CHEP Track VI

Summary

Sergei Gleyzer

July 12, 2018

Tk6 Conveners

- Sergei Gleyzer

- Michela Paganini

- Andrea Rizzi

- Sofia Vallecorsa

Keywords

- Machine Learning
- Physics Analysis
- Data Preservation

Statistics

- 47 presentations
- 14 posters

70% of contributions related to machine learning

Presentation Topics

- Machine Learning
  - Physics Analysis, Trigger, Monitoring, Reconstruction, Event and Particle Classification, Simulation, Computing/Data Management, Tools
- Non-ML
  - Analysis Tools and Formats, Data and Analysis Preservation, Fitting, GPUs

Experiments

+ Theory/Phenomenology

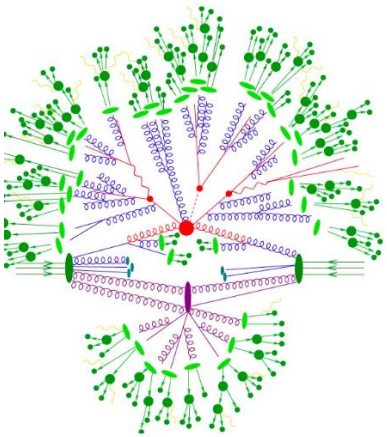
Trends

- All varieties of deep learning gaining traction
  - Convolutional, Recurrent, LSTM, GANs
  - Tree-based methods (XGBoost) still maintain some competitiveness
- Growing python ecosystem
- Machine learning models increasingly used together with low-level information

End-to-End Learning

Can we fully exploit the detectors:

- Raw data, low-level variables



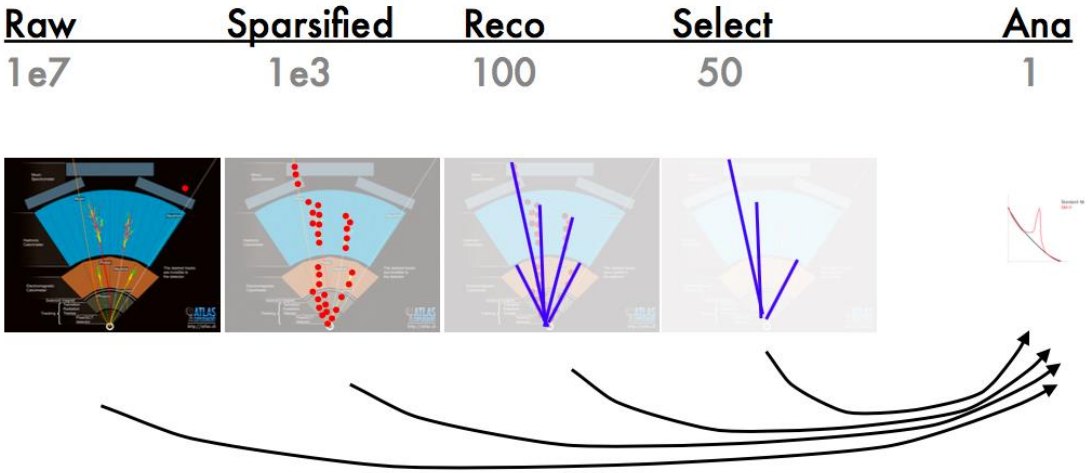


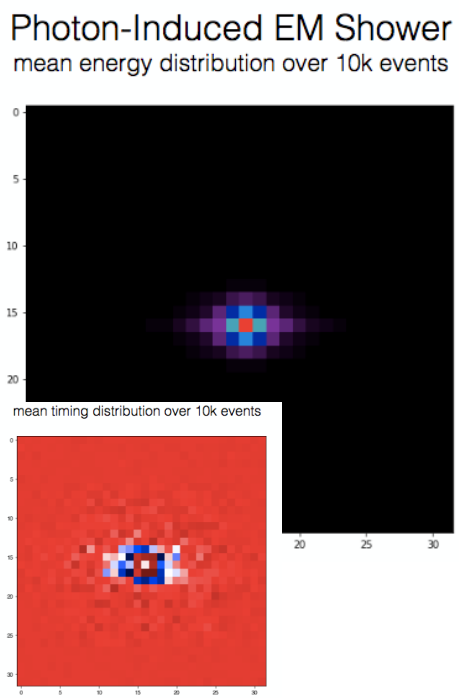
Image credit: K. Cranmer

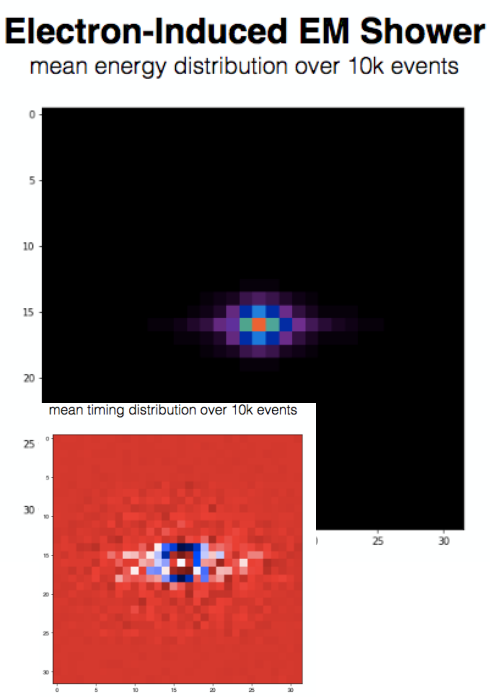
End-to-End Learning

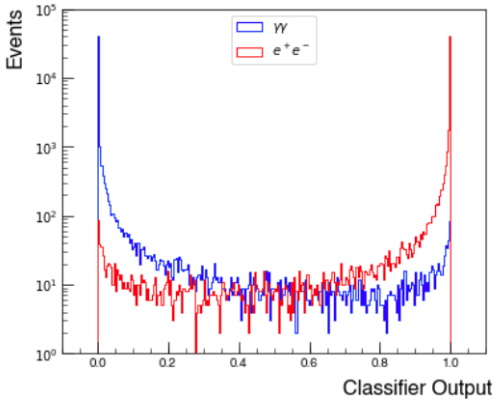
“Particle and event ID”

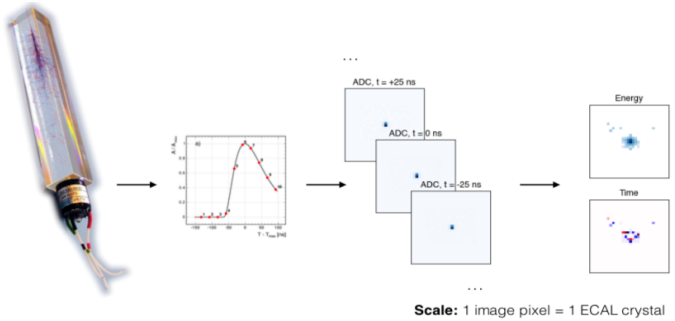
M. Andrews

- Convolutional Neural Networks





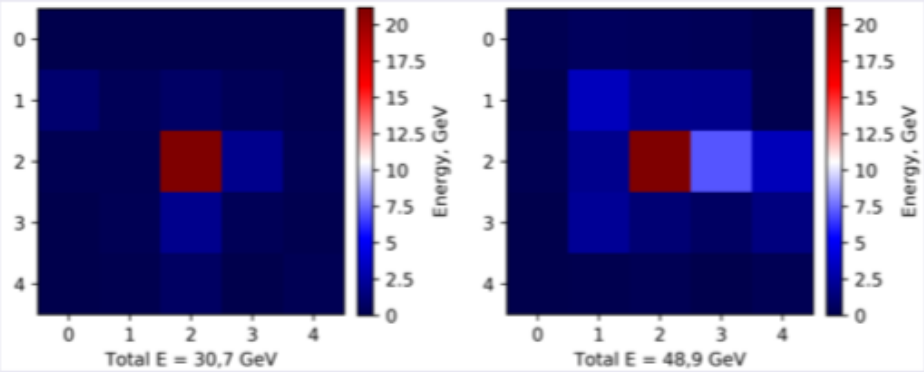




Particle ID

V. Chekalina

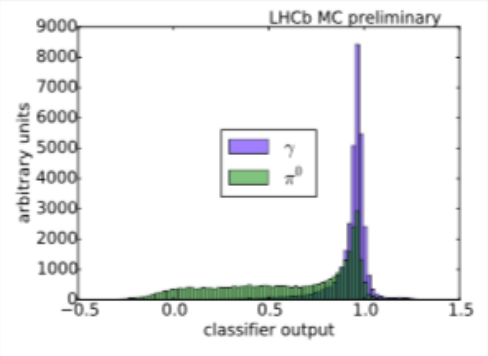
Clusters for photon and π^0photon



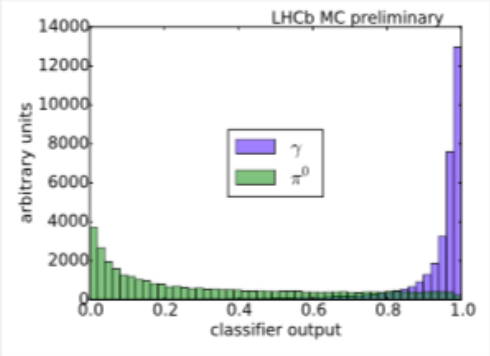
Responses from single photon(left) and merged π^0(right)

New approach

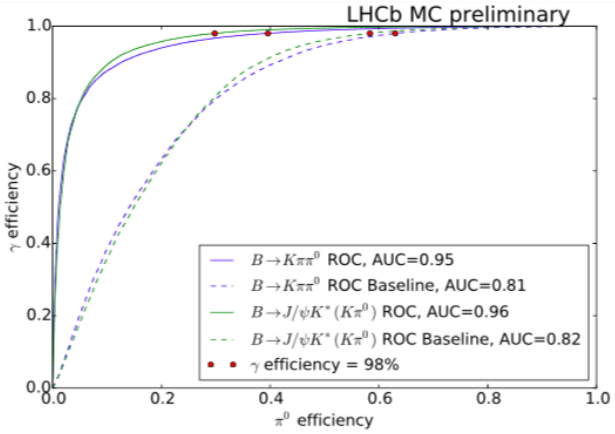
- Consider 5*5 cluster
- Use energy in each cell as a feature
- Use several models and look for the best one



