INFN FIRENZE

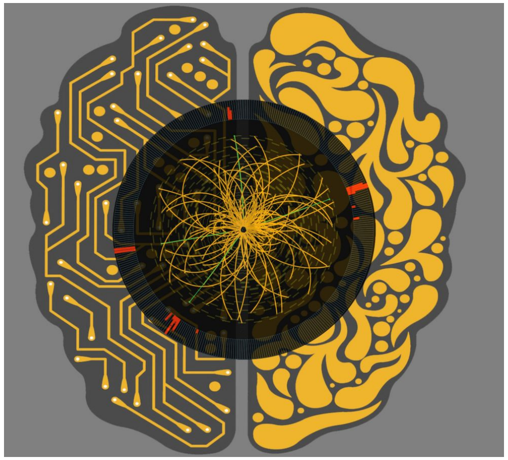
3 Febbraio 2021

Congressino CSN5

Machine Learning (ML_INFN) and Artificial Intelligence in Medicine (AIM)

Lucio Anderlini

Istituto Nazionale di Fisica Nucleare
Sezione di Firenze



Introduction

Artificial Intelligence is reshaping the way computing is done.

The most advanced machine learning libraries are designed to provide a **simplified access to GPUs in cloud environments** while hiding the complexity of adapting the code to a specific hardware.

Scientific Communities as those in astronomy, bioinformatics, computational chemistry, genetics… are using new technologies which ease the exploitation of High Performance Computing infrastructures.

While we can hardly play a leading role in the field, **dominating these technologies is crucial to effectively exploit to-date computing infrastructures.**

INFN Cloud [link]

INFN Clouds is an initiative of the INFN, supported by CCR, to setup a national cloud to give the local units for a variety of tasks and services on demand.

Cloud Computing allows to instantiate a multitude of virtual machines, container or docker on shared servers to better organize resource requirements.

Machine learning is a highly demanding task in terms of computing resources, for a limited amount of time. It fits perfectly.

Leandro

ML-INFN

ML-INFN is CSN5 initiative to enhance the culture on machine learning within the INFN communities.

A Knowledge-Base with examples from HEP, Virgo and Medical Physics is being setup (link)

Cloud experts are helping to make HPC resources accessible through INFN credentials.

Tutorials and hands-on are being organized every six months*

Lucio

*covid-19 caused some delay, tutorials to restart in 2021 in remote mode.

Deep Learning in Florence

Evolving LHC Computing platforms to speed up the pp collision simulation.

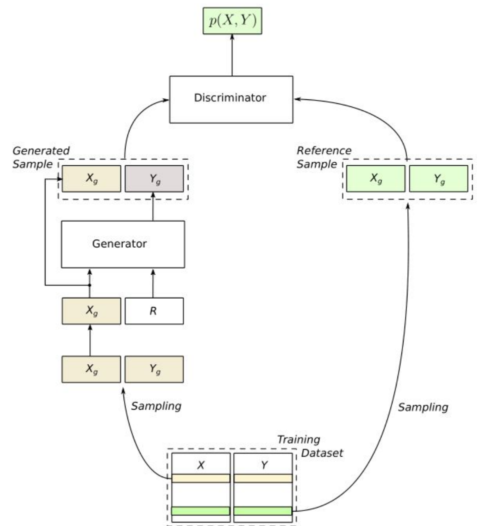
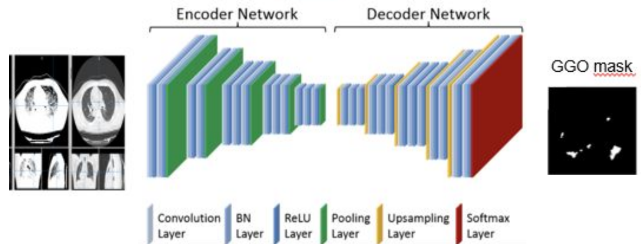


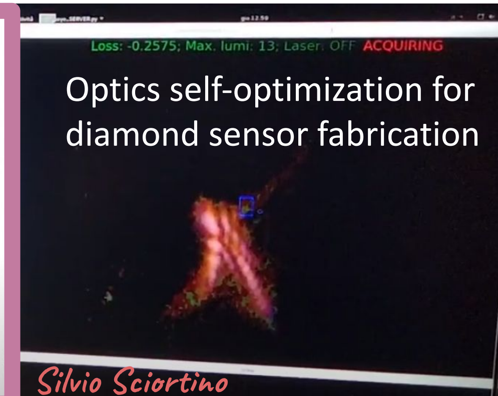
Figure 2: Schematic representation of the *Adversarial* training scheme.

Matteo Barbetti



Segmentations of Covid lesions with Active Contouring and AEs.

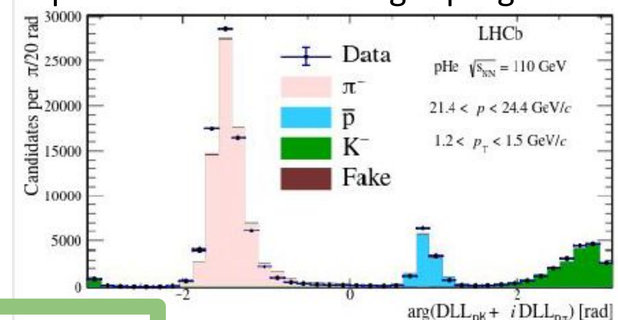
Stefano Piffer



Optics self-optimization for diamond sensor fabrication

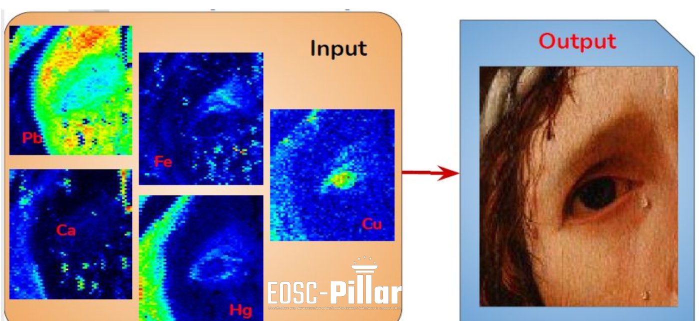
Silvio Sciortino

Precise modelling of the PID detectors response for the fixed-target program at LHCb

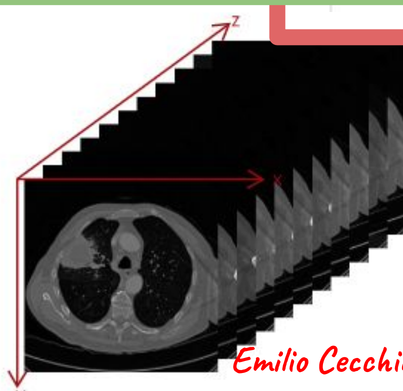


Saverio Mariani

X-ray spectrum to visible colors

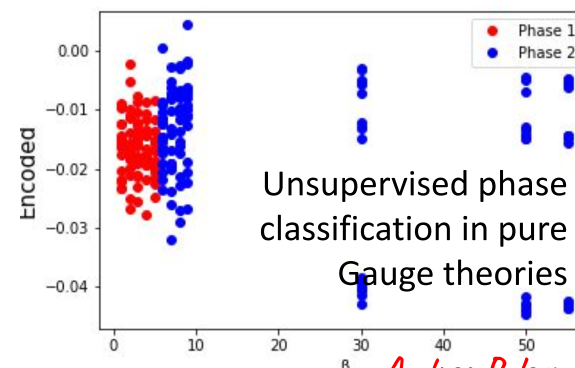


Alessandro Bombini



Emilio Cecchini

Applying 3D CNNs to small datasets with transfer learning



Unsupervised phase classification in pure Gauge theories

Andrea Palermo

Deep Learning is nothing but a very complicated fit

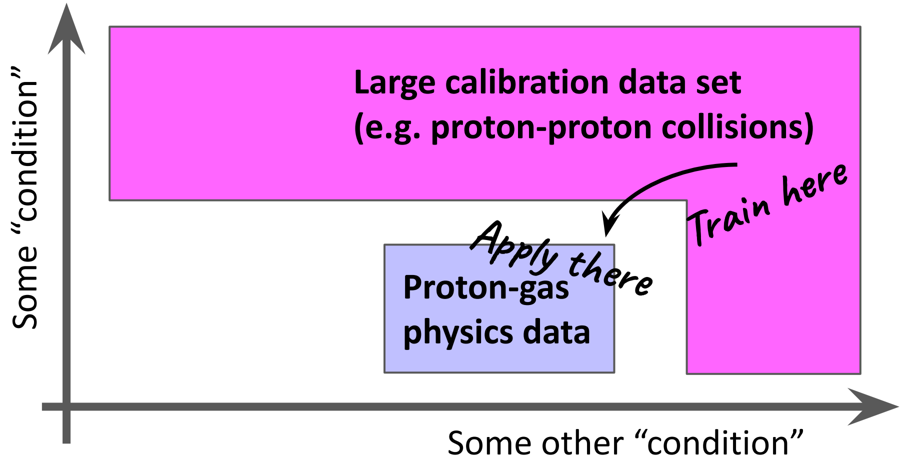
In the fixed-target program of the LHCb experiment, we are studying collisions between the accelerated protons and a gas injected in the beam pipe.

The detector conditions are very different in proton-proton and proton-gas operations. Hence, we wish to "translate" the (nearly) complete understanding of our detector built in years of high luminosity proton-proton collisions to decrypt the results of the short proton-gas runs.

We need a statistical model of the detector response trained on pp data and

robust against extrapolation

to the proton-gas condition regime.



