PAPER • OPEN ACCESS

TomOpt: differential optimisation for task- and constraint-aware design of particle detectors in the context of muon tomography

To cite this article: Giles C Strong *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 035002

View the article online for updates and enhancements.

MACHINE LEARNING
Science and Technology

PAPER

TomOpt: differential optimisation for task- and constraint-aware design of particle detectors in the context of muon tomography

Giles C Strong[1,2,*], Maxime Lagrange[1,3,*], Aitor Orio[4,5,*], Anna Bordignon[6], Florian Bury[7], Tommaso Dorigo[1,2,8], Andrea Giammanco[1,3], Mariam Heikal[9], Jan Kieseler[1,10], Max Lamparth[1,11], Pablo Martínez Ruíz del Árbol[1,12], Federico Nardi[1,13,14], Pietro Vischia[1,15] and Haitham Zaraket[1,16,17]

[1] MODE Collaboration, https://mode-collaboration.github.io/
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Italy
[3] Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, Louvain la Neuve, Belgium
[4] Muon Tomography Systems S.L., Bilbao, Spain
[5] Universidad de Cantabria (UC), Santander, Spain
[6] Department of Statistical Sciences, University of Padova, Padova, Italy
[7] University of Bristol, Bristol, United Kingdom
[8] Luleå University of Technology, Luleå 97187, Sweden
[9] American University of Beirut, Beirut, Lebanon
[10] Karlsruhe Institute of Technology, Karlsruhe, Germany
[11] Physik-Department, Technische Universität München, Munich, Germany
[12] Instituto de Física de Cantabria, Santander, Spain
[13] Dipartimento di Fisica ed Astronomia, University of Padova, Padova, Italy
[14] Université Clermont-Auvergne, Clermont-Ferrand, France
[15] Universidad de Oviedo and ICTEA, Oviedo, Spain
[16] Multi-Disciplinary Physics Laboratory, Optics and Fiber Optics Group, Faculty of Science, Lebanese University, Beirut, Lebanon
[17] Laboratoire de Physique Subatomique et de Cosmologie, Université Grenoble-Alpes, Grenoble France
[*] Authors to whom any correspondence should be addressed.

E-mail: giles.c.strong@gmail.com, maxime.lagrange@uclouvain.be and aitor.orio@muon.systems

Abstract

We describe a software package, TomOpt, developed to optimise the geometrical layout and specifications of detectors designed for tomography by scattering of cosmic-ray muons. The software exploits differentiable programming for the modeling of muon interactions with detectors and scanned volumes, the inference of volume properties, and the optimisation cycle performing the loss minimisation. In doing so, we provide the first demonstration of end-to-end-differentiable and inference-aware optimisation of particle physics instruments. We study the performance of the software on a relevant benchmark scenario and discuss its potential applications. Our code is available on Github (Strong *et al* 2024 available at: https://github.com/GilesStrong/tomopt).

1. Introduction

1.1. Differentiable programming for detector design optimisation

Over the past two decades, the availability of high-performance computing and the development of neural networks of larger capacity have conspired to fuel a revolution in the way we think at the optimisation of complex systems. When the dimensionality of the space of relevant design parameters exceeds a few units, and brute-force scans cease to be a viable option for its exploration. We nowadays, have the option of letting automated systems find their way to configurations that correspond to advantageous extrema of carefully specified objective functions. The engine under the hood of these optimisation searches is automatic differentiation, which allows computer programs to keep track of the gradient of the objective function, through the chain rule of differential calculus, as computer code performs arbitrarily complex successions of operations to model the behaviour of the system.

Crucial to a successful optimisation of the system is the inclusion in the model of all relevant effects that have an impact on the precision of the inference that the data generated by the system may produce. An

incomplete description of the inference itself, or a mock up of the reconstruction techniques performing the dimensionality reduction step which translates raw data into high-level features informing the objective function, are likely to prevent the identification of designs that maximise the true objective, as they introduce a misalignment. Instead, it is possible to introduce a substantial margin of flexibility for the components of the model responsible for explaining factors that influence the absolute magnitude of the objective function but do not affect its gradient direction within the design space. The model, in other words, can be approximate as long as it correctly captures the interdependence of the various components, as the latter is what matters in determining the gradients of the objective function.

Particle detectors are among the most complex interconnected systems humans have ever designed and built: the large experiments at the CERN LHC, for example, include tens of millions of detection elements reading out highly interdependent physical phenomena produced when protons collide in the core of the detectors, at rates of 40 million times a second. For most of present-day experimental situations, we do not yet possess the capability of producing a fully differentiable, high-fidelity end-to-end model of the whole procedure that translates raw detector readouts into final measurements of physical quantities; this prevents the application of gradient descent methods to optimise the detector design. Because of the high complexity of that goal, we approach the problem by turning, for the time being, to easier modelling tasks that nonetheless share the same basic ingredients and structure with it. The successful optimisation of particle detectors of low complexity may allow us to teach ourselves how to tackle progressively harder problems, while accumulating working solutions to sub-tasks which are liable to be reassembled into larger modeling problems as needed.

Particle detectors used in muon tomography applications (see next subsection) appear especially fit to an end-to-end modelling. They involve physical phenomena of low complexity as the data-generating process usually involves the reconstruction of only one particle at a time, and they are liable to a successful parametric modeling of all the relevant effects. They thus constitute an ideal entry point in a long-term study of the end-to-end optimisation of full-fledged experiments. In this document we describe the studies we undertook to develop a software package for the optimisation of muon tomography apparatuses. In its present implementation, the code only considers a subset of the possible three-dimensional geometries of tomography systems, and is meant to provide only a proof of principle of the accessibility of a full, end-to-end optimisation of the problems these instruments are tasked to solve—most notably, imaging of inaccessible volumes, and hypothesis testing on their composition. It is our intention to develop a future version of the software which will provide a more complete solution of the concrete problems in tomography applications.

1.2. Muon scattering tomography

Muons, elementary particles related to the electrons but about 200 times heavier, are constantly and naturally produced by cosmic-ray interactions in the atmosphere. Their flux at sea level is of the order of 100 Hz m^{-2}, and their energy spectrum is very broad, peaking at a few GeV and extending up to the TeV scale. Muons are not subject to the strong nuclear interaction, and in the energy range 1 GeV–100 GeV their electromagnetic interactions with matter are relatively mild: they mostly loose energy by ionisation, at a rate of about 200 MeV per meter of water. This makes them the most penetrating charged elementary particles.

The first known use of these cosmogenic muons for a practical application was the measurement of the overburden of a tunnel in an Australian mine, reported in 1955 [1]. A decade later, muons were famously used for searching for hidden chambers in Chephren's pyramid in Egypt [2]. While that pioneering attempt did not find any new chambers, it convincingly established their absence from the bulk of the volume of that pyramid. Half a century later, the ScanPyramids collaboration was more lucky and was able to report the actual identification of unexpected features in Khufu's Great Pyramid [3, 4]. These results have been based on the absorption of muons, i.e. on measuring the reduction of the muon flux due to their passage through matter. The same method has been applied successfully to several other use cases in archaeology and geophysics (see [5] for reports on several recent examples). However, since 2003 another process of muon-matter interaction has been used for a variety of applications: multiple scattering, where the muon undergoes several elastic electromagnetic interactions with the nuclei of the traversed material. 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 [6, 7].

The first proposal to use muon scattering for elemental classification [8] focused on the search for high-Z materials hidden in containers. Searches of illicit materials at border controls are still one of the main drivers for research in muon tomography, see [9] for a recent review. However, in the last two decades this method has been explored for a large variety of other applications, including material identification in nuclear waste drums [10], detection of nuclear warheads [11], and various industrial applications to process monitoring and preventive maintenance [12]. While most applications of the method are focusing on the identification of high-Z elements, some teams are attacking the much more challenging task of identifying relatively light

materials [13–16], which demands excellent angular resolution given the aforementioned dependence on Z of the scattering angle.

Muon imaging, considering both absorption- and scattering-based methods, is a booming research direction with tens of new publications per year at a steadily growing rate [17]. The interested reader can find more information in some recent reviews [12, 18].

Inferring the atomic number of a passive volume irradiated by a flux of muons from cosmic rays requires modelling the muon flux at the Earth surface, and the multiple scattering muons undergo in matter due to multiple interactions with the Coulomb field of medium nuclei and electrons.

A formula for the flux of cosmic muons at sea level was first given by Gaisser in 1990 [19], neglecting Earth's curvature and assuming muons do not decay while they traverse the atmosphere; the formula is therefore approximately valid only for muons arriving at sea level with a zenith angle $\theta < 70°$ and with an energy $E_\mu > 100/\cos\theta$ GeV. An improved formula was given by Guan *et al* [20], accounting for the curvature of the Earth and providing a modified term that improves the agreement of the formula with experimental data at lower muon energies. Other alternative formulas have been proposed in the literature, e.g. the model by Shukla and Sankrith [21] provides a closed form for the zenith angle distribution which gives an improved description of the data at all angles, including at $\theta > 70°$. TOMOPT, the software package we developed and describe in section 2, enables the user to choose between the model by Guan *et al* and that of Shukla *et al*.

For what concerns the scattering of muons in a material, a simple model is recommended by the Particle Data Group (PDG) [22] and used in TomOpt as the default scattering model (which we refer to as the *PDG scattering model*). It assumes a small-angle Gaussian scattering, and does not take into account non-Gaussian tails from rare large-angle scatterings. The model consists of picking two random scattering angles independently in orthogonal directions, as well as correlated transverse displacements. The initial model was proposed to predict a homogeneous material propagation, but then used by several groups in a step-by-step algorithm to model propagation in a volume discretised in voxels (*voxelised*) according to the radiation length X_0 of the voxels the muons traverse.

