Toward the End-to-end Optimization of Detector Design with Differentiable Programming

Tommaso Dorigo

INFN, Sezione di Padova

https://agenda.infn.it/event/27167/

2021

Toward the End-to-End Optimization of Particle Detectors with Differentiable Programming

by Tommaso Dorigo (INFN - PD)

Thursday 27 May 2021, 15:00 → 17:00 Europe/Rome

Online -Zoom Platform (https://infn-it.zoom.us/j/89033677040?pwd=aEl3NFpoYVBOMmVzdEcrNzI5akhaQT09)

Description Particle detectors are complex instruments, as their design involves the specification of hundreds of geometric and material parameters. That choice in the past has been informed by robust paradigms (redundancy, symmetry, "track-first, destroy later"), which allowed us to build highly performant particle physics experiments. However, in the impossibility of optimizing our design choices to the final, true goals of our instruments (the highest sensitivity to a flagship parameter, or the discovery reach to some relevant phenomenon), we have until now consistently relied on manageable proxies as figures of merit: e.g., highest resolution, lowest backgrounds. While sensible, in a high-dimensional feature space of possible choices that modus operandi corresponds to potential huge losses of performance on those true goals.

A realignment of goals and design choices may be pursued by relying on modern deep learning techniques. These allow us to produce fully differentiable pipelines where a model of the instrument, the pattern recognition procedures, the cost constraints, and the detector-related systematic uncertainties can be considered all together, and where the true experimental goals may be encoded in a carefully defined objective function. The latter can then be maximized by stochastic gradient descent, achieving a full end-to-end optimization of the design.

In this presentation the above concepts will be clarified with a few examples, and a summary of the research program of the MODE collaboration (mode-collaboration.github.io) will be offered.

Seminario_Dorigo.p...

Organized by Giuseppina Salente, Andrea Longhin, Tommaso Dorigo

Tommaso Dorigo tommaso.dorigo@pd.infn.it

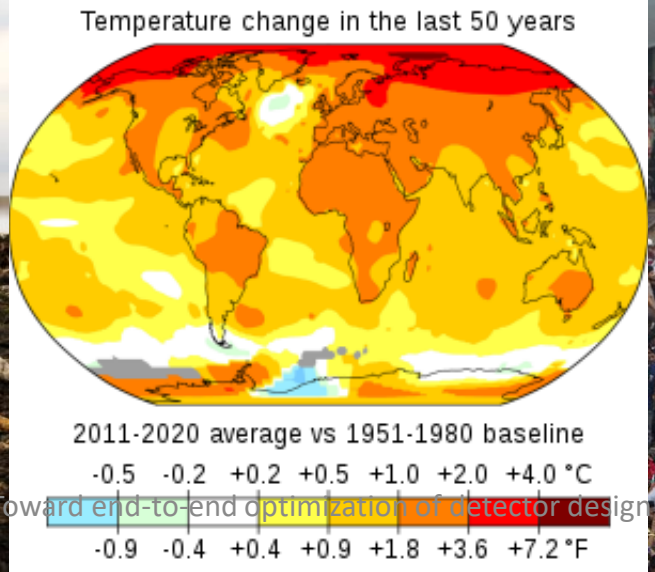
Foreword/1

The recent publication of the 2020 update of the European Strategy for Particle Physics (EUSUPP) [1] encourages feasibility studies for new large, long-term projects which will once again push our technological skills to their limits.

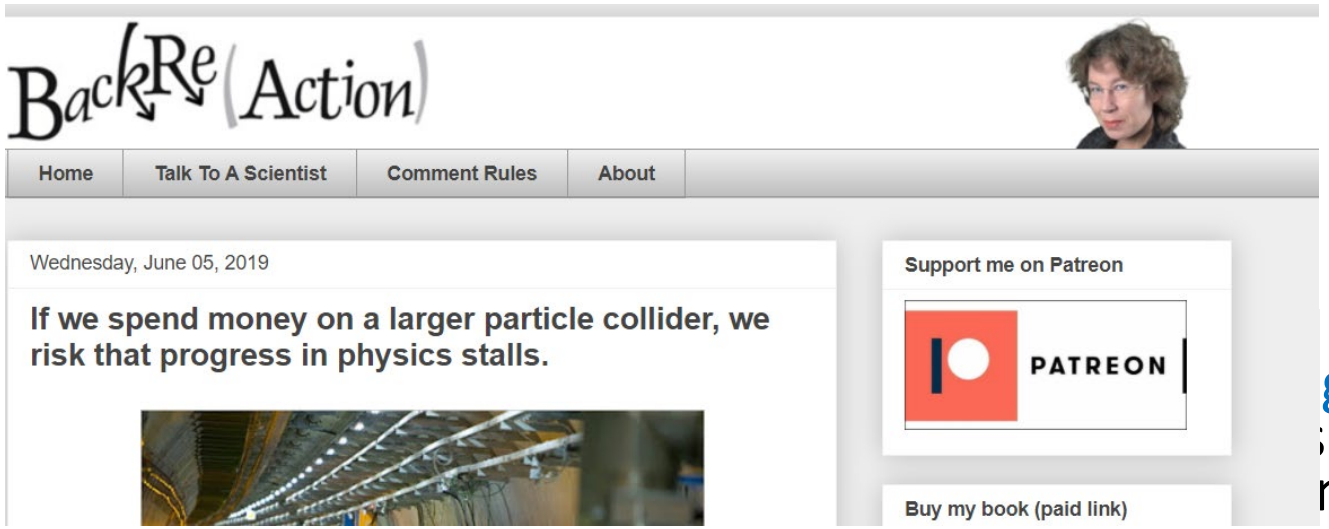
Foreword/2

The recent publication of the 2020 update of the European Strategy for Particle Physics (EUSUPP) [1] encourages feasibility studies for new large, long-term projects which will once again push our technological skills to their limits.

At the same time, humanity faces unprecedented global challenges (climate change, pandemics, overpopulation) which demand the use of our resources to seek solutions through applied science innovations, rather than investing in fundamental research.



Foreword/3



global challenges (climate change,
e of our resources to seek solutions
nvesting in fundamental research.

Furthermore, there are indications that wide sectors of society no longer consider the furthering of our understanding of matter at the smallest distance scales, or other projects that require large and coordinated effort and significant funding, a top priority [2].

In this situation, **ensuring the maximum exploitation of any resources spent on fundamental research is a moral imperative,** and it may be a key to ensure that the long-term projects envisioned by the EUSUPP may be undertaken and sustained.

The Status Quo

The design of new detectors for particle physics, astro-particle physics, neutrino physics, and HE nuclear physics applications in the past 50 years systematically leveraged the most performant available technologies for particle detection, often fostering significant further advancements and spin-offs [3].

Yet one observes that **the crucial underlying global paradigms** of experimental design **have remained mostly unchallenged** across decades.

- "**Track first, destroy later**": charged particles can be traced by their ionization in low-material-budget elements, while neutral ones must undergo destructive interactions in a calorimeter to be detected → standard setup of experiments in HEP typically involves a low-material tracker followed by a thick calorimeter.
- A focus on significant **redundancy in our detection systems**, ensuring robustness and enabling cross-calibration of the resulting measurements → this by itself is **something to grow out of**
- **Symmetrical layouts**, both respecting physics conservation laws and attempting to simplify reconstruction → **no guarantee of optimality** when symmetry breaking mechanisms are at play

While these choices have served us very well for a long time, **they are not meant to be "optimal": *i.e.***, they **do not directly maximize a high-level utility function**, such as the highest discovery reach for a physical process, or measurement precision for a given physics parameter.

Optimal for what?

The reason why detectors are complex is not only that the studied physics is complex: a lot has to do with Science being a demanding job. We want to study *everything* and do it *better* than previously

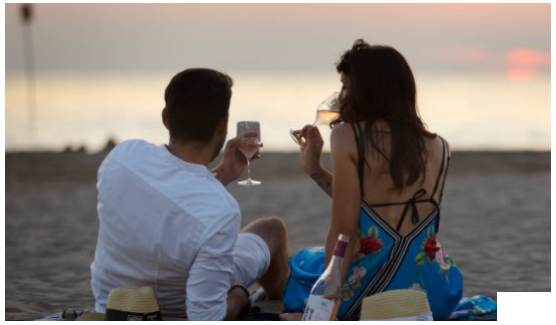
So, what does it mean for a detector to be *optimal*?

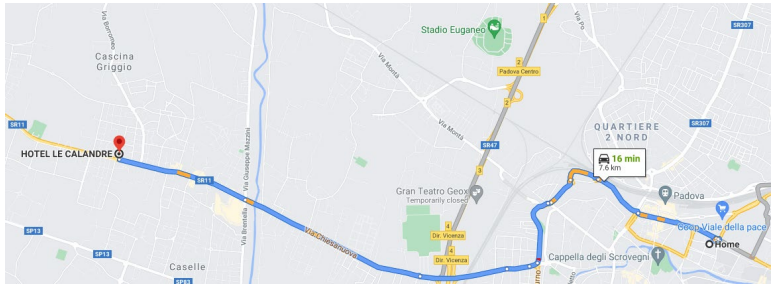
What loss function do we aim to minimize?

Does it make sense to speak of a single utility function?

Concerning the last question: I am convinced that it does, and I will try to convince you in the next few slides.

Recipe for a perfect dinner



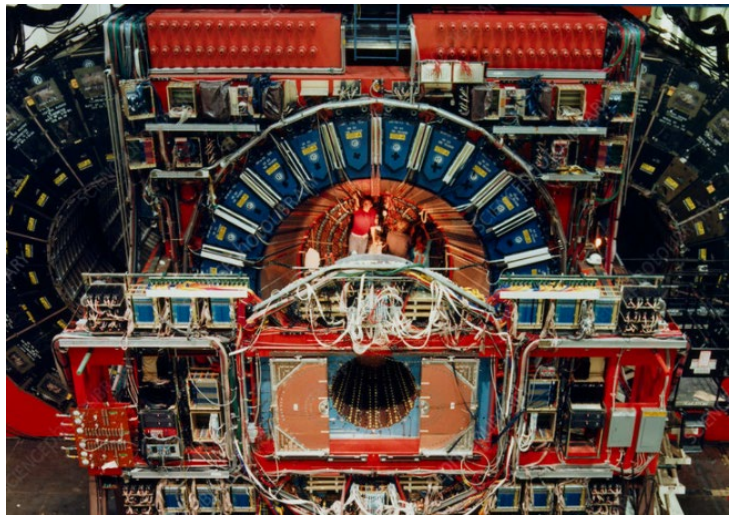


We are not alien to confidently taking complex decisions in a multi-objective space. We actually do it routinely...

