Fast and ultra-fast simulation at the LHC

Lucio Anderlini

Kick-off meeting Spoke 2
2022-10-13

Disclaimer

Activities aiming at a reduction of the computational cost of the simulation are many and diverse. This won't be a complete overview.

Priority will be given to:

- *ongoing activities within the Italian community*
- *reasons why HPC resources are precious to developments of these techniques*

Fast and ultra-fast simulation in two slides (1/2)

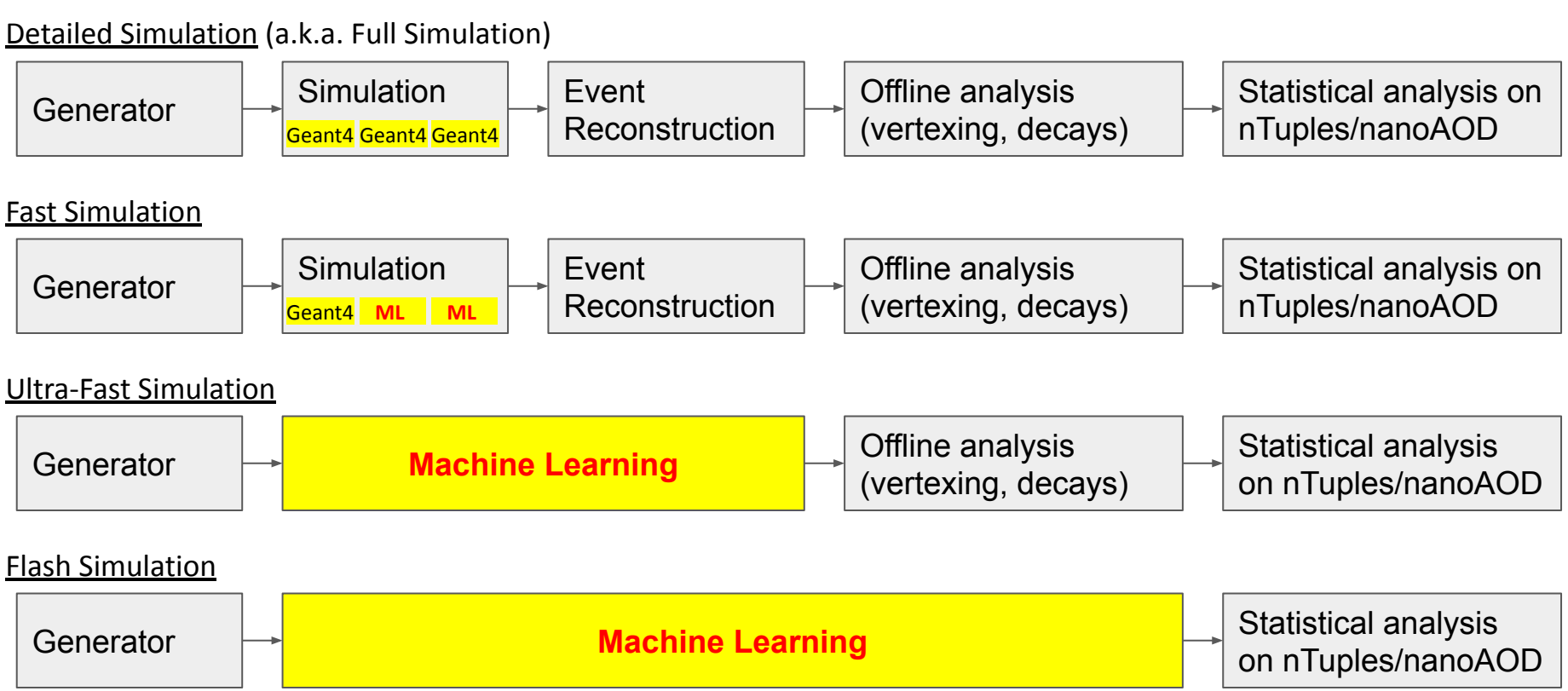
Simulating the proton-proton and ion-ion collisions and the interaction of the products with the detector is an extremely expensive process

→ it is responsible for most of the CPU requests by the LHC experiments.

Detailed simulation of the radiation-matter interactions in the detector is unavoidable for designing, commission, calibrate… an experiment, but **it may be an overkill for analyses**.

Parametrizing the detector response instead of simulating it independently for each collision event may **save up to O(MCHF)** per year.

