Reviews in Physics 13 (2025) 100120

Contents lists available at ScienceDirect

Reviews in Physics

journal homepage: www.elsevier.com/locate/revip

Long review

Progress in end-to-end optimization of fundamental physics experimental apparata with differentiable programming

Max Aehle [b,a], Lorenzo Arsini [c,d], R. Belén Barreiro [e], Anastasios Belias [f], Alexey Boldyrev [g,a], Florian Bury [h], Susana Cebrian [i], Alexander Demin [g], Jennet Dickinson [j], Julien Donini [k,l,a], Tommaso Dorigo [m,n,a,l], Michele Doro [o,n], Nicolas R. Gauger [b,a], Andrea Giammanco [p,a], Lindsey Gray [j], Borja S. González [q,r], Verena Kain [s], Jan Kieseler [t,a], Lisa Kusch [u,b,a], Marcus Liwicki [m,a], Gernot Maier [v], Federico Nardi [k,w,a], Fedor Ratnikov [g,a], Ryan Roussel [x,a], Roberto Ruiz de Austri [y], Fredrik Sandin [m,a], Michael Schenk [s], Bruno Scarpa [w], Pedro Silva [s], Giles C. Strong [n,a], Pietro Vischia [z,a,*]

[a] MODE Collaboration [1]
[b] Chair for Scientific Computing, University of Kaiserslautern-Landau, Germany
[c] La Sapienza Università di Roma, Rome, Italy
[d] Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Italy
[e] Instituto de Física de Cantabria, UC-CSIC, Spain
[f] GSI Helmholtzzentrum für Schwerionenforschung GmbH, Germany
[g] HSE University, Russia
[h] University of Bristol, United Kingdom
[i] Centro de Astropartículas y Física de Altas Energías (CAPA), Universidad de Zaragoza, Spain
[j] Fermi National Accelerator Laboratory, USA
[k] Université Clermont Auvergne, Laboratoire de Physique de Clermont, CNRS/IN2P3, France
[l] Universal Scientific Education and Research Network, Padova, Italy
[m] Luleå, University of Technology, Sweden
[n] Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Italy
[o] Dipartimento di Fisica e Astronomia, Università di Padova, Italy
[p] Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, Belgium
[q] Laboratório de Instrumentação e Física Experimental de Partículas – LIP, Lisbon, Portugal
[r] Instituto Superior Técnico – IST, Universidade de Lisboa – UL, Lisbon, Portugal
[s] CERN, Switzerland
[t] Karlsruhe Institute of Technology, Germany
[u] Computational Illumination Optics Group, Eindhoven University of Technology, The Netherlands
[v] Deutsches Elektronen-Synchrotron (DESY), Germany
[w] Universitá degli Studi di Padova, Italy
[x] Stanford Linear Accelerator Center, United States
[y] Instituto de Física Corpuscular, UV-CSIC, Spain
[z] Universidad de Oviedo and ICTEA, Spain

ARTICLE INFO

Keywords:
Particle detectors

ABSTRACT

In this article we examine recent developments in the research area concerning the creation of end-to-end models for the complete optimization of measuring instruments. The models we

[*] Corresponding author at: Universidad de Oviedo and ICTEA, Spain.
E-mail address: vischia@uniovi.es (P. Vischia).

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

https://doi.org/10.1016/j.revip.2025.100120
Received 2 September 2024; Received in revised form 15 April 2025; Accepted 26 April 2025

Available online 2 May 2025

Differentiable programming
Machine learning
Optimization
Particle physics
Nuclear physics
Astrophysics

consider rely on differentiable programming methods and on the specification of a software pipeline including all factors impacting performance — from the data-generating processes to their reconstruction and the inference on the parameters of interest — along with the careful specification of a utility function well aligned with the end goals of the experiment.

Building on previous studies originated within the MODE Collaboration, we focus specifically on applications involving instruments for particle physics experimentation, as well as industrial and medical applications that share the detection of radiation as their data-generating mechanism.

This report illustrates the most recent advancements in the area, and outlines, for each of the discussed applications as well as for automatic differentiation itself, ongoing and future work.

1. Introduction

In the course of the past few centuries, progress in fundamental science has tracked quite closely the corresponding progress in our technological skills and the capability of conceiving, constructing, and operating increasingly complex measuring instruments. An example of such synergy is offered by the history of particle physics, whose research in the past seventy years posed increasingly challenging demands on the performance of devices that base their functioning on the interaction of radiation with matter.

While this field continues to benefit from advancements in the production of high-performance gaseous and solid-state detectors and related electronics, we believe the most notable and ground-breaking technological advancement in the past two decades has been the maturation of machine learning methods, further enhanced by more affordable and powerful computing infrastructures. It is therefore only natural to exploit that advancement in the design of instruments meant to further our understanding of Nature. Besides, the sheer scale of the experiments we perform today lends itself naturally as a challenge to be addressed with solutions offered by computer science. Specifically, the non-trivial interrelation between the outputs of the large number of components and subsystems making up a modern particle detector causes a significant difference between the result of optimization procedures considering each the subsystems as separate entities, which are necessarily based on local figures of merit, and the global optimization of the system as a whole, which can directly be based on an utility function aligned to the experimental goals. An example of this sort of misalignment is well known to data acquisition specialists at hadron colliders: the fixed nature of the total bandwidth available for data collection poses luminosity-dependent constraints on the effective cross section of data selection recipes corresponding to each individual trigger stream, each of which results from careful optimization of selection strategies performed separately. During occasional higher-than-average luminosity runs, where the average accept rate is too high, the probability that accepted events fail to be stored is finite and non-negligible ("dead time"); the urgency of reducing this dead time forces to develop rate reduction strategies that cannot account for the overall scientific goals of the experiment.

By and large, the engine under the hood of most modern machine learning methods is what has come to be called *Differentiable Programming* (DP). DP relies on automatic differentiation (AD) procedures, which are nowadays offered by several common tools (PyTorch [1], TensorFlow [2], and JAX [3], among others). Benefiting from the automatic computation of derivatives of whole pieces of software simplifies greatly the search for the extremum of arbitrarily complex utility functions, by employing the standard technique of gradient descent. While DP is not the only solution to large-scale holistic optimization problems, and may not be viable in specific cases, we find it particularly suitable to allow for a unified modeling of the various parts of a global optimization task of the kinds of interest in our research area.

The intrinsically stochastic nature of the data-generating processes of interaction of radiation with matter, arising from quantum phenomena, is a source of additional, conspicuous complications in the way of the creation of a full differentiable model of the whole problem. Workarounds based on the creation of surrogate models often provide viable solutions, which are however invariably rather specific to the problem at hand [4]. This is one of the main stumbling blocks in the creation of versatile, multipurpose architectures for end-to-end optimization. Still, we believe that the solution of a significant number of different problems of low to medium complexity will empower our community to construct solutions to still harder, larger-scale experiment optimization tasks, such as those concerning detectors for a future very-high-energy particle collider.

The present document constitutes an update and an extension of the ideas and the use cases that some of us recently described in a previous work [5]. We have structured it as follows. In Section 2 we discuss the state of the art and the recent developments of computer science tools involved in the deployment of end-to-end models for optimization. The following sections discuss separate use cases for end-to-end optimization. In Section 3 we discuss applications to muon tomography, where we have made our first attempts at a full solution of the optimization problem. In Section 4 we consider the possible advancements in high-precision calorimetry by exploiting differentiable models to study the integration and hybridization of tracking and calorimetry devices. Section 5 discusses applications in accelerator optimization. Section 6 discusses several use cases from fundamental research in experimental astroparticle physics. We discuss in Section 7 the possibility of integrating neuromorphic computing devices in particle detectors and of exploiting this innovative computing paradigm for the optimization tasks at the focus of this work. Section 8 deals with progress in optimization tasks for the benefit of detectors for medical applications. Finally, we provide a brief overlook of this young but promising new field of research in Section 10.

2. Progress in AD methods

2.1. Towards algorithmic differentiation of GATE/Geant4

Simulations of the detection process are an essential step in the assessment of a proposed detector design. Often, complex Monte Carlo simulators like Geant4 [6–8] are employed for this task, as they provide the most realistic and adaptable computational models for the interactions between particles and the detector. In addition, however, many simplified simulators and surrogate models have been proposed in the literature.

Naturally, detector optimization toolchains derived from assessment pipelines contain at least one of these simulators or models. If changes in the detector design have only minor effects on the particles, as is the case e.g. in the muon tomography setup of TomOpt (described in Section 3), there is no need to differentiate through the particle simulation. Otherwise, e.g. in the case that the geometrical layout of absorber material in the detector changes, the particle simulation code becomes part of the objective function to be optimized.

One possible approach to algorithmically differentiate such a complicated objective function would be to replace any Monte Carlo particle simulators by surrogate models. Using such surrogate models has the following advantages:

- after a training phase, surrogate models can run faster;
- surrogate models might provide a “smoother” approximation that can be better suited for gradient-based optimization;
- ultimately, applying AD/DP directly to big software projects like Geant4 is generally considered a severe technical challenge, given their size and complexity.

However, the training and use of surrogate models consumes developer and computation time as well. Because of the additional modeling step, assessment pipelines based on a surrogate model may produce less realistic results and thus steer an optimization procedure towards sub-optimal or biased designs. In order to compare the optimization results achievable with surrogate models and the direct application of AD/DP to particle physics Monte Carlo codes, we first need to overcome the technical challenge associated with the latter. In this section, we present a successful application of an AD tool to a Geant4 4 application, computing derivatives of physical output variables (hit coordinates) with respect to physical input variables (kinetic energy of primary particles). Surrogate models are further discussed in Section 4.2 and Ref. [4].

To this end, we recently created the AD tool Derivgrind[2] [9,10]. AD tools identify real-arithmetic operations in the *primal program*, and create a program that computes the derivative of the primal program’s *output variables* with respect to its *input variables*, where both sets of variables are defined by the user. AD tools exist in the shape of, e.g. execution environments for domain-specific languages, programs transforming the source code, class definitions, or compiler plugins. Unlike traditional source-code-based AD tools, Derivgrind operates on the machine code of the primal program just before it runs on the processor. In the best case, machine-code-based AD reduces the developer interaction with the primal program’s source code to the necessary minimum of inserting macro calls indicating input and output variables. Derivgrind is therefore well-positioned for initial explorative studies about the direct application of AD/DP to complex, cross-language or partially closed-source software projects.

Derivgrind utilizes the dynamic binary instrumentation framework Valgrind [11] to discover floating-point instructions in the compiled primal program, and to insert the corresponding AD logic. Some real arithmetic can also be performed by binary manipulation of floating-point data. Derivgrind handles the most important constructs of this sort correctly. However, more obscure “bit-tricks” are hard to discover, and accordingly must not be present in the primal program.

In this work, we apply Derivgrind to GATE (v9.2) [12], a software package built on top of Geant4 (v11.0.0) for simulations in medical imaging and radiotherapy. Our setup is related to a proton computed tomography (pCT) scanning process with a digital tracking calorimeter (DTC) developed by the Bergen pCT collaboration [13]; in Section 8.2, we give more details on the tomographic reconstruction algorithm that makes use of the post-processed DTC measurements. As in Ref. [14], we consider the beam energy as an input variable x for AD, simulate a single proton passing through a human head and the DTC, and record the first coordinate of the hit in the ith tracking layer of the DTC as the AD output variable $f_i(x)$ for $i = 1, 2$. For all runs of GATE/Geant4 4, we kept the seed of the random number generator fixed, so random numbers are treated like constants from the perspective of differentiation. As apparent from Fig. 1, these functions f_1 and f_2 are piecewise differentiable. Fig. 2 shows the central difference quotient $\frac{f_i(x_0+h)-f_i(x_0-h)}{2h}$ around $x_0 = 230\,\mathrm{MeV}$ for various values of h. Mathematically, this difference quotient converges to the derivative $\frac{\partial f_i}{\partial x}(x_0)$ in the limit $h \to 0$. On a computer however, floating-point inaccuracies become large for small h.

Listings 1 and 2 show our insertions into GATE’s source code, which was slightly refactored beforehand for the purpose of presentation. In the first code block of either listing, GATE reads the energy x from the configuration file and sets the respective property of the beam source object. The second code block is run whenever the proton hits a layer, to assemble GATE’s output data. The calculations performed inbetween involve Geant4, which we do not modify at first. The inserted macros are defined in a header `derivgrind.h` and perform a “client request” from the primal program to the Derivgrind process running it, basically declaring the AD input and output variables.

Listing 1 shows the insertion of forward-mode client requests, which give access to the *dot value* $\frac{\partial a}{\partial x}$ of any floating-point value a. In the first block, we set the dot value $\frac{\partial x}{\partial x}$ of the beam energy x to 1.0. In the second block, we extract the particular output variable

[2] https://github.com/SciCompKL/derivgrind

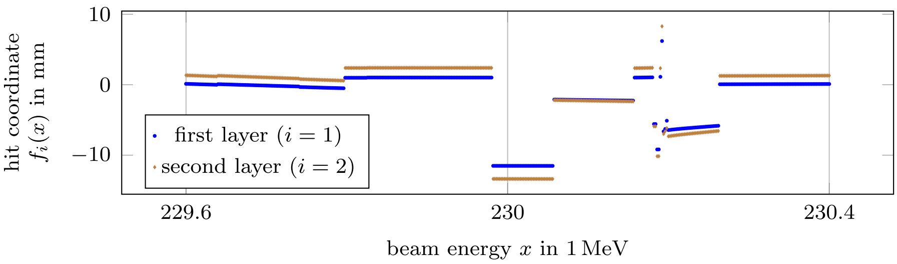


Fig. 1. Graph of the function that we apply AD to.

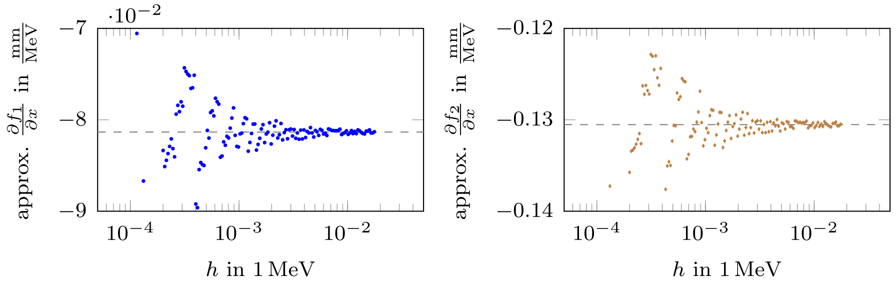


Fig. 2. Central difference quotients $\frac{f_i(x_0+h)-f_i(x_0-h)}{2h}$ (blue and brown markers) around $x_0 = 230\,\text{MeV}$, and the derivative computed with Derivgrind (dashed line) after changing G4Log to log. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Listing 1: Insertions into the source code of GATE for forward-mode differentiation. Seed the input variable (beam energy), and print the dot value of the output variables (hit positions). Additionally, the header derivgrind.h must be included.

```
   if (command == pEnergyCmd) {
     double energy = pEnergyCmd->GetNewDoubleValue(newValue);
+    double one = 1.0;
+    DG_SET_DOTVALUE(&energy,&one,sizeof(double));
     pSourcePencilBeam->SetEnergy(energy);
   }
```

```
   if (m_rootHitFlag) m_treeHit->Fill();
+  float pos = *(float*)(m_treeHit->GetBranch("posX")->GetAddress());
+  float pos_d;
+  DG_GET_DOTVALUE(&pos,&pos_d,sizeof(float));
+  std::cout << "pos_d=" << pos_d << "\n";
```

Listing 2: Insertions into the source code of GATE for reverse-mode differentiation. Declare the input variable (beam energy) and output variables (hit positions). Additionally, the header derivgrind.h must be included.

```
   if (command == pEnergyCmd) {
     double energy = pEnergyCmd->GetNewDoubleValue(newValue);
+    DG_INPUTF(energy);
     pSourcePencilBeam->SetEnergy(energy);
   }
```

```
   if (m_rootHitFlag) m_treeHit->Fill();
+  float pos = *(float*)(m_treeHit->GetBranch("posX")->GetAddress());
+  DG_OUTPUTF(pos)
```

$f_i(x)$ of interest for AD, to retrieve and print its dot value $\frac{\partial f_i}{\partial x}$. Only the two modified source files of GATE need to be recompiled. Running GATE under Derivgrind reproduces the original output of GATE, interleaved with additional output from Derivgrind and the sought derivatives.

Listing 2 shows the insertion of recording-mode client requests to mark x as an input variable and $f_i(x)$ as an output variable for $i = 1, 2$. Applied to the modified GATE program, Derivgrind records the real-arithmetic evaluation tree (*tape*) and identifiers (*indices*) for the input and output variables in the tree. Reverse-mode AD is about tracking the *bar value* $\frac{\partial f_i}{\partial a}$ of all floating-point variables a (here, for one output at a time). A simple tape evaluator program in the Derivgrind package can be used to allocate space for all bar values, set the bar value of the output variable f_i to 1, and evaluate the bar value $\frac{\partial f_i}{\partial x}$ of the input variable x according to the

Table 1
Numerical and automatic derivatives of f_1 and f_2 at $x_0 = 230\,\mathrm{MeV}$.

Differentiation method	Approximation of $\frac{\partial f_i}{\partial x}$ in $\frac{\mathrm{mm}}{\mathrm{MeV}}$	
	$i = 1$	$i = 2$
Central difference quotient		
… for $h = 0.01\,\mathrm{MeV}$	−0.0812531	−0.130463
… for $h = 0.005\,\mathrm{MeV}$	−0.0811577	−0.130272
… for $h = 0.001\,\mathrm{MeV}$	−0.0815392	−0.131130
Derivgrind, original Geant4		
… forward mode	−0.0685116	−0.113841
… reverse mode	$-1.72 \cdot 10^{8}$	$5.36 \cdot 10^{13}$
Derivgrind, `G4Log` $\rightsquigarrow$ `log`		
… forward mode	−0.0813391	−0.130524
… reverse mode	−0.0813391	−0.130524

tape. Reverse-mode AD finds the bar values of all input variables in one sweep and is therefore likely to provide a better run-time than numeric differentiation for optimization problems with many design parameters.

Table 1 lists the computed derivatives. The forward-mode automatic derivatives deviate from the difference quotients by about 15 %, while the reverse-mode derivatives are completely off. Printing all the results of intermediate floating-point operations alongside their dot values and comparing with difference quotients, we were able to identify the statement at which they start to differ significantly. Geant4 defines an alternative math function `G4Log` to numerically approximate the natural logarithm $\log z$ for $z \in \mathbb{R}_+$, using an approximation algorithm adapted from the VDT math library [15]. The algorithm starts with a range reduction step, multiplying the argument z by 2^k for a suitable integer k such that $\frac{1}{2} \leq z \cdot 2^k < 1$, by simply overwriting the exponent bits of z. Derivgrind does not recognize the real-arithmetic significance of this bit-trick.

Thus, we replaced `G4Log` by a call to the standard C `log` function. GLIBC, and other implementations of the C standard library, also use bit-tricks to implement `log`; however, Derivgrind recognizes and intercepts calls to C 95 math functions, and uses the proper analytical derivatives. After this small code change in Geant4, the automatic derivatives computed by Derivgrind's forward and reverse mode agree, and we indicated them by horizontal lines in Fig. 2. As these lines are surrounded from both sides by the markers indicating difference quotients, the automatic derivatives are either entirely correct, or at least their deviation from the true derivative is small compared to the variance of difference quotients.

We have measured the run-times for a release-mode build of GATE/Geant4 with the `time` command on an exclusive node with two 2.6 GHz Intel Xeon Gold 6126 processors at the University of Kaiserslautern-Landau's Elwetritsch cluster. The runtime goes up from around 12 s in native execution to 13 min in the forward mode, which is a factor of 65. Derivgrind's recording takes about 24 min, corresponding to a factor of 120, to record a tape of 25 MB whose reverse evaluation takes about 0.05 s.

To summarize, the AD tool Derivgrind provides accurate derivatives for the parts of Geant4 analyzed in this study. Besides applying macros to input and output variables, the only change to the source code was to replace a alternative implementation of a math function by a call to the C math library. Therefore, it now becomes possible to include realistic Monte Carlo particle simulations into differentiable pipelines, instead of using surrogate models. Further research may compare these two approaches with respect to computational performance, convergence behavior of the optimizer, and quality of the optimized designs.

2.2. Optimization of derivatives using polyhedral compiler

Derivatives, mainly in the form of gradients and Hessians, are ubiquitous in machine learning and in frequentist and Bayesian inference.

Traditionally, most AD systems have coded in high-level programs [1,16], and therefore have been unable to achieve a good performance on scalar code or memory-modifying loops. These systems have instead relied on extensive libraries of optimized kernels — e.g. for linear algebra, convolutions, or probability distributions — combined with associated adjoint rules, forcing practitioners to express their models in terms of these kernels to attain satisfactory performance.

New, low-level AD tools such as Enzyme [17] allow for differentiating kernels implemented as naive loops. However, since the derivative code is generated programmatically during the application of reverse mode AD, the reverse passes are likely to access memory in patterns suboptimal with respect to both cache and SIMD performance. This opens a new opportunity for polyhedral loop-and-kernel compilers to provide the aggressive transforms needed for high performance. Polyhedral compilers [18] are standalone compilers or libraries that can be integrated in an existing compiler: they are based on a special representation of a computer program in terms of parametric polyhedra, which then allows using techniques from geometrical optimization to provide efficient compilation.

In this section, we report on work in progress on the implementation of the LoopModels[3] automatic loop optimization library based on LLVM Intermediate Representation (IR) [19], which uses polyhedral dependency analysis methods to attain a performance competitive with vendor libraries on many challenging loop nests, with applications from linear algebra to the derivative code produced by AD.

[3] Available at https://github.com/JuliaSIMD/LoopModels

2.2.1. Techniques and applicability

For many reasons, preserving performance optimizations during differentiation is a non-trivial task, in particular in the context of reverse-mode automatic differentiation. The footprint of read/write memory accesses of the generated derivative (adjoint) program usually differs drastically from that of the original program. Implementation choices such as hand-tuned tiling strategies in the primal code are often not ideal for the derivative code. While the overhead of suboptimal derivative programs is sometimes considered to be constant and non-significant, it is often limiting in high-performance, demanding applications.

The LoopModels library will provide a compiler pass based on polyhedral modeling techniques, which will perform source code optimizations and autonomously certify the correctness of transformations. This will allow the automatic optimization of adjoint programs, and hopefully alleviate the need for expert human intervention. Ultimately, the library should allow machine generation of optimal code even for challenging cases like reverse-mode automatic differentiation of expressions containing highly-nested loops. By working at the LLVM IR level, LoopModels will naturally benefit from existing solutions in the LLVM compiler toolchain, thus generating high-performance gradients of any language going to the LLVM intermediate representation, such as C/C++, Fortran, Julia, Rust, and others, with a small overhead at compilation time.

LoopModels will rely on polyhedral methods for analyzing the loop programs. The polyhedral model techniques for compiler optimization provide a powerful mathematical framework to represent nested loop computation and its data dependences using integral points in polyhedra. This approach works by finding beneficial affine code transformations through a practical cost function that enables efficient fusion and tiling of arbitrarily nested loops in a synthesized adjoint program. This allows simultaneous optimization for coarse-grained parallelism and locality.

2.2.2. Simple example

Let us consider, as an example, a nontrivial loop transformation that LoopModels would be capable of discovering and applying. The convolution operation is key to many applications and is one of the building blocks in machine learning (ML), and in particular deep learning (DL), pipelines. When used in a convolutional neural network (CNN), the backpropagation stage in AD also requires the calculation of the gradient of the convolution with respect to its arguments during the reverse pass.

We examine the 1-D case, where given vectors $A \in \mathbb{R}^{N+I}$, $B \in \mathbb{R}^{I}$, and $C \in \mathbb{R}^{N}$, the convolution of A and B is defined for all $0 \leq n \leq N-1$ as:

$$C_n = \sum_{i=0\ldots I-1} A_{n+i} \cdot B_i \,.$$

Note that zero-based indexing is used here. The pseudocode in Listing 3 provides a basic implementation of the convolution operation.

Listing 3: Original program.

```
function convolution(C, A, B)
    N, I = length(C), length(B)
    for n in 0:N-1 do
        for i in 0:I-1 do
            C[n] += A[n + i] * B[i]
        end
    end
end
```

Listing 4: Adjoint program.

```
function adjoint($\overline{C}$, $\overline{A}$, B)
    N, I = length($\overline{C}$), length(B)
    for n in 0:N-1 do
        for i in 0:I-1 do
            $\overline{A}$[n + i] += $\overline{C}$[n] * B[i]
        end
    end
end
```

One of the possible outputs of reverse mode AD applied to the program of Listing 3 is displayed in Listing 4. The function corresponds to the backpropagation algorithm that computes the adjoint $\overline{A}$, given the adjoint $\overline{C}$ and B; in this example, we ignore differentiation with respect to B.

