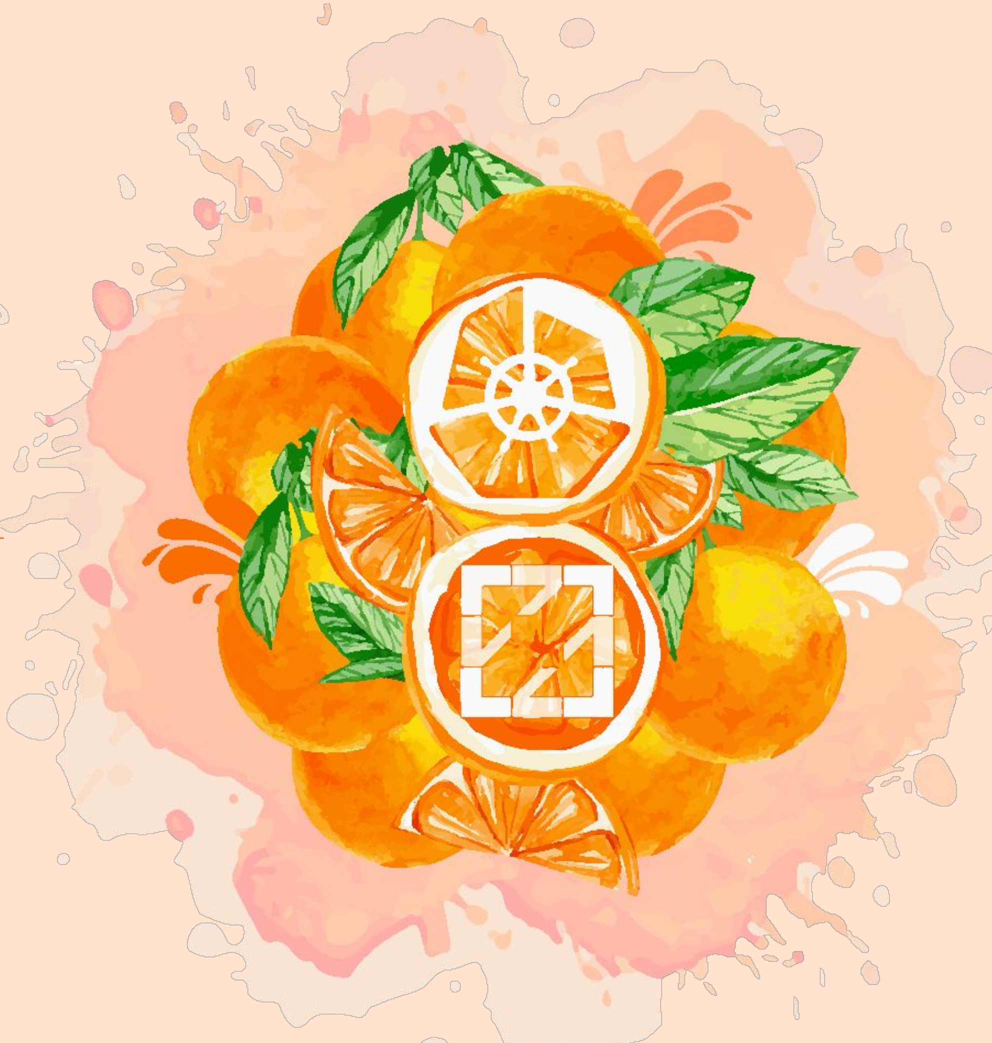
KubeCon | CloudNativeCon

Europe 2022

WELCOME TO VALENCIA



KubeCon | CloudNativeCon
Europe 2022

Jet Energy Corrections with GNN Regression using Kubeflow at CERN

Daniel Holmberg, CERN
Dejan Golubovic, CERN

Introduction

Machine Learning and Kubeflow at CERN

Jet Energy Corrections with GNN

Jet Energy Corrections with GNN - Kubeflow Demo

Conclusions

CERN

Mission - uncover **what the universe is made of and how it works**

- Study of subatomic particles

Largest particle physics laboratory in the world

International collaboration

- 17 000 employees
- 110 nationalities
- Collaboration with institutes in 70 countries