The PDG scattering model begins by computing the width of the angular distribution θ_0. For high energy muons with momentum p travelling a distance d the PDG model assumes:

$$\theta_0 = \frac{0.0136\text{GeV}}{p}\sqrt{n_{X_0}} \times (1 + 0.038\ln n_{X_0}),\tag{1}$$

where $n_{X_0} = d/X_0$, d is the distance traversed by the muon in the material in meters and X_0 is the material radiation length in meters. The logarithmic term is added in the model to account for possible low Z materials and to be able to reproduce Molière's theory [23, 24]. The angular scattering is sampled from a zero-centred Gaussian distribution of this width with random numbers z_1 and z_2 taken from a normal distribution of mean 0 and variance 1:

$$\delta\theta = \theta_0 \times z_2.\tag{2}$$

The lateral displacement due to scattering is computed as:

$$\delta x = \theta_0 \left(z_1 \frac{d}{\sqrt{12}} + z_2 \frac{d}{2} \right).\tag{3}$$

1.3. Overview of this document

The present document is organised as follows: in section 2 we describe the software package we developed, which was used to obtain all results we discuss in the following sections; section 3 describes our benchmark of choice, involving the estimation of the fill level at a metal refinery; Finally, in section 4 we summarise the results and discuss the future perspectives of this work.

2. TomOpt

The simulation and optimisation routines used for this study are structured into a software package referred to as TOMOPT. The code for TOMOPT is available open-source on Github [25]. In this section we will attempt to provide the reader with a sufficient understanding of the program.

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. Its modular design allows each stage of the process to be performed by inheriting classes written to be most suitable for the particular task being studied. The key feature of TOMOPT is the fully differentiable nature of the detector modelling and inference pipeline, implemented using the automatic differentiation framework PYTORCH [26]. In this section we describe the capabilities and workflow of TOMOPT.
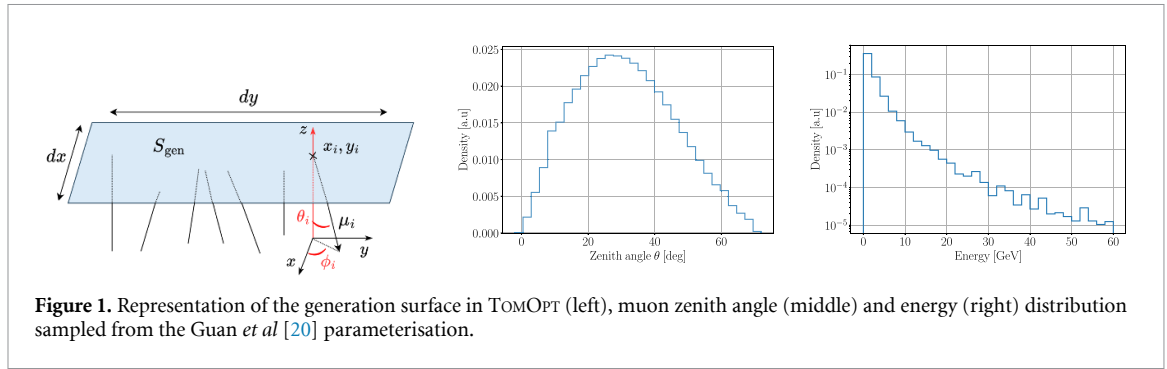
Figure 1. Representation of the generation surface in TomOpt (left), muon zenith angle (middle) and energy (right) distribution sampled from the Guan *et al* [20] parameterisation.

The TomOpt project began back in April 2021 with the aim of providing a practical demonstration of the differentiable pipeline proposed by the MODE Collaboration in [27] (and later published in [28]). Initially worked at by two of us, the project quickly gained contributors, helped by a proof-of-concept demonstration at the 1st MODE workshop on differentiable programming [29] in September that year. Over 2022, progress focussed on improving the capabilities, stability, and physical accuracy of TomOpt, as well as explorations into advanced inference using maximum likelihood methods, and deep-learning approaches. The primary contributors met twice for extended in-person development and knowledge sharing (both times kindly hosted by CP3, Belgium). Additionally, TomOpt was used as the basis for a data-challenge preceding the 2nd MODE workshop on differentiable programming [30], where further progress on the project was presented [31]. Progress in 2023 concerned the development of benchmarks for this publication. It was slowed slightly by one of the primary contributors leaving academia for industry and only able to offer evenings and weekends; but anyway, here we are at last!

2.1. Cosmic muon generation

In the context of detector optimisation, it is essential to have a realistic and flexible cosmic muon source that can adapt to the various muon scattering tomography applications and geographical locations. With that in mind, TomOpt samples muons from a horizontal generation surface S_{gen} with customisable energy and angular distributions. Users are free to provide their own parameterisation of the differential cosmic muon flux, expressed as:

$$\Phi_z(E,\theta) = \frac{dN}{dE \cdot d\theta \cdot d\phi \cdot dS_z} \ \text{m}^{-2} \cdot \text{sr}^{-1} \cdot \text{GeV}^{-1}, \tag{4}$$

where dN is number of muon crossing a horizontal surface element dS_z, with energy between E and $E + dE$, zenith angle between θ and $\theta + d\theta$, and azimuthal angle between ϕ and $\phi + d\phi$. Most of cosmic muon flux parametrisation are given as:

$$\Phi(E,\theta) = \frac{dN}{dE \cdot dS_n \cdot d\Omega} \ \text{m}^{-2} \cdot \text{sr}^{-1} \cdot \text{GeV}^{-1}, \tag{5}$$

where dN is number of muon crossing a surface element dS_n perpendicular to the muon direction, with energy between E and $E + dE$, and solid angle between Ω and $\Omega + d\Omega$. Because $d\Omega = \sin\theta \, d\theta \, d\phi$ and $dS_z = dS_n \cos\theta$, Φ_z can be obtained from Φ as:

$$\Phi_z = \Phi \cdot \sin\theta \cdot \cos\theta. \tag{6}$$

The present version of TomOpt embeds two default differential muon flux parameterisations: Guan *et al* [20] and Sankrith [21]. After providing the desired generation surface span dx, dy, energy range E_{range}, zenith angle range θ_{range}, and number of muons N_{muon}, their position is uniformly distributed on S_{gen}. Figure 1 shows the various variables associated to a generated muon as well as the energy and zenith angle distributions of a muon batch.

For computing efficiency purposes, it might be relevant to choose θ_{range} upper bound $< 70°$ since horizontal muons are unlikely to reach both the passive volume and the detector layers. Once muons are generated, they will be propagated along negative z direction. The simulation of their propagation through passive and active materials is described in sections 2.2 and 2.3 respectively.
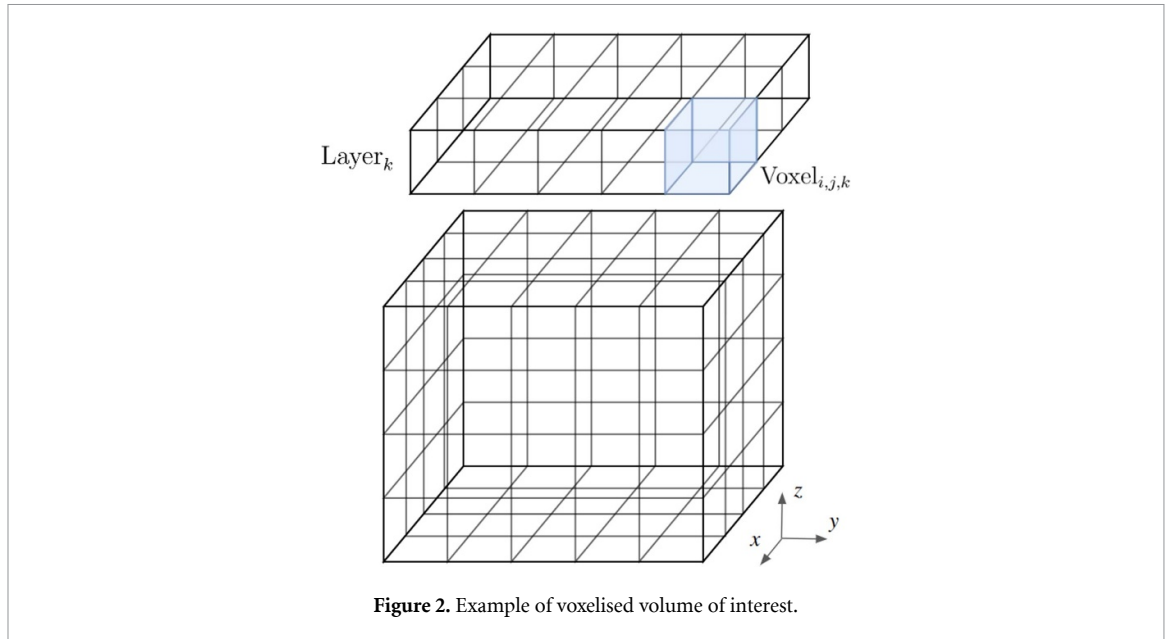
Figure 2. Example of voxelised volume of interest.

2.2. Passive volumes and multiple scattering

Once muons are generated, TOMOPT models their passage through the volume of interest (*passive volume*) and their interaction with matter. The passive volumes of interest are modelled in 3D using layers of voxels, and one can assign to each voxel a material (specified by its radiation length X_0), as shown in figure 2. TOMOPT has it own database of X_0 values for various pure elements computed assuming nominal material density at $20\,^{\circ}$C. Alternatively, users can easily provide their custom materials radiation lengths and densities, which are then plugged into equation (1). They may also manually specify the layout of materials inside the volumes of interest, or instead provide a customised generator class to provide randomly generated volumes that are representative of the current task.

For computational efficiency via vectorisation, batches of many muons traverse the volume simultaneously. Each muon is defined using its momentum, (x, y, z) position, and (θ, ϕ) trajectory, where $0 \leqslant \theta < \pi/2$[18] is the deviation from the z axis, and $0 \leqslant \phi < 2\pi$ is the azimuthal angle in the xy plane.

The PDG scattering model described in equation (1) is the default scattering model implemented in TOMOPT. We show in section 2.8 that, although embedding such model in a step-by-step algorithm affects the precision of the Coulomb scattering simulation compared to, for instance, highly accurate but slow models such as those provided in GEANT 4 [32–34], the optimisation-focussed nature of the procedure we present in this paper requires prioritising computational speed over precision.

