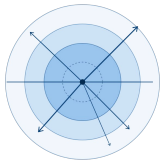
Overview of the Application of Generative Models to Fast Simulation in Large HEP Experiments

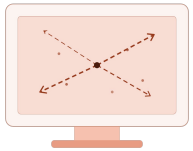
Michał Mazurek

ML4Jets 2026

14-18 September 2026 • Vienna, Austria

The big role of simulations





Experimental workflow

- Source of particles
- Experimental apparatus
- Readout system

Simulation workflow

- Event generation
- Particle transport
- Detector response

- Reconstruction
- Offline data analysis

Monte Carlo simulations are key in HEP experiments!

- new particles and interactions,
- experimental variables vs. theoretical parameters,
- design of new detectors,
- estimation of background, efficiencies, etc.

Where are we today?

Experiment upgrades bring more complexity

- **LHCb and ALICE** had their **first upgrade** for Run 3
  - LHCb: **full software trigger**, x5 higher lumi.
  - ALICE: new TPC, interaction rate x100
- **ATLAS and CMS** preparing for **HL-LHC** (Run 4)
- **LHCb and ALICE** planning **2nd upgrades** for Run 5

Detectors are getting more complex

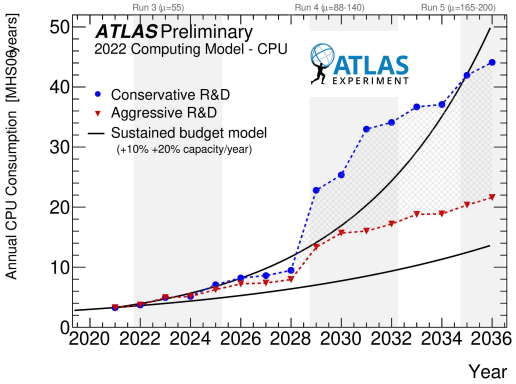
- High Granularity calorimeter in CMS
- Zero Degree Calorimeter in ALICE
- LHCb PicoCal for Upgrade II

Image courtesy: Simone Campana

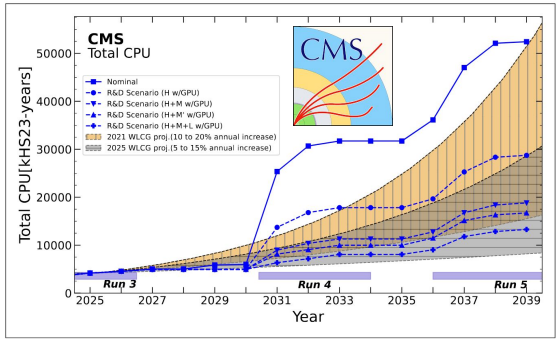
More complex events collected = more simulated samples are needed!

Universal computing challenge

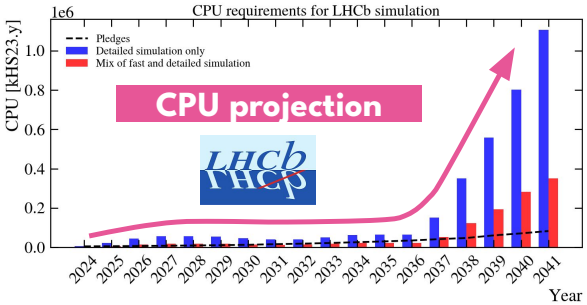
CERN-LHCC-2022-005



CERN-LHCC-2026-003



CHEP 2026

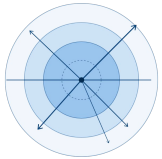


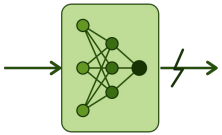
Estimated **CPU requirements** for future runs significantly **outpace computing pledges**, highlighting a **growing resource gap**.

- **ATLAS**: simulation takes **~70 %**[ref.] **of total ATLAS CPU budget**
- **CMS**: all MC simulation is **~50 %**[ref.] **of total CMS CPU budget,** with detector simulation as the primary consumer
- **LHCb**: simulation consumes **~90 %**[ref.] **of the total LHCb CPU usage**

Simulations must be faster!

What is expensive in the simulation?





Experimental workflow

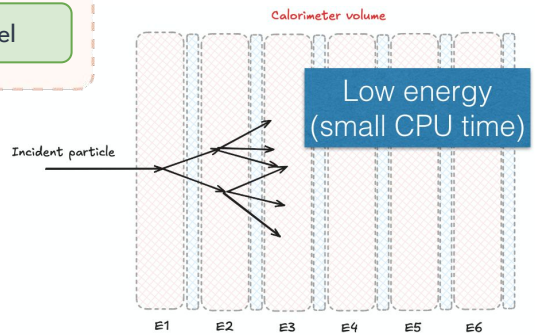
- Source of particles
- Experimental apparatus
- Readout system

Fast Simulation workflow

- Event generation + Model
- Particle transport + Model
- Detector response + Model

- Reconstruction
- Offline data analysis

- **Particle transport** remains the **main bottleneck** in MC productions, and in particular calorimeter simulations.
- In particular, **shower propagation** in the calorimeters is one of the **most time consuming tasks**.