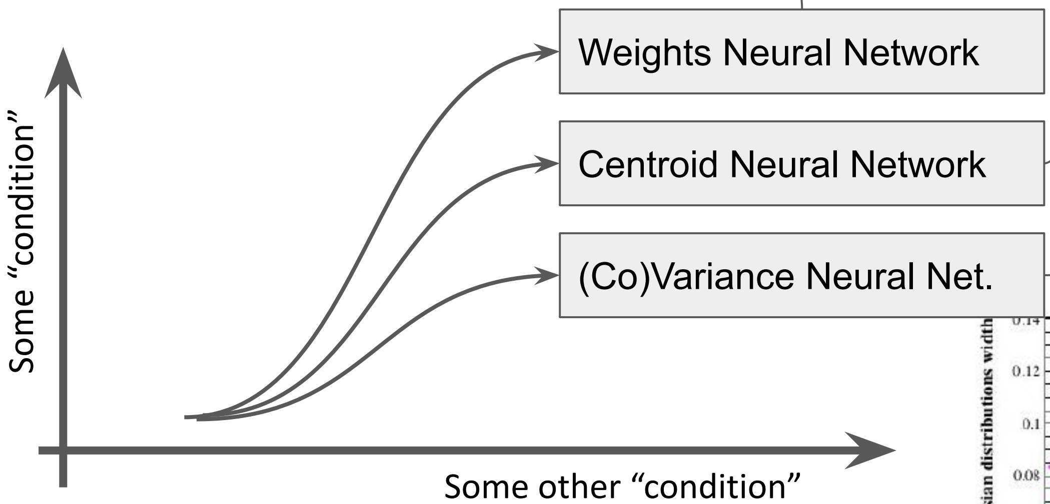
***Notice:**
today, instead of using proton-proton calibration data we use a large dataset of pNe calibration data to train a model to be used in pHe calibration data.

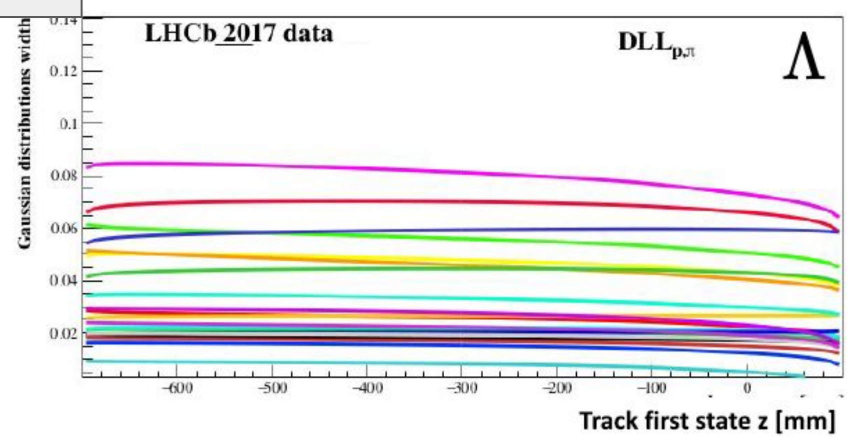
Reducing the distance to be covered with extrapolation is clearly beneficial.

A neural-network defined probability density function

$$\underline{x}_p \sim \sum_{j=1}^{N_{g,p}} \alpha_{j,p}(\underline{\theta}) \frac{\exp\left(-\frac{1}{2}(\underline{x}_p - \underline{\mu}_{j,p}(\underline{\theta}))^T \, \Sigma^{-1}_{j,p}(\underline{\theta}) \, (\underline{x}_p - \underline{\mu}_{j,p}(\underline{\theta}))\right)}{2\pi\sqrt{\det(\Sigma_{j,p}(\underline{\theta}))}}$$

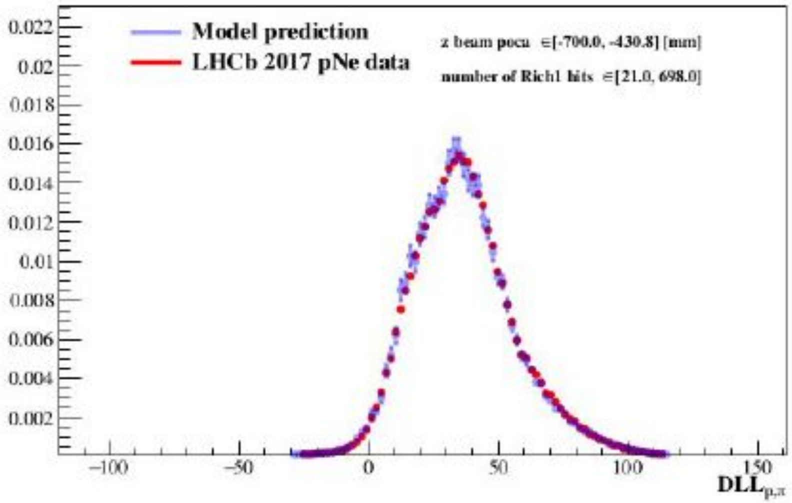
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

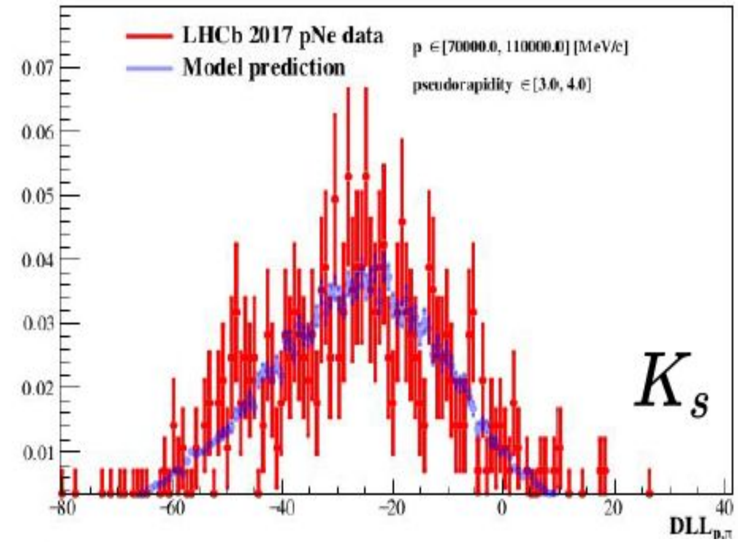


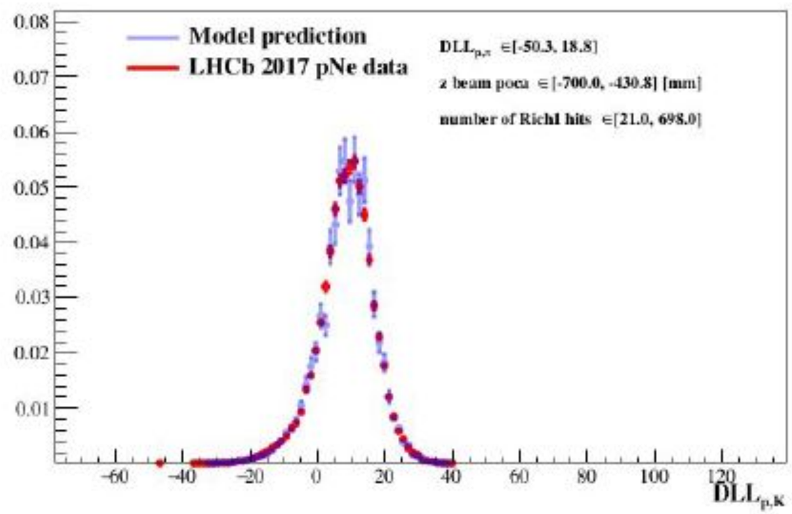


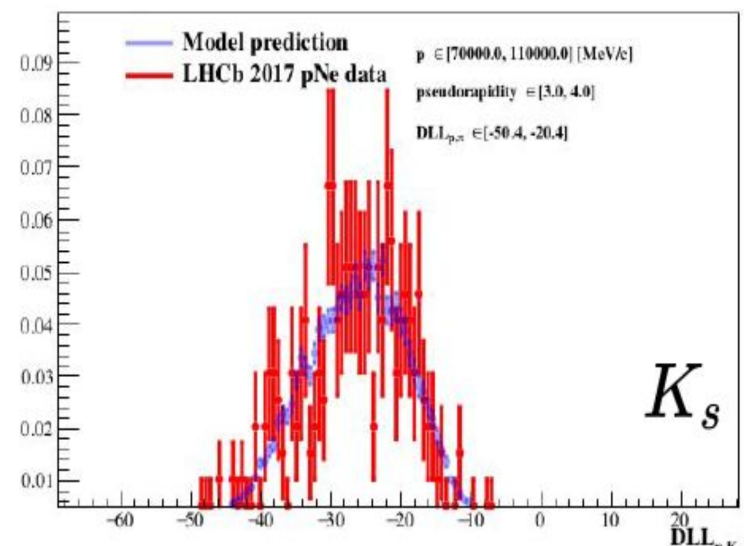
A smooth dependence of Gaussian Mixture Parameter on the "conditions" builds confidence in the extrapolation.

The trained model clearly predicts the distributions in poorly-populated regions of the condition parameter space.