While in this case the forward pass is easy to optimize via register tiling, this is not the case for the adjoint: the index into $\overline{A}$ in the innermost loop is dependent on both loop induction variables i and n, making it impossible to hoist these memory loads and stores out of any loops. This forces us to re-load and re-store memory on every iteration, requiring several additional CPU instructions per multiplication.

In Listing 5, a possible optimized adjoint program is presented. The transformation uses the observation that we can re-index the memory accesses to $\overline{A}$ so that it depends on one loop only. By introducing new loop induction variables and adjusting the loop boundaries, we can rewrite the inner loop in a way that allows a register-efficient tiled access pattern.

Listing 5: Optimized adjoint program.

```
function adjoint_optimized(C̄, Ā, B)
    N, I = length(C̄), length(B)
    for w in 0:N + I - 2 do
        for j in max(0, w - N):min(I - 1, w)
            Ā[w] += C̄[w - j] * B[j]
        end
    end
end
```

2.2.3. First experiments

To justify the potential applicability of code transformations implemented in LoopModels, we report on experimental results produced within a proof-of-concept implementation. For comparison, we consider PyTorch [1], and two ML libraries implemented in the Julia language [20]:

- `Flux.jl`, a Julia general-purpose library for ML that uses high-level AD tools;
- `SimpleChains.jl`, a Julia library that uses handwritten programs for computing adjoints and applies code transformations in the spirit of the approach proposed in LoopModels.

The performance of `SimpleChains.jl` was compared to analogues on small-dimensional datasets. Results are looking promising so far: for example, on MNIST dataset [21] with a LeNet architecture the full training pipeline in `SimpleChains.jl` took 1.5 s vs. the 50 s in `Flux.jl` and 15 s in PyTorch, respectively.[4]

3. Progress in muography optimization

In this section we describe the software package TomOpt (*Differential Optimization of Muon-Tomography Detectors*), which is the first concrete effort within the MODE Collaboration to research and develop differential optimization techniques for detector design. Rather than immediately attempting to tackle LHC-scale instruments, we instead opt for the simplified, but nonetheless useful, domain of muon tomography, where both the detectors and inference chains are more straightforward to configure and control. We described the details of TomOpt in a recent publication [22]. TomOpt is a highly modular Python-based package that provides the full suite of tools and resources required for the investigation of the general problem of optimization of a scattering tomography detector.

3.1. Muon tomography

Muons, elementary particles related to the electrons but about 200 times heavier, are produced by cosmic-ray interactions in the atmosphere. Their flux at sea level is of the order of $100\,\mathrm{Hz\,m^{-2}}$, and their energy spectrum is very broad, peaking at a few GeV and extending up to the TeV scale. In the energy range $1\,\mathrm{GeV}$ to $100\,\mathrm{GeV}$, muons mostly loose energy by ionization, at a rate of about $200\,\mathrm{MeV}$ per meter of water. This makes them the most penetrating charged elementary particles. When traversing a material, muons undergo several elastic electromagnetic interactions with the nuclei of the traversed material (*multiple scattering*), As the strength of each collision depends on the charge of the nucleus, the deflection of a muon trajectory has a known dependence on the atomic number Z [23,24] of the traversed material. This dependence can be inverted, to infer the atomic number of an unknown material by measuring the scattering angle of a batch of muons that scatter through it. The measurements are typically performed by means of two groups of layers of muon detectors, one above and one below the passive volume (we use "passive volume" to refer to the volume to be scanned). The muon trajectory above and that below are fitted using the hits generated in the layers by the muon passage, and the scattering angle between the two directions can be measured. TomOpt models this process in a differentiable pipeline where the detector parameters can be optimized by minimizing through AD-powered gradient descent a loss function that includes both the physics goal of the experiment and the cost of the detector configuration.

3.2. Package overview

TomOpt is built as a modular and user-inheritable Python package, backed by PyTorch [1]. It is currently under development, with an open-source release planned soon, along with accompanying dedicated publications.

The package implements all aspects of the simulation, detection, inference, and optimization without external heavy dependencies. However, given the wide variety of possible applications, these aspects are presented as base classes, which are designed to be inherited by users and configured for their exact use-cases.

[4] See also the talk by Chris Elrod at https://youtu.be/rfBYA1gZa6E, last visited on February 2023

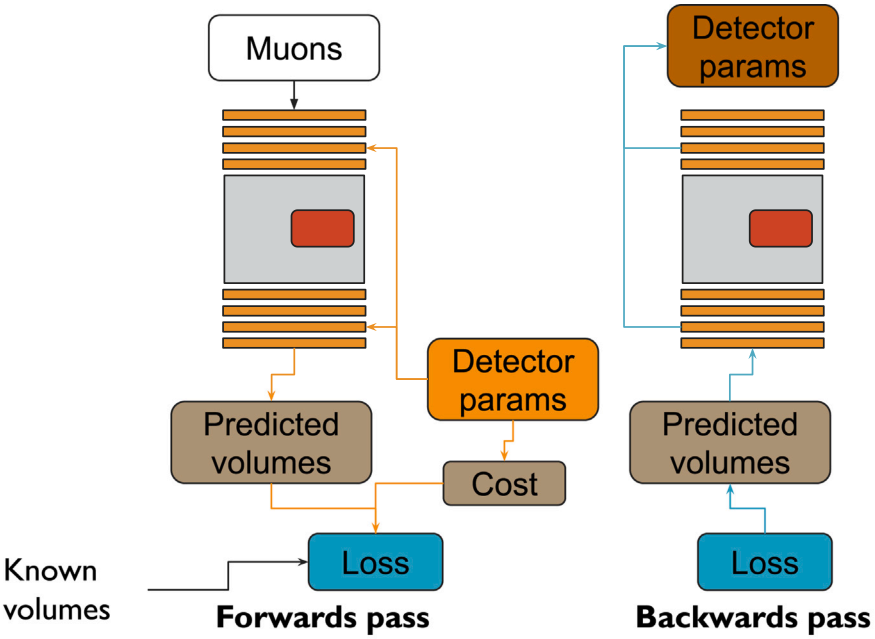


Fig. 3. Forwards and backwards passes in TomOpt.

3.3. Usage

TomOpt is designed to iteratively adjust a detector system such that it becomes optimal; where optimality is quantitatively defined though the minimal value of a task-specific loss function that depends on the detector parameters.

When setting up a problem, users define detector panels with an initial position and size, and also specify the dimensions of the passive volume to be imaged. Next, users must provide both a differentiable inference method, and a loss function. The former is used to provide predictions on properties of the passive volume, and the latter quantifies the error on these predictions. The exact nature of both of these methods will be dependent on the users' tasks, but TomOpt provides starting base classes for a range of problem categories.

In order to optimize the detector parameters (sizes and positions), typical layouts for the passive volume are sequentially loaded and inferred on. These may either be manually specified by the user, or generated by a suitable function. Inference is performed per passive volume layout using batches of many muons. The loss function may then be computed over batches of several passive volume layouts. The use of a differentiable inference method means that the analytic effect of each detector parameter may be computed via back-propagation of the loss gradient, and the parameters iteratively updated via gradient descent, as illustrated in Fig. 3

3.4. Example

To better describe the usage, let us consider a concrete task: we need to scan containers in order to search for smuggled uranium, which might be hidden amongst scrap metal.

Loss function. We can consider this a binary classification exercise, in which each container belongs to one of two classes: contains uranium or does not contain uranium. For such a task we can use binary cross-entropy as a loss function:

$$\mathcal{L}(y, \hat{y}(\theta)) = -y \ln(\hat{y}(\theta)) - (1-y)\ln(1-\hat{y}(\theta)), \tag{1}$$

where $y \in \{0, 1\}$ are the true class labels, and $\hat{y}(\theta) \in [0, 1]$ are the predicted labels based on detector parameters θ.

Passive volumes. In order to simulate the passive volumes, we can generate examples by filling a metal container with a randomly varying amount of assorted metal, inter-spaced with air. The top of the container is also filled with air. With a specified probability, we can possibly then place a block of uranium of random shape inside the container, at a random location. An example volume is shown in Fig. 4.

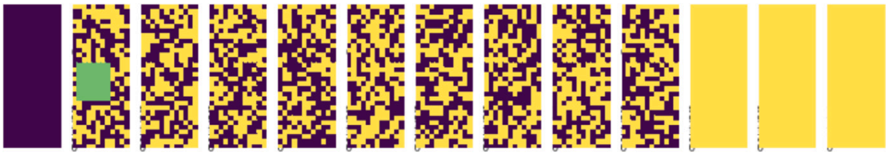


Fig. 4. Example: a passive volume in horizontal cross-sections starting from the bottom layer to the top. Yellow voxels indicate air, blue scrap metal, and green uranium. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Detectors. We define detectors as panels placed parallel above and below the passive volume. When muons pass through the panels, hits will be recorded with a certain spatial resolution. These hits can later be used to infer the trajectory of the muon before and after traversing the passive volume. The detector panels each have five learnable parameters: position in x, y, z, and span in x, y.

Since a physical detector-panel will either record a hit or not, depending on whether the muon passes through the panel, the hits would not be differentiable with respect to the x, y parameters of the panel. To circumvent this problem, we instead model the detectors such that the resolution on the recorded hits varies with distance from the centre of the panel, and the span of panels. This means that the uncertainty in the true muon position varies as a function of the detector parameters.

The realistic model for the detectors may still be used, when updates to the parameters are not required, e.g. when validating the detector configuration.

Inference. When imaging a passive volume, many muons will be used. While each muon acts independently, it is convenient to group muons together and perform their propagation and inference in parallel for computational efficiency.

In our case, full inference will be a two stage process: first we will use the muons to construct a rough 3D image of the volume by approximating the density of each voxel in the volume. This initial stage is rather task-agnostic; the next stage of inference takes this image and uses a task-specific algorithm to map the density predictions to final predictions, which in our case is a single number between zero and one representing the probability that there is uranium somewhere in the volume.

The first stage of inference involves fitting trajectories to the incoming and outgoing hits, considering the uncertainties in each hit as a weight in the fit; thus each trajectory is differentiable with respect to the detector parameters. The changes in trajectory may be used to compute a prediction of the density of material in which the muon scattered. The position of this scattering may be predicted using, e.g. the “point of closest approach” method [25], which extrapolates the incoming and outgoing trajectories inside the passive volume and assumes that a single scattering occurred at the point closest to both trajectories (they are not guaranteed to intersect). With a sufficient number of muons, this method can be used to build up an estimate of the 3D distribution of the density in the volume, although it will be highly biased due to the assumption of a single point of scattering per muon.

For the second stage of inference, we effectively need to convert a 3D tensor of floats to a single float. Here, one may choose to use e.g. a three-dimensional (3-D) CNN to classify the volume, but classical approaches are also possible. In our case, we can expect that, although the predicted densities will not correspond to the true densities of the materials in the volume, their distribution should contain several peaks, according to the number of materials present: either two (air and scrap metal); or three (air, scrap metal, and uranium). Since air and scrap metal will always be present, we actually only need to quantify the presence of a high-density peak, which can be done by considering the difference between the mean of the m highest-predicted densities, and the mean of the $n - m$ lowest predicted densities. If this difference is large, then it indicates the presence of uranium. After a suitable rescaling and sigmoid normalization, we arrive at our required single-float prediction for the whole volume.

Optimization. An estimate of the value of the loss function at the current detector-parameter points can be computed by predicting and inferring many passive volumes. Through the use of automatic differentiation, the partial derivatives of the loss can be computed with respect to each parameter. Using the standard gradient update rule, the detectors can be improved by making one step of length γ in the direction of steepest descent:

$$\theta_{t+1} = \theta_t - \gamma \nabla \mathcal{L}\left(y, \hat{y}\left(\theta_t\right)\right). \tag{2}$$

This is the most basic optimization loop: depending on the tasks and approaches used, however, it may be beneficial to augment the optimization, or to run ML models as part of the pipeline. TomOpt enables such possibilities through the use of a stateful callback system, which allows classes to interject during the optimization loop and have full read/write access to all aspects of the fit.

3.5. Status and prospects

We have released TomOpt as an open-source package, along with documentation [26]. Nevertheless, the package is under continuous development as we add new functionalities and use cases. A first accompanying publication is published and introduces the package from a more technical perspective, and demonstrates its application to an industrial example (ladle furnace) [22]; the second is in preparation and focusses more on the estimation of the muon momentum via regression [27], and three additional ones are in preparation, progressing in several aspects of the problem.

4. Progress in calorimetry optimization

Calorimetry is often the crucial part of a particle detector. By relying on destructive interaction of energetic particles with thick layers of matter, and the production of showers of secondaries, these devices are relatively simple in their functioning, yet the conversion of their output signals into physics measurements is made very complex by the stochasticity of the involved processes. Calorimeters have marked the history of particle physics in the past decades, and their continuous improvement has been a significant driver of new discoveries—it suffices to mention the detection of Higgs boson decays from measured photon pairs by the CMS and ATLAS Collaborations. Calorimeters are modular detectors made of an array of "towers", that is, individual elements that collect the energy of the incoming particles. The size of the calorimetric towers defines the spatial resolution of the calorimeter (each tower corresponds to one energy measurement): the spatial density of towers (the number of towers per area unit) is referred to as "segmentation".

Whereas the main task of both hadronic and electromagnetic calorimeters has been for a long time the one of detecting the collective energy yielded by all secondary particles produced in their interior, with relatively little emphasis on retaining or extracting precise position information about the energy depositions, these instruments started to be designed differently in recent times, when several physics-driven requirements (sticking with high-energy physics examples, it is unavoidable to mention here the separation of single photons from background-produced photon pairs in the case of the search for the Higgs boson, as well as the reconstruction of hadronic decays of boosted heavy objects in high-energy searches for new physics) have brought us to increase the transversal (i.e. along the direction perpendicular to that of the colliding beams) as well as the longitudinal (i.e. along the direction of the colliding beams) segmentation of the active detection components. Consequently, the asymmetric nature of the development of particle showers naturally raises the question of what is the optimal arrangement and segmentation of calorimeter cells. Furthermore, new technologies enable the recording of energy deposition timestamps with sufficient accuracy to serve as a fourth dimension for study and optimization. Together, these new capabilities also pose new questions on the possibility to actually exploit the difference in how different hadrons interact in dense media, with a view to extract particle identity information and further improve the particle-flow-based holistic reconstruction of complex hadronic showers typical of the big LHC experiments. In this section we consider a few use cases of relevance to the above program.

4.1. Optimization of the LHCb calorimeter for the LHC phase 2 upgrade

Optimization of a calorimeter refers to the development of a new or the modernization of an existing one. In the case of an existing calorimeter, fine-tuning also involves certain constraints, due to the reuse of already existing components, which may instead be considered free parameters when studying a new development. Examples of fine-tuning are the particular technology, geometry, and configuration of the calorimeter. Typically, *ab initio* ML approaches show that this fine-tuning can be avoided and a comparable reconstruction quality as in classical methods can be achieved [28]. When developing a model for the reconstruction of a real detector, both the classical and ML-based approaches require some preprocessing of the data to obtain geometry-agnostic inputs to the model.

As a first example of calorimeter optimization, we hereby consider the possible upgrade of the LHCb electromagnetic calorimeter (ECAL) [29]. The current ECAL is based on Shashlik-type modules with transverse dimensions of 12 × 12 cm^2 [30,31]. 3312 such modules are arranged in a rectangular wall perpendicular to the LHC beam axis and are laterally segmented into 1, 4, or 9 cells. In the upgraded calorimeter, some of the existing modules might be replaced by more granular modules that employ the SpaCal technology [32]. The response of a wall-like calorimeter in terms of reconstructed, analysis-level quantities can be represented as an image of an electromagnetic shower, where the value of each "pixel" corresponds to the energy deposited in the corresponding cell of the calorimeter. The lateral size of an electromagnetic shower can be determined by parameterization using the Molière radius specified by the technology. In such an approach, it is possible to choose in advance the size of the considered area (*window*) so that the specified fraction of all calorimeter clusters is contained within it, and the cell with the highest energy (*seed*) is located around the centre of the window. In addition, by increasing the window size, one can apply algorithms that also estimate pile-up contributions, since in this case one can compare signal clusters and clusters from background contributions. Most ML-based reconstruction algorithms in this approach are limited by the fixed dimensionality of the input array of energy deposits. And if the window size is large enough, e.g. 5 × 5 cells, when scanning the entire calorimeter, we will find that many areas of the calorimeter will be inaccessible if we require that all cells in the window contain (homogeneous) information. The calorimeter is therefore divided into three regions with three different granularities.

We consider different cases of deviations from the strict geometrical regularity of calorimeter cells arrangements. In the first case, we will consider the boundary between calorimeter regions with different cell granularity. In this case, it may be effective to divide large cells into smaller ones by interpolation. In the second case, the irregularity may arise due to the technological necessity to rotate the modules slightly around one or two axes. In this case, it is useful to use the coordinates of the cells as additional information at the input of the regressor. In a third case, due to engineering reasons, it is possible to skip a row or column of cells: one can then also interpolate the information in the missing cells (shown in Fig. 5). Finally, the irregularity at the borders of the calorimeter can be handled by padding. By using such data pre-processing algorithms, we can significantly streamline the reconstruction model architecture.

Various interpolation techniques and DL models were compared to restore the missing rows or columns within the cell matrix. The results obtained were averaged over the specific location of the missing row or column and are presented in Table 2. The best performing model, in terms of both peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), was found to

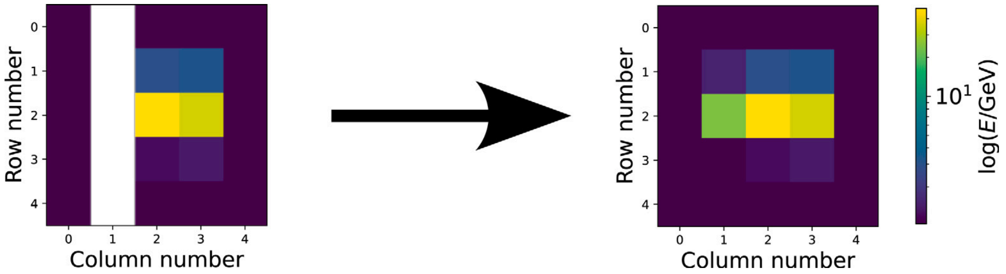


Fig. 5. An example of a calorimetric cluster with missing column #1 of cells (left) and a recovered cluster (right). The color represents $\log(E/\mathrm{GeV})$ for each cell. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
A comparison of classical interpolation methods and DL interpolation for the reconstruction of calorimetric clusters with missing information. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) metrics are used to measure the quality of cluster reconstruction. The results are averaged over the position of the missing row or column of cells. The best results are displayed in bold.

Model		PSNR↑	SSIM↑
Interpolation	Nearest-neighbor	85.3	0.74
	Cubic	91.2	0.75
	Linear	92.8	0.79
Deep Learning	Fully-Connected	**96.9**	**0.94**

be a fully connected neural network consisting of two linear layers activated by the rectified linear unit (ReLU). We consider such pre-processing as part of a geometry-agnostic reconstruction model. Such a model can be automatically trained on both simplified data sets designed for preliminary evaluation of calorimeter performance and data sets derived from detailed simulations. The use of automatic training ensures consistency and uniformity of the reconstruction results.

4.2. The challenge of beam-induced backgrounds in the electromagnetic calorimeter for a muon collider detector

The construction of a detector to study high-energy muon-muon collision poses significant challenges, many of which are entirely novel. One such case is provided by a very significant background due to in-flight decays of beam particles in the vicinity of the collision point. The detector is thus expected to be showered with a huge flux of low-energy photons and neutrons resulting from the interaction of decay products with structures around the collision area. This has been preliminarily taken into account at the machine-detector interface phase by introducing a tungsten nozzle [33], which screens a big portion of the radiation, leaving almost exclusively the background coming from the area around the interaction point. The chosen design is the one from Crilin [34]: a dodecahedron, every edge of which is made of 5 layers of arrays of PbF2 cells - each equipped with silicon photomultipliers. The modular structure of this design lends itself quite naturally to geometrical optimization studies, and therefore is chosen as reference for our work.

This “Beam-Induced Background” (BIB) invites a revisitation of the construction paradigms of the detector components. In particular, the BIB is expected to significantly affect the detection of photon and electron-induced showers in the electromagnetic calorimeter. Fig. 6 shows the deposition of a Geant4-simulated [6] BIB event inside the calorimeter at a centre-of-mass energy of 3TeV. The considerable amount of energy still left within the detector makes this a significant background that cannot be neglected. Furthermore, its asymmetric deposit distribution suggests that a uniform layout of the calorimeter cells would result in significant loss of performance with respect to a design that optimally adapted to the BIB flux geometry.

While our studies of this particular application of calorimeter optimization are still in an initial phase, we mention it here for its special interest and for the potentially large impact that a full optimization may have. The work plan includes the following tasks:

- Start with the choice of active material (PbF2 is presently suggested) and geometry of the initial design for the central electromagnetic calorimeter;
- develop a continuous model of the BIB energy deposition as a function of depth in the calorimeter material and position with respect to beam axis and detector centre. This must rely on a Geant4 simulation of the BIB particles as a function of energy and angle of incidence in the detector material;
- construct a function that returns the expected energy deposit in a given volume;
- create a model of the photon clustering and energy reconstruction;
- consider the three-dimensional size $\Delta x, \Delta y, \Delta z$ of calorimeter cells as parameters of a continuous model of the detector volume;

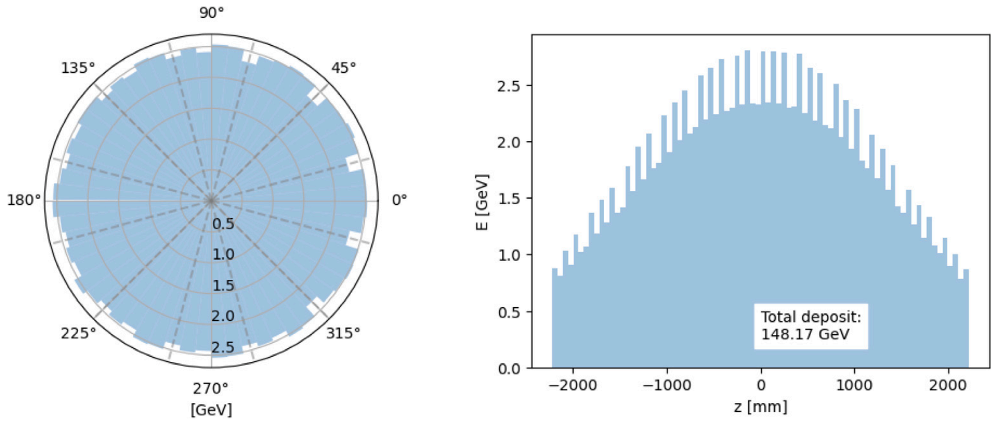


Fig. 6. BIB deposition in PbF2 <ECal cells: Left: radial distribution with respect to beam axis (here defined as z), the dashed lines mark the calorimeter edges. Right: Distribution along beam axis.

- generate real photon signals overlaid with BIB energy deposits using the parameterized model, reconstruct the signals, extract suitable metrics of utility;
- modify the size of calorimeter cells by following the gradient of the utility function, and iterate to convergence.

This procedure must, at the bare minimum, be complemented with the comparison of performance attained by a state-of-the-art photon reconstruction algorithm, given initial and final parameters. Higher accuracy can be sought for by producing maps of fast versus complete reconstruction performance.

The work is in progress: a working overview of these design optimization efforts can be consulted in Refs. [35,36].

4.3. Optimizing irregular geometries

An optimal calorimeter design, or future detectors in general, does not necessarily follow two dimensional grid structure, or structures that can be transformed trivially to fit into regular grids. Furthermore, for an optimization of the detector from the material choices, the geometry, and down to the physics output, the reconstruction algorithms need to be capable of processing and utilizing the correlations across the detector beyond regions of interest defined by seeds. On the other hand, the raw detector data is typically sparse, with only a small fraction of cells being active in each event. This is particularly true for hadronic showers in highly granular calorimeters or tracking devices.

