AI application dans HEP

David Rousseau, IJCLab-Orsay

david.rousseau@ijclab.in2p3.fr

AISSAI OnBoarding school

Octobre 2023

Outline

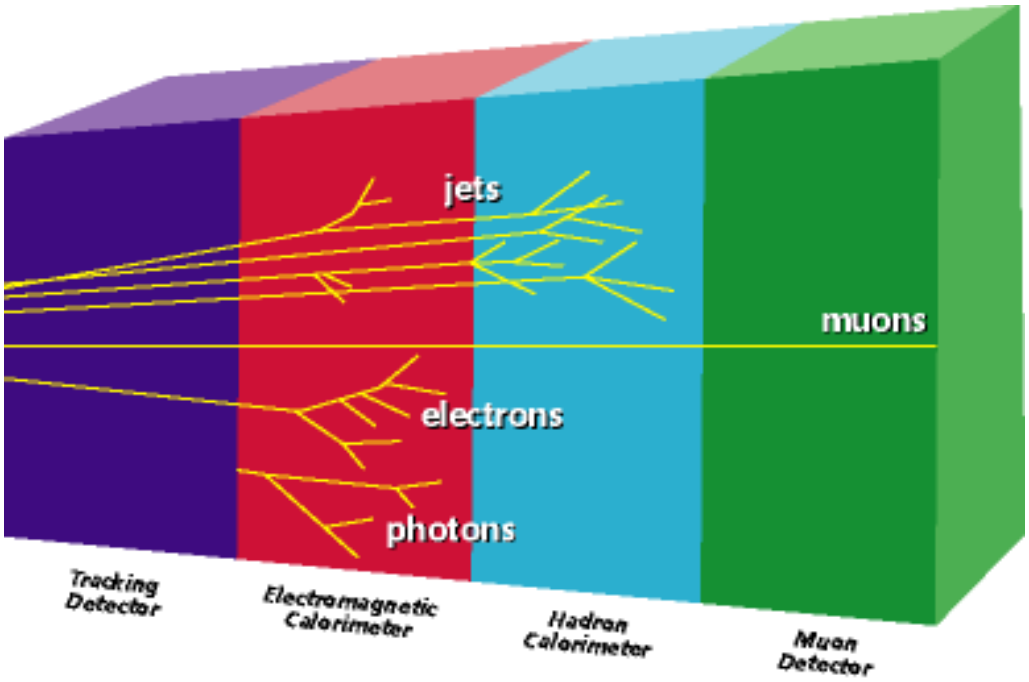
- Particle-level, event-level and experiment-level inference (e.g. ATLAS or CMS experiment on the Large Hadron Collider at CERN)
- High Energy Physics data are not images
- GAN/VAE for simulators
- Dealing with uncertainties

Event-level inference

Event level inference

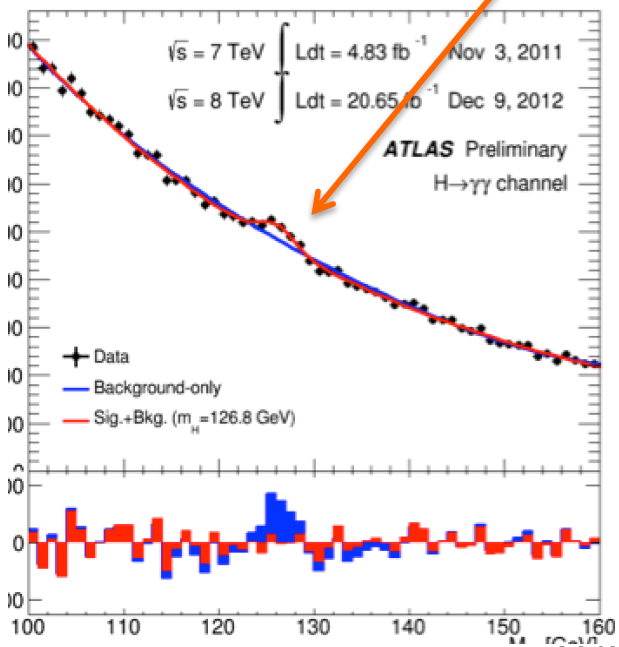
Particle-level inference

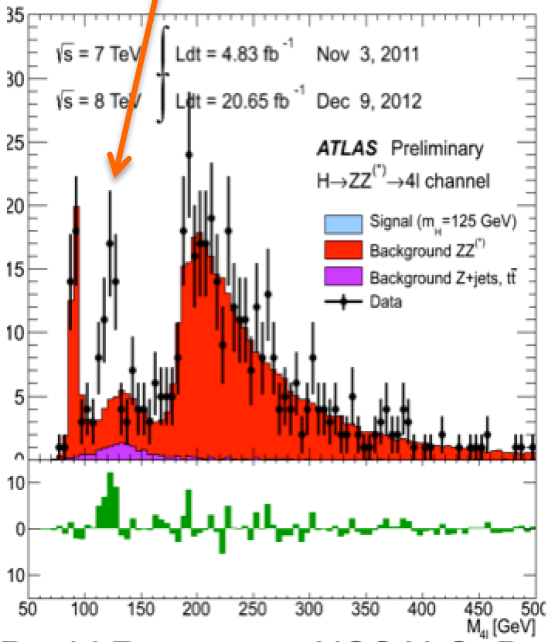
- Particles identified (pattern)
- And measured : 3D direction and energy, origin

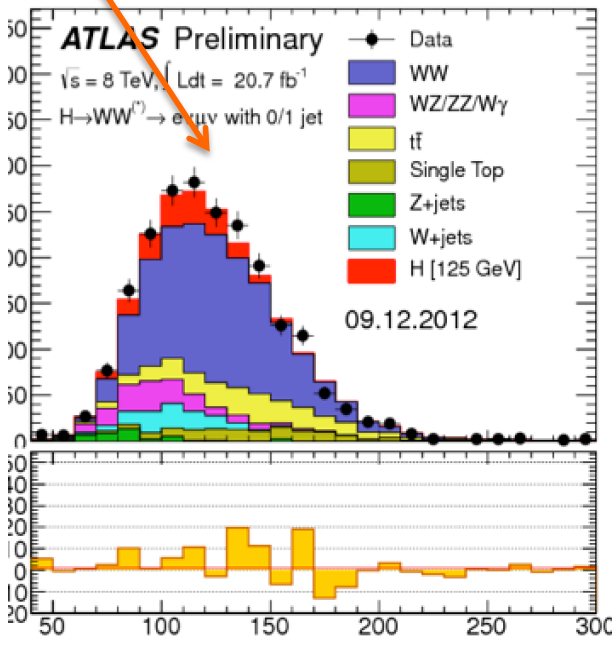


Experiment-level inference

To « see » the Higgs boson

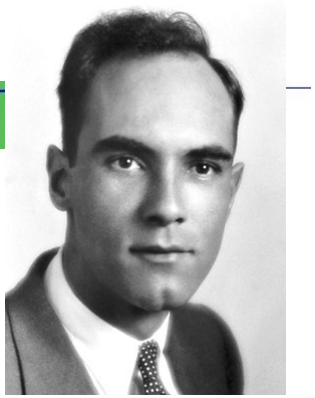






Positron discovery

- Cloud chamber picture



Carl Anderson
Nobel 1936

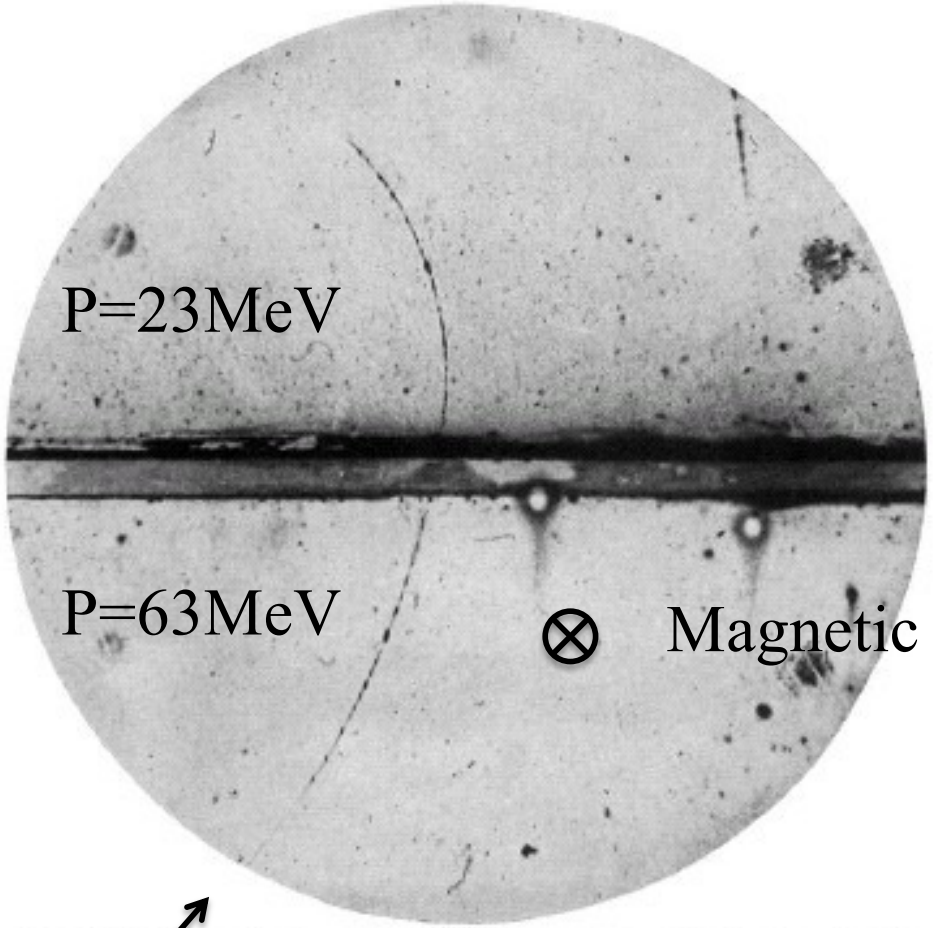
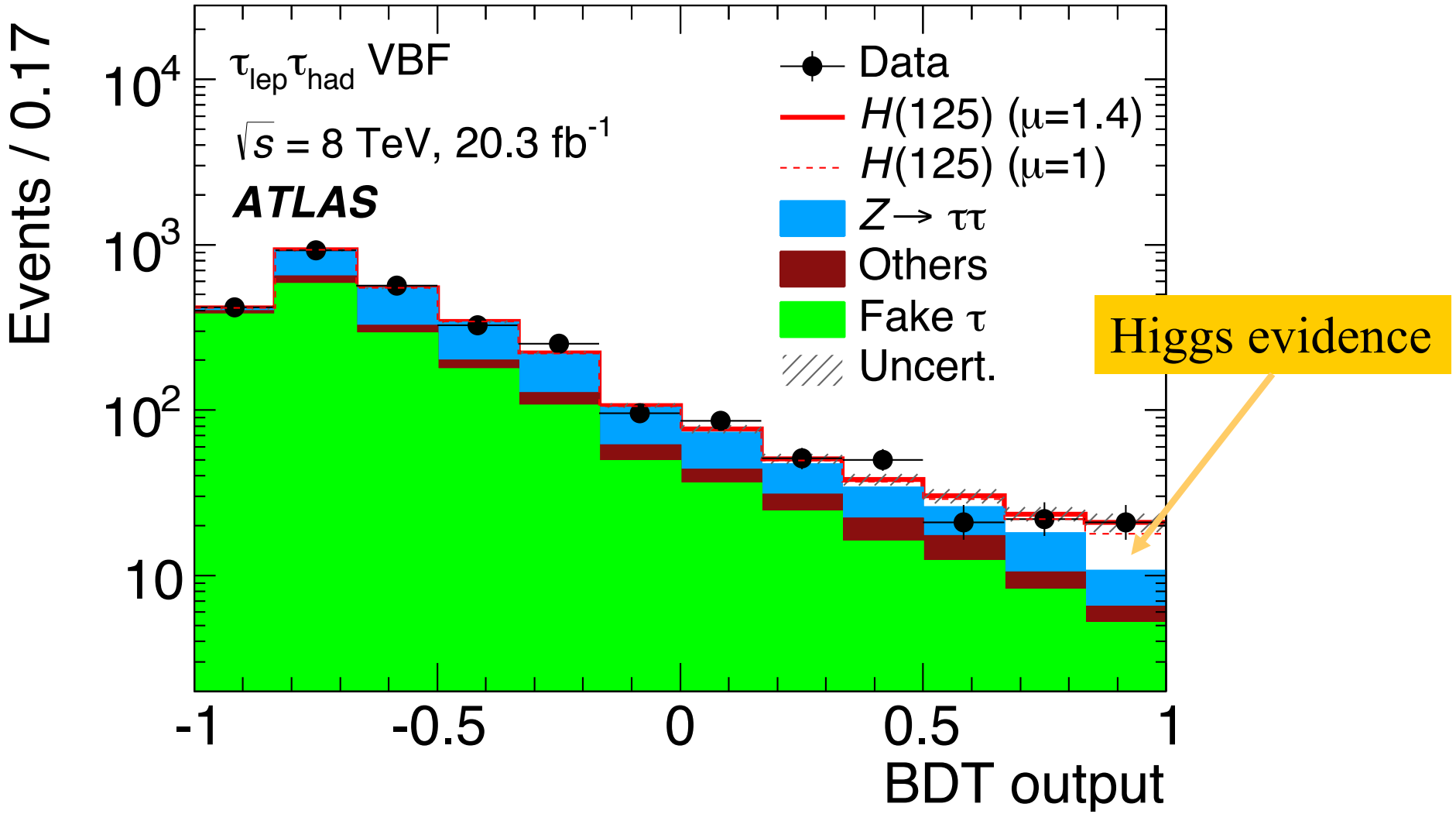