As described in section 1.2, we sample the angular scattering according to equation (2) and subsequently compute the displacement due to the scattering according to equation (3): this procedure is performed twice per scattering call, to compute scatterings in perpendicular directions (x_μ and y_μ) in the muon reference frame. The muon-frame scatterings must then be converted into the volume frame using 3D rotations matrices. As mentioned earlier in this section, muons are propagated by steps of distance dr along their flight path. The default value of dr is set to 1 cm, but can be changed to accommodate for the voxel size, the volume of interest size, and the remaining distance to the bottom of the voxel. The optimal choice of dr is discussed in section 2.8. The algorithm 1 summarises TOMOPT muon propagation implementation.

---

[18] Muons travelling upwards are removed, since all muons in the batch must share the same z position.

Algorithm 1. Muon propagation in TOMOPT.

---

for each voxel layer **do**

    for each muon **do**

        while muon inside layer **do**

            Compute $\theta_0(dr, X_0)$ equation (1), $\delta\theta_\mu$ equation (2), $(\delta x_\mu, \delta y_\mu)$ equation (3) $\triangleright \mu$ denotes variables computed in the muon reference frame

            Compute $d\theta_{\text{vol}}, d\phi_{\text{vol}}, \delta x_{\text{vol}}, \delta y_{\text{vol}}, \delta z_{\text{vol}}$ $\triangleright$ vol denotes variables computed in the volume reference frame

            $x \leftarrow x + \delta x_{\text{vol}}$ $\triangleright$ Update muon position

            $y \leftarrow y + \delta y_{\text{vol}}$

            $z \leftarrow z + \delta z_{\text{vol}}$

            $\theta \leftarrow x + \delta\theta_{\text{vol}}$ $\triangleright$ Update muon direction

            $\phi \leftarrow y + \delta\phi_{\text{vol}}$

            if $z_\mu < z_{\text{bottom}}$ **then**

            $z_\mu = z_{\text{bottom}}$ $\triangleright$ Muons that get propagated outside of the layer are moved (up) to the bottom of the layer, x, y and z positions are updated according to θ and ϕ

            end if

        end while

    end for

  end for

2.3. Detector modelling and hit recording

The detectors modelled in TOMOPT must have a flexible enough parameterisation to adapt to the use case, whilst not being constrained by *a priori* assumptions about the optimal configuration. The basic detector unit used is referred to as a *panel*. In *xy*-space, each panel has a spatial resolution σ_{max} and detection efficiency ϵ_{max}, along with a fixed cost per unit area. In this way, the user may simulate detectors that are either constructable in-house, or otherwise commercially available. Panels are free to be positioned in *xyz* space within pre-specified regions, referred to as *detector layers*. In optimisation, the aim is to learn the optimal position in *xyz* and width in *xy* (*xy*-span) per panel; five learnable parameters per panel. Currently, the number of panels per detector layer is fixed throughout the optimisation process, due the difficulties in computing gradients for integer variables.
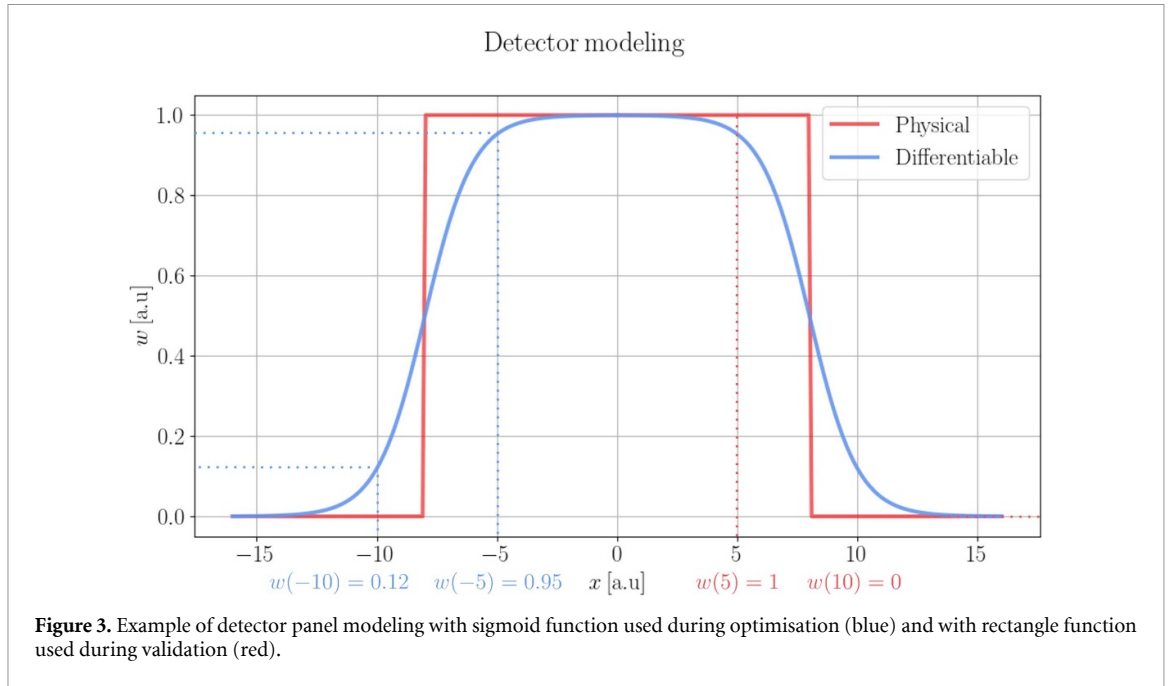
Muon positions (referred to as *hits*) are recorded in *xyz* by the panels. The effect of spatial resolution on the hits will depend on the detector technology used. For simplicity, the panels implemented in TOMOPT assume that the recorded position of a hit can be sampled from a Gaussian distribution $\mathcal{G}\left(\mu = xy_{\text{true}}, \sigma = \sigma_{xy,\text{eff}}\right)$ centered over the true position of the muon xy_{true}, the scale of which depends on the effective spatial resolution $\sigma_{xy,\text{eff}}$, as dictated by σ_{max} of the detection panel, and the position of the muon relative to the panel centre.

In reality, a muon passing through a panel would either record a hit, or not, depending on the efficiency of the panel. For differentiability reasons, TOMOPT instead takes a probabilistic interpretation: hits are always recorded, but with an efficiency factor ϵ_{hit}. The total efficiency of the muon ϵ_μ is then computed given its individual hits as the probability of at least two hits before and after the passive volume[19]. Both the hit and overall efficiency enter various parts of the inference and optimisation process, and will be discussed in later sections.

Optimisation of the panel parameters requires that the muon trajectories be differentiable with respect to them. Trajectory reconstruction (discussed next) will naturally be differentiable with respect to the *z* position of the hits, however derivatives for the *xy*-position and *xy*-span are non-trivial: these affect only whether a muon hits a panel or not, and hits are then recorded with a constant resolution and efficiency, i.e. the optimisation process would be insensitive to muons which could have been recorded but were not due to the position and size of the panel.

Instead, during optimisation mode, TOMOPT takes an unphysical approach in which 'panels' will always record a hit for every muon at its current *z* position, even if these hits are located outside the panel. The effective resolutions $\sigma_{xy,\text{eff}}$ and efficiencies ϵ_{eff} of hits, however decreases the further away the hits are from the centre of the panel, at a rate that scales with the span of the panel. They are computed by scaling the maximum panel resolution σ_{max} and efficiency ϵ_{max}, according to where the muon passes the panel's *z*-position:

---

[19] At least two hits are required to reconstruct a trajectory, but more hits will better constrain the uncertainty on the trajectory parameters.

Figure 3. Example of detector panel modeling with sigmoid function used during optimisation (blue) and with rectangle function used during validation (red).

$$\sigma_{xy,\text{eff}} = \sigma_{\max}/w\left(xy_{\text{true}}\right), \quad \epsilon_{\text{eff}} = \epsilon_{\max} \cdot w\left(xy_{\text{true}}\right), \tag{7}$$

where $w(xy_{\text{true}})$ is the panel model evaluated at the true muon-position, as presented in figure 3. This scales between zero and one, reaching a maximum when the muon hits at the centre of the panel. The current model used in TOMOPT is a pair of double-sigmoid functions in x and y. The values of this pair are used to provide resolutions in xy, and their product provides the hit efficiency. These are then scaled such that a hit at the centre receives the full resolution and efficiency of the panel. Effectively, this can provide detection performance close to the actual detector for hits within the panel, and a smooth transition to zero efficiency and resolution outside the panel. In this way, each hit and muon trajectory will be differentiable with respect to all parameters of the panel. When derivatives are not required, e.g. for validating the detector, the panels can of course be reverted to a more physical interpretation. The algorithm 2 presented below summarises TOMOPT hit recording implementation.
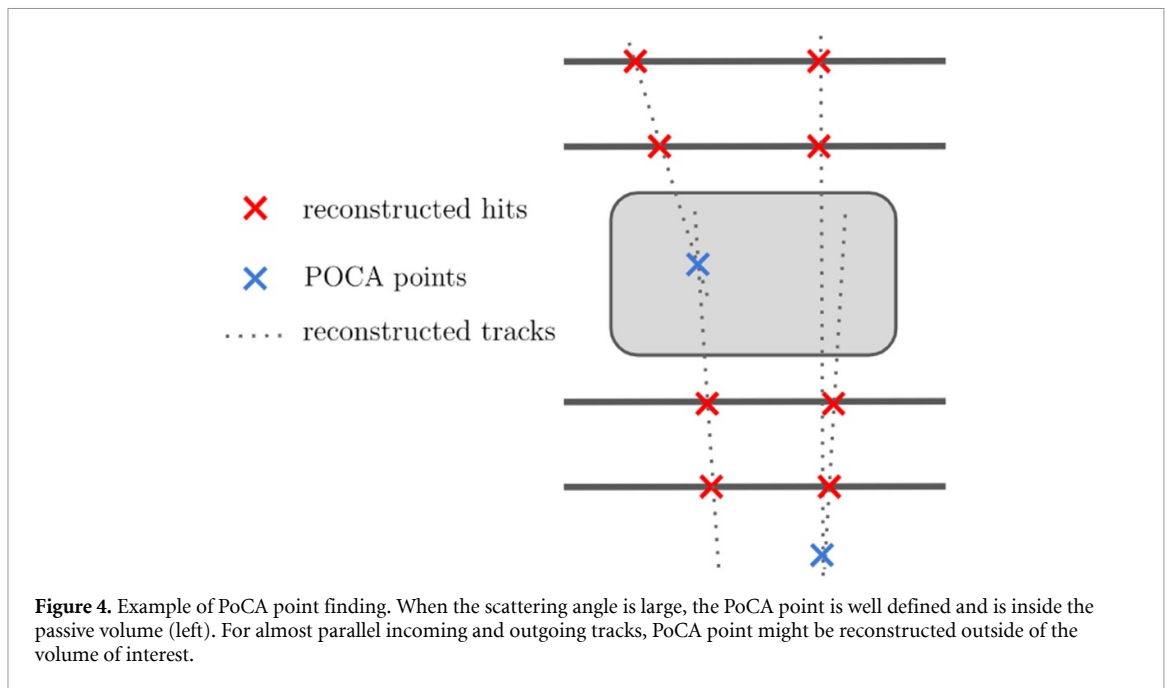
The smoothness of the sigmoid model (how slowly the hit performance transitions at the panel edge) can be adjusted, and even evolved during optimisation. Inference with smoother panels will be more sensitive to muons that are far from the panel, and so are more suitable for optimising the xy position and span of the panels: since the inference will account for the efficiency on the muons, muons far from the panel are able to affect the inference. Panels with a sharper transition have close to full hit performance within the majority of the area of the panel, and so inference is more sensitive to the z position of the panels: muons far from the panel receive diminished weights, and muons inside the panel receive a more uniform efficiency and resolution therefore the only remaining parameter to have non-negligible gradient is the z-position. Again, the user is free to implemented their own panel model w to reach the desired detector configuration. Examples could be different scalings in x and y, or a periodic function to represent detection wires with a given density.