The solution to these problems is two-fold: the data representation needs to be more generic, and the algorithms need to be capable of processing such data—conceptually, but also in terms of resource requirements. Currently, the most convenient way to represent the data is as a generic point cloud, with each point corresponding to a detector signal above threshold that can carry additional features such as the position, the shape, the deposited energy or even a full signal pulse shape. To process the point-cloud data, graph neural networks [37,38] are well suited, as they do not enforce a particular sorting of the points, yet provide information exchange across all points that can be used to infer the properties of the particles that the detector hits originated from. On the other hand, the choices are restricted by the resource constraints imposed by the hardware and the need to evaluate such models many times in a detector optimization task. For these resource reasons, fully connected graphs, where connections scale with the number of points squared, are not feasible in the context of a whole detector optimization. Therefore, graph neural network variants that are capable of learning the graph topology while keeping within resource constraints even for $\mathcal{O}(10^5)$ inputs are of particular interest [39,40].

Not only the input dimensionality poses challenges, but also the fact that the large amount of detector cells originated from an unknown number of sparsely or densely distributed particles that entered the calorimeter. In absence of regular grids, well-defined outer edges of physics objects, and the possibility of defining meaningful bounding boxes in the physical space for particle reconstruction, classic object detection techniques from computer vision are not applicable. Instead, the object condensation formalism [41] is being employed for a growing set of reconstruction tasks based on graph neural networks [42–45].

Based on these techniques, significant progress has been made in the past years, from first studies with $\mathcal{O}(10^4)$ inputs in a simplified environment [40], to the reconstruction of multiple particles in toy calorimeters with $\mathcal{O}(5 \times 10^4)$ inputs, as well as the CMS HGCAL [44], to point cloud sizes up to $\mathcal{O}(2 \times 10^5)$ [43] within seconds, while maintaining promising physics performance, often outperforming classic approaches.

These algorithms can provide the basis for a calorimeter optimization beyond grid-like structures, and open the possibility of a differentiable generic reconstruction algorithm for a full detector, allowing to jointly optimize different subsystems. The second ingredient for such an optimization is finding differentiable surrogates for point cloud generation, which is a challenging task. Exploratory studies are currently ongoing and are reaching higher complexity at each iteration [46–49]. So far, these algorithms are not capable of simulating point clouds of the size that reconstruction algorithms can process, but the progress is promising: soon, studies on optimizing more complex, not necessarily grid-structured detector designs with high granularity will become possible. Furthermore, even for grid-structured geometries, these reconstruction and generation algorithms can make very high granularity computationally feasible by exploiting the sparsity of the data.

4.4. Optimization of the CMS high granularity calorimeter

A case study is proposed for the optimization of the readout optical fibre plant of a high granularity calorimeter with over 6M channels. The High Granularity Calorimeter of CMS (HGCAL) is being designed for Phase II of the LHC and will cover the forward region of 1.5–3.0 in pseudo-rapidity [50]. Given the dense pileup environment foreseen in the forward region, resulting from up to 200 simultaneous proton–proton collisions and the high number of channels measuring energy and time of particle showers, it is expected that a large event size, of the order of 4-6 MB is produced after each bunch crossing. The readout is expected to occur at a rate of 750 kHz. A balanced throughput in the optical fibres as well as efficient aggregation of multiple fibres in bundles is required to maximize the usage of resources of the back-end electronics. Mechanical constraints are unavoidable in the routing of fibres, with break-points foreseen at the edge of each layer or the detector, where a re-arrangement (splicing) can occur. In this contribution we have detailed these constraints and summarized the results obtained by a simple scan of the phase space. This simple approach allowed us to reduce the so-called dark fibre (unused fibre) presence resulting in a significant decrease of the cost. The implementation of a ML algorithm poses however challenges due to the discrete nature of the problem. The example was presented as it could be a good use case for many other detectors with optical path continuum and discrete distribution of fibre patch panels.

4.5. Studies of granular calorimetry for future collider experiments

High-granularity in hadron calorimeters offers a significant increase in the performance of the information extraction procedures from particle interactions with dense media [51]. At particle colliders, its benefits stem from a two-pronged revolution that took place at the beginning of this century. The first prong is constituted by boosted jet tagging, which was developed when it was recognized that the signal of hadronically-decaying heavy particles (W, Z, and H bosons, top quarks, and other massive particles decaying to hadrons which may be hypothesized in new physics models) could be successfully extracted from backgrounds if sub-jets could be identified within wide jet cones. The second prong is constituted by the success of particle flow techniques, which were instrumental to e.g. increase the energy resolution of hadronic jets at the CMS experiment above the non-state-of-the-art baseline performance of its hadron calorimeter. Both boosted jet tagging and particle flow reconstruction rely on accessing fine-grained information on the structure of hadron showers.

Other observations of the benefit of high granularity for future HEP endeavours include a recent demonstration [52] that fine-grained hadron calorimeters allow the measurement of the energy of multi-TeV muons from the pattern of radiative deposits (to a 20% relative resolution that does not degrade with muon energy), offering itself as an obvious substitute to magnetic bending, which becomes impractical above a few TeV. If we want to preserve the discovery potential of energetic muons, this becomes an important aspect in the design of detectors at, e.g. a future circular collider for protons at energies above that of the LHC.

Jointly with, and independently from, the open hardware question of how far can granularity be pushed with existing or future available technologies, there remains an open question of how useful it can be, from an information extraction standpoint, to arbitrarily increase it. In principle, fine-granularity calorimeters offer more information than what we currently exploit them for. The nuclear interactions that a proton, a pion, or a kaon withstand when they traverse dense matter are different in cross section as well as in outcome, and this difference corresponds to information which until today we have never even attempted to extract. DL algorithms may today allow it, if only in probabilistic terms that are still going to be strongly useful for particle flow reconstruction and AI-based pattern recognition. The question to be investigated is whether this information extraction is feasible.

One of the first issues to address is to provide a specific quantification of the following question: What are the ultimate particle identification capabilities of an arbitrarily granular hadron calorimeter? In more quantitative terms, this question can be turned into the quantification of two sets of numbers. The first set is composed by the highest achievable Bayes factors Q of hypothesis testing for discriminating long-lived hadrons — e.g. protons from charged kaons, protons from pions, pions from kaons — assuming no lower limit on the size Δx of individually readout cells. The second set of numbers then determines the feasibility of such an instrument, through the physical size of cells above which any meaningful information (practically useful Q values) is lost (see Fig. 7). A meaningful determination of these quantities requires the deployment of state-of-the-art DL models and very careful studies, and possibly also an extrapolation to future capabilities of those models, as what is of specific interest is precisely the technical limit of such discrimination performance. Of course, the answers will depend on the details of the chosen technology for signal detection and active material of the calorimeter, which adds an additional dimension and further interest to this study.

It should be noted that despite the intrinsic interest of determining the quantities discussed above, they are only intermediate proxies to any final goal of a measuring instrument, and thus only a preliminary step (albeit a quite informative one at that) in the optimization task of a calorimeter. They would be crucial inputs to define the range of parameter values for the design of an instrument in any specific application.

Combined with the above research questions, one should also consider how much additional information gain is available by exploiting timing information in this high-granularity setup, and what are the potential performances on strange-quark tagging in a future collider by combining the time-of-flight discrimination with the spatial one. Preliminary findings have been described in

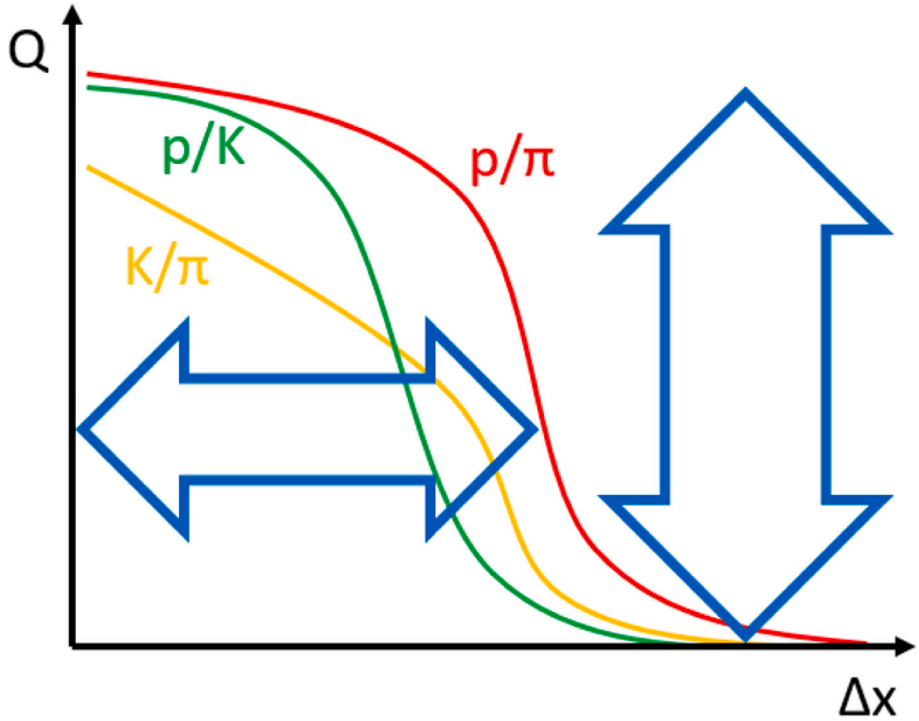


Fig. 7. Schematic view of possible discrimination power Q for pairs of hadrons as a function of the calorimeter cell size Δx. The figure illustrates the meaning of two relevant quantities (maximum discrimination power (1) and volume of cells above which usable information is lost (2)) that can be extracted by a study of the discrimination power of different hadron species (protons, positive pions, positive kaons) attainable by optimal use of the observable energy deposits in the cells of a granular calorimeter. Both the horizontal and vertical scale have arbitrary units. Of interest are both the value of Q attainable with arbitrarily high granularity (vertical arrow), and the value of Δx above which no significant discrimination can be extracted from data (horizontal arrow).

Ref. [53], where it was demonstrated that adding timing information to the CALICE detectors significantly improved performance. Fig. 8 illustrates the improved performance in calorimetry energy resolution obtained in these tests.

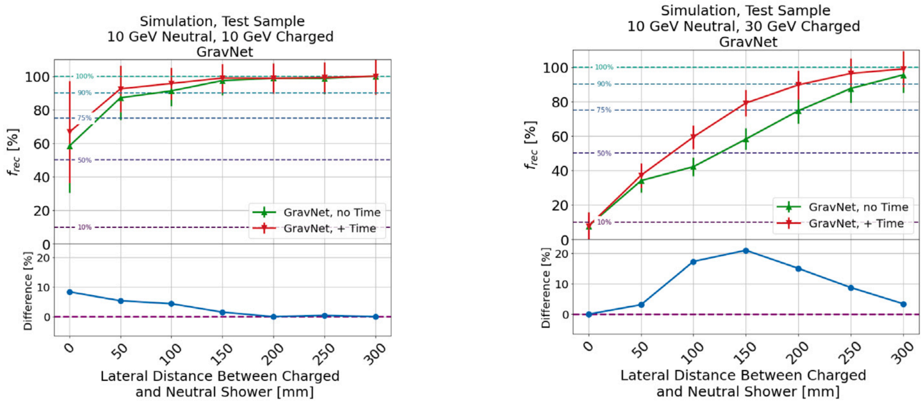
If we find out that there is valuable information in the primary particle identity, which can be mined in the fine-structure of the development of hadronic showers, and that this results in a corresponding significant gain for future instruments, the maximum granularity that is capable of retaining it becomes the gauge with which to measure whether our technology can be exploited for the task. Timing information can then be studied as the necessary complement, given recent developments in ultra-high time resolution.

The studies needed to satisfactorily address the above questions require, in a first phase, the deployment of large DL models trained and tested on large simulated datasets; and the prototyping and test-beam operation of a small-scale demonstrator if the simulations demonstrate potential for hadron ID separation for cell sizes that are — or that may potentially become in the near future — technologically feasible.

A first study of the readout for highly granular calorimeters, showing shower reconstruction capabilities, has been recently released [54].

4.5.1. Hybridization of tracking and calorimetry

A long-standing paradigm in detector design for HEP can be summarized by the motto "track first, destroy later". With no exception, particle tracking has been reliant on low-material-density to avoid as much as possible the degrading effect of nuclear interactions; and conversely, calorimetry has exploited dense materials for efficient energy conversion in limited volumes. However, the advent of DL questions the validity of that paradigm, as today's neural networks can make sense of the complex patterns resulting from nuclear interactions. It appears therefore highly desirable to investigate the possibility to trade off some of the undeniable benefits of light-weight tracking (in terms of resolution and low background) for a better reconstruction of the identity and development of hadronic jets. The abovementioned hypothesized possibility to discriminate the identity of different particles based on their behavior in traversing matter invites a study of what are the performance gains and losses of a combination of a state-of-the-art tracker followed by a fine grained calorimeter, when the density of the former and the latter do not abruptly change at their interface, but rather vary with continuity from the first to the second. Beyond the possibility of particle identification, a hybridization of tracking and calorimeter brings in a natural impedence matching with the state-of-the-art of particle flow reconstruction, in the sense that it potentially provides the algorithms with a larger and more coherent amount of information about the behavior of individual particles and their interaction history within showers.



(a) WCD design with three PMTs.

(b) WCD design with four PMTs.

Fig. 8. The calorimeter resolution as a function of the distance between showers of charged (Q) and neutral (N) hadrons in the CALICE calorimeter in mm for the test sample of simulation described in Ref. [53], for E_Q = 10 GeV (left) and 30 GeV (right). In both cases, E_N = 10 GeV. Models trained with (red lines) and without (green lines) time information are compared, showing the significant performance improvement achieved by models that account for the time information during training. For each plot, s with the blue lines indicate the additional fraction of showers, in percent, reconstructed by the model trained with time than without. Figure reproduced from Ref. [53] respecting the terms of the CC-by license. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The study of the above subject is again reliant on the deployment of highly specialized DL models, and in fact it requires to extrapolate to the future capabilities of these algorithms to the time when such a detector could become operational. It could be articulated as follows: (1) Study ultimate performances, on specific high-level benchmarks (e.g., precision of the extraction of a H→bb signal in specific FCC or Muon-collider setups) of a idealized state-of-the-art tracker plus calorimeter (e.g. starting with an existing design, such as the CMS central detector), with a developed DL reconstruction. (2) Consider increasingly hybrid scenarios when the outermost layers of the tracker are progressively embedded in the calorimeter, gauging the performance on low-level primitives (single-particle momentum resolution, fake rates) and high-level objectives. Eventually, such a study should inform the one described above in (1), to converge on a design of a future instrument capable of optimally exploiting the enhanced information extraction potential. (3) End-to-end optimization: the studies included in the above project will inform a full modeling by differentiable programming of the whole chain of procedures, from data collection to inference extraction, which allows to directly connect the final utility function of an experiment with its design layout and technology choices, such that the navigation of the pipeline by stochastic gradient descent may allow a full realignment of design goals and implementation details.

A first study of the hybridization of tracking and calorimetry for a highly granular calorimeter has been recently released [55].

5. Progress in accelerator applications

Particle accelerators are a critical part of enabling discoveries in high-energy physics. The goal of accelerator science is to advance our understanding of fundamental beam physics and particle accelerators while developing novel methods and tools to aid in the operation of current beam facilities and the development of future ones.

5.1. Six-dimensional beam distribution studies

High energy physics applications often require beam distributions in a 6-D position-momentum phase space (x, p_x, y, p_y, z, p_z) that are tailor-made for individual applications. For example, beams must be compressed longitudinally and flattened transversely to improve collider luminosity [56], controlled transversely to mitigate beam losses in high-intensity accelerators [57], or shaped longitudinally to mitigate emittance growth due to coherent synchrotron radiation [58] and improve the performance of novel acceleration techniques [59–62]. Manipulating beam distributions at a fine level represents a paradigm shift from traditional beam dynamics control goals, which only seek to control high-level beam properties. This level of control requires diagnostic techniques that provide a similarly detailed reconstruction of the beam distribution in the 6-D phase space, far beyond traditional [63] or more recent [64] approaches that infer only scalar properties of the beam.

Enabling practical detailed reconstructions of 6-D phase space distributions requires novel techniques that reduce diagnostic and numerical complexities associated with current methods. For example, pinholes [65], slits (combined with longitudinal phase space manipulations) [66], mesh grids [67] or laser wires [68] have been used to provide high-dimensional (> 2-D) information about the beam distribution. However, these techniques require specialized diagnostic elements and/or a large number of measurements to produce high-resolution reconstructions of the beam distribution. On the other hand, tomographic manipulations of the beam

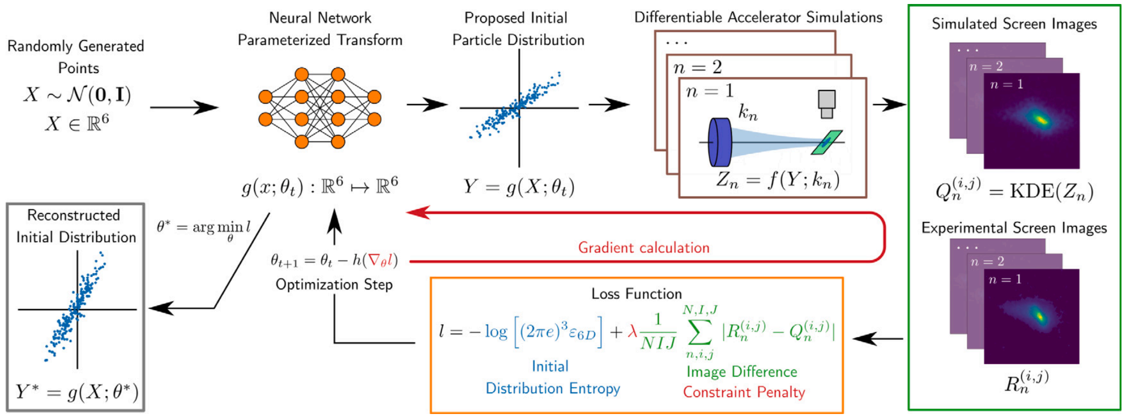

Randomly Generated Points
$X \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$X \in \mathbb{R}^6$
Neural Network Parameterized Transform
$g(x;\theta_t) : \mathbb{R}^6 \mapsto \mathbb{R}^6$
Proposed Initial Particle Distribution
$Y = g(X;\theta_t)$
Differentiable Accelerator Simulations
. . .
$n = 2$
$n = 1$
k_n
$Z_n = f(Y;k_n)$
Simulated Screen Images
$n = 2$
$n = 1$
$Q_n^{(i,j)} = \mathrm{KDE}(Z_n)$
Experimental Screen Images
$n = 2$
$n = 1$
$R_n^{(i,j)}$
Gradient calculation
Reconstructed Initial Distribution
$Y^* = g(X;\theta^*)$
$\theta^* = \arg\min_\theta l$
$\theta_{t+1} = \theta_t - h(\nabla_\theta l)$
Optimization Step
Loss Function
$l = -\log\left[(2\pi e)^3 \varepsilon_{6D}\right] + \lambda \frac{1}{NIJ} \sum_{n,i,j}^{N,I,J} |R_n^{(i,j)} - Q_n^{(i,j)}|$
Initial Distribution Entropy
Image Difference Constraint Penalty

Fig. 9. Description of the approach for reconstructing phase space beam distributions using differentiable accelerator physics simulations. *Source:* Reproduced from Ref. [74].

distribution in phase space, combined with commonly available detailed measurements of 2-D transverse beam distributions at diagnostic screens, have also been used to reconstruct high dimensional phase space distributions [69–71]. Unfortunately, these tomographic reconstruction techniques incur in significant computational costs when trying to infer high dimensional distributions from 2-D projections. Finally, while ML techniques have been implemented to reconstruct phase space distributions from experimental data [72,73], they demand significant initial investment to be effective, including the generation of large training data sets from simulation or experiment and the training of ML models. These limitations of detailed, high-dimensional measurement techniques hinder their practical usage, restricting our comprehension and control of beam dynamics within accelerators.

We have developed a novel method to provide detailed reconstructions of the beam phase space using simple and widely-available accelerator elements and diagnostics. To achieve this, we take advantage of recent developments in ML to introduce two new concepts (shown in Fig. 9): a method for parameterizing arbitrary beam distributions in 6-D phase space, and a differentiable particle tracking simulation that allows us to learn the beam distribution from arbitrary downstream accelerator measurements. This method extracts detailed 4-D phase space distributions from measurements in simulation and experiment, using a simple diagnostic beamline, containing a single quadrupole, drift and diagnostic screen to image the transverse (x, y) beam distribution. To obtain the figure, first, randomly generated points drawn from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ are transformed via a neural network, parameterized by θ_t, into a proposed initial distribution. This distribution is then transported through a differentiable accelerator simulation of the tomographic beamline. The quadrupole is scanned to produce a series of images on the screen, both in simulation and on the operating accelerator. The images produced both from the simulation $Q_n^{(i,j)}$ and the accelerator $R_n^{(i,j)}$ are then compared with a custom loss function, which attempts to maximize the entropy of the proposal distribution, constrained on accurately reproducing experimental measurements. Neural network parameters $\theta_t \rightarrow \theta_{t+1}$ are then iteratively tuned via gradient descent in order to minimize the loss function.

We demonstrated our algorithm using a synthetic example, where we attempt to determine the distribution of a 10-MeV beam given a predefined structure in 6-D phase space. The propagation of a synthetic beam distribution through a simple diagnostic beamline containing a 10 cm long quadrupole followed by a 1.0 m drift is simulated using a custom implementation of the beam dynamics code Bmad [75]. To illustrate the capabilities of our technique, the synthetic beam contains multiple higher-order moments between each phase space coordinates. To simulate an experimental measurement, we simulate particles traveling through the diagnostic beamline while the quadrupole strength k is scanned over N points. The final transverse distribution of the beam is measured at each quadrupole strength using a simulated $I \times J = 200 \times 200$ pixel screen, with a pixel resolution of 300μ m. The set of images, where the intensity of pixel (i, j) on the nth image is represented by $R_n^{(i,j)}$, is then collected with the corresponding quadrupole strengths to create the data set, which is then split into training and testing subsets by selecting every other sample as a test sample, resulting in 10 samples for each data subset.

The reconstruction algorithm begins with generating arbitrary initial beam distributions (referred to here as proposal distributions) through a neural network transformation.

5.2. Six-dimensional beam compression with AD

A neural network, consisting of only two fully-connected layers of 20 neurons each, is used to transform samples drawn from a 6-D normal distribution centred at the origin to macro-particle coordinates in real 6-D phase space (where positional coordinates are given in meters and momentum coordinates are in radians for transverse momenta). As a result, the coordinates of particles in the proposal distribution are fully parameterized by the neural network parameter set θ_t.

Fitting neural network parameters to experimental measurements is done by minimizing a fully differentiable loss function to determine the most likely initial beam distribution, subject to the constraint that it reproduces experimental measurements; this is

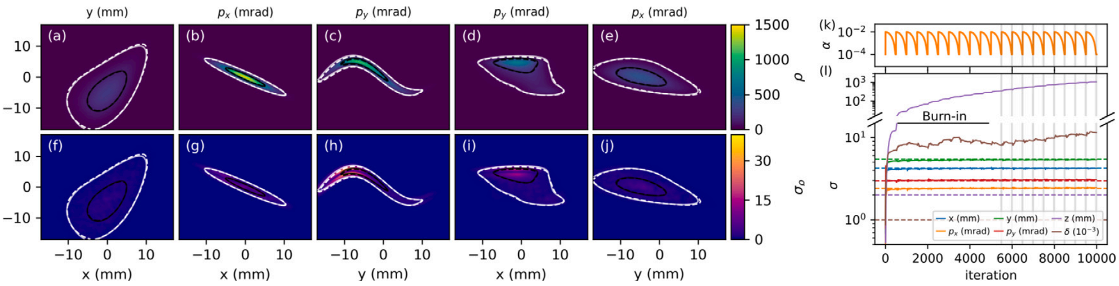


Fig. 10. Left: Comparisons between the synthetic and reconstructed beam probability distributions using our method. (a-e) Plots of the mean predicted phase space density projections in 4-D transverse phase space. Contours that denote the 50th (black) and 95th (white) percentiles of the synthetic ground truth (dashed) and reconstructed (solid) distributions. (f-j) Plots of the predicted phase space density uncertainty. Right: Evolution of the proposal distribution during training on synthetic data. (k) Learning rate schedule for snapshot ensembling. (l) Second-order moments of beam reconstruction during training for each phase space coordinate. Dashed lines denote ground truth values. Vertical lines denote snapshot locations after burn-in period. Reproduced from [74]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

similar to the MENT algorithm [76]. The likelihood of an initial beam distribution in phase space is maximized by in turn maximizing the distribution entropy, which is proportional to the logarithm of the 6-D beam emittance ε_{6D} [77]. Thus, we specify a loss function that minimizes the negative entropy of the proposal beam distribution, penalized by the degree to which the proposal distribution reproduces measurements of the transverse beam distribution at the screen location. To evaluate the penalty for a given proposal distribution, we track the proposal distribution through a batch of accelerator simulations that mimic experimental conditions to generate a set of simulated images $Q_n^{(i,j)}$, which we then compare with experimental measurements. The total loss function is given by:

$$l = -\log\left[(2\pi e)^3 \varepsilon_{6D}\right] + \lambda \frac{1}{NIJ} \sum_{n,i,j}^{N,I,J} |R_n^{(i,j)} - Q_n^{(i,j)}|\,, \tag{3}$$

where λ scales the distribution loss penalty function relative to the entropy term and is chosen empirically based on the resolution of the images.

