Development of ML for Particle Physics Training and CS collaboration

GT09 Town Hall Meeting: Calcul, Algorithmes et Données

17-18 Octobre 2019

Julien Donini
LPC / Université Clermont Auvergne

Outline

Overview of ML activities @ IN2P3

ML software and tools

ML algorithms

Computing and hardware resources

Collaborations with CS / maths

Training and schools

Conclusion

Disclaimer

This talk is ...

… based on input sent by teams : some aspects might be uncovered

… probably biased towards things that I know better

… full of acronyms (sorry !)

And thanks a lot to all who provided for material and explanations !

Categories of ML activities (HEP)

Based on classification in *Machine Learning in High Energy Physics Community White Paper*, https://arxiv.org/abs/1807.02876

1. **Detectors & accelerators**
2. **Simulation**

3. **Object Reconstruction, Identification, and Calibration**
4. **Real Time Analysis and Triggering**
5. **Uncertainty Assignment**
6. **Learning the Standard Model – searches for anomalies**

7. **Matrix Element Method with ML**
8. **Theory Applications**

9. **Computing Resource Optimization**

ML @ IN2P3

1. **Detectors & accelerators**
2. **Simulation**

Color code

Advanced
Studies
Interest

Detector design

- Use ML to optimize detector design (LPNHE)

ML for Accelerator developments

- **Accelerator** tuning, lasers, virtual detectors (LAL)
- NN for particle **accelerator** operations and optimization (LPSC)

Simulation

- Simulation of ATLAS **calorimeter** with GAN's (LAL)
- MC sample **reweighting** in ATLAS (LPNHE)
- NN to simulate **fuel evolution** in nuclear reactors (IPNO)
- BDT's for multidim **reweighting** between MC (LAL)
- Gaussian Processes to **smooth MC** stat fluctuations (LAL)

Accelerator physics

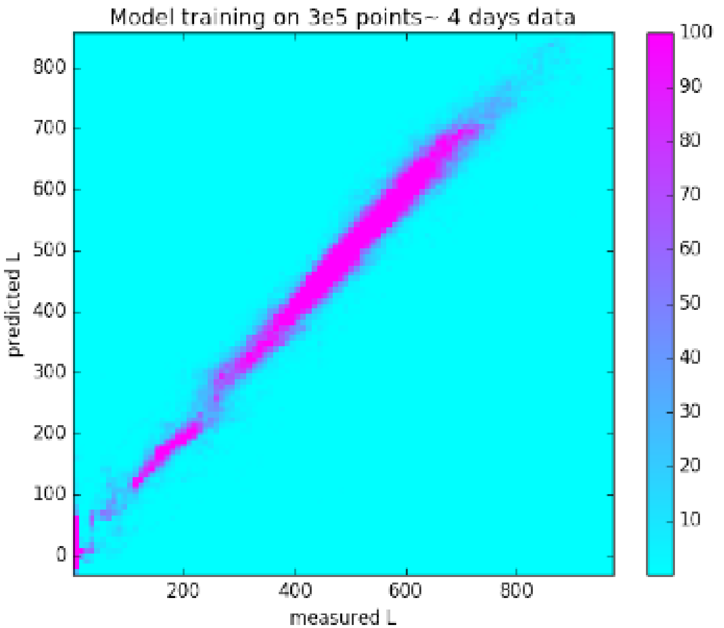
V. Kubytskyi, H. Guler, K. Cassou et al. (LAL)

Involved in local accelerators (ThomX, PHIL, PERLE, …), local Lasers (LaserX, …) and international collaborations (BELLE2@KEK, CLEAR@CERN, …)

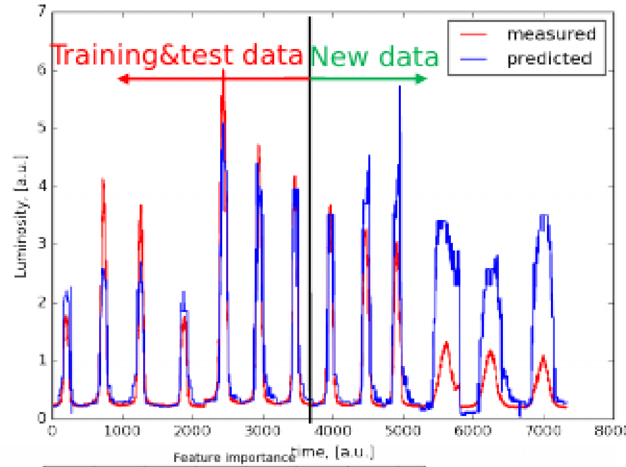
→ Needs for ML appear at different stages:

*Example 1: training a **virtual** luminosity monitor*

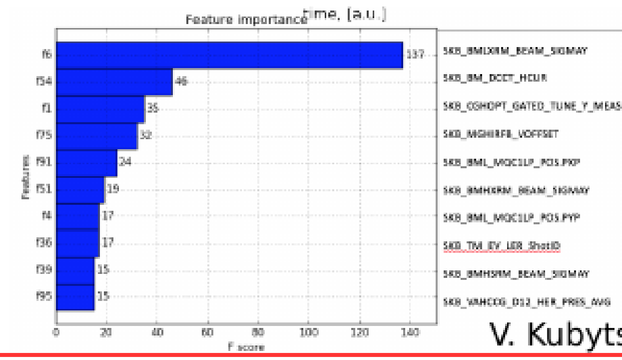
- Fast luminosity monitor LumiBelle2/SuperKEKB
- Train ML model (XGBoost) to predict luminosity from 165 machine variables parameters



Example 2: training over luminosity scans



- In total 30 scans (100* 2e5 datapoints): use first 24 for training and last 6 for prediction.
- The agreement is not perfect but number of scans available is relatively small.