Large Hadron Collider - LHC

Largest particle accelerator in the world

27 km ring of superconducting magnets

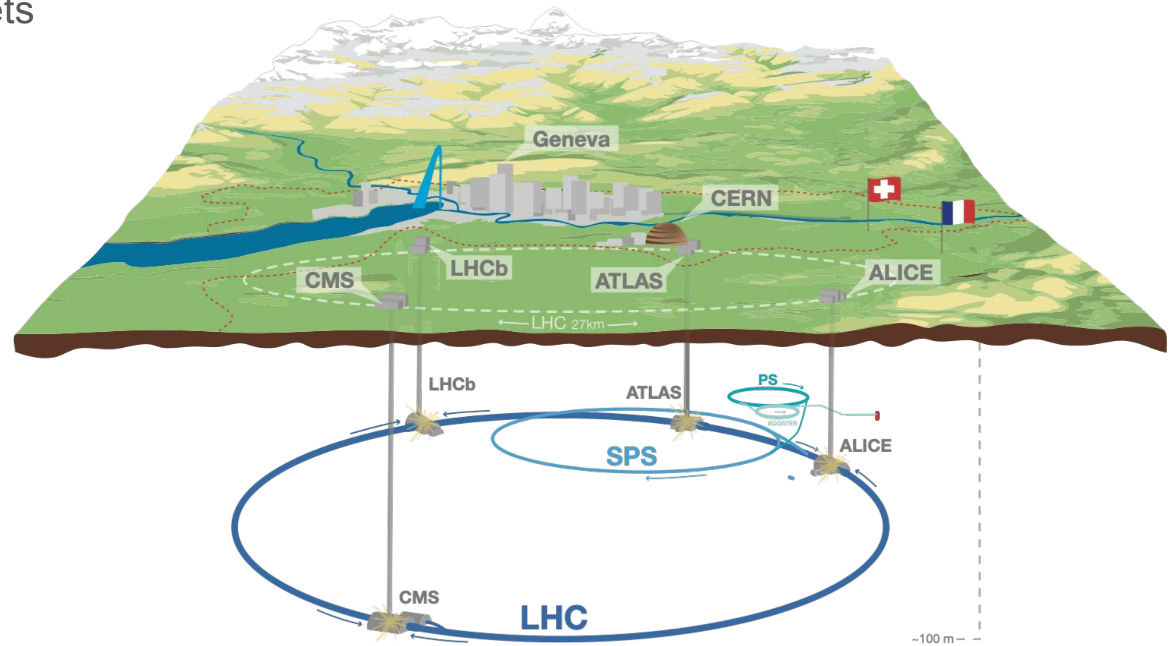
100 m underground

Accelerate particles

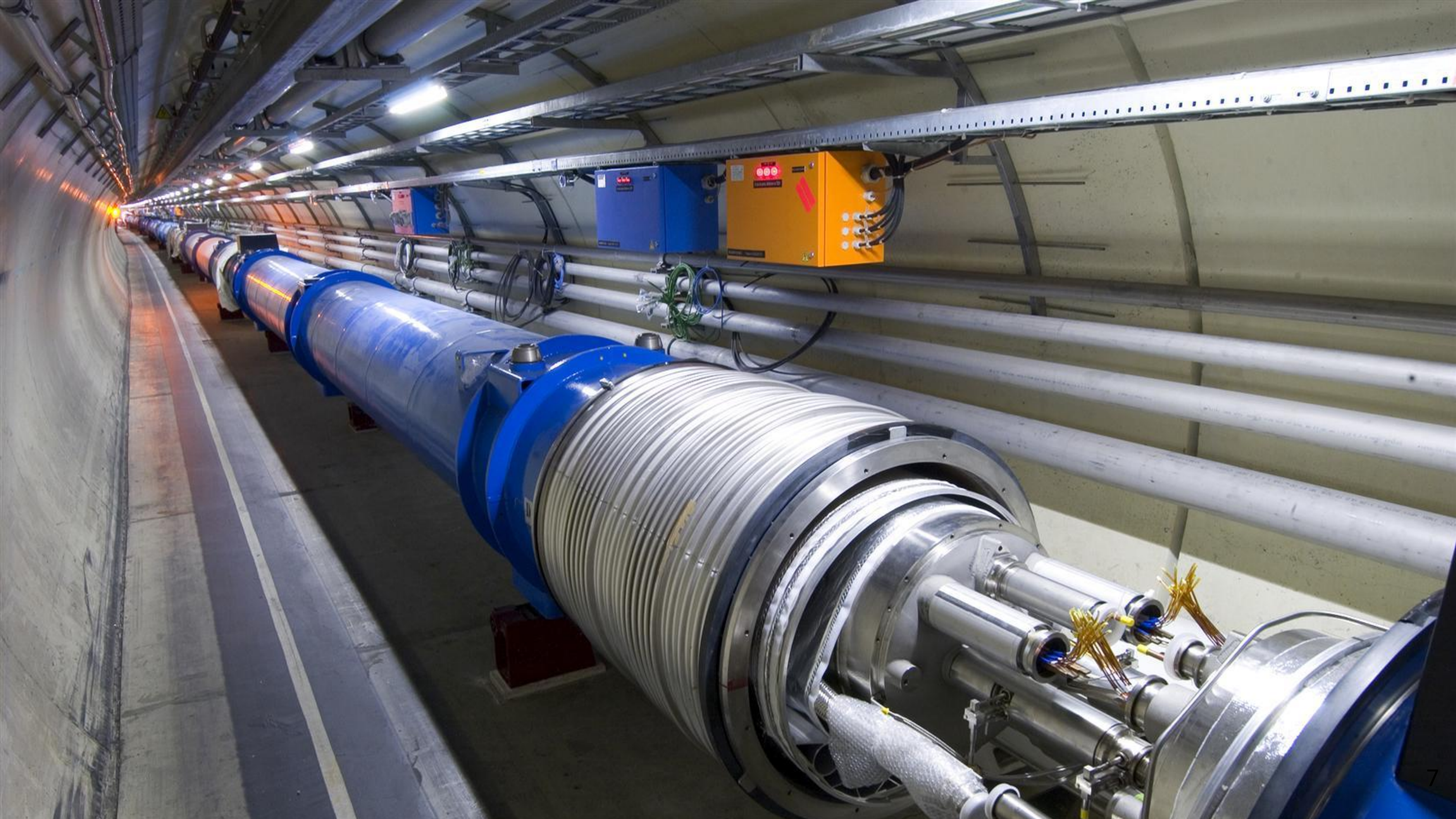
- Near the speed of light

4 collision points - detectors

- CMS, ATLAS, LHCb, ALICE




home.cern
Universe of Particles




JLG LIFTLUX 153-12
JLG
cms.cern

Introduction

Machine Learning and Kubeflow at CERN

Jet Energy Corrections with GNN

Jet Energy Corrections with GNN - Kubeflow Demo

Conclusions

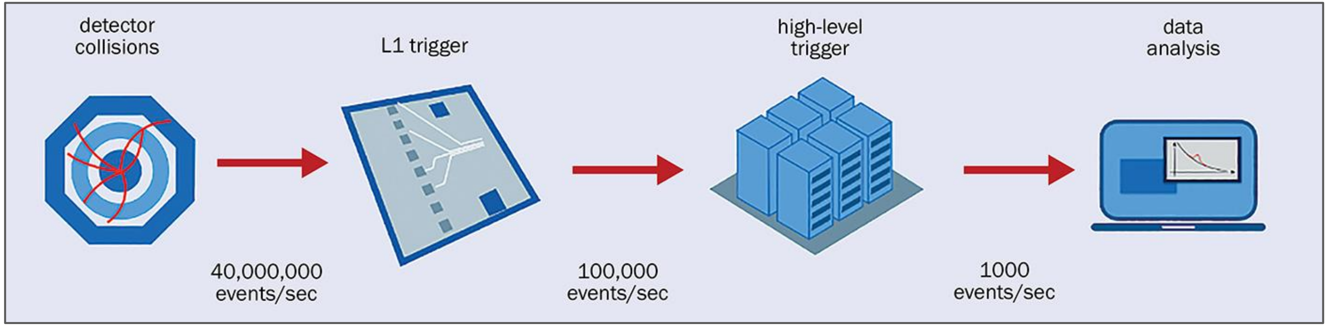
High Energy Physics Data at CERN

Around 40 million collisions per second in LHC

90 petabytes of data per year produced by all experiments

Potential for machine learning in different stages of data acquisition