FIG. 1. A 63 million volt positron ($H\rho = 2.1 \times 10^5$ gauss-cm) passing through a 6 mm lead plate and emerging as a 23 million volt positron ($H\rho = 7.5 \times 10^4$ gauss-cm). The length of this latter path is at least two times greater than the possible length of a proton path of this curvature.

Evidence using a classifier

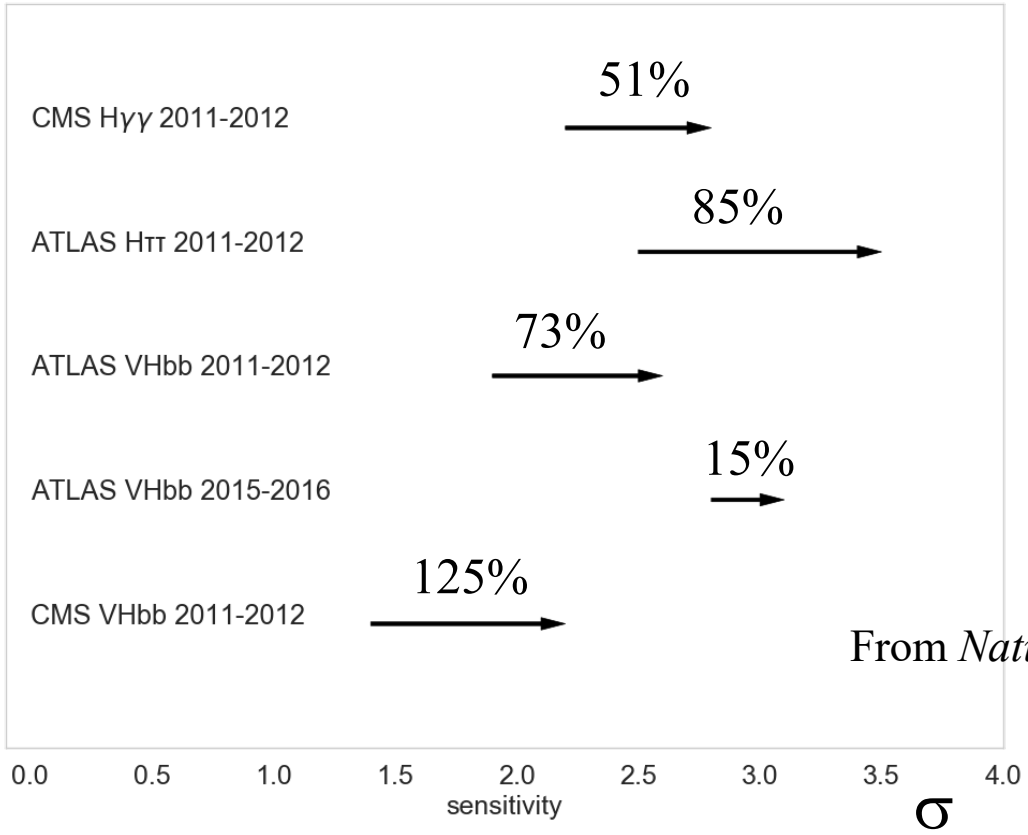
JHEP 04, 117 (2015) 1501.04943

Boosted Decision Tree using ~dozen of high level variables



ML on Higgs Physics

- At LHC, Machine Learning used almost since first data taking (2010) for reconstruction and analysis
- In most cases, Boosted Decision Tree on ~10 variables
- For example, impact on Higgs boson sensitivity at LHC:

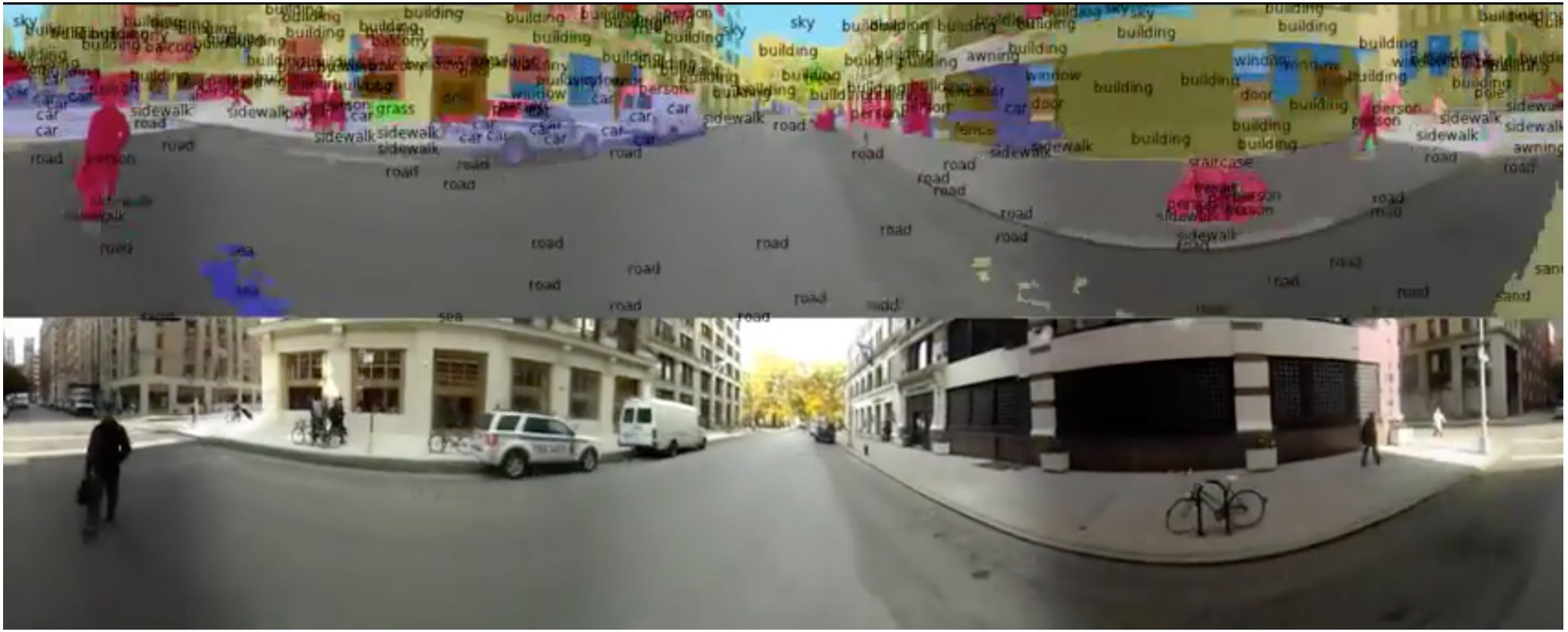


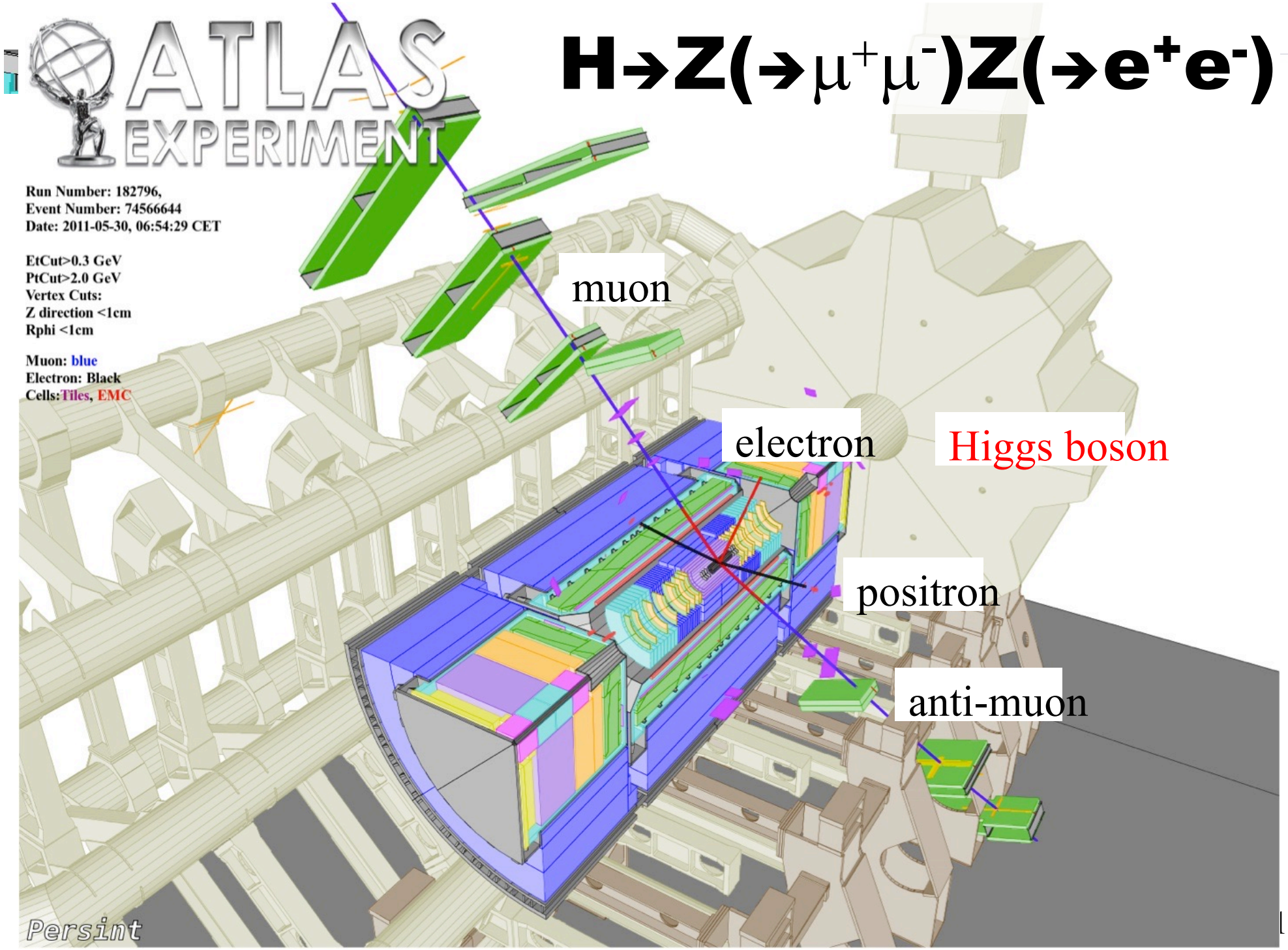
From *Nature* **560**, 41–48 (2018) |

➔~50% gain on LHC running

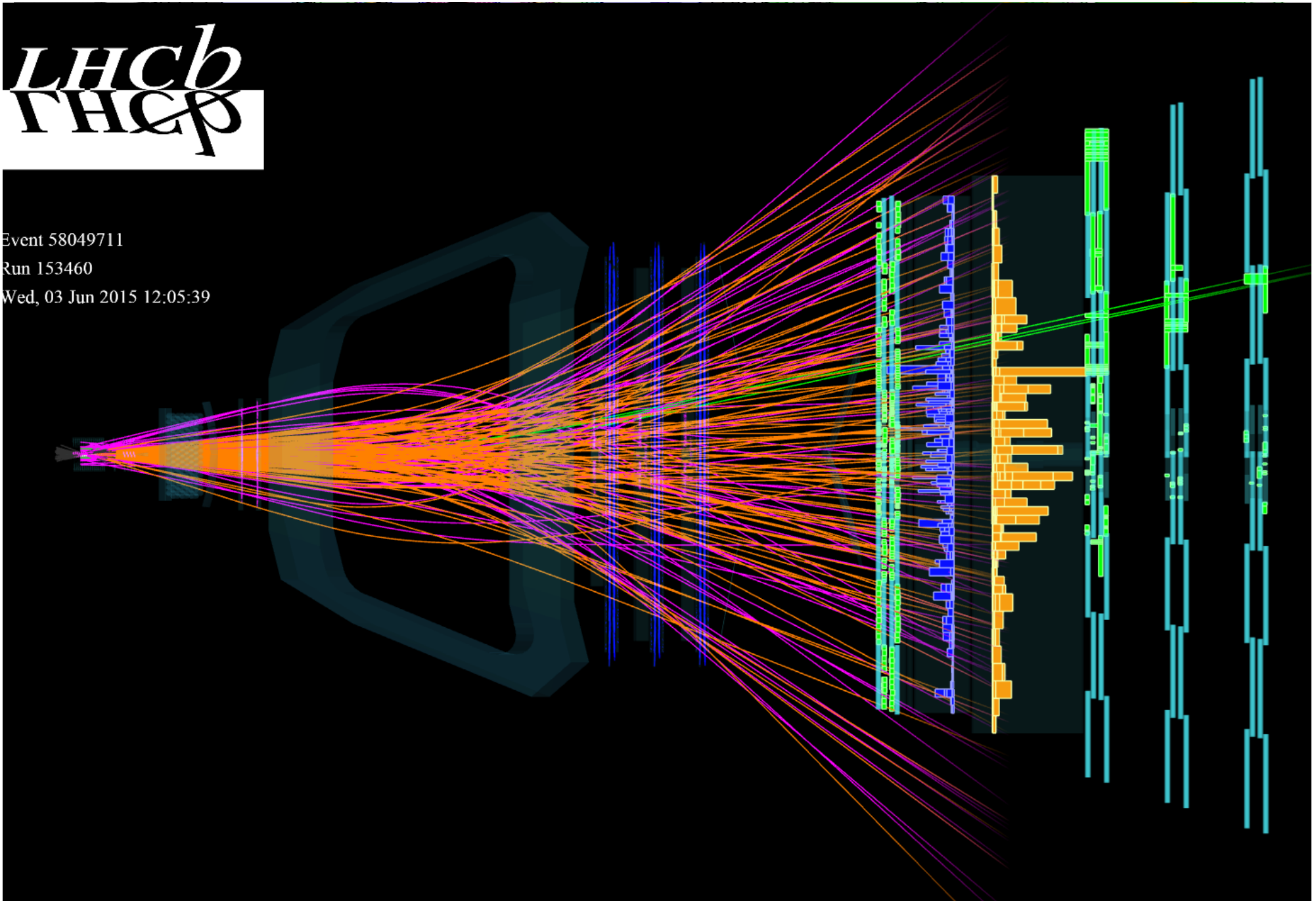
High Energy Physics data are not images

Typical Deep Learning application

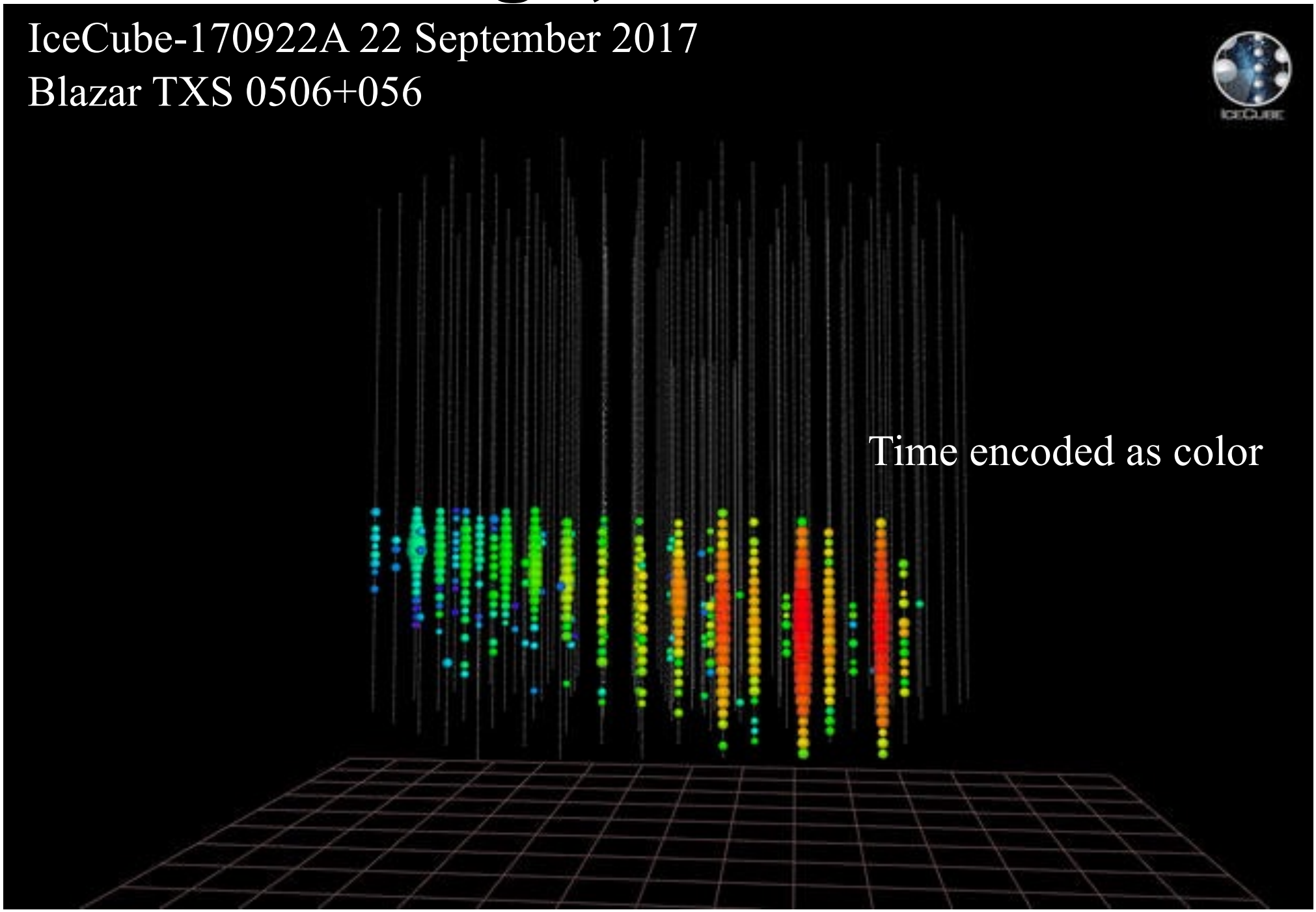



ATLAS
EXPERIMENT
Run Number: 182796,
Event Number: 74566644
Date: 2011-05-30, 06:54:29 CET
EtCut>0.3 GeV
PtCut>2.0 GeV
Vertex Cuts:
Z direction <1cm
Rphi <1cm
Muon: blue
Electron: Black
Cells:Tiles, EMC
H→Z(→μ+μ-)Z(→e+e-)
muon
electron
Higgs boson
positron
anti-muon
Persint