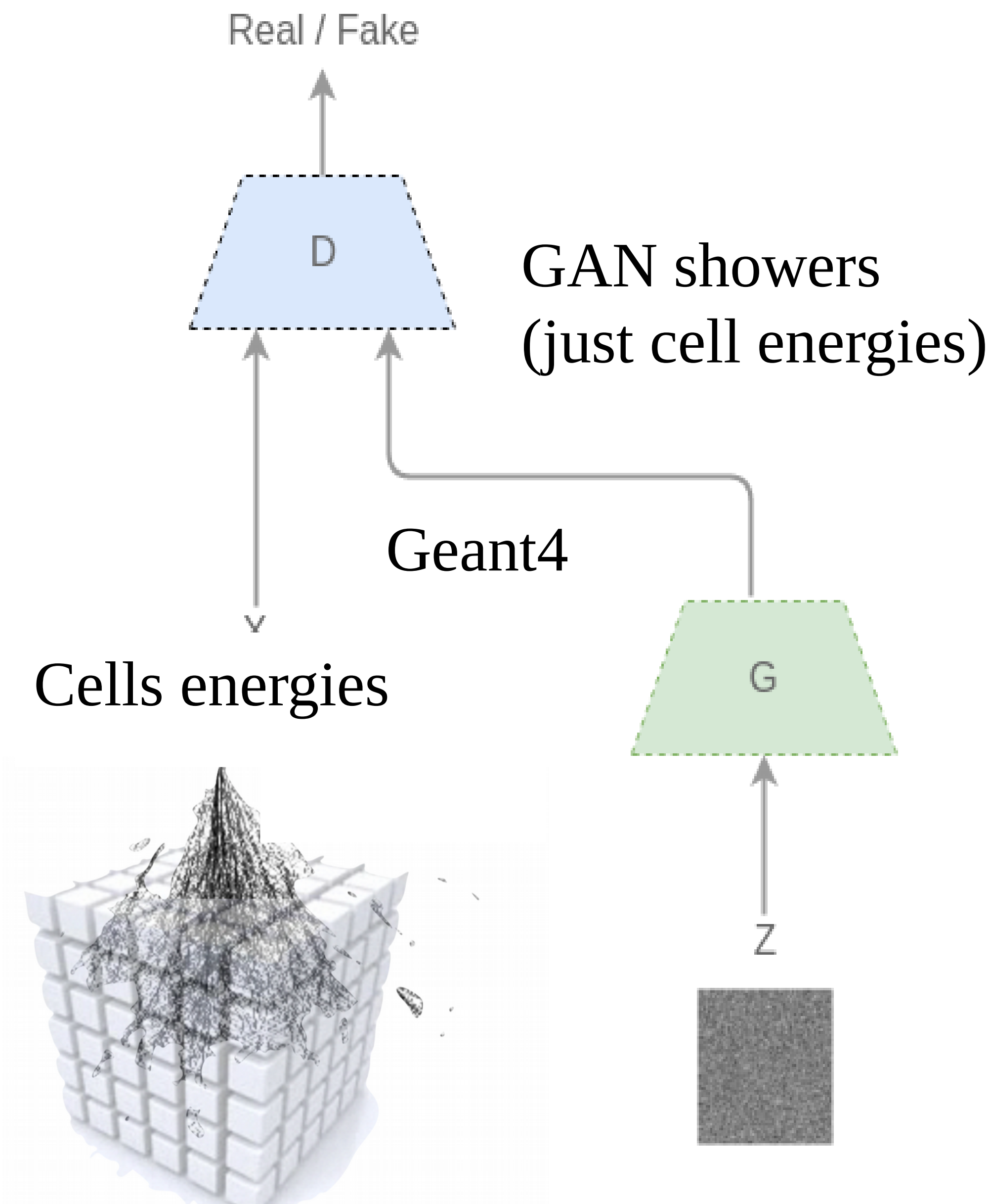
- Feature importance reveals strongest dependencies, however could be sensitive to the data sample chosen

V. Kubytskyi, H. Guler, K. Cassou et al.

Perspectives: machine **tuning** & beam dynamics, control of high intensity **lasers**, **virtual detectors** for machine monitoring purpose