The amount of data accessible can benefit machine learning algorithms



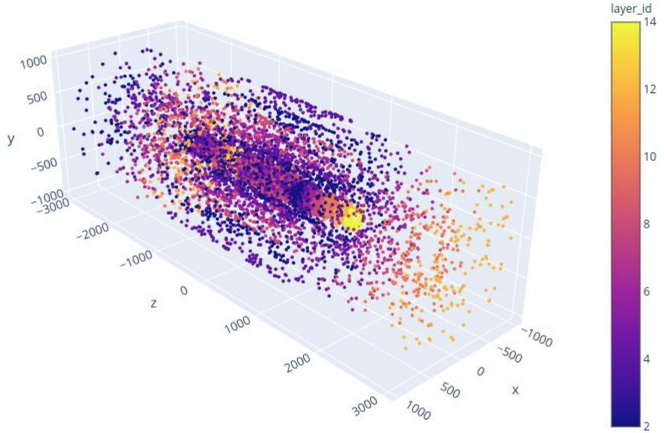
CMS Data Acquisition System

Machine Learning at CERN

Wide range of ML applications at CERN

Data acquisition

- Coarse grain event selection - fast inference on FPGAs
- Fine grain event selection - GPU inference
- Particle tracking and reconstruction



Beam calibration - reinforced learning

Simulations - 3D GANs as a faster alternative to Monte Carlo

IT infrastructures - notification delivery system, anomaly detection in cloud monitoring

https://towardsdatascience.com/particle-tracking-at-cern-with-machine-learning-4cb6b255613c

Kubeflow at CERN

Centralized ML platform to improve resource utilization across CERN

- Reduce maintenance work for researchers
- Easier access to GPUs
- Scaling capabilities

