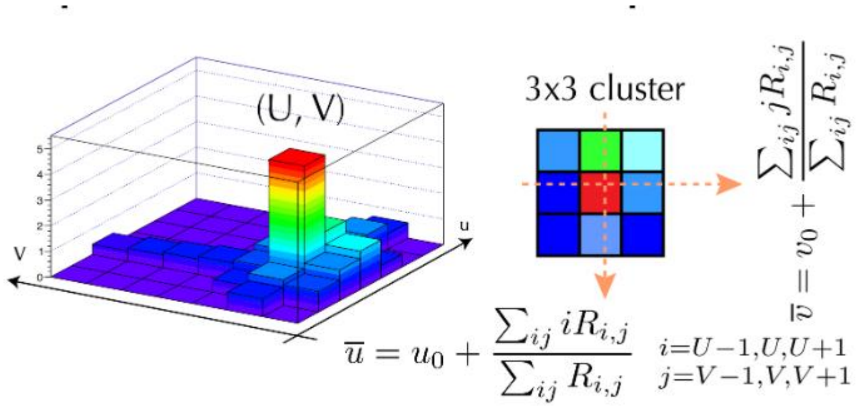
More Info

- Reconstruction complexity should be O(n^l), where l is the number of layers.
- Huge work to develop algorithms with lower time complexity. Many of them can scale to O(n^2) time complexity.
- Highly-parallel architecture for pattern recognition with O(n) complexity . Inspired to the early stages of image processing L.Ristori(2000)NIMA.



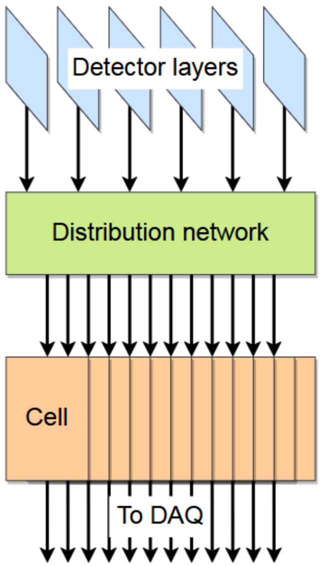


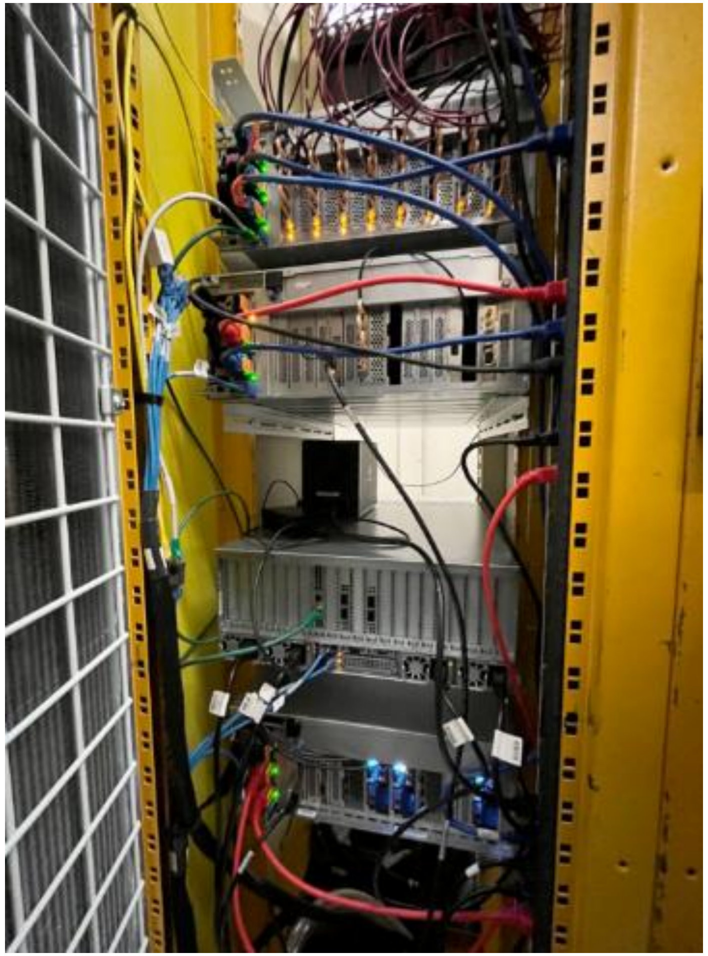


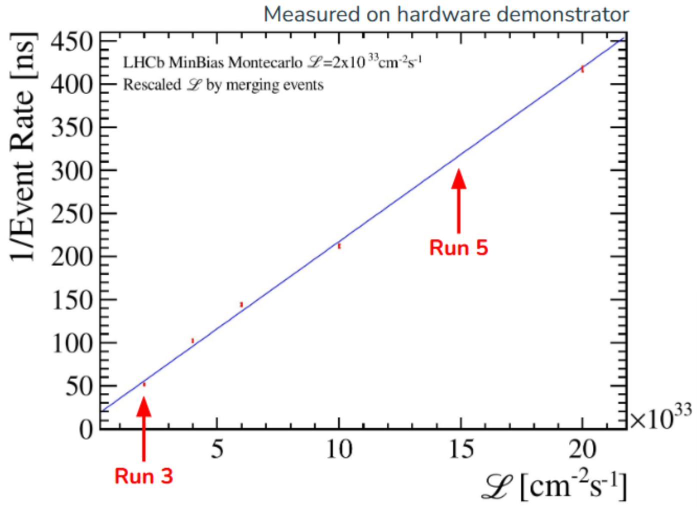


Retina - results

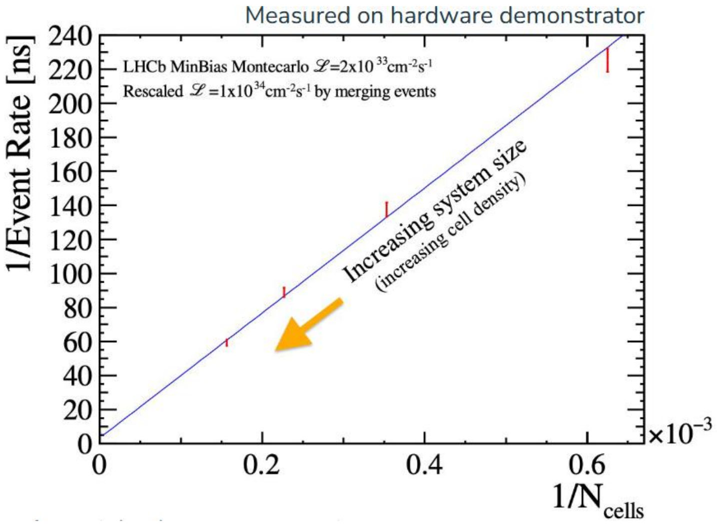
- Achieved O(n) scaling with Retina. Crucial for HL-LHC
- Fitting also performed on same FPGA, Nice performance reached on the demonstrator
- Emulate higher luminosities by event overlapping
- **Plan to implement for next LHCb run (Run 4) LHCb trigger TDR**







High luminosity emulation



Larger systems emulation

Retina – applications

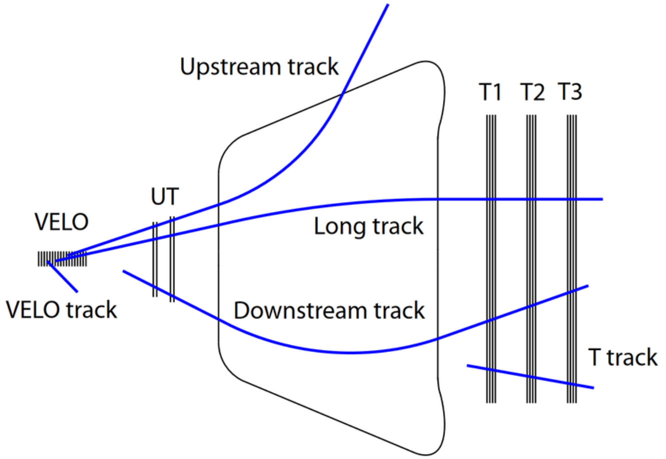
Can be creative with the new infrastructure:

Reconstruct T-tracks replacing HLT seeding algorithm

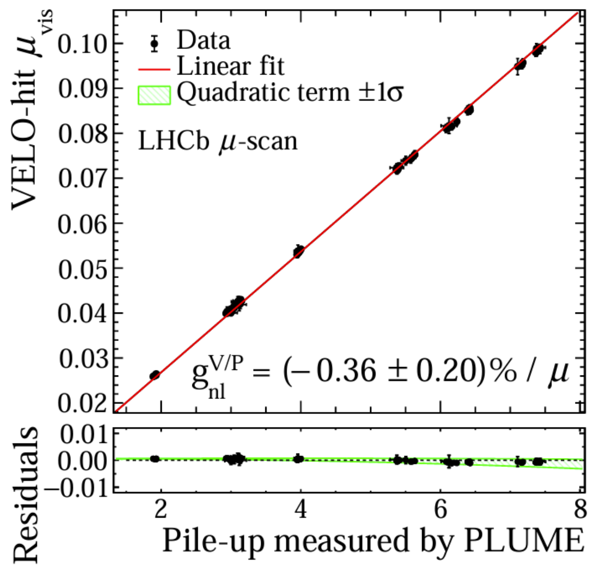
- Effect on Full HLT1 sequence, long tracks matching VELO tracks and T-tracks
- Execution time: Total: 7.2 μs; Seeding: 1.5 μs
- Execution time: Total: 5.2 μs; Primitives decoding and refitting: 0.06 μs and negligible overhead

Overall HLT1 throughput increased by 33%. Makes room for further HLT1 functionality.

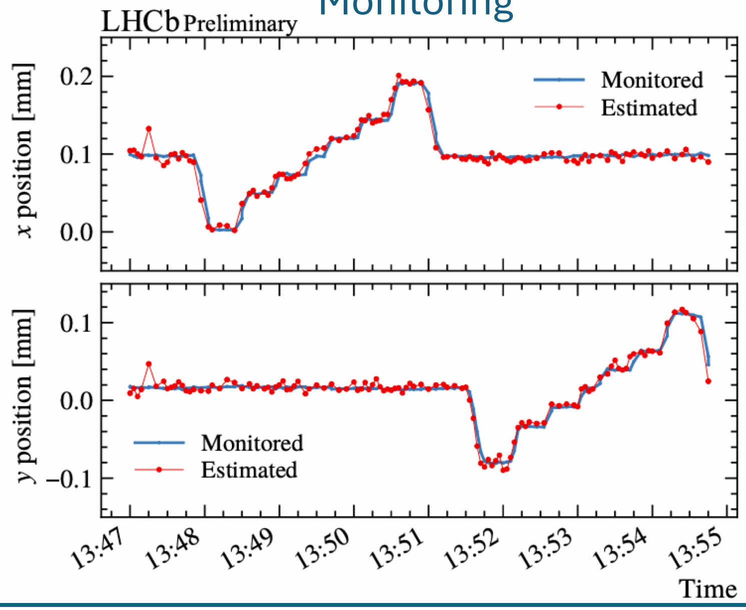
F. Lazzari, CHEP, Proposal for FPGA-based tracking in the LHCb downstream region



Luminosity



Monitoring



Inst. lumi. measurement with FPGA

Real-time monitoring of LHCb interaction region with a fast trackless methodology

A FPGA-Based Architecture for Real-Time Cluster Finding in the LHCb Silicon Pixel Detector

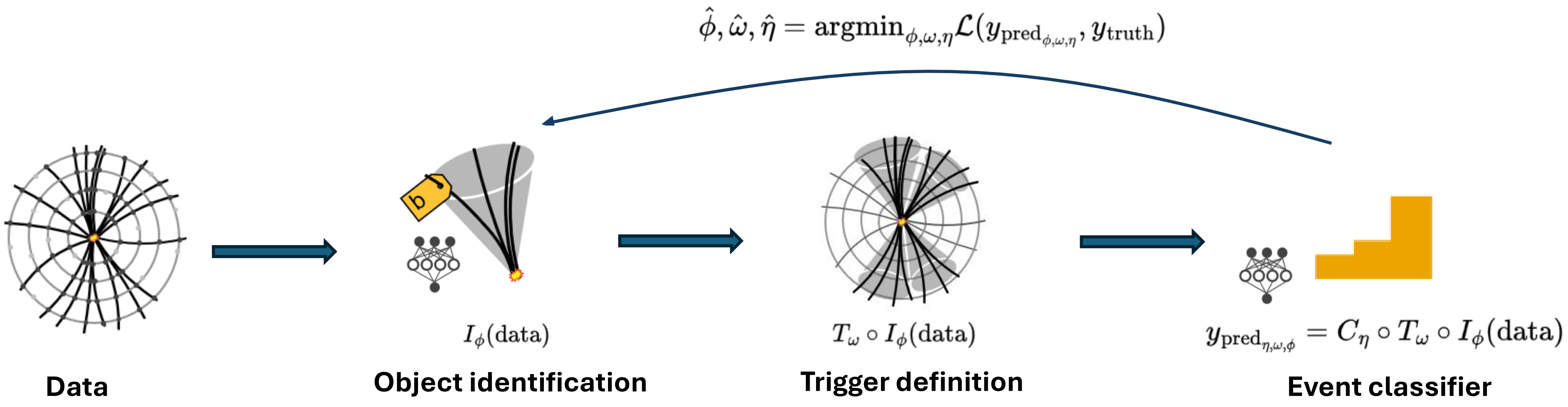
Conclusions

- Above overoptimism from the community, AI is having a real impact to the physics outcome of our experiments. I think most of the road is still unexplored.
- Hep-ex groups at UniPi strongly involved, I only scratched the surface! e.g. declarative assisted analysis
- I think next years will be devoted to further understanding of transformers, and their usage in set-to-set algorithms potentially in end-to-end hybrid approaches. Calibration technques also a big deal, i.e. how do we ensure a mismodeled feature is not learned?
- Fast and Ultra-fast inference is also a strong area of expertise.
- It is a crucial time for LHC experiments… HL-LHC is very close!

<u>F. Dyson “*new directions in science are launched by new tools more often than by new concepts. The effect of a concept-driven revolution is to explain old things in new ways. The effect of a tool-driven revolution is to discover new things that have to be explained.*”</u>