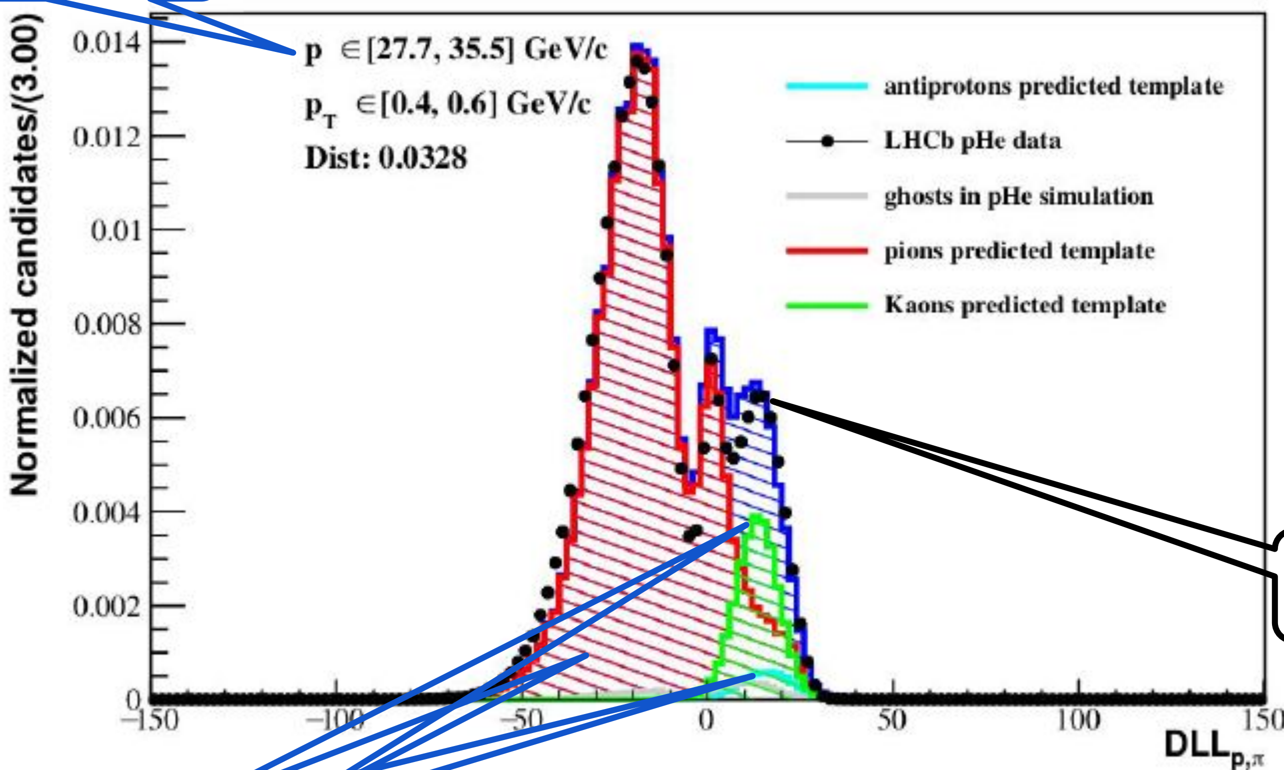






The predicted distributions are used to build precise templates to fit the experimental data.

Some conditions…



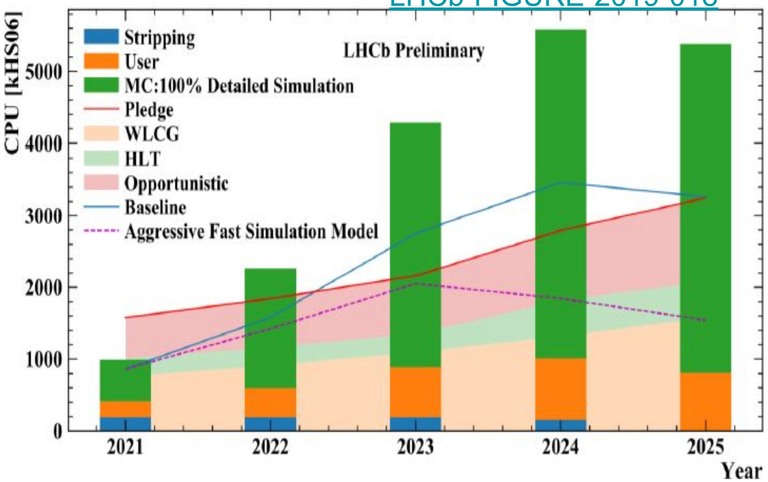
Experimental data

Extrapolated templates

The fit quality obtained combining extrapolated templates is systematically better than for fits combining Geant4-simulated templates.

Fast simulation of the LHCb detector: *the idea*

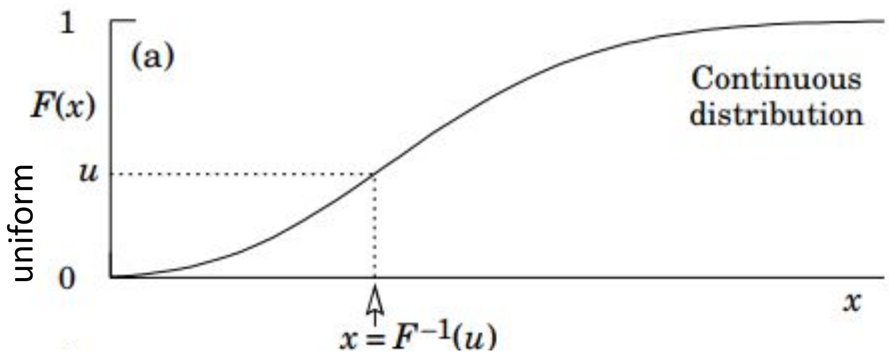
LHCb-FIGURE-2019-018



Neural networks are being explored in several field to approximate the equivalent of the "**inverse of the cumulative**" in a multivariate space.

Such an approximation allows to obtain **highly efficient MC simulations** mapping for example multinormal random variables into random variables with physical meaning and precise correlation patterns.

1D



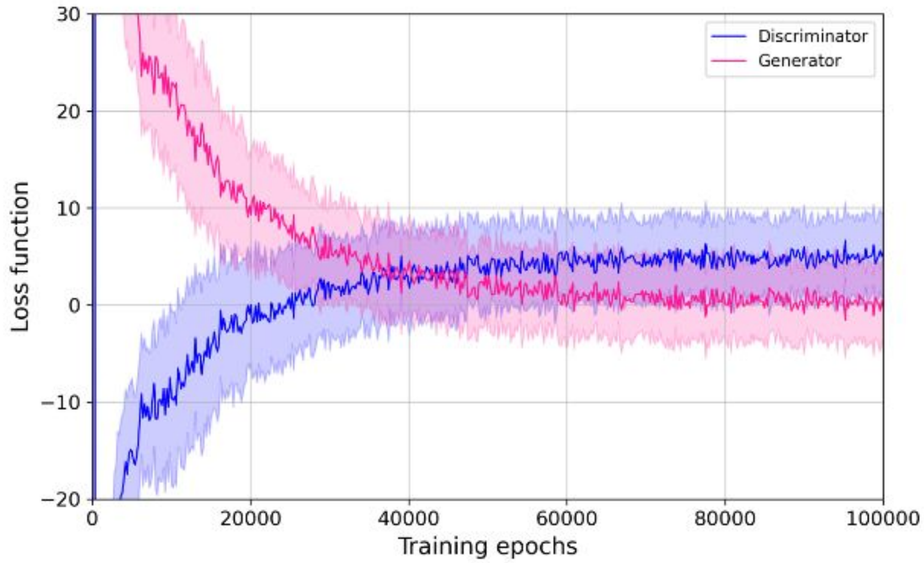
mD

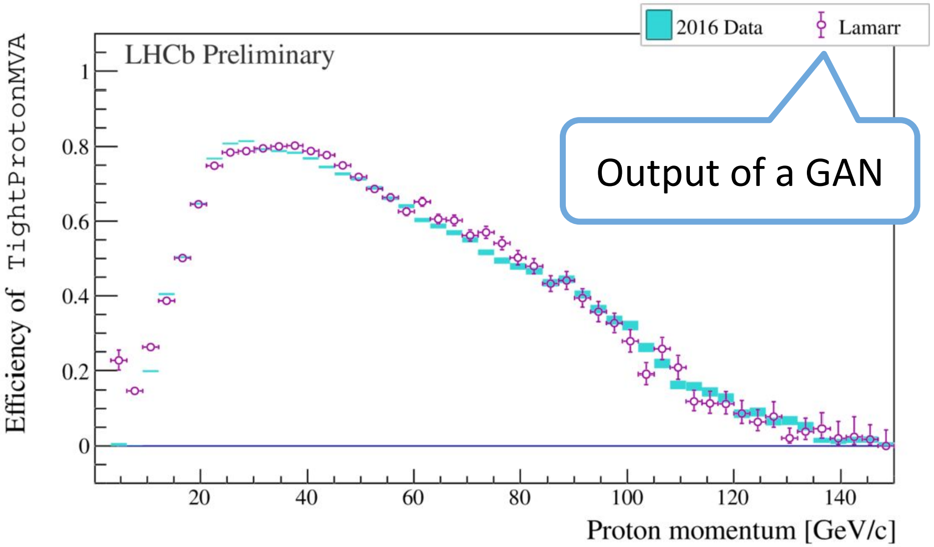
$$\mathrm{ANN} : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Multinormal
or uniform
or combinations

with $m \leq n$

Fast simulation of the LHCb detector: *training, validation*





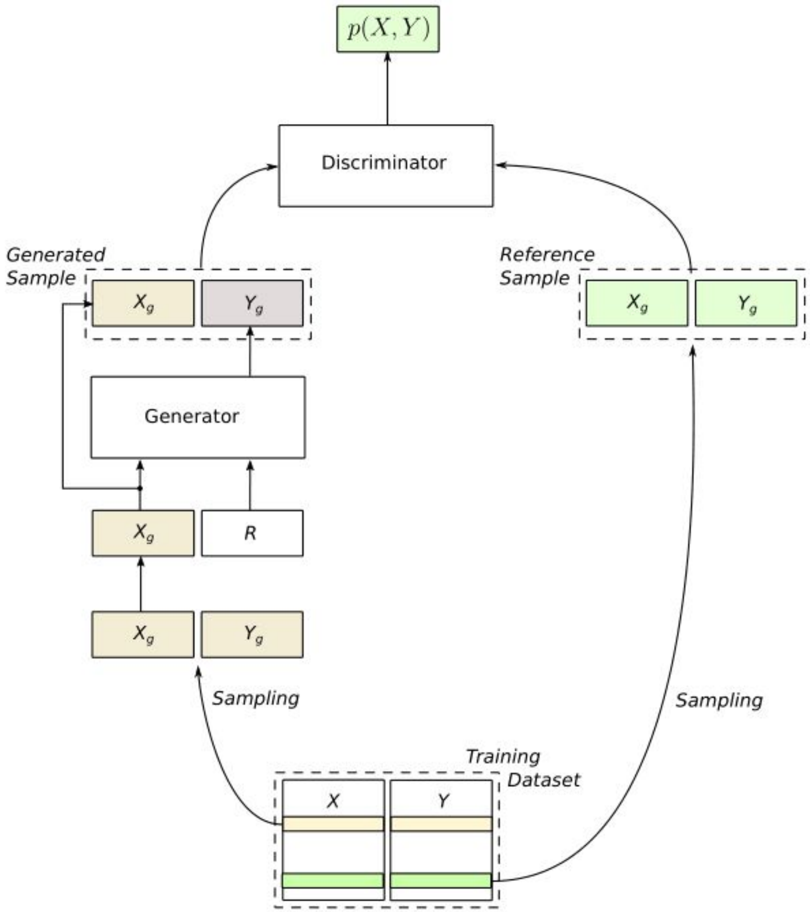
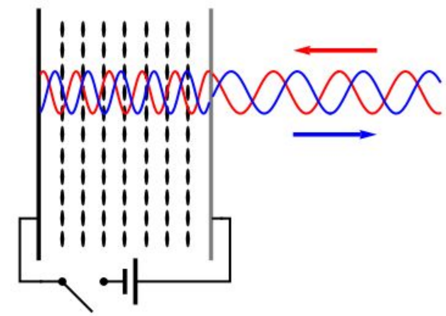


Figure 2: Schematic representation of the *Adversarial* training scheme.