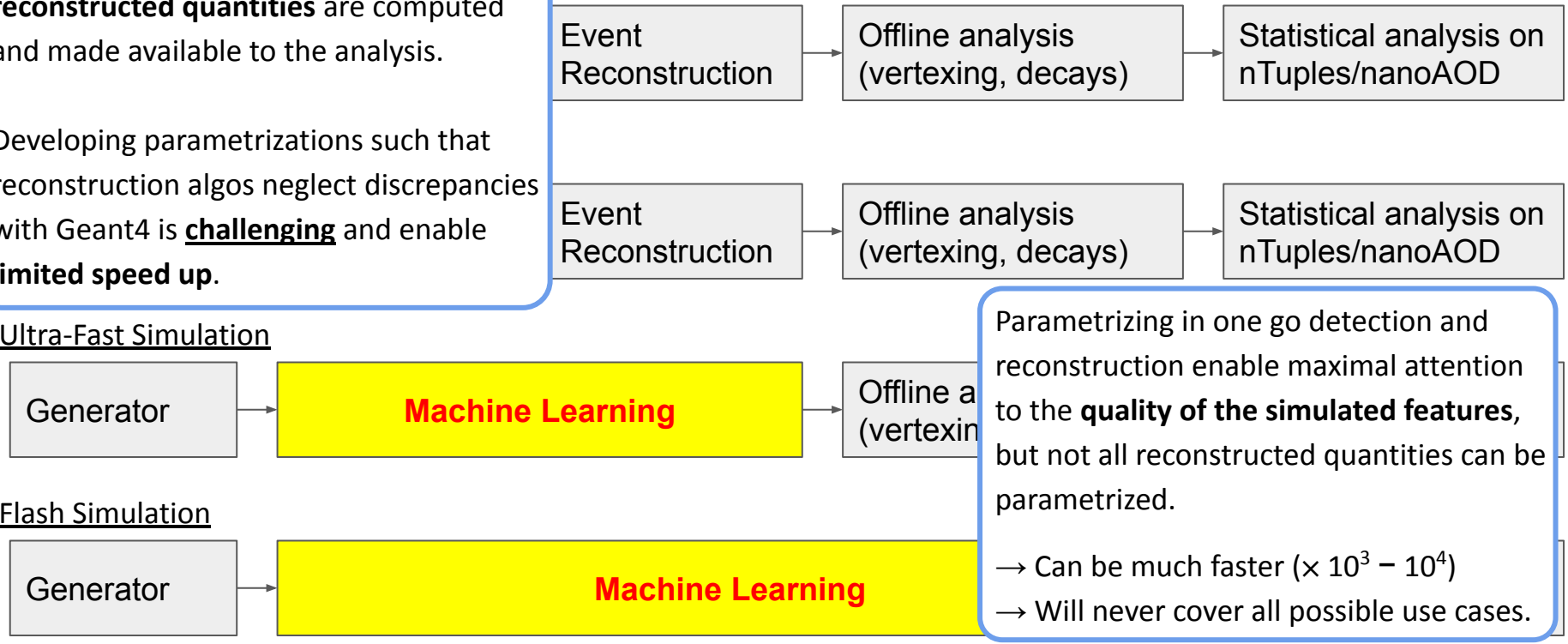
Fast and ultra-fast simulation in two slides (2/2)



Fast and ultra-fast simulation in two slides (2/2)

Sharing the reconstruction step with detailed simulation means **ALL reconstructed quantities** are computed and made available to the analysis.

Developing parametrizations such that reconstruction algos neglect discrepancies with Geant4 is **challenging** and enable **limited speed up**.



Parametrizing in one go detection and reconstruction enable maximal attention to the **quality of the simulated features**, but not all reconstructed quantities can be parametrized.

→ Can be much faster ($\times 10^3 - 10^4$)

→ Will never cover all possible use cases.

An example of Fast Simulation: AtlFast3

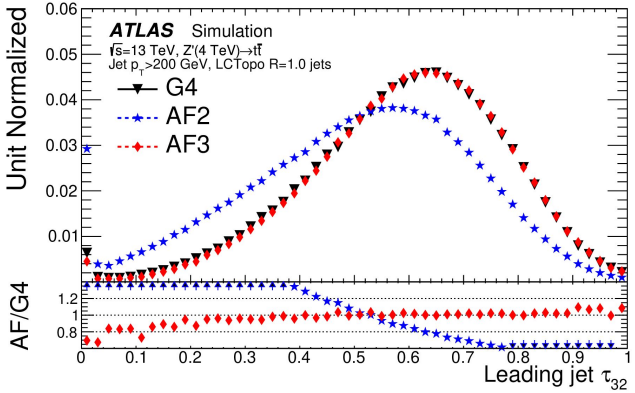
AtlFast3 combines a set of parametrizations to simulate particle showers to a level of precision with no sizable difference from Geant4.

Michele Faucci Giannelli, INFN Roma 2, ATLAS

FastCaloGAN is key component using ML to simulate calorimeters.

AtlFast3 combines Geant4, "standard" **parametrizations** and **neural networks** to predict shape and energy of the showers.

Parametrizations developed on shower shape. Model selection based on physics performance.



