Plans for End-to-End Optimization of Experiments

It has become possible, with the advent of differentiable programming, to create models of experimental apparatus that include the stochastic data-generation processes, the full modeling of the reconstruction and inference procedures, and a suitably defined objective function, along with the cost of any given detector configuration, geometry and materials. This enables the end-to-end optimization of the instruments, by using techniques developed within computer science that are currently vastly exploited in fields such as fluid dynamics. The MODE Collaboration has started to consider the problem in its generality, to provide software architectures that may be useful for the optimization of experimental design. These models may be useful in a "human in the middle" system as they provide information on the relative merit of different configurations as a continuous function of the design choices. In this short contribution we summarize the studies that have been done so far and their potential in the long term.

Contact person: Tommaso Dorigo – tommaso.dorigo@gmail.com

Exploiting Differentiable Programming for the End-to-end Optimization of Detectors

Max Aehle[4], Mateusz Bawaj[5], Anastasios Belias[26], Alexey Boldyrev[1,6], Pablo de Castro Manzano[1,2], Christophe Delaere[1,3], Denis Derkach[1,6], Julien Donini[1,7], Tommaso Dorigo[1,2], Auralee Edelen[8], Peter Elmer[24], Federica Fanzago[1,2], Nicolas R. Gauger[4], Andrea Giammanco[1,3], Christian Glaser[1,9], Atılım G. Baydin[1,10], Lukas Heinrich[1,11], Ralf Keidel[12], Jan Kieseler[1,13], Claudius Krause[1,14], Maxime Lagrange[1,3], Max Lamparth[1,11], Lukas Layer[1,2,15], Gernot Maier[16], Federico Nardi[1,2,17,7], Helge E. S. Pettersen[18], Alberto Ramos[19], Fedor Ratnikov[1,6], Dieter Röhrich[20], Roberto Ruiz de Austri[19], Pablo Martínez Ruiz del Árbol[1,21], Oleg Savchenko[2,3], Nathan Simpson[22], Giles C. Strong[1,2], Angela Taliercio[3], Mia Tosi[1,2,17], Andrey Ustyuzhanin[1,6], Pietro Vischia[1,3], Gordon Watts[25], and Haitham Zaraket[1,23]

[1]MODE Collaboration, `https://mode-collaboration.github.io/`
[2]Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Italy
[3]Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, Belgium
[4]Chair for Scientific Computing, Technische Universität Kaiserslautern, Germany
[5]Università di Perugia and INFN, Sezione di Perugia, Italy
[6]HSE University, Russia
[7]Université Clermont Auvergne, Laboratoire de Physique de Clermont, CNRS/IN2P3, France
[8]SLAC National Accelerator Laboratory, USA
[9]Department of Physics and Astronomy, Uppsala University, Sweden
[10]Department of Computer Science, University of Oxford, UK
[11]Physik-Department, Technische Universität München, Germany
[12]Center for Technology and Transfer, University of Applied Sciences Worms, Germany
[13]CERN, Switzerland
[14]Institut für Theoretische Physik, Universität Heidelberg, Germany
[15]Università di Napoli "Federico II", Italy
[16]Deutsches Elektronen-Synchrotron (DESY), Germany

[17]Università degli Studi di Padova, Italy
[18]Department of Oncology and Medical Physics, Haukeland University Hospital, Norway
[19]Instituto de Física Corpuscular, UV-CSIC, Spain
[20]Department of Physics and Technology, University of Bergen, Norway
[21]Instituto de Física de Cantabria, UC-CSIC, Spain
[22]Lund University, Sweden
[23]Multi-Disciplinary Physics Laboratory, Optics and Fiber Optics Group, Faculty of Sciences, Lebanese University, Lebanon
[24]Princeton University
[25]University of Washington
[25]University of Washington

October 25, 2022

Abstract

The coming of age of differentiable programming makes possible today to create complete computer models of experimental apparatus that include the stochastic data-generation processes, the full modeling of the reconstruction and inference procedures, and a suitably defined objective function, along with the cost of any given detector configuration, geometry and materials. This enables the end-to-end optimization of the instruments, by using techniques developed within computer science that are currently vastly exploited in fields such as fluid dynamics.