Results from our reconstruction of the initial beam phase space using synthetic images are shown in Fig. 10. We characterize the uncertainty of our reconstruction using snapshot ensembling [78]. During model training, we cycle the learning rate of gradient descent in a periodic fashion: this encourages the optimizer to explore multiple possible solutions (if they exist). After several of these cycles (known as a “burn-in” period), we save model parameters at each minimum of the learning rate cycle, as shown in Fig. 10(a). We then weight predictions from each model equally, using them to predict a mean initial beam density distribution Fig. 10(a-e) with associated confidence intervals Fig. 10(f-j). Performing this analysis by tracking 10^5 particles for each image took less than 30 s per ensemble sample using a professional grade GPU (< 60 ms per iteration, 500 steps per ensemble sample).

We see excellent agreement between the average reconstructed and synthetic projections in both transverse correlated and uncorrelated phase spaces. Furthermore, the prediction uncertainty from ensembling is on the order of a few percent relative to the predicted mean, providing confidence that the overall solution found during optimization is unique. Finally, reconstructions of the beam distribution from image data predicts transverse phase space emittances that are closer to ground truth values than those predicted from traditional measurement techniques, i.e. second-order moment measurements of the transverse beam distribution. This results from non-linearities and cross-correlations present in the 4-D transverse phase space distribution.

It is instructive to examine the evolution of the proposal distribution during model training. In Fig. 10(l) we examine second-order scalar metrics of the proposal distribution after each training iteration for each phase space coordinate. The entropy term in Eq. (3) causes the distribution to expand in 6-D phase space until constrained by experimental evidence. Phase space components that have the strongest impact on beam transport through the beamline as a function of quadrupole strength converge quickly to the true values, whereas the ones that have little-to-no impact (e.g. the longitudinal distribution characteristics) continue to grow. In other cases, there is weak coupling between the experimental measurements and beam properties; for example, chromatic focusing effects due to the energy spread σ_δ of the beam weakly affect the measured images. Here, the reconstruction can only provide an upper-bound estimate of the energy spread, since small changes in transverse beam propagation due to chromatic aberrations are overshadowed by statistically dominated particle motion. Accelerator physics imported the term “chromatic aberration” from optics. In optics, chromatic aberration refers to a lens deflecting light with different wavelengths by a different amount; in accelerator physics, it refers to quadrupoles (the “lenses” of beam control) bending particles with a slightly different energy by a different amount. Convergence of the proposal distribution thus provides a useful indicator of which phase space components can be reliably reconstructed from arbitrary sets of measurements.

Recent work [79] studied the problem of optimizing simultaneously two quantities of interest (heat load and trip rates) in the control of a continuous electron beam at the Continuous Electron Beam Accelerator Facility. The work compared solutions obtained with Bayesian optimization, with model-free reinforcement learning, and with gradient-based deep reinforcement learning optimization of a differentiable model. The work demonstrated that gradient-based optimization (DDRL) converges faster than the other methods on smaller scale (in terms of dimension of the parameter space) problems, as illustrated in Fig. 11, and that for larger

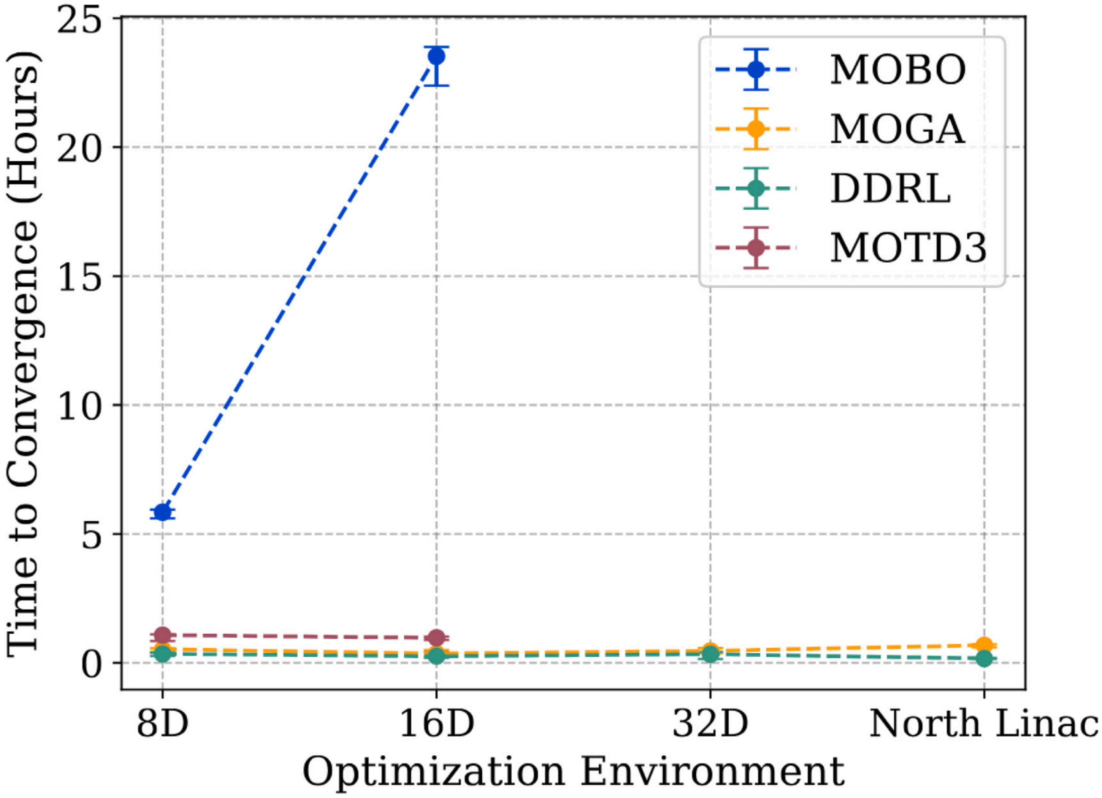


Fig. 11. Time taken to converge in the joint optimization of heat load and trip rates for a continuous electron beam at the Continuous Electron Beam Accelerator Facility, for different algorithms. Each scatter dot represent median and the error bars cover two standard deviation confidence bounds over 16 trials. MOBO refers to Multi-Objective Bayesian Optimization; MOGA refers to Multi-Objective Genetic Algorithms; MOTD3 refers to Conditional Multi-Objective Twin-Delayed Deep Deterministic Policy Gradient (a model-free reinforcement learning policy); DDRL refers to Conditional Multi-Objective Deep Differentiable Reinforcement Learning (a gradient-based differentiable reinforcement learning policy). Reproduced from Ref. [79].

scale problems it is the only feasible method. Although this is a known property of gradient-based optimization, which, because of its mathematical formulation, is naturally designed to efficiently solve the empirical risk minimization problem [80,81], this work further empirically confirms this behavior when applied to the field of beam dynamics and accelerator control.

The development of tools and techniques that leverage differentiable accelerator physics modeling has the potential to revolutionize key aspects of how experimental data is interpreted in the field of accelerator physics. Differentiable beam dynamics modeling directly addresses fundamental limitations facing traditional analysis approaches in accelerator physics. By enabling the combination of physical laws with the flexibility of ML-based function representations and detailed experimental measurements of accelerator beams, it can provide an unparalleled understanding of detailed beam dynamics inside accelerators.

6. Progress in the optimization of experiments for astrophysics research

6.1. Dark matter direct detection experiments

Dark Matter (DM) particles pervading our Galactic halo could be directly detected by measuring their scattering in a suitable detector. The rare and small expected signal requires ultra-low background conditions and low energy detection thresholds. After summarizing the features of this possible DM signal and briefly describing the experimental efforts to detect it, we will outline the application to DM direct searches of ML techniques: these allow the discrimination of the expected signal from radioactive backgrounds or other noise events, being typically more effective than conventional filtering protocols. This capability is very important, taking into account the increasing demands to lower backgrounds and thresholds in future experiments.

6.1.1. Dark matter signals

The presence of DM is required to explain an important fraction of the energy-mass budget of the Universe following different cosmological and astrophysical observations, although its nature is unknown [82]. Thermal Weakly Interacting Massive Particles (WIMPs) are a type of DM candidates that are supposed to have been produced in the early Universe via a freeze-out mechanism when Standard Model (SM) and DM particles were in thermal equilibrium, producing a constant relic density, reproduced for a wide range of masses from 1 $\mathrm{eV/c^2}$ to 120 $\mathrm{TeV/c^2}$.

Different complementary strategies are being attempted for WIMP detection [83]. DM candidates could be produced at colliders and indirectly detected by identifying an excess of SM particles like gamma-rays, neutrinos, positrons or antiprotons produced by the annihilation of DM particles. In the direct detection of DM in the Galactic halo, the goal is to register the elastic scattering of WIMPs off target nuclei or electrons in a detector [84]. Taking into account the expected signal from this interaction, the direct detection of DM is really challenging: the interaction has an extremely low probability, and large exposures and low background conditions (operating deep underground to suppress the effect of cosmic rays) are mandatory; the signal is concentrated at very low energies (below tens of keV), which requires the use of low energy threshold detectors; and the signal has a continuum energy spectrum, which would appear entangled with background, therefore distinctive signatures would be helpful to assign a DM origin to a possible observation.

Direct detection experiments can be focused on different physics cases [84]; many of them just look for an excess of events over the known backgrounds, considering different ranges of candidate masses, Nuclear or Electronic Recoils (NR/ER) or different types of interactions between the dark matter particles and the nuclei (Spin-Independent, SI or Spin-Dependent, SD). Other experiments search specifically for distinctive DM signatures, like the annual modulation in the interaction rate or the directionality. Several physics cases are described in the following.

- There are particular requirements to probe DM candidates with masses at sub-GeV/c^2 scale: lighter targets must be used to keep kinematic matching between WIMPs and nuclei, lower threshold are necessary to detect smaller signals, and new search channels (absorption or scattering off by electrons, ER) are being considered, as light WIMPs cannot transfer sufficient momentum to generate detectable NR. Following the proposed Migdal effect,[5] the DM-nucleus interaction could lead to excitation or ionization of the recoiling atom, being for low mass DM this additional signal above threshold (unlike the NR alone) and then enhancing sensitivity [85]; for this reason, this effect is already being considered by many collaborations to release results exploring sub-GeV masses.
- The movement of the Earth around the Sun makes the relative velocity between detectors and DM particles in the Galactic halo oscillate in time, which produces a modulation in the expected DM interaction rate with defined features like a one-year period; this signature would allow identifying a possible DM signal [86]. The DAMA/LIBRA experiment at the Laboratori Nazionali del Gran Sasso in Italy is observing for more than 20 years an annual modulation in the measured rate compatible with DM [87]; this modulation signal has not been confirmed nor refuted at high confidence level by other experiments.
- The average direction of DM particles through the solar system comes from the constellation of Cygnus, as the Sun is moving around the Galactic centre; the measured track direction of NR could be therefore used to distinguish a DM signal from background events (expected to be uniformly distributed) and to prove the Galactic origin of a possible signal [88]. The main difficulty is to reconstruct the very short tracks (~1 mm in gas, $\sim 0.1\ \mu\mathrm{m}$ in solids) expected for keV scale NRs [89].

6.1.2. Detection

Many different and complementary technologies are being applied or under consideration in experiments attempting the direct detection of DM, e.g. solid-state cryogenic detectors, time projection chambers based on noble liquids, scintillating crystals, and purely ionization detectors using semiconductors or gaseous targets; detectors measure the heat, light or charge produced or a combination of two of them in hybrid detectors. A discussion of advantages and disadvantages for each technique and relevant results obtained in the field can be found at [84]. The properties of the DM candidates are constrained under different scenarios for the interaction.

For high mass DM, experiments using large liquid noble detectors (Xe and Ar) provide now the best limits on cross-sections for SI DM-nucleus interaction and will explore regions of cross-sections where solar and atmospheric neutrinos become an irreducible background with projects using even larger detectors starting at the end of the decade [90]; for SD DM-proton interaction, bubble chambers provide the best limits.

For low mass DM, the best sensitivity comes from a combination of experiments based on different detection techniques: solid-state cryogenic detectors (using scintillating bolometers or small mass Ge and Si semiconductor crystals), purely ionization detectors (Ge diodes, CCDs or gaseous detectors) and liquid noble detectors when operated in low threshold mode; candidates with increasingly lower masses could be investigated thanks to the development of novel technologies to further reduce the energy threshold [91].

It is also worth noting that important results from NaI(Tl) experiments [92,93] to solve the long-standing conundrum of the DAMA/LIBRA annual modulation result have been presented and that studies for a DM detector with directional sensitivity are underway to prove the Galactic origin of a possible signal [94,95].

6.1.3. Machine-learning techniques

In the context of DM direct detection experiments, ML techniques are being applied mainly to improve the discrimination of the expected signal from radioactive background or other type of noise events. A few examples of application of these techniques in this context are highlighted here.

[5] Atomic physics effect that leads to the emission of a bound-state electron from atomic or molecular systems when the atomic nucleus is suddenly perturbed. It has been observed for radioactive decays; there is no evidence for NR yet, although attempts are in progress.

- The sensitivity for low-mass DM searches of detectors of the EDELWEISS experiment (made of germanium cryogenic bolometers and operated in the Modane Underground Laboratory in France) has been studied in Ref. [96]. Using a data-driven background model, frequentist and multivariate analysis approaches (profile likelihood and boosted decision tree) are used to compute exclusion limits on cross-sections.
- The LUX experiment used a dual-phase Xe TPC in the SURF laboratory in the US; backgrounds from the wire grid electrodes near the top and bottom of the active target were found to be particularly pernicious, limiting the sensitivity to low-mass DM. A ML technique based on ionization pulse shapes to specifically identify and reject these background events was developed and sensitivity improved [97]. Moreover, results combining ML with the profile likelihood fit procedure using LUX data have been presented in Ref. [98] as a fast and flexible analysis of DM data; it is considered that this technique can be exploited by future DM experiments to make use of additional information and reduce computational resources needed for signal searches and simulations.
- A new type of analysis for the DRIFT-IId directional DM detector (operated in the Boulby laboratory in UK) using a Random Forest classifier has been presented in Ref. [99]. Events are labeled as signal or background based on a series of selection parameters, rather than solely applying hard thresholds, allowing an increased efficiency at lower energies and a projected sensitivity enhancement of even one order of magnitude for some WIMP masses.
- For experiments using NaI(Tl) scintillators to search for a possible annual modulation in the interaction rate, raw data below $\sim$10 keV are fully dominated by non-scintillation events mainly related to the photomultipliers (PMTs) coupled to crystals; multiparametric thresholds on parameters deduced from the pulse shapes of the PMT signals (like first momentum or ratios of different fractions of the areas) are mandatory for careful event selection. But an unavoidable leakage of noise in the lowest energy bins can happen. Boosted Decision Trees (BDT) have been developed to improve the rejection of PMT-related noise with a multivariate analysis combining several weak discriminating variables into a single powerful discriminator. Decision trees have a binary structure with two classes, in this case associated to signal and noise, respectively. This approach has been followed by the COSINE-100 [100], SABRE [101,102] and ANAIS-112 [103] experiments, obtaining lower energy threshold or background levels just above the threshold. In particular, in the case of ANAIS-112 the use of ML techniques has allowed to improve the sensitivity to the DAMA/LIBRA signal, as reported in Ref. [104]; training populations independent of background data have been obtained from dedicated onsite neutron calibrations (producing genuine NR) for signal-like events and from data taken from a blank module (identical to the ANAIS-112 detectors but without a scintillating NaI(Tl) crystal) for noise-like events. The optimal thresholds on the BDT parameter (being -1 for noise and $+1$ for signal) have been defined for each energy bin and for each detector independently and the efficiencies for the selection of bulk scintillation events in the region of interest carefully estimated. Thanks to the improvement in background rejection in that region (a background level reduction of around 20% has been achieved between 1 and 2 keV) and the increase in detection efficiency with this analysis with respect to the previous filtering, the ANAIS-112 sensitivity to test the DAMA/LIBRA annual modulation result has been pushed, being possible to reach five standard deviations by extending the data taking a few more years than the accumulated five years in August 2022.

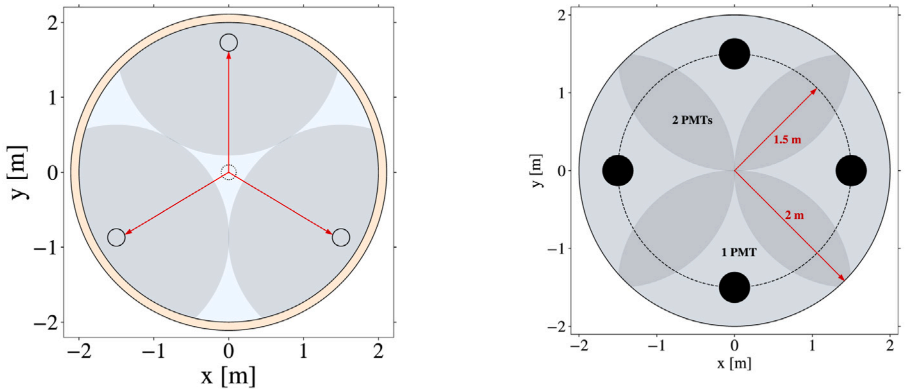
In summary, the direct detection of DM particles is really challenging due to the small and rare signal expected and is being attempted by complementary experiments based on different detection technologies and targets and exploring different interactions, mass ranges of candidate particles and possible signatures. Multivariate machine-learning techniques are being applied in this field mainly to improve discrimination capabilities between signal-like events and different types of background or noise events.

6.2. Muon identification and gamma/hadron discrimination using compact single-layered water Cherenkov detectors powered by ML techniques

6.2.1. Introduction: indirect detection of gamma rays with extensive air shower arrays

Very high-energy (VHE) gamma rays, ranging from 100 GeV to a few hundred TeV, can be used to investigate some of the most extreme non-thermal events taking place in the Universe, such as Active Galactic Nuclei (AGNs) and gamma-ray bursts (GRBs) [105–107]. Their indirect detection with a wide field-of-view (FoV) and a high-duty cycle is possible using Extensive Air Shower (EAS) arrays placed at high altitudes [108]. Observatories of this kind use dense arrays of detectors to observe the EAS of particles produced when gamma rays interact with the Earth's atmosphere. However, distinguishing the EAS induced by gamma rays from those by the vast cosmic-ray background is a challenge. Water Cherenkov detectors (WCDs) have proven to be effective for these purposes, as they allow having a duty cycle near to 100% and detect muons, which are much more prone to appear in hadronic showers. Currently, despite the potential for mapping of large-scale emissions as the Fermi bubbles in the centre of our galaxy [109,110], no such wide FoV experiment is operating in the Southern Hemisphere. Such an observatory would be complementary to the major facility CTA-South [111] and enable full-sky coverage for transient and variable multi-wavelength and multi-messenger phenomena at this energy range.

In this work, it is discussed the use of an EAS array composed of single-layered WCDs with multiple photomultipliers tubes (PMTs) for such purpose. By analyzing the signals from the PMTs with a CNN, we aim to identify muons in the stations, and thereby discriminate between gamma-induced and hadron-induced showers in subsequent analyses.



(a) WCD design with three PMTs.

(b) WCD design with four PMTs.

Fig. 12. Scheme of the single-layered WCDs. A dark circle with a 1.5 m radius around each of the eight-inch PMTs was drawn. If a muon travels vertically through one of these circles, the direct Cherenkov light should be picked up by the corresponding PMT.

6.2.2. *WCD concepts*

The rationale behind the design of the WCDs used in this work is that muons will cross the whole detector, creating a direct Cherenkov light that reaches only a portion of the WCD floor, while photons and electrons will produce electromagnetic showers inside the station, creating a broader Cherenkov light pool. This asymmetry in the signal provides a mean to distinguish muons from other particles. The Cherenkov angle of a relativistic muon with an energy of ∼2 GeV in water is approximately 41°. Thus, to ensure complete signal coverage and a maximal signal asymmetry caused by vertical muons, the WCDs have a base diameter of 4 m and a water height of 1.7 m (see Fig. 12).

The number of PMTs, and thus the dimension of the detector, can be optimized to balance the physics performance and the cost of the detector. This design evolved from four (Fig. 12(b)) to three PMTs (Fig. 12(a)), as the distance of PMTs to the centre is adjusted to optimize the detector while minimizing the overlap of the areas covered PMTs, as well as the number of PMTs used. Both station concepts use eight-inches PMTs and white diffusive walls to maximize signal collection, which is essential to lower the energy threshold of the experiment.

This approach offers a cost-effective alternative to other methods that may increase the cost of the project. For example, the HAWC experiment [112] uses WCDs with a large volume of water and black walls to identify muons. Alternatively, other dedicated muon detectors can be buried or shielded as it is being done in the LHAASO experiment [113] or placed below the WCDs as in the MARTA project [114].

6.2.3. *Simulation framework and sets*

The data sets used in this study were created using a simulation framework that combines the use of CORSIKA (v7.5600) [115] for simulating extensive air showers and the Geant4 toolkit (v4.10.05.p01) [6–8] for simulating the detector's response.

The array configuration uses a dense layout, comprising 5 720 WCDs, covering an area of about $\sim 80\,000$ m^2 with a fill factor of 85%. The experiment observation level was set at 5 200m above sea level, which corresponds to the altitude of the ALMA site in Chile.

The gamma-induced showers were generated with energies E_0 in the range 1–1.6 TeV and a zenith angle of $\theta_0 = 10°$, while the proton-induced showers were simulated with E_0 in the range 0.6–6 TeV and a zenith angle in the range 5–15°. For both primary particles, the showers were generated following a E_0^{-1} spectra (the events are afterwards weighted to ensure a realistic power-law spectrum of energies), and the azimuth angle was uniformly distributed. For each detector concept, more than 3 000 events were generated for gamma-induced showers, while for proton primaries more than 17 000 showers were simulated.

Finally, a threshold on the total measured WCD signal at the ground is introduced to emulate a typical energy reconstruction, keeping events whose total signal at the ground is within one sigma around the mean of the gamma events.

6.2.4. *Single station performance*

The objective of this study is to determine the probability, $P_\mu^{(i)} \in [0, 1]$, that a muon has passed through a given WCD with index i based on the signal it has recorded. To achieve this, we use variables derived from the WCD signal time traces, as previously proposed in other studies [116,117]. These variables aim to explore both temporal (patterns in the signal time traces) and spatial (asymmetry in the PMTs' integrals) features. The used variables include:

- normalized signal time trace of each PMT;
- integral of each PMTs signal time trace;

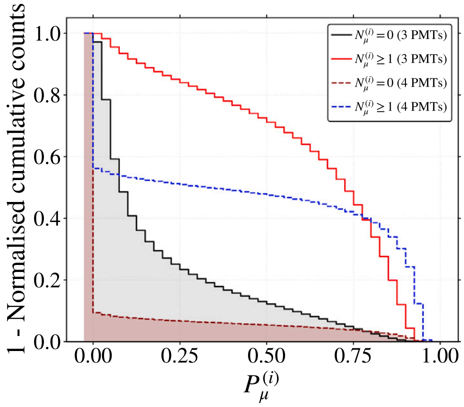


Fig. 13. Inverse cumulative function of the $P_\mu^{(i)}$ variable for stations with and without muons using both detector concepts: three PMTs (full lines) and four PMTs (dashed lines).

- sum of the PMTs' signal trace integrals;
- normalized integral of each PMTs signal time trace.

The PMT signal is normalized to the sum of the signals recorded by all PMTs during the time window considered for the variable that will be normalized. Note that the normalized signal traces contain the first 30 nanoseconds to explore features from both direct Cherenkov light and the reflections, while the rest of the variables considered only the direct Cherenkov light (first 10 nanoseconds). Given the typical signal for vertical muons, only stations with more than 200 p.e. for the station with three PMTs and 300 p.e. for the station with four PMTs are considered.