Algorithm 2. Hit recording in TOMOPT.

---

for each muon **do**
    for each detector panel **do**
        Propagate muon to the next detector panel at z_j
        Record true hit: $(x_{\text{true}}, y_{\text{true}}, z_j)$
        Compute effective hit resolutions: $\sigma_{x,\text{eff}} = \sigma_{\max}/w(x_{\text{true}}), \; \sigma_{y,\text{eff}} = \sigma_{\max}/w(y_{\text{true}})$
        Compute reconstructed hit: $(x_{\text{true}} + \delta_x, y_{\text{true}} + \delta_y, z_j)$ ▷$\delta_{x,y}$ sampled from $\mathcal{G}\left(0, \sigma_{x,y,\text{eff}}\right)$
        Compute hit efficiency $\epsilon_{\text{eff}} = \epsilon_{\max} \cdot w(x_{\text{true}}) \cdot w(y_{\text{true}})$
    end for
end for

Figure 4. Example of PoCA point finding. When the scattering angle is large, the PoCA point is well defined and is inside the passive volume (left). For almost parallel incoming and outgoing tracks, PoCA point might be reconstructed outside of the volume of interest.

2.4. Trajectory fitting and Point-of-Closest-Approach finding

The aim of trajectory fitting is to determine the paths of each muon as they enter and exit the passive volume using the hits recorded by the two sets of detector panels. Each trajectory is computed via analytic minimisation of a likelihood function that takes the *xyz* hits as an input. When computing the fit, a weight is associated to each of the hit positions, that depends on the effective *xy* resolution of the hits $\sigma_{xy,\text{eff}}$, and their efficiencies ϵ_{eff}, i.e. low-resolution, or low efficiency, hits have less of an impact on the trajectory fit than higher resolution, or more probable hits. No uncertainty in *z* is considered. From an optimisation perspective, the true trajectory of the muon is most likely to be recovered when the detector panels are positioned such that all hits for the muon are recorded with high efficiency and resolution. Potentially this would come at the cost of a poor reconstruction of other muon tracks, and so a global compromise must be found, as determined by the loss function of the muon tomography task being performed.

Once the incoming and outgoing trajectories of the muons are computed, scattering variables such as the incoming and outgoing θ and ϕ angles, and the total amount of angular scattering may be differentiably computed. By exploiting the differentiable computation of these variables, their associated uncertainties can also be easily computed using the product of their partial derivatives with respect to the recorded hits and the *xy* uncertainties on the hits.

Once muon tracking is performed, TOMOPT proceeds to a first inference phase in order to associate a region of space with the measured muon deflection. It is done using an extended version of the Point-of-Closest-Approach algorithm, which is a standard in muon scattering tomography. In this approach, the entirety of the muon scattering is assumed to have occurred at a single point in the passive volume, located by extrapolating the inferred trajectories, as shown in figure 4. Because the PoCA points are computed from the incoming and outgoing reconstructed tracks, they also bear an uncertainty in *x, y* and *z*. The PoCA points may then be used for a variety of purposes, such as: directly producing a 3D image for human interpretation; or computing hit densities in voxels, perhaps scaling their weights by their associated angular scattering. Exactly which variables are required, and the way in which they are used depends on the task-specific inference approach used. We will show an example of this in section 3.3.1.

Algorithm 3. Track fitting and PoCA finding in TOMOPT.

---

for each muon **do**
    for incoming and outgoing tracks **do**
        Fit track vector
    end for
    Compute PoCA point from the track vectors
    Compute PoCA point uncertainty via automatic differentiation
end for

2.5. Volume inference

The purpose of 'volume inference' is to convert the muon-trajectory variables into a (set of) prediction(s) for the volume. The form of these predictions and the inference method used will depend on the muon tomography task being performed. A variety of methods may be used here, provided they allow the predictions to be differentiable with respect to the muon-trajectory variables, and a range of these will be explored in a second publication.

Our implementation of PoCA in TomOpt aims to compute estimates of the voxel X0 values by inverting the PDG scattering model. The basic approach for a given batch of muons with their PoCA located in a given voxel of height δz is:

$$\theta_0^{\mathrm{rms}} = \frac{\theta_{\mathrm{tot.}}^{\mathrm{rms}}}{\sqrt{2}}, \tag{8}$$

$$\theta_0^{\mathrm{rms}} = \frac{0.0136}{GeV} p^{\mathrm{rms}} \sqrt{n_{X_0}}^{\mathrm{rms}},$$

$$n_{X_0}^{\mathrm{rms}} = \frac{\delta z}{X_0 \cos\left(\bar{\theta}^{\mathrm{rms}}\right)},$$

$$\bar{\theta}^{\mathrm{rms}} = \frac{\theta_{\mathrm{in}}^{\mathrm{rms}} + \theta_{\mathrm{out}}^{\mathrm{rms}}}{2}, \tag{9}$$

$$X_0 = \left(\frac{0.0136}{GeV} p^{\mathrm{rms}}\right)^2 \frac{\delta z}{\cos\left(\bar{\theta}^{\mathrm{rms}}\right)} \frac{2}{\theta_{\mathrm{tot.}}^{\mathrm{rms}\,2}}. \tag{10}$$

where $\theta_{\mathrm{tot.}}^{\mathrm{rms}}$ is the root-mean-square (RMS) of the measured scattering angles. In equation (9), $\theta_{\mathrm{in}}^{\mathrm{rms}}$ and $\theta_{\mathrm{in}}^{\mathrm{rms}}$ respectively refer to the incoming and outgoing muon zenith angle. An example of volume inference is summarized in algorithm 4. Note that the natural log term present in equation (1) has been ignored to simplify the inversion. The effects of efficiency and resolution may be further included in equation (10) by instead aggregating over the muons with a weighted RMS, where muon-wise weights are computed as the efficiency divided by the variance of the squared value of the variable being aggregated.

In the above inference, we only considered muons whose PoCA was located inside the voxel in question, however we have an uncertainty on the PoCA position, and some muons may have a PoCA located outside the passive volume due to mis-reconstruction. An extension is then to consider that muons can contribute to inference in multiple voxels by considering that the PoCA could have occurred in voxels surrounding its nominal location. The PoCA uncertainty PoCA_unc in x, y and z is computed via automatic differentiation. Then the probability density function (PDF) for a PoCA to be located within a voxel is modelled using 3 independent Gaussian distributions centered over the nominal PoCA location and scaled in the three directions according to PoCA_unc. A 'scattering probability' is then computed per voxel for each muon, by integrating the PoCA PDFs over ever voxel in the passive volume. An illustration of this procedure is presented on figure 5. This probability then enters into the weighted RMS as a multiplicative coefficient.

Algorithm 4. Volume inference in TomOpt.

---

   for each PoCA point **do**
      Compute scattering probability per voxel
      Compute muon weights per voxel as muon efficiency times scattering probability in each voxel
   end for
   for each voxel **do**
      Compute weighted θ_0^{rms} from equation (8)
      Compute X_0 from equation (10)
   end for