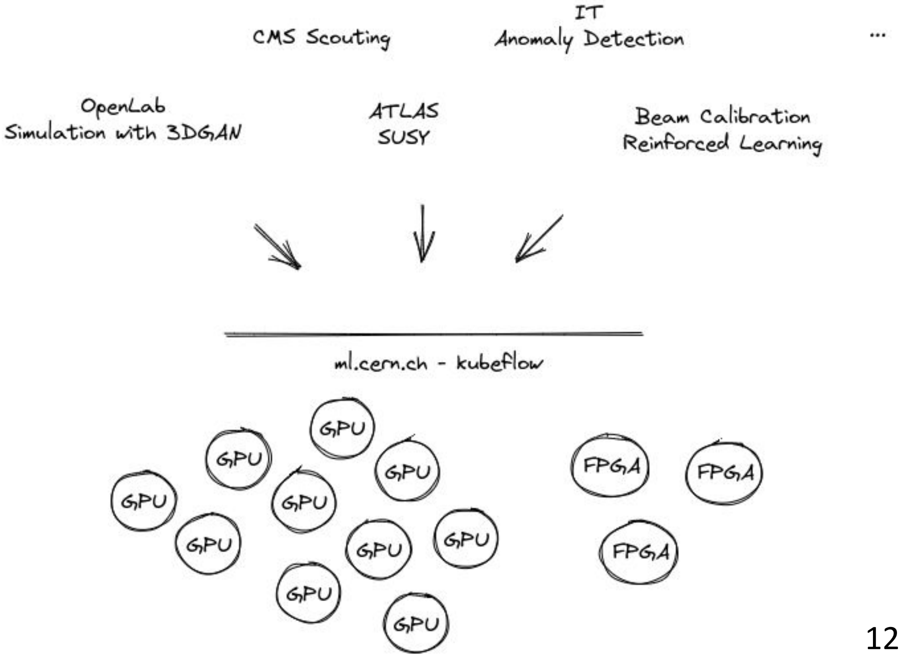
On-premise cluster, using Openstack

GitOps with ArgoCD

Integration with CERN services

- SSO, Harbor registry, CSI, Gitlab CI

In production since April 2021



Kubeflow at CERN

KubeCon CloudNativeCon Europe 2022

Previous talks

- [A Better and More Efficient ML Experience for CERN Users]
- [Building and Managing a Centralized ML Platform with Kubeflow at CERN]
- Focus on infrastructure and admin workflows

Today - focus on a specific use case from CERN

- Show **application of machine learning in high energy physics**
- Demonstrate **utilization of Kubeflow to scale ML workloads**

Introduction

Machine Learning and Kubeflow at CERN

Jet Energy Corrections with GNN

Jet Energy Corrections with GNN - Kubeflow Demo

Conclusions

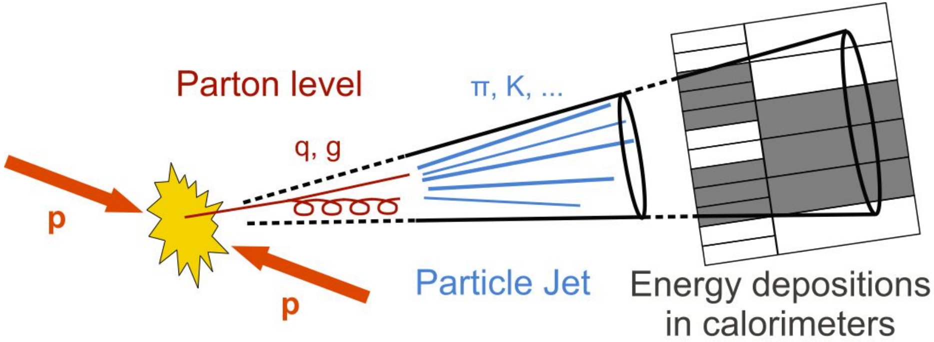
Jet Energy Corrections

Colliding **protons** at high energies produces color-charged **partons**

Hadronization gives rise to a spray of color-neutral particles that are clustered into a **jet**

Measured **energy** differs from theory due to detector inaccuracies, invisible particles etc.

Can machine learning help with energy calibration?

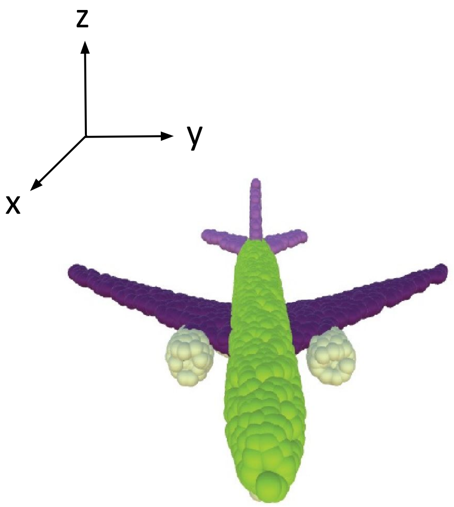


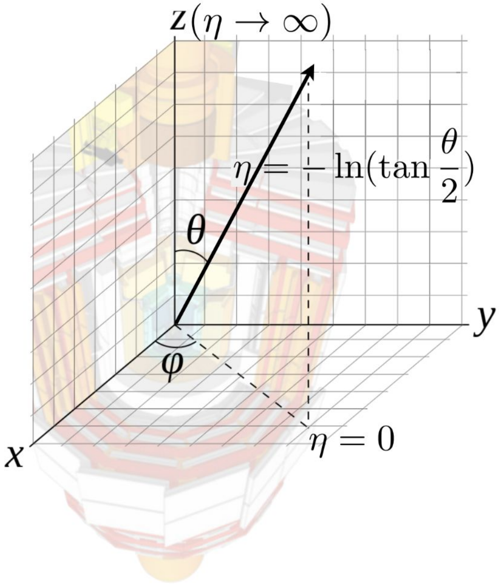
https://cms.cern/news/jets-cms-and-determination-their-energy-scale