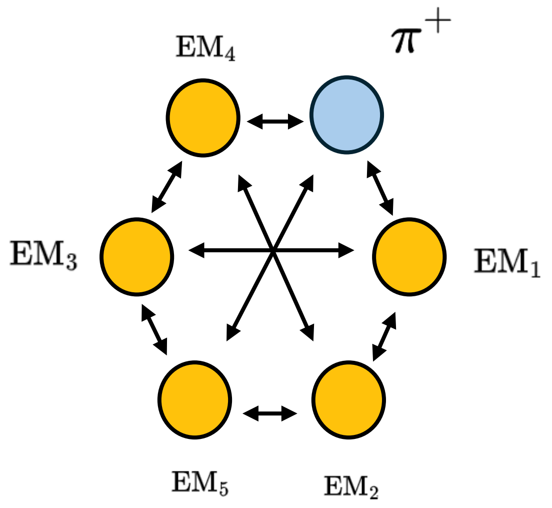
Flash sim

- We are not yet saturating yet…

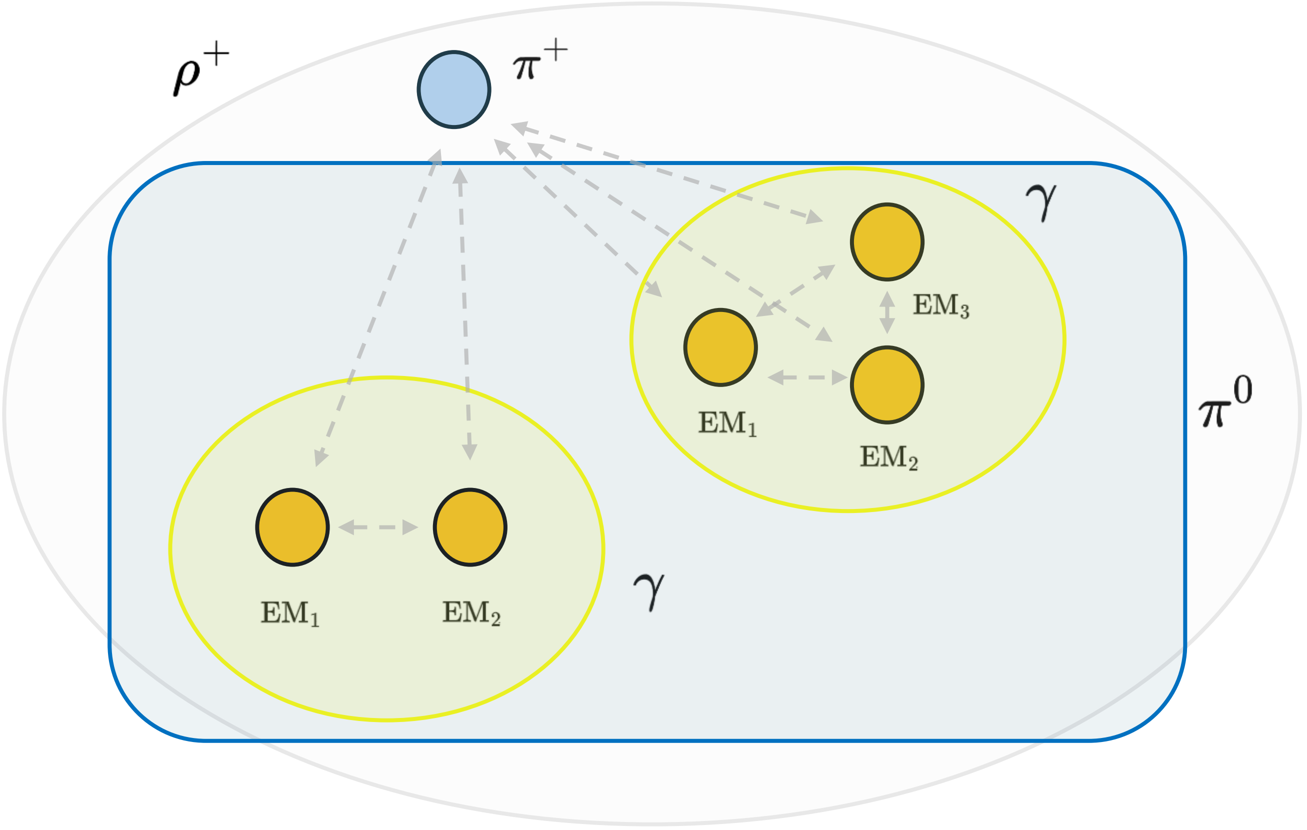


Flash sim

Traditional ML based reconstruction



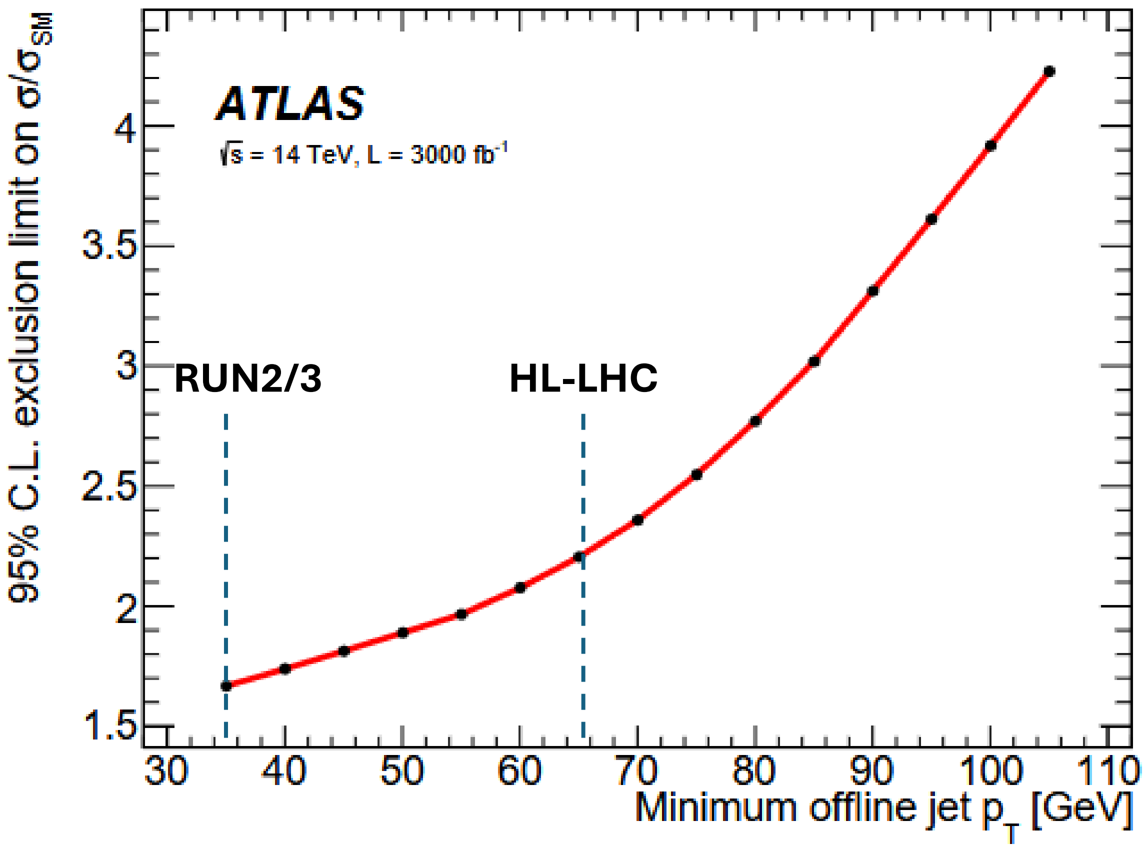
Novel ML hierarchical reconstruction



$$\tau^+ \to \rho^+ + \nu \to \pi^+ + \pi^0 + \nu \to \pi^+ + \gamma + \gamma + \nu$$