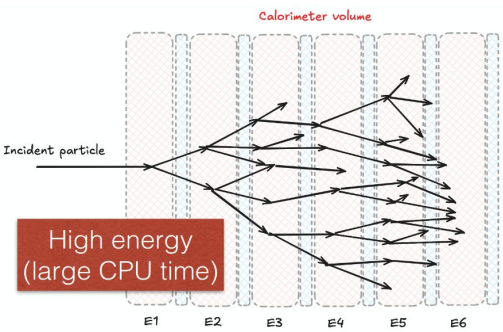
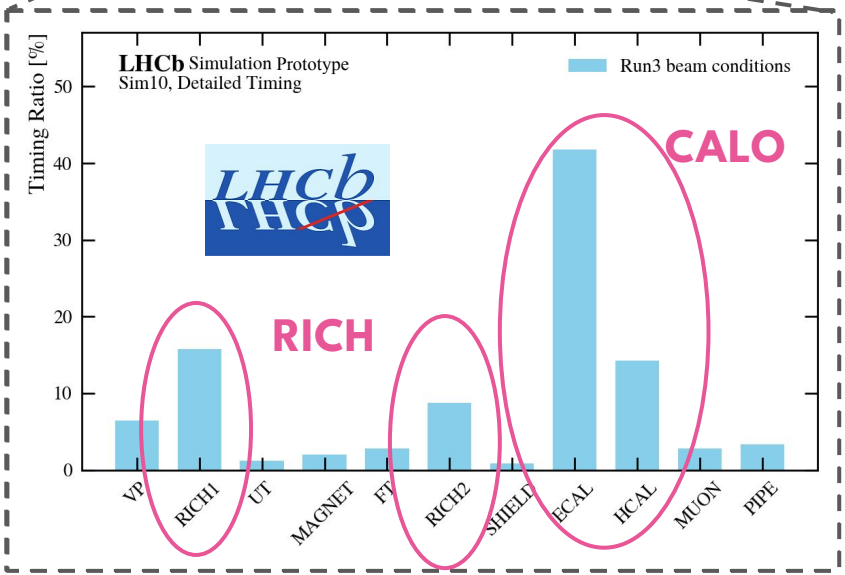


Image courtesy: Marco Valante

Calorimeter remains the main bottleneck

Most of the work done (for now) by the experiments is spent on speeding up the calorimeter simulation

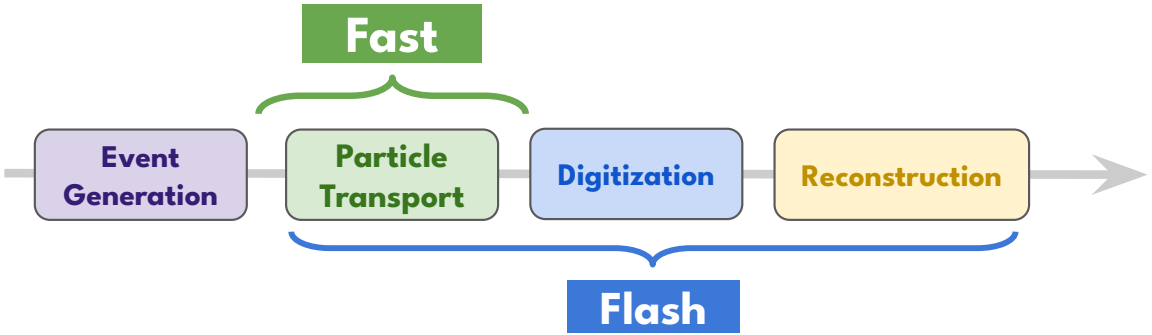
- **ATLAS:** calorimeter simulation takes **~80 %** [ref.] of the whole simulation time
- **LHCb:** calorimeter simulation takes **~50 %** [ref.] of the whole simulation time