An example of Fast Simulation: AtlFast3

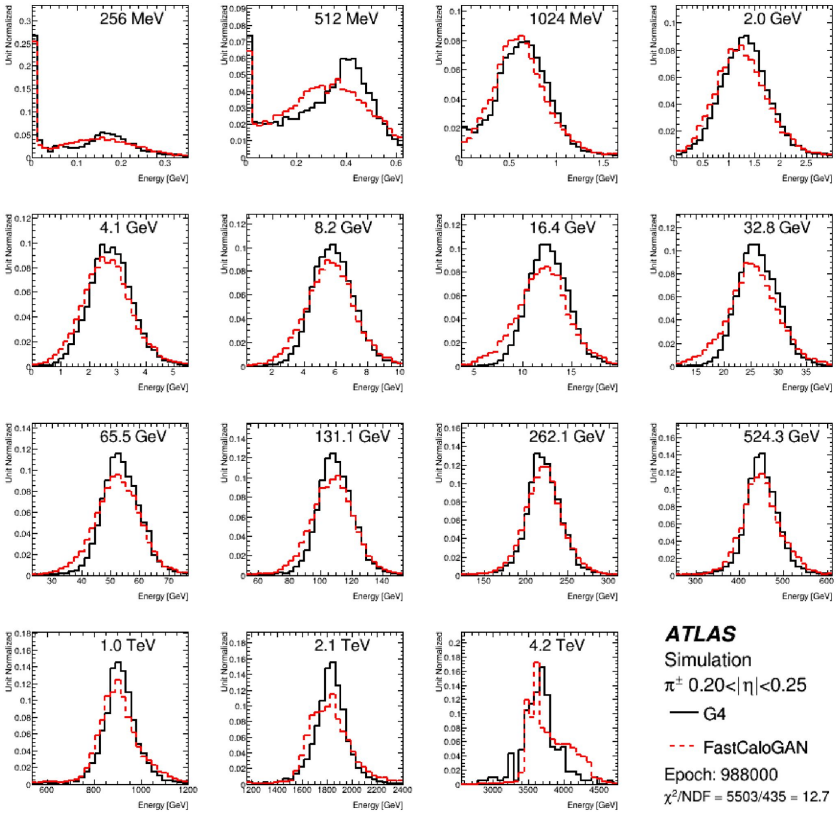
AtlFast3 combines a set of parametrizations to simulate particle showers to a level of precision with no sizable difference from Geant4.

Michele Faucci Giannelli, INFN Roma 2, ATLAS

FastCaloGAN is key component using ML to simulate calorimeters.

AtlFast3 combines Geant4, "standard" **parametrizations** and **neural networks** to predict shape and energy of the showers.

Parametrizations developed on shower shape. Model selection based on physics performance.

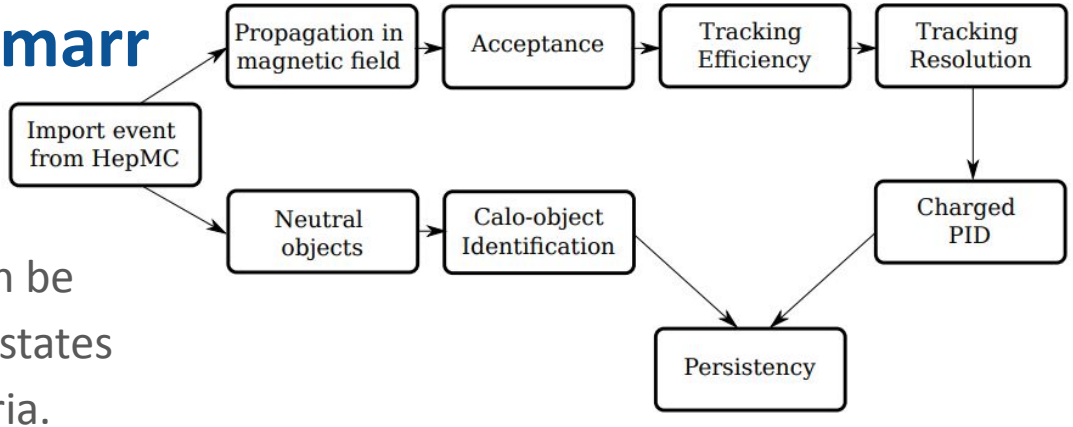


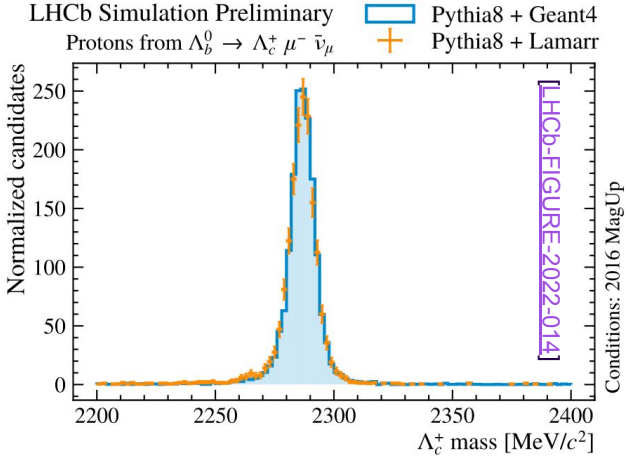
Follow and contribute to developments: **#calochallenge**