This method of using PoCA points, however, is biased to underestimate X_0, since the entirety of the scattering is assumed to happen in one scattering event. The muons of course undergo many scattering events as they traverse through the volume. Additionally, it converts muon-wise data into voxel-wise predictions, however the muon tomography task may not require voxel-wise predictions, but instead perhaps predictions concerning the volume as a whole. Nonetheless, it can serve as a useful and generic first-stage of inference to serve as input to a second, more task-specific, stage of inference; indeed both of the benchmarks
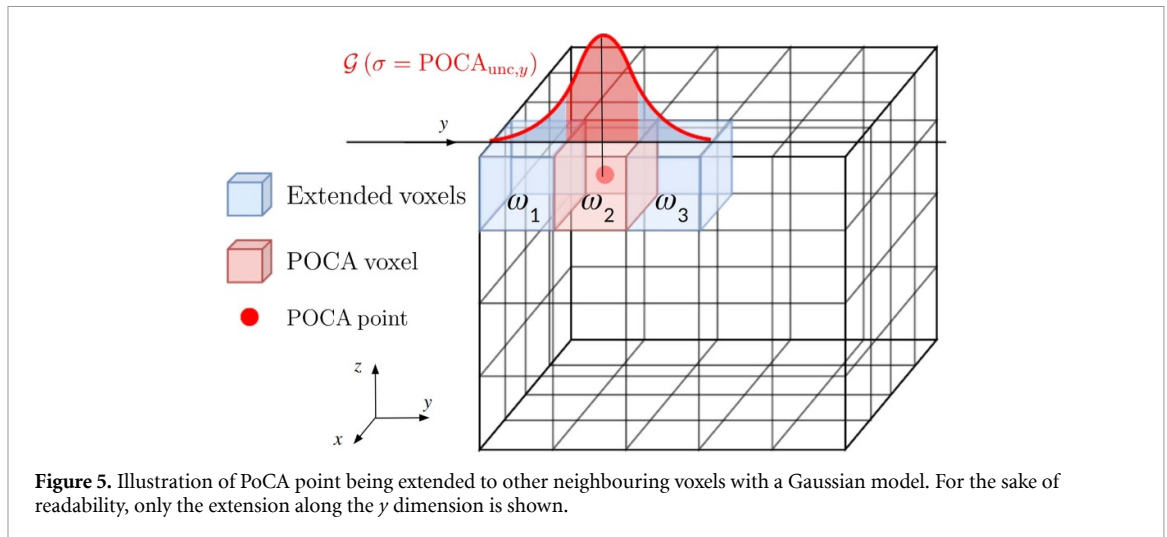
Figure 5. Illustration of PoCA point being extended to other neighbouring voxels with a Gaussian model. For the sake of readability, only the extension along the y dimension is shown.

presented in this paper take this approach. TomOpt implements more advanced methods such as an Expectation-Maximisation algorithm, where the strong (and often unrealistic) assumption of a single scattering centre is dropped and all most likely paths of the muon are taken into account with a proper weight [35].

Alternatively, deep-learning based methods can be applied to muon-level data. Here, latent-space representations of the relevant muons in each voxel can be learnt using graph-neural-networks that are able to learn their own graphs in a latent-coordinate-space, such as those proposed in [36]. These representations can then be refined per voxel, based on the surround voxels through a learnable higher-level graph. An initial investigation of this is presented in [37]: the representation in each voxel is then refined based on the surrounding voxels. More simply, deep-learning methods can be used as a second stage to X_0 inference, to conveniently convert the dimensionality of voxel-level data to volume-wise predictions. Convolutional neural networks are appropriate for this task, as demonstrated during the Second MODE Workshop data-challenge [30].

2.6. Loss functions and optimisation

Similar to the volume inference, the most suitable loss function for the optimisation will depend on the muon tomography task being performed, e.g. cross-entropy for classification tasks. The total loss function, however, should not only consider the performance of the detector, but also its budget, for example the cost in currency units that the detector would take to build given the surface area of its detectors. The cost could enter directly into the loss via a scaling coefficient, however it is more likely that the user already has in mind a particular target budget, and wants to design the best detector that gets close to this budget.

TomOpt implements two approaches: fixed-budget, where the user specifies the maximum budget allowed and the optimisation is constrained to not surpassing that value; and budget-penalisation, where the user specifies a loose target budget.

In the budget-penalisation case, a cost component of the loss added according to a function that is low up to the target budget, and then begins to rapidly increase if the detector exceeds the target budget. In this approach, the detector design can go over budget, but only if the increase in performance vastly surpasses the budget excess, as exemplified in figure 6. This mode introduces hyper-parameters for the weight of the new component in the total loss, and the functional form of the penalisation given the cost and budget.

Alternatively, in 'fixed-budget optimisation mode', the user still specifies a total and fixed cost for the detector, and a fractional assignment of this cost to each panel is learnt, starting from a uniform assignment. The panels still learn a separate xy span, however the the area of the panel is then rescaled such that the cost of the panel is equal to its assigned fraction of the cost of the detector. This mode of optimisation has two main advantages: the user is able to ensure that they meet their budget exactly; there is no balancing coefficients required in the loss function between performance and cost, the loss is purely based on performance.

Once the value of the loss function has been computed, the analytic effects of each detector parameter can be computed through back-propagation [38–40] of the the derivative of the loss through the inference and hit-recording steps. The parameters may then be updated via stepping in the direction of steepest descent of the loss surface. For this process, TomOpt leverages the standard optimisers built into PyTorch (SGD [41], Adam [42], etc).
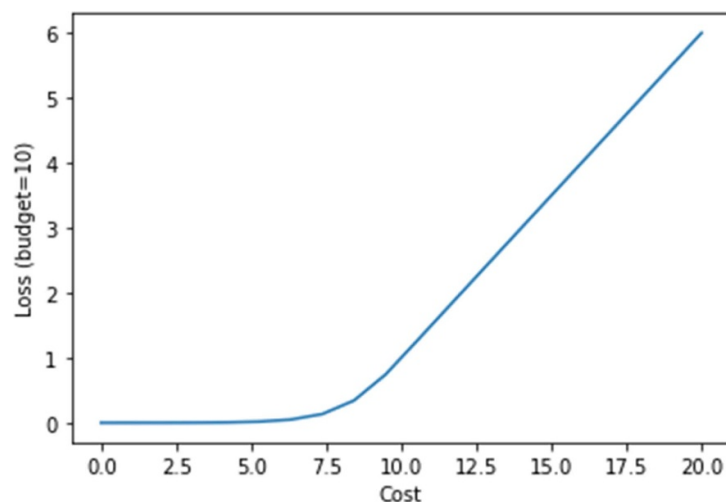
Figure 6. Example of the cost component of the loss as a function of detector cost, for a target budget of 10 A.U.. A sigmoid component below the budget provides a slowly increasing gradient as the cost approaches the target budget, beyond which the loss increases linearly. The function is fully differentiable.

Training is supported by a callback system, which allows aspects of the modules and the training procedure to be adjusted without requiring new modules to be written.

2.7. Update cycle

When computing a value for the loss function, it is important to ensure that it is computed under conditions that are generally representative of those that can be encountered during actual operation, in terms of e.g. number of muons and passive volumes. In effect, this means that optimisation should be performed using a range of passive volumes (whether generated on demand, or supplied by the user as a dataset), and that these should each be scanned using a number of muons that would be reasonable to expect given the position and exposure time of the proposed detector. Additionally, rather than computing the loss using just a single passive volume, it should instead be averaged over many passive volumes in order to provide a more generalising update to the detector parameters.

The update cycle in TOMOPT is performed thus in the following way:

1. Figure 7(c): scanning of a given passive volume involves passing a batch of muons through the volume and inferring their tracks. For computational reasons, the total number of muons may be split into smaller batches. Once all muons have been processed, the volume predictions can be made and the loss computed;
2. Figure 7(b): this process is repeated for a batch of volumes, and the sum (or average) of their losses can be used to update the detector parameters. This is referred to as an *epoch*, in keeping with Machine Learning terminology[20];
3. Figure 7(a): the fit runs for a specified number of epochs before finishing. It is also possible to run epochs in validation mode, where the detector is not updated: the detector panels can be therefore be considered, more realistically, as discrete detector panels rather than with an associated distributed resolution and efficiency, as discussed in section 2.3.
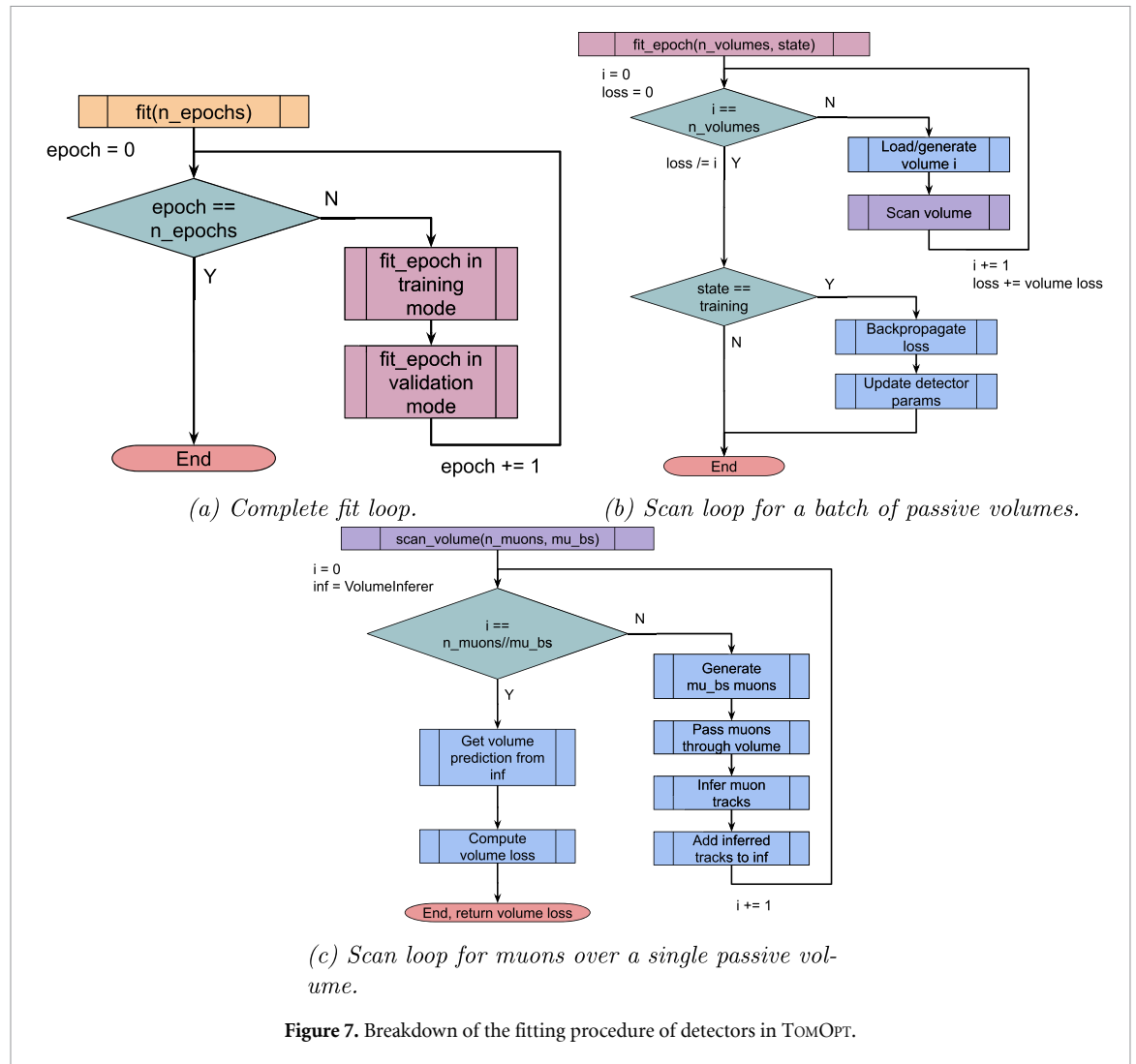
The figures 21 to 23 in appendix show an overview of each of the steps presented above. The main variables and learnable parameters are presented in table 2.