A CNN was used to extract complex features from the signal time traces of the PMTs and combine them with the spatial features (integrals). With it, we define a regression problem to provide the probability $P_\mu^{(i)}$ that a muon has passed through the station. The algorithm's configuration was optimized using a validation data set for vertical showers. The network contains three convolutional layers to study the signal time traces and use ReLU activation function. The input of the first convolutional layer is set as one channel of data for each normalized PMT signal time trace. Afterwards, three dense layers are introduced to perform the regression using the previous signal features and the spatial variables. The model was trained with the Adam optimizer [118] during 200 epochs with a batch size of 512 and a learning rate of 10^{-3}. Python 3.7 and Keras [119] were used as the framework of the entire study. A discussion of the CNN configuration and the use of various other approaches to this problem can be found in Ref. [116].

In Fig. 13, the normalized inverse cumulative function for the probability $P_\mu^{(i)}$ is shown for both station concepts using proton-induced showers. The results were separated into stations with and without muons, being possible in the first case to have stations with both muons and electromagnetic contamination. Most stations with muons were correctly identified by requiring a probability $P_\mu^{(i)} \geq 0.5$. Roughly 70 and 50% of the events passed the threshold for stations with three PMTs and four PMTs, respectively. Besides that, a low false positive rate was found for stations without muons, with approximately 15 and 10% of them having a probability $P_\mu^{(i)} < 0.5$ for stations with three PMTs and four PMTs, respectively. A similar muon tagging capability can be achieved for inclined showers with an incident angle $\theta_0 \sim 30°$, as long as the CNN is retrained for such events [120].

6.2.5. Gamma/hadron discrimination

In this section, we describe a gamma/hadron discrimination strategy that relies on the use of the muon information extracted in the previous section at the single detector level, $P_\mu^{(i)}$. This information is combined at the shower event level to create a simple and intuitive observable, $P^\alpha_{\gamma h}$, which is defined as the sum of the probabilities $P_\mu^{(i)}$ to the power α as proposed in a previous study [121]. Hence, we introduce the following discriminator quantity:

$$P^\alpha_{\gamma h} = \sum_{i=1}^{N_S} P_\mu^{(i)\alpha}(r_i \ \geq 40\text{ m})\,, \tag{4}$$

where N_S denotes the total number of WCD stations passing the threshold on the signal in the shower event. In addition to the threshold on the station signal, only stations far away from the shower core are sampled, to avoid the bulk electromagnetic shower component, which is much higher near the shower core.

As described in Ref. [121], the introduction of the α power in the discriminator is motivated to reduce the relative importance of the large number of stations without muons in gamma and hadron high-energy showers ($E_0 \geq 40$ TeV). In this work, since the relative importance of the number of stations without muons is not as high at ~ 1 TeV energies, an α value of 1 is used.

This observable was tested for proton and gamma-induced showers with an equivalent total signal at the ground. It was possible to efficiently separate the gamma-induced showers from the ones generated by protons by adequately choosing a threshold on $P^\alpha_{\gamma h}$.

A proxy to a gamma-ray experiment flux sensitivity can be obtained by evaluating $S/\sqrt{B}$, where S and B are the selection efficiency for gamma rays and the background (protons induced events), respectively. As shown in Fig. 14, a $S/\sqrt{B} \sim 4$ was found

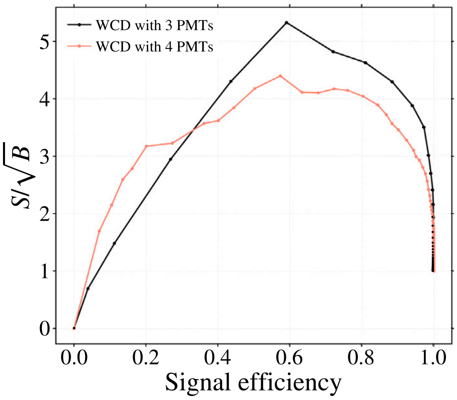


Fig. 14. Ratio of the selection efficiency of the discriminator $P^{\alpha}_{\gamma h}$ for gammas (S) over the square root of the selection efficiency for protons (B) as a function of S. The red line corresponds to the result obtained with four PMTs [120], while the black line corresponds to the results obtained with three PMTs [122]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for both detector concepts when fixing the selection efficiency for gammas to $S = 0.8$. The obtained value is similar to the one quoted in other experiments such as HAWC [112] and LATTES [110], but using significantly smaller WCDs than HAWC.

6.2.6. Discussion on the performance of muon tagging with PMTs

In this section, we have shown that a WCD with a reduced water volume of $12\,\mathrm{m}^2 \times 1.7\,\mathrm{m}$ and multiple PMTs placed at the bottom can be used to efficiently tag muons. A ML-based analysis was developed to process the PMT-acquired signals. The results showed that the identification of muons in stations, given as a probability, $P^{(i)}_{\mu} \in [0; 1]$, can be used to build a simple gamma/hadron discrimination observable able to distinguish between gamma and proton-induced showers with a $S/\sqrt{B} \sim 4$ for shower energies of $E_0 \sim 1$ TeV and $\theta_0 \sim 10°$ and a similar total signal at the ground.

6.3. End-to-end optimization of the SWGO array layout

The Southern Wide-Field Gamma Observatory (SWGO) is planned to be built at high altitude in south America (possible sites being considered for its construction are in Peru, Chile, and Argentina). The observatory will consist of several thousand Cherenkov detectors in the guise of PMT-endowed water tanks, deployed in an array spanning a footprint of up to a few square kilometers; alternatives include the deployment of photodetectors in excavated pools or in submerged bladders on the surface of a lake. The scientific interest of such a detector, which will be sensitive to primary cosmic gamma rays in the multi-GeV to multi-PeV range, lays in its capability of studying a number of sources in the southern sky, with continuous operation, excellent pointing and energy resolution, and very high discrimination of background from hadronic primaries.

At the time of writing, a number of different configurations for the precise design of the water tanks are being considered. They share the need to count both the collective number and energy of electrons, positrons and photons produced in the electromagnetic shower development, as well as the number and energy of muons, which are very rare in gamma ray-originated showers and thus constitute a powerful discriminator of hadron backgrounds. Due to the higher penetration power of muons, Cherenkov tanks may be able to distinguish these particles by the shape of the light signals collected by properly placed photomultiplier tubes on the bottom of the tank, or by having PMTs pointing both upwards and downwards in the middle of the tank. Optimization studies of these designs are ongoing, with the aim of finding the right compromise between cost, logistics, and performance of these units.

A separate issue concerns the arrangement of the tanks on the ground. The footprint of energetic air showers may span several hundred meters or even kilometers at the height above sea level of SWGO, and the capability of the array to appropriately measure the energy of the primary particle in the low-energy range (tens-hundreds of GeV) requires the collection of as large a fraction of the collective signal of particles on the ground as possible; this mandates that tanks be laid out as tightly packed as possible—i.e. with a “fill factor” (in terms of fraction of ground covered by the active area of the detectors) of 50% or more. On the other hand, in order to be sensitive to the very low flux of photons of extremely high energy (1 PeV and above), the total instrumented area (including the void between units) needs to be maximized with detectors distributed with a low fill factor, well spaced from one another. However, too low fill factors will hinder both a precise energy reconstruction and a successful separation of hadron backgrounds, whose rate at those high energies exceeds that of gamma rays by a factor exceeding 10^4.

The above contrasting requirements have led the SWGO collaboration to propose as benchmarks a set of different layouts that share as a common feature a dense core and an extended region more sparsely populated. Investigations are ongoing to evaluate in quantitative terms the relative benefit of the proposed layouts for the various astrophysics cases of interest to the collaboration. The more principled question however concerns the investigation of the way more high-dimensional space of all possible configurations of N detectors. As a detector position is specified by its ground coordinates x, y, the space of configurations lives in $\mathbb{R}^{2N-3}$, accounting for the azimuthal symmetry of the problem. Therefore, e.g. the layout of three units involves the choice of three real parameters:

by setting the first unit in (0,0), the second can be set at (x_2,0) without loss of generality; the third can then be specified by (x_3, y_3), when y_3 can be chosen in the positive semi-axis without loss of generality. With over 6000 units to be deployed, the space of possibilities is thus prohibitively large to be probed with discrete methods.

The problem of constructing an optimization pipeline that scans such a large parameter space with differentiable programming can be addressed only if we endow ourselves with a parameterized approximation of the density of particles on the ground as a function of the distance from the shower axis, resulting from a cosmic ray shower of given energy E and polar angle θ. Ideally this should be available in closed form for both gamma and proton-originated showers (the background to discriminate against), and for the particle types to which the detectors are differently sensitive: electrons, positrons, and photons (which can be in first approximation be considered together, due to their similar behavior in interacting with water at energies much above the critical energy E_{crit}, which is of 76.2 MeV for positrons and 78.3 MeV for electrons [123]), and muons. As atmospheric showers arising from high-energy primaries involve a number of complex stochastic processes, precise parameterizations are difficult to produce; yet even an imprecise choice may be able to successfully inform a continuous scan of the parameter space, if a full simulation is then used to validate the results in the vicinity of interesting configuration points.

We concentrated our initial efforts on the high-energy part of the spectrum of photons that SWGO wants to study. For this, we used large CORSIKA [115] simulations of gamma and proton showers at different energies (from 10 TeV to 10 PeV) to extract a model of the density of secondary particles on the ground as a function of the distance from shower axis. These were the basis of an optimization pipeline, which we now describe.

1. The starting point is the generation of an initial ground configuration of N detector units, $(x_i, y_i)_{i=1,\dots,N}$.
2. A set of gamma and proton showers are simulated with an intersection of the shower axis with the ground, at a position x_0, y_0 randomly chosen within a region suitably exceeding on all sides the footprint chosen for the detector tanks, such that showers at the edge of this region have a negligible probability of being detected by the array. Given shower energy E and polar angle θ, the position of the shower centre on the ground, together with the azimuthal angle ϕ_0 defined as the angle between the shower axis and the positive x direction, determines an estimate of the average number of particles of different kinds that will hit the sensitive area of each of the N detector units. These numbers can be sampled from Poisson distributions of means equal to the estimated averages.
3. Through a likelihood maximization, estimates of the shower parameters are obtained by considering the true value of the parameters of the model, and the number of particles detected for each species in each tank under, for the time being, the assumption that tanks have 100% detection efficiency and perfect discrimination power between the various particle species. The likelihood is maximized under both hypotheses (gamma or proton) for each shower, regardless of the true primary particle species.
4. The obtained maximum likelihood values, maximized over shower parameters E, θ, ϕ, X_0, and Y_0 are used to compute a likelihood ratio test statistic T for each generated shower. The sampling of a large number of showers of different parameters allows to construct a PDF of T for both hypotheses.
5. A new batch of gamma and proton showers is generated, and a distribution of T values obtained. This is fit as the sum of the two distributions, and the uncertainty of the fraction of gammas in the batch is extracted as the Rao-Cramer-Frechet bound.
6. Using the uncertainty on the gamma fraction, a utility function is computed as the weighted sum, for different energy points, of the inverse of the relative uncertainty in the flux of gammas in the batch. An additional piece of the utility can be added to account for the energy and angle resolution of the measured photon showers.
7. A propagation of derivatives through the whole pipeline allows to extract the derivatives of the utility function over displacements δx, δy for each of the N detector units. These are used to update the detector positions, by using adaptive learning rates as multipliers in the position updates.
8. The cycle can be continued by generating a new set of gamma and proton showers and deriving new PDF of T, then fitting a batch of showers and recomputing the utility function U.

The pipeline described above allows to converge to layouts that optimize a simple utility function focusing on the precision of the gamma flux. Of course, more complex utility functions could be conceived, to model the real objectives of the wide scientific program of the SWGO collaboration. This modeling task has not yet been attempted. In the near future the code will be expanded to improve the way it represent the full inference-extraction procedures, with a view to providing useful input to the collaboration on the most advantageous layouts of tanks on the ground. The relative gains of those layouts over baseline ones can then be appraised by exploiting full simulation of atmospheric showers and detection of secondaries by the detector units.

As an example of the performance gains that a full optimization of the SWGO array may provide, we show in Fig. 15 a comparison of the utility measured for the 13 benchmarks with the utility of an optimized layout composed of 300 “macro-tanks”, each made up of 19 tanks packed together. The macro-tanks allow for a quicker scan of the parameter space, which reduces from 10397 to 597 dimensions. A further reduction is operated by coupling together in triplets the macro-tanks to form concentric equilateral triangles on the ground. Each triangle may rotate or scale in dimension, but the three macro-tanks at its vertices retain the triangular symmetry around the centre of the array. This allows to reduce by a further factor of three the dimensionality of the problem, to 199, and retains some generality, as it does not violate (if not by the discreteness of the number of tanks) the cylindrical symmetry of the problem. With 199 free parameters the optimization run, which considers a total of 2000 cycles over batches of 3000 showers, takes about one week of CPU to run on a 30-core cluster. The resulting optimization provides some 30 percent improvement in utility over the most performant of the 13 benchmarks, while using a 15% smaller number of detection units. More detail on this work is available in Ref. [124]

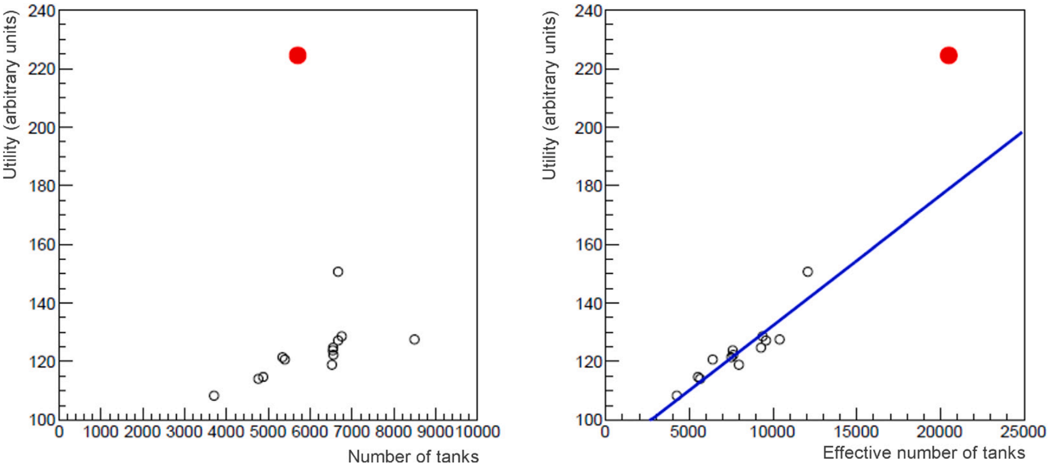


Fig. 15. Comparison of the utility function of an optimized layout of SWGO (red point) to that of 13 benchmarks proposed by the collaboration (black circles). On the left the utility is shown as a function of the number of detection units; on the right, a correction of the abscissa value is operated to account for the effective area covered by the arrays, such that the effect of the area is removed from the utility. A 30% improvement of the optimized solution is thus seen to be due to the internal arrangement of the units. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7. Progress in neuromorphic computing optimization

Neuromorphic computing (NC) is an emerging computing paradigm that exploits brain-inspired, time-encoded signal processing to allow for highly energy-efficient, highly parallelizable, and decentralized information processing including fast inference of data-driven models [125,126]. In addition to other novel models of computation, such as quantum computing (which also holds the promise of making differentiable programming at scale possible), NC promises a major path for beyond-Moore computing and artificial intelligence. Unlike quantum computing, NC relies on less exotic hardware and focuses on designing electronic systems, such as processors and memory devices, that are inspired by the structure and function of biological neural networks.

The primary motivation for this approach is to overcome the limitations of conventional computing technologies, particularly in cognitive tasks such as low-power perception and learning, which the human brain excels at. By leveraging biomimetic principles, NC systems offer the potential for real-time processing, low power consumption, and enhanced efficiency, paving the way for the next generation of intelligent machines and technologies that can potentially transcend Moore's law [127–129].

Neuromorphic hardware can run natively artificial neural networks where the perceptrons are replaced with neurons whose mathematical expression consists in a differential equation (typically the "integrate and fire" model) inspired by the structure of biological neurons: this "spiking" networks are then trained with gradient-descent-like algorithms powered by automatic differentiation. Spiking networks implemented in NC systems can therefore be considered a step of differentiable programming that goes towards low-power consumption and enhanced computational efficiency.

Below we provide a brief introduction and an overview of the status of NC and its applications at large, after which we discuss a few ideas of potential exploitation in fundamental science of the specific functionalities and prerogatives that NC offers.

7.1. Neuromorphic computing introduction

It is helpful to recap some basic neuroscience concepts [130] in order to understand how NC is different from conventional deep neural networks and related hardware accelerators. Biological neurons have ion pumps that maintain a Nernst potential of about −70 mV with respect to the exterior of the cell membrane. The time-varying potential depends on the flow of ions (Na^+, K^+, Cl^-, etc.) through channels in the cell membrane, which open and close depending on the electric field gradient and concentration of neurotransmitter molecules etc. Neurons have rich temporal dynamics, signaling patterns, and adaptation mechanisms at different temporal and spatial scales that mediate processing and storage of information. AD-powered differentiable models of neuronal dynamics are compared with data from electric fishes in Ref. [131]. Neuromorphic devices mimic these processes either at the physical level through mixed-signal electronic circuits, which for example emulate the ion diffusion processes in neurons with energy-efficient analog CMOS circuits operating in the subthreshold regime, or through custom digital systems offering flexible simulation capabilities. Due to the high variability of naturally evolved biological computation mechanisms, there is a remarkable diversity of approaches falling into the bucket of NC, such as:

- *Memristive Systems*: Memristive devices are a class of electronic components that exhibit memory and resistive switching characteristics. These properties enable them to mimic synaptic and neuronal dynamics, providing an alternative way to design neural networks and learning in neuromorphic systems [132].

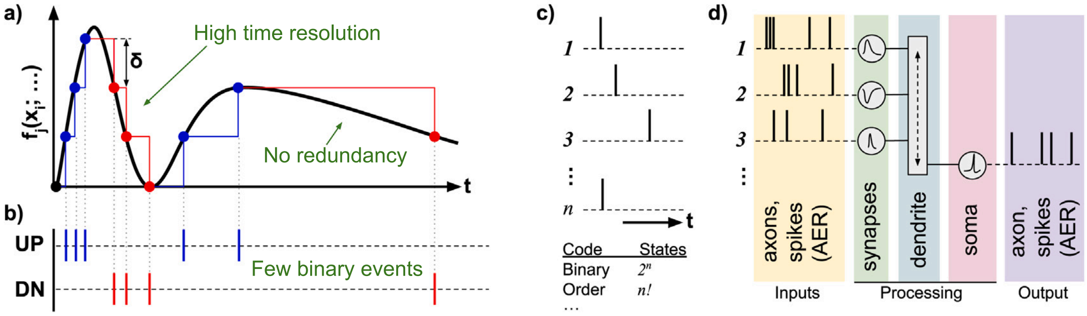


Fig. 16. Illustration of (a) Lebesgue sampling of a signal, x_i, transduced from the environment, (b) the resulting signal encoding as unary *spikes* on two channels representing changes of $+\delta$ (UP) and $-\delta$ (DN), (c) spike-based encoding of information across n channels representing axons, and (d) a schematic of a spiking neural network unit with input and output channels. In general, the performance of a system comprising such parts is subject to optimization of the sensors, $f_j(x_i, \ldots)$, sampling parameters such as δ and noise shielding mechanism, spiking neural network hyperparameters, and hardware realizations for the application at hand.

- *Dendritic Integration*: Neurons are observed to perform a variety of different forms of synaptic integration on their inputs, emphasizing the critical role of dendrites in the brain and for understanding how neurons learn and integrate the thousands of synaptic inputs. The central role of dendrites suggests that in the search for more energy-efficient NC solutions we should move beyond learning with synapses to learning with dendrites [133].
- *Recurrent Networks*: This approach includes large, randomly connected recurrent neural networks known as ‘reservoirs’ with nonlinear recurrent dynamics that act as a temporal kernel for the input signals, which can be further processed/classified using a trained readout network. This research domain includes early approaches such as Liquid State Machines (LSM). Recurrent networks are particularly relevant for tasks involving time series data and pattern recognition [134,135].
- *Brain-inspired Cognitive Architectures*: These entail developing microelectronic circuits, high-level modular architectures and algorithms that capture the cognitive features of the brain, such as memory, perception, learning, and problem-solving [136].
- *Efficient algorithms*: NC opens new algorithmic opportunities for efficient solutions to some conventional hard computational problems such as constrained optimization, graph algorithms, kernels for composition, and signal processing [137].

While each of these directions possesses unique strengths and limitations, they collectively contribute to the progress and diversity of NC research, offering a variety of paths towards emulating the computational efficiency and cognitive capacity of the brain. For reviews, see Refs. [126,129,137,138] and references therein. In the following we focus on some general aspects of NC architectures, which can influence the future of fundamental science computing systems and design optimization solutions.

7.2. Spiking neural networks

A Spiking Neural Network (SNN) aims to mimic the behavior of biological neurons and neural networks more closely than conventional artificial neural networks. In SNNs, neurons communicate using time-encoded events called *spikes*, similar to how neurons in our brain encode and transmit information using (mostly stereotype) electrical pulses in the form of action potentials. Spikes are asynchronous, unary “one-or-nothing” events where the physical timing of the spikes in relation to the dynamic states of neuron units and the environment encode information, as illustrated in Fig. 16. Time-based signal encoding is a key characteristic that differentiates SNNs from traditional artificial neural networks, making them more biologically plausible and in some cases more resource-efficient information processors. Instead of sampling and processing signal amplitudes at some time interval (the conventional Nyquist/Riemann paradigm) each sensor channel asynchronously generates one bit of information if and only if the corresponding signal feature has changed by some amount (Lebesgue sampling, Fig. 16a). This way input signals are transformed into patterns of precisely timed spikes, with the information being encoded in the relative timing of the spikes (Fig. 16b). Level-crossing analog to digital converters are basic sampling devices of this type. Spike-based representations of information allow SNNs to process and propagate information efficiently across the network: see Ref. [139] for an introduction. For example, in comparison with a synchronous binary encoding on n channels an asynchronous spike-ordering code can represent $\log_2(n!)$ bits of information given sufficient timing precision (Fig. 16c). Synapse conductances, dendritic processes and neuron voltages are modeled with time- and sometimes also space-dependent differential equations under the assumption that action potentials can be approximated with spikes that are characterized by the spike time (Fig. 16d). In neuromorphic systems, spikes are typically encoded and communicated using address event representation (AER) to improve efficiency, where the spike time is represented by physical time while only the source neuron address has an explicit binary representation. Given an SNN neuron connectivity table the AER spike representation includes the minimum binary information needed to determine the receiving neuron addresses.

There are no graded activation functions in SNNs (such as ReLU, sigmoid etc.), but the mean spiking frequency corresponds in a quasistationary sense to the activation of an artificial neural network (ANN) unit. Thus, ANNs resemble “mean-field models” in the sense that they model mean firing rates of neurons, hence the commonly used term “rate-based” neural networks, while SNNs model the time-dependence of neuron potentials (with varying spatial resolution through point-like, multi-compartment or finite-element models) under the assumption that action potentials can be represented with spikes. This spiking mechanism introduces an additional layer of complexity and dynamics to the network, allowing it to efficiently encode and process temporal information. SNNs are therefore well-suited for tasks requiring real-time processing and adaptation, such as sensory processing, motor control, and decision-making. As a result, they have become a major focus in the field of NC, offering the potential to unlock new paradigms in artificial intelligence and algorithm development.

Recent literature was dedicated to devising SNN learning policies that enable scaling up SNN models while keeping the energy cost relatively low and simultaneously accelerating inference speed. Ref. [140] proposes a SNN-based large language model with the additional features of: detecting channels with high activation potential (therefore important for the learning) and focussing on them; and a generalized integrate-and-fire model that reduces quantization errors in the quantization of outliers. The resulting model outperforms traditional ANN-based models and paves the way towards deeper studies on the scaling properties of SNN.

7.3. Neuromorphic computing optimization challenges

With DL computing needs doubling every few months the foundational optimization algorithms and digital technology drive the energy and resource requirements of artificial intelligence at an unsustainable rate [129,141], making this vital technology inaccessible for broad and safe societal benefit. The development of neuromorphic technologies provide hints to what may become the future of ubiquitous computing, optimization and artificial intelligence.

In principle it is possible to perform end-to-end optimization of SNNs using surrogate gradients for the non-differentiable parts [142]. However, training of SNNs using state-of-the-art back-propagation through time (BPTT) is costly and can lead to numerical instabilities if not handled carefully [143]. Furthermore, the task of systematically optimizing the design of a system, including sensors, spike based encoding, spiking/artificial neural network modules, and the cognitive architecture given specific actuators and application constraints is an open challenge. This is an opportunity for research and development in NC design optimization using differential and probabilistic programming. Progress in this direction can enable valuable advances in the co-design and optimization of the hardware–software–algorithm stack [138] and edge-to-cloud computing continuum [144]. Furthermore, considering the knowledge gaps and broad interdisciplinary NC efforts to understand and mimic information processing and optimization in the brain, bridging these communities can lead to radically new insights in resource-efficient optimization and design of experiments and systems in general.