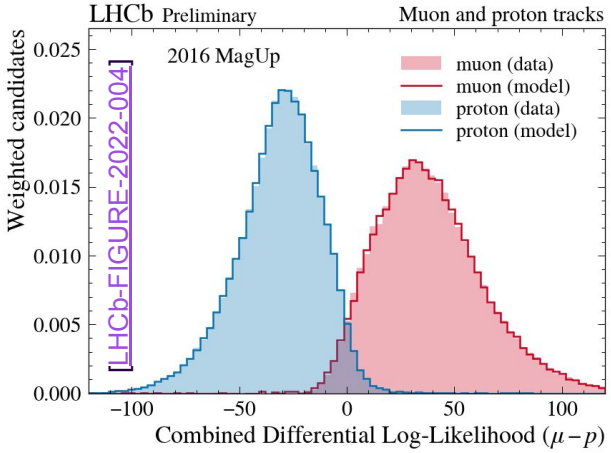
LA, M. Barbetti, INFN Firenze, LHCb

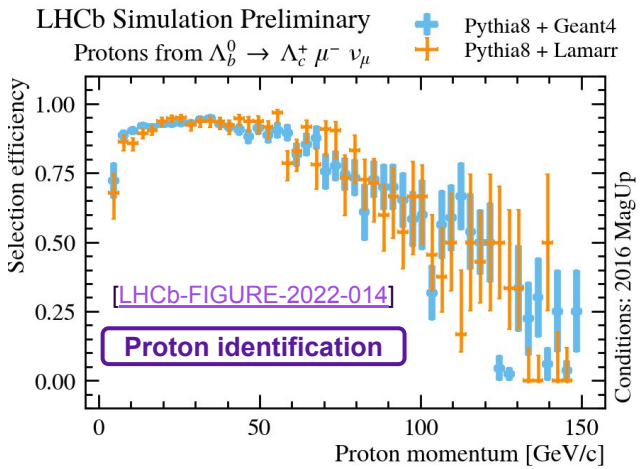
Ultra-fast simulation: Lamarr

LHCb aims at reproducing tracks and calorimeter objects in a format that can be used to combine particles into heavier states with user-defined algorithms and criteria.







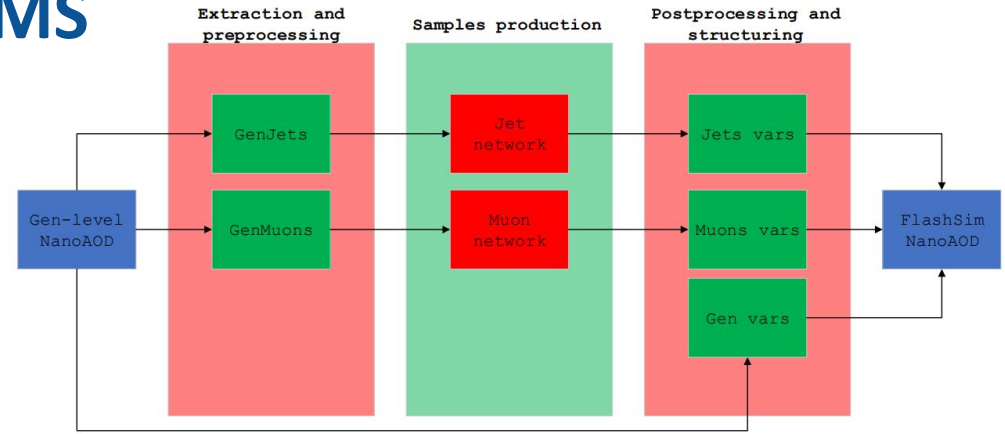


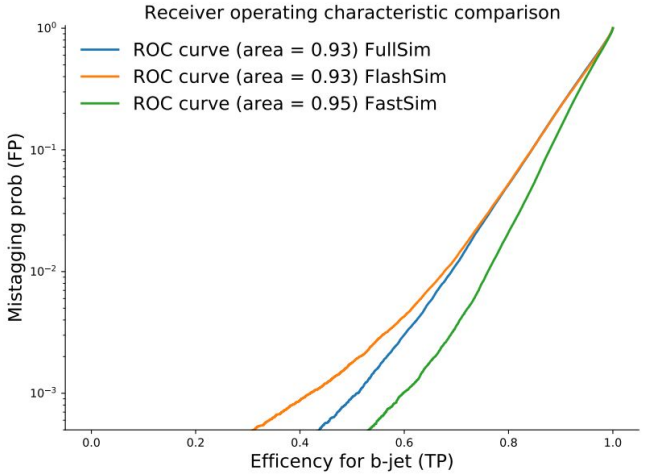
F. Vaselli, A. Rizzi INFN/Uni Pisa, CMS

Flash Simulation: Jets at CMS

CMS has recently introduced the **nanoAOD** data format retaining higher-level information sufficient for O(50%) analysis effort.