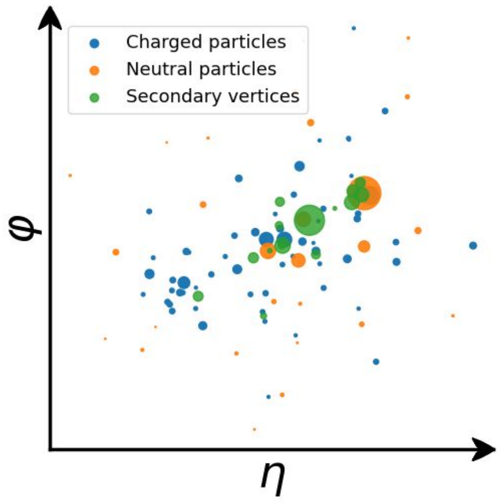
Representing Jets as Particle Clouds

Use detector coordinates to represent jets as **particle clouds**

Analogous to **point clouds** in computer vision problems

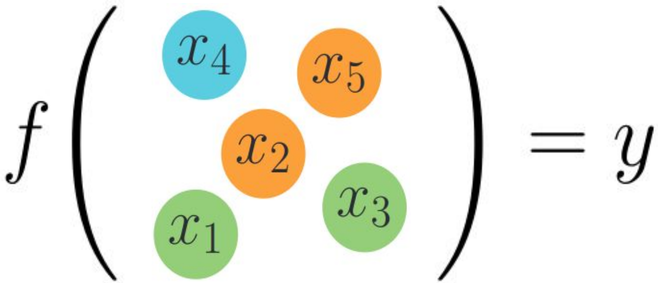






Learning on Particle Clouds

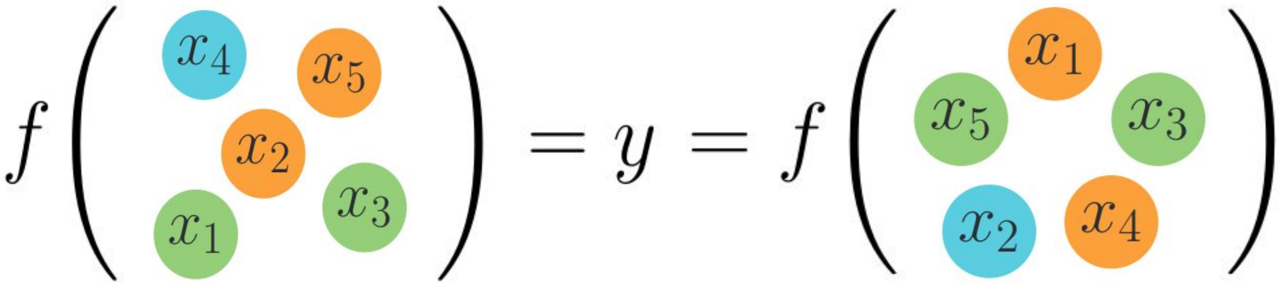
Map set of particle feature vectors x_i towards energy target y