Focusing the laser beam in diamond

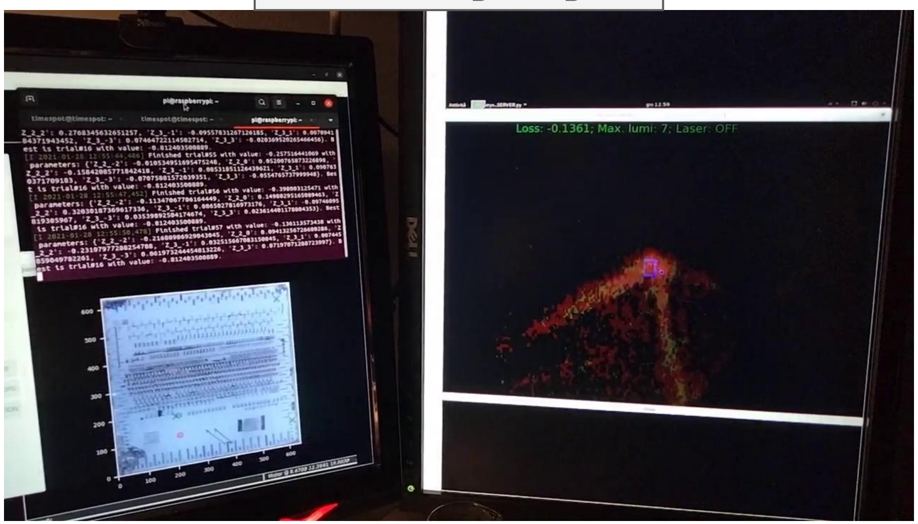
To make diamond sensors, a **femtosecond-pulsed laser is focused in diamond** to modify the carbon bindings from diamond to graphite, creating conducting wires in a semiconductor material. The **quality of the laser focus is critical**.



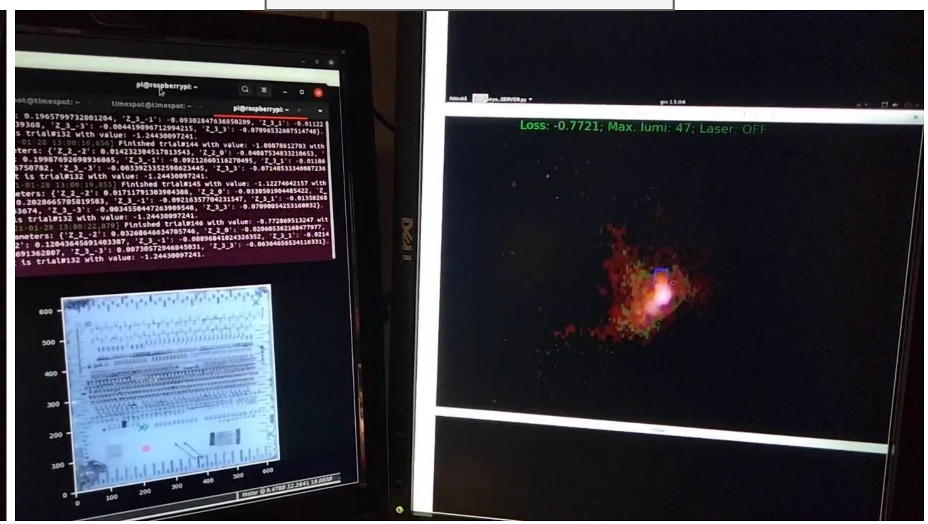
A computer-controlled **Spatial Light Modulator** is used, the wave phase is modified on a matrix of 800x600 pixels by rotating a liquid crystal array and the resulting **laser spot is acquired with a camera**.

A **Bayesian Optimization** algorithm is integrated in the loop to iteratively search for the optimal SLM configuration during lunch time.

At the beginning



5 minutes after



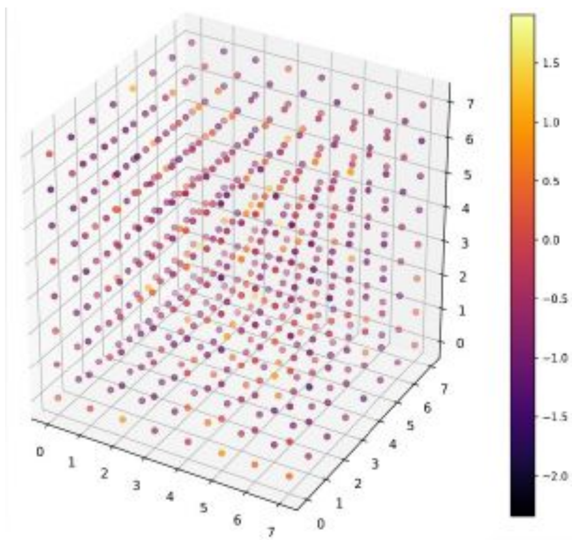
Studying phase transitions in gauge theories

Pure gauge theories are studied on lattice, for example, as a simplification of Lattice QCD.

The level of *“disorder”* in these lattice representations is a measure of the phase, which is different, for example for Quark-Gluon Plasma (disordered) or hadrons (ordered).

While it is easy to say whether a lattice is completely ordered or completely disordered, to study the phase transition an order parameter is needed.

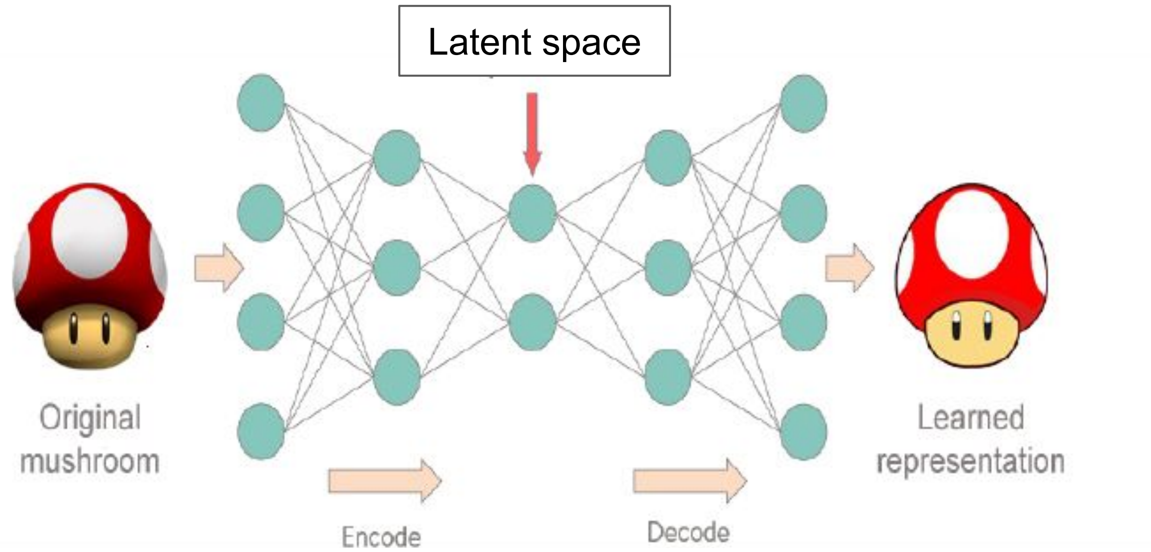
Unfortunately, the definition of this parameter in QCD is not trivial.



How much is this ordered?

Autoencoder architecture: *extract the key feature(s)*

Autoencoders are a widely explored class of algorithms for unsupervised machine learning applications.



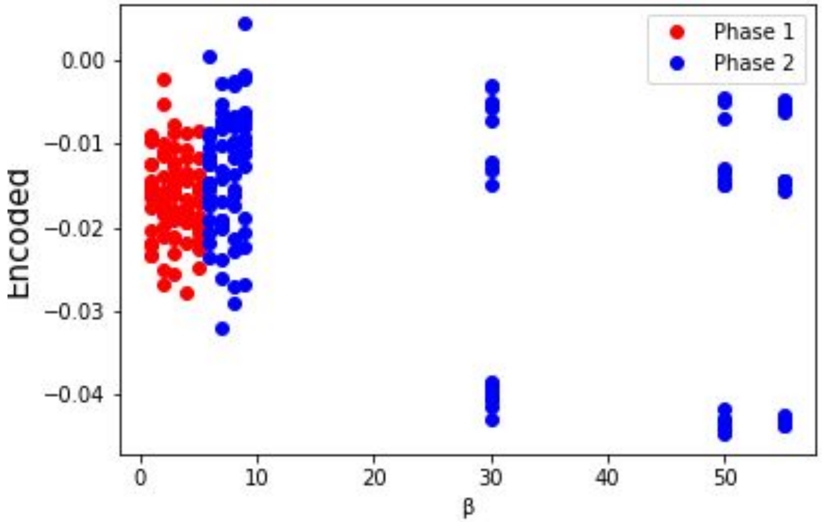
A neural network is trained to “reproduce” the input under some kind of constraint, for example synthesizing a (or few) variables that **condense the whole information** to “reproduce” a given data entry.