Baseline output



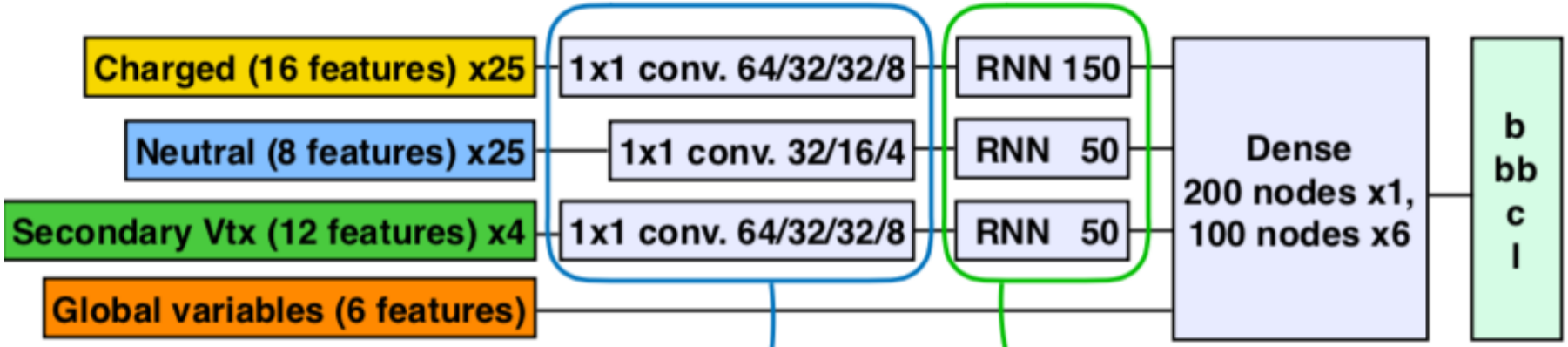
XGBoost approach output



Flavor Tagging

Particle-based NN architecture

M. Verzetti

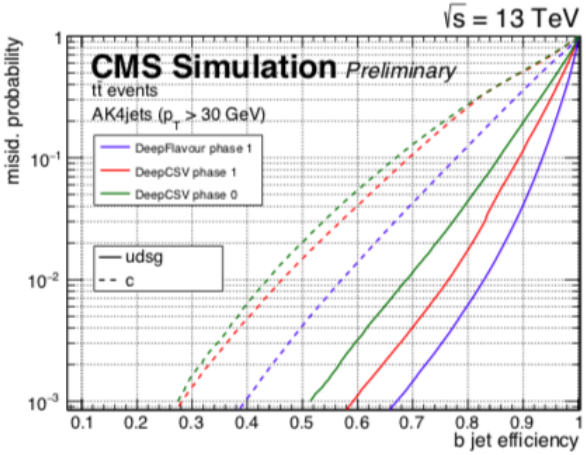


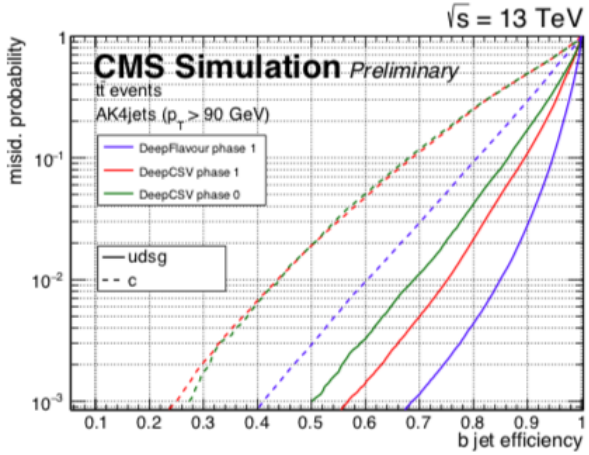
Convolutional layers progressively learn a more compact feature representation (automatic feature engineering)

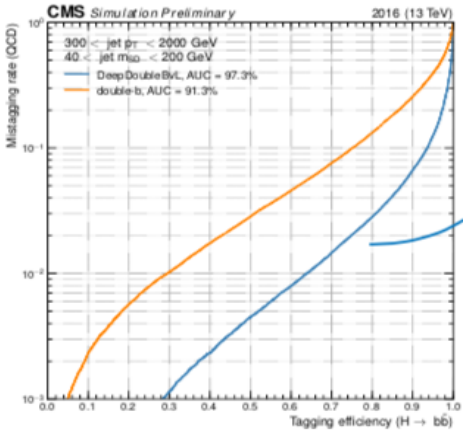
The recurrent layers (LSTM) builds a “summary” of the information contained in each set of feature types

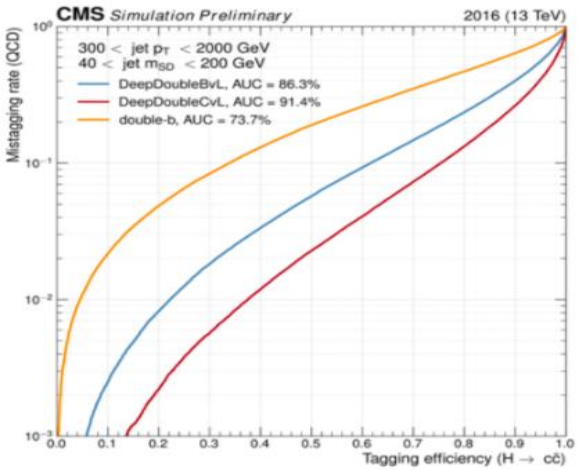
Flavor Tagging

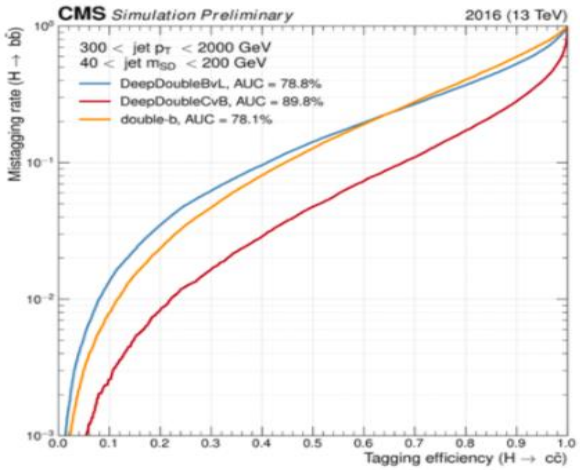
M. Verzetti

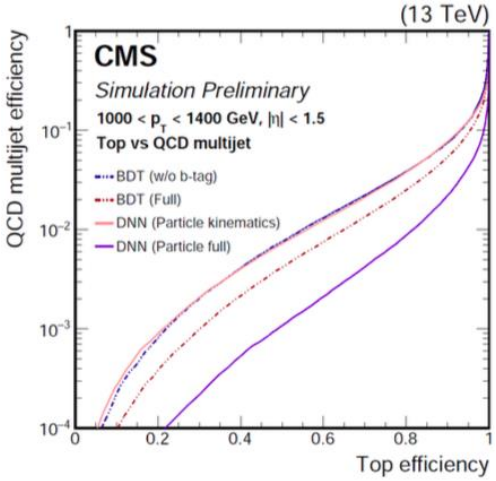




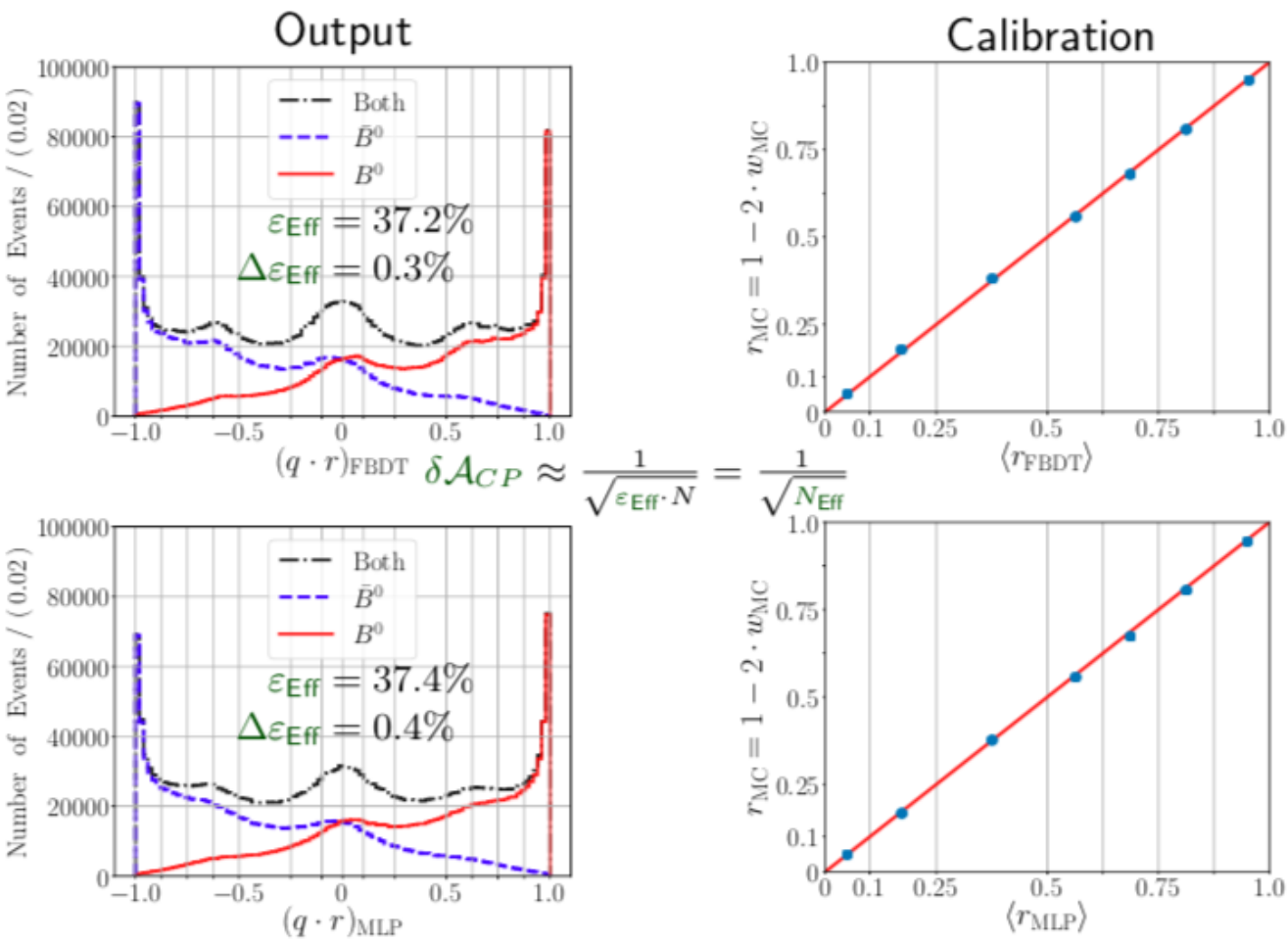








Flavor Tagging



$$\delta\mathcal{A}_{CP} \approx \frac{1}{\sqrt{\varepsilon_{\text{Eff}} \cdot N}} = \frac{1}{\sqrt{N_{\text{Eff}}}}$$