An image, not the data

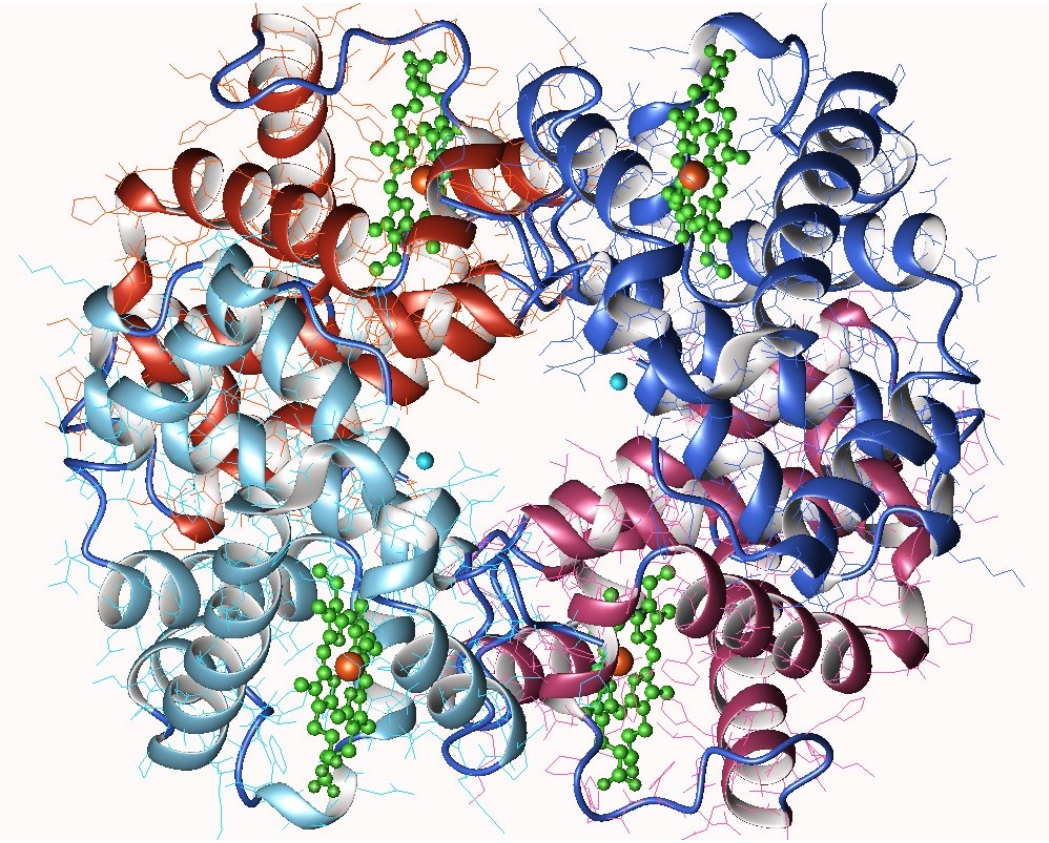


An image, not the data



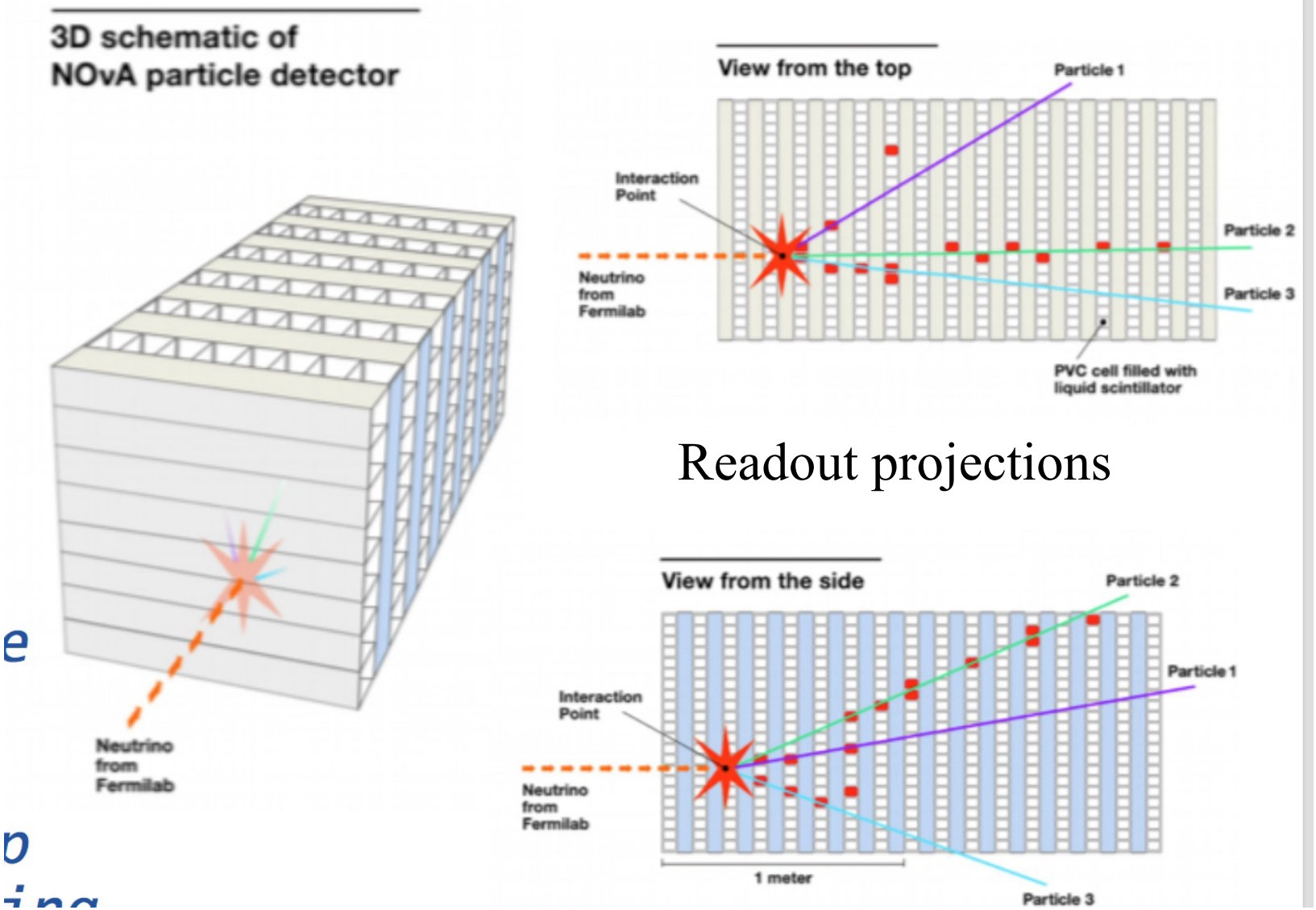
Une image, pas les données

Structure de l'hémoglobine



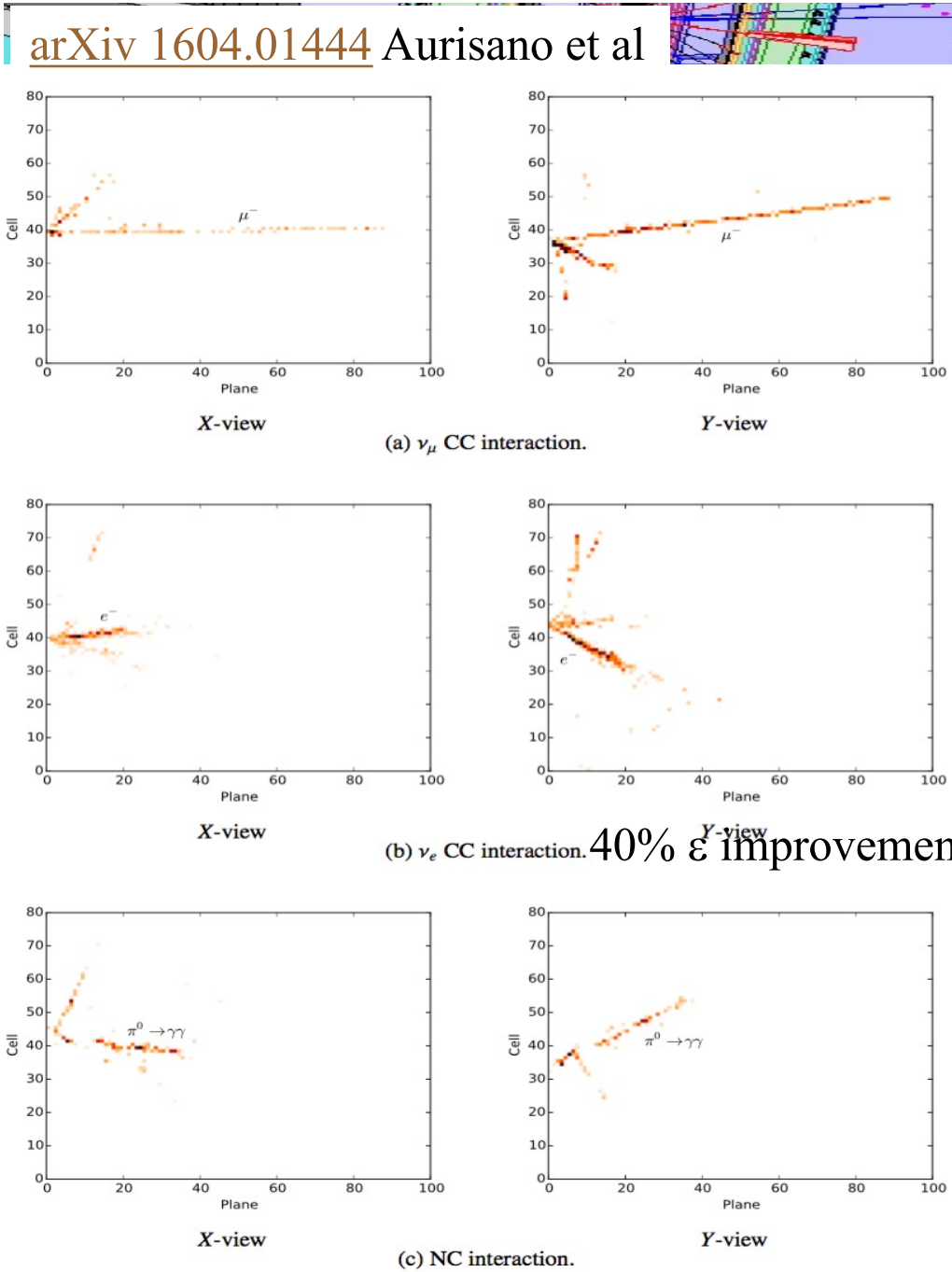
Par Deposition authors: Fermi, G., Perutz, M.F.; visualization author: User:Astrojan — https://www.rcsb.org/structure/3hhb, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=48714519

An exception : NOVA



An exception: NOVA

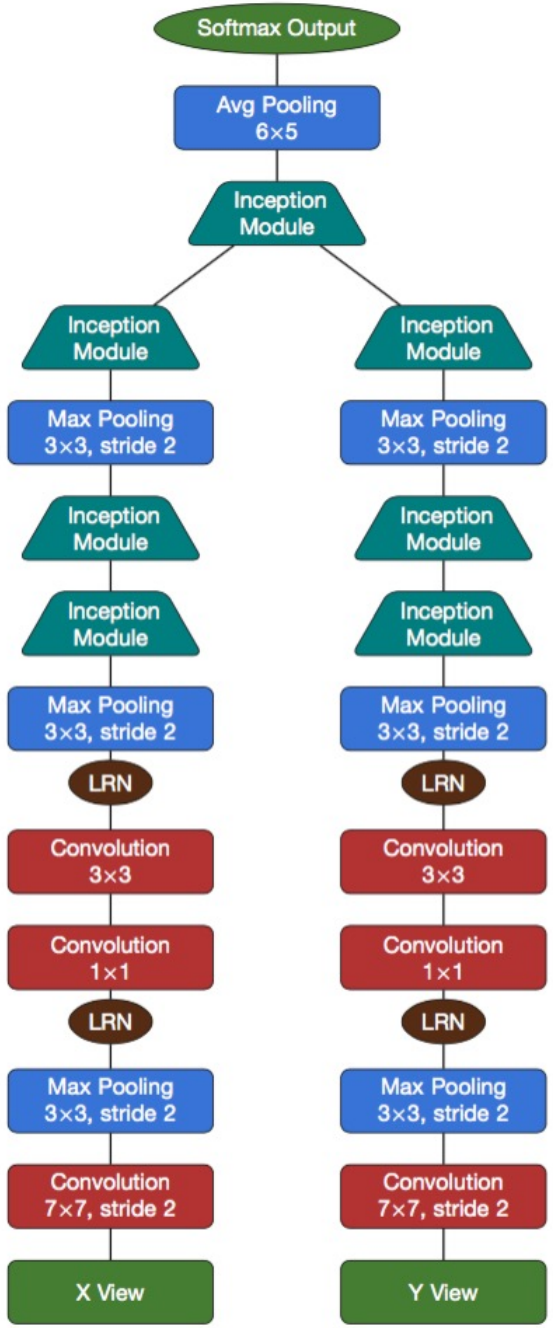
arXiv 1604.01444 Aurisano et al



40% ε improvement

Neutrino interaction classification
Using Convolutionnal Neural Network (GoogleNet)

Actually used in physics results 1703.03328 and 1706.04592

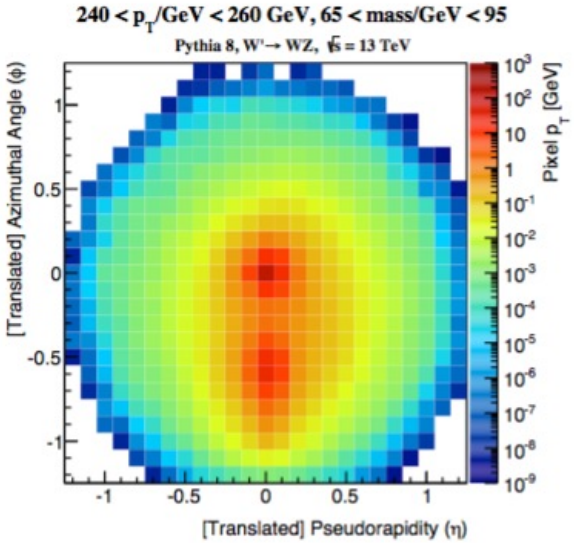


Jet Images

arXiv 1511.05190 de Oliveira, Kagan, Mackey, Nachman, Schwartzman

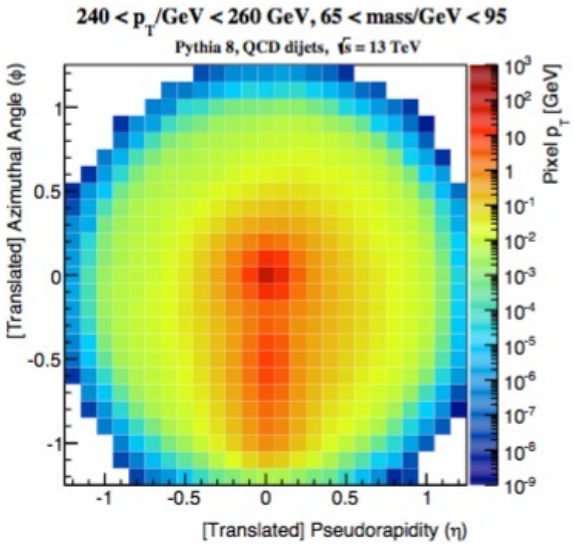
- Early attempt at image-like simulation
- ➔promising results, but not really applicable