source: CHEP 2026

Trade-off: time vs. quality

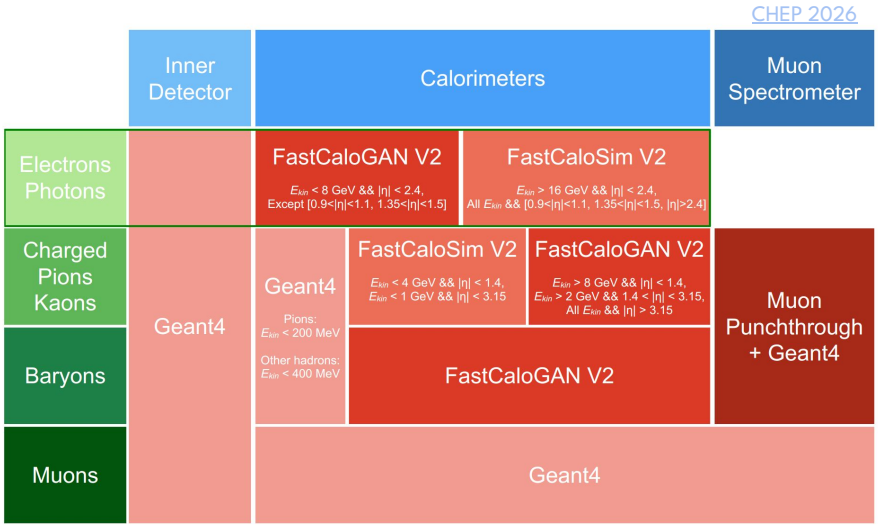
Not all of the analyses have the same requirements!
👉 multiple options needed to satisfy the needs!



In this talk

- ➥ review of selected applications of generative models to fast and flash simulations
- ➥ lessons learned from the deployment process

ATLAS: AtlFast3



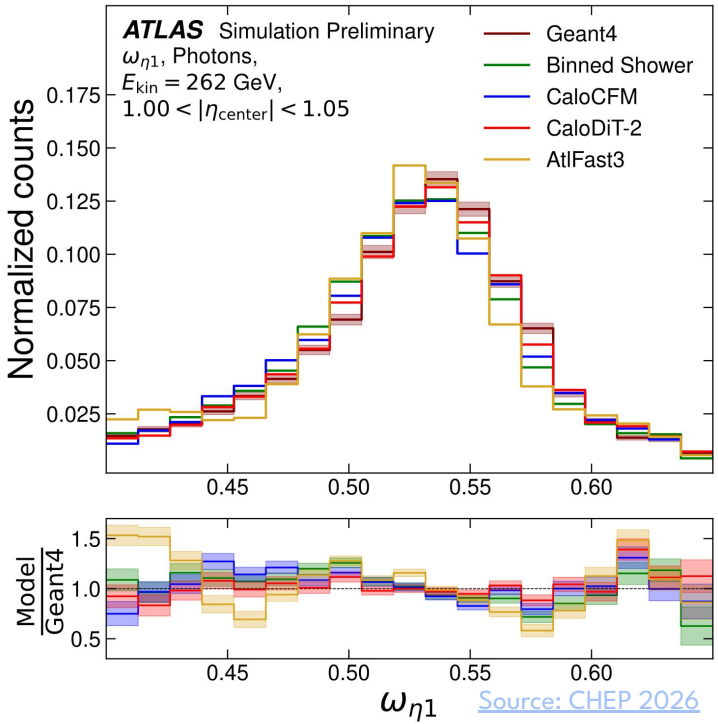
Main idea

- **Goal:** provide fast simulation option for the calorimeter
- **AtlFast3 consists of 2 components**:
  - **FastCaloSim V2** doing longitudinal and lateral parametrization of showers,
  - and **FastCaloGAN,** a ML based approach (300 independent Wasserstein GANs) where the correlated longitudinal and lateral shower fluctuations matter most

Current status

- AtlFast3 has the **same CPU cost as the older AtlFastII** but **much better fidelity**,
- **Currently used as Run3 baseline and used to resimulate the bulk of Run 2 MC**

ATLAS: calorimeter simulation beyond GANs



Source: CHEP 2026

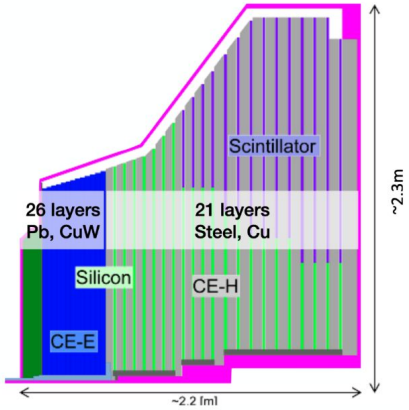
Main idea

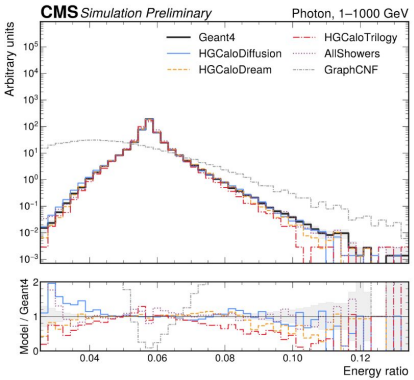
- **FastCaloGAN's** coarse voxelization and shower-shape modeling are **known to underperform for electrons and photons**.
- New two-stage pipeline per detector region:
  - a **layer-energy model, CaloINN**,
  - then a **shower-shape model** — either **CaloCFM** (conditional flow matching) or **CaloDiT-2** (a Diffusion Transformer operating on shower patches).