Of course, we are not deterred by knowing that our optimization target is not universal

Recipe for a perfect trigger

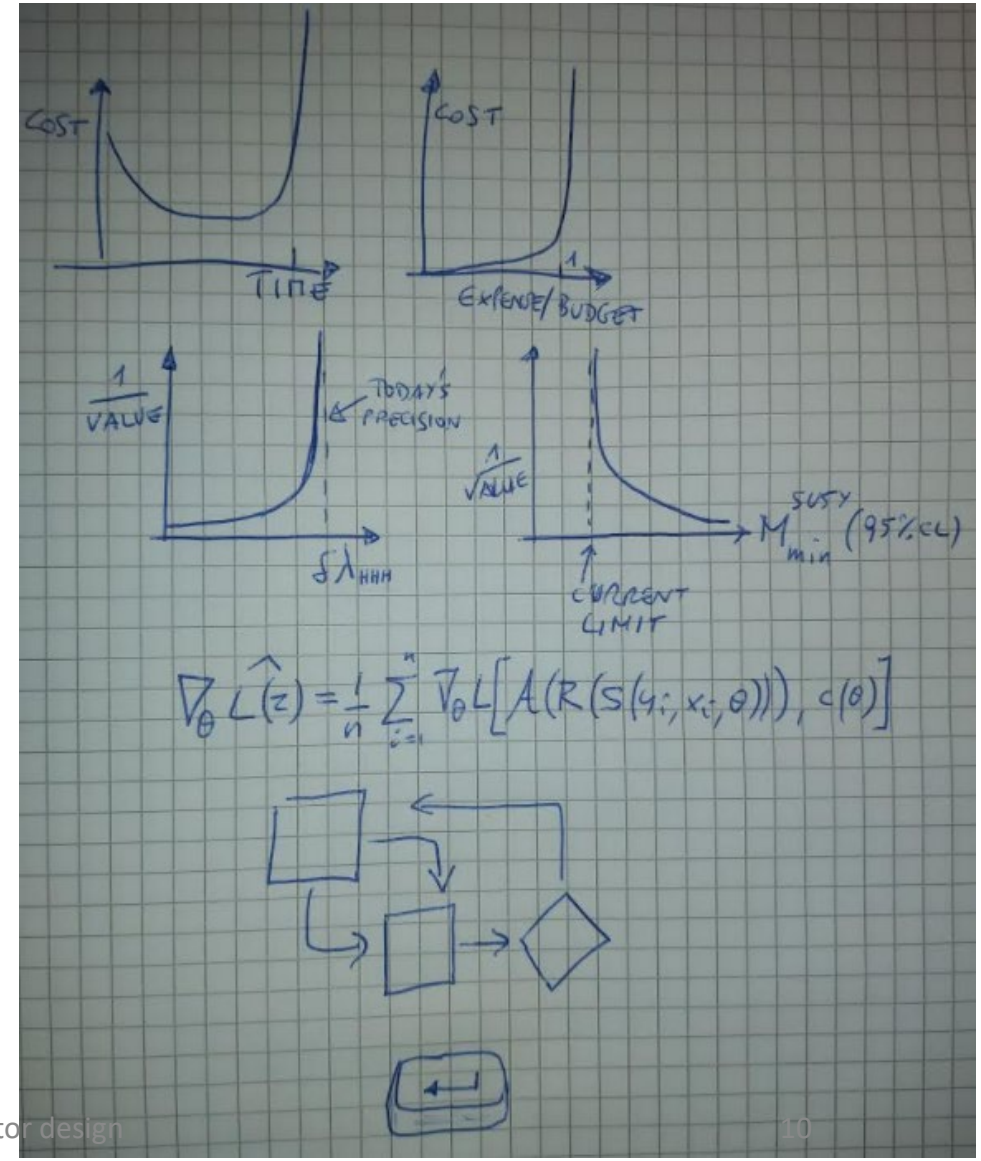
Similarly, we are actually *used* to create multi-target optimization strategies, e.g. when we allocate resources for the trigger menu of a collider detector.



Consider CDF, Run 1 (1992-96): taking in a rate of 300 kHz of proton-antiproton collisions and having to select 50 Hz of writable data created some of the most heated scientifically-driven, rationally motivated, painfully well argumented debates I ever listened to. The top quark had to be discovered, but it was not the only goal of the experiment...

Recipe for a perfect detector

1. Assess your total **budget** and **time**-to-completion
2. Model as a steep function the **cost** of overriding budget or time
3. Assess the **scientific impact** of each achievable scientific results, optionally *as a continuous function of their precision*
4. **Create a differentiable model** of the geometry, the components, the information-extraction procedures, and the utility function
5. Construct a pipeline with those modules, enabling backpropagation and gradient descent functionality
6. Let the chain rule of differential calculus do the hard work for you



Makeshift surrogates of objectives

When we design the sensors for a tracking device, operate choices on budget allocations, define requirements for the various resolutions of detection elements, or choose composition and layout of active and passive material of calorimeter cells, **we are implicitly trying to find an optimal working point in a loosely-constrained feature space of hundreds of dimensions**. Such a task is clearly super-human.

Because of that, we set our aim on makeshift surrogates of our real objectives.

- E.g., we might desire our objective to be "*the highest precision on the Higgs boson self-couplings our budget can ensure*", but all we can do is stick to useful proxies suggested by past experience, and rather focus on the "*highest achievable energy resolution for isolated photons*", ignoring the rest of the parameter space
- In a neutrino detector this would instead sound as "the highest precision on $\boldsymbol{\theta_{13}}$ we can get", when the focus becomes instead maximizing the number of interactions and reducing the background level.
- Our simulations only allow us to probe the result of specific choices, not to map interdependencies.

Makeshift surrogates of objectives

When we design the sensors for a tracking device, operate choices on budget allocations, define requirements for the various resolutions of detection elements, or choose composition and layout of active and passive material of calorimeter cells, **we are implicitly trying to find an optimal working point in a loosely-constrained feature space of hundreds of dimensions**. Such a task is clearly super-human.

Because of that, we set our aim on makeshift surrogates of our real objectives.

- E.g., we might desire our objective to be "*the highest precision on the Higgs boson self-couplings our budget can ensure*", but all we can do is stick to useful proxies suggested by past experience, and rather focus on the "*highest achievable energy resolution for isolated photons*", ignoring the rest of the parameter space
- In a neutrino detector this would instead sound as "the highest precision on θ_{13} we can get", when the focus becomes instead maximizing the number of interactions and reducing the background level.
- Our simulations only allow us to probe the result of specific choices, not to map interdependencies.

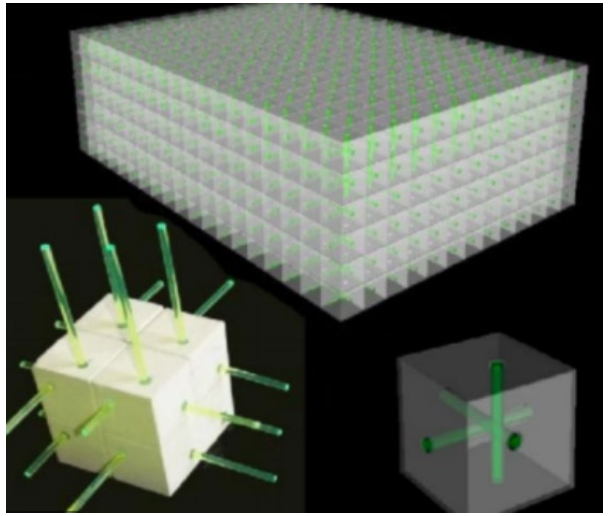
Evolving from this modus operandi to directly goal-informed decisions enables potentially enormous performance gains.

The design space is large – no, larger

New technological advancements are crucially enabling a better optimization of our instruments by reducing the cost of complex layouts.

- 3D printing of scintillation detectors are being explored for neutrino physics [4]
- Very thin layouts of resistive AC-coupled silicon detector elements may provide large gains in spatial and temporal resolution [5].

The geometry space has become larger and more complex to explore.



Higher-performance demands are also arising:

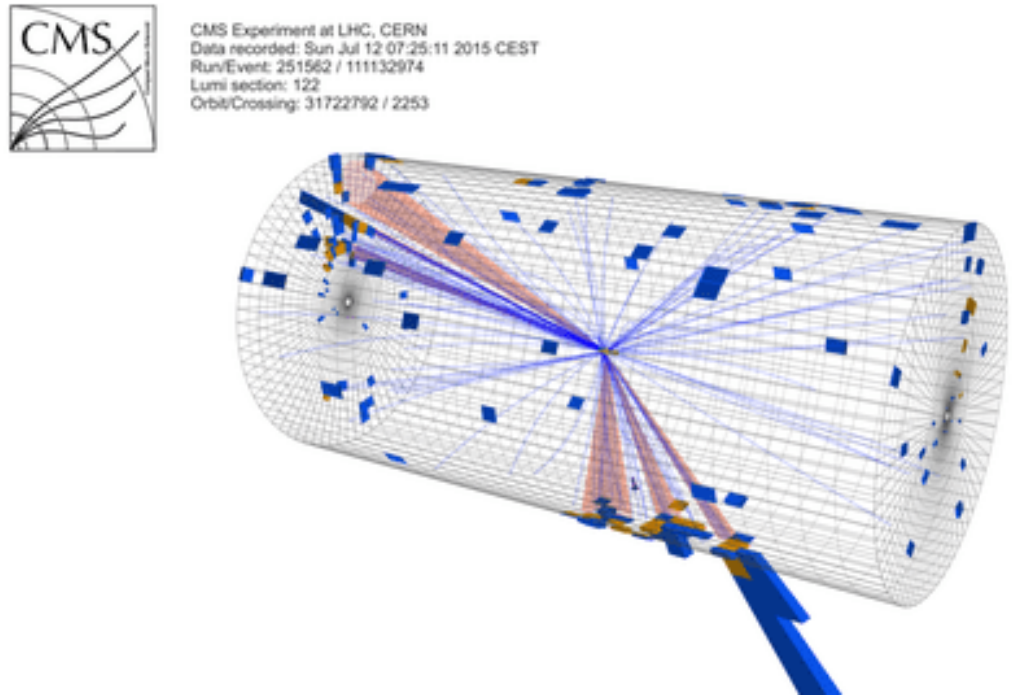
- Tracking in dense environments requires AI solutions
- Hitting the neutrino floor (*e.g.* at SuperCDMS) may require new paradigms for DM searches
- As we move fundamental physics research to space, payload and power consumption become major constraint, making applications especially sensitive to tough design decisions
- Boosted jet tagging at high p_T –all the rage for NP searches at the LHC –demands us to invest in more granular, higher-performance hadron calorimeters

(Background: boosted decays and fat jets)

Energetic LHC collisions may produce heavy objects with large momentum (top quarks, or W, Z, H bosons). When these decays, they usually yield a collimated stream of particles – a single hadron jet.

A number of techniques allow the extraction of features sensitive to the heavy object decay

The point is however that high granularity and effective identification of constituents in dense environments has become unavoidable



Above: a top-pair decay produces two fat jets, where the individual subjects are visible

Speaking of calorimeters…

High granularity has thus become a compelling requirement.

- The CMS HGCAL detector [6] is a step in that direction; its design will improve by a large margin usable information about the showers, development, pointing, and composition.
- For different reasons, similar developments and improvements are planned for other projects (*e.g.,* CALICE [7] or CaloCUBE [8]).

However, an end-to-end optimization of the design of such instruments has not been attempted yet; **nor have models of the future potential of machine learning in pattern recognition been considered so far in the design phase**

As a telling example, the HGCAL detection elements are arranged in a hexagonal symmetry which offers construction benefits but significantly complicates the most common imaging techniques employing convolutional neural networks (CNN) to shower reconstruction. While solutions to this issue do exist (*e.g.*, see [9]), this is an example of misalignment between design and potential exploitation.

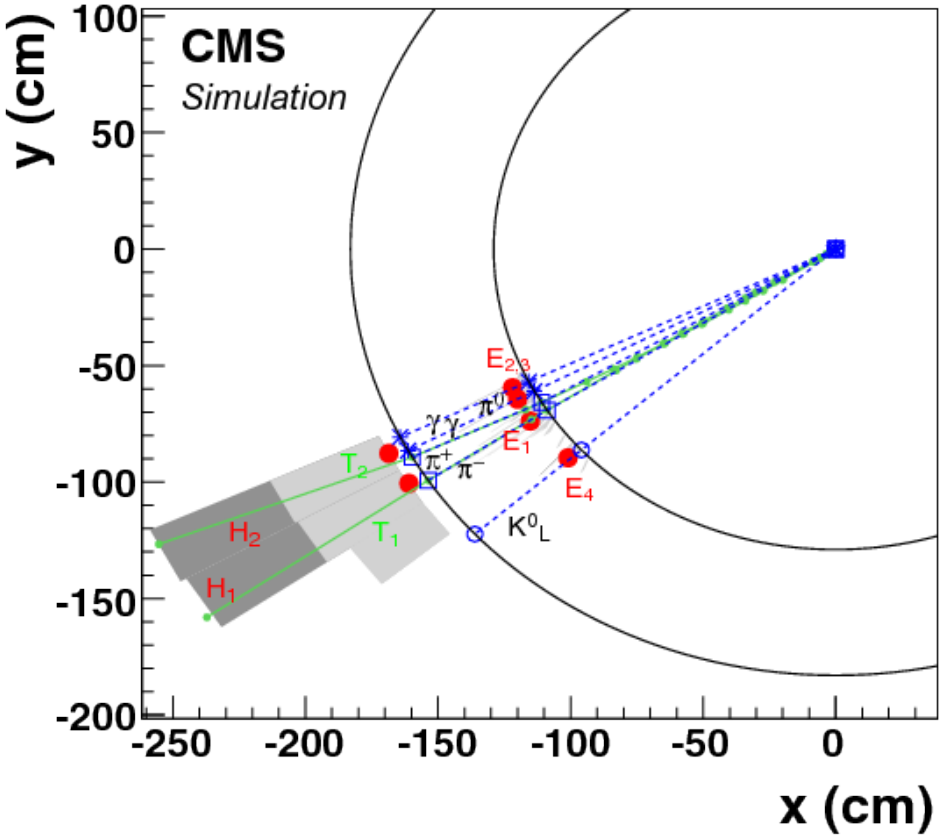
One further note on calorimetry

Take the LHC experiments for a telling example. CMS was originally endowed with a less performant hadron calorimeter than ATLAS. But hadron calorimetry ended up being crucial for a number of new physics searches involving boosted jets.