Event-driven architectures can significantly reduce the average power of sensor systems that operate in random-sparse-event physical environments, for example in the form of an always-on wake-up circuit that triggers a conventional high-performance computing system. Note that this is also possible when the relevant frequencies in the environment are much higher than the spike rates and time scales in the SNN domain [145]. However, there are also challenges related to noise management and precision, see for example Refs. [146,147]. In the presence of noise, a level-crossing sampler of the type illustrated in Fig. 16a-b generates false events that waste energy and downstream processing performance. Different methods have been investigated to overcome this difficulty, such as introducing a hysteresis window around each amplitude level or a noise-shielding time window after each sample. When high amplitude precision is required a conventional uniform-sampling ADC is more efficient because the number of level-crossing samples tend to increase exponentially with resolution, while a level-crossing sampler can offer relatively high dynamic range and temporal resolution. Further work is required to enable a systematic approach to optimizing the design of systems incorporating both Nyquist/Riemann and Lebesgue sensor degrees of freedom, as well as hybrid SNN-ANN cognitive architecture and NC hardware co-design optimization.

7.4. Spiking neural networks for low-momentum tracks removal in CMS pixel layers

Silicon pixel detectors allow for precise measurements of charged particle tracks and vertices at collider experiments. Next-generation detectors will require a reduction in pixel size, leading to unprecedented data rates exceeding those foreseen at the HL-LHC. Signal processing that performs a physics-motivated filtering of the data within the pixelated region of the detector *at the collision rate* will enhance physics performance in a high luminosity environment.

The shape and time evolution of charge clusters deposited in an array of small pixels encodes information about the physical properties of the traversing particle. These can be extracted with locally customized neural networks implemented directly in the front-end electronics. A first demonstration takes the form of a filtering algorithm that rejects tracks with transverse momentum p_T below the threshold that is useful for physics analysis.

Spiking neural network models offer several advantages for this machine learning task: the time evolution of the charge cluster is naturally encoded in the input data, and the minimal number of parameters and low power requirements are well suited for implementation on an ASIC. The incoming charge waveforms can be converted to streams of binary-valued events which are processed by the SNN. Studies show that an SNN trained using an evolutionary algorithm and with optimized set of hyperparameters shows a signal efficiency more than 90% for particles with $p_T > 2$ GeV, using a factor of two fewer parameters than a similarly trained Deep Neural Network [148].

7.5. Integration of neuromorphic computing in online triggers at hadron collider experiments

The data rate yielded by high energy physics colliders such as the LHC is of the order of the tens of megahertzs. Online triggers systems are responsible for bringing the rate down to $\mathcal{O}(1\,\mathrm{kHz})$, a rate compatible with data storage systems. A recent study [149] has demonstrated that it is possible to reduce the trigger rate from 40 MHz to about 75 kHz trigger using neuromorphic chips when considering anomaly-detection-based triggers. In comparison to CPU and GPU benchmarks, neuromorphic chips offer throughputs larger by over a factor 20 for the largest networks that were tested.

These results open the road to further testing and to the deployment of “Level-1” trigger menus implemented in neuromorphic chips.

7.6. Integration of neuromorphic computing in fine-grained calorimeter readout

In parallel to the investigations of granular calorimetry discussed in Section 4, it appears interesting to study the possibility to integrate edge computing elements based on neuromorphic processors in the readout of the device. While the typical aim of NC are applications where extremely low power consumption and computing at the data-generating end are required, we speculate that the potential of localized preprocessing and extraction of information within the core of a highly granular calorimeter could be quite significant. A calorimeter made of millions of individual cells, whose complete readout would pose significant challenges, might strongly benefit from a localized, ultra-fast, and ultra-low-power pre-processing of the information at the highest spatial resolution, before higher-level primitives can be transferred to the back-end for reconstruction. In addition, the possible timing information that new generation detectors are starting to enable would find a perfect match with the time-encoded specific capabilities of timeconstant-modified NC hardware processors based on, e.g., state-of-the-art CMOS, nanowire memristor networks, and photonic and opto-electronic technologies.

The work plan in this case involves: (1) A demonstration of the unsupervised learning of specific patterns in macro-cells, via *in-situ* neuromorphic processing of full granularity information via emulation in digital processors; (2) Study of possible designs and hardware implementations that incorporate that functionality, and their effectiveness in terms of latency, information extraction, power consumption, and heat generation with respect to ordinary digital computing solutions; (3) Prototyping, if the solutions developed in (1) and (2) prove effective.

The lack of existing hardware solutions to the problem of increasing the processing speed of NC devices to a level suitable for fundamental science applications such as the ones we discussed in this section demands a wide-ranging parallel study of photonic and opto-electronic solutions for the emulation of neuronal decoding of signals which are natively encoded as light pulses in conceivable homogeneous calorimeter elements.

8. Progress in medical physics

In this Section we outline recent progress in medical physics made possible by AD/PD techniques.

8.1. Fast emulation of deposited dose distributions by means of graph neural networks

Cancer is a leading cause of death worldwide, accounting for nearly 20 million cases and 10 million deaths only in 2020. Among the various existing therapies, one of the most effective and frequently used treatments is external beam radiotherapy (RT). For every RT treatment, an essential part of the process that precedes the effective delivery of radiation is the treatment plan optimization. This phase involves choosing the therapeutic beam energies and fluences based on their orientation to fit the medically prescribed dose, with particular attention on the dose that should be received by the tumor and the maximum deliverable dose to organs at risk (OARs). Currently, this optimization is done using reliable but standard sequential algorithms that optimize energies and fluencies in two different steps. Often, such algorithms suffer the need of reaching a trade-off between the optimality of the solution and the computing time and in most cases, they are not suited to handle new complex tasks in reasonable times. Novel therapies, such as electron FLASH RT [150] or Volumetric Modulated Arc Therapy (VMAT) [151], offer much more freedom in the choice of the entry angles of the beam. The dose can be shot to the patient from a wide set of angles, reducing significantly collateral damage to healthy tissues, while delivering the right amount of dose to the tumor. However, this possibility entails a significant increase in complexity in the optimization process which can lead to a significant increase in computing time. In principle, the dose should be computed and optimized for all the possible orientations of the beam accelerator, with continuity.

In this context, DL algorithms, based on new tools such as tensorization, GPU acceleration, and AD, can represent a way to overcome such limitations. The aim of our study is to develop a DL model that can generate energy deposition distributions as a function of beam parameters and medium density in negligible time and with high precision. Constructing a differentiable *dose engine* — in this case, using a surrogate model — represents a critical first step in bringing differentiable programming paradigms to RT treatment plan optimization. Such a model could serve as the foundation for a new generation of Treatment Planning Systems (TPS) that use gradient-based optimizers, enabling the simultaneous optimization of all beam parameters as continuous degrees of freedom. At this stage, we developed a Variational AutoEncoder (VAE), based on Graph Neural Networks, that can generate deposited energy distributions inside a simplified voxelated material. Voxelization refers to discretizing a three-dimensional space into small cubes (voxels), obtaining a discrete-element representation of the material. In the following, we show a representative sample of the studies already published in Ref. [152], where the full description of the DL model and some ablation studies are presented. Moreover, we will show some unpublished results on the VAE latent space and sample generation.

8.1.1. Dataset

The dataset employed in this work was built running simulations using Geant4 [6–8], a toolkit in C++ for Monte Carlo simulations in particle physics. We simulated electron beams, made up of 10 000 primaries each, fired toward a voxelated material. Such material consists of a water cube, with a side length of 80 cm, containing a 3.5 cm thick slab of with variable density, perpendicular to the beam. For each run of the simulation, we sampled uniformly the particle energy between 50 and 100 MeV and the slab density between 0 and 5 $\mathrm{g\,cm^{-3}}$.

Energy deposition data are collected in a cylindrical scorer, aligned with the electron beam, of 50 cm height and 5 cm radius, divided in 28 × 28 × 28 voxels along the z, θ, and r axes. This choice brings two main advantages. First, without any loss of generalization, we are reducing the complexity of the generation task. We do not want the network to generate the deposited energy distribution in the whole volume, which may depend on beam orientation and, in real applications, on the organs' shapes and dimensions. We just consider a cylindrical volume around the beam, which still can be made large enough to contain all the energy released by the beam. The second advantage is that with a cylindrical shape, we gain more precision near the beamline, where more energy is deposited and a precise prediction is most important.

The dataset is made up of 6 239 examples of energy distributions and is divided into train, validation, and test set. The test set represents a relevant central subset in the parameter space. It only contains examples with both particle energies ranging between 70 and 80 MeV and slab densities ranging between 2 and 3 $\mathrm{g\,cm^{-3}}$, which account for the 4% of total data samples. These examples, are removed from the dataset and used for testing the network's ability to interpolate between samples. The remaining data is used for training and validation. In this way, we reduce the probability of generating artifacts in the model predictions at the edges of the test set. In particular, we used ∼ 5 400 examples for training and ∼ 600 examples for validation.

8.1.2. Deep learning model

To emulate deposited energy distributions we employed the generative network described in Ref. [152]. It is a VAE [153] in which both encoding and decoding are based on Graph convolutional layers to fully take advantage of the cylindrical structure of our data. A VAE is a generative model made up by two networks: the encoder and the decoder. The encoder takes the input data and maps them to a distribution in a low-dimensional space, called latent space. The decoder is then trained to retrieve the input data starting from a point sampled from that distribution. The latent space is forced to be continuous; in this way, it is possible to generate new data with continuity, sampling from the latent space.

In our case, first, a suitable graph structure is imposed on the data. A node is associated to each voxel and nodes are connected to each other with nearest-neighbor connectivity. Graph data is then processed by the encoder which applies GraphConv layers [154], and lowers the graphs dimensionality using ReNN-Pool, a geometry-based pooling operation we developed for this task [152]. The encoded data is then processed as in standard VAE architecture with Gaussian prior, using the reparameterization trick, mapping the input data to Gaussian distributions in the latent space, which, in this case, is set as two-dimensional. The decoding uses the same graph representations employed for the encoding, but in reverse order. Nodes from a lower dimensional representation are up-sampled using the inverse operation of ReNN-Pool, and further processed by graph convolutions until the original graphs are recovered. The model is trained using the standard VAE loss with binary cross-entropy as the reconstruction term, and the Kullback–Leibler divergence term. We employed the Adam optimizer for weight updates, starting with a learning rate of 0.003. Additionally, we utilized an exponential scheduler with a decay factor of $\lambda = 0.9$.

8.1.3. Results and discussion

Reconstruction. The DL model was trained for 200 epochs on a Tesla V100 SXM2 GPU and the best set of learnable parameters was chosen as the one that minimizes the validation loss. Training was stopped at 200 epochs observing that the validation loss ceased decreasing in the last 40 epochs.

In Fig. 17 we show the energy profiles along the z and r axes, obtained by integrating the deposited energy distribution data along the other two dimensions, r, θ and z, θ, respectively. In each panel, the orange line corresponds to the reconstructed data from our Graph VAE, while the blue line refers to the ground truth, i.e. Monte Carlo simulated data. In Table 3 we report the mean relative errors on profiles and total energy deposition, with their standard deviation over the test set. Note that most of the errors are relative to the tails of the energy distribution, where the fluctuations of the Monte Carlo simulation are not always negligible. To quantify the node-per-node reconstruction quality we employ the δ index [155] taking inspiration from the global gamma index [156], used for clinical validation of treatment plans. It is defined as:

$$\delta = \frac{D_{gen} - D_{MC}}{max(D_{MC})}\,, \tag{5}$$

where D_{gen} is the deposited energy predicted by the VAE, while D_{MC} is the deposited energy obtained by the Monte Carlo simulation. As reconstruction performance measure we consider the 3% passing rate, which is the percentage of voxels with a delta index smaller than 3%. On the test set our Network reaches $98.6 \pm 0.3\%$ of voxels with 3% passing rate.

Latent space and generation. Once the network is trained, it is possible to generate new samples of energy distributions by picking and decoding a point from the latent space. Thus, to generate new samples conditioned to desired physical parameters, a relation between such parameters and the latent space variables is needed. Looking at the latent space, we found that the network has well decoupled and reconstructed the information about those physical parameters. Indeed, after applying a 2-D rigid rotation, which, by definition, does not change the latent space structure, a strong correlation between the two physical parameters of the simulated data, i.e. particle energy and slab density, and the two latent variables emerges. Such a result is shown in Fig. 18.

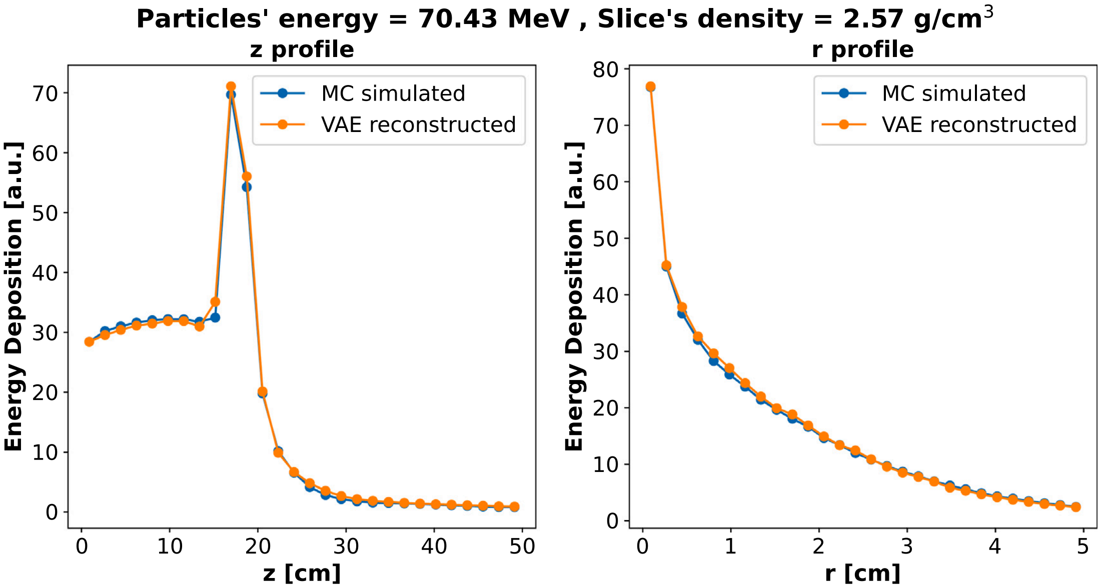


Fig. 17. Energy profiles reconstruction. Distribution of energy deposition along z and r axes. The blue line correspond to the Monte Carlo simulated data, while the orange line refers to the reconstructed data from our Network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Mean relative errors on energy profiles and total energy on test set. values are reported along with their standard deviations on the test set.

	Z profile	R profile	Total energy
Relative error	6.9 ± 3.4%	3.0 ± 1.2%	2.2 ± 1.6%

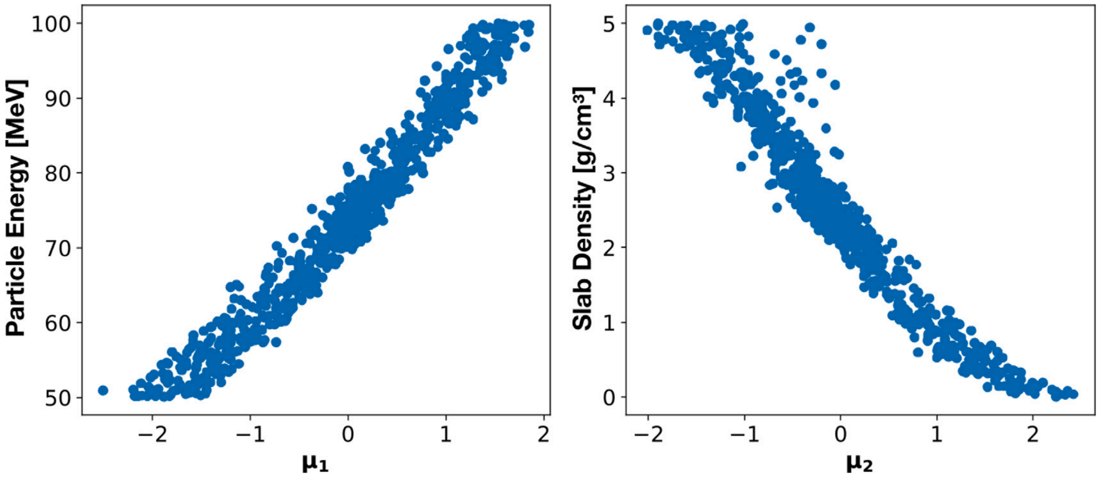


Fig. 18. Correlation between physical parameters and latent space variables. The latent variables (μ_1, μ_2) are highly correlated with particle energy and slice density. Thanks to this relation it is possible to generate new samples conditioned to those parameters.

Thanks to these relationships, the generation of deposited energy distribution for a given particle energy and slice density is straightforward. Once the desired physical parameters are chosen, it is possible to link them to a point in the latent space using a fit and then, decode that point to obtain the corresponding deposited energy distribution. The relation between latent variables and physical parameters presents some scattering, especially near the boundaries of the parameter space. This effect is due to the non perfect decoupling of the latent space in the whole parameters' range. The width of the scattered relations represents, indeed, the residual inter-correlations. This effect on large parameters' range is expected for a VAE, but in fact we designed our model to generate samples in the limited parameters' range of the test set, where a quasi-perfect decoupling is possible.

In order to test the generative capabilities of the Network we performed a third degree polynomial fit on the relations between latent variables and parameters. Then we considered the physical parameters of the examples in our test and validation sets. Using the results of the fit, we converted physical parameters to latent variables and used the decoder to generate the corresponding samples. We finally computed the δ index between the original and the generated samples. The 3% passing rate on the test set reaches $98.4 \pm 0.5\%$, a value that is compatible with the reconstruction result.

Discussion. We presented a Graph VAE that can generate deposited energy distributions inside a voxelized material. There are two main advantages to our deposited energy distribution generation technique. First, the described generation pipeline represents a way to write the deposited energy distribution in a volume as a differentiable function of beam parameters and medium density. Secondly, there is a huge gain from the point of view of computing time. Generating one example using Monte Carlo simulation with 4 threads and 10 000 primaries on our CPU (HP Z2 Tower G5 Workstation) took around 82 s, while our network, on the same device, needs only 0.02 s, which corresponds to a speedup of a factor of $10^3 - 10^4$. It is also worth considering that the deposited energy distributions for treatment planning are generally simulated with 1 million primaries, which make them even more computationally demanding. In contrast, the computing time for our network is independent of the number of primaries, because the DL model is trained to reproduce the per-particle dose distribution. In addition, the DL model's computing time can be further reduced on GPUs, thus the speed up with respect to Monte Carlo simulation is even greater. Future works will be focused on testing the network on more complex materials and finally, on a patient's Computed Tomography (CT) scan.

8.2. Quantification of uncertainties in a model-based iterative reconstruction algorithm for proton computed tomography

Proton computed tomography (pCT) is a novel medical imaging modality that provides the three-dimensional distribution of *relative stopping power* (RSP) in the scanned object ("phantom"). As the RSP is the measure of how the proton beams deposit their energy in e.g. a human body, RSP images form the basis of treatment planning procedures in proton radiation therapy. List-mode pCT establishes the RSP image from the energy losses of millions of individual protons, shot through the phantom on various paths. Several research groups are currently working on pCT detector hardware as well as simulation, tracking and tomographic reconstruction software [157,158]. This contribution presents our work to quantify uncertainties in the tomographic reconstruction step of the Bergen-pCT system [13].

The uncertainty quantification technique we developed is natively based on AD and can therefore be inserted as a step to a fully differentiable pipeline with optimization, thus resulting in an overall optimization loop that accounts for uncertainties, achieving uncertainty-aware design optimization.

8.2.1. Proton CT and model-based iterative reconstruction

In the setup of the Bergen pCT collaboration [13], a digital tracking calorimeter (DTC) is used to measure the position $x_{i,\mathrm{out}}$, direction $\dot{x}_{i,\mathrm{out}}$, and energy of the individual protons $i = 1, 2, \dots, m$ after they left the phantom. The position $x_{i,\mathrm{in}}$, direction $\dot{x}_{i,\mathrm{in}}$ and energy before entering the object are statistical properties of the proton beam source. The energy difference before and behind the phantom can be converted to a *water-equivalent path length* (WEPL) w_i.

Here, we substitute the measurement by a Monte Carlo simulation using GATE [12] and Geant4 [6–8] in a similar setup as in Section 2.1, and merely focus on the list-mode tomographic reconstruction task given the proton positions, directions and energies.

This task can be modeled as a least-squares problem in the following way. First, the unknown path of each proton through the phantom must be estimated from $x_{i,\mathrm{in}}$, $\dot{x}_{i,\mathrm{in}}$, $x_{i,\mathrm{out}}$, and $\dot{x}_{i,\mathrm{out}}$. The (extended) most likely path (MLP) formalism [159,160] provides such estimations based on the model of multiple Coulomb scattering by Lynch and Dahl [161] and Gottschalk [162].

Then, given the estimated path of length l_i, the (measured or simulated) WEPL w_i should satisfy the equation:

$$w_i = \int_0^{l_i} r(x_i(s))\,\mathrm{d}s\,, \tag{6}$$

where $x_i(s)$ is the position of the ith proton along the path and $r(x_i(s))$ is the RSP at this position. For every proton $i = 1, \dots, m$, an equation like (6) is obtained in this way, with a known WEPL w_i and an unknown (and sought) RSP distribution r. Discretizing r as a 3D image with n voxels, we obtain a linear system of equations $w = A_{\mathrm{x}} r$ with $A_{\mathrm{x}} \in \mathbb{R}^{m \times n}$, where the matrix entry $(A_{\mathrm{x}})_{ij}$ is related to the intersection length between the ith proton path (estimated based on $x_{i,\mathrm{in}}, \dot{x}_{i,\mathrm{in}}, x_{i,\mathrm{out}}, \dot{x}_{i,\mathrm{out}}$) and the jth voxel. Here, x is the collection of $x_{i,\mathrm{in}}, \dot{x}_{i,\mathrm{in}}, x_{i,\mathrm{out}}, \dot{x}_{i,\mathrm{out}}$ for all protons $i = 1, \dots, m$. The linear system is sparse, overdetermined ($m > n$) and can only be solved in a least-squares sense. Typically, this is done by iterative algorithms like ART or DROP [163] on a GPU.

8.2.2. Model of input uncertainty

Various types of uncertainties may arise during the pCT process and, thus, influence the process of the reconstruction of the 3-D image from proton positions and directions. As a long-term scope of this work, uncertainties shall be included in the process of designing a tracking calorimeter for pCT. For this purpose, we investigate on the effect of local perturbations of the input positions of protons by means of a differentiability analysis as well as on the effect of random inputs, focussing on proton positions, with the help of strategies for uncertainty propagation in a probabilistic description.

Epistemic and aleatoric uncertainties caused by calibration errors, measurement errors, or errors in the track reconstruction process may affect the inputs w and x to the RSP reconstruction. It is then of interest to quantify their effect on the reconstructed RSP image r. In the following, we focus on a scenario with an uncertain but uniform shift Δx_{in} of all proton positions $x_{i,\mathrm{in}}$ in front of the DTC (aleatoric uncertainty). Such a systematic measurement error may, e.g. arise from inaccuracies in the calibration of the magnets directing the proton beam and the positioning of the DTC, and lie in the range of a few millimeters. We model this error by a normally distributed random variable ΔX_{in} with mean 0 and standard deviation given by $3\sigma = 4\,\text{mm}$. This is only a modeling assumption and not based on actual measurements, but it is based on the experience that variations in the range of $\pm 4\,\text{mm}$ are observed with a high probability (99.7% corresponding to the 3σ-region of 4 mm). Realizations Δx_{in} of the random variable shift the proton positions, i.e. instead of the hit positions $x_{i,\text{in}}$ the DTC measures the perturbed positions $x_{i,\text{in}} + \Delta x_{\text{in}}$. To model more complex scenarios, like non-uniform random variations in the $x_{i,\text{in}}$ or the other measured quantities, a moderately larger number of random variables would have to be used.