Back to Gauge Theory.

Assuming the most important “feature” distinguishing phases in the lattice is its degree of “order”,
we expect some kind of “order metric” to appear in the “latent space”.

Results for the pure gauge theory *(work in progress towards LQCD)*

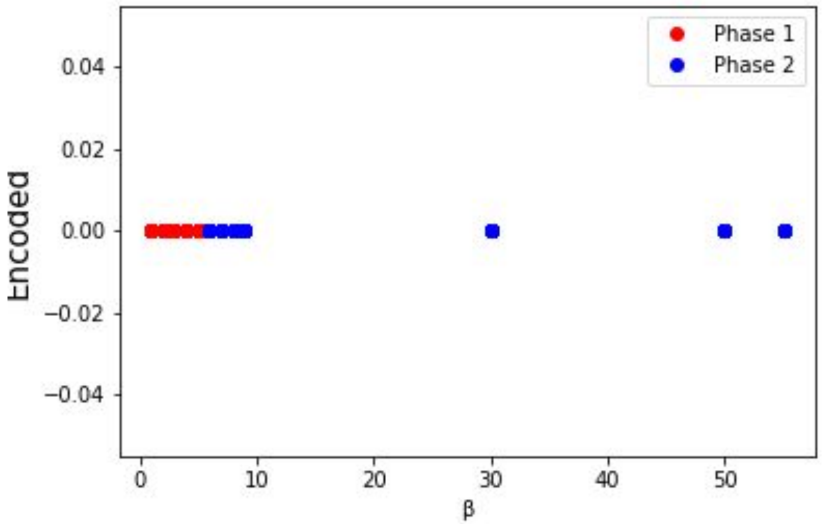
Here is the training of a simple autoencoder with one variable in the latent space, named "Encoded".



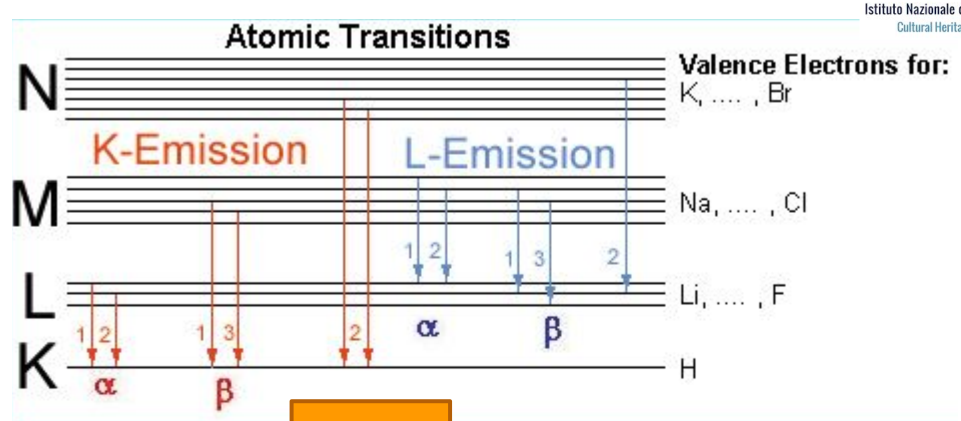
Higher temperature

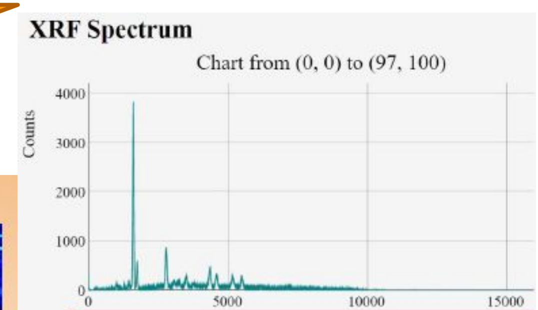
Once the network is able to distinguish the different phases it is sufficient to pin down the "parameter order" for the extreme cases, and the network does the rest.

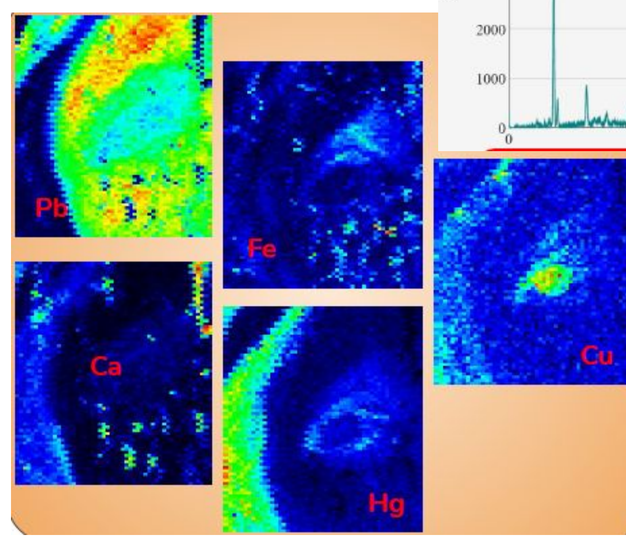
This is an example of "semi-supervised" neural network.

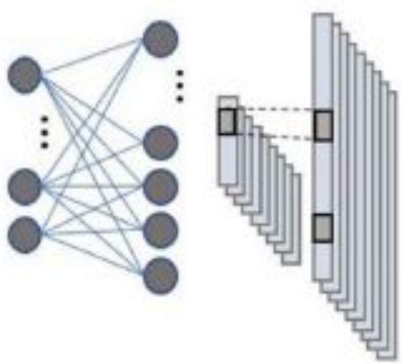


Digital reconstruction of peintures from XRF spectra









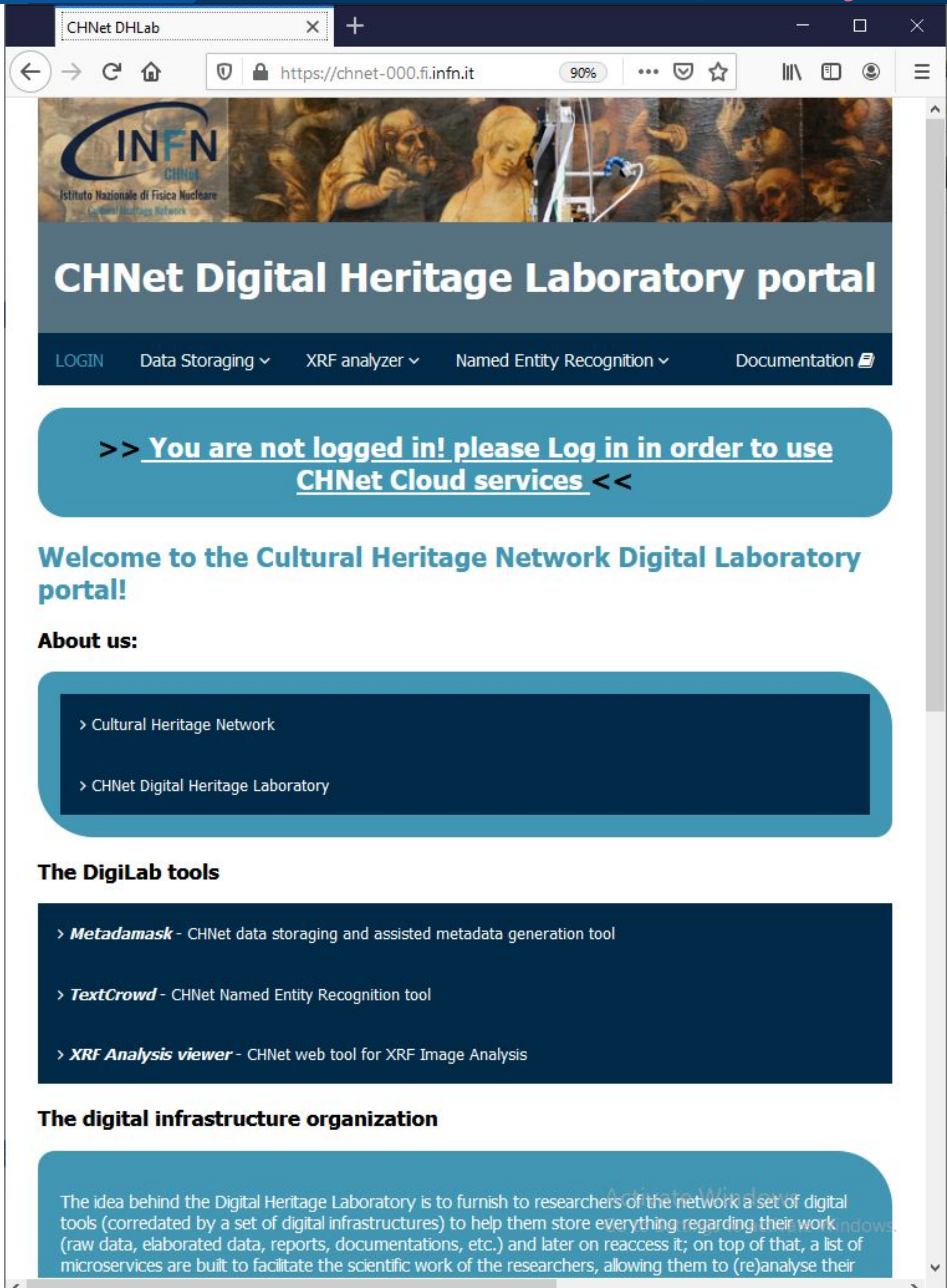
Once trained, this neural network will be able to reconstruct peinture layers coated in later interventions.

Cloud deployment

The digital reconstruction tool is planned to be released as a service in the Cloud-hosted “**CHNet Digital Heritage Laboratory**”.