Such a **well defined target** opens to the development of FastSim aiming at reproducing the information in the AODs to be indistinguishable from *detailed simulation.*



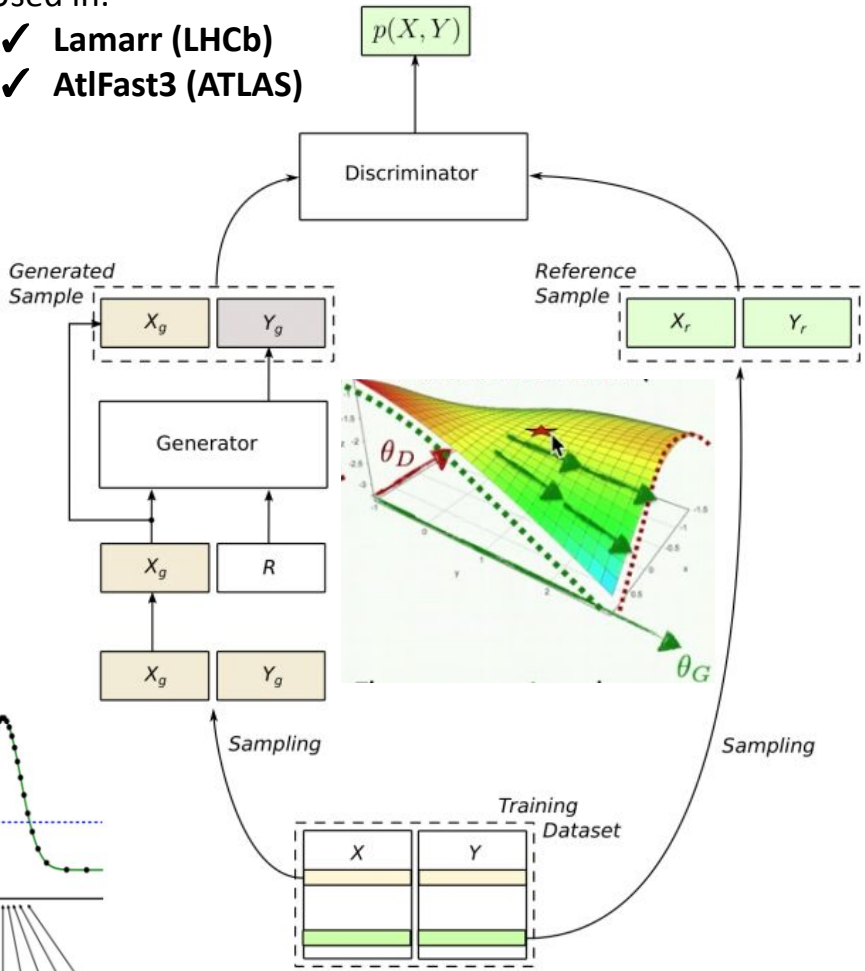


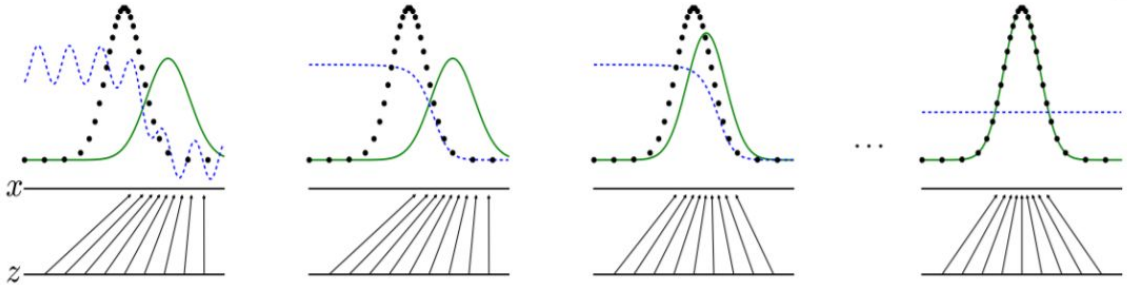
Generative models 1/2: GANs

Used in:
- ✓ **Lamarr (LHCb)**
- ✓ **AtlFast3 (ATLAS)**

GANs represented a major revolution for the quality of generative models, providing a fast algorithms generating random samples with very complicated distributions learnt from data.

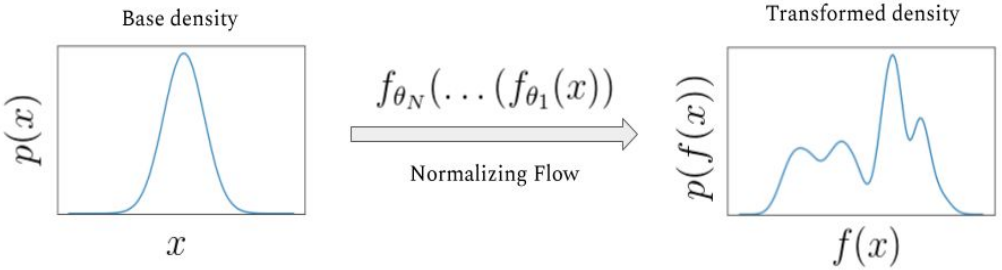
Being based on an equilibrium on a saddle point, they are **rather difficult to train** and the value of the loss is rather meaningless.