8.2.3. Analysis of local perturbations

The image of a Gaussian distribution with mean μ and covariance matrix Σ under an affine-linear function

$$f : \mathbb{R}^q \to \mathbb{R}^{q'},\ z \mapsto J \cdot z + s \tag{7}$$

is again a Gaussian distribution, with mean $f(\mu)$ and covariance matrix $J \cdot \Sigma \cdot J^T$. In our setup, the output r depends on the uncertain shift Δx_{in} in a non-linear way. If the dependency was sufficiently smooth, its linearization (7) could still be used to approximately quantify the uncertainty in r, by Taylor's theorem; s would be the reconstructed image based on the measured x and w, and J would be the Jacobian matrix, which could be computed by AD/DP. However, our previous work [14] indicated that the usual way of defining $(A_{\text{x}})_{ij}$ as a mean chord length if the ith proton path and the jth voxel intersect, and zero otherwise, leads to a piecewise differentiable dependency that achieves most of its evolution through jumps. Methods relying on linearization are not applicable under these circumstances. Further work might employ the "fuzzy voxels" approach [14] that led to a piecewise differentiable dependency whose evolution between the jumps better resembled the global behavior, at the price of increased reconstruction time and blurrier solutions.

8.2.4. Probabilistic analysis

Apart from a local analysis that is only valid for small perturbations, it is of interest to observe and quantify the effects of uncertainties in a probabilistic setting. This may also help to analyze modeling strategies that try to include uncertain inputs.

There exist different methods to propagate input uncertainties in modeling. As already described above, we may express calibration errors using a moderate number of random variables Z with realizations z_i. This makes the use of projection methods like non-intrusive discrete projection computationally much more affordable than sampling-based strategies like classical Monte Carlo sampling. In the following, we consider a non-intrusive polynomial chaos approach (see e.g. Ref. [164]), which is also referred to as pseudo-spectral approach. In this approach, the probabilistic outcome q is expanded in terms of polynomials ϕ_i that are orthogonal with respect to the probability density function ρ_Z of the input random variables Z (compare with Ref. [165]). For a one-dimensional random input Z results in the expression:

$$f(Z) = \sum_{i=0}^{\infty} \hat{f}_i \phi_i(Z), \tag{8}$$

with deterministic coefficients $\hat{f}_i = \gamma_i^{-1} \int_{\mathbb{R}} f(z)\phi_i(z)\rho_Z(z)\,\mathrm{d}z$ and $\int_{\mathbb{R}} \phi_i(z)\phi_j(z)\rho_Z(z)\,\mathrm{d}z = \gamma_i \delta_{ij}$. When applied to find statistical quantities or to sample from this distribution, the infinite expansion is truncated. In the non-intrusive approach, the coefficients are approximated by employing a quadrature rule that is suitable for the used polynomials, e.g. Hermite polynomials for normally distributed variables. The computational effort of the quadrature can be reduced by using sparse grid quadrature rules.

For this study, we measure the quality of the reconstruction by observing the reconstruction error of the RSP values r_{recon} in the voxels, i.e.:

$$\Delta r := \|r_{\text{recon}} - r_{\text{org}}\|_2^2, \tag{9}$$

with r_{org} being a approximation of the original RSP representation of the phantom used in the simulation with GATE. The aim is to observe the influence of uncertainties on this reconstruction error. We restrict the error calculation to regions of interest for a specific test case. For this case we have two circular air holes in a phantom (see Fig. 19).

Uncertainties in the measurements may be specifically addressed in modeling when using an extended most likely path formalism [160]. Here, we provide the corresponding covariance matrices alongside $x_{i,\text{in}}, \dot{x}_{i,\text{in}}$ and $x_{i,\text{out}}, \dot{x}_{i,\text{out}}$ and extend the expression for the maximum likelihood estimate. With the strategies of Monte Carlo sampling (MC) and the non-intrusive polynomial chaos method (PC) we may compare the effects on the quality of the reconstruction for both path estimation formalism, the original MLP [159] and the extended MLP (EMLP) [160]. The resulting expected reconstruction quality and its variance are shown for both formalisms in Table 4.

In general, we observe that the non-intrusive polynomial chaos approach will provide a good approximation of the expected value and the variance with less function evaluations (60) than the Monte Carlo method (2 000). A more rigorous convergence analysis shows that, as expected, the polynomial chaos approach will converge faster than the Monte Carlo approach. However, we do not observe exponential convergence. This can usually only be observed for sufficiently smooth function which is not the case for the reconstruction error.

In Fig. 20, which shows the sampled reconstruction error (black), we observe a high amount of fluctuations in the function. The sampled polynomial chaos expansion (blue) is smoother and it may only reflect a global behavior. This can also be observed in the corresponding histograms. The general trend for observations with a high probability is similar. However, the values of a low reconstruction error around 0.0038 are not observed for the expansion. This shows that the expansion can be used to analyze average behavior and outcomes with high probabilities (which is of interest for detector design), but it is not feasible for analyzing, e.g. outliers.

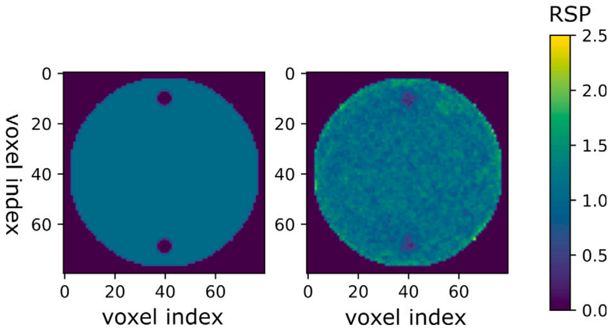


Fig. 19. RSP values for a slice (80 × 80 voxels, each sliced voxel is 2 mm × 2 mm) of the approximated original phantom (left) and a reconstructed image (right) with circular air holes in the upper and lower part. Note that the original phantom has more holes with less contrast.

Table 4
Expected value E and variance Var of the reconstruction error for the different formalisms.

Formalism	Method	Points	$\Delta r(E\Delta x)$	$E(\Delta r)$	Var(Δr)
MLP	MC	2000	0.0047410	0.0047158	3.1852e–08
MLP	PC	60	0.0047410	0.0047251	3.7064e–08
EMLP	MC	2000	0.0039905	0.0039281	6.3499e–09
EMLP	PC	60	0.0039905	0.0039128	6.9032e–09

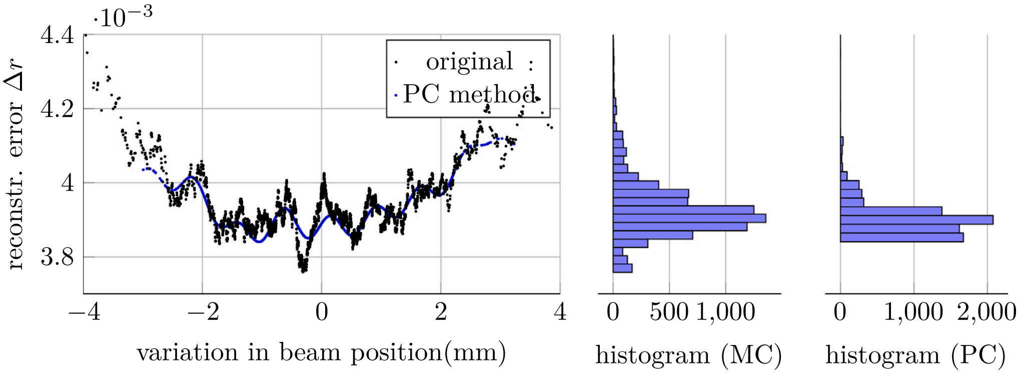


Fig. 20. Sampled the distribution of the reconstruction error (left) for the original distribution (Monte Carlo sampling, black) and the PC expansion (blue) and corresponding histograms (right) sharing the y-axis with the left plot (reconstruction error). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

8.2.5. Summary of uncertainty estimation for pCT reconstruction techniques

In this work, we have considered the effects of uncertainties on the image reconstruction in pCT using different modeling approaches. Regarding local perturbations, we observed that the algorithmic description of the reconstruction process is only piecewise differentiable. Jumps arise from the discrete computation of the set of voxels intersected by a proton path. For propagating the input uncertainty in the calibration, we could reduce the computational costs by using a non-intrusive discrete projection method. The modeling and propagation of uncertainties allowed us to assess the overall impact of the extended most-likely path approach on the quality of the reconstruction. Quantities like the expected value or the variance could be approximated with satisfying quality. In future work, it is intended to include the quantification of uncertainties in the design process of the calorimeter to find solutions that are more robust under perturbations.

9. Limitations of gradient-based optimization of end-to-end pipelines

Optimizing end-to-end models with gradient-based optimization is not without limitations and difficulties. The derivatives need to be integrated into an optimizer workflow in a way that ensures that the optimizer does not diverge. If the loss function is not well-behaved, the optimizer would easily diverge.

Handling design and outcome constraints also needs special care. In differentiable optimization, outcome constraints can be accounted for using optimizers such as “Method of Moving Asymptotes” or “Null-Space Optimizer”. Design constraints can be handled within the parameter-space definition, or geometry generation process. For non-differentiable optimization, constraints can be handled by some sub-classes of optimizers directly, e.g. bayesian optimization.

Requiring multi-objective optimization (MOO) is an indication that either the problem has not been thought about enough to reduce it to a single most important objective, or that a single design is attempting to appeal to multiple-use-cases. Often, single-objective optimization (SOO) can be achieved via constraining other metrics, that are of concern.

Whilst differentiable optimization is able to handle MOO, via multiple loss functions in a weighted combination, its nature as a local search will not attempt to populate/trace a pareto-optimal front (although one could be extracted a posteriori). If MOO really is required, then global searches, such as Bayesian Optimization and or Unified evolutionary optimization algorithms (UNSGA3) [166] may be used without significant change to the software.

10. Conclusions and future prospects

The end-to-end optimization of detectors for fundamental physics measurements constitutes a class of challenging and multi-faceted problems whose solution requires the development of specialized models performing a number of unique tasks. For instance, an experiment tasked with detecting high-energy cosmic rays and an experiment built to study particle collisions have very little in common, despite the fact that the underlying physics responsible for the data-generating processes is the same.

This heterogeneity has until recently discouraged a synergistic effort in the direction of providing a general solution. Nevertheless, the common traits these problems share, and the large potential value of the optimization of instruments whose price tag ranges in the several million to several hundred million euros, motivated us to invest our research time in this area. Our goal is to create a library of solutions and a toolkit that may empower the tackling of even harder use cases, by considering problems of comparatively small or moderate complexity, yet already at the edge of our capability.

Following the above plan, this document collects discussions of a range of use cases on which we are progressing toward end-to-end modeling and optimization solutions for experiments and instruments in fundamental physics. As a work in progress, each individual project discussed here only offers a snapshot of the current state of the art, rather than final solutions or crystallized results which, for the considered use cases, may or may not be far behind.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Maxime Lagrange and Zahraa Zaher for the fruitful discussions and the technical work that resulted in publications that sadly could not be part of this draft. This work was supported by the U.S. Department of Energy, under DOE Contract No. DE-AC02-76SF00515, Office of Science, Office of Basic Energy Sciences. This work was partially supported by the EU Horizon 2020 Research and Innovation Programme under grant agreement No. 101021812 (“SilentBorder”). Some of the research leading to these results has received funding from the Basic Research Program at the National Research University Higher School of Economics. Max Aehle gratefully acknowledges funding from the research training group SIVERT by the German federal state of Rhineland-Palatinate. Max Aehle, Lisa Kusch and Nicolas R. Gauger contributed on behalf of the Bergen pCT collaboration, whose current members include: Johan Alme, Gergely Gábor Barnaföldi, Tea Bodova, Vyacheslav Borshchov, Anthony van den Brink, Viljar Eikeland, Gregory Feofilov, Christoph Garth, Nicolas R. Gauger, Ola Grøttvik, Håvard Helstrup, Sergey Igolkin, Ralf Keidel, Chinorat Kobdaj, Tobias Kortus, Viktor Leonhardt, Shruti Mehendale, Raju Ningappa Mulawade, Odd Harald Odland, George O’Neill, Gábor Papp, Thomas Peitzmann, Helge Egil Seime Pettersen, Pierluigi Piersimoni, Maksym Protsenko, Max Rauch, Attiq Ur Rehman, Matthias Richter, Dieter Röhrich, Joshua Santana, Alexander Schilling, Joao Seco, Arnon Songmoolnak, Ganesh Tambave, Ihor Tymchuk, Kjetil Ullaland, Monika Varga-Kofarago, Lennart Volz, Boris Wagner, Steffen Wendzel, Alexander Wiebel, RenZheng Xiao, Shiming Yang, Hiroki Yokoyama, Sebastian Zillien. Jennet Dickinson (also supported by the DOE Early Career Research Program) and Lindsey Grey are supported by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the Department of Energy (DOE), Office of Science, Office of High Energy Physics. Borja S. González is grateful for the financial support by the FCT PhD grant PRT/BD/151553/2021 under the IDPASC program. Fredrik Sandin’s work was supported by the Jubilee and Kempe Foundations, project No. JCSMK JF-2303. Pietro Vischia’s work was supported by the “Ramón *y* Cajal” program under Project No. RYC2021-033305-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

References

[1] Adam Paszke, et al., PyTorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, pp. 8024–8035, URL https://dl.acm.org/doi/10.5555/3454287.3455008.

[2] M. Abadi, et al., TensorFlow: A system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 16, 2016, pp. 265–283.

[3] J. Bradbury, et al., JAX: composable transformations of Python+NumPy programs, 2018, http://github.com/google/jax. Version 0.2.5.

[4] M. Pereira, X. Cid, P. Vischia, Automatic optimisation of a parallel-plate avalanche counter with optical readout, Particles 8 (1) (2025) http://dx.doi.org/10.3390/particles8010026], URL https://www.mdpi.com/2571-712X/8/1/26.

[5] Tommaso, Andrea Giammanco, Pietro Vischia (eds), et al., Toward the end-to-end optimization of particle physics instruments with differentiable programming, Rev. Phys. 10 (2023) 100085, http://dx.doi.org/10.1016/j.revip.2023.100085, URL https://www.sciencedirect.com/science/article/pii/S2405428323000047.

[6] S. Agostinelli, et al., GEANT4 — a simulation toolkit, Nucl. Inst. Meth. A 506 (2003) 250, http://dx.doi.org/10.1016/S0168-9002(03)01368-8.

[7] J. Allison, et al., Geant4 developments and applications, IEEE Trans. Nucl. Sci. 53 (1) (2006) 270–278, http://dx.doi.org/10.1109/TNS.2006.869826.

[8] J. Allison, et al., Recent developments in Geant4, Nucl. Instr. Meth. A 835 (2016) 186–225, http://dx.doi.org/10.1016/j.nima.2016.06.125.

[9] Max Aehle, Johannes Blühdorn, Max Sagebaum, Nicolas R. Gauger, Forward-mode automatic differentiation of compiled programs, 2022, URL arXiv:2209.01895[cs].

[10] Max Aehle, Johannes Blühdorn, Max Sagebaum, Nicolas R. Gauger, Reverse-mode automatic differentiation of compiled programs, 2022, URL arXiv:2212.13760[cs].

[11] Nicholas Nethercote, Julian Seward, Valgrind: A framework for heavyweight dynamic binary instrumentation, SIGPLAN Not. 42 (6) (2007) 89–100, http://dx.doi.org/10.1145/1273442.1250746.

[12] S. Jan, et al., GATE - Geant4 application for tomographic emission: a simulation toolkit for PET and SPECT, Phys. Med. Biol. 49 (19) (2004) 4543–4561.

[13] J. Alme, et al., A high-granularity digital tracking calorimeter optimized for proton CT, Front. Phys. 8 (2020) 460, http://dx.doi.org/10.3389/fphy.2020.568243.

[14] Max Aehle, et al., Exploration of differentiability in a proton computed tomography simulation framework, Physics in Medicine & Biology 68 (24) (2023) 244002, http://dx.doi.org/10.1088/1361-6560/ad0bdd.

[15] Danilo Piparo, Vincenzo Innocente, Thomas Hauth, Speeding up HEP experiment software with a library of fast and auto-vectorisable mathematical functions, J. Phys.: Conf. Ser. 513 (5) (2014) 052027, http://dx.doi.org/10.1088/1742-6596/513/5/052027, URL https://iopscience.iop.org/article/10.1088/1742-6596/513/5/052027.

[16] Dougal Maclaurin, Modeling, Inference and Optimization With Composable Differentiable Procedures, (Ph.D. thesis), 2016.

[17] William Moses, Valentin Churavy, Instead of rewriting foreign code for machine learning, automatically synthesize fast gradients, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), in: Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 12472–12485, URL https://dl.acm.org/doi/abs/10.5555/3495724.3496770.

[18] C. Bastoul, et al., Putting polyhedral loop transformations to work, in: Lawrence Rauchwerger (Ed.), Languages and Compilers for Parallel Computing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, p. 209.

[19] Chris Lattner, Vikram Adve, LLVM: A compilation framework for lifelong program analysis and transformation, 2004, pp. 75–88.

[20] Jeff Bezanson, Alan Edelman, Stefan Karpinski, Viral B. Shah, Julia: A fresh approach to numerical computing, SIAM Rev. 59 (1) (2017) 65–98, http://dx.doi.org/10.1137/141000671, URL https://epubs.siam.org/doi/10.1137/141000671.

[21] Li Deng, The MNIST database of handwritten digit images for machine learning research, IEEE Signal Process. Mag. 29 (6) (2012) 141–142.

[22] G.C. Strong, M. Lagrange, A. Orio Alonso, A. Bordignon, F. Bury, T. Dorigo, A. Giammanco, M. Heikal, J. Kieseler, M. Lamparth, P. Martínez Ruíz del Árbol, F. Nardi, P. Vischia, H. Zaraket, Tomopt: differential optimisation for task- and constraint-aware design of particle detectors in the context of muon tomography, Mach. Learn.: Sci. Technol. 5 (3) (2024) 035002, http://dx.doi.org/10.1088/2632-2153/ad52e7.

[23] E. Rutherford, The scattering of α and β particles by matter and the structure of the atom, Lond. Edinb. Dubl. Phil. Mag. 21 (125) (1911) 669, http://dx.doi.org/10.1080/14786440508637080.

[24] Gerald R. Lynch, Orin I. Dahl, Approximations to multiple Coulomb scattering, Nucl. Inst. Meth. B 58 (1991) 6, http://dx.doi.org/10.1016/0168-583X(91)95671-Y.

[25] L.J. Schultz, et al., Image reconstruction and material z discrimination via cosmic ray muon radiography, Nucl. Instruments Methods Phys. Res. Sect. A: Accel. Spectrometers, Detect. Assoc. Equip. 519 (3) (2004) 687–694, http://dx.doi.org/10.1016/j.nima.2003.11.035, URL https://www.sciencedirect.com/science/article/pii/S0168900203028808.

[26] G.C. Strong, M. Lamparth, M. Lagrange, P. Vischia, F. Nardi, A. Giammanco, A. Orio Alonso, H Zaraket, Gilesstrong/tomopt: v.0.1.0: 1st publication version, 2024, http://dx.doi.org/10.5281/zenodo.10673885.

[27] M. Lagrange, Tomopt: Muon tomography experiment optimization, J. Adv. Instrum. Sci. 2024 (1) (2024) http://dx.doi.org/10.31526/jais.2024.492, URL https://jais.andromedapublisher.org/index.php/JAIS/article/view/492.

[28] Alexey Boldyrev, Denis Derkach, Fedor Ratnikov, Andrey Shevelev, ML-assisted versatile approach to calorimeter R&D, JINST 15 (09) (2020) C09030, http://dx.doi.org/10.1088/1748-0221/15/09/C09030, arXiv:2005.07700.

[29] R. Aaij, et al., LHCb Collaboration Collaboration, Physics case for an LHCb upgrade II - opportunities in flavour physics, and beyond, in the HL-LHC era, 2018.

[30] A. Alves, et al., LHCb Collaboration Collaboration, The LHCb detector at the LHC, JINST 3 (2008) S08005, http://dx.doi.org/10.1088/1748-0221/3/08/S08005.

[31] Roel Aaij, et al., LHCb Collaboration Collaboration, LHCb detector performance, Internat. J. Modern Phys. A 30 (07) (2015) 1530022, http://dx.doi.org/10.1142/S0217751X15300227, arXiv:1412.6352.

[32] Yu. Guz, Crystal Clear Collaboration Collaboration LHCb, The phase 2 upgrade of the LHCb calorimeter system, JINST 15 (09) (2020) C09046, http://dx.doi.org/10.1088/1748-0221/15/09/C09046.

[33] N. Bartosik, et al., Detector and physics performance at a muon collider, J. Instrum. 15 (05) (2020) P05001, http://dx.doi.org/10.1088/1748-0221/15/05/P05001.

[34] S. Ceravolo, et al., Crilin: A crystal calorimeter with longitudinal information for a future muon collider, J. Instrum. 17 (09) (2022) P09033, http://dx.doi.org/10.1088/1748-0221/17/09/p09033.

[35] L. Castelli, Machine learning approach to beam induced background shield at 3 tev muon collider, in: Presentation at the Fourth MODE Workshop, Valencia, Spain, 2024, https://indico.cern.ch/event/1380163/contributions/6013467/.

[36] F. Nardi, Towards the end-to-end design optimization of a muon collider calorimeter, 2023, https://indico.in2p3.fr/event/31308/.

[37] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, Maosong Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81, http://dx.doi.org/10.1016/j.aiopen.2021.01.001, URL https://www.sciencedirect.com/science/article/pii/S2666651021000012.

[38] Petar Veličković, Everything is connected: Graph neural networks, Curr. Opin. Struct. Biol. 79 (2023) 102538.

[39] Yue Wang, et al., Dynamic graph CNN for learning on point clouds, 2019, arXiv:1801.07829.

[40] S.R. Qasim, J. Kieseler, Y. Iiyama, M. Pierini, Learning representations of irregular particle-detector geometry with distance-weighted graph networks, Eur. Phys. J. C79 (7) (2019) 608, http://dx.doi.org/10.1140/epjc/s10052-019-7113-9, arXiv:1902.07987.

[41] 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, http://dx.doi.org/10.1140/epjc/s10052-020-08461-2, arXiv:2002.03605.

[42] Shah Rukh Qasim, et al., Multi-particle reconstruction in the high granularity calorimeter using object condensation and graph neural networks, EPJ Web Conf. 251 (2021) 03072, http://dx.doi.org/10.1051/epjconf/202125103072, arXiv:2106.01832.

[43] Shah Rukh Qasim, et al., End-to-end multi-particle reconstruction in high occupancy imaging calorimeters with graph neural networks, Eur. Phys. J. C 82 (8) (2022) http://dx.doi.org/10.1140/epjc/s10052-022-10665-7.

[44] S. Bhattacharya, et al., GNN-based end-to-end reconstruction in the CMS phase 2 high-granularity calorimeter, J. Phys.: Conf. Ser. 2438 (1) (2023) 012090, http://dx.doi.org/10.1088/1742-6596/2438/1/012090.

[45] Francesco Armando Di Bello, et al., Reconstructing particles in jets using set transformer and hypergraph prediction networks, 2022, arXiv:2212.01328.

[46] Raghav Kansal, et al., Particle cloud generation with message passing generative adversarial networks, 2022, arXiv:2106.11535.

[47] Erik Buhmann, Gregor Kasieczka, Jesse Thaler, Epic-GAN: Equivariant point cloud generation for particle jets, 2023, arXiv:2301.08128.

[48] Matthew Leigh, et al., PC-JeDi: Diffusion for particle cloud generation in high energy physics, 2023, arXiv:2303.05376.

[49] Erik Buhmann, et al., Caloclouds: Fast geometry-independent highly-granular calorimeter simulation, 2023, arXiv:2305.04847.

[50] CMS Collaboration, The Phase-2 Upgrade of the CMS Endcap Calorimeter, Technical Report CERN-LHCC-2017-023. CMS-TDR-019, CERN, 2017, URL https://cds.cern.ch/record/2293646.

[51] F. Torales Acosta, et al., The optimal use of segmentation for sampling calorimeters, J. Instrum. 19 (06) (2024) P06002, http://dx.doi.org/10.1088/1748-0221/19/06/P06002.

[52] Jan Kieseler, et al., Calorimetric measurement of multi-TeV muons via deep regression, Eur. Phys. J. C 82 (1) (2022) 79, http://dx.doi.org/10.1140/epjc/s10052-022-09993-5, URL https://cds.cern.ch/record/2776531, arXiv:2107.02119.

[53] Lai. S., et al., Shower separation in five dimensions for highly granular calorimeters using machine learning, 2024, URL https://arxiv.org/abs/2407.00178, arXiv:2407.00178.

[54] E. Lupi, et al., Neuromorphic readout for hadron calorimeters, Particles 8 (2) (2025) http://dx.doi.org/10.3390/particles8020052], URL https://www.mdpi.com/2571-712X/8/2/52.

[55] A. De Vita, et al., Hadron identification prospects with granular calorimeters, 2025, URL https://arxiv.org/abs/2502.10817.