Current status

- **CaloCFM/CaloDiT-2 reproduce Geant4 more accurately than AtlFast3** across nearly the full phase space, at similar memory footprint
- Pions and electrons are next to check
- **Talk at ML4Jets on this topic tomorrow**

CMS: Tackling High-Granularity Calorimeter





Source: CHEP 2026

Model	Paradigm	Backbone	Representation
HGCaloDiffusion	Diffusion (EDM/DDIM)	Cyl.-conv U-Net + attention	Voxel (GLaM)
HGCaloTrilogy	MeanFlow + GMM prior	Shared with HGCaloDiffusion	Voxel (GLaM)
HGCaloDREAM	Flow matching	Vision Transformer (DiT)	Voxel (patched)
AllShowers	Flow matching	PointCountFM + CNF-Transf.	Point cloud
GraphCNF	Riemannian flow	Graph NF + Riemannian FM	Graph (per cell)

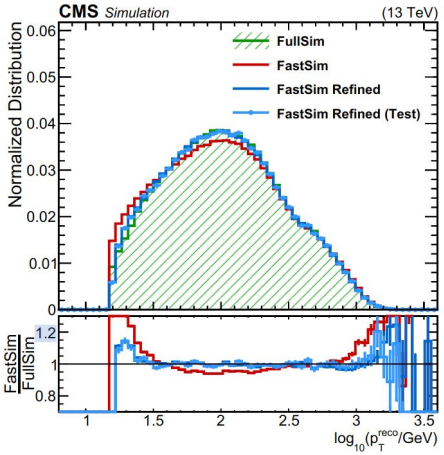
Main idea

- **CMS Upgrade** brings new calorimeter HGCal:
  - roughly **20× the complexity** of the public CaloChallenge datasets.
- **Goal:** substitute Geant4 response with advanced ML models
  - Various architectures: diffusion, flow matching & continuous NFs
  - Various representations: voxel, patched, point-cloud

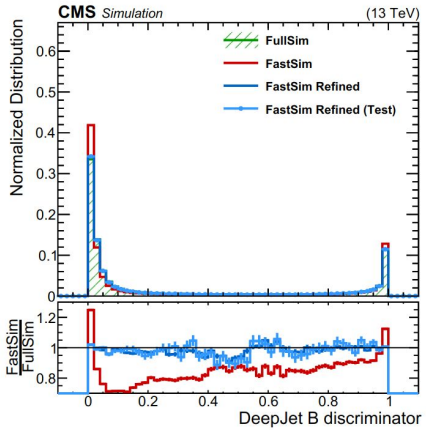
Current status

- **No single model dominates**
  - most of the showers reproduced with high fidelity
- **Further validation in progress**
- **Talk at ML4Jets on this topic tomorrow**

CMS: Refining fast simulation with ML



Refining Jets paper and CHEP 2026



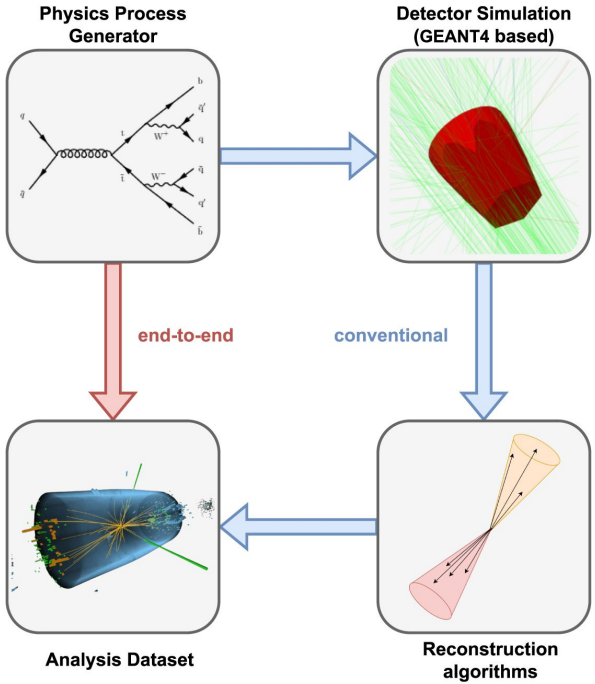
Main idea

- **Idea**: introduce a refinement method to improve the accuracy of fast simulation
- **Solution:** a post-processing ML correction applied to the fast simulation output - neural network that learns the residuals between fast and full simulation