The MODE Collaboration has started to consider the problem in its generality, to provide software architectures that may be useful for the optimization of experimental design. These models may be useful in a "human in the middle" system as they provide information on the relative merit of different configurations as a continuous function of the design choices. In this short contribution we summarize the plan of studies that has been laid out, and its potential in the long term for the future of experimental studies in fundamental physics.

Scientific context

The optimal choice of layout, characteristics, materials, and information-extraction procedures of a measuring instrument constitutes a loosely constrained problem, featuring a very large number of free parameters related by non-obvious correlations. Although typically quite complex, similar problems may sometimes still be tractable by standard means, in the sense that a parameterized model of the system allows the definition of a likelihood function $L = p(x|\theta)$, given simulated data x, and a solution by minimization of $-\ln L$ with respect to the modelling parameters θ. If, however, the instrument bases its functioning on quantum phenomena such as those governing the interaction of radiation with matter, the optimization problem becomes intractable: the probability $p(x|\theta)$ of observing data x given underlying parameters θ may not be written explicitly. In such circumstances, one has access at best to the generating function of the observed data only through forward simulation, a setting commonly referred to as likelihood-free or simulation-based inference [1].

Over the course of the past eighty years, the intractability of the design optimization problems commonly encountered in fundamental physics has not prevented us from successfully conceiving, commissioning, and operating detectors of huge complexity. The development of increasingly performing instruments followed a robust strategy that, while systematically leveraging technological advancements in electronics and material science, duly exploited well-tested paradigms proven to work by previously acquired experience. For example, a long-standing paradigm for the detection of particles in collider physics experiments has always been the need to measure the momentum of all electrically charged particles by magnetic bending in gaseous or light materials, before exploiting the electromagnetic and hadronic showers produced by both charged and neutral particles in dense matter. Another paradigm common to endeavours in nuclear, particle, and astroparticle physics is the requirement of significant redundancy in the detection systems, to enable cross-calibration of the different components and offer robustness of the resulting inference. A further typical default of such instruments is the choice of a symmetric layout of the detection components, such as the equal spacing of scintillating and passive elements along the depth of a calorimeter, or the spacing of photomultiplier tubes observing a water vessel. While the above mentioned paradigms have a strong motivation in the past successes of particle detection, they are not meant to guarantee the optimality of the devices. In fact, redundancy is the very opposite of optimality; and symmetry of layouts is certainly not the most performant choice when one has to cope with the non-symmetrical nature of energy-dependent processes such as the shower development in a medium. What we argue is that the time is ripe to move away from some of those paradigms, armed with new powerful tools that computer science may now provide.

Objectives and Methodologies

The fast progress of computer science in the past twenty years, together with the development of deep neural networks and optimization software based on differentiable programming, offers us an unprecedented opportunity to rethink the foundations of our design strategies, and to identify and investigate novel, possibly revolutionary solutions we have been unable to figure out by ourselves. The typical design problems we face involve the choice of hundreds, if not thousands of parameters defining the placement and geometry of materials and detection units, their specifications and performance, and their monetary cost. The full exploration of this high-dimensional space of design solutions is a wholly super-human task, and the discrete sampling of the space with full-fledged simulation tools has become completely impractical. To move forward, we must turn to the differentiable programming