GAN for simulation for ATLAS

D. Rousseau, A. Ghosh (LAL), G. Louppe (U Liège)

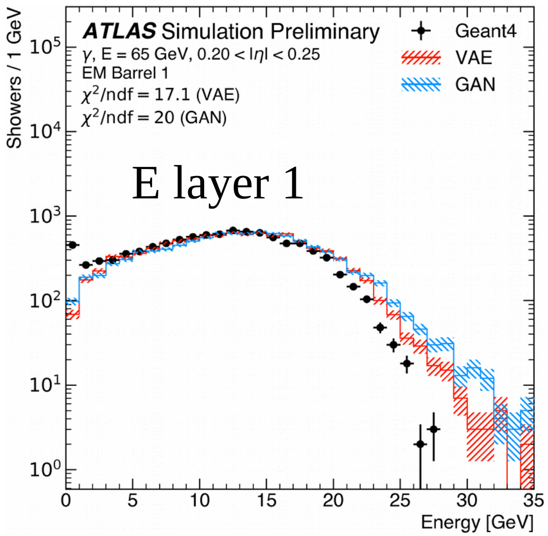


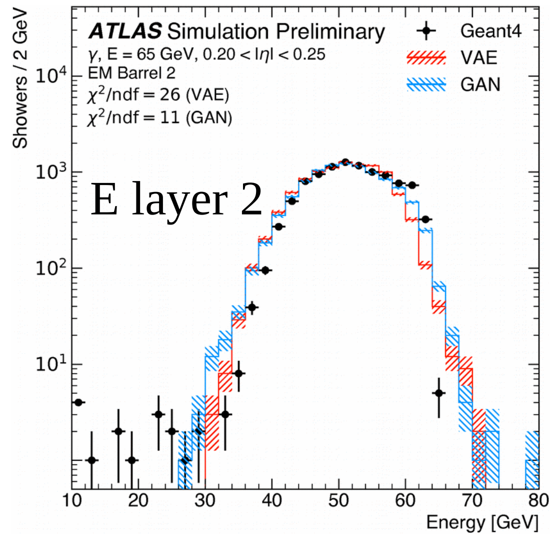
- Half of LHC grid computers (~300.000 cores) are crunching Geant4 simulation 24/24 365/365
- ...while LHC experiments are collecting more and more events
- reducing CPU consumption of simulation is very important
- training a GAN on single particle showers of all types and energies
- Then when an event is simulated it would ask for GAN showers on request (superfast by 3-4 order of magnitude)
- Would replace current fast simulation, frozen shower libraries....
- If/when it works, would require large GPU clusters

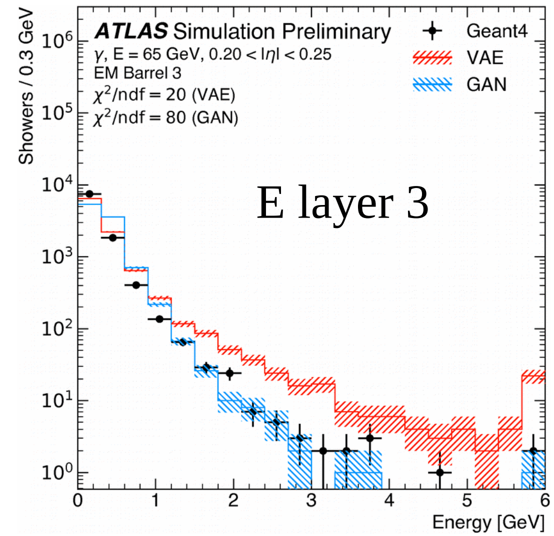
GAN for simulation for ATLAS

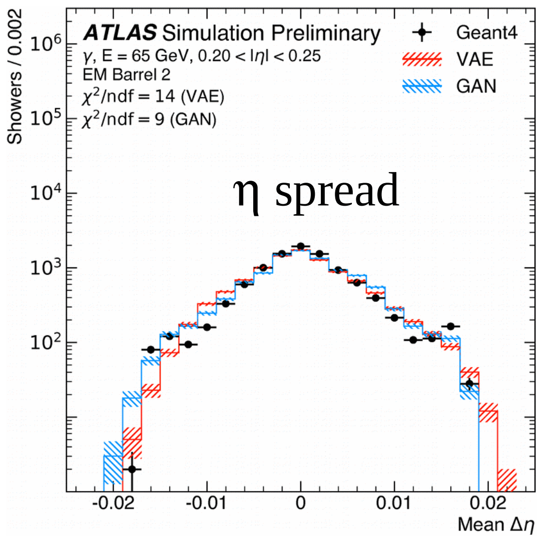
D. Rousseau, A. Ghosh (LAL), G. Louppe (U Liège)

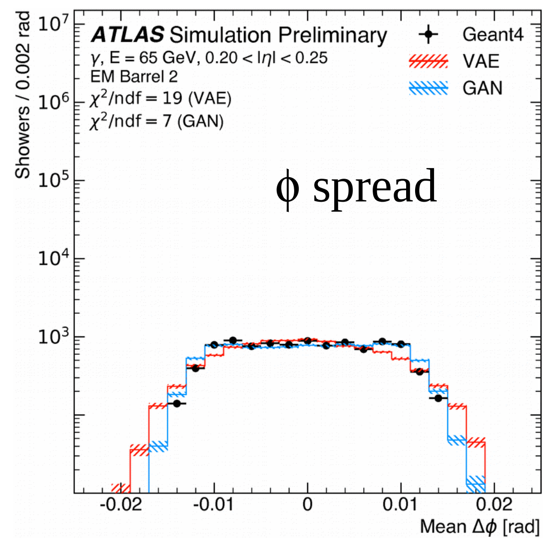
ATL-SOFT-PUB-2018-001 and update ATLAS-SIM-2019-004