Current status

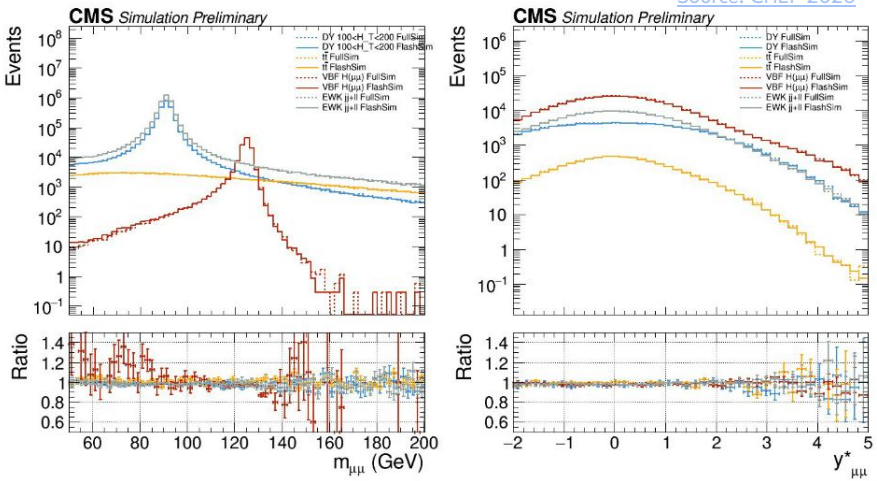
- Originally applied to jet-flavor tagging and now extends to Run 3 jet pT and its propagation to missing transverse momentum
- **Integrated with the CMS software**

CMS: Flash simulation



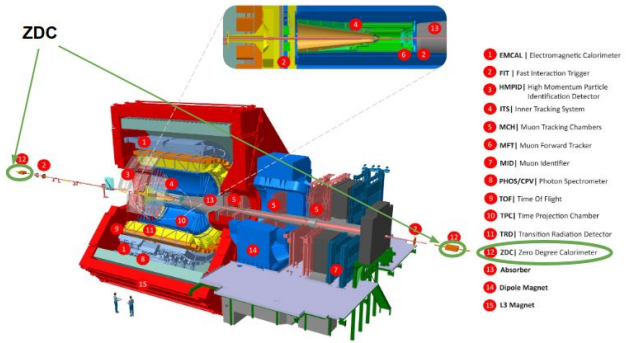
Source: CHEP 2026

Source: CHEP 2026



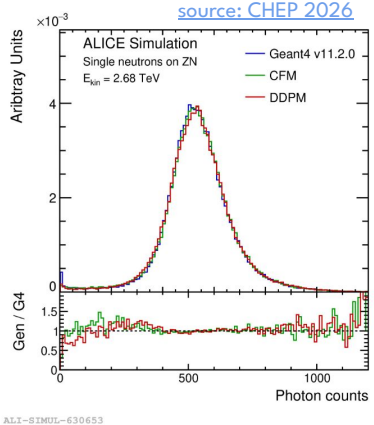
- **Goal:** skip the simulation and reconstruction steps
- Modeled by a **pair of networks**:
  a. an ***efficiency*** **model**
  b. and a ***properties*** **model** generating its kinematics/ID variables.
- **Validated on entire physics-analysis chains**
- **Speed-up:**
  a. 2–4 orders of magnitude faster than full simulation
  b. 1–3 orders of magnitude faster than fast simulation
- **Current status:** ongoing development, validation in progress

ALICE: speeding up the Zero Degree Calorimeter



	CFM	DDPM
SWD on ZN	**1.204**	1.434
SWD on ZP	1.555	**1.426**
Total SWD	1.451	**1.417**
RB for neutrons	0.85%	**0.37%**
RB for protons	**0.64%**	1.11%
Training Time	43 s/epoch	46 s/epoch
Generation time	**2.9** μs/sample	46.0 μs/sample

source: CHEP 2026



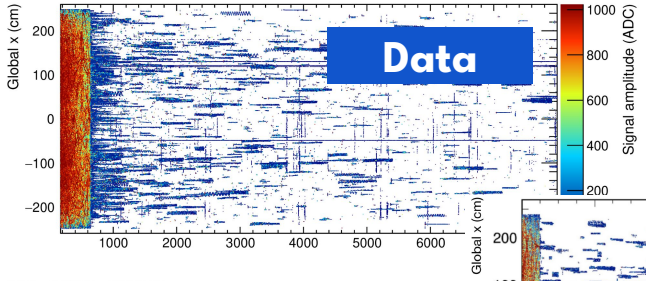
Main idea

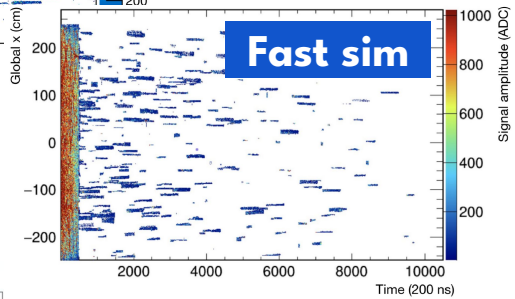
- **Zero Degree Calorimeter**:
  a. two identical stations located 112 m on both sides from the interaction point
  b. measures forward spectator nucleons for centrality determination
  c. when switched on → **doubles** the simulation time
- **Goal:** generative model to replace Geant4 transport and digitization entirely

Current status

- **Conditional Flow Matching** and a **DDPM diffusion model** were trained and **reproduce Geant4 results with a very good agreement**
- **5–7 orders-of-magnitude speedup**
- **In advanced prototype stage**