Generative models 2/2: NFs and GMMs

For a small (< 100) number of dimensions, an **explicit model of the underlying *pdf*** function may provide better quality.

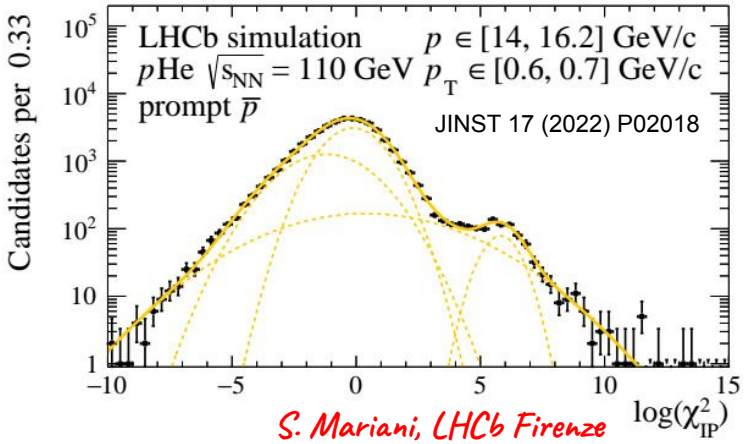


Normalizing Flows describe invertible transformations of a multinormal distribution into a generic *pdf* by controlling the Jacobian of the transform.

Proposed for:
✓ **CMS FlashSim**

(Gaussian) Mixture Models provide a quasi-parametric description of the pdf, possibly describing the dependence of the parameters on conditions with DNNs.

Used for:
✓ **LHCb ParticleID**

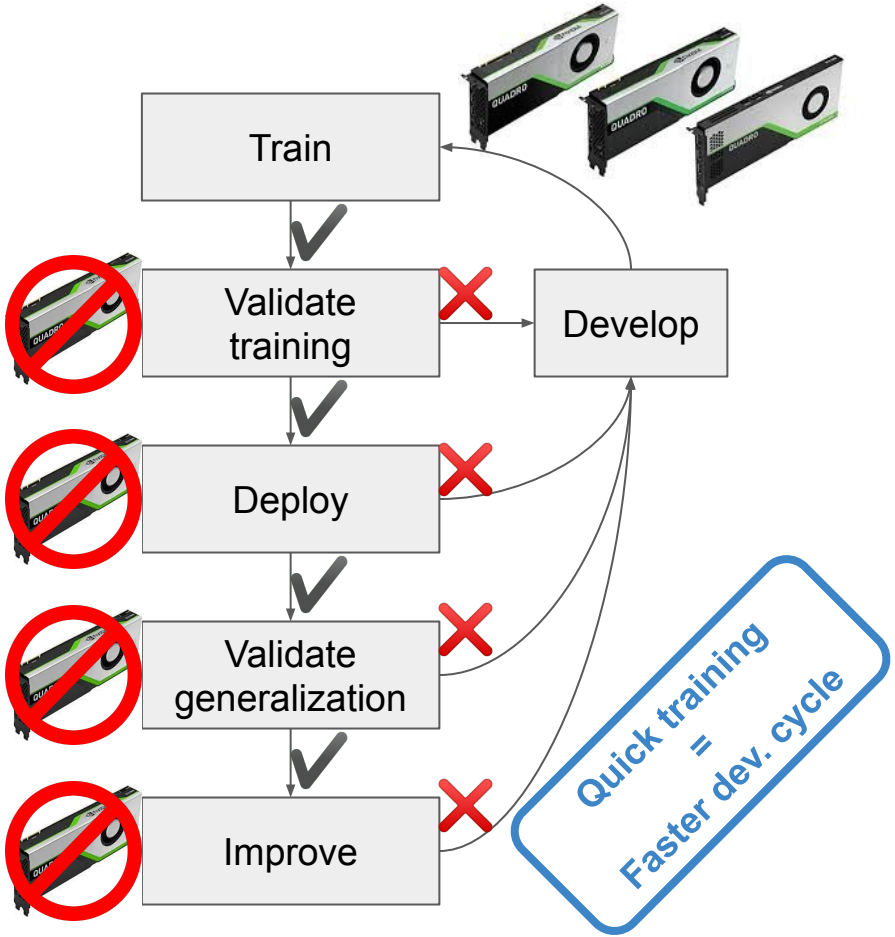


Why GPUs? And how.

The datasets used to train Generative Models are usually not very large and often fit the memory of a GPU or require splitting in few batches.