Speed: <1ms compared to 10s

Not accurate enough yet in all corners of phase space

Will need much larger network to simulate all parts of the calorimeters

3. Object Reconstruction, Identification, and Calibration

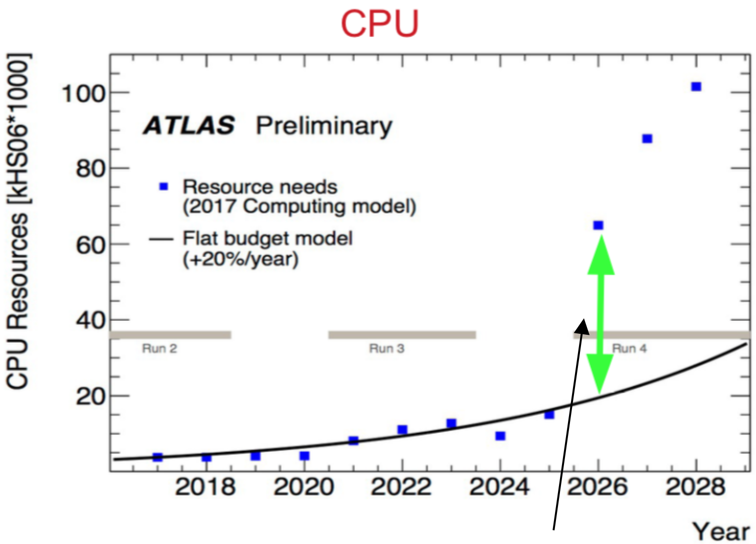
Several contributions:

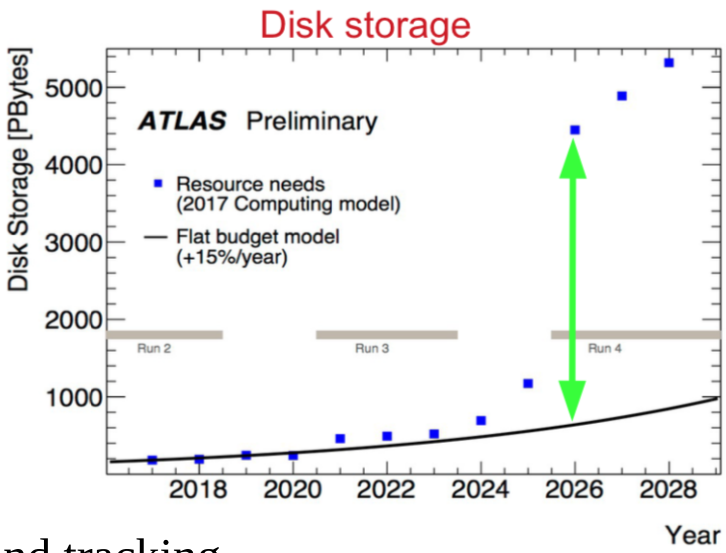
- **Tracking** ML challenge for LHC (LAL)
- **b-tagging** algorithms with BDT’s for ATLAS (CPPM)
- **Particle identification** for LHCb (LPNHE)
- **Position reconstruction** of particles for med app (IMNC)
- Reconstruction **calorimeter** objects with CNN, RNN for LHCb (LAL)
- DNN to optimize **jet reconstruction** using RNN for ATLAS (LPSC)
- RNN for **tau ID** and QCD rejection for CMS (IP2I)
- Reco position, tracking **gamma** for nuclear app. (IP2I)
- Full **Event interpretation** algorithm with DNN, Belle 2 (IPHC)
- DNN for **calo reco** and transfert to FPGA for L1 ATLAS trigger (CPPM)

Color code

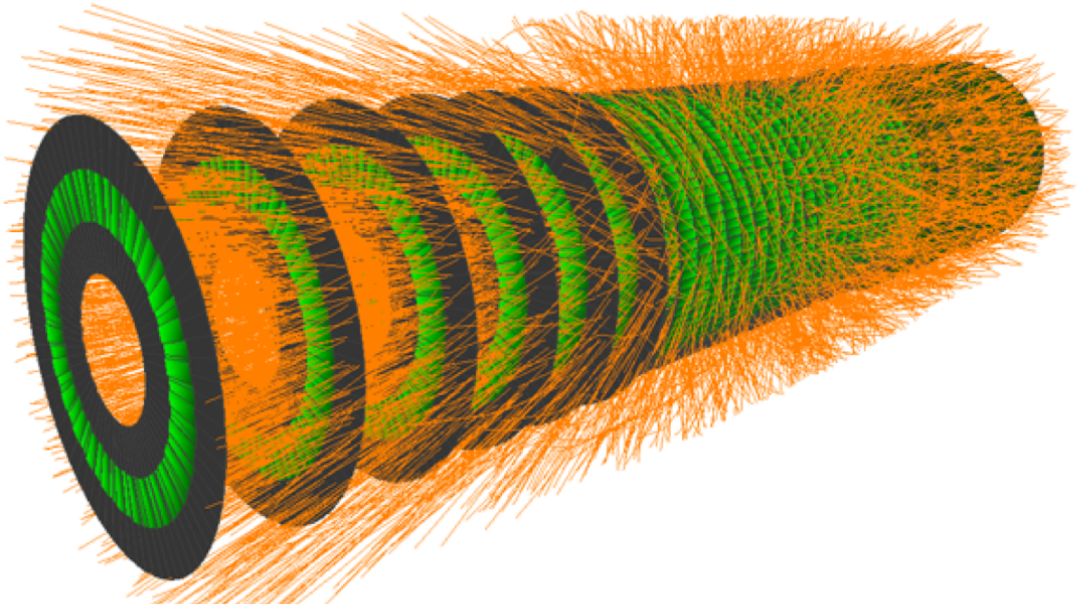
Advanced
Studies
Interest

Resources for High-Lumi LHC





Dominated by : calorimeter simulation and tracking



HL-LHC tracking becoming difficult due to pile-up reaching 200

Track ML challenge

D. Rousseau + many others

TrackML is a competition to expose new algorithms for pattern recognition:

- 3D points are given and participate connect the dots

Links:

- https://sites.google.com/site/trackmlparticle/
- @trackmllhc

Two phases (Accuracy and throughput) → Superfast (0.5s, 1s, compared to state of the art 10-50s) and accurate solutions submitted.