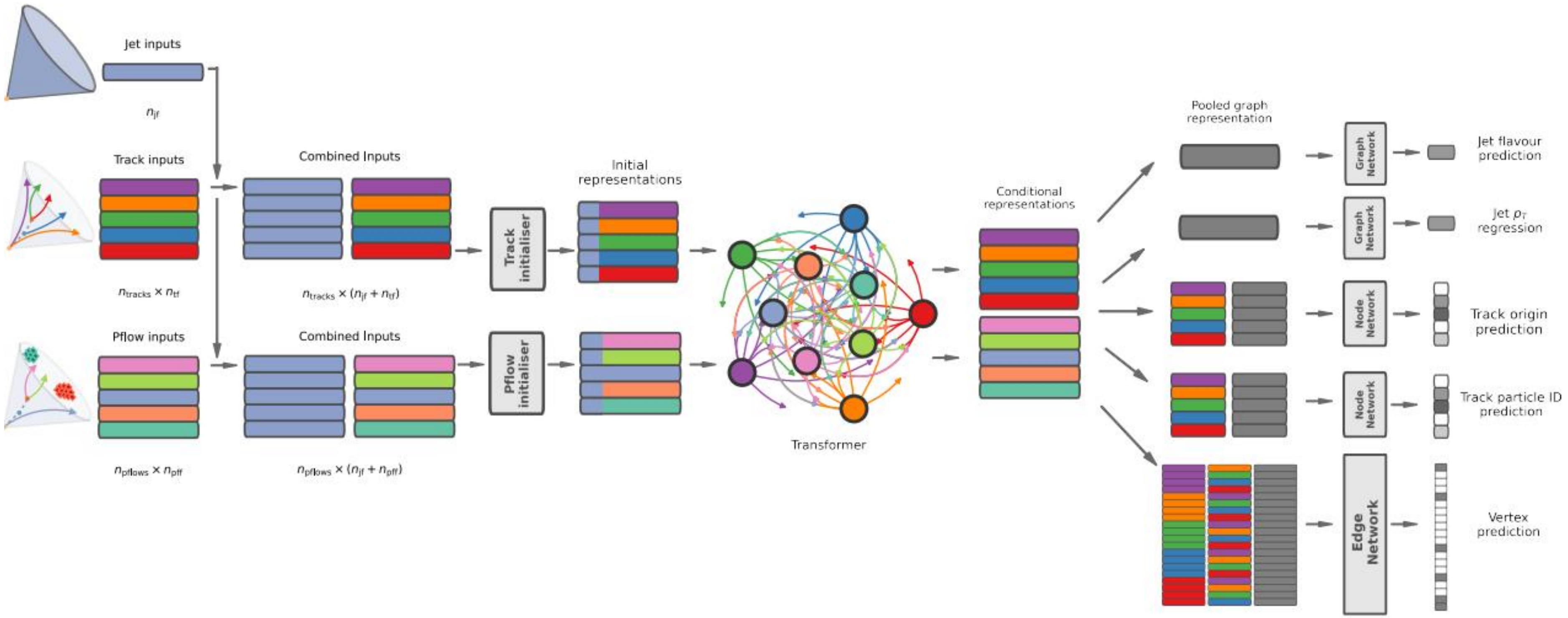
Flash sim

- We are not yet saturating yet…



The state-of-the art now

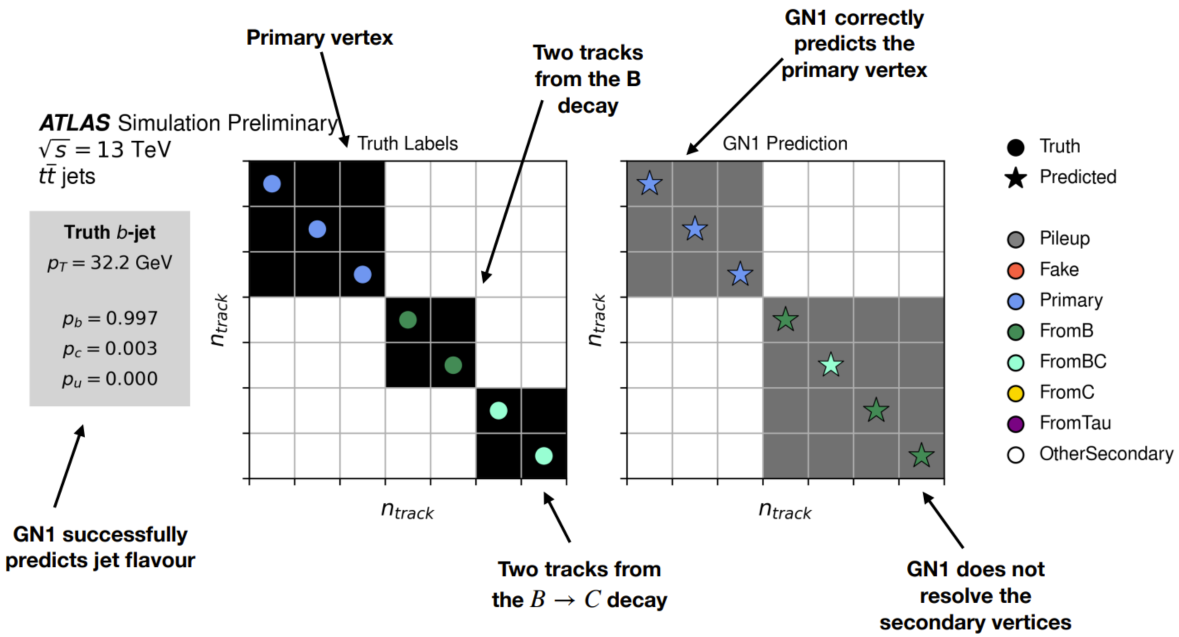
- State of the art flavour tagging algorithm. This start looking like a foundation model (yet a selected one)

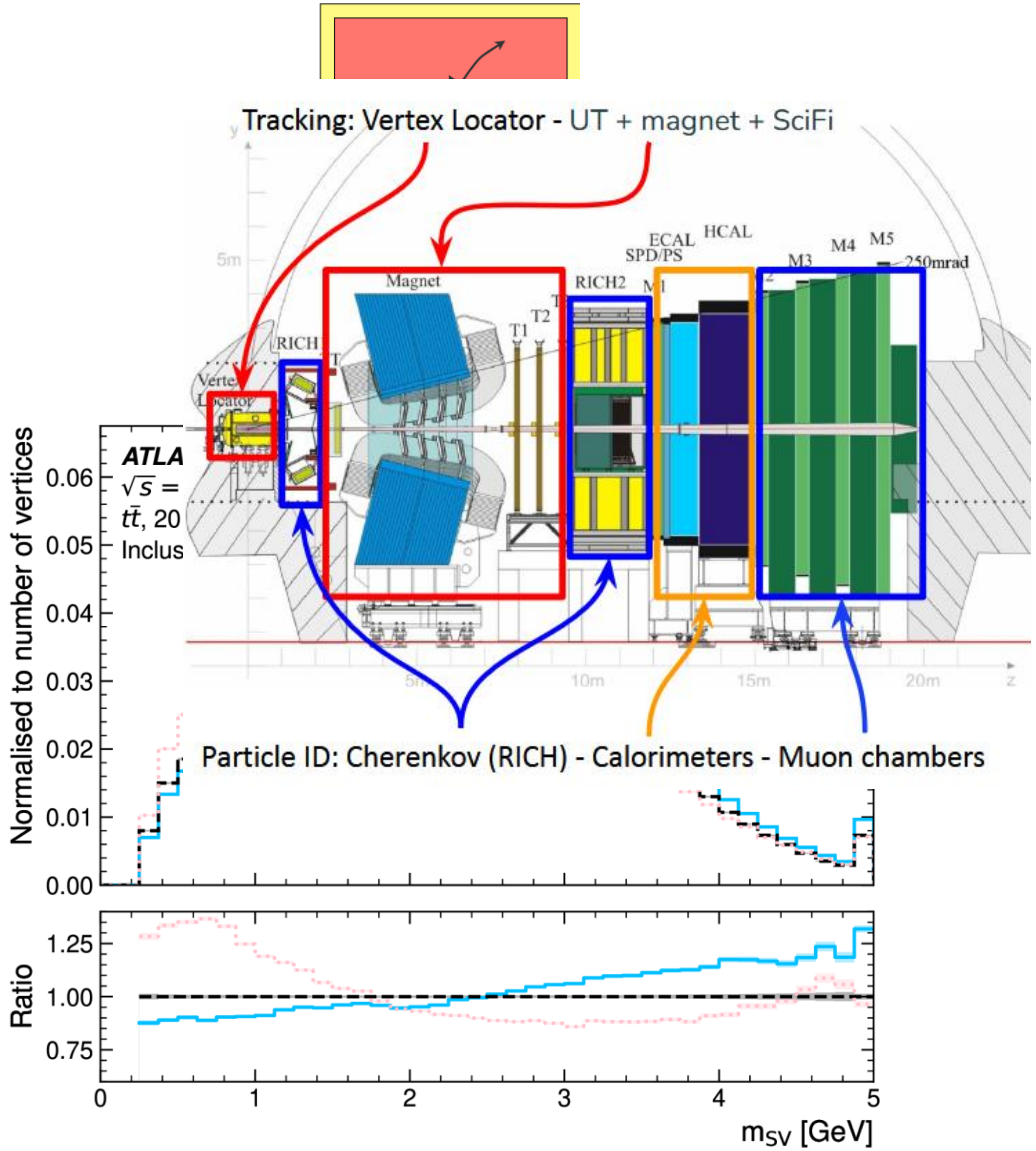


Energy regression to correct for jet momentum embedded into the model.
Does it still make sense to do tau- and b-tagging separately? ML helps not only performance but also to streamline workflow