ALICE: Generative Modeling of TPC Loopers





pp 13.6 TeV collisions with Pythia8
(Monash 2013)

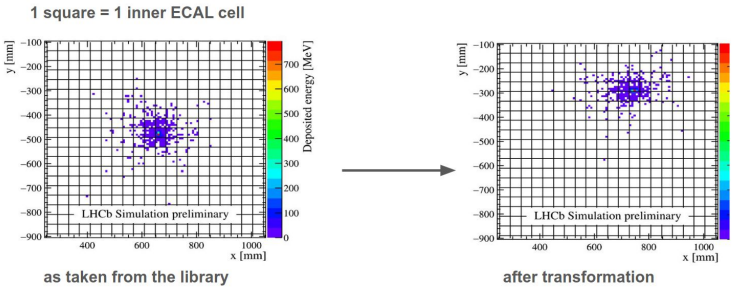
Main idea

- **TPC Loopers:** Slow neutrons captured in the TPC enclosure produce e^+e^- pairs and electrons from Compton scattering that **spiral in the gas volume** and are **costly to simulate via Geant4**
- **Solution:** WGAN is trained to bypass slow-neutron transport → only final loopers are injected in Geant4

Current status

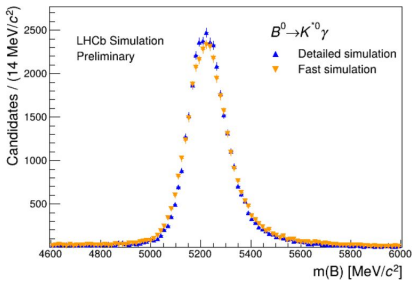
- The fast-looper generator is **fully validated and integrated** into the ALICE simulation framework
- **10x faster than full simulation**
- Faster/more-precise alternative architectures and a full hits-level (Geant4-skipping) ML approach are under study.

LHCb's Point Library



Source: CHEP 2026

Reconstructed B^0 mass



Main idea

- point library = lookup table
- replace calorimeter simulation with a **non-AI based fast simulation**
- provides a **collection of points** generated previously with Geant4 that are **transformed to recreate showers**

Current status

- simulates EM showers with up to **3 orders of magnitude speedup**
- fast simulation option **already deployed** and to be **turned on by default for all physics analyses without electrons/photons in the final state**
- developing a **new ML-based extension** to **fast simulate punch-through particles**

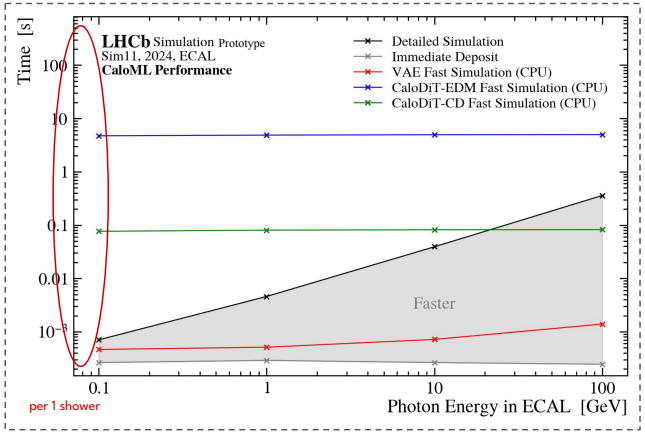
LHCb: CaloML

Main idea

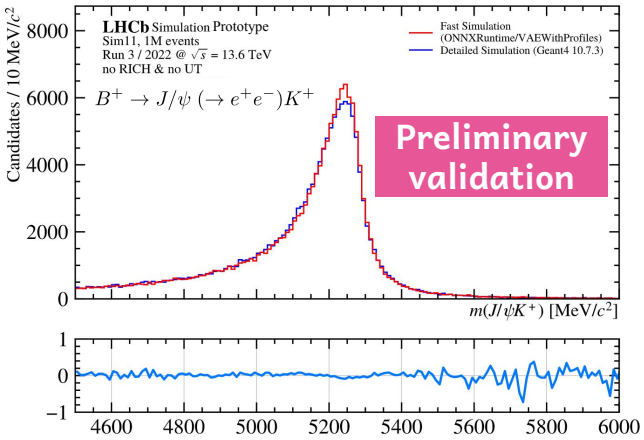
- replace calorimeter simulation with generative models
- **implementation in Gaussino** (core simulation framework) that allows to run **CaloChallenge-compatible models**

