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

Tommaso Dorigo
INFN, Sezione di Padova
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.

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 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.

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 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 through projects that require large effort and significant funding [2].

Furthermore, there are indications that wide sectors of society no longer consider a top priority the furthering of our understanding of matter at the smallest distance scales through projects that require large effort and significant funding [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 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 the crucial underlying paradigms of experimental design remained mostly unchallenged across decades, e.g.

• “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 and astro-HEP typically involves a low-material tracker followed by a thick calorimeter.

• Focus on significant redundancy of detection systems, ensuring robustness and enabling cross-calibration of resulting measurements

• Symmetrical layouts leveraging conservation laws, also simplifying reconstruction

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 so damnedly complex is not only that the studied physics is hard: a lot has to do with Science being a competitive job. *We want to study everything* and do it *better* than our predecessors.

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

*What loss function do we aim to minimize?*

*Does it make sense of speaking of a single utility function?*

Concerning the last question: I am convinced that it does, and I will try to convince you, too, 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 have actually grown *used* to create multi-target optimization strategies, e.g., to 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

So - what does an end-to-end detector optimization looks like in blueprint?

1. Model as steep functions the cost of overriding budget and commissioning time

2. Assess the scientific impact of each achievable scientific results, optionally as a continuous function of their precision

3. Create a differentiable model of geometry, components, information-extraction procedures, and utility function

4. Construct a pipeline with those modules, enabling backpropagation and gradient descent functionality

5. Let the chain rule of differential calculus do the hard work for you!

Experiment proposals undergo a review, wherein the utility function is well explicitated to the funding agents... So, we invariably know what we are shooting at!
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 sound as “the highest precision on theta_13 we can get”, when the focus becomes instead maximizing the number of reconstructed 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.

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 \( \theta_{13} \) 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.

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.

New performance demands are also arising:

- HL-LHC studies made clear that tracking in dense environments **requires AI solutions**
- In space, payload and power consumption remain **major constraints**, making applications especially sensitive to hard-to-make design decisions
- Boosted jet tagging at high $p_T$ – all the rage for NP searches at the LHC – demands us to invest in **better calorimeters**
Energetic LHC collisions may produce heavy objects with large momentum (top quarks, or W, Z, H bosons). When these decay, they yield a collimated stream of particles – a single (fat) hadron jet.

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

High granularity and effective identification of individual constituents within 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. Only later did hadron calorimetry get recognized as crucial for 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] must be pursued

• 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 will become imperative in future colliders at the energy frontier (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)
  
  → 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, $\sigma(p)/p = 0.2 \ p$

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

![Graph showing mass stopping power](image)

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 $\Delta E_\mu$, which corresponds to the fractional muon energy loss $\nu$. Note that the ordinate is the logarithmic derivative $d\sigma/d(\ln E_\mu) = d\sigma/d(\ln\Delta E_\mu)$.

(From a CCFR study [xx])
Energy regression with CNNs

We studied how a granular calorimeter may determine the energy of multi-TeV muons using 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 show that one can recover 20-25% resolution for muons of up to 4 TeV by combining the radiation loss information (we assumed a relative momentum resolution of 20% from magnetic bending performed by ATLAS in mid-rapidity region).

See Lukas Layer’s talk Aug 25!
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

Virtual hadronic loop
By optimizing layout with a discrete sampling, I proved how an improvement of a factor of 2 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.
Sample results

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

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

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 10\(\mu\)m 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 kEuro

My 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 \(O(1\mu)\) 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)+(reconstruction method)+(inference extraction) can provide.
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._
To show what differentiable programming can do for our data analyses, 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 incorporates 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 only 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

Top: Control flow of INFERNO algorithm, which extracts optimal NN parameters given a final analysis objective, such that the resulting measurement becomes maximally robust to nuisance parameters. Data $x$ generated by a simulator $g$ (left block) which depend on nuisances $\theta_{MC}$ are used to train a NN $h$ (second block), producing output $y$. The output is made differentiable by a softmax function (third block) and used to construct a summary statistic $t$, which is histogrammed to compute a saturated likelihood $L_A$ used to perform inference (right block). The expected variance $U$ on the parameter of interest may be derived from the information matrix $I$, and used as NN loss, whose parameters $\varphi$ are optimized by backpropagation.
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 goal:

The target of the **MODE** collaboration is to design and offer to the community a scalable, versatile architecture that can provide end-to-end optimization of particle detectors, proving its performance 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:
- Optimization of Cherenkovs for SWGO
- 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...

«Simple» 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 m³ inside a 0.6x1x1m³ 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 purposely crude model – but once we have something that «breathes», we may start building into it functionality and detail.
*VERY* preliminary results

The code and results shown have been produced by Giles Strong

The 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.

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

Still a looong way to go, but an important milestone for this use case.
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.
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 $p(\cdot)$ on parameters $\theta$ that describe the physical properties of the detector and its geometry.
The observations \( z \) are used by a reconstruction model \( R(\ ) \) that produces high-level features
\[
\zeta(\theta) = R[z, \theta, v(\theta)]
\]
(e.g. particle four-momenta), by employing knowledge of the detector parameters as well as a model of the detector-driven nuisance parameters \( v(\theta) \) which affect the pattern recognition task.

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

Once properly trained for the task at hand, the network produces a low-dimensional summary statistic
\[
s = \text{NN}[\zeta(\theta)]
\]
with which inference can finally be carried out to produce the desired goal of the experiment.
In general, one may formally specify the problem of identifying optimal detector parameters as that of finding estimators $\hat{\theta}$ that satisfy

$$\hat{\theta} = \text{arg min}_\theta \int L[NN(\zeta), c(\theta)]p(z|x, \theta)f(x)dxdz$$

$c(\theta)$ is a function modeling the cost of the considered detector layout of parameters $\theta$, 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 $p(z|x, \theta)$ is not available in closed form –the considered models are implicit–, we must rely on forward simulation: we approximate $\hat{\theta}$ with a sample of $n$ events:

$$\hat{\theta}_a = \text{arg min}_\theta \frac{1}{n} \sum_{i=1}^{n} L[NN(R(z_i)), c(\theta)]$$

where $z_i$ is distributed as $F(x_i, \theta)$ to emulate $p(z|x, \theta)$ as $x_i$ 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.
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,

\[
z = S(y, x, \theta),
\]

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 \( \bar{L}(z) \) by following its surrogate gradient

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

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

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

You are most welcome to participate to MODE activities, or propose to become a member
THANK YOU FOR YOUR ATTENTION!
References


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)
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[xx].

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