Boosted W➔qq jet

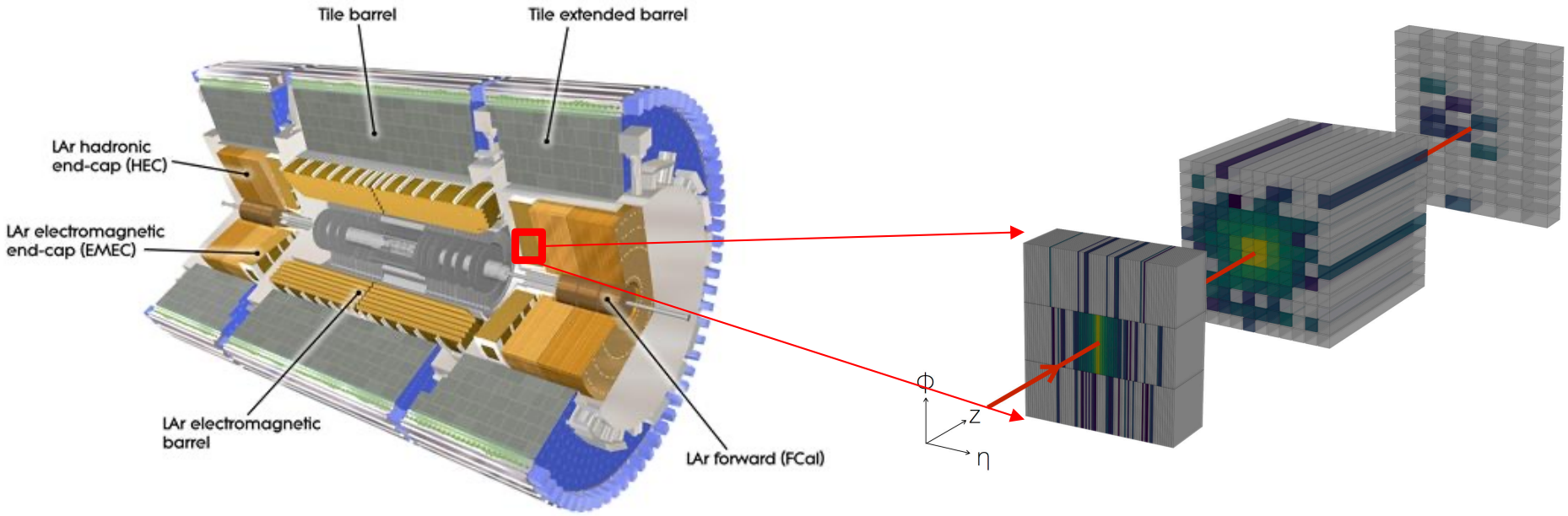


Average images

QCD



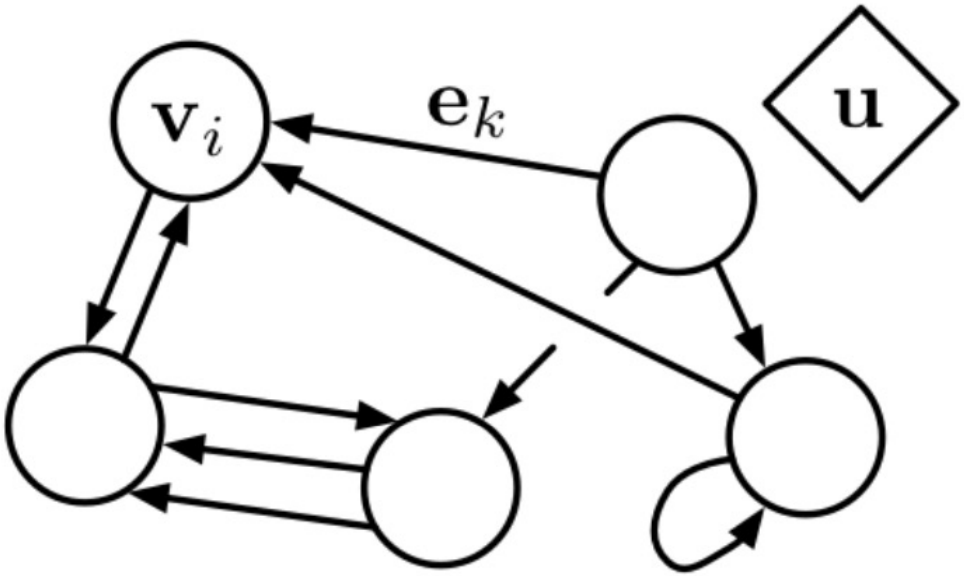
Detectors are complex 3D objects



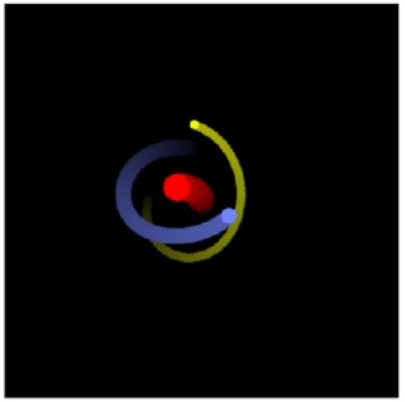
Graph Networks

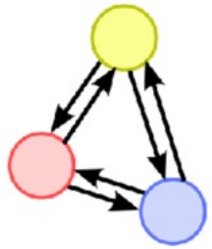
GNN

- Now some structure:
  - v_i : nodes
  - e_k : edges
  - u : global



n-body





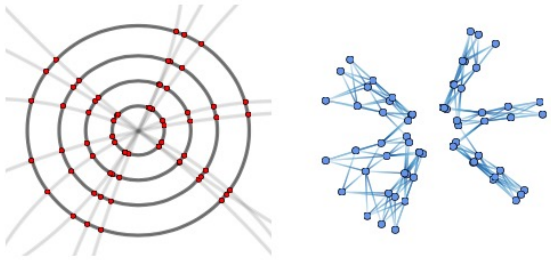
Nodes: bodies

Edges: gravitational forces

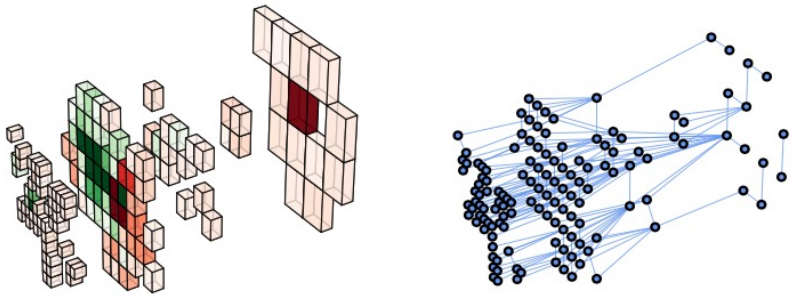
Global : potential energy

Graph on HEP data

from 2007.13681



(a)

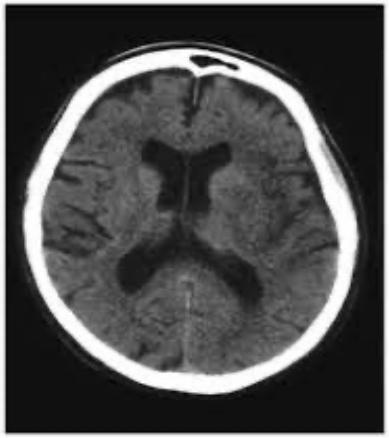


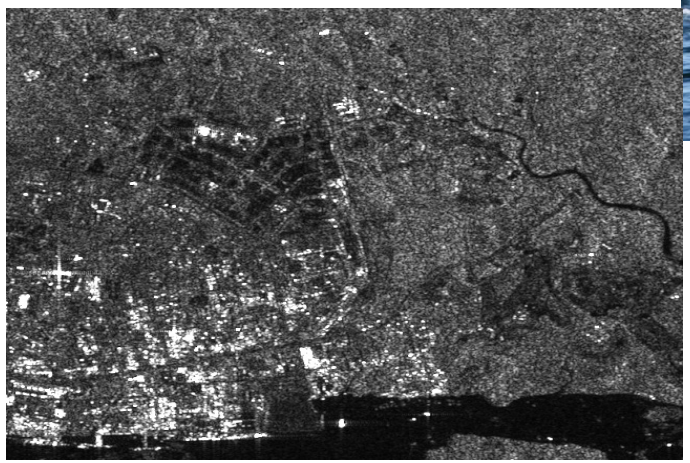
(b)

(c) (d)