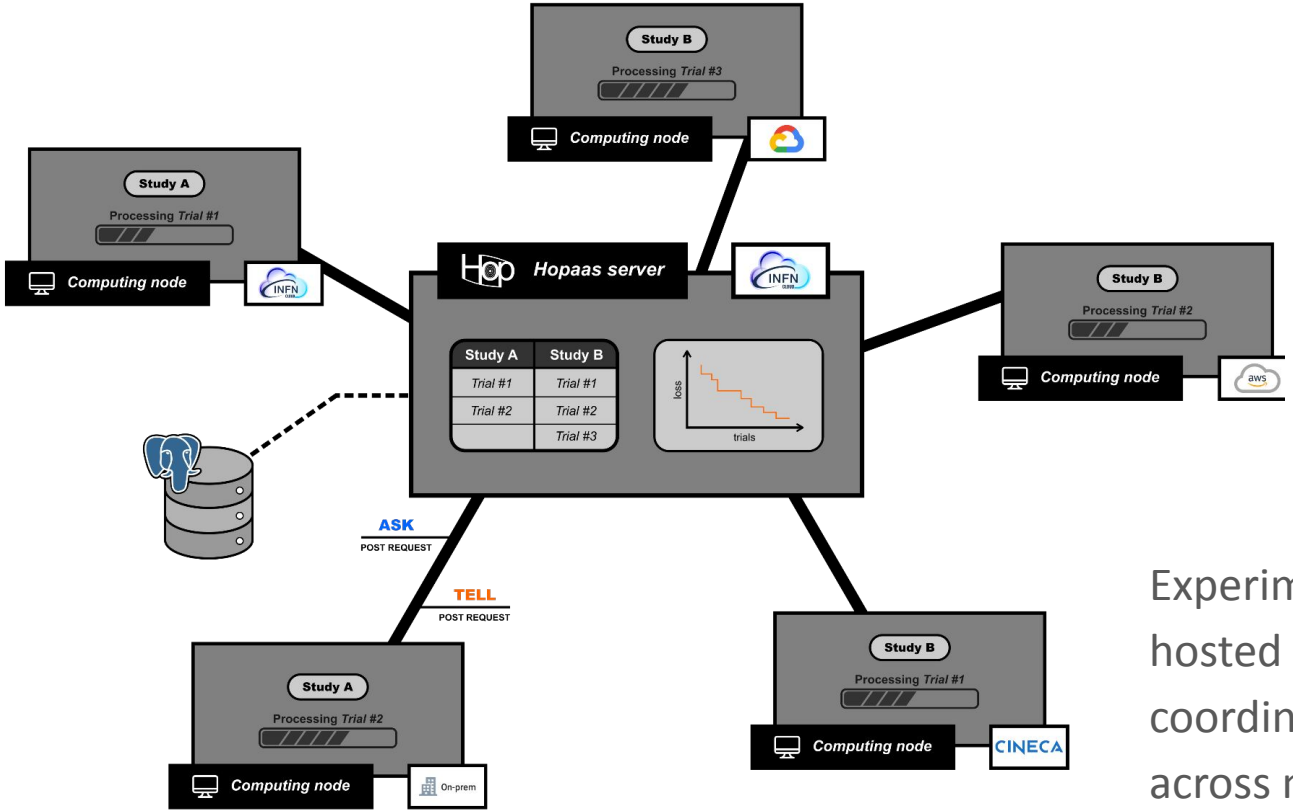
In contrast, a large number of optimization steps is required to "**fit**" (with implicit or explicit models) such large dimensionality space.

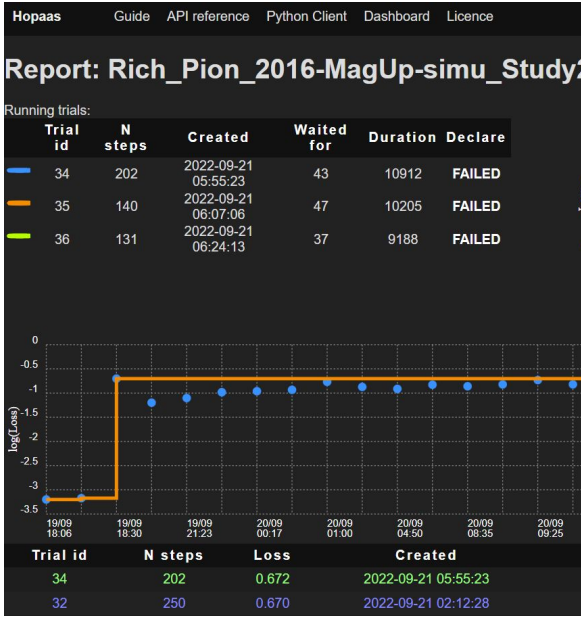
Hyper-parameter optimization is crucial and makes ease parallelization:
many GPUs needed for a limited time:
→ **Perfect for sharing resources**



M. Barbetti, INFN Firenze, LHCb

An exercise on M100 resources: *hopaas*





Experimental in-house service hosted on INFN Cloud to coordinate optimization studies across multiple sites via REST APIs.