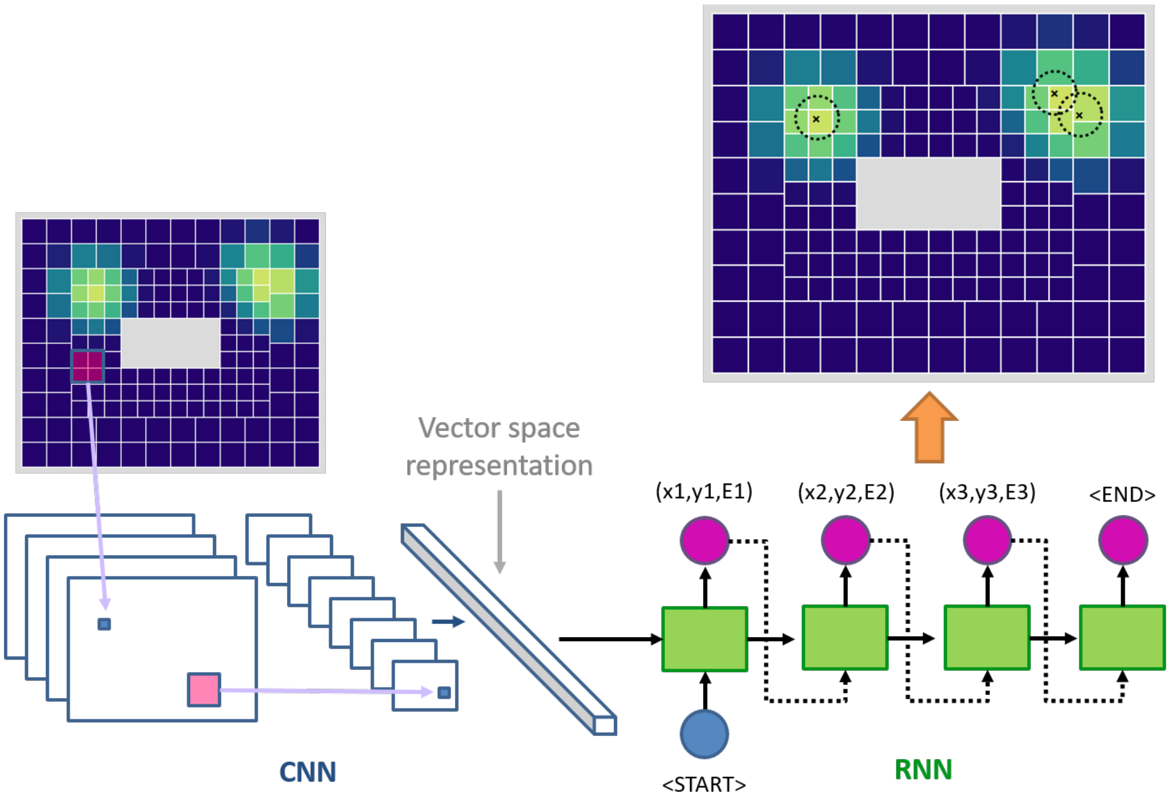
Winners are two experts from the community, **was it worth it** ? Definitely yes:

- The **ML techniques** are now on the table
- The **experts** themselves liked this competition
- The **dataset will be released** on CERN Open Data Portal and serve as a **reference**. Already used for e.g. R&D on Quantum Computing

Calorimeter Reconstruction with DL

Joao Coelho et al. (LAL, LHCb)

- Studying possible Deep Learning solution to shower reconstruction
- Use computer vision techniques (image captioning and object detection)
- Calorimeter hits are processed as an image and encoded in vector space
- Vector representation fed into Neural Networks to output shower candidates



Restricted to a regular detector geometry

Realistic irregular geometry: potentially through Graph Neural Networks (GNN).