Convolution Neural Networks application

CNN applications

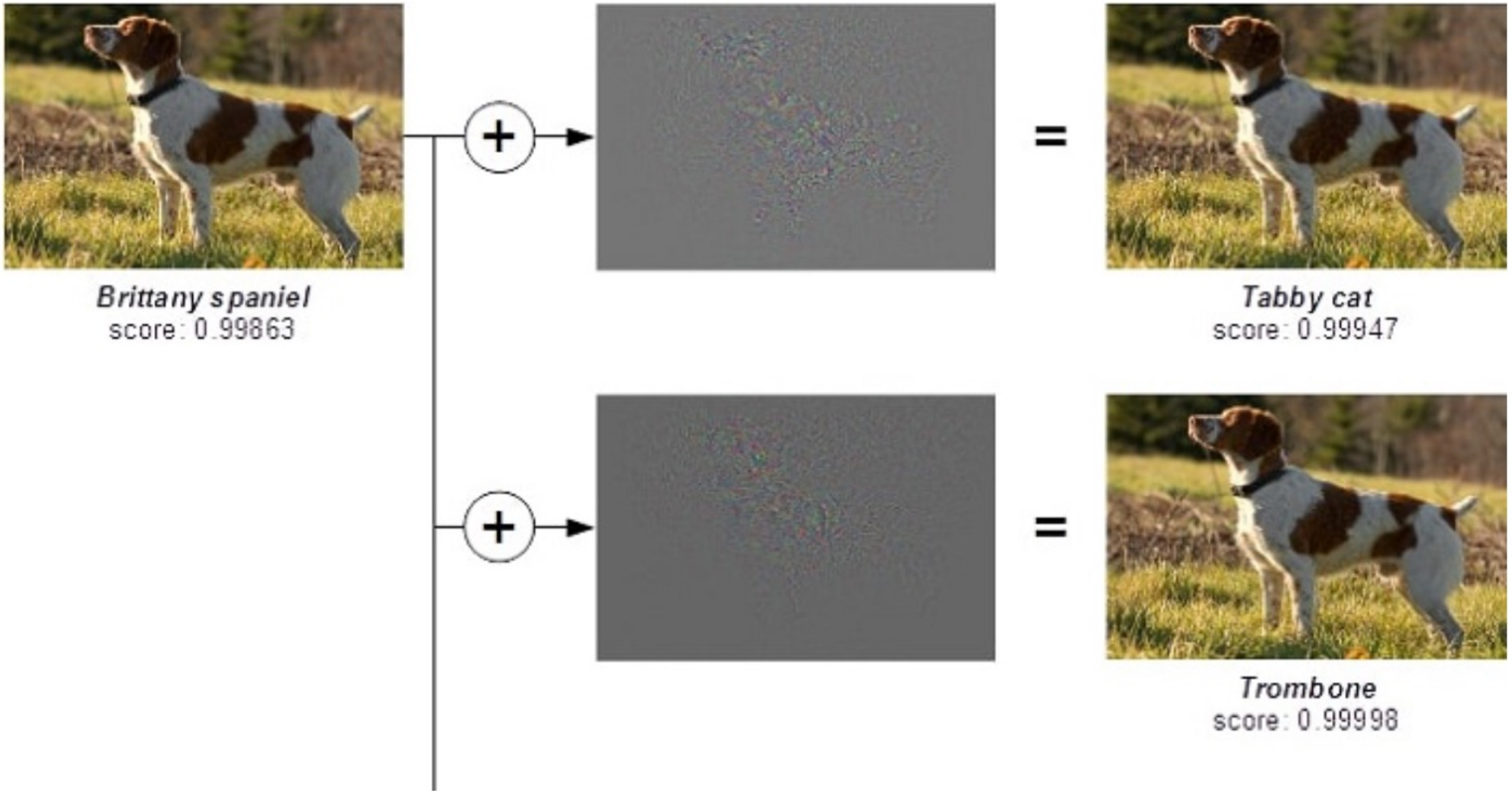




Adversarial examples

Review 1911.05268

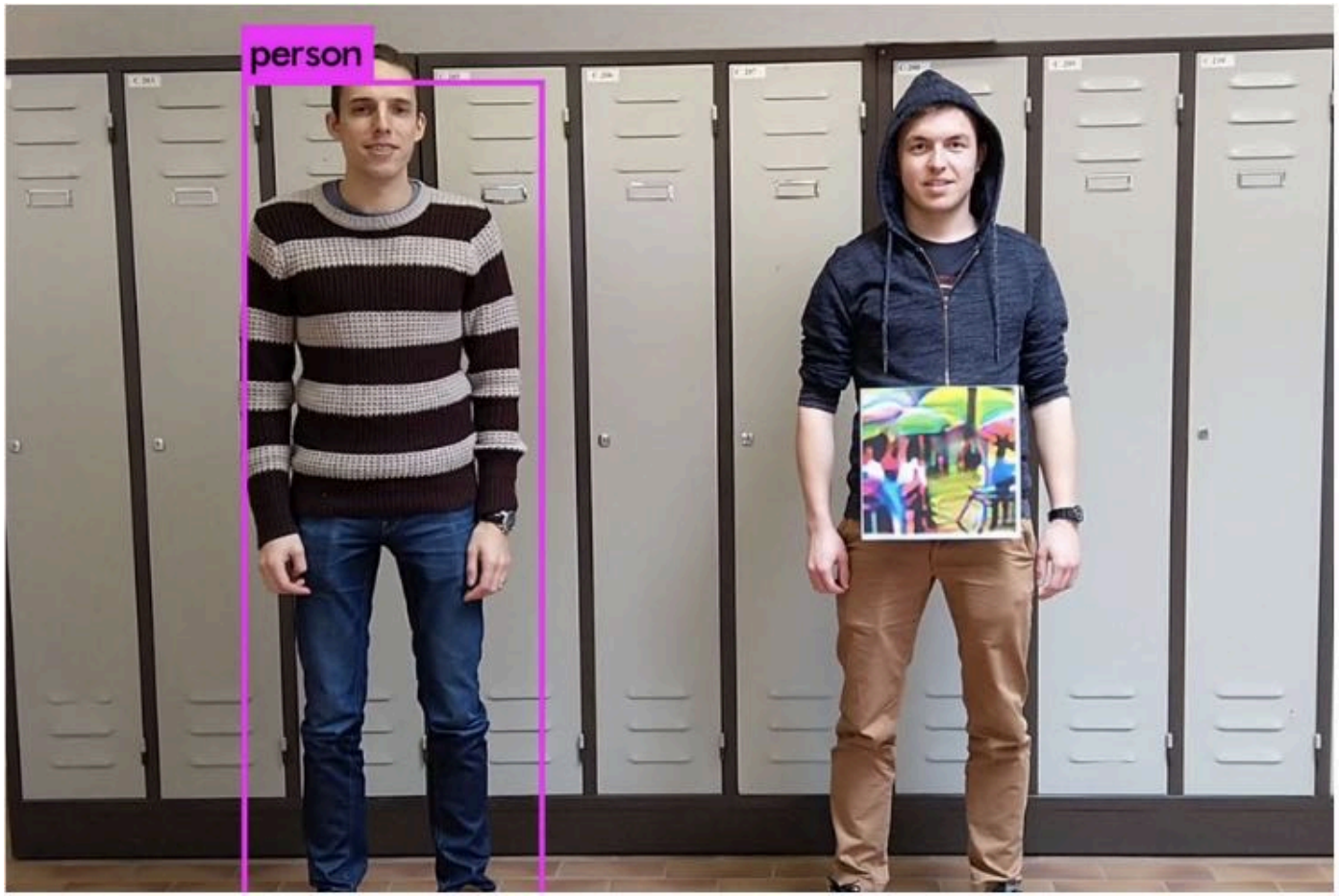
- Subtle alteration of an image fooling a classifier



Adversarial examples (2)

the verge 2019/04/23 (see also video)

- Extraneous object
- ➔more worrying, for HEP it would be e.g. a glitch in the data which is not simulated

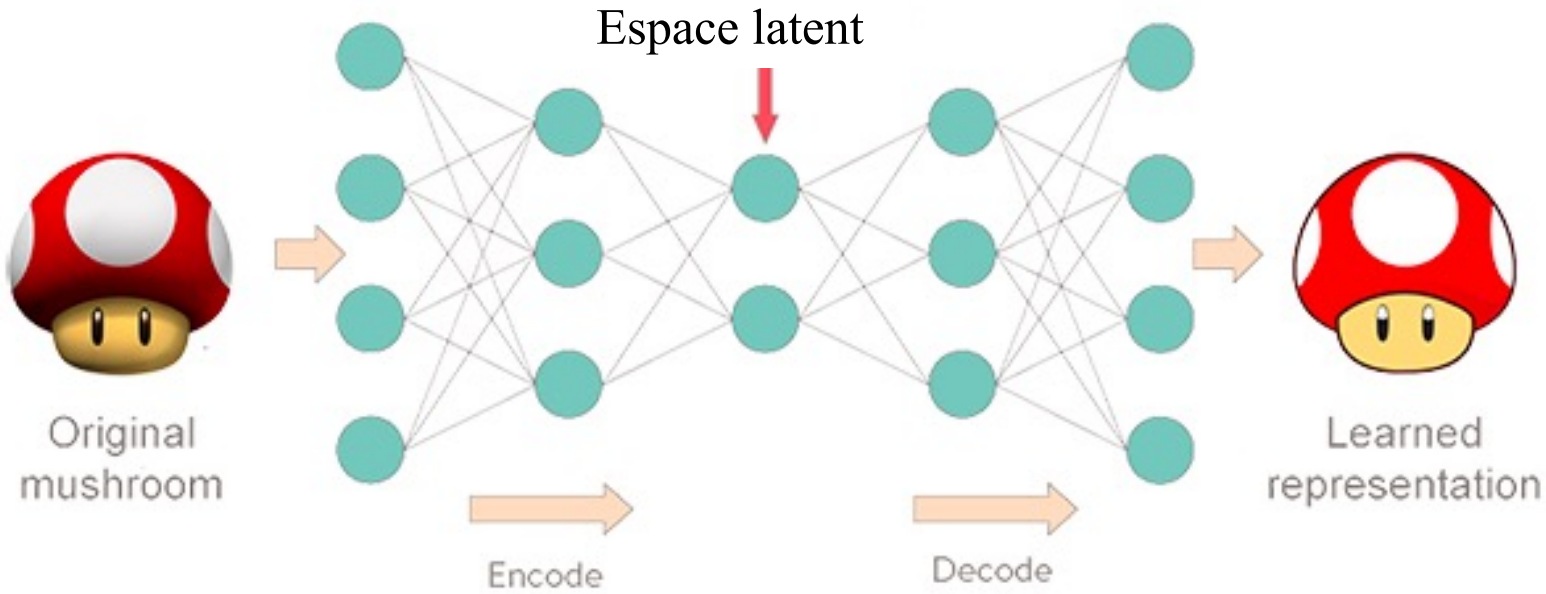


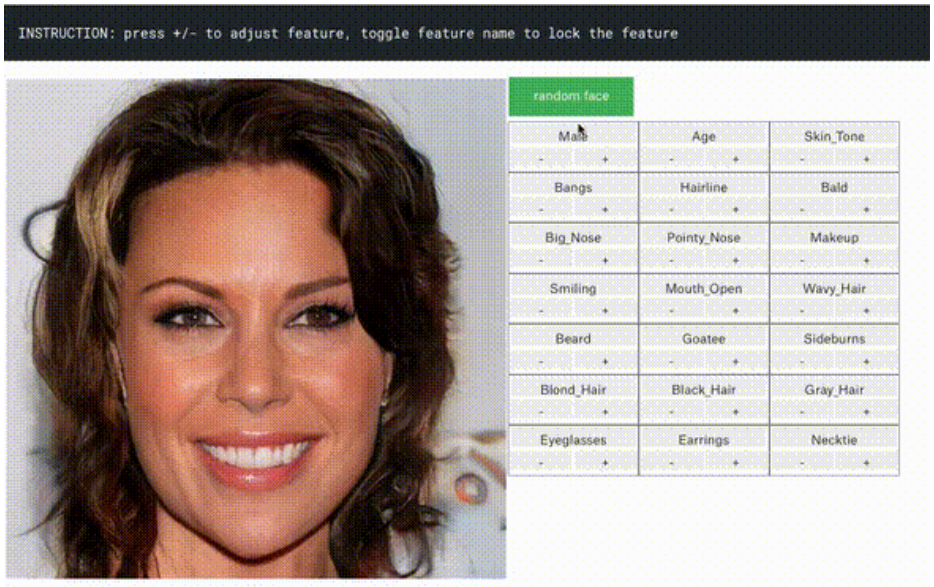
Are Adversarial Examples a worry for ML in HEP ?

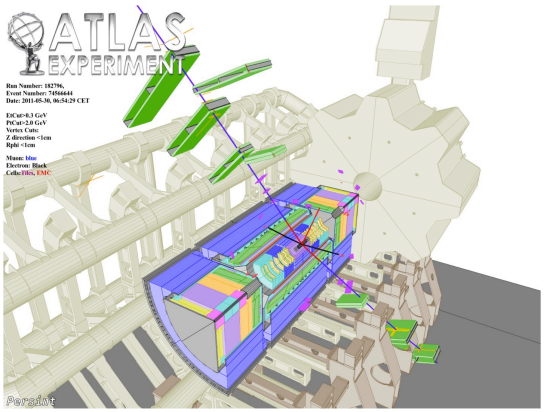
- Adversarial Examples are « inputs that are specifically designed to cause the target model to produce erroneous outputs »
  - ⇔Adversarial Examples are not accidental
  - ⇔Volume of feature space minuscule
- we could not think of a case a detector or physics glitch would fulfill the definition above

GAN/VAE to accelerate simulators

Auto encodeurs



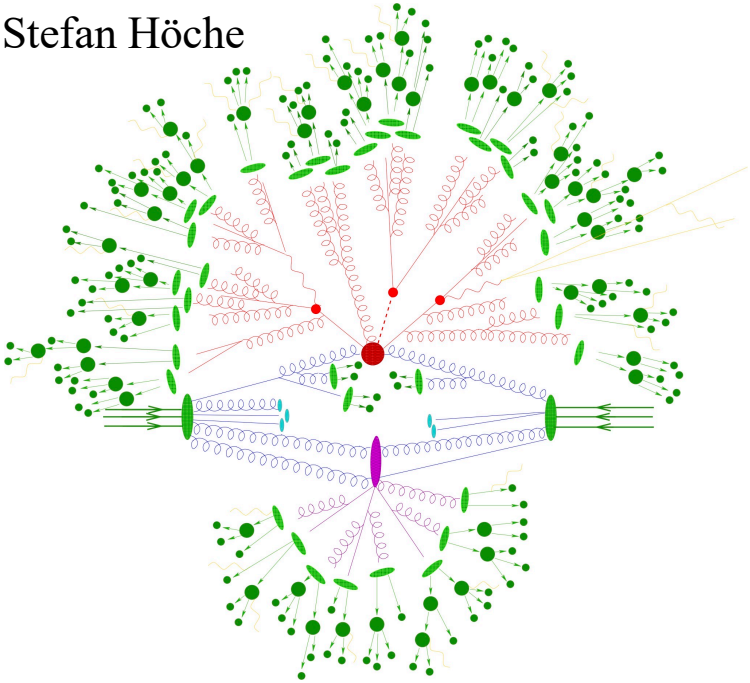




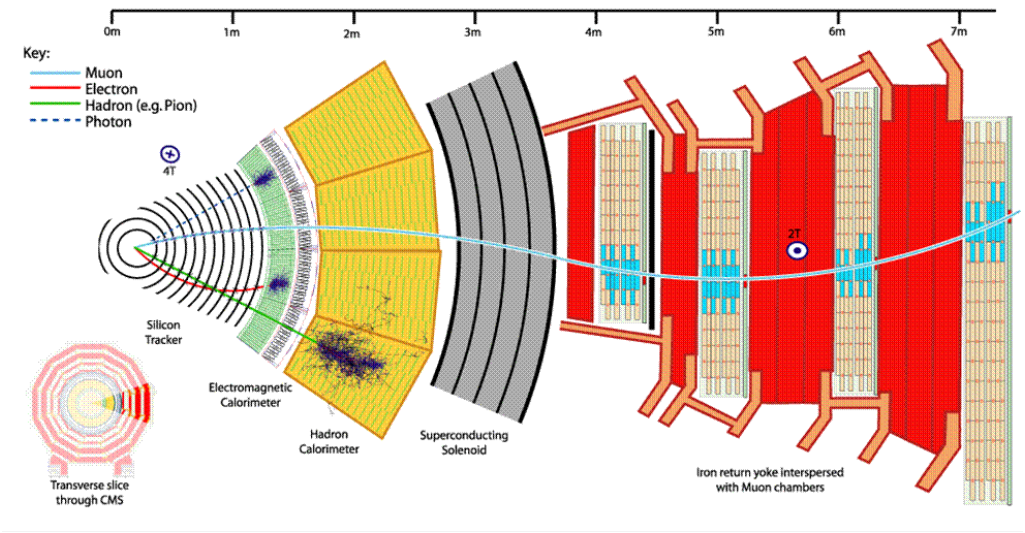
- Variables latentes==variables abstraites ~peu nombreuses caractérisant les différences entres images présentées
- Applications:
  - Classification/évaluation
  - Génération rapide
  - Explicabilité➔relation entre variable latente et concept scientifique
  - Problème inverse
  - Détection d'anomalies