Current status

- **modified VAE up to 3 orders of magnitude faster** for EM showers → extra modifications to the head were necessary to produce realistic showers
- preliminary **physics results show very good agreement** with reconstructed observables
- **new architectures tested: CaloDiT**, but performance is an issue → more work to be done to speed it up
- early tests for **Upgrade 2 calorimeter** are in progress



source: CHEP 2026



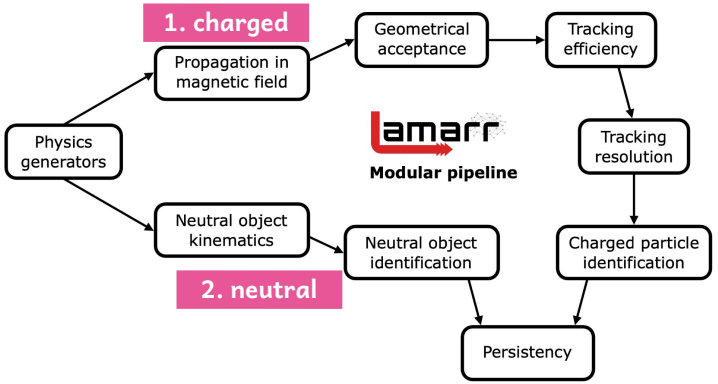
LHCb's flash simulation: Lamarr

Main idea

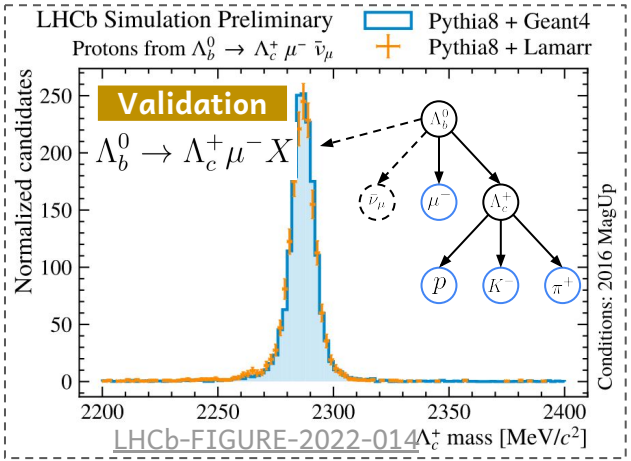
- introduce a **pipeline of ML-based modular parametrizations** to replace both the **simulation *and* reconstruction** steps
- 2 branches:
  - a branch treating **charged particles** to provide tracking and particle identification parameterizations
  - a branch treating **neutral particles** that require an accurate parameterization of the calorimeter

Current status

- **Speed-up:** allows to significantly reduce the CPU cost for the simulation phase by at least **two orders of magnitude**
- **Physics results show very good agreement for the charged particle branch**
- Current focus on moving Lamarr to production

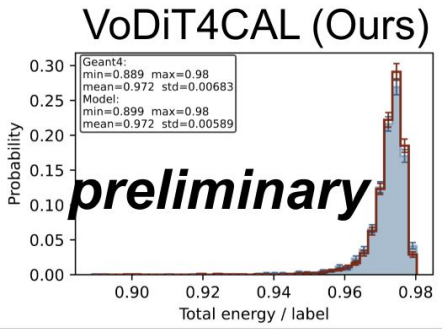


source: CHEP 2026

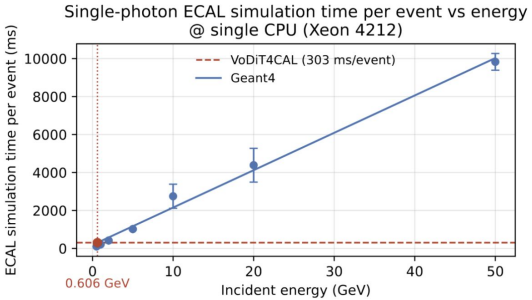


LHCb-FIGURE-2022-014

Beyond LHC



source: CHEP 2026



Need for faster simulations isn't LHC-specific future colliders and lower-energy experiments face similar challenges