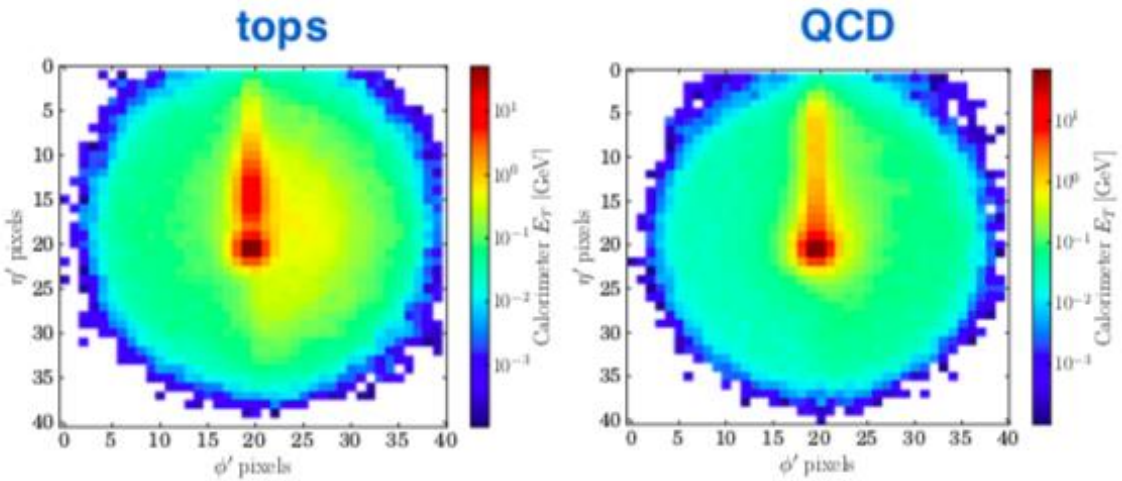
F. Abudinen

Jet Images

Can recent advances in DNNs benefit jet physics?

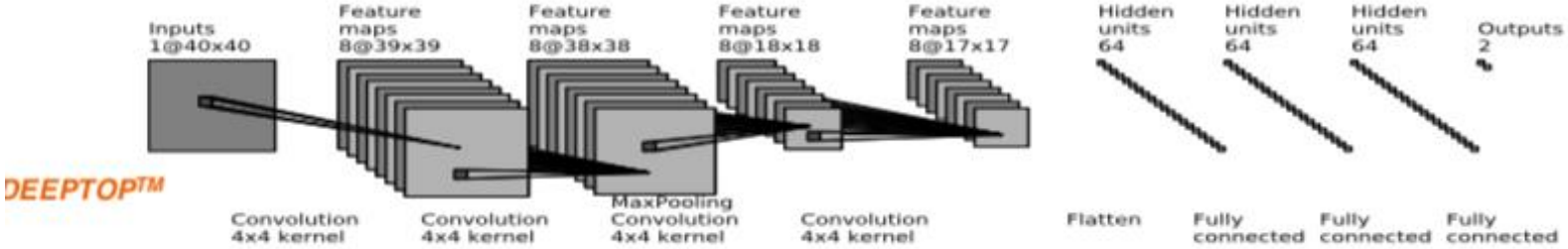
Kasieczka, Plehn, MR, Schell '17

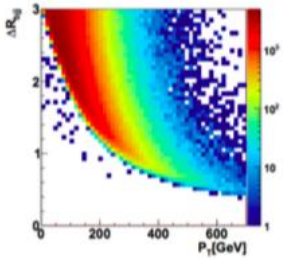
- View calorimeter plane as 2-d "image" with energy deposits as pixels
- After some pre-processing, train a *convolutional neural network* (no details here) on sample of top jets and QCD background



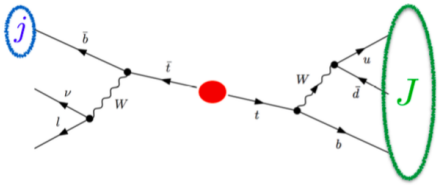
- Last layer of network converts weights for each image into probability of it being either top or QCD

M. Russell





LoLa



Butter, Kasieczka, Plehn, MR '17

Beyond images: LoLa

Why not use the jet constituent 4-vectors directly?
Two ingredients:

1. CoLa* - learns the jet clustering history

$$k_{\mu,i} \xrightarrow{\text{CoLa}} \tilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij}$$

$$C = \begin{pmatrix} 1 & 0 & \cdots & 0 & C_{1,N+2} & \cdots & C_{1,M} \\ 0 & 1 & & \vdots & C_{2,N+2} & \cdots & C_{2,M} \\ \vdots & \vdots & \ddots & 0 & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & C_{N,N+2} & \cdots & C_{N,M} \end{pmatrix}$$

- Test on-shell conditions

$$\tilde{k}^2_{\mu,1} = (k_{\mu,1} + k_{\mu,2} + k_{\mu,3})^2 = m_t^2$$
$$\tilde{k}^2_{\mu,2} = (k_{\mu,1} + k_{\mu,2})^2 = m_W^2 \,.$$

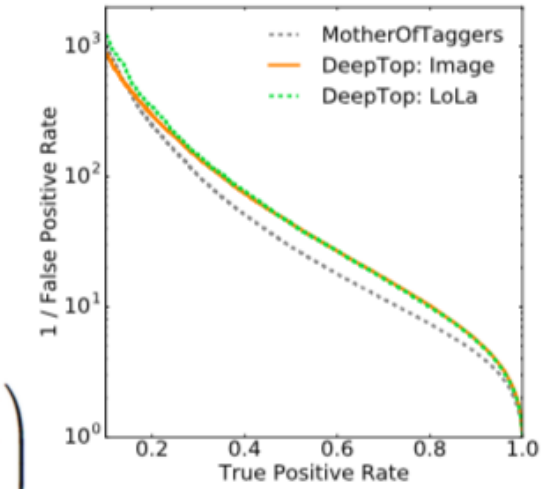
2. LoLa** - learns the kinematics

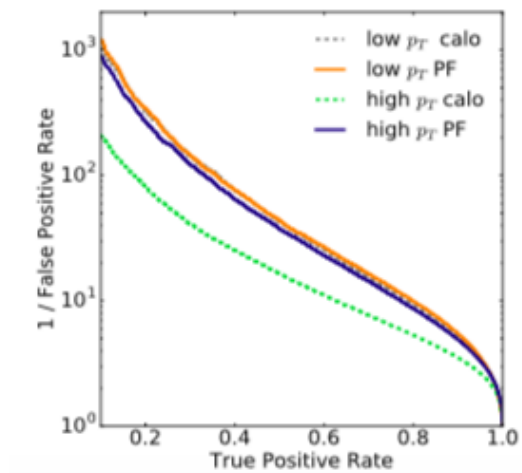
$$\tilde{k}_j \xrightarrow{\text{LoLa}} \hat{k}_j = \begin{pmatrix} m^2(\tilde{k}_j) \\ p_T(\tilde{k}_j) \\ w^{(E)}_{jm}\, E(\tilde{k}_m) \\ w^{(d)}_{jm}\, d^2_{jm} \end{pmatrix}$$

transform 4-vectors into: invariant mass, pT, energy and Minkowski distance
effectively a rotation in observable space

M. Russell

** CoLa = Combination Layer*
*** LoLa = Lorentz Layer*





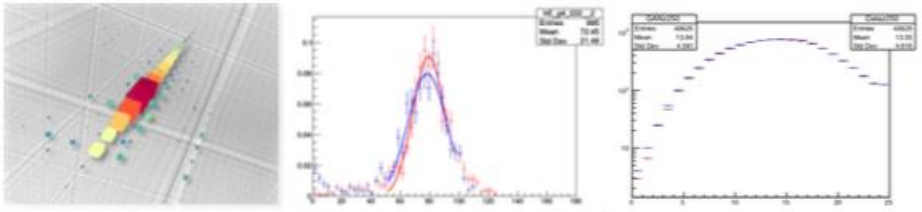
Fast Simulation

Event Simulation

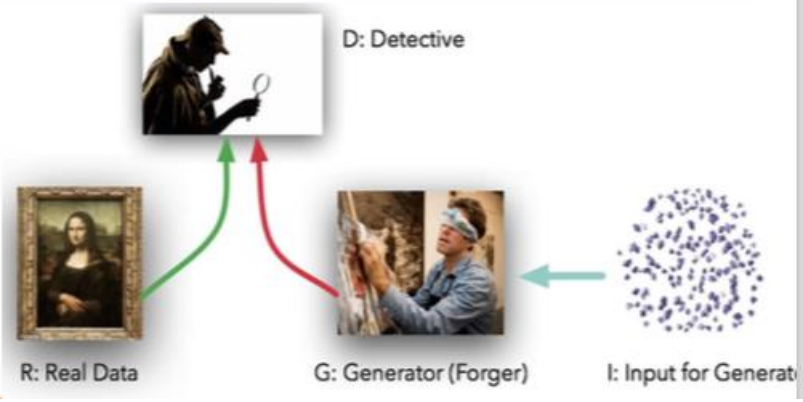
- Simulation is one of the most resource-intensive computing applications.
- Main R&D areas:
  - Adapting the existing code to new computing architectures
  - Replacing complex algorithms with deep-learning approaches (FAST SIMULATION)