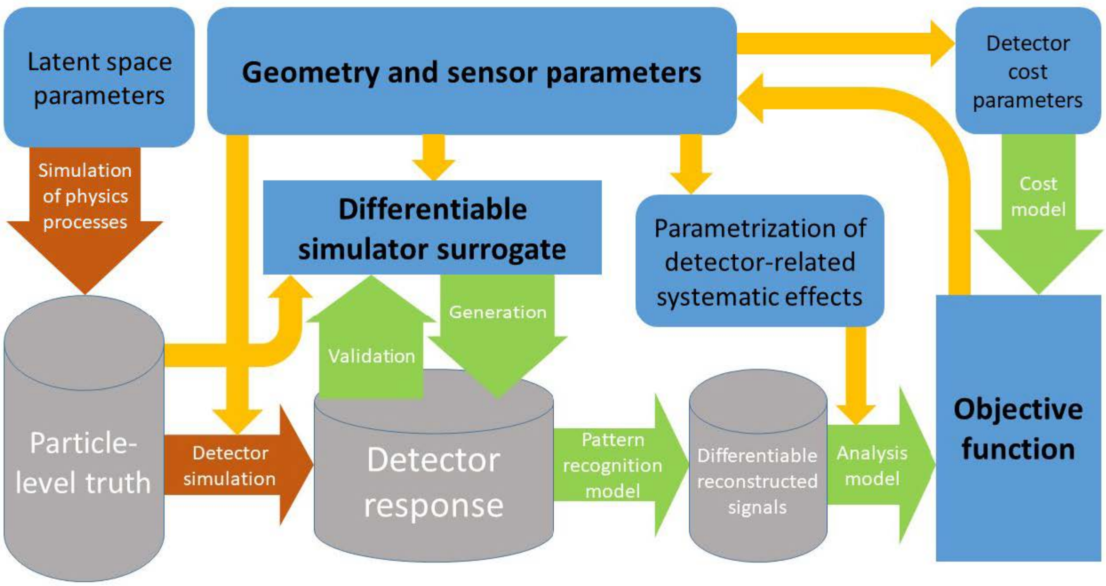

Figure 1: Block diagram for the optimization of a generic detector. Data from a simulator (left, cans labeled "Particle-level truth" and "Detector response") are used to train and validate a differentiable model ("Differentiable simulator surrogate") of the relevant physical processes. Models of pattern recognition, inference extraction, cost of components, and a loss function may then become a function of detector geometry and construction layout parameters. A back-propagation loop of loss derivatives through the functional elements of the system allows their optimization. The figure is adapted from Ref. [2].

tools that make this exploration possible.

It must be noted that the breadth of the space of design solutions has also been increasing with our technological advancements. Nowadays we can 3D print scintillation detectors [3], as well as design more complex detection elements with thin layers of AC-coupled resistive silicon sensors [4]. These advancements can best be exploited if we endow ourselves with the capability of performing continuous scans of the geometry space of the devices we wish to construct: this is something we achieve by developing differentiable programming pipelines.

Another reason for revisiting our detector design paradigms while accounting for the availability and development of new computer science tools is the evolution of the pattern recognition and inference procedures we have been adopting in the extraction of information from raw detector readouts. The demands posed to our instruments are continuously increasing, as we move, e.g., toward the high-luminosity (HL) phase of the Large Hadron Collider (LHC), or toward larger and larger detection volumes in cosmic ray and neutrino physics. At the HL-LHC, in a few years the ATLAS and CMS experiments will be reconstructing high-energy particle collisions within $\mathcal{O}(200)$ pileup interactions taking place during the same bunch crossing; the performance of standard reconstruction algorithms for charged tracks will be strongly reduced in the presence of an exponential increase of the combinatorial background. If deep learning methods will be employed for those pattern recognition tasks, the question arises of whether the detectors have been conceived to be optimal for those tools. Such a potential misalignment between design and exploitation is even more evident if we look further into the future, when new larger experiments are being planned in all fields of fundamental physics. Given that we are currently sitting on a rapidly growing curve of performance of artificial-intelligence-powered methods [5], in order for our future detectors to be most effective we need to consider their design as an optimization problem that includes a model of the pattern recognition and inference extraction procedures available at operation time, however hard it may be to envision their power today.