Jet identification

- Aux. tasks helps both performance and intepretability

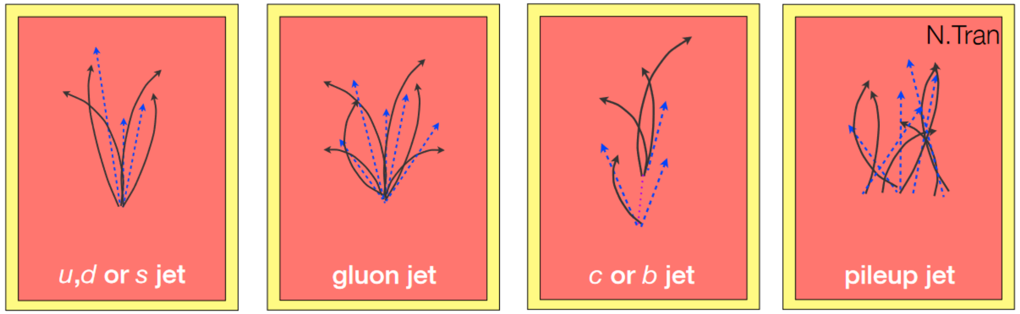


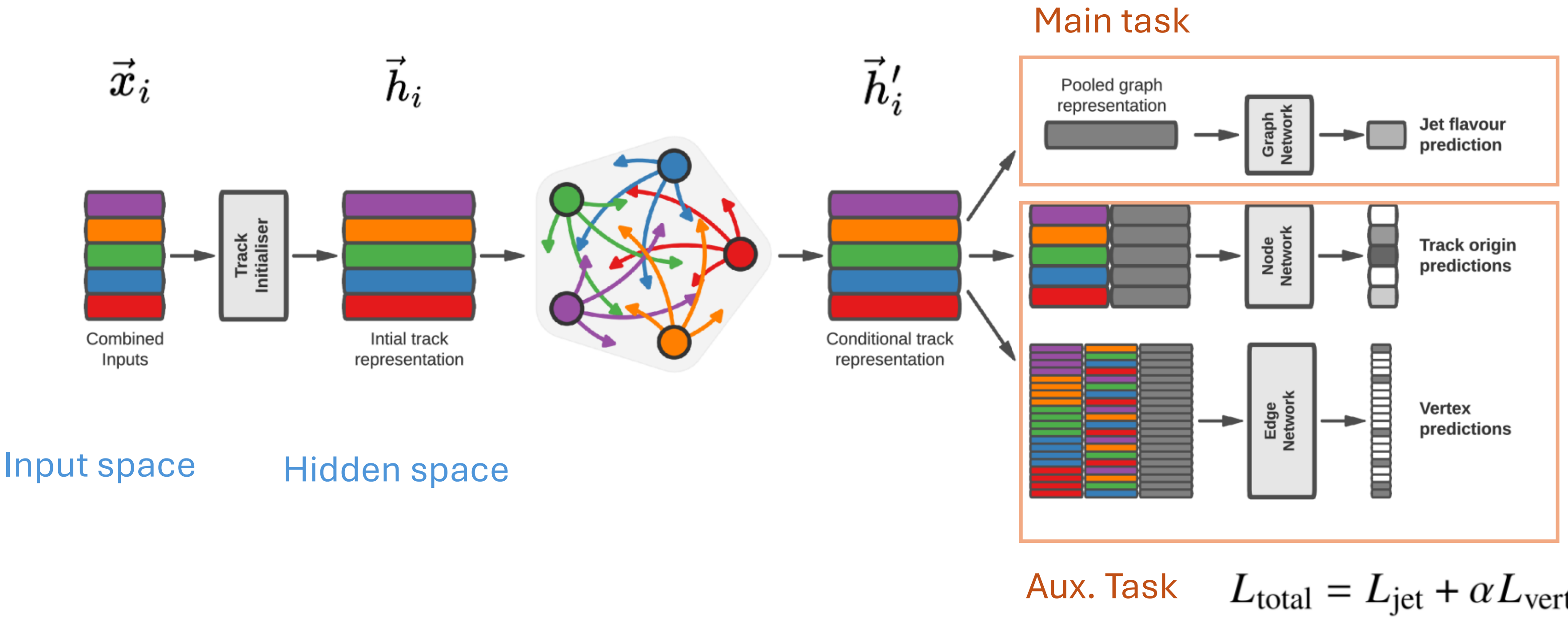


Examples from aux taks

Jet identification

- Inputs are recontructed low-level object: no vertex information
- Hidden dim. h is O(100)

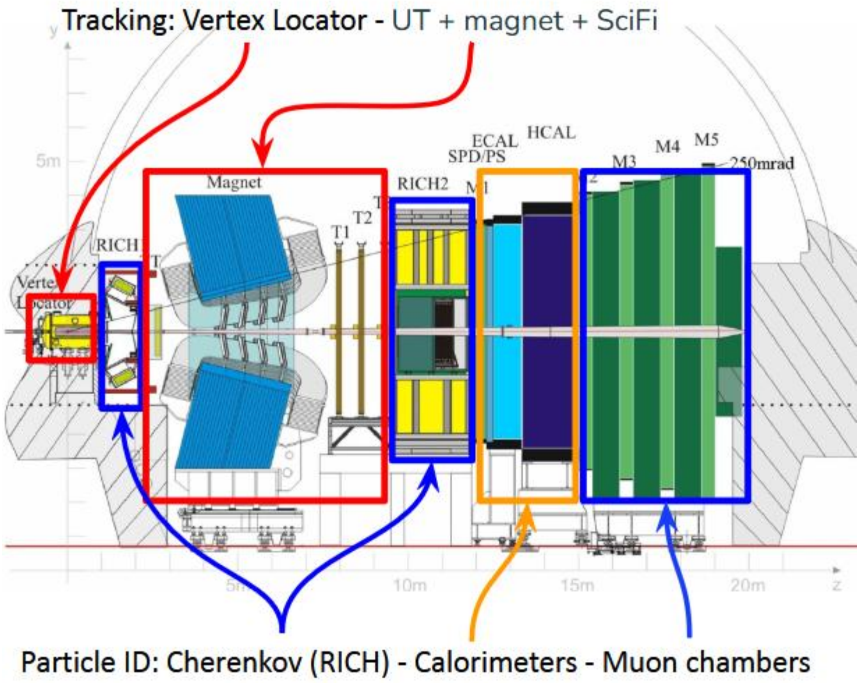




$$L_{\text{total}} = L_{\text{jet}} + \alpha L_{\text{vertex}} + \beta L_{\text{track}}$$

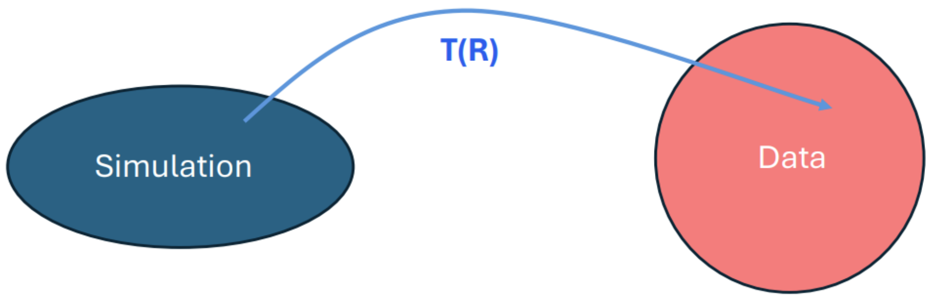
A set of particle encoded into an high dimensional fixed size space