Looking at generative adversarial networks to improve speed, without giving up accuracy of simulated events

- One network attempts to simulate events that match a data distribution (Generator G)
- While a second network tries to distinguish data and simulation (Discriminator D)



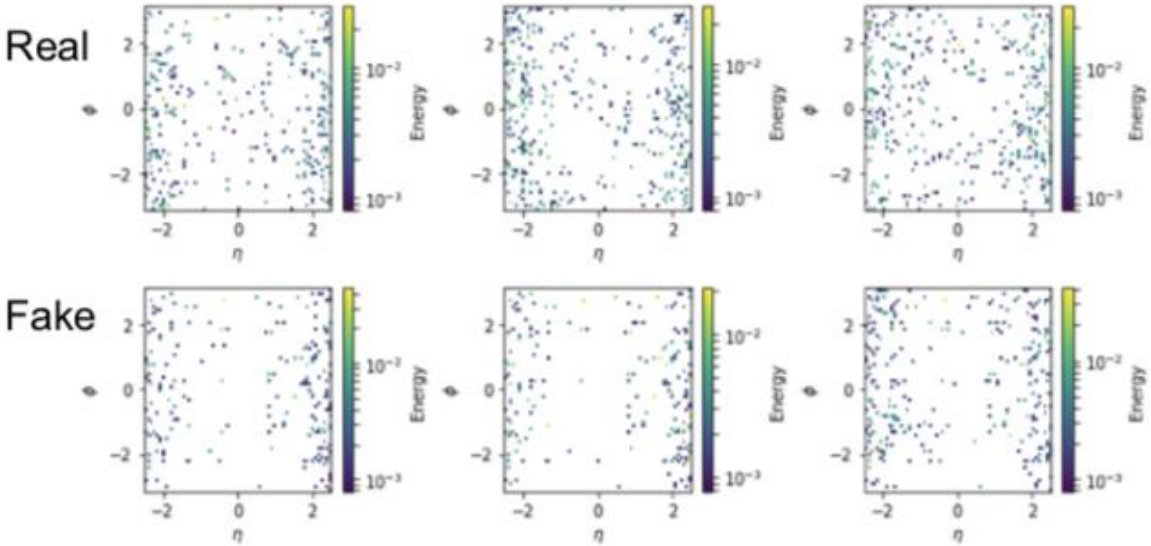
S. Vallecorsa: A machine learning tool for fast simulation

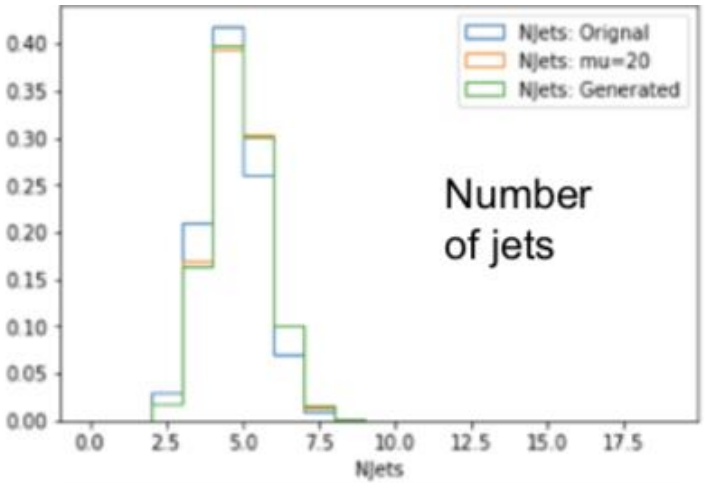


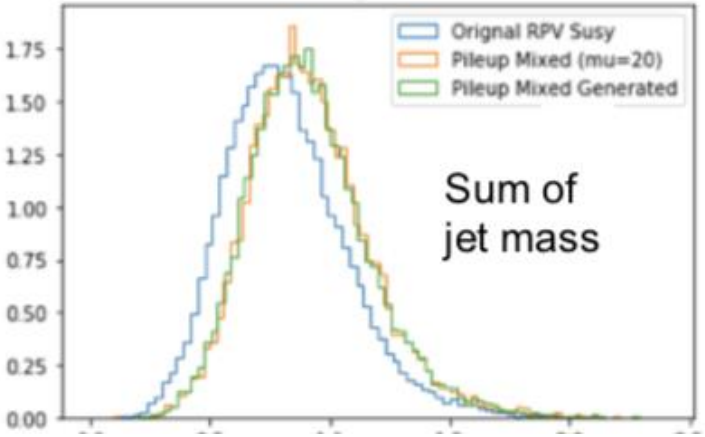
M. Girone

Maria Girone, CERN openlab CTO

Pileup GAN





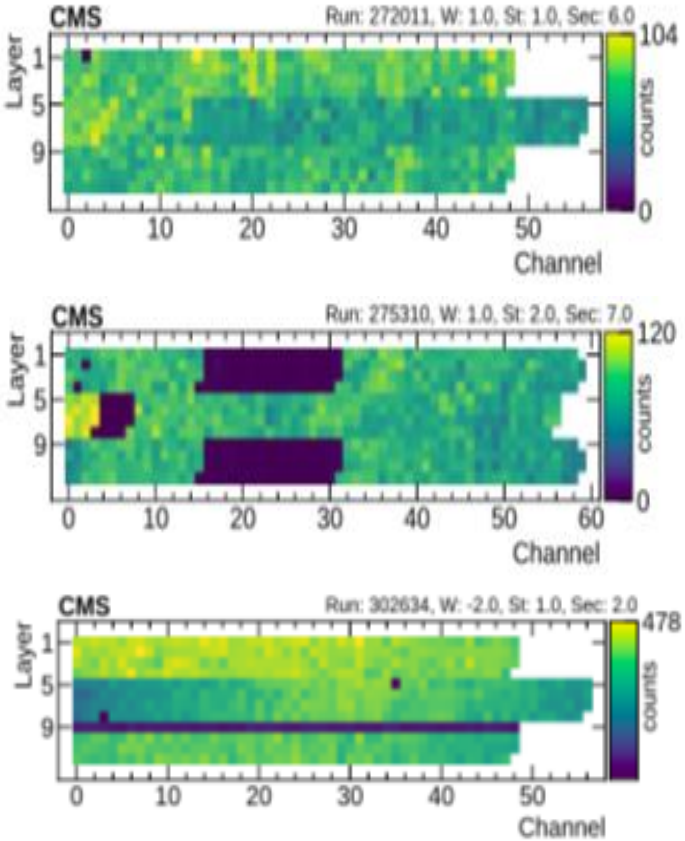


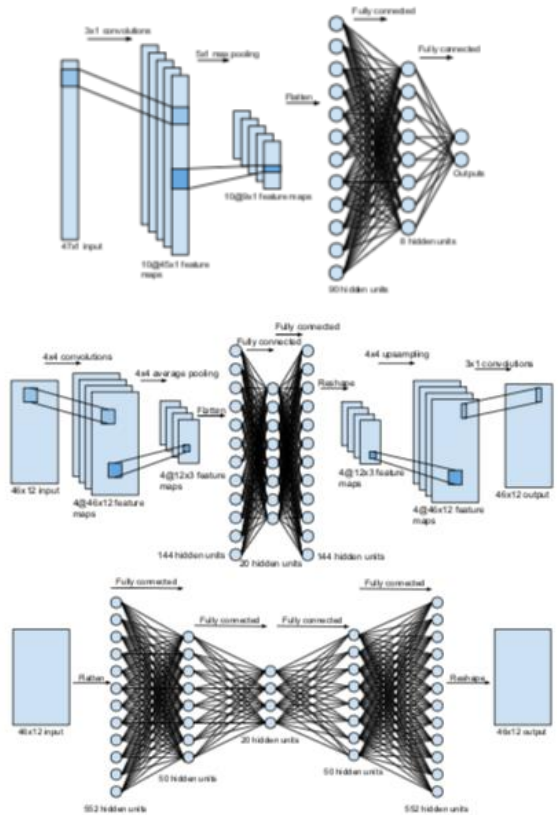
- **The GAN gives realistic looking pileup images**
- **When overlayed onto RPV events, we see realistic shifts in the distributions**