CMS regained the lost ground through the high performance of its “particle flow” reconstruction algorithm [10].

This was only possible thanks to the high magnetic field integral of CMS, which spreads out charged particles of different momenta within jets, easing their matching to calorimeter deposits.

This *post-hoc* exploitation of the solenoid characteristics, whose original specifications were rather driven by compactness and transverse momentum resolution of charged particles, is a **striking example of how the combined search of hardware and software solutions may be proficuous** to inform the optimization of a modern particle detector.



A hybrid calorimeter?

As particle flow techniques allow the tracing of individual particles and the complete reconstruction of dense, collimated jets, we must have more of that:

- Optimizing the design of a detector for a long-timescale project based on reconstruction capabilities which will be available in the future [11] seems the right thing to do, if we can pull that off (more on this *infra*)
- Integrating tracking and calorimetry layers may improve the «image» reconstruction of energetic hadronic jets, shown to be crucial for high-mass new particles
- Measuring muons from their radiative loss in a dense environment using convolutional neural networks was first shown to be viable (J. Kieseler, CERN) by «accident», as a tangential observation of NN outputs from shower reconstruction in the phase-2 CMS endcap calorimeter.
  → This needs to be pursued for high-energy collisions (see *infra*)
- Nuclear interactions have always been dreaded in a tracker, but in combination with calorimetry they may strengthen particle-ID (using probabilistic information coming from nuclear cross sections of different species)
  → this may be of special interest to a number of applications

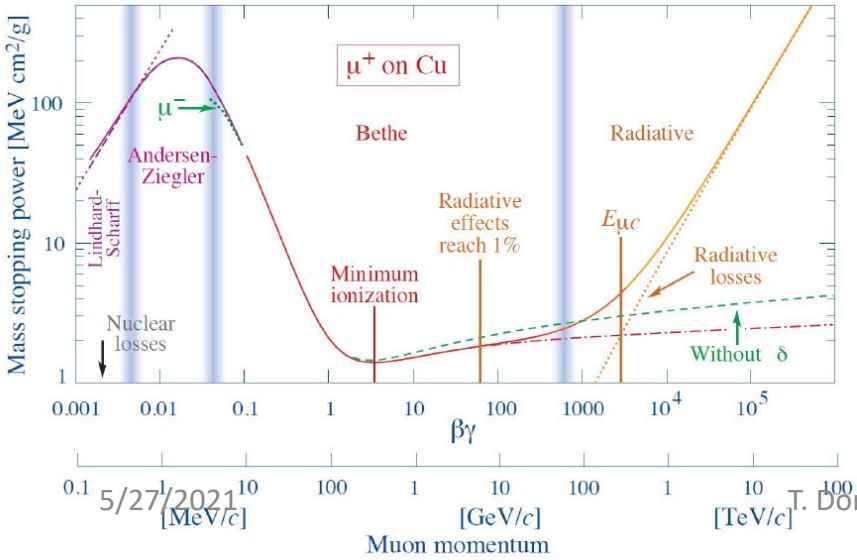
Muon energy measurement in a calorimeter?

Muons interact with matter by ionization, pair production, bremsstrahlung, and photonuclear reactions. The E loss is dominated by the high-end of the Landau distribution (knock-on electrons).

The total release is very modest and stochastic, so we have to rely on magnetic bending for inference on muon momentum

Bending measurements break down for TeV energies: in 2T, a 1 TeV muon traversing 2m of field is deflected by less than a mm

→ resolution scales: *e.g.,* in ATLAS **σ(p)/p = 0.2 p**



Left: mass stopping power for positive muons in Cu,showing the radiative energy loss onset above 1 TeV

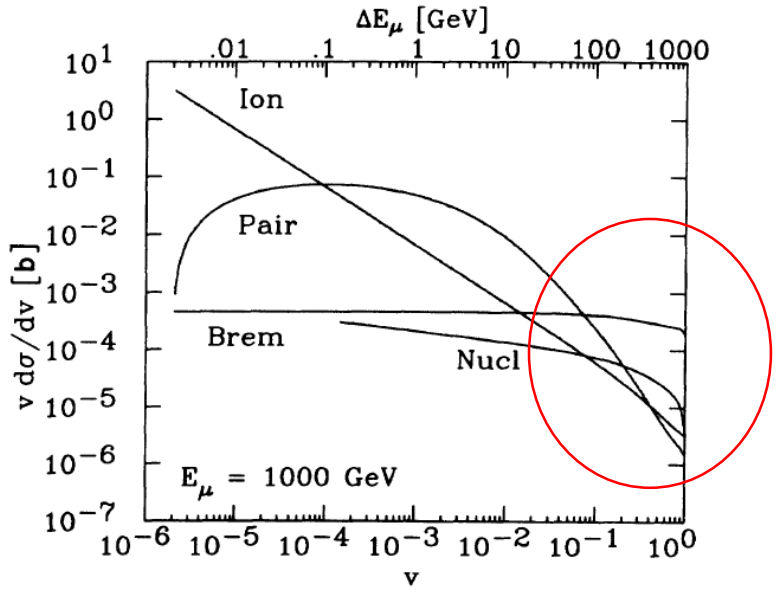


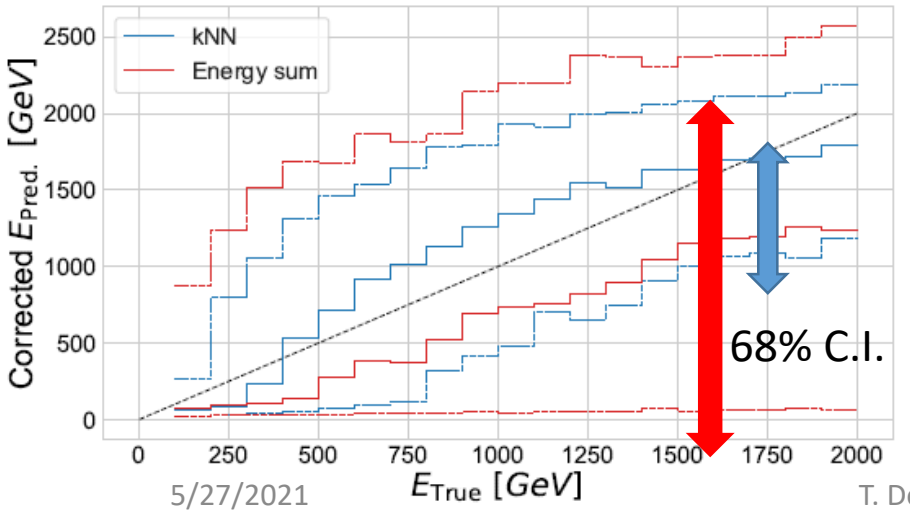
FIG. 3. The ionization, bremsstrahlung, pair production and photnuclear cross sections of a 1000-GeV muon incident on an iron atom. The top scale gives the muon energy loss ΔE_μ, which corresponds to the fractional muon energy loss v. Note that the ordinate is the logarithmic derivative $d\sigma/d(\ln v) = d\sigma/d(\ln \Delta E_\mu)$.

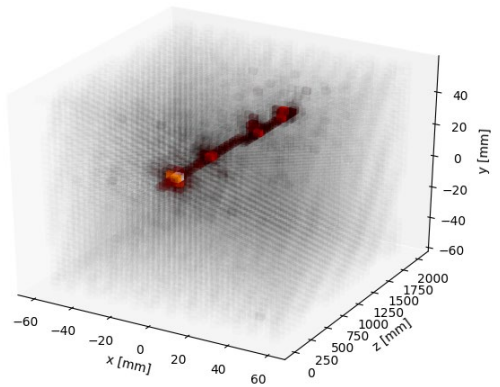
(From a CCFR study [22])

Muon energy measurement in a calorimeter?

In a preliminary study, we showed that resolutions of 30-35% are achievable for 2 TeV muons in a highly granular, homogeneous calorimeter [12]

- Genetically breeded kNN learners used for this study
- Spatial information proven to be crucial for the task (blue vs red double arrow for 68.3% CI, see below, left)





arXiv:2008.10958v1 [physics.ins-det] 25 Aug 2020

Muon Energy Measurement from Radiative Losses in a Calorimeter for a Collider Detector

Tommaso Dorigo[1], Jan Kieseler[2], Lukas Layer[3], and Giles C. Strong[4]

[1,3,4]INFN, Sezione di Padova
[2]CERN
[3]Università di Napoli "Federico II"
[4]University of Padova

August 25, 2020

Abstract

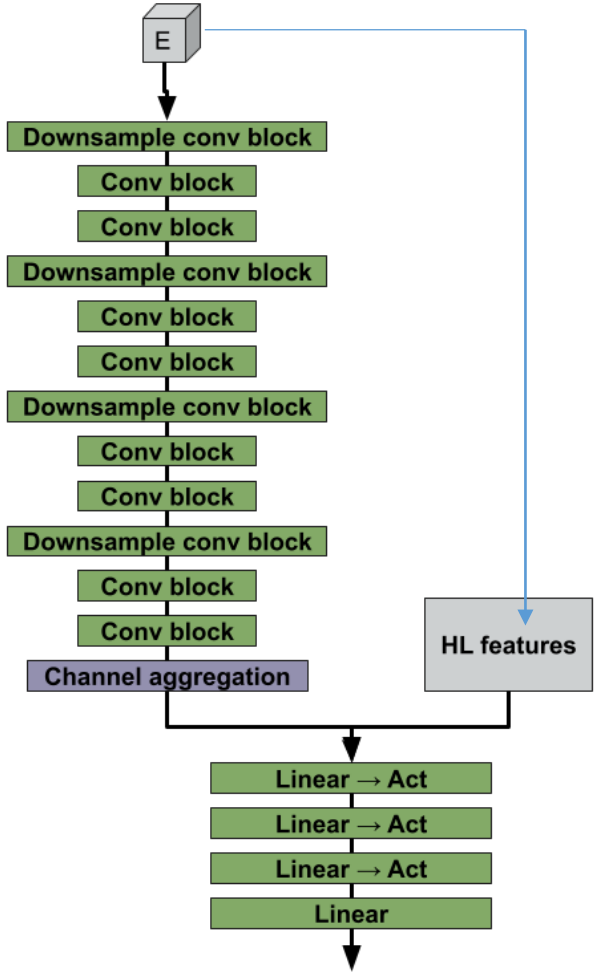
The performance demands of future particle-physics experiments investigating the high-energy frontier pose a number of new challenges, forcing us to find new solutions for the detection, identification, and measurement of final-state particles in subnuclear collisions. One such challenge is the precise measurement of muon momenta at very high energy, where the curvature provided by conceivable magnetic fields in realistic detectors proves insufficient to achieve the desired resolution.

In this work we show the feasibility of an entirely new avenue for the measurement of the energy of muons based on their radiative losses in a dense, finely segmented calorimeter. This is made possible by the use of the spatial information of the clusters of deposited photon energy in the regression task. Using a homogeneous lead-tungstate calorimeter as a benchmark, we show how energy losses may provide significant complementary information for the estimate of muon energies above 1 TeV.