Accurate simulators

Stefan Höche



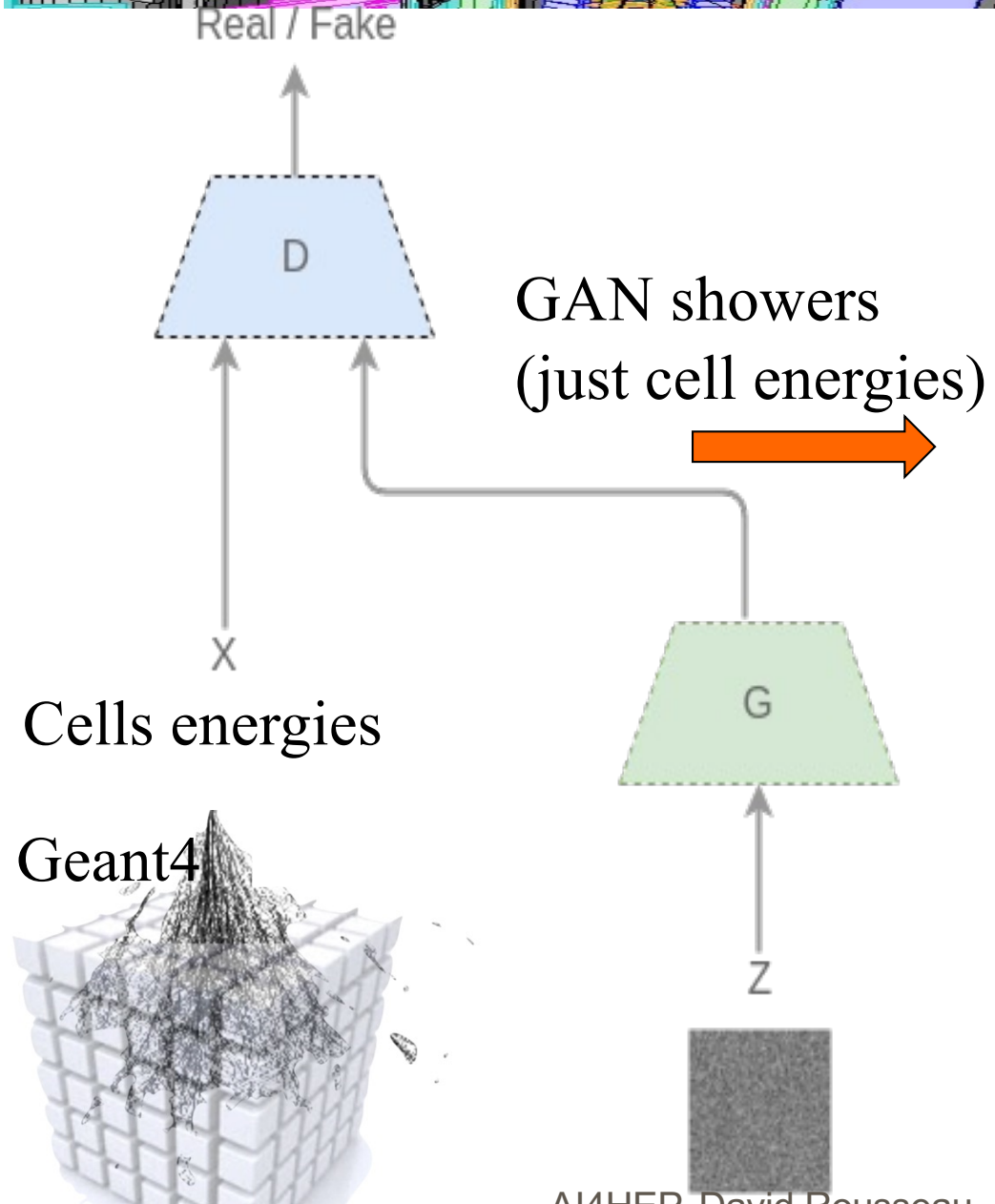
- Proton collision



- Interaction of particles with the detector

- ➔data very similar to real data from the experiment
- + ground truth
- This has been in HEP culture since the seventies, and developped through huge efforts in resource and manpower

GAN for simulation

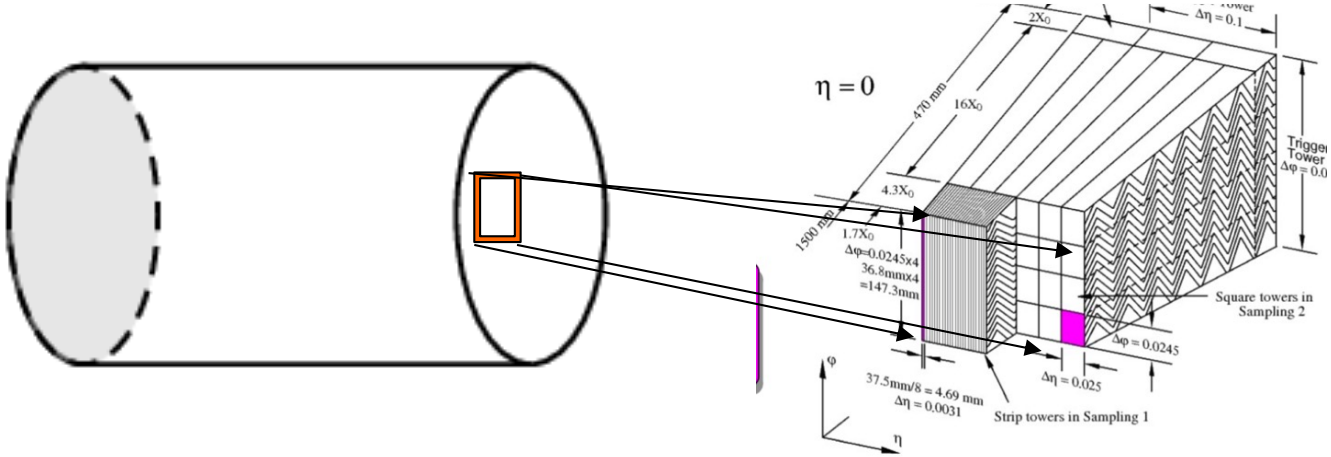


- Half of LHC grid computers (~1.000.000 cores) are crunching Geant4 simulation 24/24 365/365
- …while LHC experiments are collecting more and more events
- ➔reducing CPU consumption of simulation is very important

- Geant4 is the standard in particle simulation
- Imagine training a GAN on single particle showers of all types and energies
- Then when an event is simulated it would ask for GAN showers on request (superfast by 3-4 order of magnitude)

ATLAS calo simulation

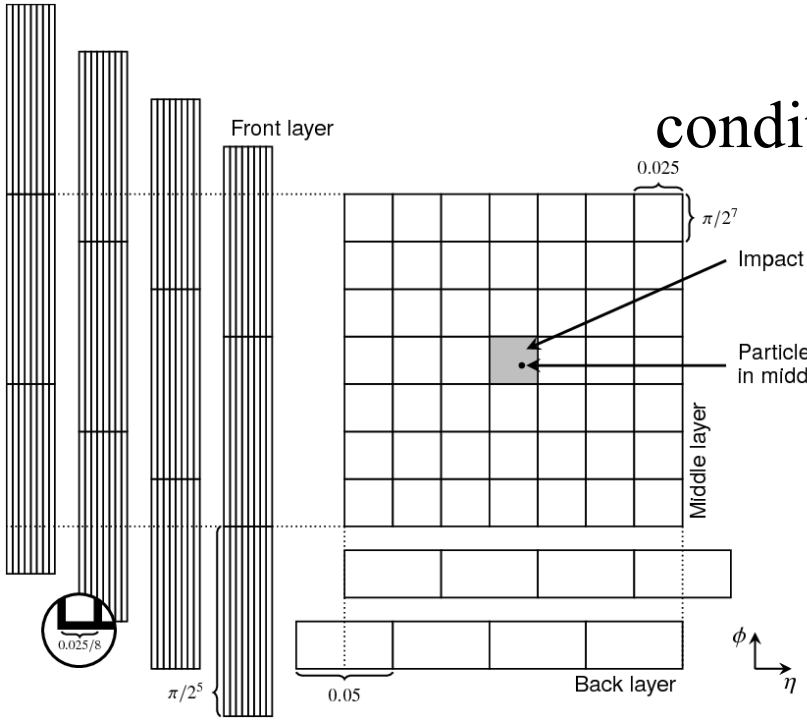
ATL-SOFT-PUB-2018-001 and update ATLAS-SIM-2019-004 proper paper to be published real soon now