5. Uncertainty Assignment

Color code

Advanced
Studies
Interest

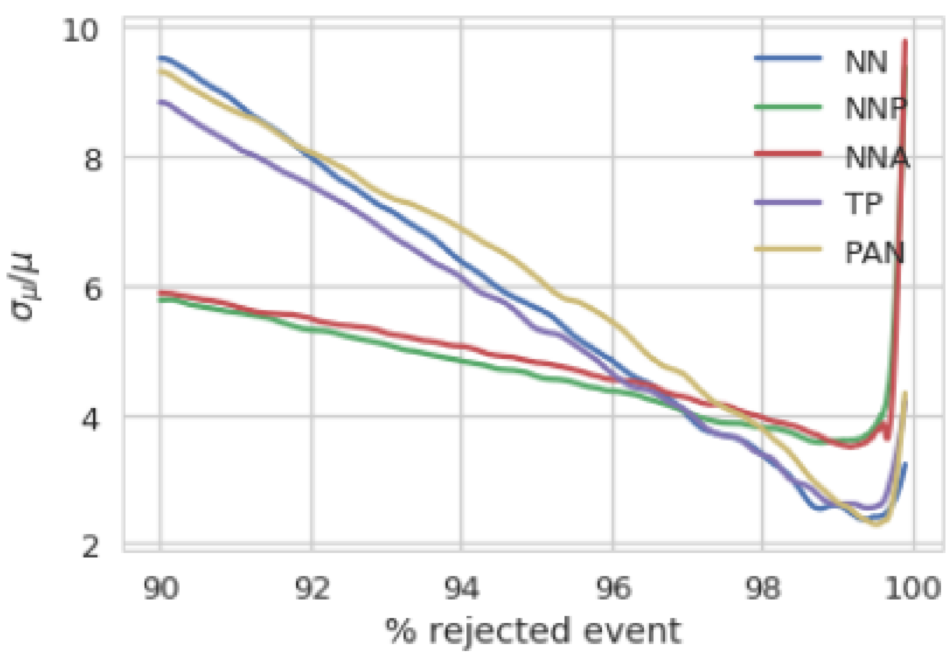
Contributions:

- **Systematic** aware training (LAL)
- ML tools for handling **uncertainties** ATLAS (LPNHE)

Systematic Aware Training

David Rousseau LAL + collaboration Victor Estrade PhD student, Cécile Germain, Isabelle Guyon Laboratoire Recherche Informatique Orsay

- Typical ML classifier (BDT, NN) training is minimizing the *statistical* uncertainty. However *systematic* uncertainty is an important aspect (!)
  → how can an ML classifier take into account a model of systematics at training time, to optimize the ***total*** uncertainty ?
- Several studies done using HiggsML H→tautau public sample
- No clear recommendation yet



Systematics aware learning:
a case study in High Energy Physics
V. Estrade et al.

https://hal.inria.fr/hal-01715155

6. Learning the Standard Model – searches for anomalies

Color code

Advanced
Studies
Interest

Contributions:

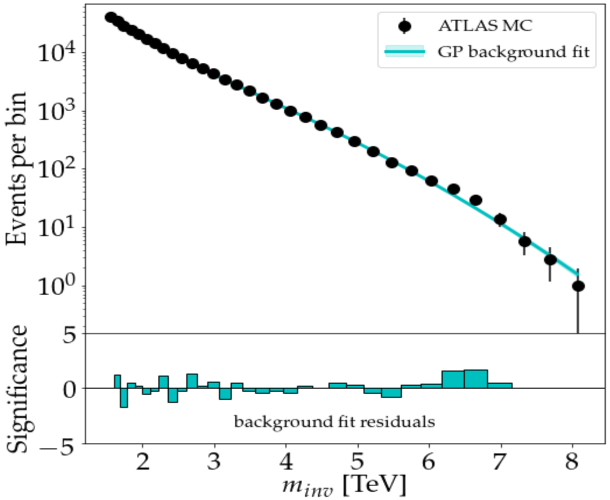
- Search for **anomalies** (LPC)

Searches for anomalies

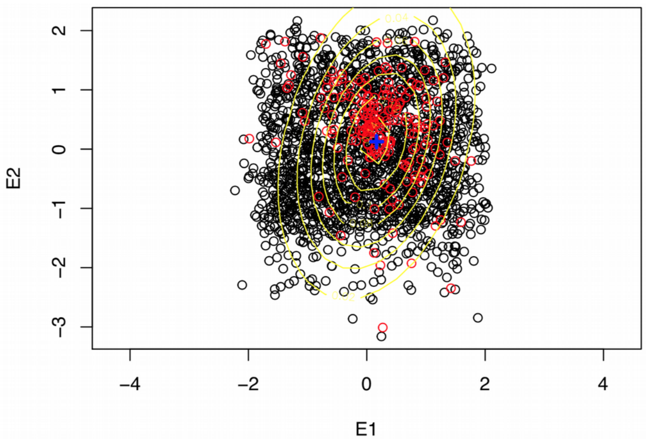
F. Jimenez, L. Vaslin, I. Dinu, JD (LPC) + ITN + LIMOS

Methods for **model independent searches** (ie not relying on a theory in particular)

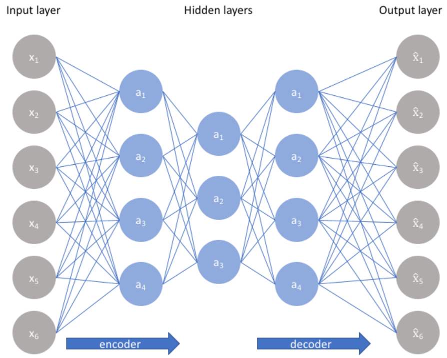
Modeling with Gaussian Processes
- Searches for resonances