(Soon to be published) results with CNNs

The study shown *supra* has now been extended to higher energy and by use of a customized deep learning architecture, which combines convolutional blocks and dense layers using both high-level features and raw «image-like» energy deposits in 3D space

Of relevance is the point that the pattern of radiation deposits contains information useful to regress to true muon energy



(Soon to be published) results with CNN

The study shown *supra* has now been extended to higher energy and by use of a customized deep learning architecture, which combines convolutional blocks and dense layers using both high-level features and raw «image-like» energy deposits in 3D space

Of relevance is the point that the pattern of radiation deposits contains information useful to regress to true muon energy

Results (right) show that one can recover 20-25% resolution for muons of up to 4 TeV by combining the radiation loss information with curvature information (here assumed is a relative momentum resolution of 20% from magnetic bending at 1 TeV, as *e.g.* quoted by ATLAS in mid-rapidity region)

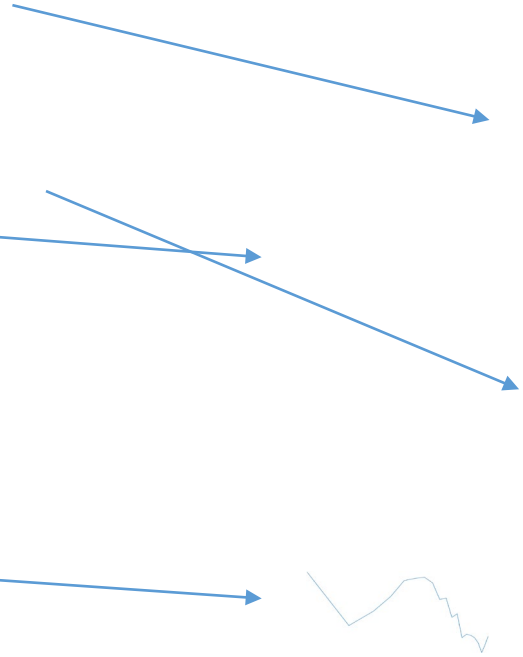
How good is that, BTW?

Measuring multi-TeV muons has been a Group-2 issue before LHC experiments started to consider it

The resolution of muons traversing 1.5km of ice (=3850 X_0) in IceCUBE has been determined with three different methods in[23].

Although of course the problem is very different, I have not resisted the temptation to overlay to the graph on the right the ballpark of the resolution we achieve with a 2m-long lead tungstate calorimeter (=225 X_0) + CNN reconstruction

→ **3x better**



How large are the gains of a full optimization?

I recently provided a clear example [13] of how experimental design as is carried out today leaves ample room for improvement from the systematic study of even seemingly irrelevant choices for, *e.g.*, the placement of active and passive material in a simple detector.

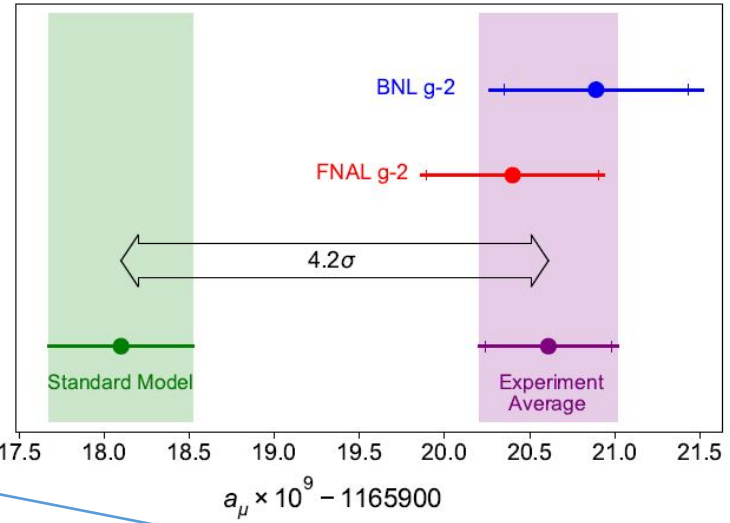
The chance of doing so was offered by my refereeing work of the detector proposed by the MUonE collaboration [14], which aims at determining with high precision the cross section of elastic muon-electron scattering.

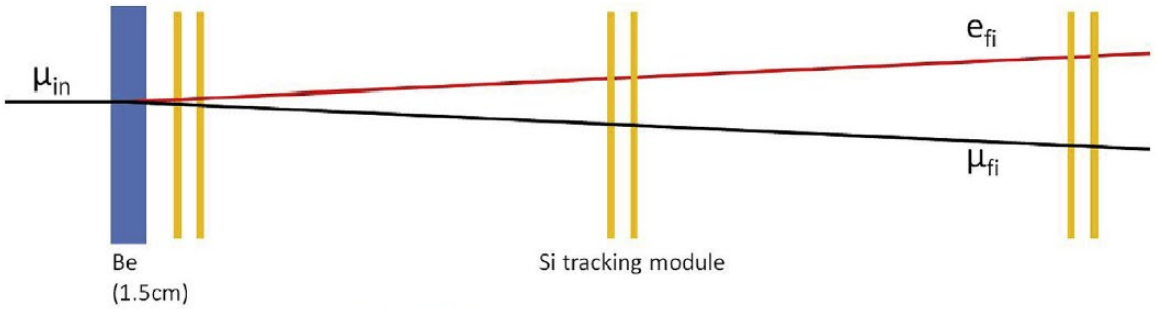
In the cited study I demonstrated, through the direct exploration of the parameter space of detector geometry, how **large gains** in suitable utility functions (related to the resolution in the event q^2) can be obtained by moving away from choices dictated by past experience

One example of geometry optimization: MUonE

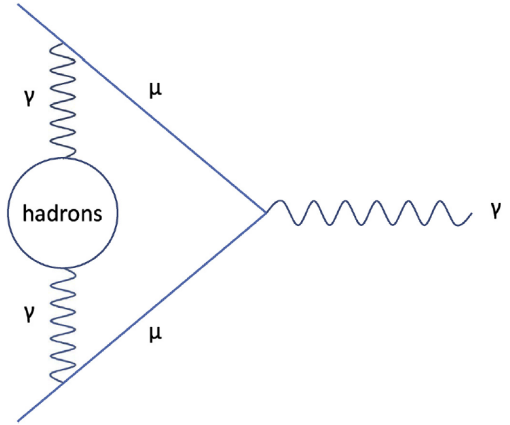
MUonE [14] aims to determine with high precision the muon-electron elastic scattering differential cross section, to extract hadronic contributions and reduce the systematics of the g-2 muon anomaly

The experiment must be sensitive to hadronic loop effects particularly at high q^2, where a 10^{-4} measurement may substantially improve the theoretical understanding of the g-2 value

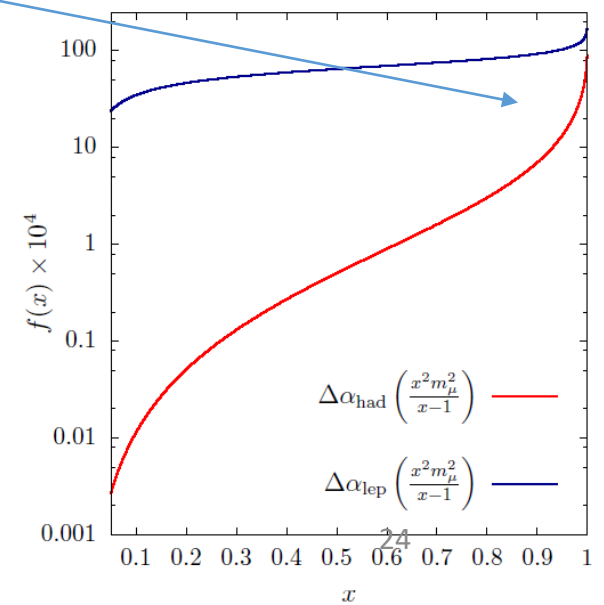




Above: layout of one of 40 1m-long stations



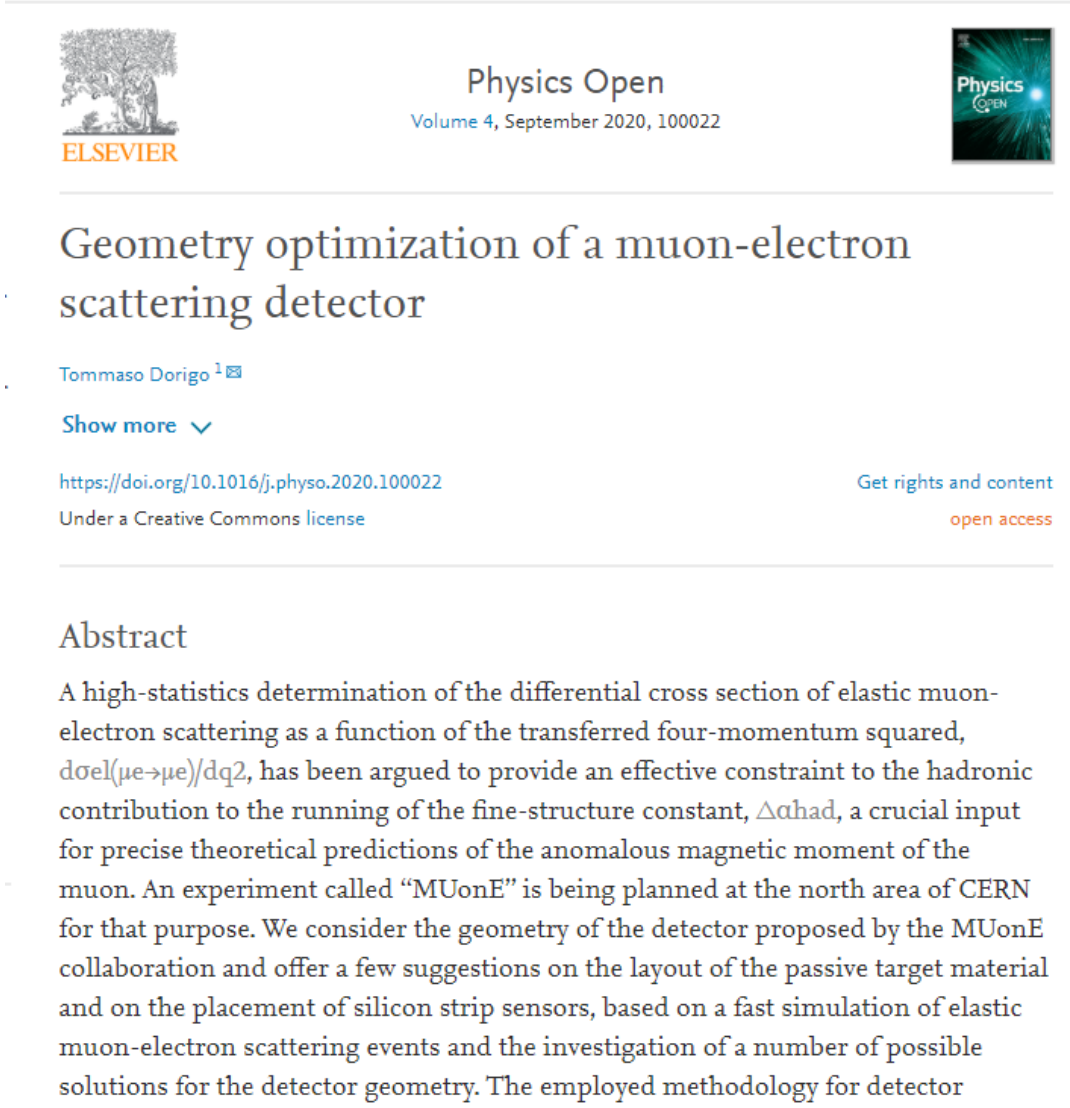
A virtual hadronic loop



MUonE optimization

By optimizing layout with a discrete sampling, I proved how **a factor of 2 improvement in the relevant metric could be achieved** without increase in detector cost

The study also proved how dreaded systematic effects from positioning uncertainties could be nullified by software means