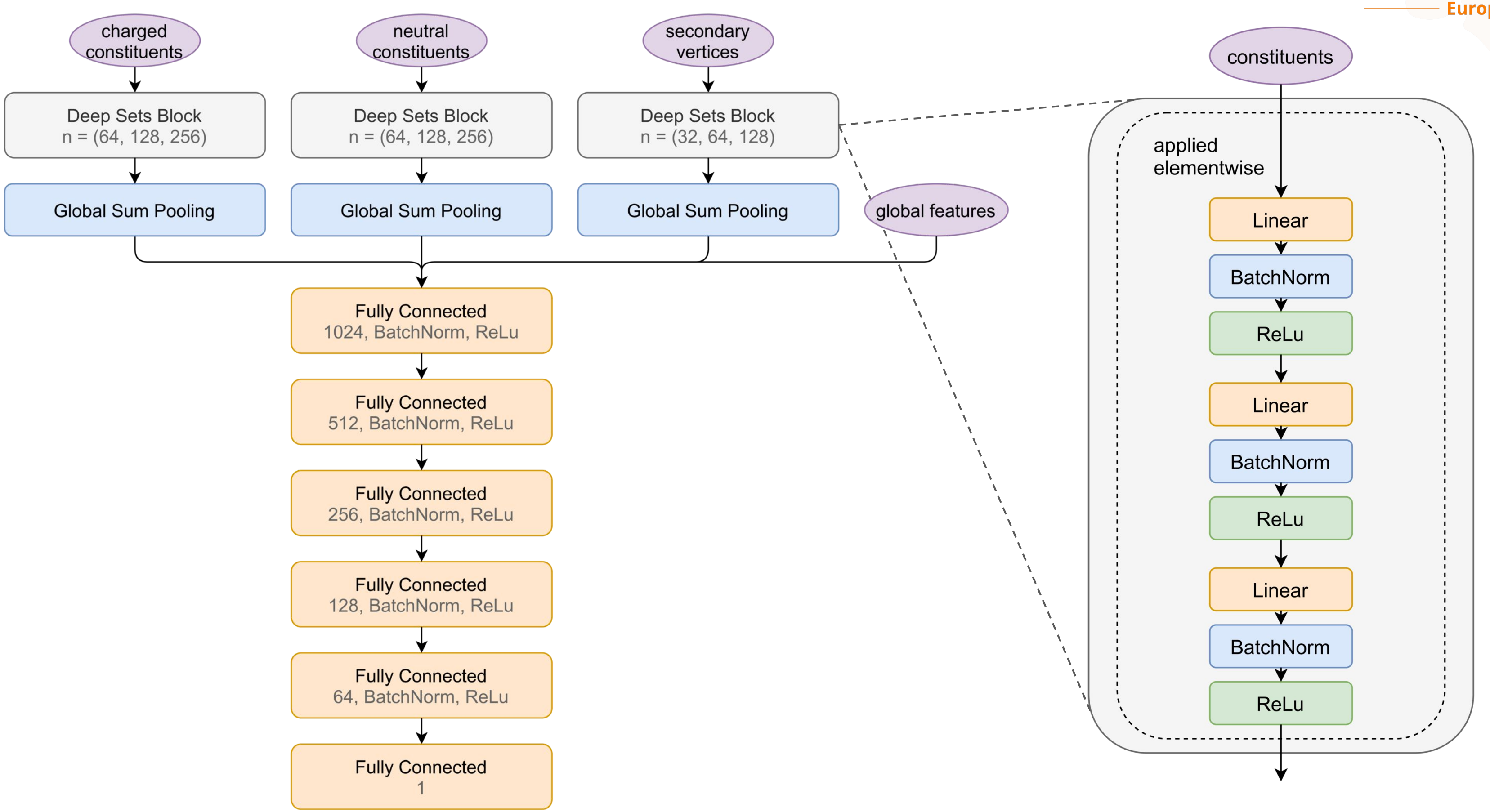
Learning on Particle Clouds

Map set of particle feature vectors x_i towards energy target y

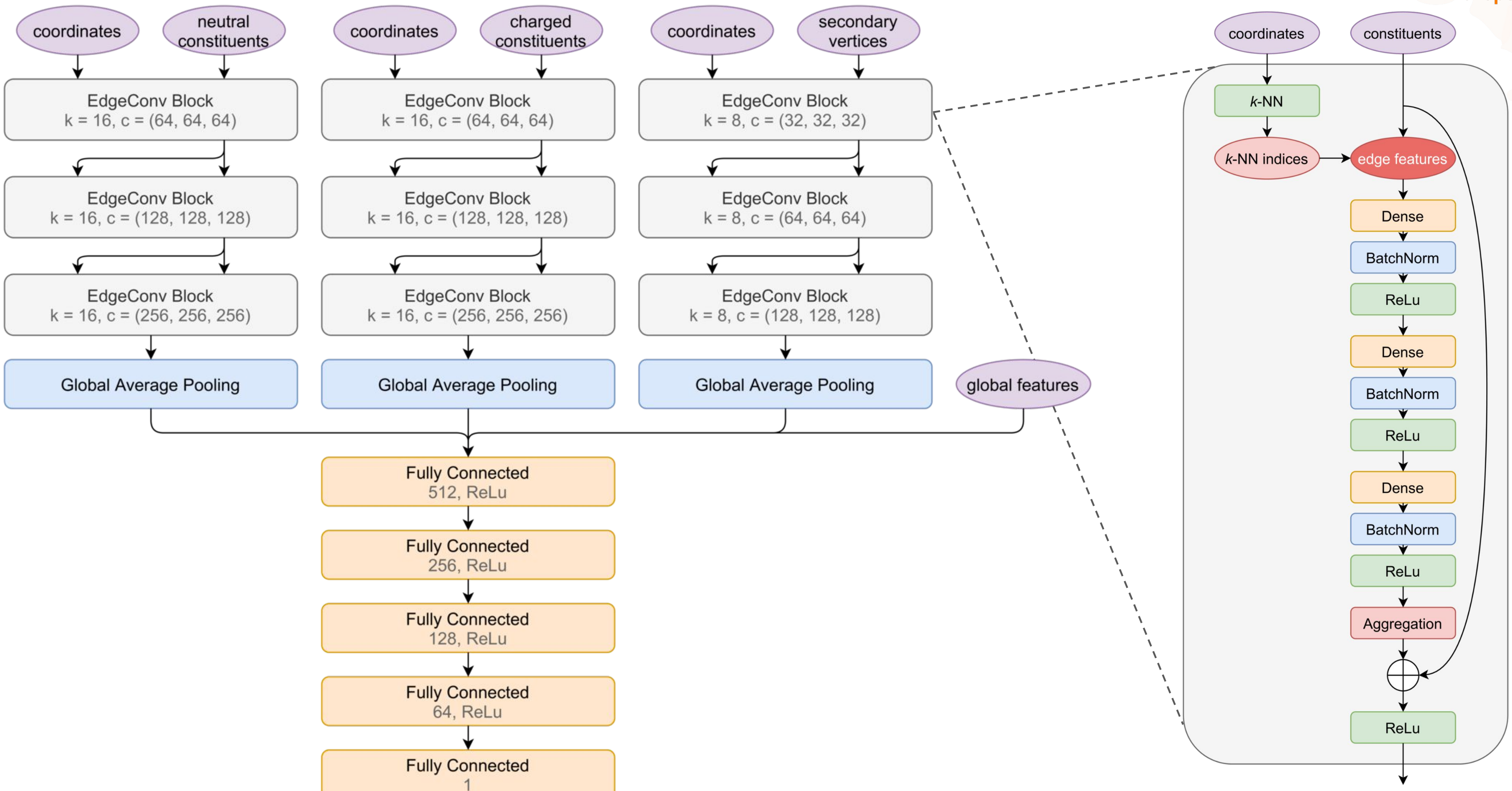
The model must be invariant to the order of the particles