This will require a wide set of important “machine learning” skills complementary to those used for model design and training.

Hence, this project will deepen our understanding of the **deployment techniques** and build experience with **Cloud and Web technologies**.



Focus on Medical Physics

Radiomic & Dosiomic Analysis

Apply a retrospective exploratory MR-CT-based radiomics and dosiomic analysis based on ML, to investigate imaging biomarkers of clinical outcomes in paediatric patients affected by medulloblastoma.
Annotated dataset was extracted from the **Careggi and Mayer databases**.

Features from MR-CT scans and dose distribution is tested as predictor for the overall survival, recurrence-free survival, and loco-regional recurrence-free survival after IMRT.

Database (MRI, Dose, Structures) → Feature extraction → Feature selection (PCA) → Random Forest Classification

Any evidence for radio-induced damage?

24 positive

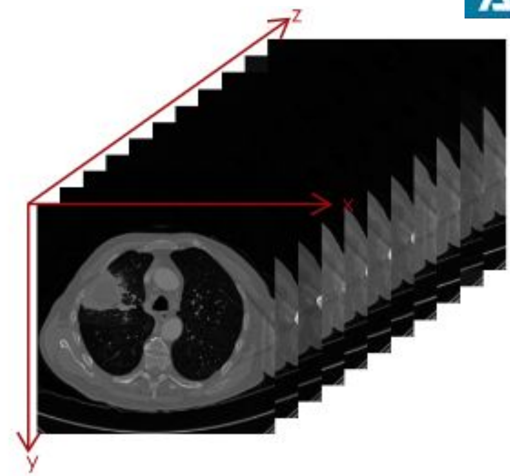
21 negative

pyradiomics
python + RADIOMICS

Feature set	**Accuracy** (5-folding)
Dose	0.44 ± 0.21
T1	0.56 ± 0.16
T2	0.59 ± 0.16
FLAIR	0.41 ± 0.07
Dose + T1 + T2 + FLAIR	0.62 ± 0.09

Accuracy is clearly limited by the statistics of the training sample.
Need to explore ***Data Augmentation*** techniques to enhance the performance.

Exploiting the 3D nature of small datasets of CT scans



Computer Vision has been revolutionized by 2D Convolutional Neural networks, the application of these algorithms to medical images opens to computer-assisted diagnosis and improved statistical predictions of treatment outcome.

But,

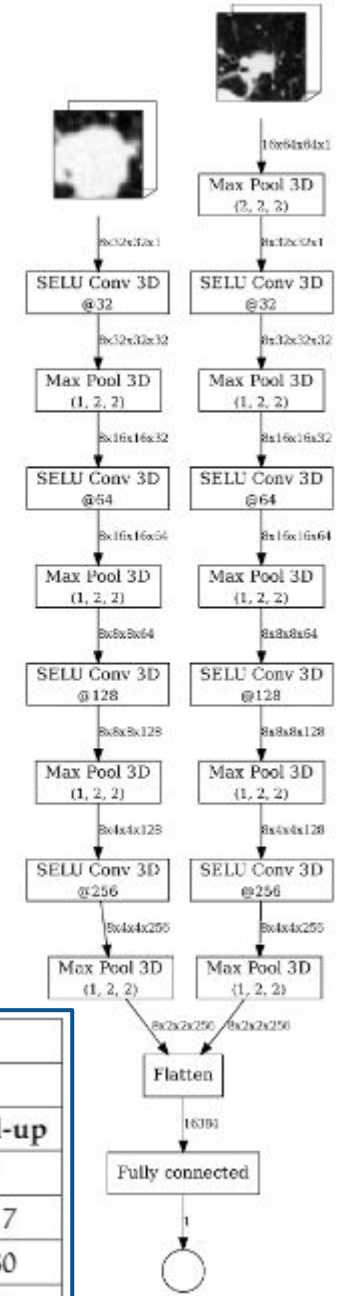
- CT scans are **3D objects**, the texture information on the third (longitudinal) dimension may be relevant to the diagnosis.
- **Labeled CT scans are extremely expensive**, training samples are therefore much smaller than in traditional Computer Vision.

The keywords are therefore:

3D CT scans, **Transfer Learning**, and **Unsupervised pre-training**

3D CT scans: *requires specialized hardware*

- Developed a 3D CNN for the classification of lung nodules into benign and malignant lung injuries.
- Tested the training procedure on a variety of hardware solutions spanning from multi-CPU architecture to the CINECA Marconi-100 supercomputer.
- Compared the performance with 2D CNN of similar depth.

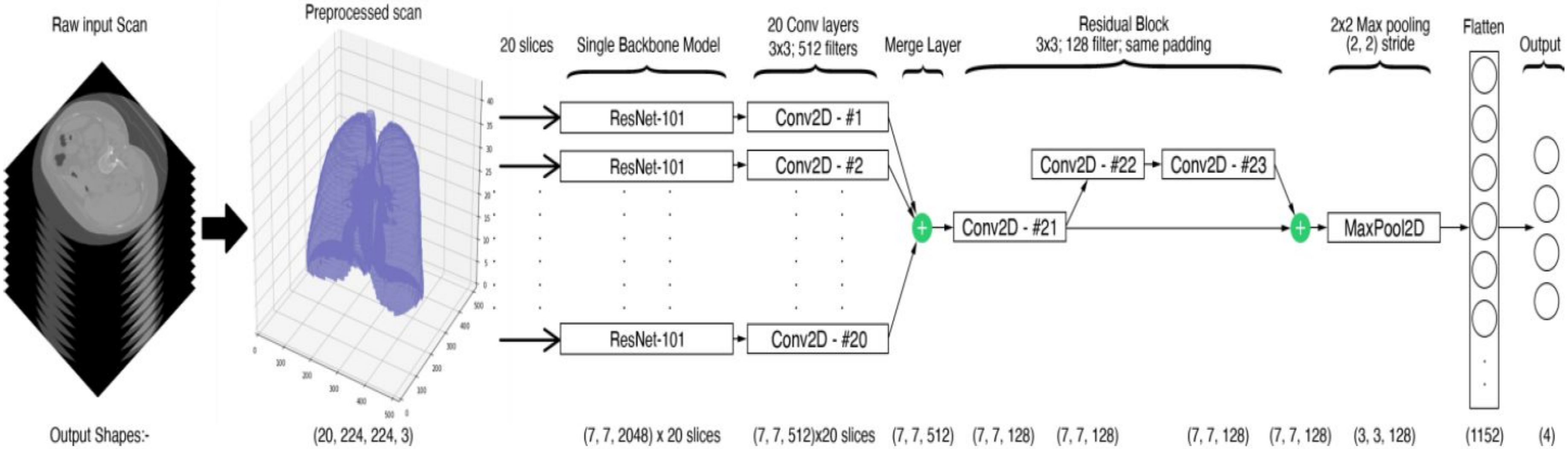


Conclusion

Training neural networks 3D CT scans require a huge amount of GPU RAM. A good single GPU with large RAM (~ 5 k€) performs almost as Marconi-100, and combining several GPUs on the same node does not help.

The systematic performance improvement over 2D CNNs is too small to justify the large increase in computing resources.

	Batch size = 2			
	Filters = 8		Filters = 16	
	Exec. time	Speed-up	Exec. time	Speed-up
acer-1	1194.03s	-	2027.96s	-
M100 (1 GPU)	5.34s	×224	6.4s	×317
M100 (4 GPUs)	7.95s	×150	8.82s	×230
Lb7	6.32s	×189	7.19s	×282