LHCb

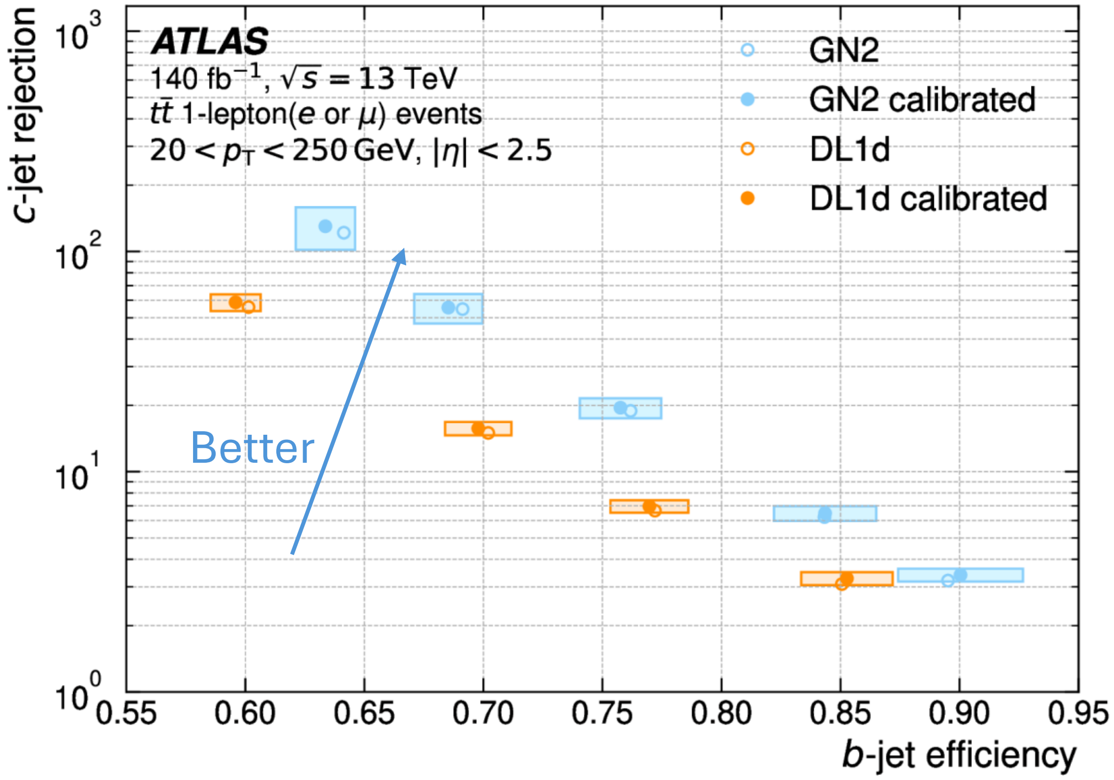

Tracking: Vertex Locator - UT + magnet + SciFi
Magnet
RICH1
RICH2
Vertex Locator
T1
T2
SPD/PS
ECAL
HCAL
M1
M3
M4
M5
250mrad
5m
10m
15m
20m
Particle ID: Cherenkov (RICH) - Calorimeters - Muon chambers

Domain shifts: calibrations

- How do we know it is well modeled in data? $$p_{\text{calib}}(x|\theta) = \int dR'\, dR\, dH\, dT\, dz\, \delta(x - g(R'))\boxed{\delta(R' - T(R))}\delta(R(H) - R)\, p_{\text{sim}}(H|z) p_{\text{gen}}(z|\theta)\,.$$

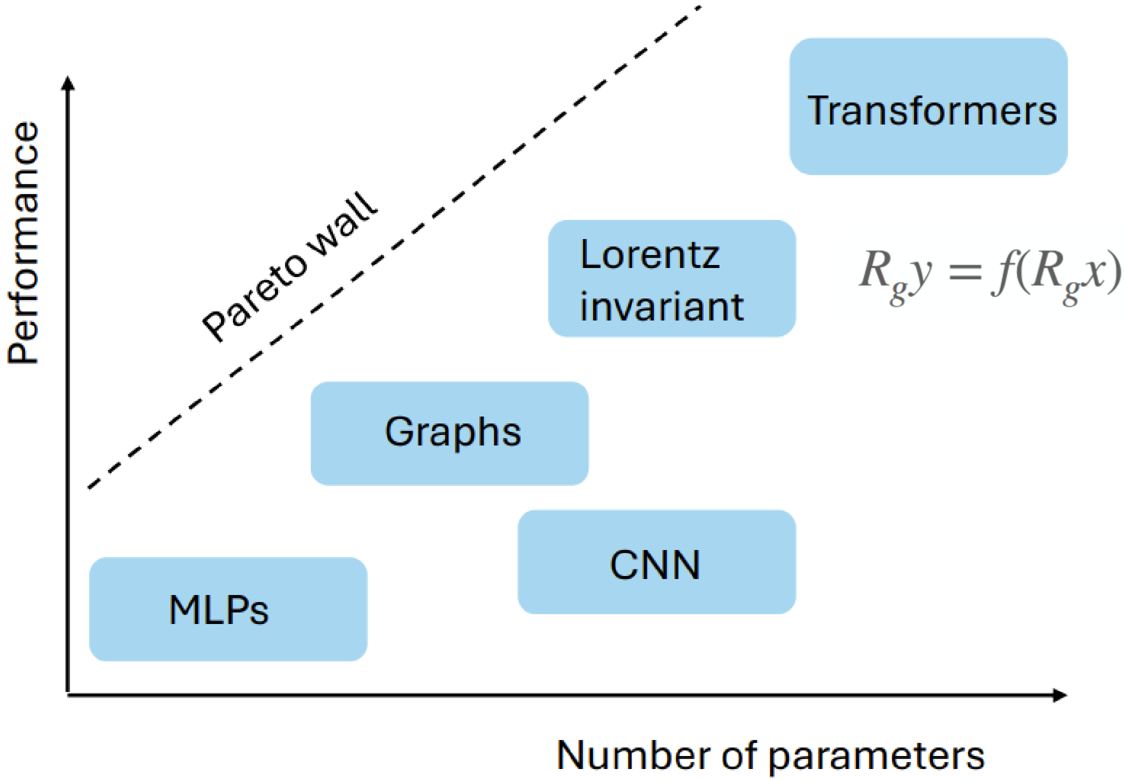


Calibrated performance



On Set-to-vector and symmetries

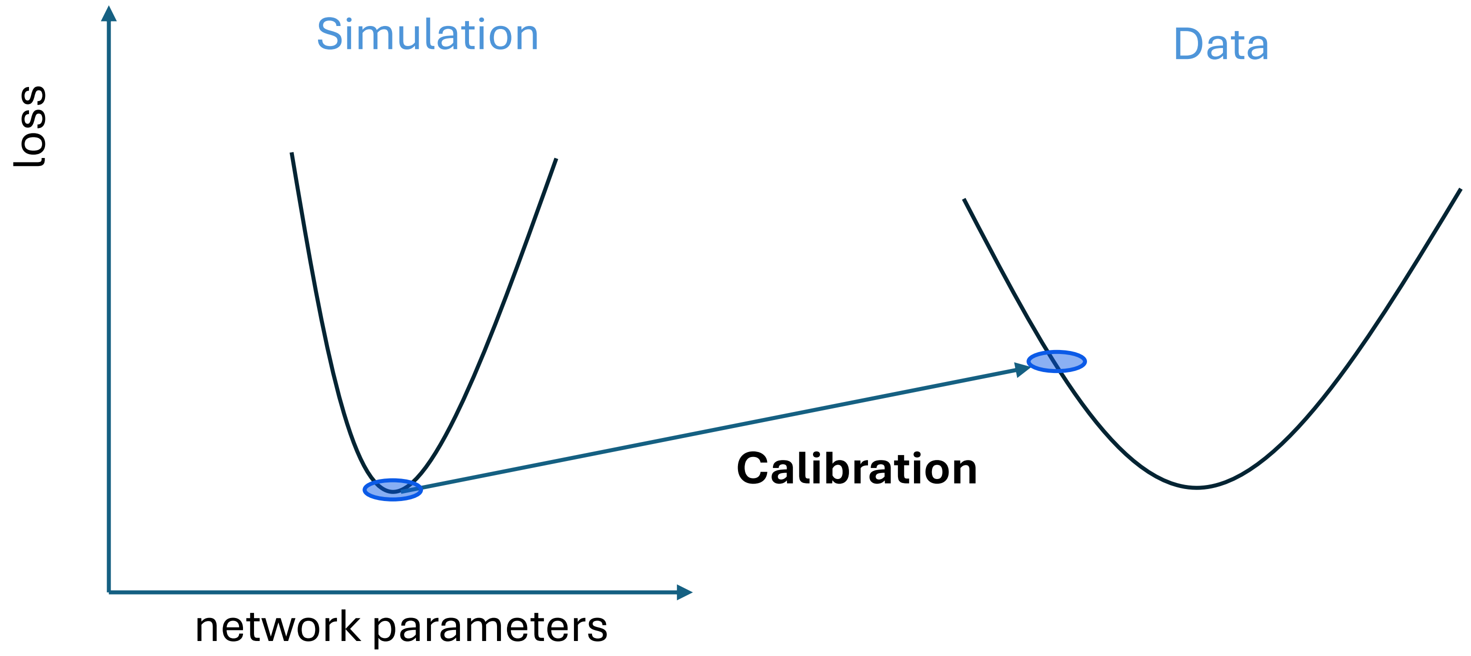
- From experience, nowadays set-to-vector is mostly just some variation of pure transformers.
- Embeding symmetries into the model (e.g. Lorentz) still needs some work