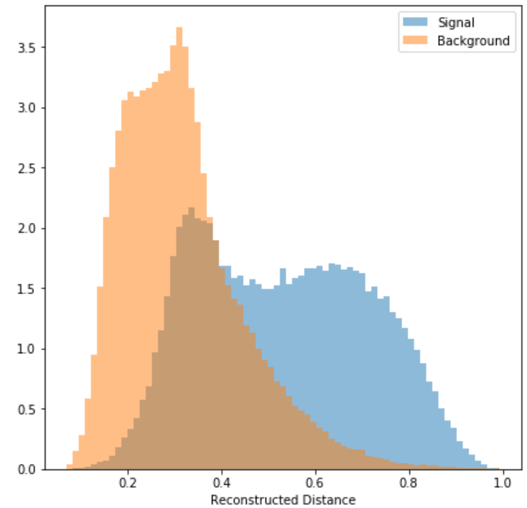


Penalized Anomaly Detection
- Based on Gaussian mixture model
- Multiple dimensions (variable selections)



Learning background model
- ex: with AutoEncoders





ML @ IN2P3

7. Matrix Element Method with ML → **Uncovered ?**

8. Theory Applications

- LPSC: ML activities for HEP **phenomenology** (LPSC)

7. Computing Resource Optimization

- **CCIN2P3**

Color code

Advanced
Studies
Interest

ML software, tools and interfaces

Internal (HEP) tools
- ROOT framework for data storage and processing
- Multivariate Analysis: TMVA for mostly BDT and (deep) NN
- Specific for Neural Networks: NeuroBayes

External tools
- Data format: text, csv, images, HDF5, ...
- ML libraries: Keras+TensorFlow, Pytorch, scikit-learn (no DL), ...
- All kinds of popular algorithms: CNN, GAN, RNN, LSTM, AE, VAE ...

Interfaces and middleware
- PyMVA: Interface TMVA and Keras
- Several middleware file format conversion solutions:

arxiv:1807.02876

PyROOT	Python extension module that allows the user to interact with ROOT data/classes. [69]
root_numpy	The interface between ROOT and NumPy supported by the Scikit-HEP community. [65]
root_pandas	The interface between ROOT and Pandas dataframes supported by the DIANA/HEP project. [70]
uproot	A high throughput I/O interface between ROOT and NumPy. [71]
c2numpy	Pure C-based code to convert ROOT data into Numpy arrays which can be used in C/C++ frameworks. [72]
root4j	The hep.io.root package contains a simple Java interface for reading ROOT files. This tool has been developed based on freehep-rootio. [73]
root2npy	The go-hep package contains a reading ROOT files. This tool has been developed based on freehep-rootio. [73]
root2hdf5	Converts ROOT files containing TTrees into HDF5 files containing HDF5 tables. [74]

Computing and Hardware resources

ML computing @ IN2P3

- Mostly **CPU**, sometime **GPU**, and some attempts with **FPGA**
- **Local** resources: laptop, lab/university clusters
- **CCIN2P3** resources: lots of CPU, less GPU

Any other resources ?

- Tensor Processing Units (**TPU**)
- Vision Processing Units (**VPU**)
- Calculation on **cloud** from industry ?
  - Amazon Web Services machines
  - Google colab notebook with GPU support
  - ...

Computing and Hardware resources

Bogdan Vulpescu (LPC)

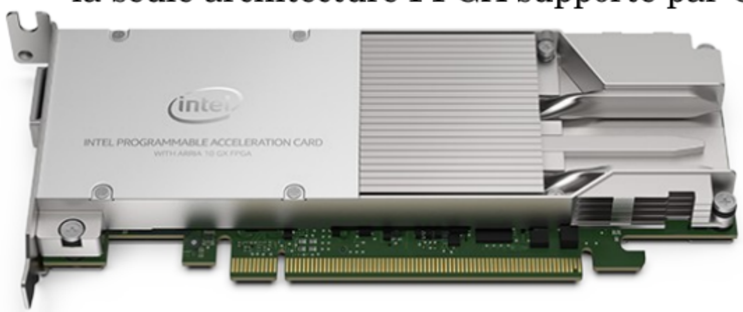
Le toolkit OpenVINO d'Intel

- à partir d'un modèle de **réseau déjà entraîné**
(plusieurs frameworks : TensorFlow, Caffe, etc.)
génère un micro-code pour être exécuté sur une architecture parallèle :
  - CPU (OpenCL)
  - GPU (OpenCL)
  - VPU (processeurs vectoriels et haute granularité de la mémoire, adapté au flux de données)
  - FPGA (OpenCL)
- **exécute l'étape d'inférence !**

Le VPU **NCS2** (Neural Computing Stick 2)

<u>Prospective LPC en 2019</u>

- accélérateur FPGA Intel PAC Arria 10 GX
- la seule architecture FPGA supporté par OpenVINO



Démo TensorFlow :

- type véhicule
- plaque minéralogique

Collaborations with CS/maths

ML collaborations @ IN2P3

- Common project, co-supervision of PhD, post-doc
- Example of **local** collaborations :
  - **LPC** and LIMOS/ISIMA (CS), LMBP (maths)
    - LSST (astronomical time series), ATLAS (anomaly detection), LHCb (bayesian learning)
  - **LPNHE** and Sorbonne (maths): ATLAS (fuzzy number systems)
  - **LAL** and LRI (CS): ATLAS (TrackML, Syst. Aware Training)
  - **CPPM** and LIS (CS): ATLAS (ttH), Cosmology (deep learning)
  - **LAPP** and LISTIC (CS): CTA (deep learning)
  - …
- **International** collaborations: EU-funded **ITN** with non-academics partners, ...