Scaling to production

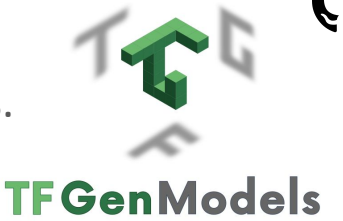
Development of fast simulation in Italy is basically **ready to scale to production**.

Needs for storage resources are negligible (few TB).
Massive GPU resources available for a limited amount of time would shorten the development cycle.

Test-beds on **CERN HTCondor** batch system, **CINECA M100**, **INFN Cloud** and local resources were already performed by Italian members of ATLAS, CMS and LHCb.

Limited access to GPUs is one of the factors limiting developments of ML-based simulation.

Better inter-experiment networking and resource sharing may also greatly contribute to improve future simulations.
ICSC is an opportunity to enhance it at national level.



mbarbetti/tf-gen-models

Early-state effort to abstract Generative Models developed in HEP from specific tasks experiment

Summary and conclusion

Generating simulated events costs O(10^6 €) per year, future upgrades will produce more data, requiring larger simulated samples.

Generative Models may provide much cheaper simulated samples.
The main challenge is to **push quality and completeness of the ML-based simulation.**

Several very interesting approaches are being studied. Also within the Italian community.

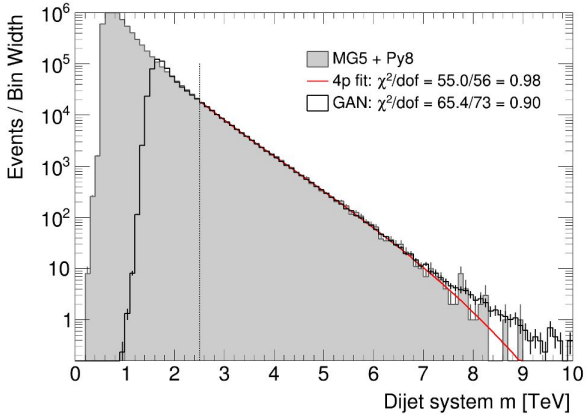
ICSC is a great opportunity for providing resources and networking.

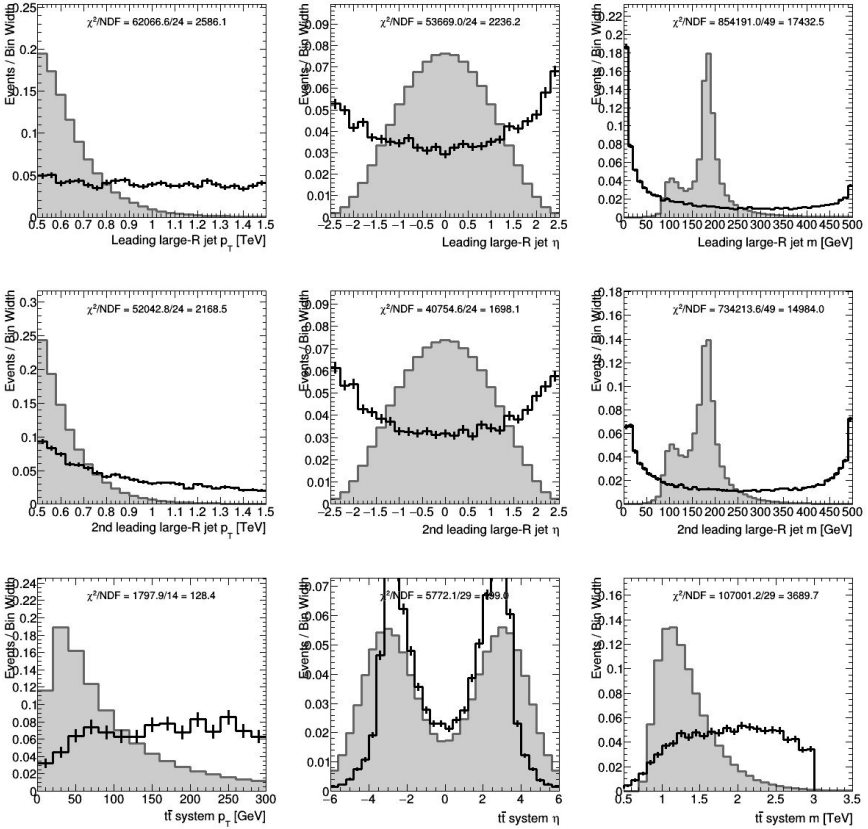
Backup

Michele Faucci Giannelli, INFN Roma 2, ATLAS

Flash Simulation: Jets at ATLAS

Combining **physics generator** of DiJet events, **detection** and **reconstruction** effects in a unique Generative Model allowed ATLAS to drastically simplify the problem arising from conditioning.





Public repository disipio/DiJetGAN