ELSEVIER

Physics Open
Volume 4, September 2020, 100022

Geometry optimization of a muon-electron scattering detector

Tommaso Dorigo [1] ✉

Show more

https://doi.org/10.1016/j.physo.2020.100022

Get rights and content
Under a Creative Commons license
open access

Abstract

A high-statistics determination of the differential cross section of elastic muon-electron scattering as a function of the transferred four-momentum squared, dσel(μe→μe)/dq2, has been argued to provide an effective constraint to the hadronic contribution to the running of the fine-structure constant, Δαhad, a crucial input for precise theoretical predictions of the anomalous magnetic moment of the muon. An experiment called "MUonE" is being planned at the north area of CERN for that purpose. We consider the geometry of the detector proposed by the MUonE collaboration and offer a few suggestions on the layout of the passive target material and on the placement of silicon strip sensors, based on a fast simulation of elastic muon-electron scattering events and the investigation of a number of possible solutions for the detector geometry. The employed methodology for detector

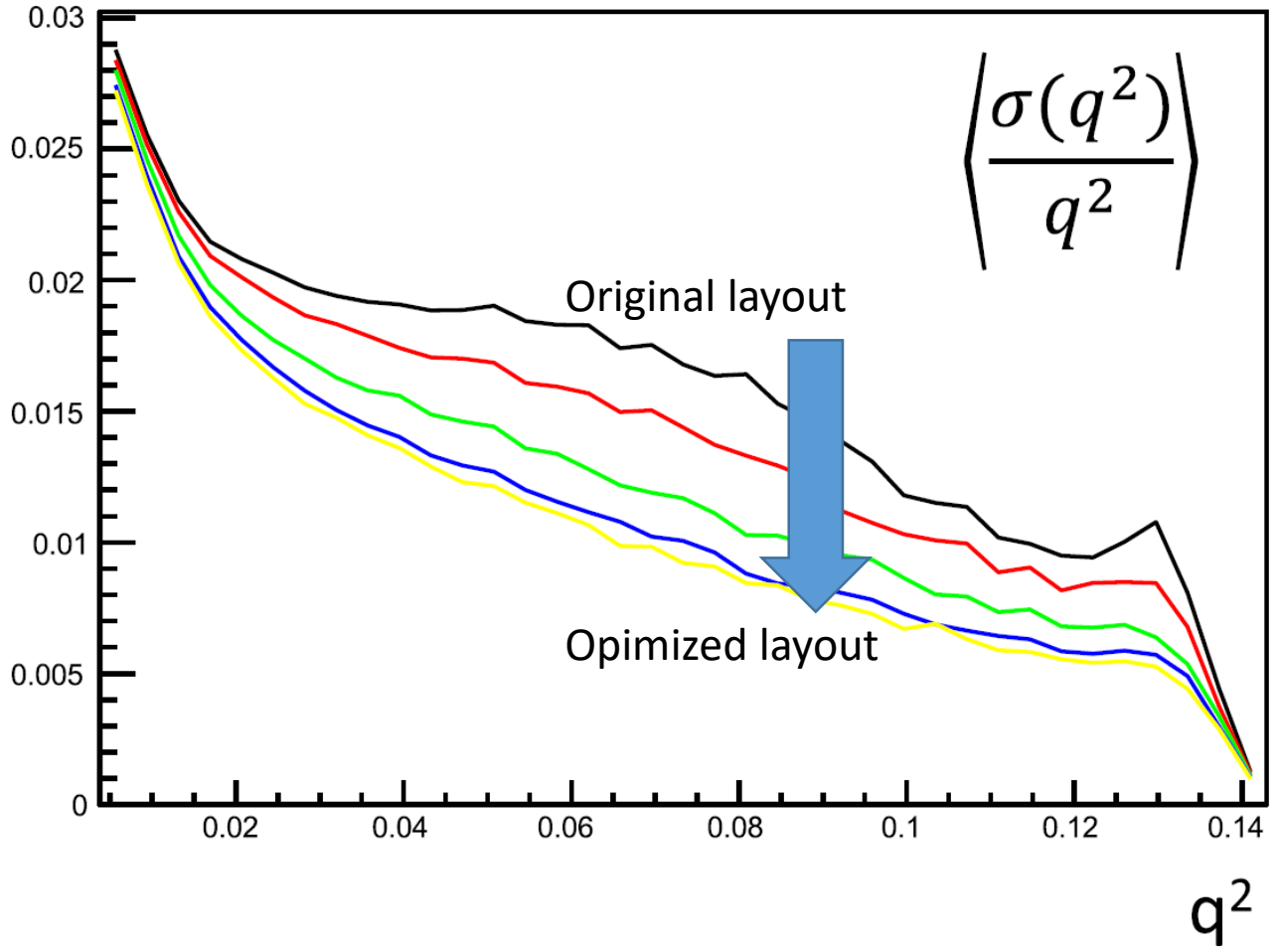
Sample results

The study was *not* performed with deep learning technologies, as it was not strictly necessary given the reduced space of design choices I wished to investigate.

The results prove that design optimization is not something alien to our reach, but rather, something we should pay more attention to!

We can only guess how large are the gains in the final experimental objectives possible if a fully differentiable model is created for detectors of significantly higher complexity than MUonE.

My guess: huge.



Above: relative resolution in event q^2 for different configurations (the higher, black line is the original proposal by the MUonE coll.)

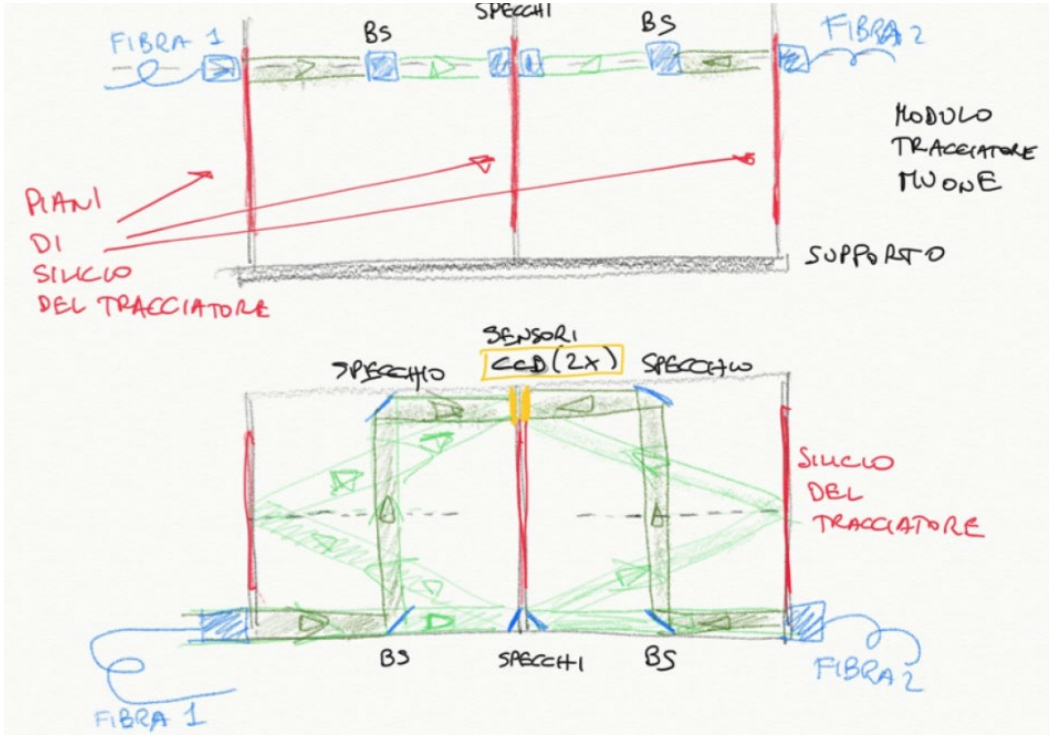
Speaking of systematic uncertainties,

MUonE correctly identified the need for locating the scattering vertex to within 10μm along the beam axis (it has a strong impact on the q^2 resolution), and proceeded to design a very fancy holographic laser system, to be mounted on each station (=40 systems) to monitor the sensors locations

Cost: **several hundred kEuros**

As a by-product of the modeling of detector + information extraction process, the optimization study showed that with 5' of muon beam data, the location, tilt and bow of all detector and target elements **can be determined with O(1μ) accuracy by a global fit to the vertex**!

This is an example of the dividends that the study of a full model of (physics)+(detector)+(reco method)+(inference extraction) can provide



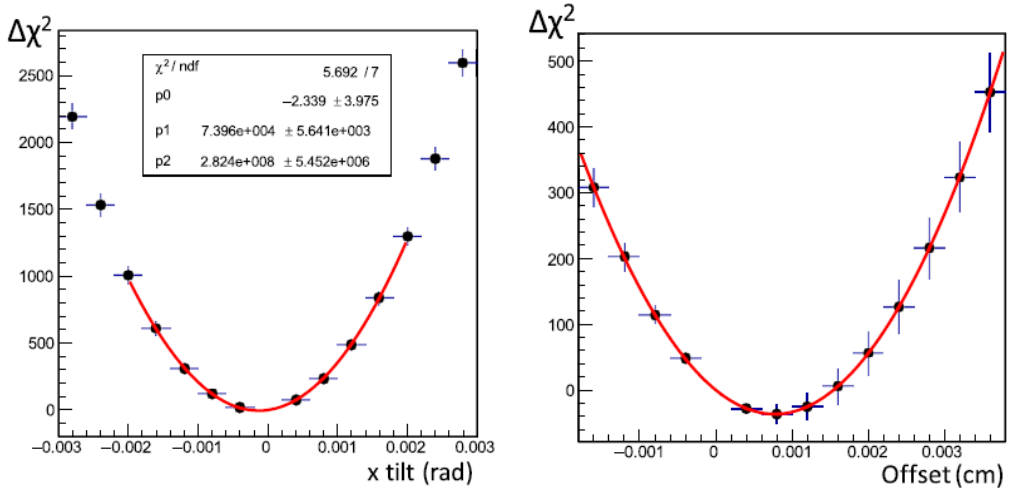
Speaking of systematic uncertainties,

MUonE correctly identified the need for locating the scattering vertex to within 10μm along the beam axis (it has a strong impact on the q^2 resolution), and proceeded to design a very fancy holographic laser system, to be mounted on each station (=40 systems) to monitor the sensors locations

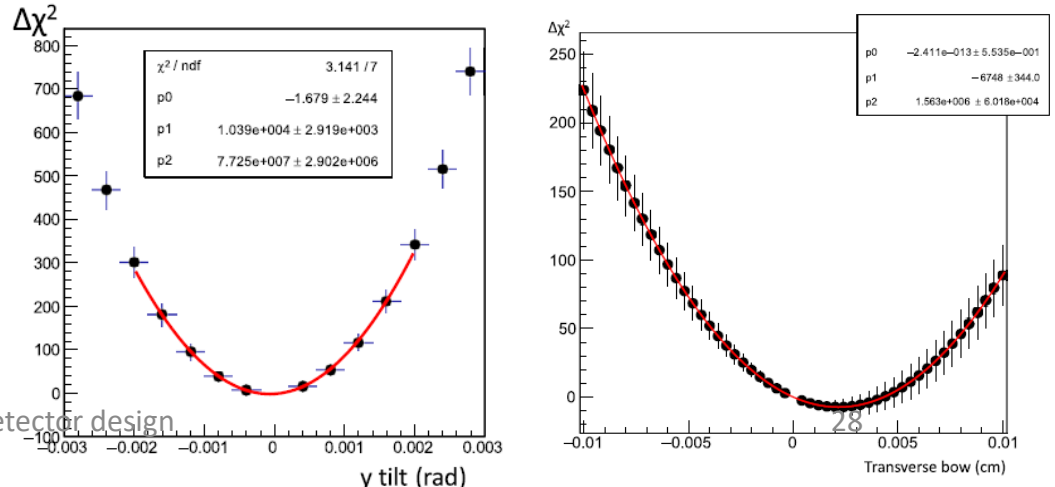
Cost: **several hundred kEuros**

As a by-product of the modeling of detector + information extraction process, the optimization study showed that with 5' of muon beam data, the location, tilt and bow of all detector and target elements **can be determined with O(1μ) accuracy by a global fit to the vertex**!