Particle Flow Network



ParticleNet

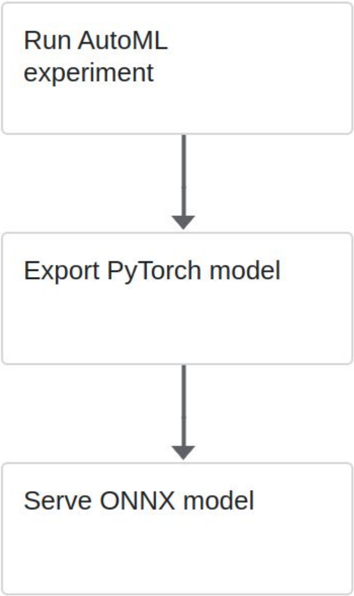


ML Pipeline

Kubeflow Pipelines: “Engine for scheduling multi-step ML workflows”

Define end-to-end ML pipeline as a directed graph

- Start with running a Katib AutoML experiment
- Export the optimal model
- Finally serve using KServe



Training

Dataset with 14 million jets = 10GB stored on S3

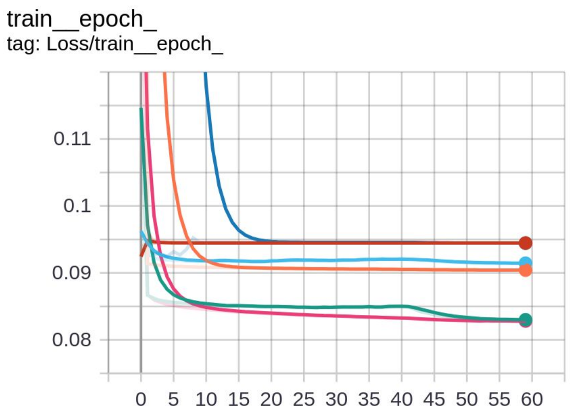
Minimize mean absolute error (MAE) loss

Tune hyperparameters using Random Search to reach a lower loss

Scalability

- Multi-node training by using the [PyTorchJob](#) operator
- Multiple CPU workers can read data simultaneously
- Additionally, many Katib trials can be run in parallel

Monitor training with Tensorboard component



Inference

Export best PyTorch model to [ONNX]

Serve model with [Nvidia Triton inference server]

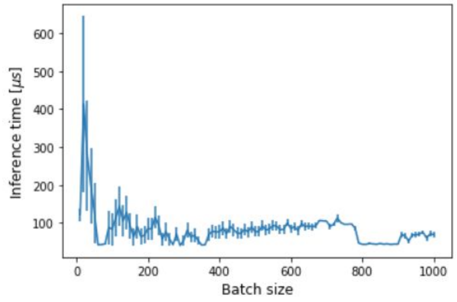
Use Triton's Python client to request predictions and get usage statistics