Obvious advantage in collaborating with ML experts but some caveats:

- Speaking same **language** & getting familiar with vast stat **literature**
- Question of access to **confidential** experimental data and **authorship**
- **Publication** in journal of CS/math field
- Produce outcome **relevant** to collaborator

Training and schools

Being able to apply ML to practical problems requires understanding underlying statistical concepts and ML algorithms.

- **Target**: students (Master, PhD), staff IN2P3

Training courses exist in several universities / labs

- In general Master degree level some also open to staff for continuous training
  - Ex: Diplome Universitaire Data Scientist
- Training CNRS formation entreprise
  - Ex: Introduction to ML and Deep learning

Schools / workshops

- IN2P3 School of Statistics (organized every 2 years since 2008)
- Workshop CCIN2P3: GPU and deep learning

Uncovered needs should trigger specific training actions.

To encourage access to these training courses it would be beneficial to identify and list the existing ones within a catalog, and to have them included in the "plan de formation" of CNRS.

Conclusions

Usage of “traditional” **ML** since many years within IN2P3

More recently moved to **modern software and algorithms**

Expertise from (local) CS/statistician is available and valuable

Publishing with them is also essential

Lots of potential **opportunities** with these new approaches

Gain needs to be well assessed (i.e is it worth w.r.t simpler approaches ?)

Techniques can be deployed in many (other) **sectors**

Threats: scalability and optimization, integration to experimental software

Training: important to list offers and survey needs

Announcements

Please register to machine-learning-l@in2p3.fr

Foreseen event :

→ IN2P3+CEA ML HEP workshop at CCIN2P3

2 days end january/early february 2020

ML@IN2P3: received contributions

Black: expression of interest
Blue: ongoing studies
Red: preliminary results

Detector design
- LPNHE: use ML to optimize detector design

Simulation
- LPNHE: MC sample reweighting in ATLAS
- LAL: simulation of ATLAS calo with GAN's
- LAL: BDT's for multidim reweighting between MC (ATLAS)
- LAL: GP to smooth MC stat fluctuations (ATLAS)
- IPNO: NN to simulate fuel evolution in nuclear reactors

Real time analysis
- LPNHE: real time ML (LHCb)

Uncertainties
- LPNHE: ML tools for handling uncertainties ATLAS, collab with Sorbonne maths
- LAL: systematic aware training

Object Reconstruction, Identification, and Calibration
- IMNC: position reconstruction of particles for med app
- IP2I: RNN for tau ID and QCD rejection for CMS
- IP2I: reco position, tracking gamma for nuclear app. (LSTM)
- LAL: track ML challenge (LHC)
- LAL: reco calo objects with CNN, RNN (LHCb)
- IPHC: Full Event interepretation algorithm with DNN (Belle 2)
- CPPM: b-tagging algorithms with BDT's (ATLAS)
- CPPM: DNN for calo reco and transfert to FPGA for L1 trigger (ATLAS)
- LPSC: DNN to optimize jet reconstruction using RNN (ATLAS)
- LPNHE: particle identification (LHCb)

ML for Accelerator developments
- LAL: accelerator automatic tuning, HI Laster diagnostics, virtual detectors, laser laserix
- LPSC: NN for particle accelerator operations and optimization.

Data analysis
- Lots of expertise from different (LHC) groups with BDT and MLP NN mostly
- CPPM: usage of RNN (ATLAS), collab with LIS
- LPC: anomaly detection using AE/VAE, background modeling using GP (ATLAS), collab with LIMOS

Theory
- LPSC: ML activities for HEP phenomenology: fast xs calculator, classification for NP models, recasting

Hardware and ML
- LPC: Bodgan

A Bibliography: ML in HEP

Reviews/guides

Machine Learning in High Energy Physics Community White Paper, https://arxiv.org/abs/1807.02876

Deep Learning and its Application to LHC Physics, https://arxiv.org/abs/1806.11484

Supervised deep learning in high energy phenomenology: a mini review, https://arxiv.org/abs/1905.06047

A guide for deploying Deep Learning in LHC searches: How to achieve optimality and account for uncertainty, https://arxiv.org/abs/1909.03081

Machine learning and the physical sciences, https://arxiv.org/abs/1903.10563

GAN

How to GAN LHC Events, https://arxiv.org/abs/1907.03764

Machine Learning Templates for QCD Factorization in the Search for Physics Beyond the Standard Model, https://arxiv.org/abs/1903.02556

DijetGAN: A Generative-Adversarial Network Approach for the Simulation of QCD Dijet Events at the LHC, https://arxiv.org/abs/1903.02433

MEM

Effective LHC measurements with matrix elements and machine learning, https://arxiv.org/abs/1906.01578

AE/VAE

Variational Autoencoders for New Physics Mining at the Large Hadron Collider, https://arxiv.org/abs/1811.10276

A robust anomaly finder based on autoencoder, https://arxiv.org/abs/1903.02032

Novelty Detection Meets Collider Physics, https://arxiv.org/abs/1807.10261

Bump hunt

Extending the Bump Hunt with Machine Learning, https://arxiv.org/abs/1902.02634

Other

Machine Learning Pipelines with Modern Big Data Tools for High Energy Physics, https://arxiv.org/abs/1909.10389

The Metric Space of Collider Events, https://arxiv.org/abs/1902.02346