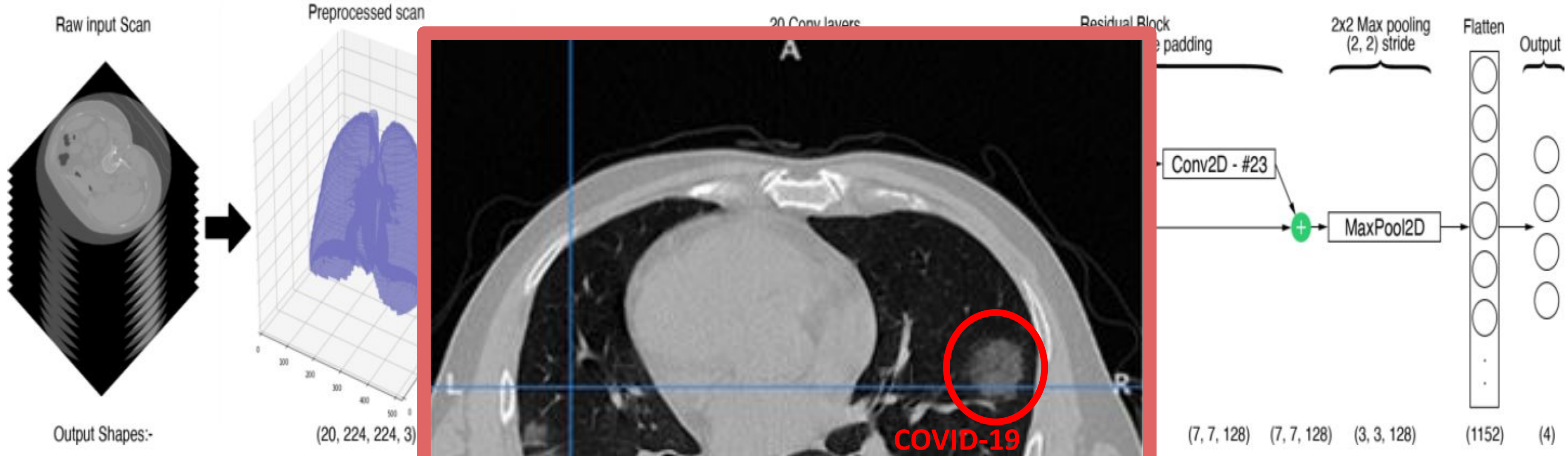


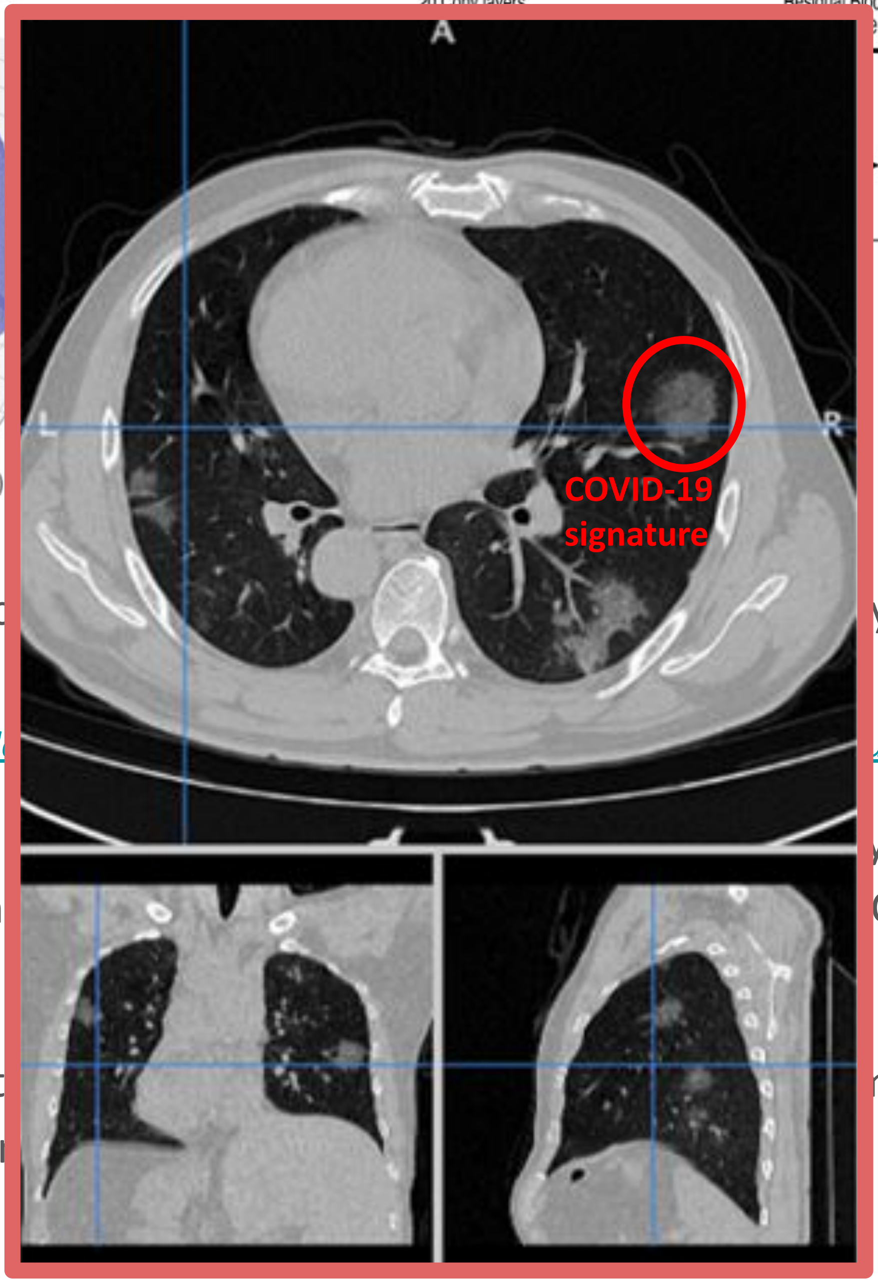
During Emilio's research project, an alternative architecture to analyse lung CT scans has been published in the context of Covid-19 detection.

[(Singhroihla et al. Deep Learning Assisted Covid-19 Detection using full CT-scans, November 2020)]

Instead of considering the texture in the longitudinal direction, they combine the high level feature of the single slice as determined using pre-trained 2D CNN networks with a hundred of layers.

This architecture and training strategy outperforms 3D CNNs with much looser constraints on the training hardware architecture.





During Emilio's researc … yse lung CT scans has been published in the

(Singhroihl … ull CT-scans, November 2020)

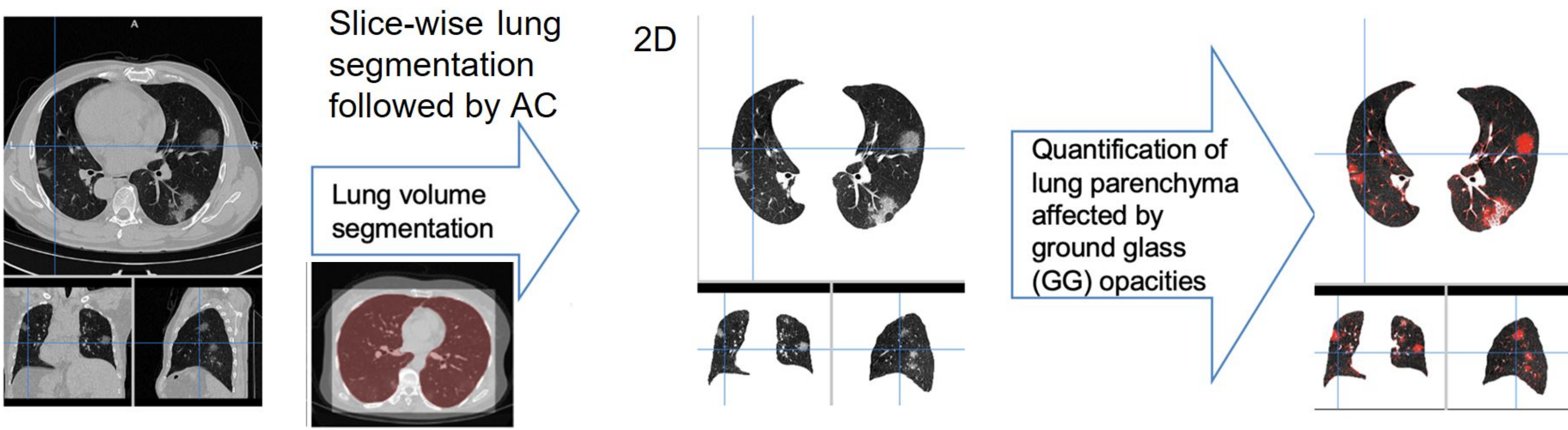
Instead of considering … y combine the high level feature of the sin … CNN networks with a hundred of layers.

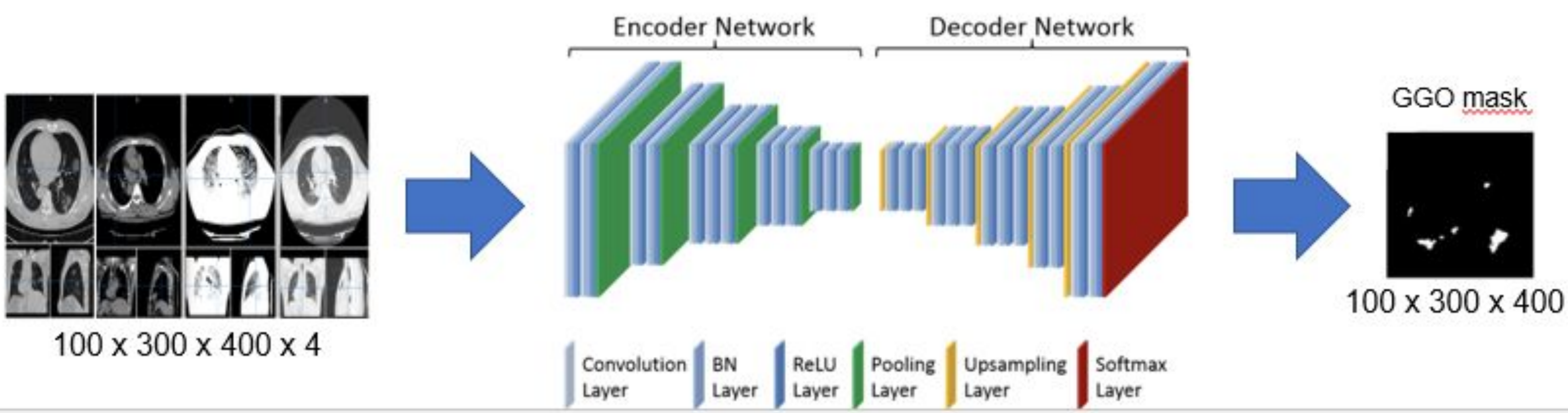
This architecture and t … much looser constraints on the trai

Covid-19 *Lung CT Lesion Segmentation Challenge*

Identification and segmentation of lung lesion due to Covid-19.

Dataset and context: https://covid-segmentation.grand-challenge.org





Performance

Dice score (DSC) is a statistic used to gauge the similarity of two samples X and Y.

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

Dice score obtained:

- 0.77 (training set)
- 0.62 (test set)

Not far from the world's best score!

(though generalization may improve)

Challenge Validation Phase Leaderboard

1st	qiaoyuanfang	8 Dec. 2020	0.7709 ± 0.1450
2nd	mfutrega	8 Dec. 2020	0.7687 ± 0.1427
3rd	LCSBmedAI	8 Dec. 2020	0.7687 ± 0.1700
4th	NWPUjpzhang	7 Dec. 2020	0.7678 ± 0.1493
5th	wangliwen1994	6 Dec. 2020	0.7677 ± 0.1698
6th	gx375047	6 Dec. 2020	0.7664 ± 0.1567
7th	timothy	3 Dec. 2020	0.7647 ± 0.1696
8th	Jsy	4 Dec. 2020	0.7645 ± 0.1597
9th	shaun	8 Dec. 2020	0.7628 ± 0.1419
10th	King_HAW	2 Dec. 2020	0.7627 ± 0.1756

Conclusion

Machine Learning and Cloud infrastructures are reshaping the world of scientific computing.

While we have certainly not the strengths to compete with the gigantic computing companies developing these technologies, we may profit from these emerging technologies in our daily scientific activities.