Domain shift: calibrating ML algorithms

A question rises naturally: if we train on simulation and correct them to look like the data, are we at the optimal performance in the data?

Clearly not in general...

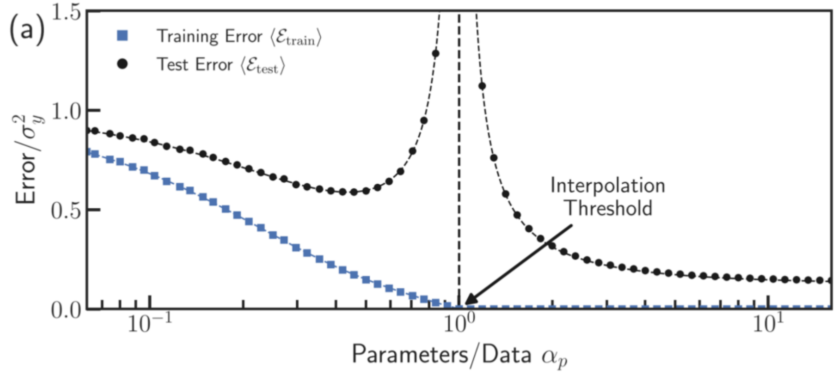


This is an active area of developements, still no clear strategy available.

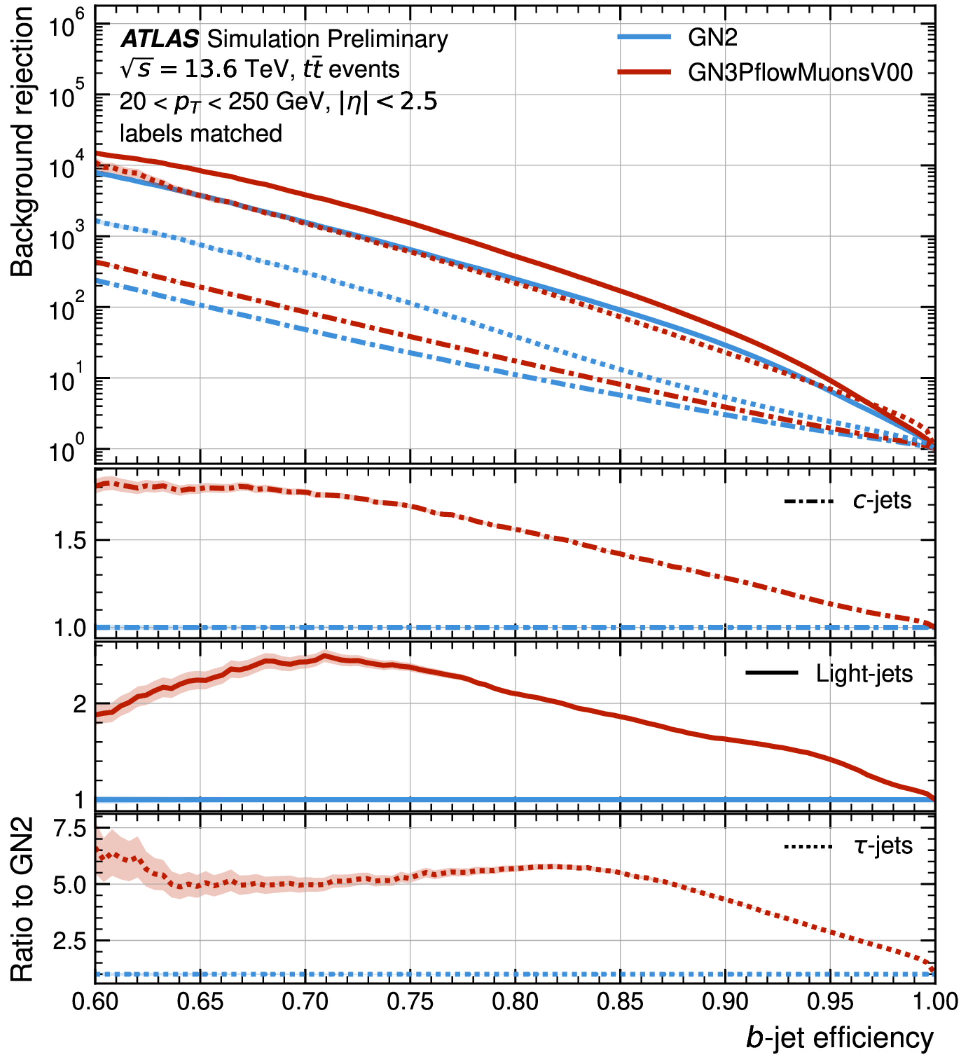
Jet identification – scaling up the models

- We are not yet saturating yet…
- Main improvements from:
  - Larger architecture and data sample
  - Additional inputs (calo+soft leptons)

Where is the limit? Where is double-descent? Generalization?



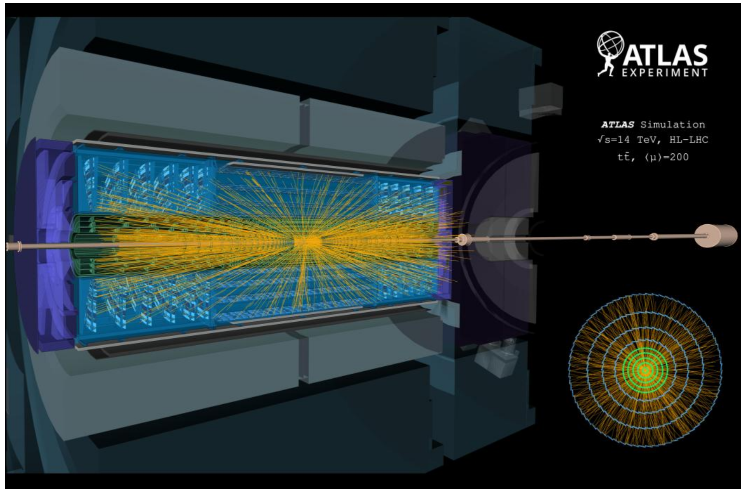
[2509.01397] L. Henirich et al.