The above considerations motivated a group of particle physicists, nuclear physicists, and astrophysicists to join forces with computer scientists interested in our scientific use cases and create the MODE Collaboration [1]. The aim of MODE is to pursue a wide-ranging plan of investigations that has the primary purpose of educating ourselves and our communities on how to best integrate all the elements of a detector design problem—from the modeling of the stochastic quantum phenomena to the description of detector layout, geometry, and performance; from the pattern recognition to the inference extraction procedures; and from the interplay of geometry and systematic uncertainties to the physical and economic constraints—into a single optimization problem, as schematized in Fig. 1. We believe that the capability to compute derivatives of the objective function with respect to any one of the parameters of the system, provided by implementing the whole pipeline using differentiable programming, will be key to enable the successful exploration of the large space of design choices, and the discovery of innovative solutions.

At the core of any optimization procedure lays a carefully defined objective function, which should encode as closely as possible the explicit goals of the instrument we are designing. For a large scientific endeavor, specifying this function may at first sight appear an impossible task, given the multi-purpose nature of the detectors, the breadth of physics studies they enable, and the arbitrariness of the relative value of different scientific objectives of the experiment. However, we argue that the exercise of appraising those goals and proposing an evaluation metric *can* be beneficially carried out, and an objective function—or a family of objective functions that address different points of view—can proficuously be specified. Indeed, such an exercise is not altogether different from the one of defining a trigger menu for a collider physics experiment, which produces a list of triggers with relative selection strategies, bandwidths, and prescaling factors: however painful the allocation of bandwidths to different physics datasets may be, this choice is strictly necessary and is routinely operated by the experiments, based on an appraisal of the different goals that those datasets enable.

One recent example of a goal-informed optimization of a detector layout is the one studied by MODE members who participate in the LHCb experiment at the CERN LHC. The upgrade of the electromagnetic calorimeter of LHCb involves the arrangement of photomultiplier modules with three different granularities in a two-dimensional grid (see Fig.2). A differentiable pipeline including a generative adversarial network modeling electromagnetic showers in the detector was developed to optimize the layout, given number of detection units, for the significance of the extractable signal of a B hadron decay involving photons. This allowed the identification of optimized configurations for given detector cost.

Readiness and Challenges

What we are facing is an extremely tall order if we consider a detector of the scale of collider experiments such as ATLAS or CMS, neutrino facilities such as JUNO, nuclear physics experiments such as ALICE, or similar large facilities for astroparticle physics studies. In fact, it is doubtful that we have today the resources, expertise, and skills required to attack problems of that scale of

[1]https://mode-collaboration.github.io

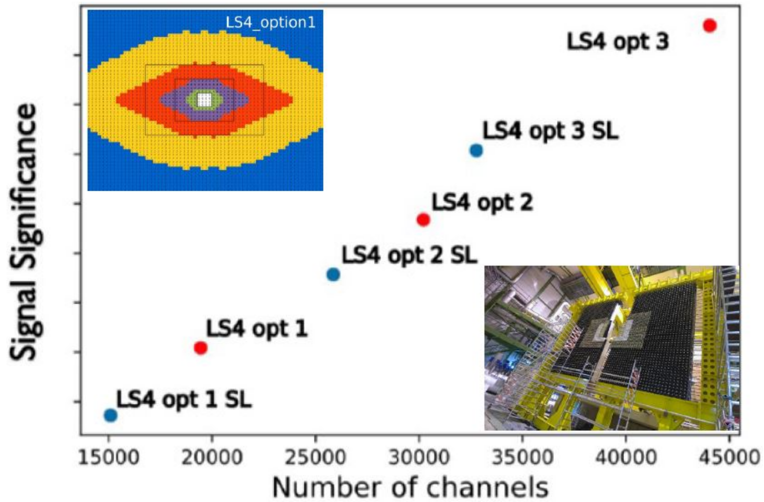

Figure 2: Signal significance (in arbitrary units) of the extractable $B_s \to J/\psi\pi^0$ signal as a function of the number of channels in the electromagnetic calorimeter for the LHCb upgrade. The left inset shows an arrangement of different granularity of photomultipliers; the right picture shows the current front face of the calorimeter.