[56] Nan Phinney, Nobukasu Toge, Nicholas Walker, ILC reference design report volume 3 - accelerator, 2007, arXiv:0712.2361.

[57] M. Ball, et al., The PIP-II conceptual design report, 2017, http://dx.doi.org/10.2172/1346823, URL https://www.osti.gov/biblio/1346823.

[58] Chad Mitchell, Ji Qiang, Paul Emma, Longitudinal pulse shaping for the suppression of coherent synchrotron radiation-induced emittance growth, Phys. Rev. ST Accel. Beams 16 (2013) 060703, http://dx.doi.org/10.1103/PhysRevSTAB.16.060703, URL https://link.aps.org/doi/10.1103/PhysRevSTAB.16.060703.

[59] Eric R. Colby, L.K. Len, Roadmap to the future, Rev. Accel. Sci. Technol. 09 (2016) 1–18, http://dx.doi.org/10.1142/S1793626816300012.

[60] Andrei Seryi, et al., A concept of plasma wake field acceleration linear collider (PWFA-LC), 2009, URL https://www.osti.gov/biblio/968518.

[61] M. Tzoufras, et al., Beam loading in the nonlinear regime of plasma-based acceleration, Phys. Rev. Lett. 101 (2008) 145002, http://dx.doi.org/10.1103/PhysRevLett.101.145002.

[62] Ryan Roussel, et al., Single shot characterization of high transformer ratio wakefields in nonlinear plasma acceleration, Phys. Rev. Lett. 124 (4) (2020) 044802, http://dx.doi.org/10.1103/PhysRevLett.124.044802, URL https://link.aps.org/doi/10.1103/PhysRevLett.124.044802.

[63] Helmut Wiedemann, Particle accelerator physics, 3rd ed., Springer, Berlin, 2007, http://dx.doi.org/10.1007/978-3-540-49045-6.

[64] Eduard Prat, Masamitsu Aiba, Four-dimensional transverse beam matrix measurement using the multiple-quadrupole scan technique, Phys. Rev. ST Accel. Beams 17 (2014) 052801, http://dx.doi.org/10.1103/PhysRevSTAB.17.052801.

[65] M. Gordon, et al., Four-dimensional emittance measurements of ultrafast electron diffraction optics corrected up to sextupole order, Phys. Rev. Accel. Beams 25 (2022) 084001, http://dx.doi.org/10.1103/PhysRevAccelBeams.25.084001.

[66] Brandon Cathey, Sarah Cousineau, Alexander Aleksandrov, Alexander Zhukov, First six dimensional phase space measurement of an accelerator beam, Phys. Rev. Lett. 121 (2018) 064804, http://dx.doi.org/10.1103/PhysRevLett.121.064804.

[67] D. Marx, et al., Single-shot reconstruction of core 4D phase space of high-brightness electron beams using metal grids, Phys. Rev. Accel. Beams 21 (2018) 102802, http://dx.doi.org/10.1103/PhysRevAccelBeams.21.102802.

[68] Jonathan C. Wong, et al., 4D transverse phase space tomography of an operational hydrogen ion beam via noninvasive 2D measurements using laser wires, Phys. Rev. Accel. Beams 25 (2022) 042801, http://dx.doi.org/10.1103/PhysRevAccelBeams.25.042801.

[69] O. R. Sander, G. N. Minerbo, R. A. Jameson, D. D. Chamberlin, Beam tomography in two and four dimensions, in: 8th Particle Accelerator Conference, 1980, pp. S5–10.

[70] Minwen Wang, et al., Four-dimensional phase space measurement using multiple two-dimensional profiles, Nucl. Instrum. Methods Phys. Res. Sect. A: Accel. Spectrometers, Detect. Assoc. Equip. 943 (2019) 162438, http://dx.doi.org/10.1016/j.nima.2019.162438.

[71] A. Wolski, et al., Transverse phase space characterization in an accelerator test facility, Phys. Rev. Accel. Beams 23 (2020) 032804, http://dx.doi.org/10.1103/PhysRevAccelBeams.23.032804.

[72] A. Scheinker, Adaptive machine learning for time-varying systems: low dimensional latent space tuning, J. Instrum. 16 (10) (2021) P10008, http://dx.doi.org/10.1088/1748-0221/16/10/P10008.

[73] A. Wolski, et al., Transverse phase space tomography in an accelerator test facility using image compression and machine learning, Phys. Rev. Accel. Beams 25 (2022) 122803, http://dx.doi.org/10.1103/PhysRevAccelBeams.25.122803.

[74] R. Roussel, et al., Phase space reconstruction from accelerator beam measurements using neural networks and differentiable simulations, Phys. Rev. Lett. 130 (2023) 145001, http://dx.doi.org/10.1103/PhysRevLett.130.145001.

[75] J. Gonzalez-Aguilera, et al., Towards fully differentiable accelerator modeling, in: Proc. IPAC'23, in: IPAC'23 - 14th International Particle Accelerator Conference, (no. 14) JACoW Publishing, Geneva, Switzerland, 2023, pp. 1725–1729, http://dx.doi.org/10.18429/JACoW-14thInternationalParticleAcceleratorConference-WEPA065, URL https://indico.jacow.org/event/41/contributions/2122.

[76] K.M. Hock, M.G. Ibison, A study of the maximum entropy technique for phase space tomography, J. Instrum. 8 (02) (2013) P02003, http://dx.doi.org/10.1088/1748-0221/8/02/P02003.

[77] J.D. Lawson, P.M. Lapostolle, R.L. Gluckstern, Emittance, entropy and information, Part. Accel. 5 (1973) 61–65.

[78] Gao Huang, et al., Snapshot ensembles: Train 1, get M for free, 2017, arXiv:1704.00109.

[79] K. Rajput, et al., Harnessing the power of gradient-based simulations for multi-objective optimization in particle accelerators, 2024, URL https://arxiv.org/abs/2411.04817, arXiv:2411.04817.

[80] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1999.

[81] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning; From Theory to Algorithms, Cambridge University Press, 2014.

[82] Gianfranco Bertone, Dan Hooper, History of dark matter, Rev. Modern Phys. 90 (2018) 045002, http://dx.doi.org/10.1103/RevModPhys.90.045002, URL https://link.aps.org/doi/10.1103/RevModPhys.90.045002.

[83] Jodi Cooley, et al., Report of the topical group on particle dark matter for snowmass 2021, 2022, arXiv:2209.07426.

[84] Julien Billard, et al., Direct detection of dark matter—APPEC committee report*, Rep. Progr. Phys. 85 (5) (2022) 056201, http://dx.doi.org/10.1088/1361-6633/ac5754.

[85] M. Ibe, W. Nakano, Y. Shoji, et al., Migdal effect in dark matter direct detection experiments, J. High Energ. Phys. 194 (2018) 2018, http://dx.doi.org/10.1007/JHEP03(2018)194.

[86] Katherine Freese, Joshua Frieman, Andrew Gould, Signal modulation in cold-dark-matter detection, Phys. Rev. D 37 (1988) 3388–3405, http://dx.doi.org/10.1103/PhysRevD.37.3388, URL https://link.aps.org/doi/10.1103/PhysRevD.37.3388.

[87] R. Bernabei, et al., Further results from DAMA/LIBRA-phase2 and perspectives, Nucl. Phys. At. Energy 22 (2021) 329, http://dx.doi.org/10.15407/jnpae2021.04.329.

[88] David N. Spergel, Motion of the earth and the detection of weakly interacting massive particles, Phys. Rev. D 37 (1988) 1353–1355, http://dx.doi.org/10.1103/PhysRevD.37.1353, URL https://link.aps.org/doi/10.1103/PhysRevD.37.1353.

[89] J.B.R. Battat, et al., Readout technologies for directional WIMP dark matter detection, Phys. Rep. 662 (2016) 1–46, http://dx.doi.org/10.1016/j.physrep.2016.10.001, arXiv:1610.02396.

[90] D.S. Akerib, et al., Snowmass2021 cosmic frontier dark matter direct detection to the neutrino fog, 2022, arXiv:2203.08084.

[91] R. Essig, et al., Snowmass2021 cosmic frontier: The landscape of low-threshold dark matter direct detection in the next decade, 2022, arXiv:2203.08297.

[92] J. Amaré, et al., Annual modulation results from three-year exposure of ANAIS-112, Phys. Rev. D 103 (2021) 102005, http://dx.doi.org/10.1103/PhysRevD.103.102005, URL https://link.aps.org/doi/10.1103/PhysRevD.103.102005.

[93] The COSINE-100 Collaboration, Three-year annual modulation search with COSINE-100, Phys. Rev. D 106 (2022) 052005, http://dx.doi.org/10.1103/PhysRevD.106.052005, URL https://link.aps.org/doi/10.1103/PhysRevD.106.052005.

[94] S.E. Vahsen, et al., CYGNUS: Feasibility of a nuclear recoil observatory with directional sensitivity to dark matter and neutrinos, 2020, arXiv:2008.12587.

[95] C.A.J. O'Hare, et al., Recoil imaging for directional detection of dark matter, neutrinos, and physics beyond the standard model, 2022, arXiv:2203.05914.

[96] The EDELWEISS Collaboration, Optimizing EDELWEISS detectors for low-mass WIMP searches, Phys. Rev. D 97 (2018) 022003, http://dx.doi.org/10.1103/PhysRevD.97.022003, URL https://link.aps.org/doi/10.1103/PhysRevD.97.022003.

[97] The LUX Collaboration, Improving sensitivity to low-mass dark matter in LUX using a novel electrode background mitigation technique, Phys. Rev. D 104 (2021) 012011, http://dx.doi.org/10.1103/PhysRevD.104.012011, URL https://link.aps.org/doi/10.1103/PhysRevD.104.012011.

[98] The LUX Collaboration, Fast and flexible analysis of direct dark matter search data with machine learning, Phys. Rev. D 106 (2022) 072009, http://dx.doi.org/10.1103/PhysRevD.106.072009, URL https://link.aps.org/doi/10.1103/PhysRevD.106.072009.

[99] The DRIFT collaboration, Improved sensitivity of the DRIFT-IId directional dark matter experiment using machine learning, J. Cosmol. Astropart. Phys. 2021 (07) (2021) 014, http://dx.doi.org/10.1088/1475-7516/2021/07/014.

[100] G. Adhikari, et al., Lowering the energy threshold in COSINE-100 dark matter searches, Astropart. Phys. 130 (2021) 102581, http://dx.doi.org/10.1016/j.astropartphys.2021.102581, URL https://www.sciencedirect.com/science/article/pii/S0927650521000256.

[101] M. Antonello, et al., Characterization of SABRE crystal nai-33 with direct underground counting, Eur. Phys. J. C 81 (2021) 299, http://dx.doi.org/10.1140/epjc/s10052-021-09098-5.

[102] F. Calaprice, et al., Performance of a SABRE detector module without an external veto, 2022, arXiv:2205.13876.

[103] I. Coarasa, et al., Machine-learning techniques applied to three-year exposure of ANAIS–112, J. Phys.: Conf. Ser. 2156 (1) (2021) 012036, http://dx.doi.org/10.1088/1742-6596/2156/1/012036.

[104] I. Coarasa, et al., Improving ANAIS-112 sensitivity to DAMA/LIBRA signal with machine learning techniques, J. Cosmol. Astropart. Phys. 2022 (11) (2022) 048, http://dx.doi.org/10.1088/1475-7516/2022/11/048.

[105] The IceCube Collaboration, Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A, Science 361 (6398) (2018) eaat1378, http://dx.doi.org/10.1126/science.aat1378, arXiv:1807.08816.

[106] B.P. Abbott, et al., Multi-messenger observations of a binary neutron star merger, Astrophys. J. Lett. 848 (2) (2017) L12, http://dx.doi.org/10.3847/2041-8213/aa91c9, arXiv:1710.05833.

[107] B.P. Abbott, et al., Gravitational waves and Gamma-rays from a binary neutron star merger: GW170817 and GRB 170817a, Astrophys. J. Lett. 848 (2) (2017) L13, http://dx.doi.org/10.3847/2041-8213/aa920c, arXiv:1710.05834.

[108] Andrea Albert, et al., Science case for a wide field-of-view very-high-energy Gamma-ray observatory in the southern hemisphere, 2019, arXiv preprint arXiv:1902.08429.

[109] Meng Su, Tracy R. Slatyer, Douglas P. Finkbeiner, Giant gamma-ray bubbles from Fermi-LAT: active galactic nucleus activity or bipolar galactic wind? Astrophys. J. 724 (2) (2010) 1044.

[110] P. Assis, et al., Design and expected performance of a novel hybrid detector for very-high-energy gamma astrophysics, Astropart. Phys. 99 (2018) 34–42, http://dx.doi.org/10.1016/j.astropartphys.2018.02.004.

[111] B.S. Acharya, et al., Introducing the CTA concept, Astropart. Phys. 43 (2013) 3–18.

[112] A.U. Abeysekara, et al., Observation of the crab nebula with the HAWC Gamma-ray observatory, Astrophys. J. 843 (1) (2017) 39, http://dx.doi.org/10.3847/1538-4357/aa7555, arXiv:1701.01778.

[113] Xiong Zuo, et al., LHAASO Collaboration Collaboration, Design and performances of prototype muon detectors of LHAASO-KM2A, Nucl. Instrum. Meth. A 789 (2015) 143–149, http://dx.doi.org/10.1016/j.nima.2015.04.010.

[114] P. Abreu, et al., MARTA: a high-energy cosmic-ray detector concept for high-accuracy muon measurement, Eur. Phys. J. C 78 (4) (2018) 333, http://dx.doi.org/10.1140/epjc/s10052-018-5820-2, arXiv:1712.07685.

[115] D. Heck, J. Knapp, J.N. Capdevielle, G. Schatz, T. Thouw, A Monte Carlo code to simulate extensive air showers, Rep. FZKA 6019 (1998).

[116] B.S. González, et al., Tackling the muon identification in water cherenkov detectors problem for the future southern wide-field Gamma-ray observatory by means of machine learning, Neural Comput. Appl. (2022) 1–14, http://dx.doi.org/10.1007/s00521-021-06730-z, arXiv:2101.11924.

[117] B.S. González, et al., Using convolutional neural networks for muon detection in WCD tank, J. Phys.: Conf. Ser. 1603 (2020) 012024, http://dx.doi.org/10.1088/1742-6596/1603/1/012024.

[118] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[119] François Chollet, et al., Keras, 2015, https://keras.io.

[120] R. Conceição, et al., Muon identification in a compact single-layered water cherenkov detector and gamma/hadron discrimination using machine learning techniques, Eur. Phys. J. C 81 (6) (2021) 542, http://dx.doi.org/10.1140/epjc/s10052-021-09312-4, arXiv:2101.10109.

[121] R. Conceição, B.S. González, M. Pimenta, B. Tomé, $P^{\alpha}_{\gamma h}$: A new variable for γ/h discrimination in large gamma-ray ground arrays, Phys. Lett. B (2022) 136969, http://dx.doi.org/10.1016/j.physletb.2022.136969, URL https://www.sciencedirect.com/science/article/pii/S0370269322001034, arXiv:2108.13954.

[122] Pedro Assis, et al., The mercedes water cherenkov detector, Eur. Phys. J. C 82 (10) (2022) 899.

[123] R.L. Workman, et al., Particle Data Group Collaboration Collaboration, Review of particle physics, PTEP 2022 (2022) 083C01, http://dx.doi.org/10.1093/ptep/ptac097.

[124] T. Dorigo, et al., End-to-end optimization of the layout of a Gamma ray observatory, 2023, arXiv e-prints, arXiv:2310.01857.

[125] Carver Mead, How we created neuromorphic engineering, Nat. Electron. 3 (7) (2020) 434–435.

[126] Dennis V. Christensen, et al., 2022 roadmap on neuromorphic computing and engineering, Neuromorphic Comput. Eng. 2 (2) (2022) 022501, http://dx.doi.org/10.1088/2634-4386/ac4a83.

[127] John Shalf, The future of computing beyond Moore's law, Phil. Trans. R. Soc. A 378 (2166) (2020) 20190061.

[128] Charles E. Leiserson, et al., There's plenty of room at the top: What will drive computer performance after Moore's law? Science 368 (6495) (2020) eaam9744, http://dx.doi.org/10.1126/science.aam9744.

[129] Adnan Mehonic, Anthony J. Kenyon, Brain-inspired computing needs a master plan, Nature 604 (7905) (2022) 255–260.

[130] Wulfram Gerstner, Werner M. Kistler, Richard Naud, Liam Paninski, Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition, Cambridge University Press, 2014, URL https://neuronaldynamics.epfl.ch/.

[131] Pietro Vischia, Angel A. Caputi, Modelling the neurons of the electrosensory lobe in gymnotus omarorum with differentiable programming, 2023, http://dx.doi.org/10.5281/zenodo.8394819.

[132] Melika Payvand, et al., Self-organization of an inhomogeneous memristive hardware for sequence learning, Nat. Commun. 13 (1) (2022) 5793, http://dx.doi.org/10.1038/s41467-022-33476-6.

[133] Kwabena Boahen, Dendrocentric learning for synthetic intelligence, Nature 612 (7938) (2022) 43–50, http://dx.doi.org/10.1038/s41586-022-05340-6.

[134] Arjun Rao, Philipp Plank, Andreas Wild, Wolfgang Maass, A long short-term memory for AI applications in spike-based neuromorphic hardware, Nat. Mach. Intell. 4 (5) (2022) 467–479, http://dx.doi.org/10.1038/s42256-022-00480-w.

[135] Rui-Jie Zhu, Qihang Zhao, Jason K. Eshraghian, SpikeGPT: Generative pre-trained language model with spiking neural networks, 2023, arXiv preprint arXiv:2302.13939.

[136] Chiara Bartolozzi, Giacomo Indiveri, Elisa Donati, Embodied neuromorphic intelligence, Nat. Commun. 13 (1) (2022) 1024, http://dx.doi.org/10.1038/s41467-022-28487-2.

[137] James Aimone, et al., A review of non-cognitive applications for neuromorphic computing, Neuromorphic Comput. Eng. (2022).

[138] Catherine D. Schuman, et al., Opportunities for neuromorphic computing algorithms and applications, Nat. Comput. Sci. 2 (2022) 10–19, http://dx.doi.org/10.1038/s43588-021-00184-y.

[139] Simon Thorpe, Arnaud Delorme, Rufin Van Rullen, Spike-based strategies for rapid processing, Neural Netw. 14 (6) (2001) 715–725, http://dx.doi.org/10.1016/S0893-6080(01)00083-1.

[140] X. Xing, et al., SpikeLLM: Scaling up spiking neural network to large language models via saliency-based spiking, 2024, URL https://arxiv.org/abs/2407.04752, arXiv:2407.04752.

[141] Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, Gabriel F. Manso, Deep learning's diminishing returns: The cost of improvement is becoming unsustainable, IEEE Spectr. 58 (10) (2021) 50–55, http://dx.doi.org/10.1109/MSPEC.2021.9563954.

[142] Jason K. Eshraghian, et al., Training spiking neural networks using lessons from deep learning, 2023, arXiv:2109.12894.

[143] Felix C. Bauer, Gregor Lenz, Saeid Haghighatshoar, Sadique Sheik, EXODUS: Stable and efficient training of spiking neural networks, Front. Neurosci. 17 (2023) http://dx.doi.org/10.3389/fnins.2023.1110444.

[144] Mattias Nilsson, et al., Integration of neuromorphic AI in event-driven distributed digitized systems: Concepts and research directions, Front. Neurosci. 17 (2023) http://dx.doi.org/10.3389/fnins.2023.1074439.

[145] Karen Adam, Adam Scholefield, Martin Vetterli, Sampling and reconstruction of bandlimited signals with multi-channel time encoding, IEEE Trans. Signal Process. 68 (2020) 1105–1119, http://dx.doi.org/10.1109/TSP.2020.2967182.

[146] Le Ye, et al., The challenges and emerging technologies for low-power artificial intelligence IoT systems, IEEE Trans. Circuits Syst. I. Regul. Pap. 68 (12) (2021) 4821–4834, http://dx.doi.org/10.1109/TCSI.2021.3095622.

[147] Ali Safa, et al., Exploring information-theoretic criteria to accelerate the tuning of neuromorphic level-crossing ADCs, in: Proceedings of the 2023 Annual Neuro-Inspired Computational Elements Conference, NICE '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 63–70, http://dx.doi.org/10.1145/3584954.3584994.

[148] Shruti R. Kulkarni, et al., On-sensor data filtering using neuromorphic computing for high energy physics experiments, 2023, arXiv:2307.11242.

[149] Dominique J. Kösters, et al., Benchmarking energy consumption and latency for neuromorphic computing in condensed matter and particle physics, APL Mach. Learn. 1 (1) (2023) 016101, http://dx.doi.org/10.1063/5.0116699.

[150] Emil Schüler, et al., Very high-energy electron (VHEE) beams in radiation therapy; treatment plan comparison between VHEE, VMAT, and PPBS, Med. Phys. 44 (6) (2017) 2544–2555.

[151] Karl Otto, Volumetric modulated arc therapy: IMRT in a single gantry arc, Med. Phys. 35 (1) (2008) 310–317.

[152] Lorenzo Arsini, et al., Nearest neighbours graph variational AutoEncoder, Algorithms 16 (3) (2023) http://dx.doi.org/10.3390/a16030143, URL https://www.mdpi.com/1999-4893/16/3/143.

[153] Diederik P. Kingma, Max Welling, Auto-encoding variational Bayes, 2014, arXiv:1312.6114.

[154] Christopher Morris, et al., Weisfeiler and leman go neural: Higher-order graph neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 4602–4609.

[155] Florian Mentzel, et al., Fast and accurate dose predictions for novel radiotherapy treatments in heterogeneous phantoms using conditional 3D-UNet generative adversarial networks, Med. Phys. 49 (5) (2022) 3389–3404, http://dx.doi.org/10.1002/mp.15555, URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.15555.

[156] Daniel A. Low, William B. Harms, Sasa Mutic, James A. Purdy, A technique for the quantitative evaluation of dose distributions, Med. Phys. 25 (5) (1998) 656–661.

[157] G. Poludniowski, N.M. Allinson, P.M. Evans, Proton radiography and tomography with application to proton therapy, Br. J. Radiol. 88 (1053) (2015) 20150134, http://dx.doi.org/10.1259/bjr.20150134.

[158] R.P. Johnson, Review of medical radiography and tomography with proton beams, Rep. Progr. Phys. 81 (1) (2018) 016701, http://dx.doi.org/10.1088/1361-6633/aa8b1d.

[159] R.W. Schulte, S.N. Penfold, J.T. Tafas, K.E. Schubert, A maximum likelihood proton path formalism for application in proton computed tomography: Maximum likelihood path formalism for proton CT, Med. Phys. 35 (11) (2008) 4849–4856, http://dx.doi.org/10.1118/1.2986139.

[160] N. Krah, et al., A comprehensive theoretical comparison of proton imaging set-ups in terms of spatial resolution, Phys. Med. Biol. 63 (13) (2018) 135013, http://dx.doi.org/10.1088/1361-6560/aaca1f.

[161] Gerald R. Lynch, Orin I. Dahl, Approximations to multiple Coulomb scattering, Nucl. Instruments Methods Phys. Res. Sect. B: Beam Interactions Mater. Atoms 58 (1) (1991) 6–10, http://dx.doi.org/10.1016/0168-583X(91)95671-Y.

[162] B. Gottschalk, et al., Multiple Coulomb scattering of 160 MeV protons, Nucl. Instruments Methods Phys. Res. Sect. B: Beam Interactions Mater. Atoms 74 (4) (1993) 467–490, http://dx.doi.org/10.1016/0168-583X(93)95944-Z, URL https://linkinghub.elsevier.com/retrieve/pii/0168583X9395944Z.

[163] Scott Penfold, Yair Censor, Techniques in iterative proton CT image reconstruction, Sens. Imaging 16 (1) (2015) 19, http://dx.doi.org/10.1007/s11220-015-0122-3, URL http://link.springer.com/10.1007/s11220-015-0122-3.

[164] Dongbin Xiu, Efficient collocational approach for parametric uncertainty analysis, Commun. Comput. Phys. 2 (2) (2007) 293–309.

[165] Roger G. Ghanem, Pol D. Spanos, Stochastic Finite Elements: A Spectral Approach, Springer New York, New York, NY, 1991, http://dx.doi.org/10.1007/978-1-4612-3094-6, URL http://link.springer.com/10.1007/978-1-4612-3094-6.

[166] H. Seada, K. Deb, U-NSGA-III: A unified evolutionary optimization procedure for single, multiple, and many objectives: Proof-of-principle results, in: Evolutionary Multi-Criterion Optimization, Springer International Publishing, Cham, 2015, p. 34.