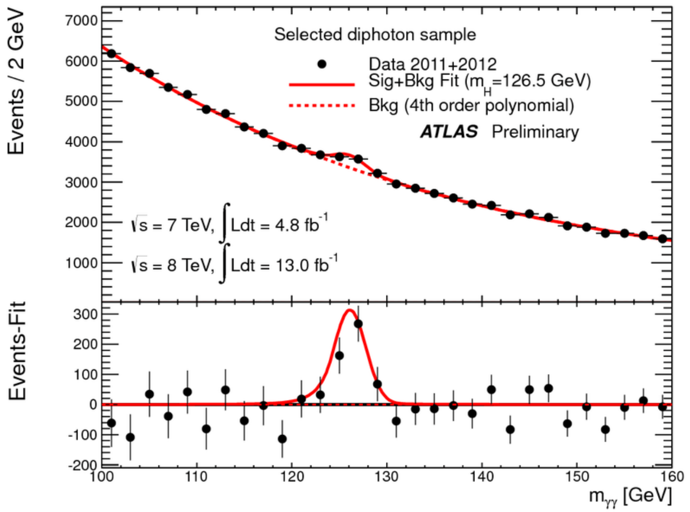
GN3, ATLAS plots

Dimensionality reduction

- Machine learning helps to solve problems with high dimensionality
- Can’t do science in high dimensions, we must find good low-dimensional representations of the data



Pattern reco, tracking, clustering, reconstructions



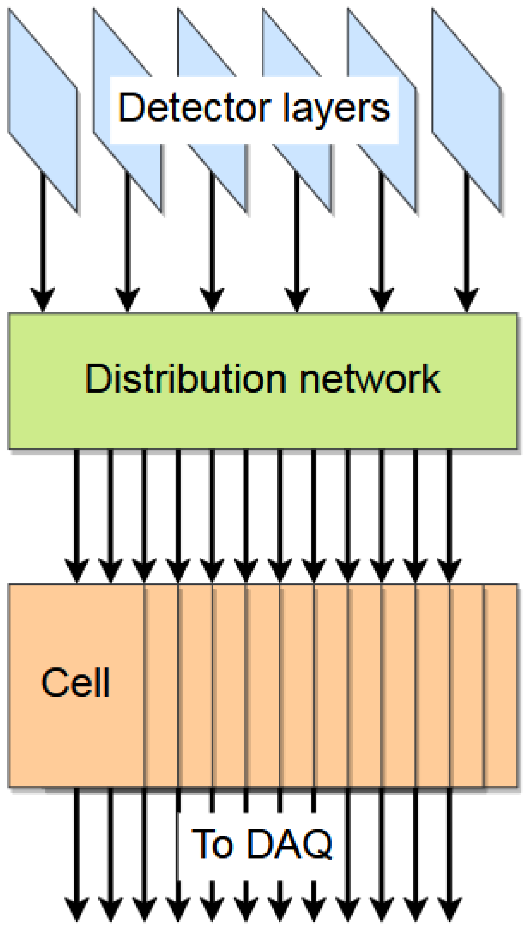
O(100M)

O(10)

Ideally loosing as little information as possible

Retina

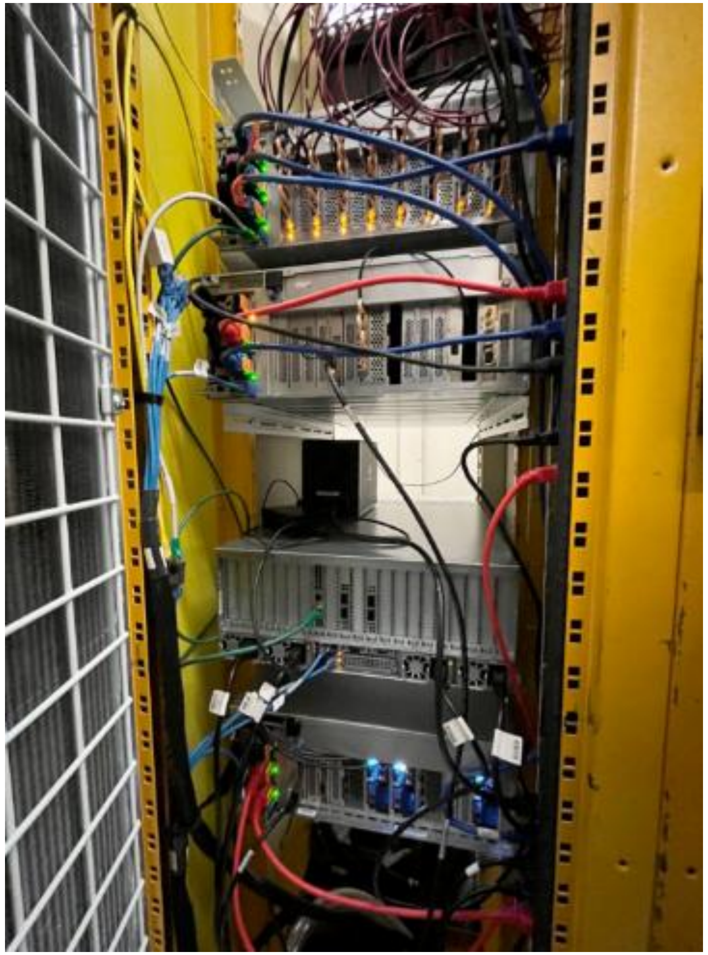
A complete Retina demonstrator was installed and tested at the LHCb TestBed facility. Succesfully!



Detector input

Distribution network. Custom switch to route signal and optimal communication to exchange across boards

Evaluates presence of primitive in specific region



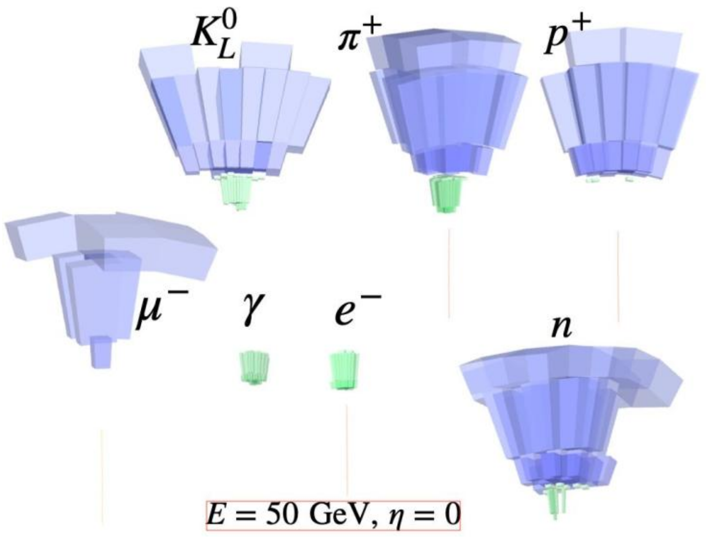
Achieved 20 MHz event rate (LHCb rate ~27 MHz). No buffering, sub-μs latency. Low power consumption 550 W

Particle-Flow reconstruction

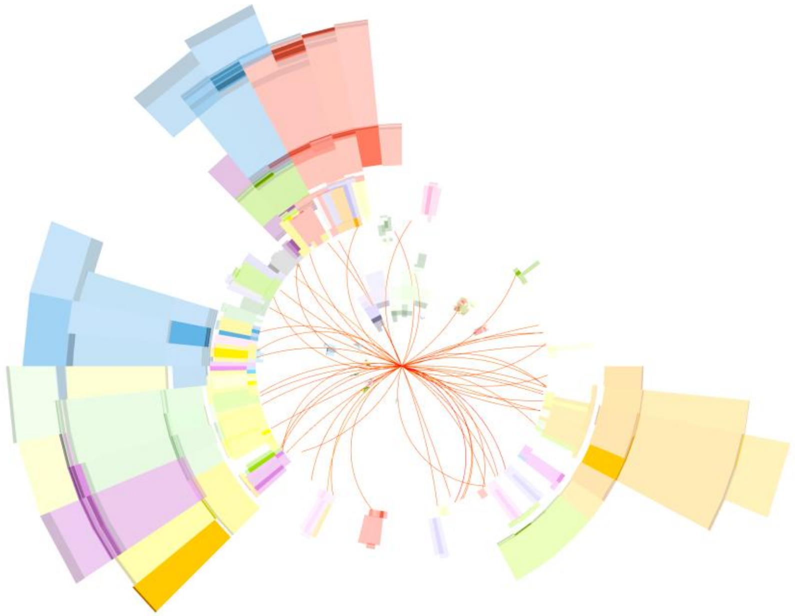
Particle flow can be phrased as follows:

Given the set of detector outputs, reconstuct the set of particles (and their properties) that generated them. This is in inverse problem.

Particle signatures



Detector output



Solving the inverse problem