This is an example of the dividends that the study of a full model of (physics)+(detector)+(reco method)+(inference extraction) can provide



Top left: global χ^2 vs x tilt; right: vs z offset;
Bottom left: vs y tilt; right: vs transverse bow.



Computer science to the rescue

Progress in CS redefined performance standards of our technologies, and reshaped the way we think about optimization, by providing us with deep learning algorithms that revolutionize common tasks and surpass human performance. We can today identify AI ingredients in, *e.g.*, language translation, speech recognition, self-driving vehicles.

→ Of course, that AI is not general but application-specific: its potential of providing new solutions to old tasks **depends on our ability to create the right interfaces**.

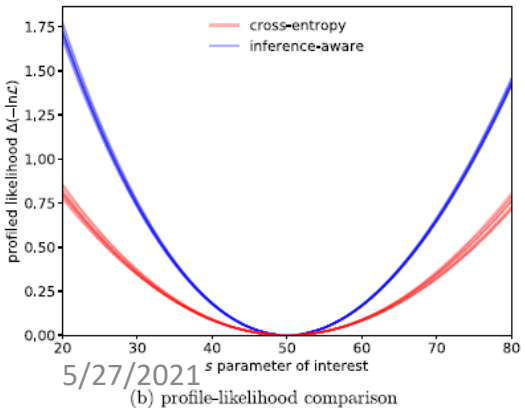
In HEP, ML applications caught up rather slowly, but NNs and gradient boosting techniques eventually operated a paradigm shift, improving the performance of our measurements by large amounts.

A new paradigm shift is now offered by differentiable programming [15], which eases the systematic search of minima of arbitrarily complex multi-dimensional functions; by casting the whole problem in a differentiable framework a full end-to-end optimization becomes possible.

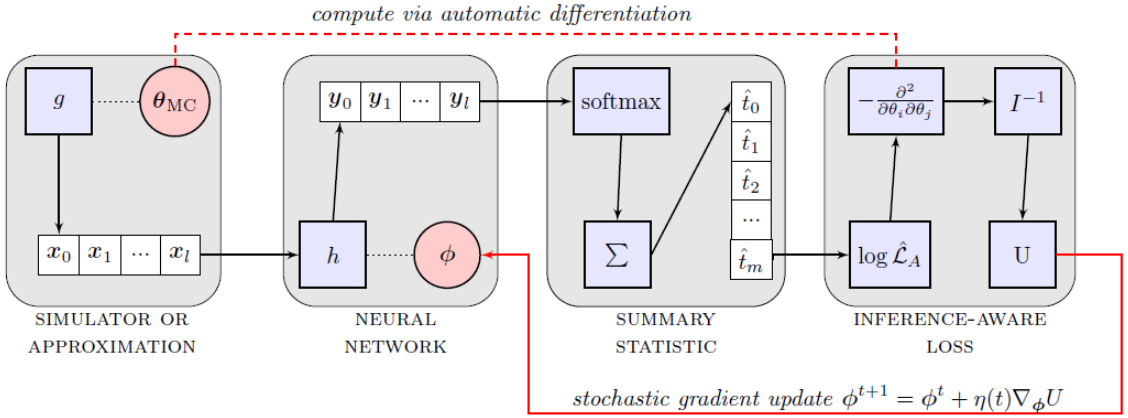
INFERNO

As an example of what differentiable programming can do for us, I designed with **P. de Castro** an innovative algorithm [16] using automatic differentiation to construct a loss function that directly targets the information content of the statistical summary produced by the neural network.

If the loss function is constructed to incorporate the effects of nuisance parameters on the measurement objective, virtual optimality of the classification task and large improvements in precision can be achieved over procedures that account for nuisance parameters downstream of the NN training.



Left: profile likelihood on the parameter of interest for a neural network with (blue) and without (red) the feedback on effect of nuisances provided by INFERNO



The code of INFERNO (in TF 1.0) is available on github.

Giles C. Strong also made available a PyTorch implementation and is demonstrating its performance on astro-HEP use case.

Lukas Layer ported the code to TF2.0

Presently, this technology is being tested in a real CMS analysis, using open Run1 data to replicate a top cross section measurement, by **Lukas Layer**

A study of muon shielding in SHIP

In another seminal work [17], local generative surrogates of the gradient of the objective function were proven to allow for the minimization by SGD and a strong reduction in muon background fluxes in the SHIP experiment

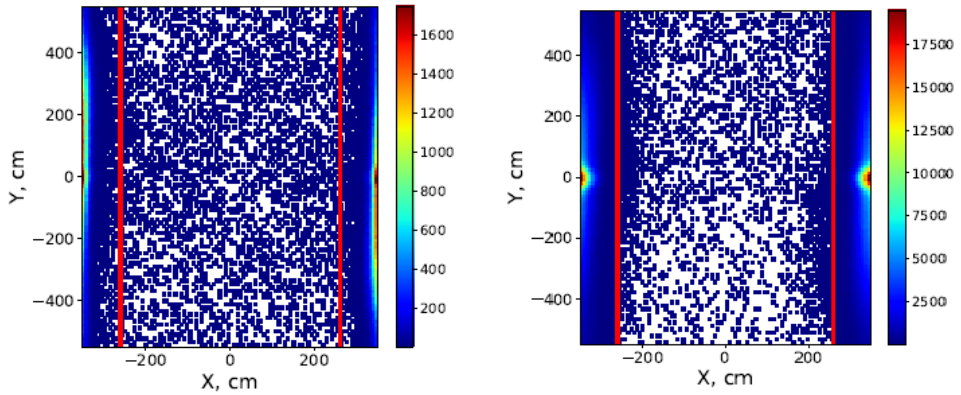
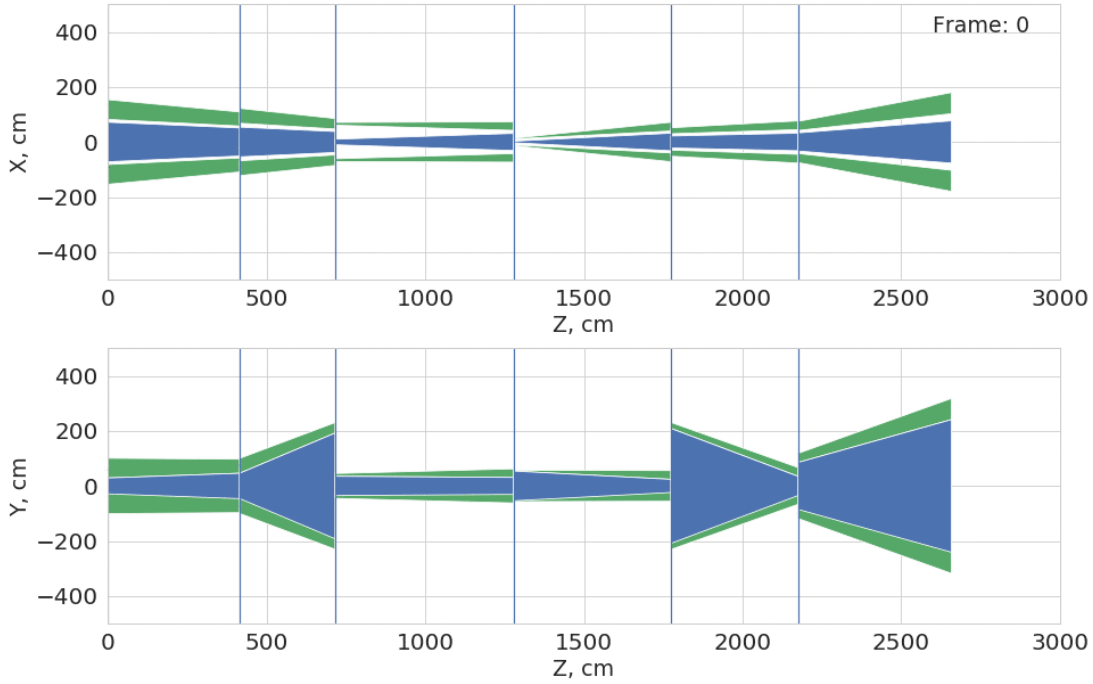


Figure 7. Muon hits distribution in the detection apparatus (depicted as red contour) obtained by Bayesian optimization (Left) and by L-GSO (Right), showing better distribution. Color represents number of the hits in a bin.



Geometry optimization at work in **real time**!

Realigning design choices and ultimate goals

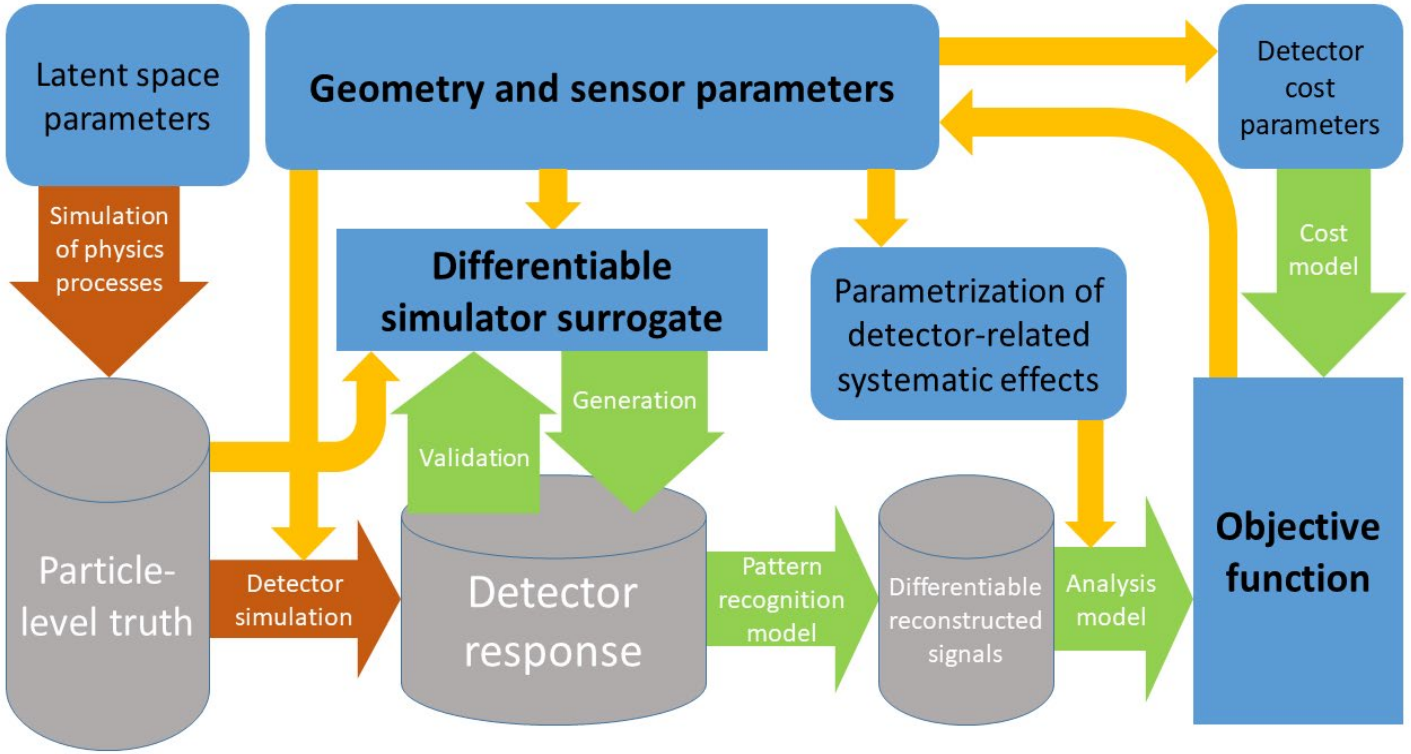
The target of **MODE** is to design and offer to the community a scalable, versatile architecture that can provide end-to-end optimization of particle detectors, proving it on a number of different applications across different domains.