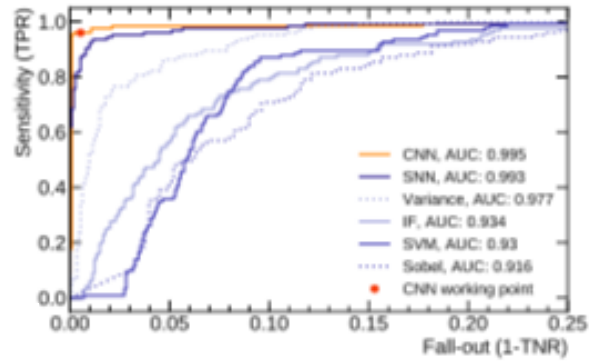
S. Farell

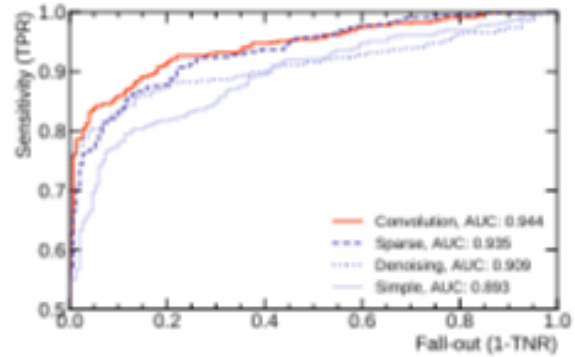
Online DQM

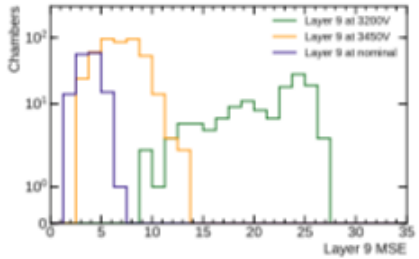
A. Pol

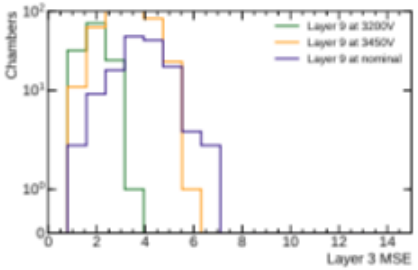






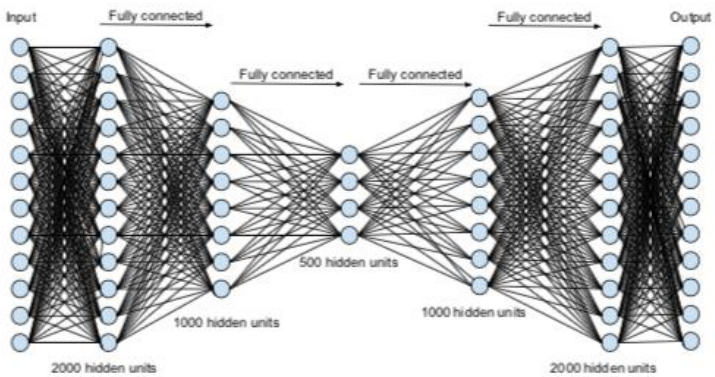




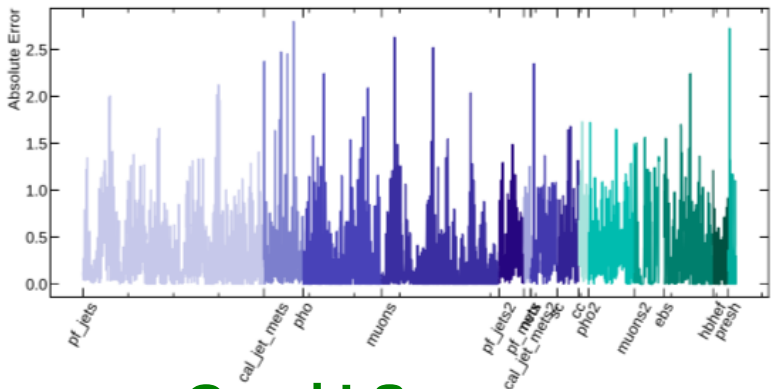


Offline DQM

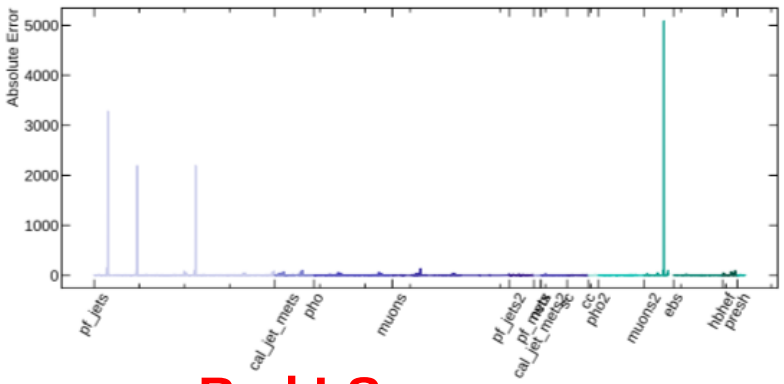
AutoEncoder



A. Pol

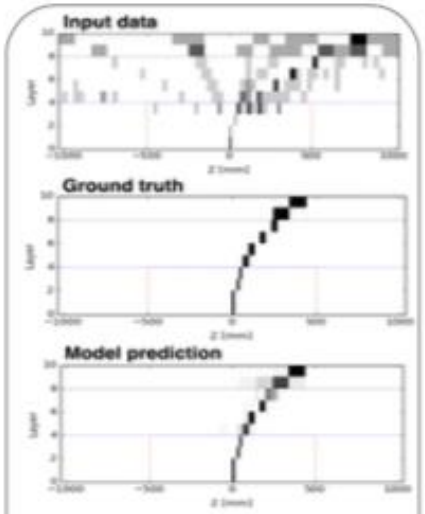


Good LS

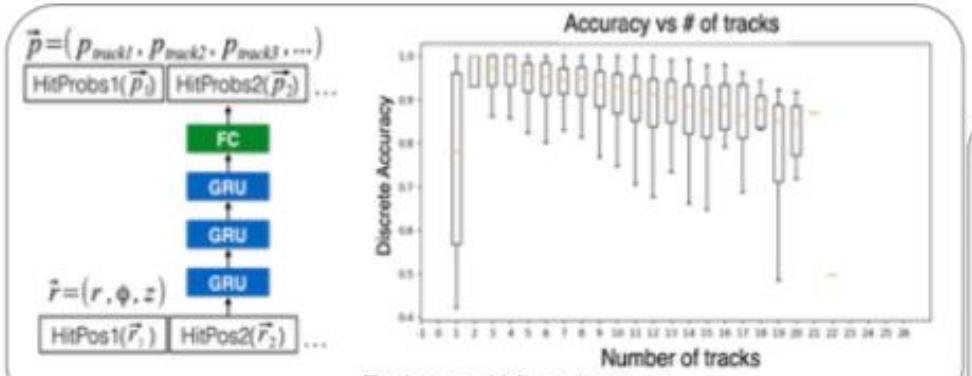


Bad LS

Tracking

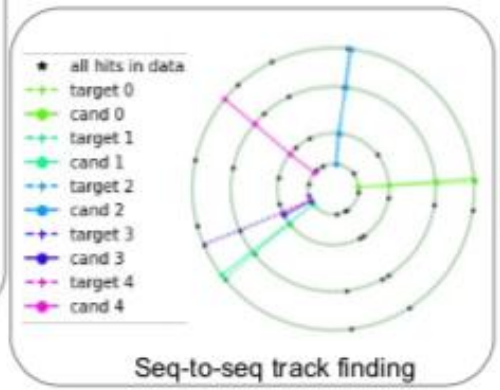


Track following with RNN



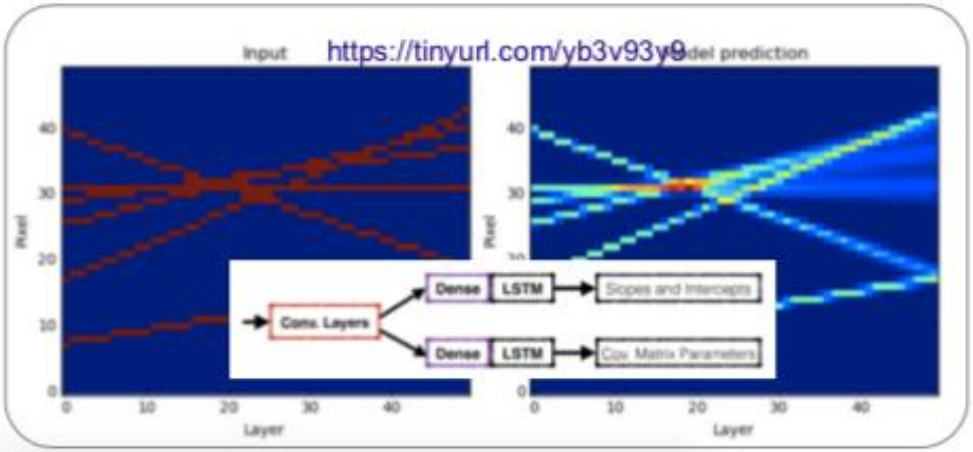
End-to-end hit assignment

JR Vlimant

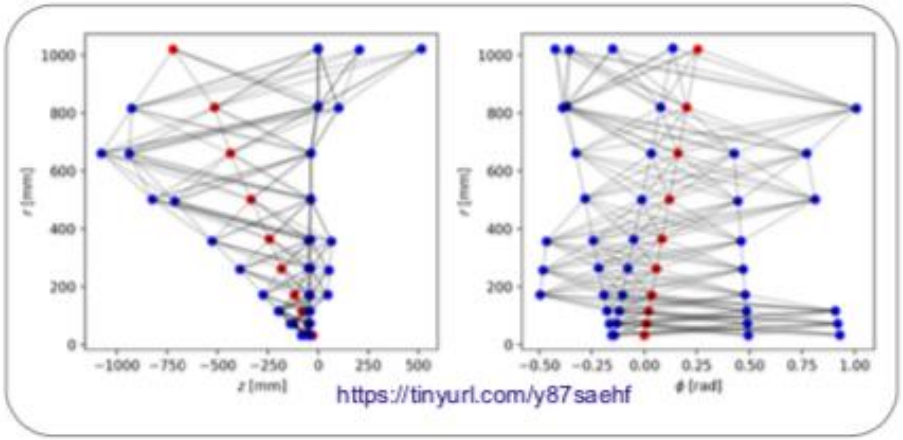


Seq-to-seq track finding

https://heptrkx.github.io/



https://tinyurl.com/yb3v93y9



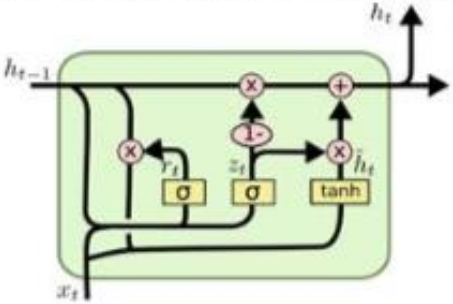
https://tinyurl.com/y87saehf

Two-step tracking

Our last solution - two step tracking procedure:

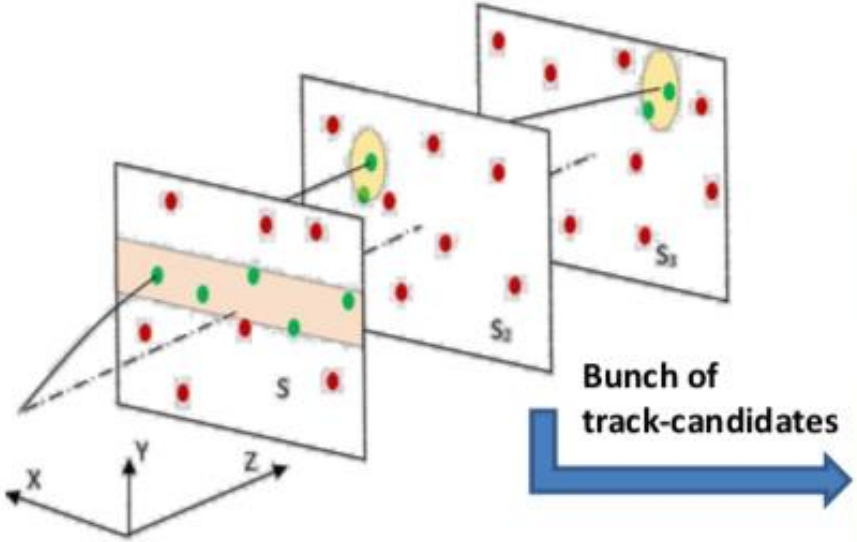
1. **Preprocessing by directed K-d tree search** to find all possible track-candidates as clusters joining all hits from adjacent GEM stations lying on a **smooth curve**.
2. **Deep recurrent network** trained on the big simulated dataset with 82 677 real tracks and 695 887 ghosts **classifies track-candidates in two groups: true tracks and ghosts**.

Gated recurrent unit (GRU) is a simplified version of LSTM networks

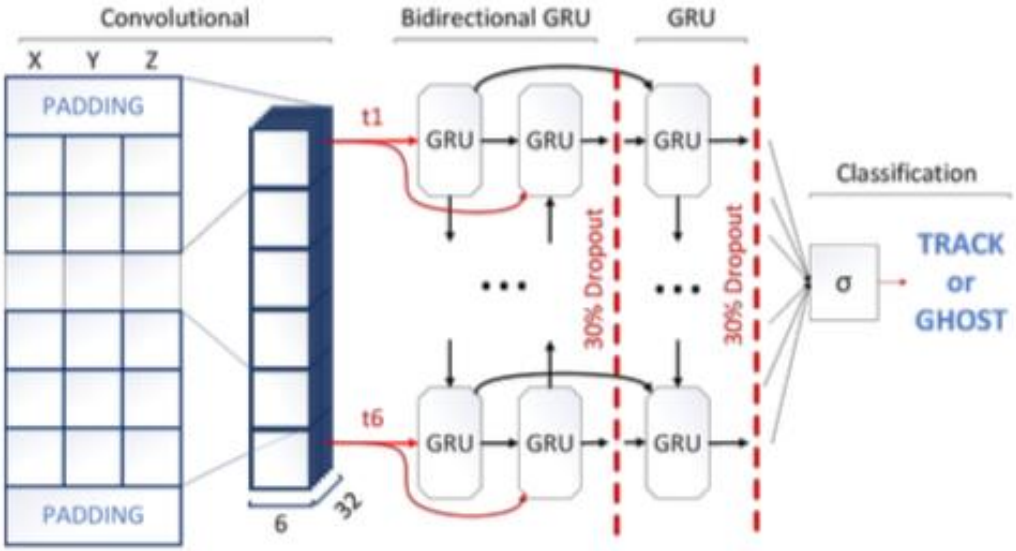


GRU with 3 layers is able to write or forget information by gates with a trainable degree of selectivity to operate on problems going through time

1) Directed K-d Tree Search

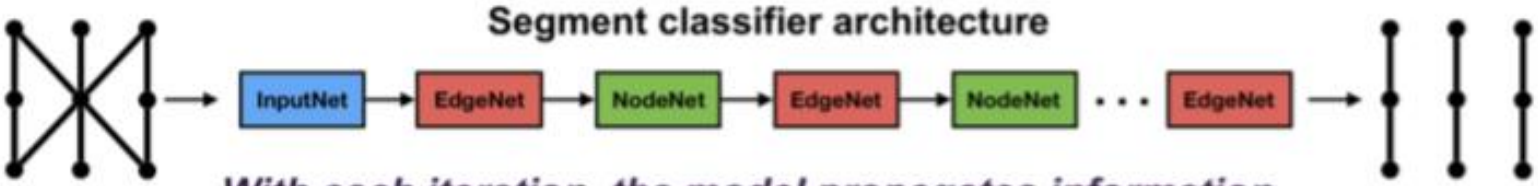


2) Deep Recurrent Neural Network Classifier

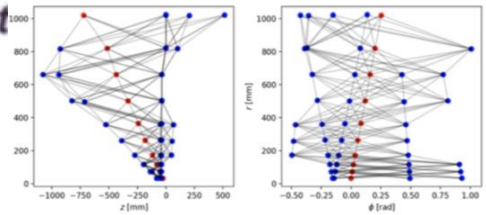


G. Ososkov

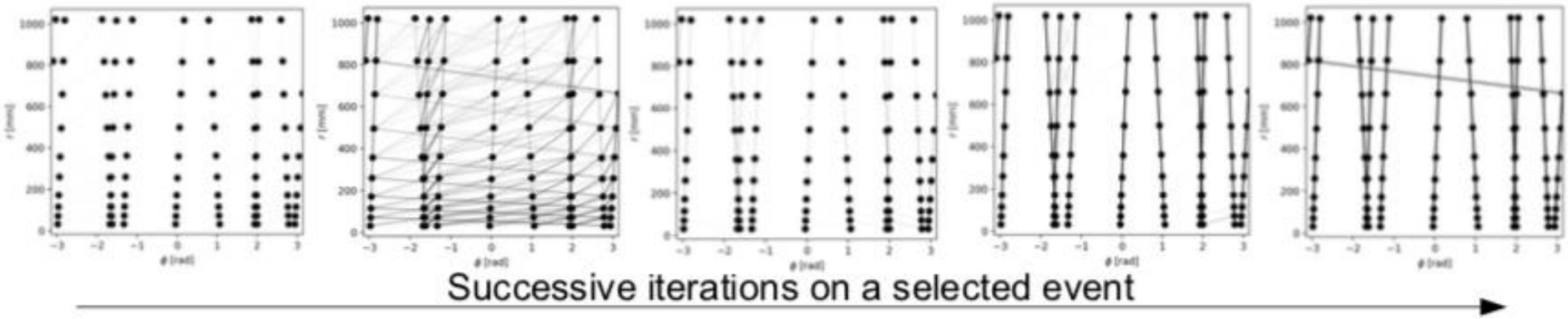
Track Building With GNN



With each iteration, the model propagates information through the graph, strengthens important connect and weakens useless ones.

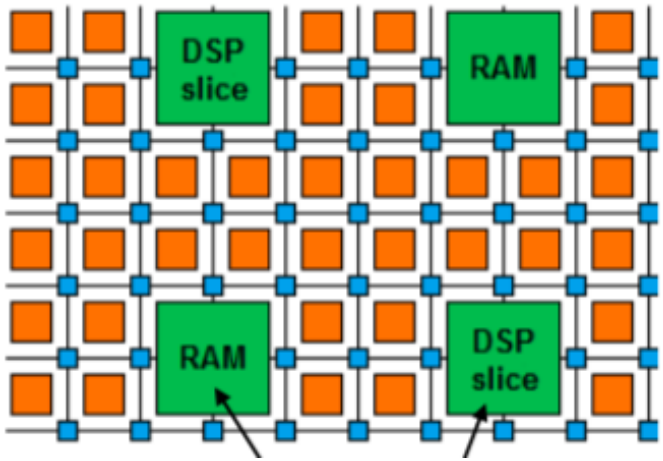


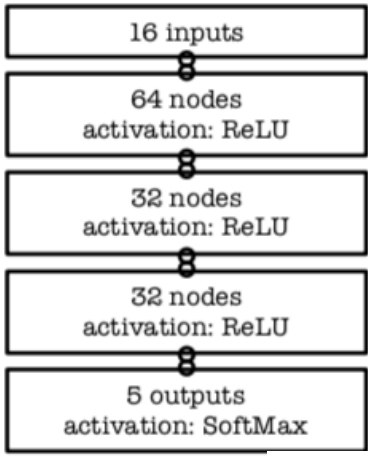
- Unseeded hit-pair classification
- Model predicts the probability that a hit-pair is valid