complexity. But that does not mean that we should give up this challenge; rather, we must proceed in steps, by considering at first less ambitious and more achievable goals. We have therefore laid out a plan [6] for a series of design optimization tasks that are interesting for the field in their own right, and whose solution via the above plan may enable us to build a framework of methods and software tools that together may constitute the building blocks for solving harder problems. While the specificity of a detector leaves little room for reuse of the differentiable surrogate models of particle interaction with active and passive components that may have been developed to study them, there is instead significant device independence in recently developed reconstruction algorithms empowered by deep learning [7], and a clear possibility of reusing the models developed for the monetary cost of the components, for the interaction between geometry- and detector-related systematic uncertainties, and for inference extraction.

In terms of readiness, the discussed technology is in a development stage; we expect that the demonstration of its performance will be produced on a significant number of medium-size tasks in the next three to five years. Existing challenges include computing availability and cost, complexity of the software architecture and its use for a global optimization task, and precision of the generative models employed as surrogates for the full simulation of the physical processes. None of them appears to be a show-stopper, as the example shown in Fig. 2 above exemplifies.

Outlook

The optimization study which may result from the development of the full model of an experiment, along with the specification of desirable experimental goals, cannot be expected to produce a final answer for the absolute “best” configuration –something which is hardly well-defined or practical, given the existence of external constraints and details that do not belong to a computer model. Rather, such a study may indicate advantageous combinations of design choices and “sweet spots” in the space of design parameters which provide invaluable information to guide our hand toward robust yet effective decisions. The focus, in other words, is to empower the human in the middle –the detector builder– with tools that while very complex to develop have a great potential of revolutionizing the performance of our instruments.

In summary, the paradigm shift that we envision is constituted by moving away from the discrete sampling of a necessarily very limited number of possible detector configurations, and into the fully continuous mapping of the utility function in the very high-dimensional space of design choices which is enabled by complete differentiable models. In a world increasingly plagued with global challenges (pandemics, overpopulation, famine, climate change) and the consequent higher demand for applied-science solutions and lower appeal of investments on fundamental science, we believe that the NuPECC community should welcome the efforts we have described in this short document, recognize their potential in increasing the performance-per-euro that future detector designs may produce, and support this long-term plan which is meant at benefiting nuclear, neutrino, particle, and astroparticle physics experiments across the board.

References

[1] K. Cranmer, J. Brehmer and G. Louppe, The frontier of simulation-based inference, Proc. of the National Academy of Sciences 117, 48 (2020) 30055, arXiv:1911.01429[stat.ML].

[2] A.G. Baydin *et al.*, Toward Machine Learning Optimization of Experimental Design, Nuclear Physics News 31, 1 (2021) 25, https://doi.org/10.1080/10619127.2021.1881364.

[3] Y. Mishnayot *et al.*, Three-dimensional printing of scintillating materials, Rev. Sci. Instrum. 85 (2014) 085102, arXiv:1507.4817[cond-mat.mtrl-sci], doi.org/10.1063/1.4891703.

[4] G. Giacomini, W. Chen, G. D'Amen, and A. Tricoli, Fabrication and performance of AC-coupled LGADs, JINST 14,09 (2019) 9004, arXiv:1906.11542[physics.ins-det], https://doi.org/10.1088/1748-0221/14/09/p09004.

[5] D. Hernandez and T.B. Brown, Measuring the Algorithmic Efficiency of Neural Networks, arXiv:2005.04305[cs.LG] (2005).

[6] T. Dorigo *et al.*, Toward the End-to-end Optimization of Particle Physics Instruments with Differentiable Programming: a White Paper, arXiv:2203.03818[physics.ins-det], https://doi.org/10.48550/arXiv.2203.13818.

[7] J. Kieseler, Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph and image data, Eur. Phys. J.C. 80, 9 (2020) 886, arXiv:2002.03605[physics.data-an], https://doi.org/10.1140/epjc/s10052-020-08461-2.