Under this perspective we have:

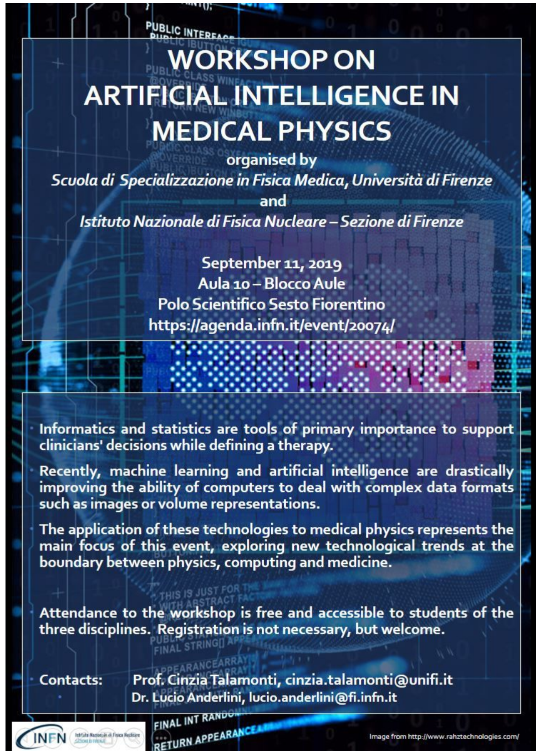
- a local discussion group (machine.learning@fi.infn.it)
- next meeting on **February 10th, 9 AM –** agenda.infn.it/event/25698
- a national coordination organizing the access to HPC resources, tutorials and building a knowledge base of applications of machine learning to scientific problems
- an INFN-dedicated Ph.D. student within the Smart Computing doctoral program (currently, Matteo Barbetti)

Don't hesitate to get in touch!

Thanks to all the people contributing, and in particular to Leandro (Servizio Calcolo) for the great support (e.g. *JupyterHub, Zammad and INFN Cloud liaison*)

Training activities, workshops and dissemination

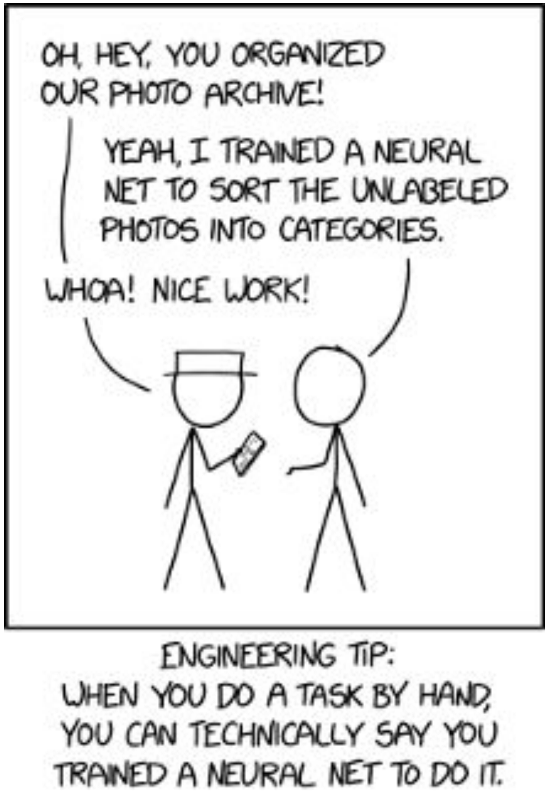
Ampio programma di collaborazioni con Istituzioni ed Associazioni interessate alla divulgazione ed alla formazione in ambito machine learning (CdL, CdLM in Fisica; CdLM Informatica; Ingegneria Informatica, DINFO; AISF; AIFM)...



12 febbraio 14.30 - 16.30
26 febbraio 14.30 - 16.30
15 marzo 14.30 - 16.30
24 marzo 14.30 - 16.30

ml INFN
Hackathons

Spare slides

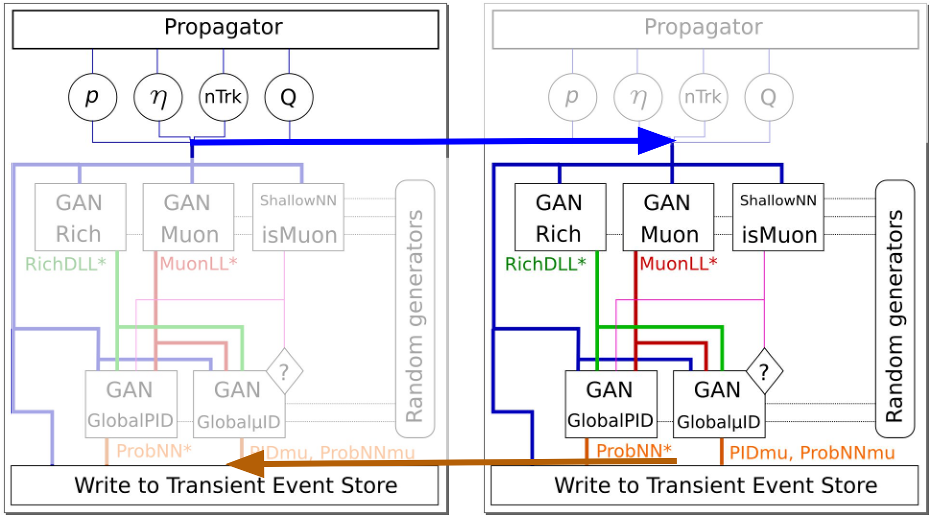


Fast simulation of the LHCb detector: *deployment*

Integrating the deployed models in the C++ of the LHCb project is challenging.

We integrated Tensorflow C APIs with Gaudi, the LHCb architecture, and with those we are able to run whatever tensorflow **model from whatever LHCb application**.
The different threading system is however a bottleneck.

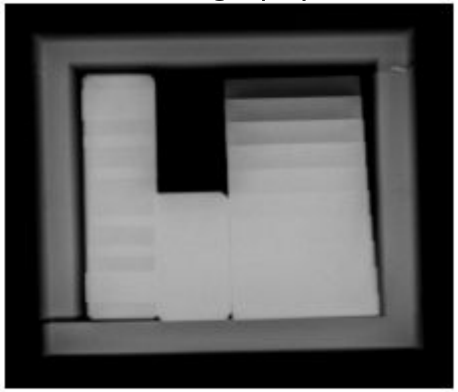
A solution developed by the ATLAS colleagues and that we are reinventing at LHCb (with good reasons?) **is to rewrite the logic of the NN algorithm in C or C++ and compile it together with Gaudi**, picking only the trained weights from Tensorflow.
While much faster, it only allows to import models for which a “translation” exists.



The main difference between ATLAS and LHCb approaches is that **LHCb is targeting plain C** with disabled name mangling to allow a transparent usage of NN and BDTs algorithms but introducing limitations on the shape of input and output tensors.

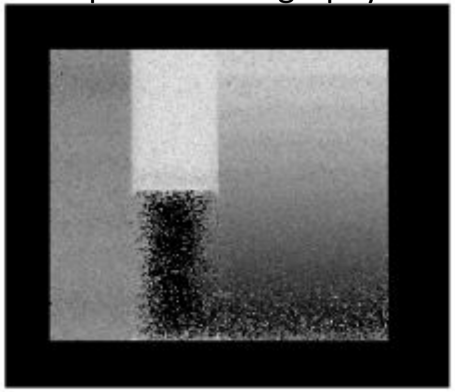
Enhancing readability of proton CT scans

Photon radiography

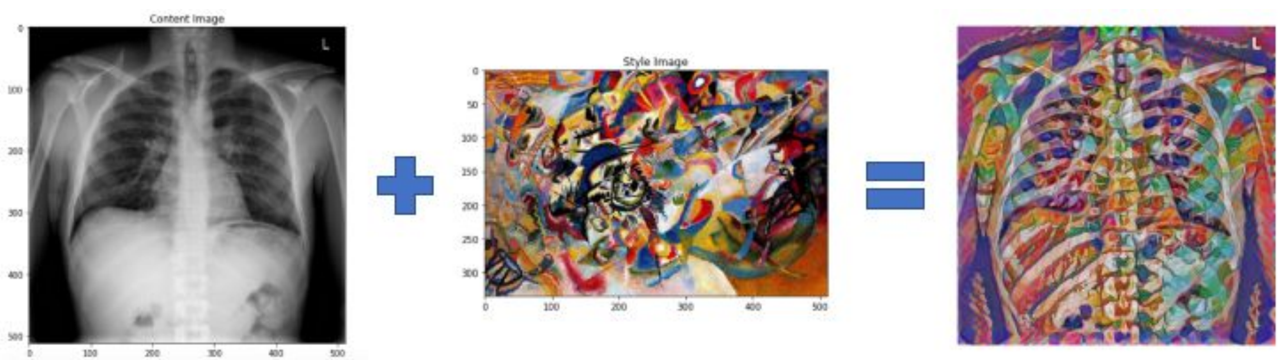


Proton CT scans (pCT) allow for enhanced accuracy in the dose calculation for proton-therapy treatments, but because of multiple scattering, the acquired images are nosier, and more difficult to interpret for clinician's eye trained on standard CTs.

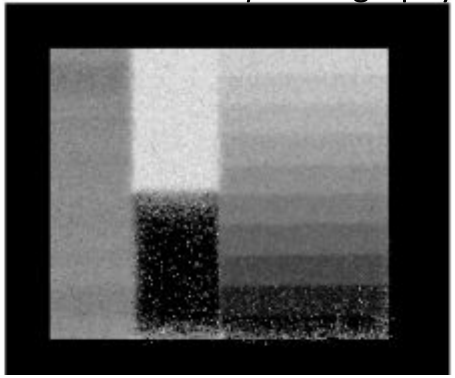
Raw proton radiography



Style transfer may allow to clean the acquired picture and, more importantly, to convert it into an image as readable as a photon radiography.



Reconstructed *p* radiography



Style training is however computationally challenging and very demanding in terms of model tuning.