Analyze inference time and plot physics result in a notebook server on Kubeflow

Model Servers + NEW MODEL SERVER

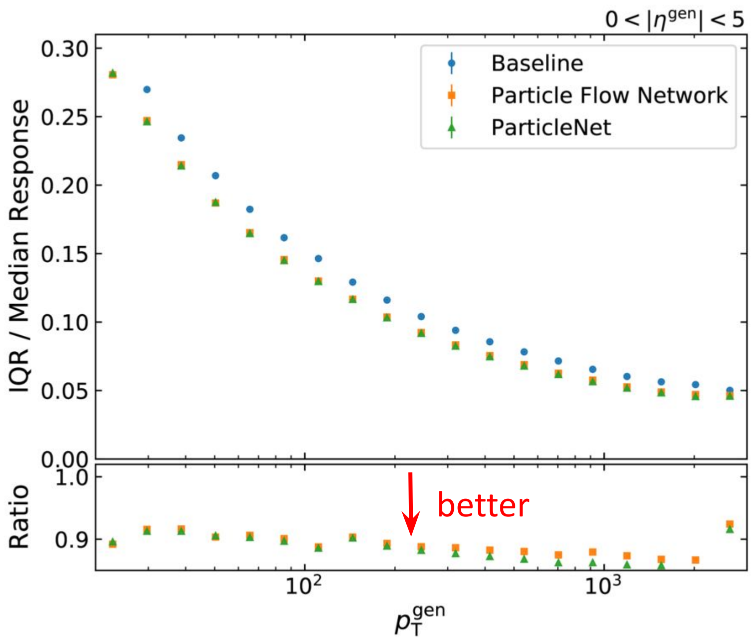
Status	Name	Age	Predictor	Runtime	Protocol	Storage URI
✓	pfn-regressor-ea37f4	2 hours ago	Triton	21.09-py3		s3://jec-data/pfn-regressor-ea37f4

```python
plt.errorbar(x=batch_sizes, y=y, yerr=yerr)
plt.xlabel('Batch size', fontsize=12)
plt.ylabel('Inference time [$\mu s$]', fontsize=12)
plt.savefig('inference_time.png')
plt.show()
```

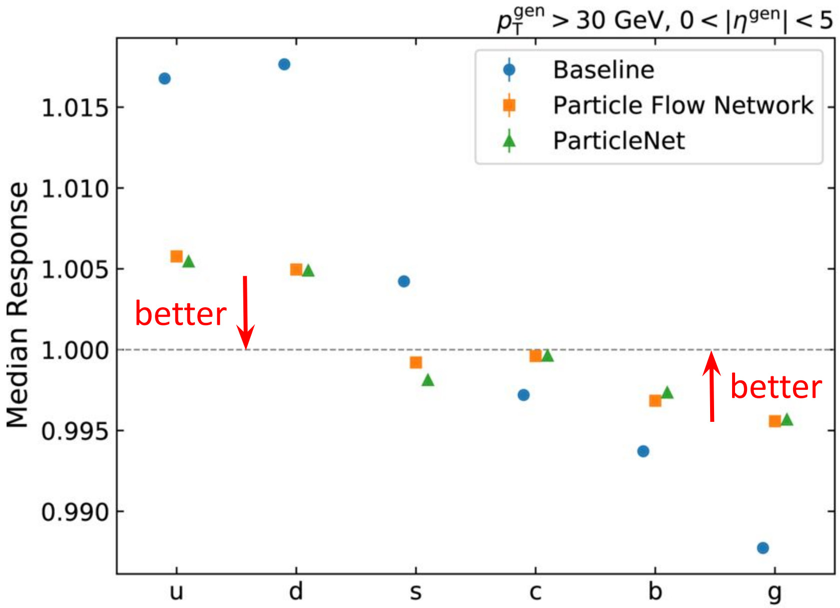


Physics Results

Energy resolution: 10% improvement



Flavor dependence: factor 3 improvement



Challenges

KubeCon CloudNativeCon Europe 2022

Finding a correct version of the triton server image

Tensorboard S3 integration

- Tensorboard controller customized downstream to pick up S3 credentials from a secret
- Issue - https://github.com/awslabs/kubeflow-manifests/issues/118

Be careful with model logs outputs

- Can make obtaining StdOut metrics difficult
- Better option - use file metrics collector

Katib UI could be better suited for multiple metrics

Conclusions

ML can provide **significant improvements** in high energy physics use cases

Jet tagging example

- Energy resolution improved by 10%
- Flavour dependance improved by factor of 3

Kubeflow greatly **facilitates the scalability** of large-scale workloads

- Excellent mutual integration of components (Pipelines, AutoML, operators, KServe)
- Customizable and reproducible environments
- Well accepted across CERN scientific community

Thank you for the attention!

Questions?

Backup

Particle Flow Network [arXiv:1810.05165]

An MLP ϕ is applied to every particle x_i

$$\mathbf{h}_i = \phi(\mathbf{x}_i)$$

Aggregate latent features h_i using sum pooling (order invariant operation)

Feed into another MLP ρ mapping to the regression target

$$f(\mathbf{X}) = \rho\left(\sum_{i \in \mathcal{V}} \phi(\mathbf{x}_i)\right)$$

ParticleNet [arXiv:1902.08570]

Initial graph in (η, φ) space — updated after each edge convolution

Local patch for every particle using *k*-nearest neighbors

Define edge features for each center-neighbor pair

$$\mathbf{e}_{ij} = \psi(\mathbf{x}_i,\ \mathbf{x}_j)$$

Aggregate using average pooling and concatenate with skip connection

$$\mathbf{h}_i = \phi\left(\mathbf{x}_i, \frac{1}{k}\sum_{j\in\mathcal{N}_i^k}\psi(\mathbf{x}_i,\ \mathbf{x}_j)\right)$$

Pool outputs and feed into another MLP mapping to the target

$$f(\mathbf{X},\mathbf{A}) = \rho\left(\frac{1}{n}\sum_{i\in\mathcal{V}_i^n}\phi(\mathbf{x}_i,\mathbf{X}_{\mathcal{N}_i^k})\right)$$