Study cases:

- Demonstration of muon energy measurement in optimized calorimeter → article in preparation
- Muon tomography detector optimization [18] → in progress
- Hybrid calorimeter design integrating tracking layers → activity starting

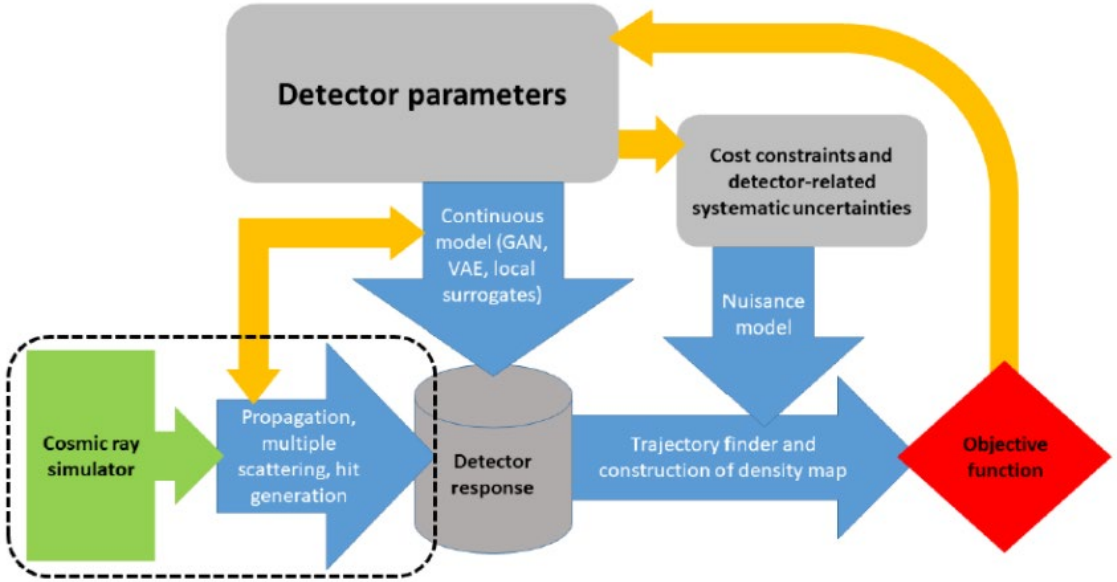
Other use cases being considered include:

- Hadron therapy (iMPACT project [19]);
- Muon collider detector shielding [20];
- Optimization of MUonE calorimeter;
- Optimized search for long-lived signatures at FCC-ee



And for the time being…

«Simpler» use case: muon tomography. We need no surrogate of a simulator, yet all other pieces of the puzzle still need to be carved and set in.

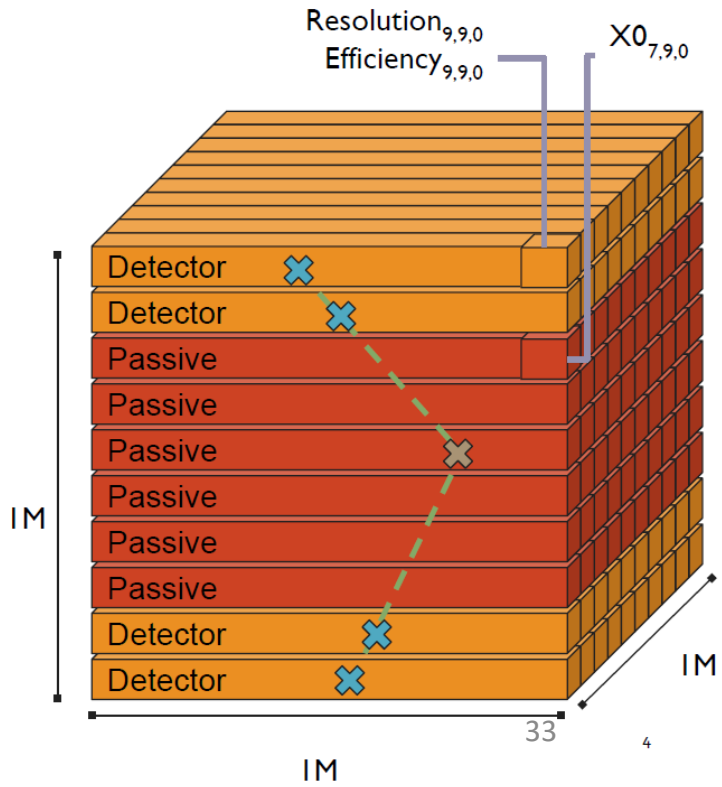


For a simple test, we model a scanned volume including a Pb block of 0.5x0.1x0.1 m^3 inside a 0.6x1x1m^3 of lower-Z material

The system «learns» how to compromise cost and precision, and where detector elements are less useful

A number of shortcuts have been taken to develop this purposedly crude model – but once we have something that «breathes», we may start building into it functionality and detail

Right: scheme of the modeled apparatus (graph courtesy **G. C. Strong**)



VERY preliminary (yesterday's) results

The code and results shown have been produced by **Giles C. Strong**

These graphs show the result of a run of 100 epochs training, followed by a prediction with 100k muons

First proof of principle (very low statistics) of correct training of a differentiable model of a schematic muon tomography apparatus.

The **loss** is a combination of detector cost (itself a function of sensors efficiency and resolution) and RMSE on rad length estimate

Still a looong way to go, but an important milestone for this use case

Above, top to bottom: loss, loss composition, resolution map, and efficiency map of detection elements after minimization.

Right: predicted and true X_0 of passive volume

Realigning our design choices to future AI

A point which cannot be stressed enough is that if we design today something that will operate 10 or 20 years in the future, we need to account for the pattern recognition capabilities of future automated systems

In 20 years, will we use a Kalman filter to reconstruct trajectories in our trackers, or photon energy and direction in our calorimeters?

No, we won't. We will employ AI technology, streamlined by a decade of consolidation in similar tasks.

Shouldn't we then build those devices by considering how AI technology could best exploit them? If we do not, we will suffer a misalignment of our design choices and the future capabilities of the software we will end up using.

How to get around this problem?

We can and should try to model increasingly performant pattern recognition in our optimization loops, and verify whether there are discontinuities in the solutions space.

It is not going to be easy, but it is IMHO absolutely necessary to start getting equipped.

Four-slide description of a differentiable model for design optimization - 1

An end-to-end detector design optimization task can be briefly formalized in the following way.

We start with a simulation of the physics processes of relevance for the considered application, which generates a multi-dimensional, stochastic input variable **x**, distributed with a PDF **f(x)**.

The input is turned by the simulation of the detection apparatus into sensor readouts **z** distributed with a PDF

$$p(z|x,\theta),$$

which constitute the observed low-level features of the physical process; readouts **z** depend through **p( )** on parameters **θ** that describe the physical properties of the detector and its geometry.

Four-slide description of a differentiable model for design optimization - 2

The observations **z** are used by a reconstruction model **R( )** that produces high-level features

$$\zeta(\theta) = R[z,\theta,\nu(\theta)]$$

(*e.g.* particle four-momenta), by employing knowledge of the detector parameters as well as a model of the detector-driven nuisance parameters **ν(θ)** which affect the pattern recognition task.

In turn, high-level features **ζ(θ)** constitute the input of a further, less dramatic, dimensionality reduction, the data analysis step: this is typically performed by a classifier or regressor **NN( )** powered by a neural network.

Once properly trained for the task at hand, the network produces a low-dimensional summary statistic

$$s = NN[\zeta(\theta)]$$

with which inference can finally be carried out to produce the desired goal of the experiment.

Four-slide description of a differentiable model for design optimization - 3

In general, one may formally specify the problem of identifying optimal detector parameters as that of finding estimators $\widehat{\boldsymbol{\theta}}$ that satisfy

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{arg}\min_{\boldsymbol{\theta}} \int \boldsymbol{L}[\boldsymbol{NN}(\boldsymbol{\zeta}), \boldsymbol{c}(\boldsymbol{\theta})]\boldsymbol{p}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})\boldsymbol{dxdz}$$

c(θ) is a function modeling the cost of the considered detector layout of parameters **θ**, and the loss function **L[NN,c]** is constructed to appropriately weight the result of the measurement in terms of its desirable goals, as well as to obey cost constraints and other use-case-specific limitations.

Since in the cases of interest the PDF **p(z|x,θ)** is not available in closed form –the considered models are implicit–, we must rely on forward simulation: we approximate $\widehat{\boldsymbol{\theta}}$ with a sample of n events:

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{a}} = \boldsymbol{arg}\min_{\boldsymbol{\theta}} \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{L}\left[\boldsymbol{NN}\left(\boldsymbol{R}(\boldsymbol{z_i})\right), \boldsymbol{c}(\boldsymbol{\theta})\right]$$

where $\mathbf{z_i}$ is distributed as $\mathbf{F(x_i,\theta)}$ to emulate **p(z|x,θ)** as $\mathbf{x_i}$ is sampled from its PDF **f( )** by the simulator. One may thus obtain an estimate of the loss function and the detector parameters which minimize it.

Four-slide description of a differentiable model for design optimization - 4

It has been shown how in applications such as those of our interest it is viable to approximate the non-differentiable stochastic simulator **F( )** with a local surrogate model,

z = S(y,x,θ),

that depends on a parameter **y** describing the stochastic variation of the approximated distribution[49]. This allows to descend to the minimum of the approximated loss $\widehat{L(z)}$ by following its surrogate gradient

$$\nabla_\theta \widehat{L(z)} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta L[NN\left(R\left(S(y_i, x_i, \theta)\right)\right), c(\theta)]\,.$$

The above recipe requires one to learn the differentiable surrogate **S( )**: this is a task liable to be carried out independently from the optimization procedure.

The modular structure of a differentiable pipeline modeling the optimization cycle allows the user to turn on and off specific parts of the chain, helping the system in its exploration of the feature space.

Machine-Learning Optimized Design of Experiments

MODE Collaboration

https://mode-collaboration.github.io

A. G. Baydin[5], A. Boldyrev[4], K. Cranmer[8], P. de Castro Manzano[1], T. Dorigo[1], C. Delaere[2], D. Derkach[4], J. Donini[3], A. Giammanco[2], J. Kieseler[7], G. Louppe[6], L. Layer[1], P. Martinez Ruiz del Arbol[9], F. Ratnikov[4], G. Strong[1], M. Tosi[1], A. Ustyuzhanin[4], P. Vischia[2], H. Yarar[1]

1 INFN, Sezione di Padova (and associates from Padova and Naples Universities), Italy
2 Université Catholique de Louvain, Belgium
3 Université Clermont Auvergne, France
4 Laboratory for big data analysis of the Higher School of Economics, Russia
5 University of Oxford
6 Université de Liege
7 CERN
8 New York University
9 IFCA

The strategy of MODE

We are fully aware that the dream of informing the design of a complex detector for a fundamental physics endeavour (be it HE, astro-, Neutrino, or Nuclear physics) entails a walk in the desert

Yet we must start it, as in 20 years the shortcomings of having designed experiments that are misaligned with goals and information-extraction procedures will otherwise be paid dearly.

The strategy is thus to start with easy use cases, where further proof may be brought of the gains of using DL architectures to parametrize the essential ingredients of the design problems

Hopefully, we will be able to convince the community, or else, we'll have to wait for a generation change.

The important observation is that the developed architectures for optimization are modular, hence we will be able to recycle part of the work for one application when we move to the next one.

Status and next steps

- An article describing the MODE program has been published last month in Nuclear Physics News International [21]
- A white paper on differentiable programming for detector design optimization is being drafted
- We are organizing a workshop on "*Differentiable Programming for Design Optimization*" on September 6-8 2021 in Louvain-la-Neuve, to allow interested scientists to join and discuss together the means and the possible applications
  - Extra support for this activity is provided by IRIS-HEP and JENAA

Group 2 members of INFN-PD or UNIPD are welcome to propose use cases of interest and share expertise / collaborate

Every other solved application = a publication AND added knowledge base on solving these hard problems!