2.8. Current limitations

2.8.1. Muon propagation through matter

As described in section 2.2, the implementation of multiple Coulomb scattering is rather straight-forward since the computation efficiency is prioritised over precision. Scattering angle and displacement distributions are assumed to be a 100% Gaussian, where in reality only 98% of the bulk is Gaussian and the tails follow a $1/\theta^4$ law corresponding to large scattering events. In addition, The PDG model used is not designed for a

---

[20] This assumes that volumes are being generated. When working with a dataset of volumes, instead, this process can be repeated several times using a volume batch size that is less than the total number of volumes in the dataset.

(a) *Complete fit loop.*   (b) *Scan loop for a batch of passive volumes.*

(c) *Scan loop for muons over a single passive volume.*

Figure 7. Breakdown of the fitting procedure of detectors in TOMOPT.

step-by-step algorithm: when it is applied to such an algorithm, it results in underestimating the scattering and displacement amplitudes more and more with each step, as shown in figure 8.

It becomes clear that the choice of step length *dr* affects the precision of the multiple coulomb scattering, and thus has to be chosen carefully. In this context, it is advised to choose a step length *dr* which represent about 1% of the passive volume size.

In addition to multiple Coulomb scattering, muons lose energy through ionisation thus sub-GeV muons might decay in the passive volume. Energy loss has not been implemented in TOMOPT yet, which results in a overestimation of the detected muon flux. Both of these limitations could be overcome by building a parametric model of GEANT 4. Creating a data-base of muon energy loss, scattering and displacement for various material and at several energy scales from GEANT 4 simulations, TOMOPT could simulate muon transport more accurately and effortlessly.
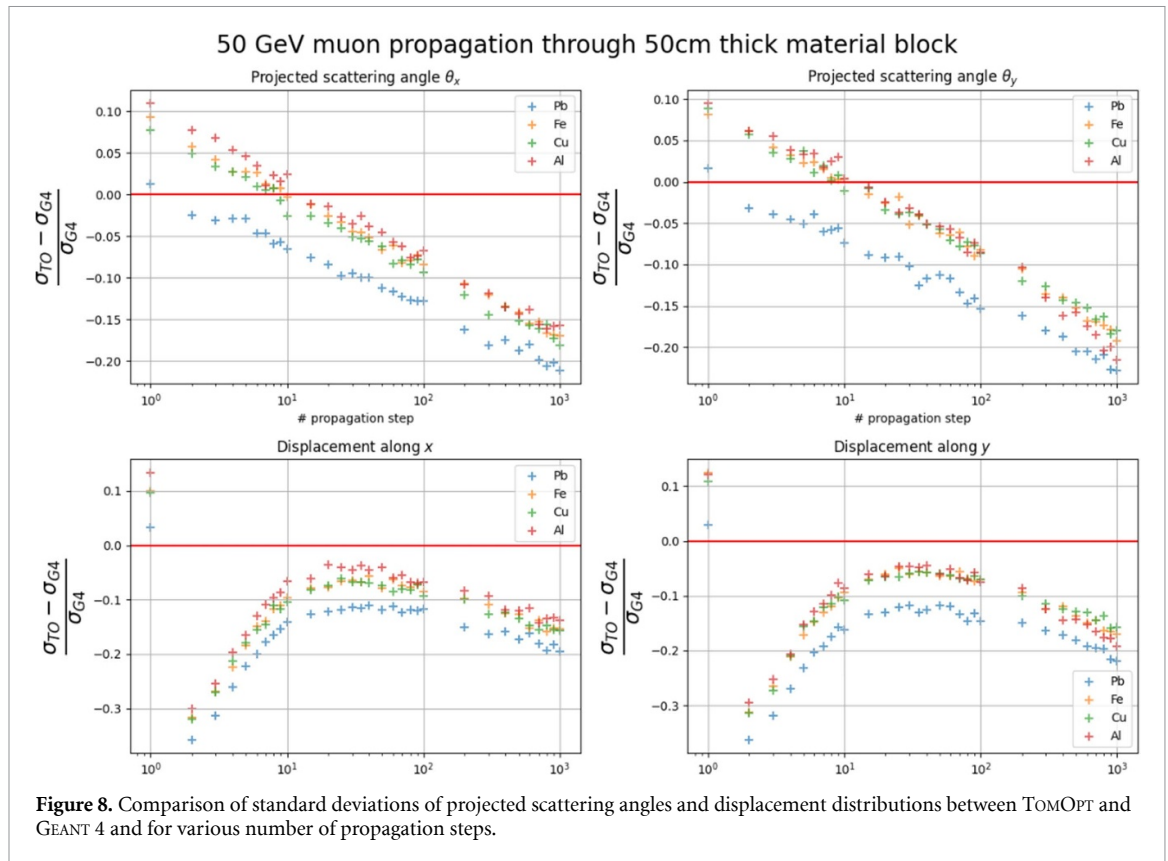
2.8.2. Detector and cost modeling

TOMOPT's ambition is to guide the user through the choice of detector technology and not to provide accurately simulation of particle detection, signal generation or light transport. Thus, the difference between various detector technologies will be implemented at the level of cost modelling, meaning how the cost scales with the active detection surface and number of readout channel. Such features have not been implemented yet, but are planned to be included in future updates.

Regarding detector modelling, TOMOPT is still limited to horizontal panels with continuous spatial resolution and uniform granularity.This limits the phase space of possible detector geometry.

2.8.3. Volume inference

For now, TOMOPT only provides a basic implementation of radiation length inference based on the inversion of the PDG scattering model utilising the Point of Closest Approach. The inversion of the scattering model

Figure 8. Comparison of standard deviations of projected scattering angles and displacement distributions between TomOpt and Geant 4 and for various number of propagation steps.

equation (10) assumes muon momentum to be known with a 100% precision, which is quite unrealistic in the context of a muon scattering tomography experiment. If a momentum estimation is available, the user can still modify the inversion model and replace the nominal muon momentum by a the estimated momentum. If not, it can be replaced by the RMS computed over the whole cosmic muon momentum spectrum.

3. Benchmark: fill-level estimation at metal refinery

The following section describes a benchmark involving the estimation of the fill level at a metal refinery, and aims at demonstrating a typical inference and optimisation chain using TomOpt. After evaluating the performance of an initial sub-optimal detector configuration, the latter will be optimised, and its performance will be compared to two human baseline detector configurations.

3.1. Description

Furnace ladles are structures used to transport melted metal produced in an industrial plant, as shown in figure 9. A description of a generic industrial furnace ladle model used to simulate muography measurements can be found in [43]. A typical example of the part of the manufacturing process which involves the use of the ladle could be described in a general way as follows: firstly the metal is melted in a furnace, then it is transported with a furnace ladle, and finally it is poured into moulds to create metal parts with a defined shape. This parts can later be post-processed, in order to get the final product.

In this process, the amount of melted metal inside the furnace is a key parameter. The lack of metal in the ladle can result in moulds that are not completely filled. Therefore, they may not be used to produce the final product correctly. On the other hand, an excessive charge of metal in the ladle leaves remnants and scraps, causing problems and losses in the manufacturing process. It has to be mentioned that usually, in the top of the melted metal inside the ladle, there is a lighter layer of slag produced in the melting process itself. This layer, prevents the use of optical methods to determine the liquid metal level. The resolution needed by the industry to optimise the manufacturing process is of the order of a few centimetres, although it depends on the distinctive features of each case.

Figure 9. Example of a furnace ladle used in foundry industry. Image by Boris Bukovský, via Pixabay.
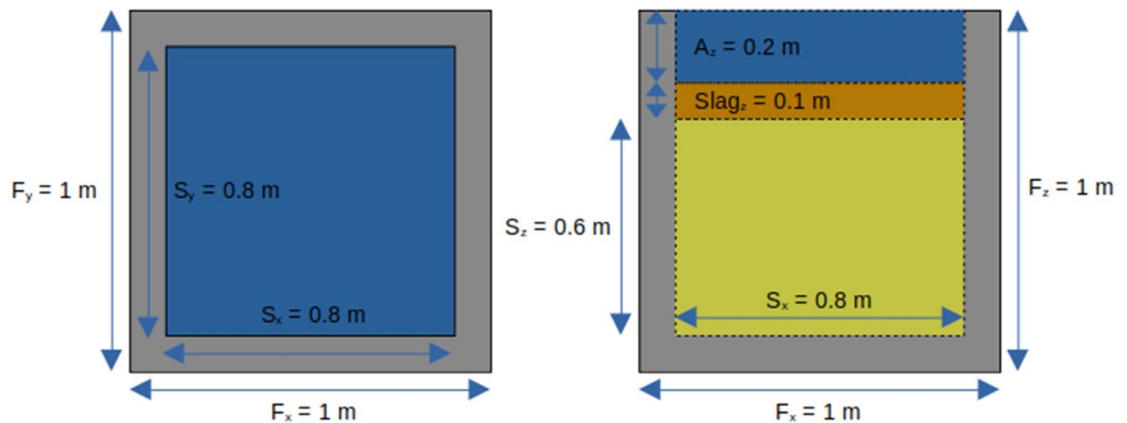