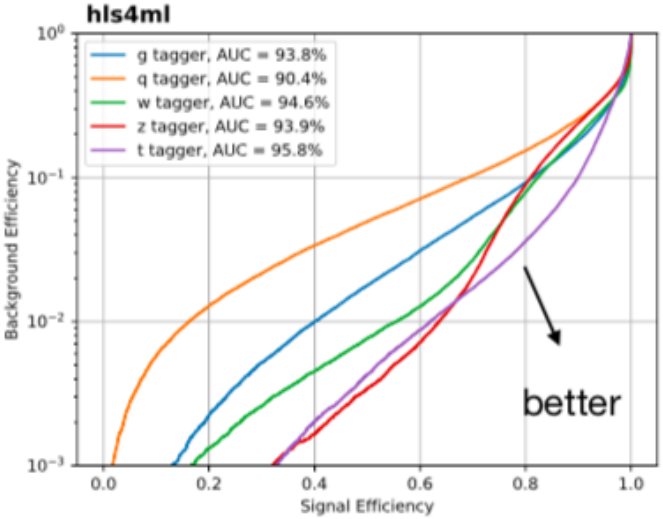


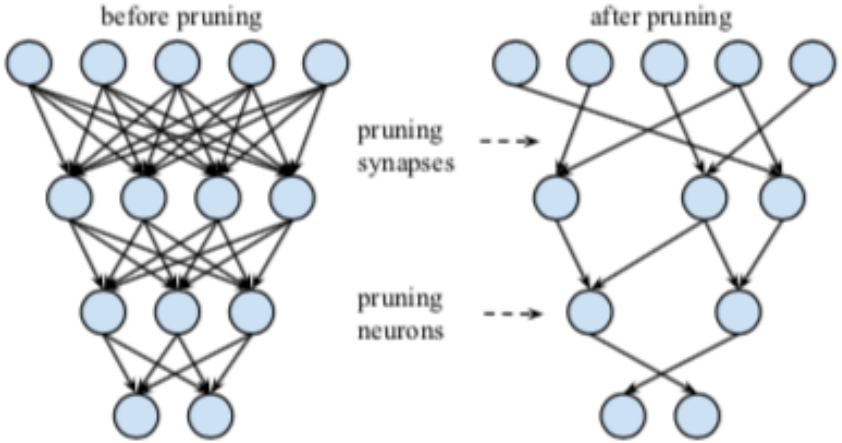
JR Vlimant

ML on FPGA

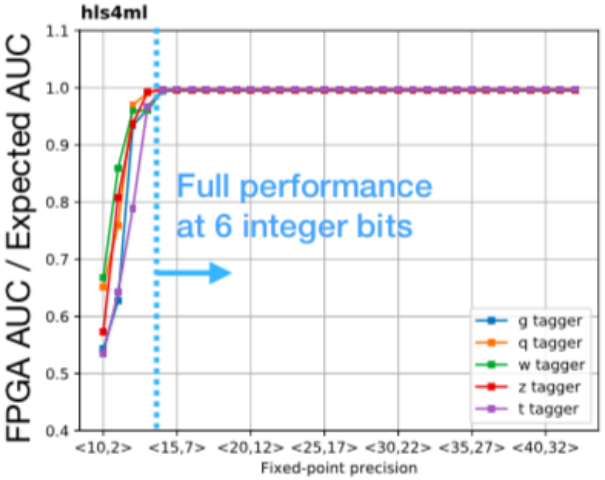








Fractional bits fixed to 8

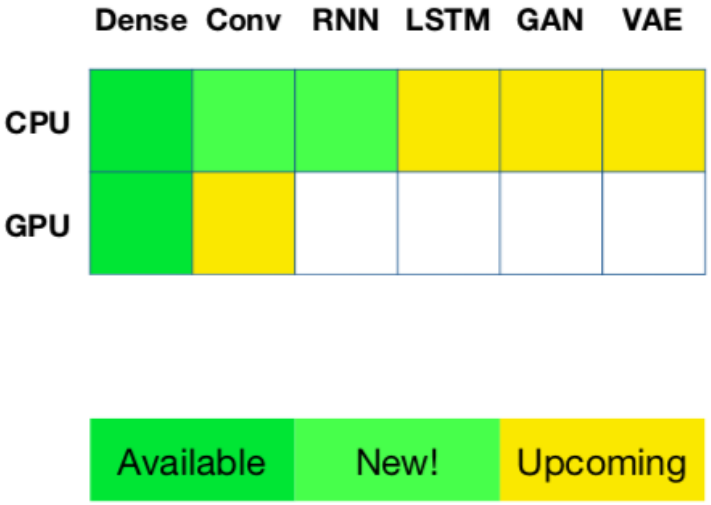


J. Ngadiuba

TMVA

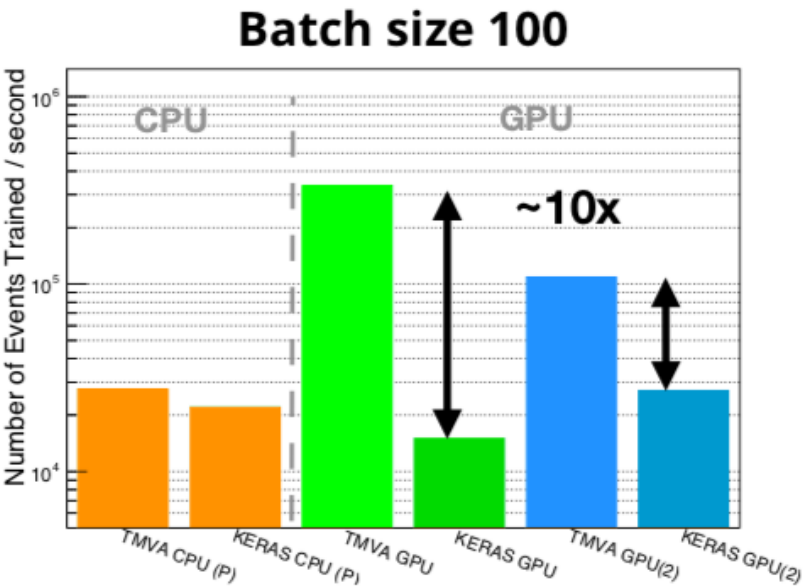
Status deep learning library

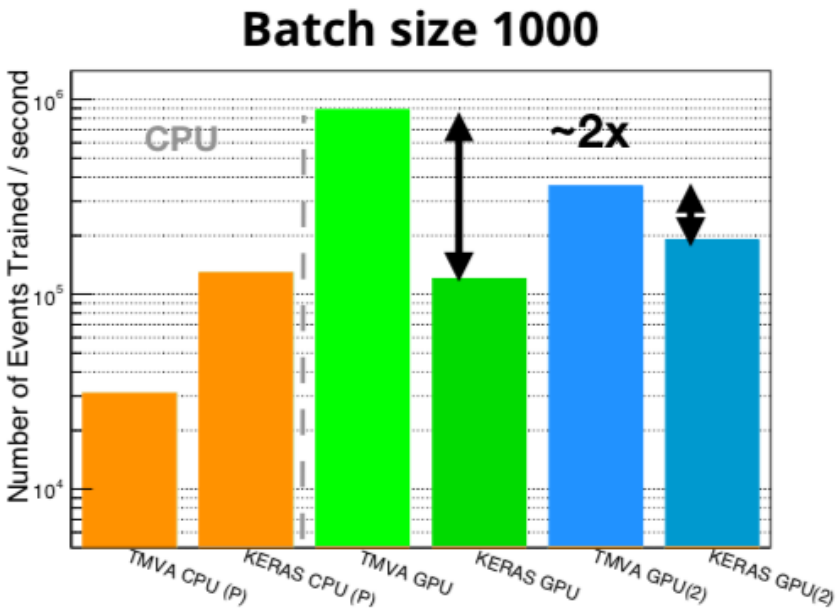
- Deep learning library since 2016
- Recent additions
  - Convolutional and recurrent layers
- Development ongoing!

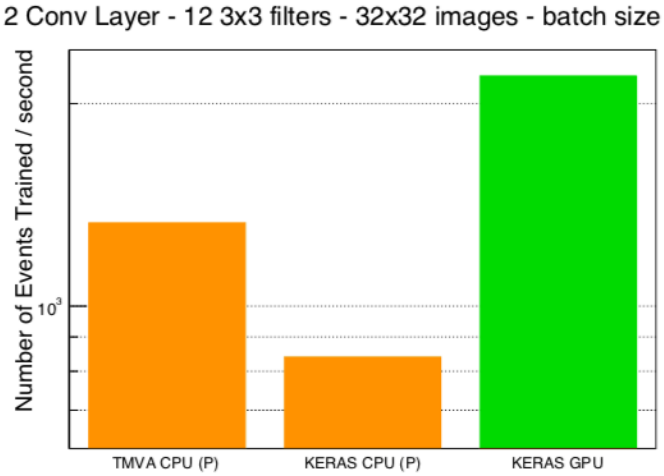


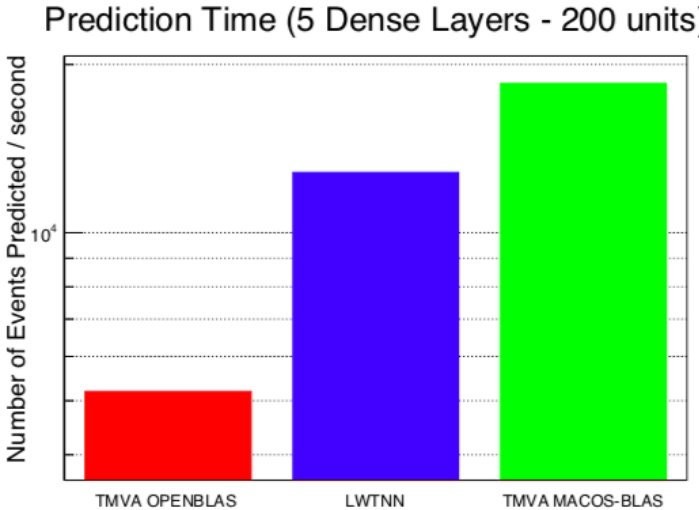
K. Albertsson

TMVA


Batch size 100
Number of Events Trained / second
CPU
GPU
~10x
TMVA CPU (P)
KERAS CPU (P)
TMVA GPU
KERAS GPU
TMVA GPU(2)
KERAS GPU(2)


Batch size 1000
Number of Events Trained / second
CPU
~2x
TMVA CPU (P)
KERAS CPU (P)
TMVA GPU
KERAS GPU
TMVA GPU(2)
KERAS GPU(2)


2 Conv Layer - 12 3x3 filters - 32x32 images - batch size = 32
Number of Events Trained / second
TMVA CPU (P)
KERAS CPU (P)
KERAS GPU


Prediction Time (5 Dense Layers - 200 units)
Number of Events Predicted / second
TMVA OPENBLAS
LWTNN
TMVA MACOS-BLAS

Scikit-HEP

The Scikit-HEP project

The idea, in just one sentence

The Scikit-HEP project (http://scikit-hep.org/) is a community-driven and community-oriented project with the aim of providing Particle Physics at large with a Python package containing core and common tools.

What it is NOT …

- ❑ A replacement for ROOT
- ❑ A replacement for the Python ecosystem based on NumPy, scikit-learn & co.