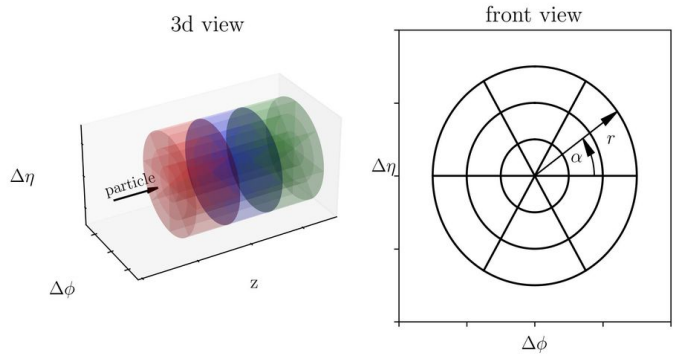
Selected work

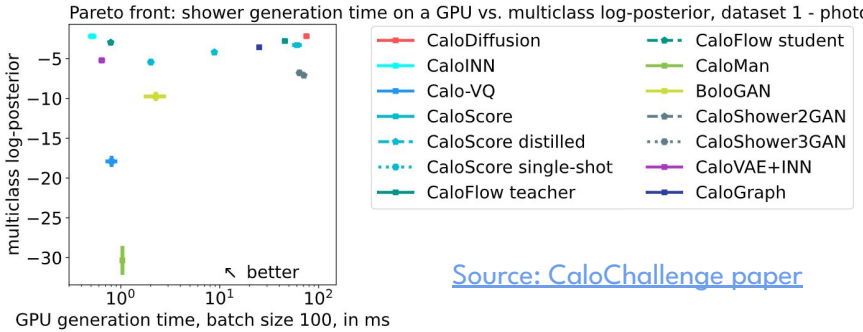
- **CEPC**[ref.]: VoDiT4CAL (extension of CaloDiT-2) for the CEPC crystal ECAL -> ~100× speedup for 50 GeV events
- **BESIII**[ref.]: GAN & diffusion models for EM showers
- **EIC**[ref.]: normalizing flows / flow matching / diffusion suite replacing Geant4 photon transport for Cherenkov PID

Lessons learned

- **Fast vs. flash:** it is still easier to deploy generative components **only where they are needed the most** rather than replacing the whole chain with one network, although flash simulations are getting more widely adopted in the experiments
- **Difficulty of fast simulation operations:** it is easy to show that generative models work on toy datasets, but scaling up to the complexity of real detectors remains a challenge
- **Physics validation is essential:** fast simulation methods must be validated on the reconstructed observables actually used in physics analyses before a large experimental collaboration will adopt them
- **State-of-the-art models are more accurate**: models based on flow matching, diffusion, and normalizing flow approaches bring a noticeable impact on accuracy; however, they are slower than VAEs/GANs
- **Geometry and physics drive the choice:** models perform differently depending on the experimental setup and what kind of physics they aim to replace

CaloChallenge





Source: CaloChallenge paper

Main idea

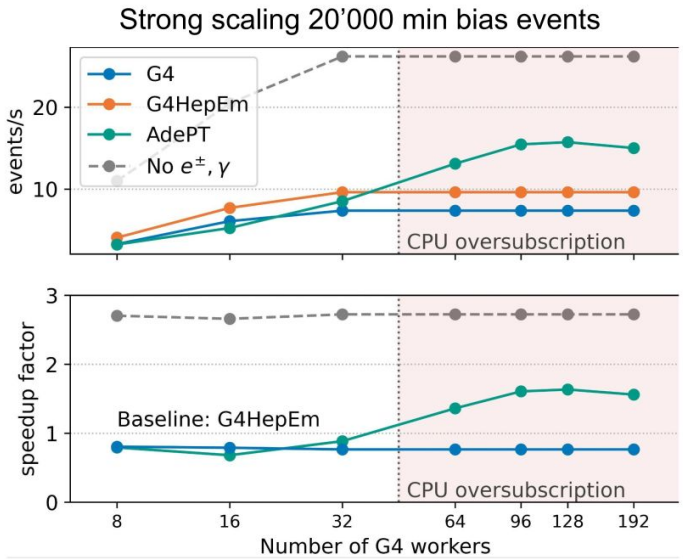
- **Problem:** performance of fast-sim models depends on the physics and geometry of each detector, making cross-experiment comparison impossible
- **Solution:** a challenge launched in 2022 with a common task: learn to sample voxel energies conditioned on incident energy on common datasets, and a common evaluation pipeline
- **Three datasets of increasing difficulty (photons/pions/electrons)**

Outcome

- **60+ participants, 31 submitted models**, covering every major architecture family: GANs, VAEs, normalizing flows, diffusion, conditional flow matching
- **A genuinely collaborative effort** — also driven by people outside the large HEP collaborations, spanning academia and industry
- The datasets now serve as a **standing benchmark for newer models** (e.g. CaloDiT-2, CaloCFM) and give experiments a **common reference** when choosing an architecture

Detailed simulation is speeding up

Adept in LHCb



Source: CHEP 2026

Main idea

- Everything so far assumed Geant4 on CPU stays fixed, but that baseline is also moving
- **AdePT** and **Celeritas** offload EM shower transport directly onto GPUs **speeding up detailed simulation**

Current status

- Production-quality GPU detector simulation now demonstrated for **ATLAS (~1.4× speedup), CMS (~1.9× speedup), and LHCb (~1.8× speedup)**

Takeaway for generative models

- **On GPUs the bar is raised for what counts as a compelling trade-off for fast simulations**

Conclusions

- Increased luminosity and complexity of events place growing demands on computing → **more accurate simulated data** is needed **at the same computational cost**
- Generative fast and flash simulations **are now in production or in advanced prototype** stage across LHC experiments
- **Similar architectures** (GANs, VAEs, normalizing flows, diffusion, and flow matching) recur across every experiment despite different detectors
- **No universal winner:** architecture choice depends on detector geometry, physics, and the speed/accuracy trade-off needed
- **Software integration and physics validation are the hard part**: moving from an early prototype to a model fully validated within an experiment's simulation framework is far harder than showing results on toy datasets, but it is clearer now how to get there