Figure 10. Top view (left), and front view (right) of the rectangular furnace ladle simulated. It is composed of walls (grey), liquid steel (yellow), slag (orange) and the air volume (blue) over the liquid steel and slag.
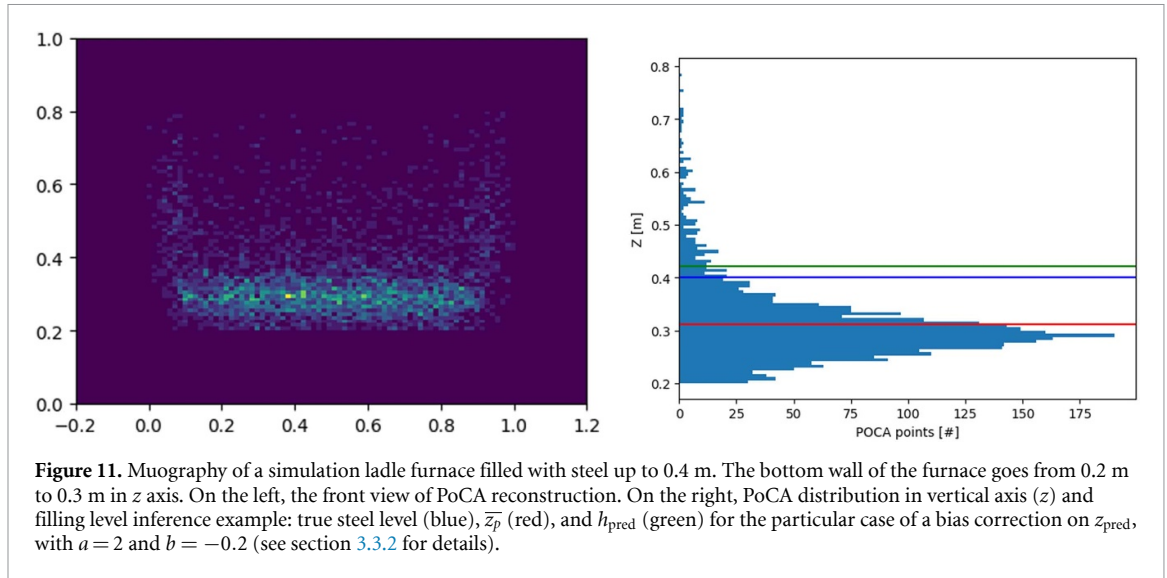
Table 1. Density, radiation length and composition of the simulated furnace ladle materials.

Material	Density, d (g cm^{-2})	Radiation length, X_0 (cm)	Composition
Solid steel	7.818	1.782	Fe (99%), C (1%)
Liquid steel	7.000	1.991	Fe (99%), C (1%)
Slag	2.700	8.211	CaO (53.5%), Al$_2$O$_3$ (33.5%)
Air	0.001 205	30 390	N$_2$ (78%), O$_2$ (20%)

3.2. Detector and volume

Active layers, i.e. the detector panels, are located above and below the measurement target, in an usual scattering muography configuration. At least two panels at different z locations providing $x - y$ positions of muons are required per tracking detector (upper, and lower detectors). As specified in section 2.3, they are modelled considering several parameters, such as position resolution σ_{max}, detection efficiency ϵ_{max}, and cost per unit area.

Within the volume of interest between the detectors, a rectangular furnace ladle filled with liquid steel has been simulated (passive volume). In the figure 10(a) sketch of the furnace ladle is shown. The grey parts correspond to ladle walls, made of solid steel, the molten steel is represented in yellow, the slag layer on top of it in orange, and the remaining void volume over them in blue. The definition of materials is detailed in table 1.

Figure 11. Muography of a simulation ladle furnace filled with steel up to 0.4 m. The bottom wall of the furnace goes from 0.2 m to 0.3 m in *z* axis. On the left, the front view of PoCA reconstruction. On the right, PoCA distribution in vertical axis (*z*) and filling level inference example: true steel level (blue), $\overline{z_p}$ (red), and h_{pred} (green) for the particular case of a bias correction on z_{pred}, with $a = 2$ and $b = -0.2$ (see section 3.3.2 for details).

3.3. Inference method and loss function

The aim of this task is to design a detector that can provide an accurate measurement for the fill-heights of furnaces. This is essentially a single-variable regression task, and a natural choice for the loss function could be the mean squared-error (MSE) between the predicted and true fill-heights. However, as will be discussed below, MSE is not necessarily the most suitable loss for optimisation of this task.

3.3.1. Inference

In the spirit of an end-to-end optimisation pipeline, the inference must output the fill-height prediction and receive optimisation loss gradients based on this. In developing a suitable inference method, we noticed that the mean value of the *z*-positions of PoCAs ($\overline{z_p}$) was monotonic with the fill-height, due to hard scatterings occurring inside the liquid-metal part of the furnace volume, rather than the air and slag regions. This meant that it could serve as an indicator of the mass density distribution centre in the vertical Z axis (see figure 11). As a benefit, this inference approach does not require computing X_0 predictions for the voxels, and thus is more computationally efficient.

This can be adapted to an inference method suitable for detector optimisation by computing a weighted average of the PoCA *z*-positions, according to the muon-trajectory properties. The basis for this weight is the muon-trajectory efficiency (ϵ_p), divided by the squared uncertainty of its PoCA *z*-position ($(\Delta z_p)^2$).

However, we also observed that PoCAs located in the furnace walls could bias the prediction, and reduce precision. Additionally, we know we have uncertainties on the *xy* positions of the PoCAs. In a similar manner to the X_0 inference described in section 2.5, we define Gaussian distributions in *xy* centred on the PoCAs and scaling with their *xy* uncertainties, and integrate them over each voxel in the transverse area of the furnace. We then use a pair of double-sigmoid-based functions in *xy* to ascribe a weight to each voxel, that smoothly decreases to zero at the furnace walls, and maximises to one at the centre. This is used to multiply the integral in each voxel per PoCA. The *xy* weight of each PoCA ($w_{xy,p}$) is defined as sum of its weighted voxel integrals, and this enters as a multiplicative factor of the overall PoCA weight:

$$w_p = \frac{\epsilon_p}{\left(\Delta z_p\right)^2} w_{xy,p}. \tag{11}$$

This weight has the effect of emphasising the *z* positions of PoCAs that are well measured in *z*, located near the centre of the furnace, and have a high trajectory efficiency.

The predicted height is then taken as the weighted average of PoCA *z* positions:

$$z_{\text{pred.}} = \frac{\sum_p w_p z_p}{\sum_p, w_p}. \tag{12}$$

Whilst the mean of PoCA *z* positions is actually more likely to correspond to the centre of the fill-material, this redefinition to the fill-height can be accomplished by the linear transformation described in the next section.

3.3.2. Bias correction

The method described above, does not perfectly predict the fill-height: as mentioned, the mean of PoCA z positions is likely to correspond to the centre of the fill material, but with some finite precision; the PoCA method is inherently biased through its ascription of the entire muon scattering to a single point; additionally, the bias on the prediction is detector-configuration-dependent.

Provided that the biased prediction increases with the true fill-height, bias correction can be performed for a given detector configuration using many volumes of different fill-heights and adjusting the predictions (e.g via a continuous function, or bin-based corrections) such that on average, the predictions match the targets. Knowing this, what we really want the inference and detectors to produce is a prediction that has low in standard deviation when repeated multiple times for the same volume, whilst still increasing in mean value with the true fill-height.

Eventually, though, we will need the detector to be used for actual measurements of fill-heights. At this point, the predictions from the inference must be de-biased as well as possible, in order to correspond to the true targets. The de-biasing can take the form of a parametric function of the predictions, and for simplicity we opt for a linear correction of the form:

$$h_{\text{pred.}} = a \times z_{\text{pred.}} + b, \tag{13}$$

where parameters a and b are computed based on the predictions and targets from a large set of known example volumes at various fill-heights. The effect of this de-biassing can be seen in figure 14.

In order to improve monitoring capabilities during the optimisation, we refine the values of a and b after each update to the detector and display the resulting MSE of de-biased predictions. In order to avoid having to re-scan a large number of volumes, this refinement is performed using predictions and targets collected during the current batch of volumes.

3.3.3. Loss function

Even with the de-biasing, the detectors can still produce predicted heights that contain a residual bias, and optimisation under MSE risks being sensitive to these residual biases, resulting in unstable optimisation. As mentioned, the aim of the detector optimisation is to eventually allow the inference to produce precise predictions that increase with the true fill-height. A loss function that is suitable for this is one that captures the distribution of predictions for each target fill-height, i.e.:

$$\mathcal{L} = \frac{1}{N_l} \sum_l \sigma(p_l), \tag{14}$$

where $\sigma(p_l)$ is the standard deviation of predictions for fill-height index l. If the spread of predictions becomes smaller, the loss will decrease, and vice versa, thus the loss encourages precise measurements of the fill-heights.

However, there is nothing to ensure that the predictions increase in value with the fill-height: given sufficient flexibility in the detector (and potentially inference), it may be possible that the system outputs the same value for all predictions, thus proving zero spread, and so a perfect loss value. In order to penalise such a collapse, we need to include a factor that encourages a large distribution in predictions for different fill-heights, thus the final loss function becomes:

$$\mathcal{L} = \frac{1}{N_l} \frac{\sum_l \sigma(p_l)}{\sigma(\bar{p}_l)}, \tag{15}$$

where $\sigma(\bar{p}_l)$ is the standard deviation of the mean prediction for each fill-height.

If the system were to output the same value for every volume (thus minimising the spread of predictions within each fill-height), the spread of mean predictions for each fill-height would also collapse to zero and act to increase the loss value. The loss function is therefore minimised when the distributions of predictions in each fill-height is small, and when the mean predictions between fill-heights is as large as possible. Thus the predictions become easier to de-bias: if the predicted height for a given volume is within a certain range, we can be more sure it corresponds to a particular true fill-height. We refer to equation (15) as the 'spread over range' loss function.