… and what IT IS

- ❑ An initiative to improve the interoperability between HEP tools and the Python ecosystem
  - Expand the typical toolkitset for HEP physicists
  - Set common APIs and definitions to ease "cross-talk"
- ❑ An initiative to build a community of developers and users
- ❑ An effort to improve discoverability of relevant tools

© dianahep

E. Rodrigues

Scikit-HEP

Scikit-HEP project – toolset / packages overview

J. Pivarski

uproot

Minimalist ROOT I/O in pure Python and Numpy

scikit-hep

Starting point of project. Contains tools for maths, kinematics, units, etc.

histbook

Versatile, high-performance histogram toolkit for Numpy

numpythia

Interface between PYTHIA and NumPy

Vega Scope

Minimal viewer of Vega / Vega-Lite plots in your web browser from local or remote Python processes

formulate

Easy conversions between different styles of expressions

pyjet

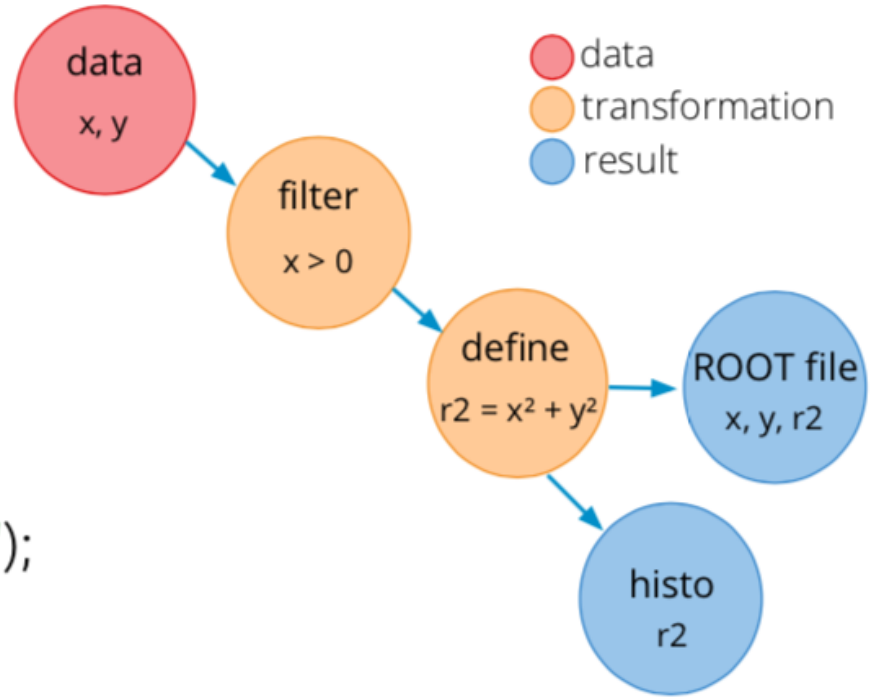
Interface between FastJet and NumPy

E. Rodrigues

And other packages, which tend to be now superseded, hence deprecated …

RDataFrame

```
ROOT::RDataFrame df(dataset);
auto df2 = df.Filter("x > 0")
             .Define("r2", "x*x + y*y");
auto rHist = df2.Histo1D("r2");
df2.Snapshot("newtree", "newfile.root");
```



E. Guiraud

Analysis Preservation

reana

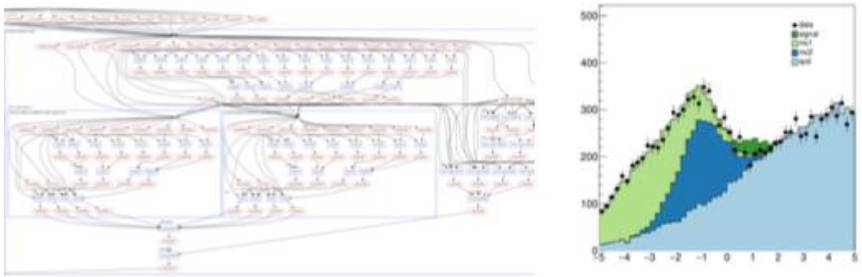
Reproducible research data analysis platform

T. Simko

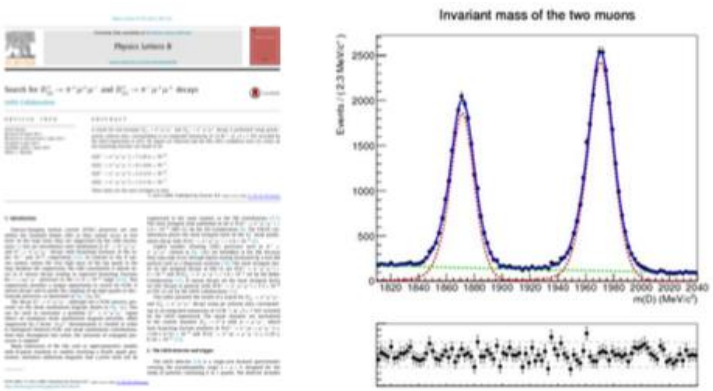
Flexible	Scalable	Reusable	Free
Run many computational workflow engines.	Support for remote compute clouds.	Containerise once, reuse elsewhere. Cloud-native.	Free Software. GPL licence. Made with ♥ at CERN.

L. Heinrich

http://www.reana.io/



https://github.com/reanahub/reana-demo-bsm-search/



https://github.com/reanahub/reana-demo-lhcb-d2pimumu/

Search for $D^+_{(s)} \to \pi^+\mu^+\mu^-$ and $D^+_{(s)} \to \pi^-\mu^+\mu^+$ decays

GitLab

docker

wflow spec

workflow agent

wflow spec

kubernetes

Job

Other Topics

- How to train DNNs distributively and efficiently
- Non-standard images and kernels
- Domain adaptation and systematics
- Bayesian optimization
- Auto-categorization
- Updates of CMS BigData project
- Pandas in HEP
- NanoAOD
- Parallel Fitting
- MEMs on GPUs

HEPML-CWP

Machine Learning in High Energy Physics Community White Paper

July 10, 2018

Abstract: Machine learning is an important applied research area in particle physics, beginning with applications to high-level physics analysis in the 1990s and 2000s, followed by an explosion of applications in particle and event identification and reconstruction in the 2010s. In this document we discuss promising future research and development areas in machine learning in particle physics with a roadmap for their implementation, software and hardware resource requirements, collaborative initiatives with the data science community, academia and industry, and training the particle physics community in data science. The main objective of the document is to connect and motivate these areas of research and development with the physics drivers of the High-Luminosity Large Hadron Collider and future neutrino experiments and identify the resource needs for their implementation. Additionally we identify areas where collaboration with external communities will be of great benefit.

arXiv:1807.02876

Editors: Sergei Gleyzer[26], Paul Seyfert[11], Steven Schramm[28]

Contributors: Kim Albertsson[1], Piero Altoe[2], Dustin Anderson[3], Michael Andrews[4], Juan Pedro Araque Espinosa[5], Adam Aurisano[6], Laurent Basara[7], Adrian Bevan[8], Wahid Bhimji[9], Daniele Bonacorsi[10], Paolo Calafiura[9], Mario Campanelli[8], Louis Capps[2], Federico Carminati[11], Stefano Carrazza[11], Taylor Childers[12], Elias Coniavitis[13], Kyle Cranmer[14], Claire David[15], Douglas Davis[16], Javier Duarte[17], Martin Erdmann[18], Jonas Eschle[19], Amir Farbin[20], Matthew Feickert[21], Nuno Filipe Castro[5], Conor Fitzpatrick[22], Michele Floris[11], Alessandra Forti[23], Jordi Garra-Tico[24], Jochen Gemmler[25], Maria Girone[11], Paul Glaysher[15], Sergei Gleyzer[26], Vladimir Gligorov[27], Tobias Golling[28], Jonas Graw[2], Lindsey Gray[17], Dick Greenwood[29], Thomas Hacker[30], John Harvey[11], Benedikt Hegner[11], Lukas Heinrich[14], Ben Hooberman[31], Johannes Junggeburth[32], Michael Kagan[33], Meghan Kane[34], Konstantin Kanishchev[7], Przemysław Karpiński[11], Zahari Kassabov[35], Gaurav Kaul[36], Dorian Kcira[3], Thomas Keck[25], Alexei Klimentov[37], Jim Kowalkowski[17], Luke Kreczko[38], Alexander Kurepin[39], Rob Kutschke[17], Valentin Kuznetsov[40], Nicolas Köhler[32], Igor Lakomov[11], Kevin Lannon[41], Mario Lassnig[11], Antonio Limosani[42], Gilles Louppe[14], Aashrita Mangu[43], Pere Mato[11], Helge Meinhard[11], Dario Menasce[44], Lorenzo Moneta[11], Seth Moortgat[45], Meenakshi Narain[46], Mark Neubauer[31], Harvey Newman[3], Hans Pabst[36], Michela Paganini[47], Manfred Paulini[4], Gabriel Perdue[17], Uzziel Perez[48], Attilio Picazio[49], Jim Pivarski[50], Harrison Prosper[51], Fernanda Psihas[52], Alexander Radovic[53], Ryan Reece[54], Aurelius Rinkevicius[40], Eduardo Rodrigues[6], Jamal Rorie[55], David Rousseau[56], Aaron Sauers[17], Steven Schramm[28], Ariel Schwartzman[33], Horst Severini[57], Paul Seyfert[11], Filip Siroky[58], Konstantin Skazytkin[39], Mike Sokoloff[6], Graeme Stewart[59], Bob Stienen[60], Ian Stockdale[61], Giles Strong[5], Savannah Thais[47], Karen Tomko[62], Eli Upfal[46], Emanuele Usai[46], Andrey Ustyuzhanin[63], Martin Vala[64], Sofia Vallecorsa[65], Mauro Verzetti[66], Xavier Vilasís-Cardona[67], Jean-Roch Vlimant[3], Ilija Vukotic[68], Sean-Jiun Wang[26], Gordon Watts[69], Michael Williams[70], Wenjing Wu[71], Stefan

Summary

- New CHEP track with many exciting results put forward by experiments
- Machine learning applications expanding everywhere, HEP not an exception
- Great progress and an opportunity to re-examine things for Run-2
- Thanks to all for making Track 6 a success