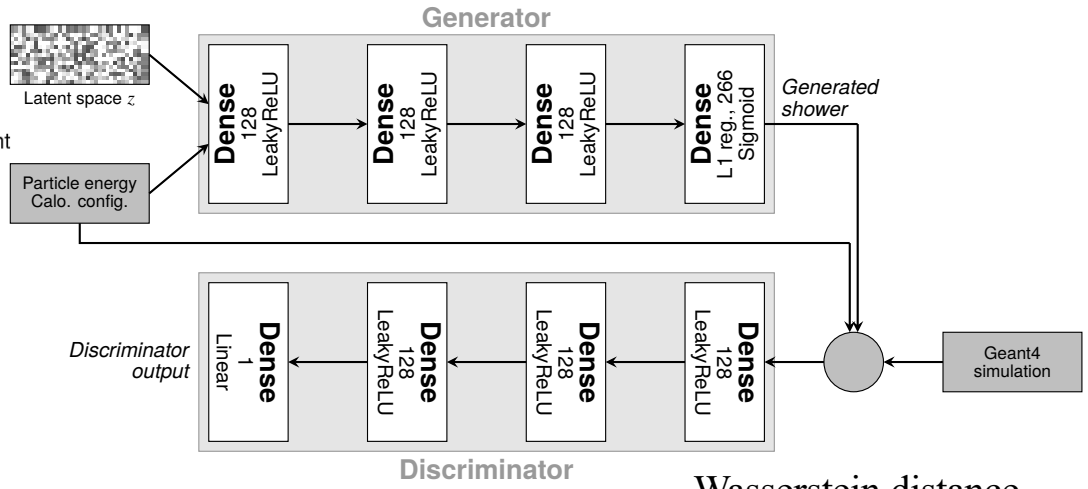


conditions

+ η, ϕ translation

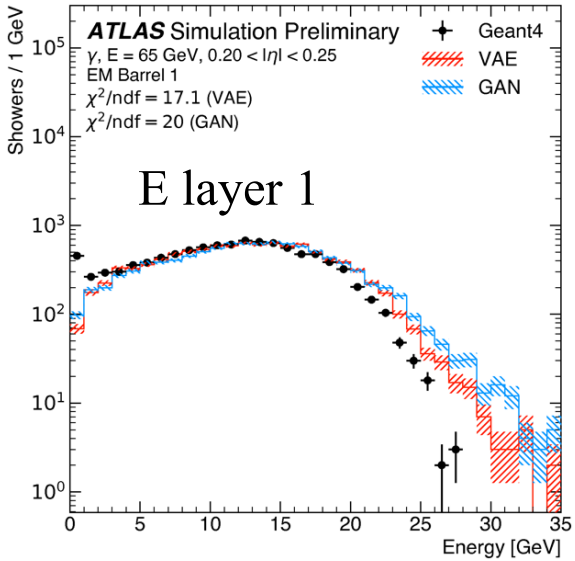
177000 cells ➔266 cells

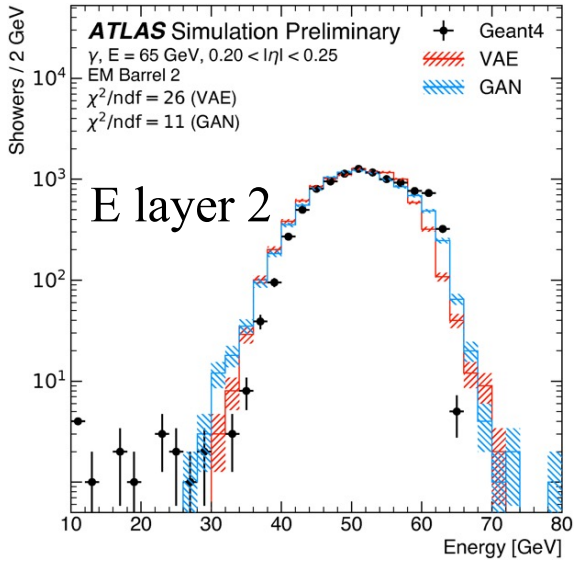


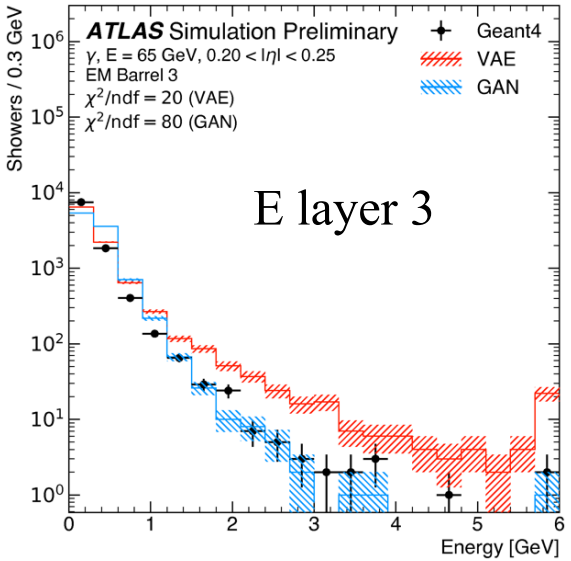


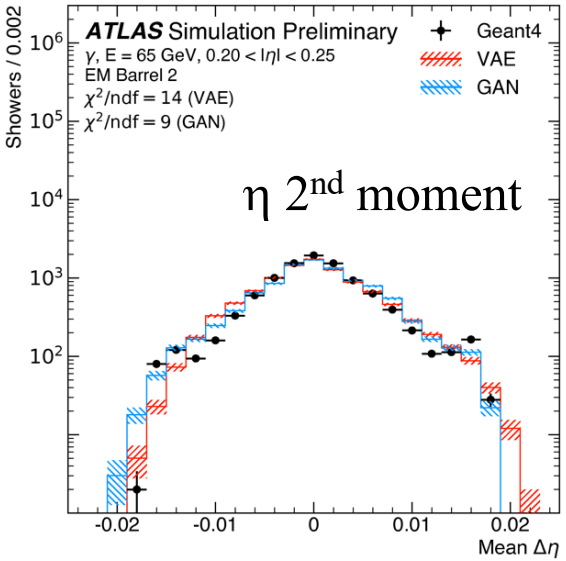
Wasserstein distance

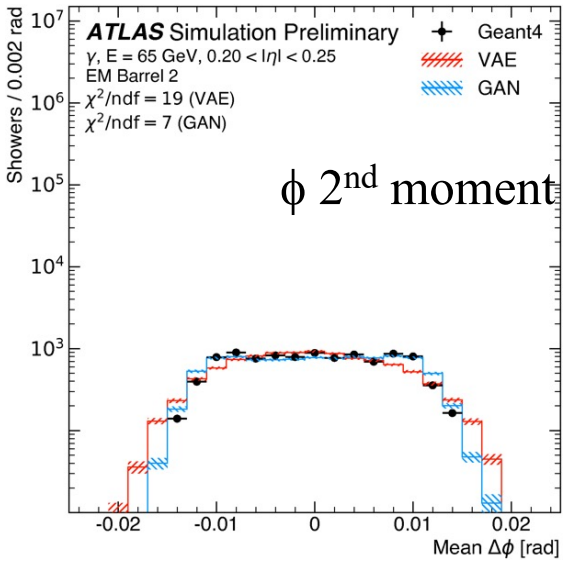
Results











- Speed: <1ms compared to 10s
- Sufficient accuracy for physics ?
- Handling of (more) awkward geometries ? (➔ **graph based GAN)**

Dealing with Uncertainties

Phys.Rev.Lett. 114 (2015)191803

Most complex measurement ever ?

Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments

(ATLAS Collaboration)[†]
(CMS Collaboration)[‡]

(Received 25 March 2015; published 14 May 2015)

A measurement of the Higgs boson mass is presented based on the combined data samples of the ATLAS and CMS experiments at the CERN LHC in the $H \to \gamma\gamma$ and $H \to ZZ \to 4\ell$ decay channels. The results are obtained from a simultaneous fit to the reconstructed invariant mass peaks in the two channels and for the two experiments. The measured masses from the individual channels and the two experiments are found to be consistent among themselves. The combined measured mass of the Higgs boson is $m_H = 125.09 \pm 0.21\,(\text{stat}) \pm 0.11\,(\text{syst})$ GeV.

Dealing with Uncertainties

- Our experimental measurement papers typically end with
  - measurement = m ± σ(stat) ± σ(syst)
  - σ(syst) systematic uncertainty : known unknowns, unknown unknowns… Convincing oneself, co-authors, the whole community that we know what we are doing➔trust !
- Name of the game is to minimize quadratic sum of :

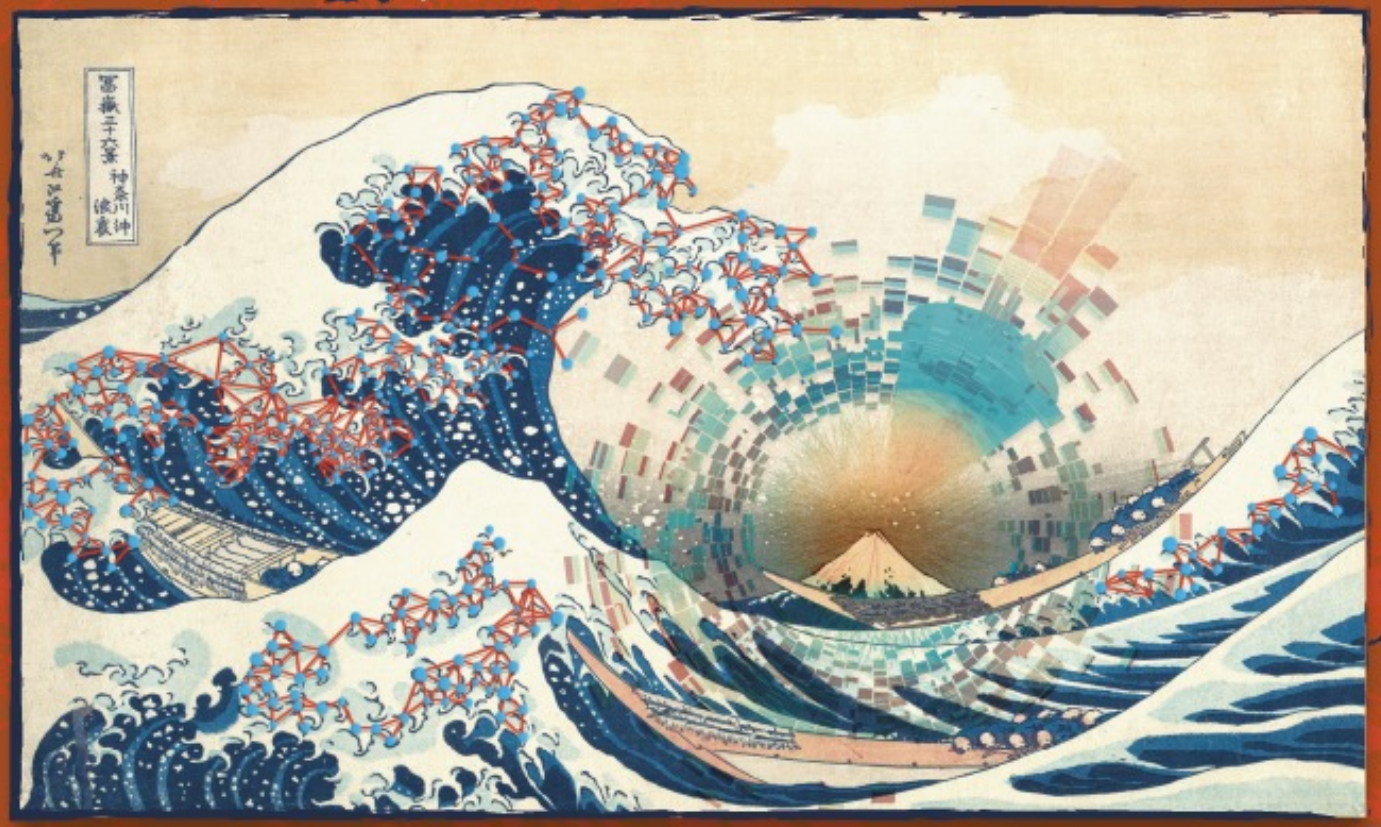
$$\sigma(\text{stat}) \oplus \sigma(\text{syst})$$

- … while ML techniques are usually trained to minimise σ(stat)
- Two challenges:

1. Maintain trust (σ(syst)) while using AI more and more
2. Include somehow (various techniques) $\sigma(\text{stat}) \oplus \sigma(\text{syst})$ in the loss in order to minimise overall uncertainty
   - However : Goodhart's Law "When a measure becomes a target, it ceases to be a good measure."

ARTIFICIAL INTELLIGENCE AND THE UNCERTAINTY CHALLENGE IN FUNDAMENTAL PHYSICS

27 NOV - 2 DEC 2023



ORGANISING COMMITTEE

Corentin Allaire (IJCLab-Orsay)

Olaf Behnke (DESY)

Anja Butter (LPNHE-Paris)

Sylvain Chevallier (LISN, Université Paris-Saclay)

Valérie Gautard (CEA Saclay)

Louis Lyons (Imperial College & Oxford)

David Rousseau (IJCLab-Orsay, Université Paris-Saclay)

Jean-Roch Vlimant (CalTech)

Thomas Vuillaume (LAPP-Annecy)

UNCERTAINTY QUANTIFICATION

EXPLAINABLE/TRUSTWORTHY AI

DATA-FRUGAL/DATA-CENTRIC AI

SIMULATION-BASED INFERENCE, UNFOLDING,...

ARCHITECTURES: ADVERSARIAL, BAYESIAN,...

CONTROLLING UNCERTAINTIES IN GENERATIVE MODELS

BENCHMARKS DATASET AND CHALLENGES

INDICO.IN2P3.FR/EVENT/30589

cnrs

AI SSAI

AI for science, science for AI

Conclusion

- We (in High Energy Physics) are analysing data from multi-billion € projects➔should make the most out of it!
- Dedicated representations (often Graph NN based) are being developped to deal with our semi-structured data
- Generative Models are accelerating our existing accurate but heavy simulators
- The bottom line is always a measurement with uncertainties which sum up the trust of the community, which should be maintained