You are also most welcome to participate to MODE activities, or propose to become a member

Day 1, morning:

8.30-9.00 Introduction and goals of the workshop (T. Dorigo, INFN-Padova)
9.00-11.00 Progress in Computer Science (chair: Atilim Gunes Baydin, U. Oxford)
- Keynote lecture on deep learning
- Keynote lecture on differentiable programming applications to design optimization
- other contributions by CS experts

11.00-11.30 break
11.30-13.30 Applications in muon tomography and hadron therapy (chair: Andrea Giammanco, U. Cath. Louvain)

Day 1, afternoon:

14.30-16.30 Applications and requirements for particle physics (chair: Pietro Vischia, U. Cath. Louvain)
16.30-17.00 break
17.00-19.00 Applications and requirements in astro-HEP (chair: Roberto Ruiz de Austri Bazan, IFIC Valencia)

Day 2, morning:

9.00-11.00 Applications and requirements for neutrino detectors (chair: Kazuhiro Terao, Columbia U.)
11.00-11.30 break
11.30-13.30 Applications and requirements in nuclear physics experiments (chair: Gian Michele Innocenti, CERN)

Day 2, afternoon:

14.30-15.30 Round table discussion, formation of working groups
15.30-17.00 (Parallel) Specification of tasks and goals of the working groups
17.00-17.30 break
17.30-19.00 Organization of future activities and closing of the workshop (chair: T.Dorigo, INFN-Padova)

MODE_INFN : a new Group 5 endeavour

In order to create a community of experts within INFN, who can participate in the process started within MODE and bring up for modeling new use cases of interest to INFN, we are proposing a new INFN experiment (or better, a collaboration) within Group 5

Why Group 5?

Because our activities cut diagonally into GR1, GR2, and GR3, and our focus is R&D for the development of tools, rather than their exploitation for specific applications. A synergy with experimental groups that have use cases of interest will hopefully be easy to establish

Tentative members and sites of MODE_INFN

Padova: Azzi, Collazuol, Conti, Dorigo (RN), Lucchesi, Rossin, Strong, Tosi (RL), Verlato, ...

Firenze: Anderlini, Barbetti, Bonechi, Borselli, Ciulli, D'Alessandro, Lenzi, Viliani (RL)

Napoli: Cimmino (RL), D'Errico, Saracino, Ambrosino, Vitiello

Roma: Giagu (RL), Ippolito + PhD

Bari: Maggi, Venditti (RL), ...

- **MODE** members who have been asked and agreed to collaborate on WPs brought forth by **MODE_INFN**:
  Andrea Giammanco (UcL); Pietro Vischia (UcL); Jan Kieseler (CERN); Andrey Ustyuzhanin (HSE); Fedor Ratnikov (HSE); Alexey Boldyrev (HSE); Pablo Martinez Ruiz del Arbol (IFCA); Julien Donini (UCA)
- Plus all other MODE members, on the general goals

Activities:
WP1: Software development
WP2: Tomography optimization
WP3: Calorimetry optimization
WP4: Other applications

MODE_INFN and You

If you are doing experimental research in HEP, astro-HEP, neutrino physics, or high-energy nuclear physics, or if you are working at spin-offs involving, *e.g.,* muon tomography, hadron therapy, or other endeavours which operate with instruments that extract information from the interaction of energetic radiation with matter, you are very likely to have a use case – a system liable to benefit from a study with differentiable programming.

The idea of MODE_INFN is to bring together ML experts who are developing the interfaces for these applications, with the researchers who have problems to solve in their area of interest

We cannot offer a solution to any given problem (we lack the personpower to work on-demand), but together we may work toward it

→ Consider joining MODE_INFN, or MODE, and bring your use case!

Do I have a use case checklist:

Are you involved in the design, assembly, or upgrade of an instrument?

Can you specify one or a set of desirable scientific goals from its use?

Are those goals achieved through information processing?

If your answers to all are «yes», **you have something to optimize** and chances are this can't be done without a deep learning model of the full information extraction chain.

THANK YOU FOR YOUR ATTENTION!

References

1. The European Strategy Group, "*2020 Update of the European Strategy for Particle Physics*," CERN, Geneva, https://europeanstrategy.cern/, doi: http://dx.doi.org/10.17181/ESU2020Deliberation, 2020.

2. The topic has been discussed widely in blogs and other social media. *E.g.*, see Sabine Hossenfelder's discussion here: http://backreaction.blogspot.com/2019/06/if-we-spend-money-on-larger-particle.html, or see Alessandro Strumia's guest post at the same site: http://backreaction.blogspot.com/2020/06/guest-post-who-needs-giant-new-collider.html. Also see the New York Times op-ed by S. Hossenfelder, https://www.nytimes.com/2019/01/23/opinion/particle-physics-large-hadron-collider.html

References / cont'd

3. F. Hartmann, *Evolution of Silicon Sensor Technology in Particle Physics*, Springer Tracts in Modern Physics, 2017, ISBN 978-3-319-64436-3; P. Delpierre, *A history of hybrid pixel detectors, from high energy physics to medical imaging*, J. INST. 9 (2014) C05059, doi: 10.1088/1748-0221/9/05/C05059; P. Allport, *Applications of silicon strip and pixel-based particle tracking detectors*, Nature Reviews Physics 1 (2019) 567–576, doi: https://doi.org/10.1038/s42254-019-0081-z; G. Case, *New trends in silicon detector technology*, JINST 15 (2020), https://iopscience.iop.org/article/10.1088/1748-0221/15/05/C05057; M. Garcia-Sciveres and N. Wermes, *A review of advances in pixel detectors for experiments with high rate and radiation*, Rep. Prog. Phys. 81 (2018) 066101, doi: http://doi.org/10.1088/1361-6633/aab064.

References / cont'd

4. Y. Mishnayot *et al*., "*3D Printing of Scintillating Materials*", arXiv:1406.4817[cond-mat.mtrl-sci] (2014); Y. Abreu *et al*. (SoLID Collaboration), "*A novel segmented-scintillator antineutrino detector*", JINST 12 (2017) P04024, doi: 10.1088/1748-0221/12/04/P04024.

5. M. Mandurrino *et al*., "*Demonstration of 200 --, 100 --, and 50 micron Pitch Resistive AC Coupled Silicon Detectors (RSD) With 100% Fill Factor for 4D Particle Tracking*," IEEE Electron Device Letters 40, 11 (2019) 1780, arXiv:1907.03314[physics.ins-det]; M. Mandurrino *et al*., "*Analysis and numerical design of Resistive AC Coupled Silicon Detectors (RSD) for 4D particle tracking*", Nucl. Instr. Meth. A959 (2020) 163479, doi: 10.1016/j.nima.2020.163479; M. Tornago et al., "*Resistive AC-Coupled Silicon Detectors: Principles of operation and first results from a combined analysis of beam test and laster data*", NIM A 1003 (2021) 165319, https://www.researchgate.net/publication/343095980 .

6. CMS Collaboration, "*The Phase-2 Upgrade of the CMS Endcap Calorimeter*", CERN-LHCC-2017-023 (2018), https://cds.cern.ch/record/2293646?ln=en .

References / cont'd

7. J. Repond *et al.*, "*Hadronic Energy Resolution of a Combined High Granularity Scintillator Calorimeter System*", JINST 13 (2018) P12022, doi: 10.1088/1748-0221/13/12/P12022; C. Adloff *et al.*, "*Tests of a particle flow algorithm with CALICE test beam data*", JINST 6 (2011) P07005, doi: 10.1088/1748-0221/6/07/P07005.

8. P.W. Cattaneo *et al.*, "*CaloCube: a novel calorimeter for high-energy cosmic rays in space*", JINST 12 (2017) 06, C06004, doi: 10.1088/1748-0221/12/06/C06004 .

9. E. Hoogeboom *et al.*, "*HexaConv*", arXiv:1803.02108[cs.LG] (2018)).

10. M. Sirunyan *et al.* (CMS Collaboration), "*Particle-flow reconstruction and global event description with the CMS detector*", JINST 12 (2017) P10003, doi: 10.1088/1748-0221/12/10/P10003 .

References / cont'd

11. See *e.g.* S.R. Qasim, **J. Kieseler**, Y. Iiyama, and M. Pierini, "*Learning representations of irregular particle-detector geometry with distance-weighted graph networks*", Eur. Phys. J. C79 (2019) 7,608, doi: 10.1140/epjc/s10052-019-7113-9; **J. Kieseler**, "*Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph and image data*", arXiv:2002.03605[physics.data-an] (2020); J. Alimena, Y. Iiymana, **J. Kieseler**, "*Fast convolutional neural networks for identifying long-lived particles in a high-granularity calorimeter*", arXiv:2004.10744 [hep-ex] (2020).

12. **T. Dorigo, J. Kieseler, L. Layer** and **G. Strong**, "*Muon Energy Measurement from Radiative Losses in a Calorimeter for a Collider Detector*", http://arxiv.org/abs/2008.10958 (2020).

13. **T. Dorigo**, "*Geometry Optimization of a Muon-Electron Scattering Detector*," Physics Open 4 (2020) 100022, arXiv:200200973[physics.ins-det], doi: 10.1016/j.physo.2020.100022.

14. G. Abbiendi *et al.*, "*Letter of Intent: The MUonE Project*", CERN-SPSC-2019-026/SPSC-I-252 (2019).

References / cont'd

15. **A. Güneş Baydin**, B.A. Pearlmutter, A.A. Radul, and J.M. Siskind, "*Automatic Differentiation in Machine Learning: a Survey*", Journal of Machine Learning Research (JMLR) 18 (153) (2018) 1, http://jmlr.org/papers/v18/17-468.html; Y. LeCun, Y. Bengio, and G. Hinton, "*Deep learning*", Nature, 521(7553) (2015) 436, doi: 10.1038/nature14539.

16. **P. de Castro Manzano** and **T. Dorigo**, "*INFERNO: Inference-Aware Neural Optimization*", Comp. Phys. Commun. 244 (2019) 170; Arxiv:1806.04743v2 [stat.ml] (2018), doi: 10.1016/j.cpc.2019.06.007.

17. S. Shirobokov, **A. Ustyuzhanin, A. Güneş Badyin** *et al.*, "*Differentiating the Black-Box: Optimization with Local Generative Surrogates*", arXiv:2002.04632v1[cs.LG] (2020).

18. S. Wuyckens, **A. Giammanco**, P. Demin, and **E. Cortina Gil**, "*A Portable muon telescope based on small and gas-tight Resistive Plate Chambers*", Phil. Trans. Royal Soc. A377 (2019) 2137, arXiv:1806.06602v2[physics.ins-det] (2018), doi: 10.1098/rsta.2018.0139.

References / cont'd

19. P. Giubilato *et al*., "*iMPACT: innovative pCT scanner*", IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC) IEEE (2015), https://ieeexplore.ieee.org/abstract/document/7581240; S. Mattiazzo, *et al*., "*Advanced proton imaging in computed tomography*", Radiation protection dosimetry 166.1 (2015) 388, https://academic.oup.com/rpd/article/166/1-4/388/1612689; S. Mattiazzo *et al*., "*The iMPACT project tracker and calorimeter*", Nucl. Instr. Meth. A845 (2017) 664, https://www.sciencedirect.com/science/article/pii/S0168900216303412; N. Pozzobon *et al*. "*Calorimeter prototyping for the iMPACT project pCT scanner*", Nucl. Instr. Meth. A936 (2019) 1, https://ui.adsabs.harvard.edu/abs/2019NIMPA.936....1P/abstract.

20. N. Bartosik *et al*., "*Preliminary Report on the Study of Beam-Induced Background Effects at a Muon Collider*," arXiv:1905.03725[hep-ex] (2019); N. Bartosik *et al.*, "*Detector and Physics Performance at a Muon Collider*," J. Inst. 15 (2020) P05001, doi: 10.1088/1748-0221/15/05/P05001.

References / cont'd

21. A.G. Baydin et al., *"Toward Machine Learning Optimization of Experimental Design"*, submitted to Nuclear Physics News International, 2021.

22. W. Sakumoto et al., "*Measurement of TeV muon energy loss in iron*",Phys. Rev. D45, 9 (1992) 3042.

23. The IceCUBE Collaboration, "*Muon energy reconstruction and atmospheric neutrino spectrum unfolding with the IceCube detector"*, proc. 30th International Cosmic-Ray Conference (2007), Merida, Mexico, arXiv:0711.0353v1