In the figure 12, the evolution of the proposed loss function with relation to the span of detection panels is shown. The centre of the panels in the horizontal axes X and Y, is the same that the one of the furnace ladle. A sharp loss decrease is noticed when detection panel surface is increased within the limits of the furnace ladle interior, where the liquid steel is located, i.e. the muography target. However, if the panel span exceeds the width of the ladle interior, the loss reduction starts to be less noticeable. We found that this effect is due
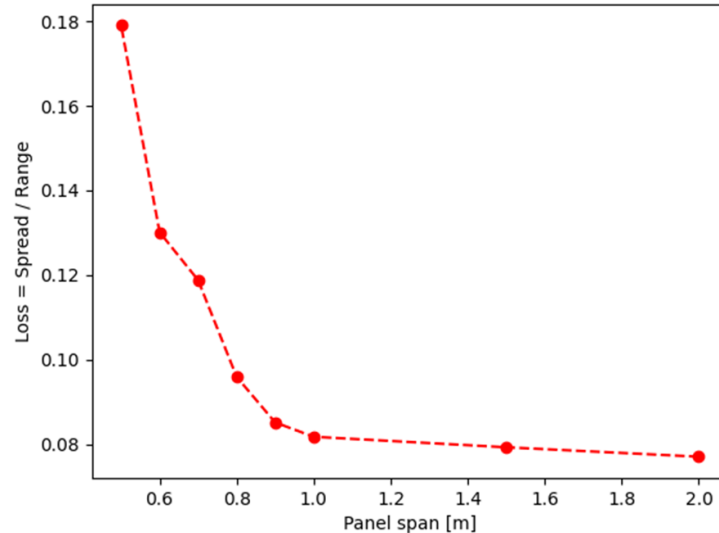
Figure 12. 'Spread over range' loss function tested for different detection surfaces, and using the h_{pred} inference method with unweighted z_{pred} and a linear de-biasing of $a = 2$ and $b = -0.2$.

to the smaller proportion of muons that cross the target when detectors are larger than the furnace interior width, and also because of the contribution of furnace walls to muon deflection. Their presence modifies the density distribution in z axis, distorting our predictions, although this effect is smoothed by the weighted PoCA inference (see section 2.5). Either way, the 'spread over range' loss is continuously and endlessly reduced when increasing the detector surface, the expected behaviour regarding the performance of a muography, since a larger surface allows to collect a higher number of meaningful muons, even if they traverse the target in a transversal direction. This generally improves the precision of inference algorithms. Note that in the 'spread over range' part of the loss function is not included the loss component of detection surface cost (see section 2.6).

3.4. Optimisation process and results
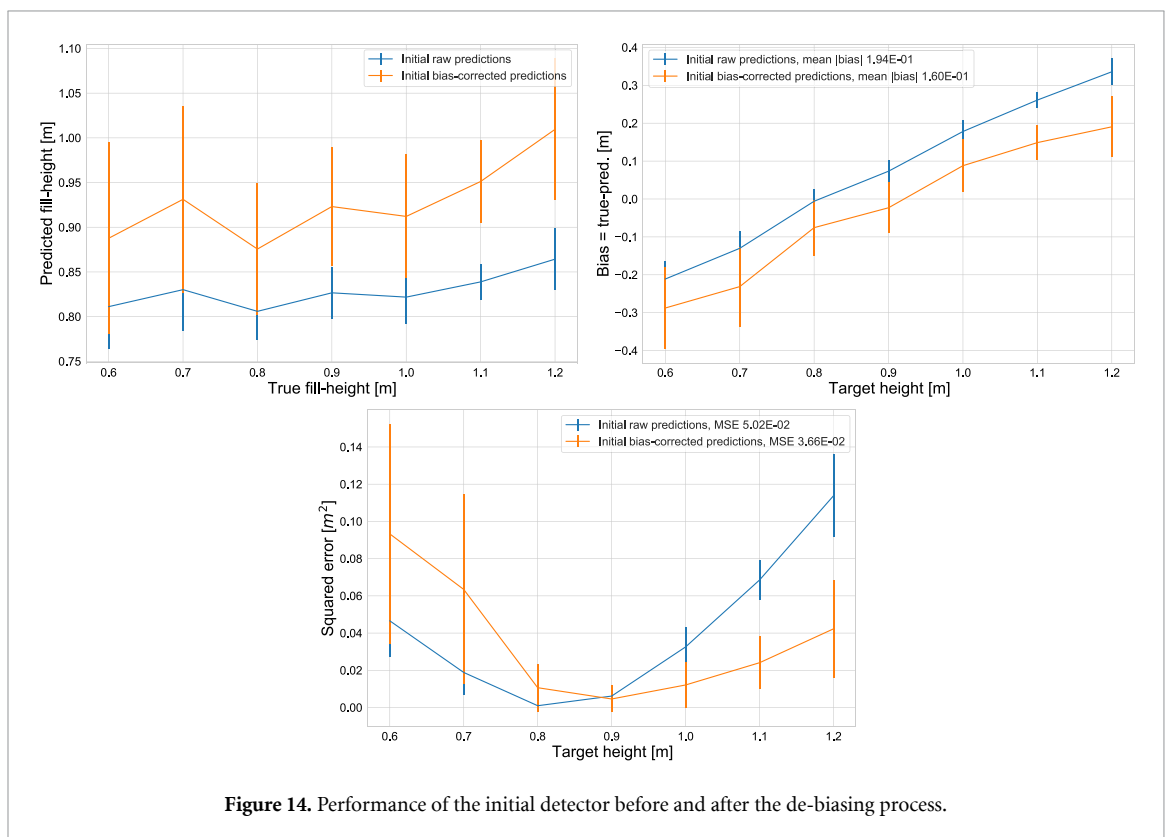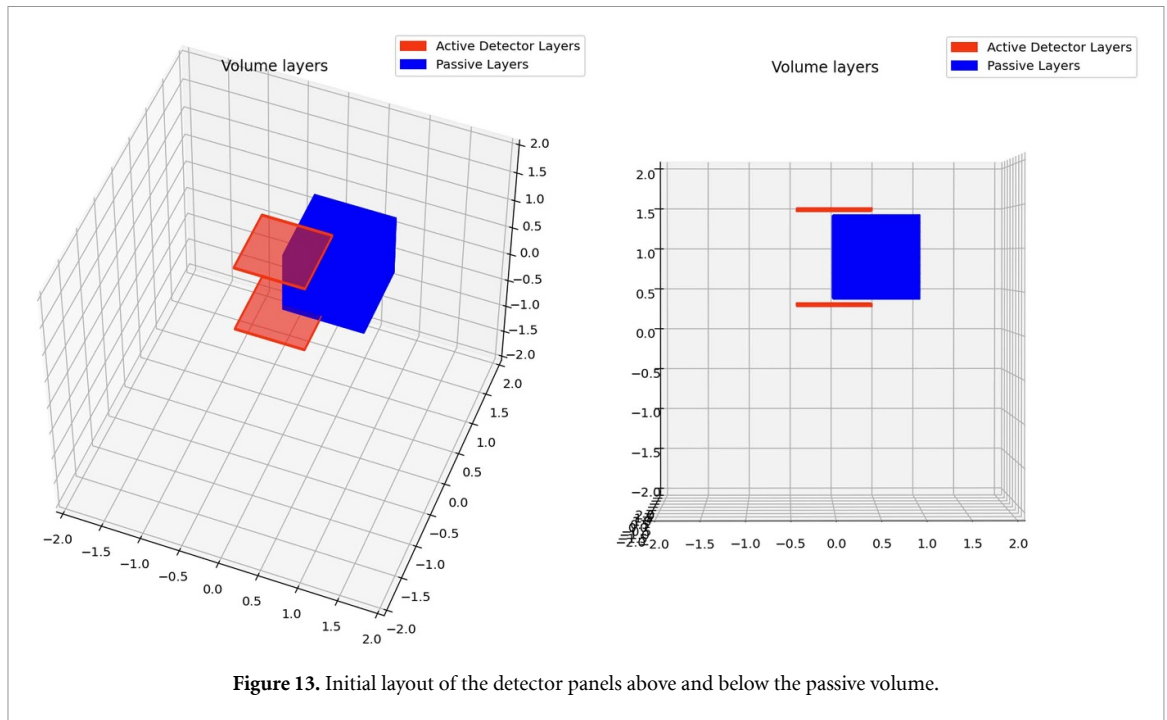
3.4.1. Passive-volume data

Using 10 cm voxels, we generate volumes at eight different fill-heights to simulate from ladles that are almost empty to ones that are almost full. In every case, we add a 10 cm layer of slag on top. When optimising the detector, each of these volumes will be evaluated five times (40 volumes per *volume batch*) – whilst the volumes will always be the same, stochasticity in prediction will still be present due to the muon generation. When evaluating the performance of the detectors, each volume is evaluated 12 times. 1000 muons will be used per evaluation.

3.4.2. Initial configuration

In order to test the optimisation capabilities of TomOpt, we initially set the detector in a knowingly sub-optimal configuration: one in which the detector panels are off-set from the passive volume in $x - y$, thus lowering the flux muons that are well detected, and interact with the passive volume; and closely space in z, thus increasing the uncertainty in PoCA reconstruction.

The detector itself (illustrated in figure 13) consists of four equally sized square panel above the passive volume, and another four panels below. These panels begin centred in $x - y$ over the corner of the passive volume, and with an (unrealistic) separation in z of 1 cm between each panel. There is a 5 cm separation between the passive volume to the nearest panels. The total cost of the detector is fixed at six arbitrary units, and panels cost one arbitrary unit per m^{-2}. The available budget is initially distributed equally to all panels, thus each panel measures 86.6 cm in both x and y length. Panels are defined to have nominal efficiencies of 90 %, and $x - y$ resolutions of 0.1 mm.

The initial performance for the detector is shown in figure 14. The 'raw' predictions are the result of the inference with no de-biasing. We can observe a slight slope to their values w.r.t. the fill-height, however the large spread in predictions for the same target, means that it is difficult to distinguish different fills. The de-biasing process via a linear correction is able slightly reduce the MSE, but cannot do much more that shift

Figure 13. Initial layout of the detector panels above and below the passive volume.

Figure 14. Performance of the initial detector before and after the de-biasing process.

the predictions. Averaging across the range of fill-heights considered the initial detector has a mean MSE of 0.037 m^2, and an average absolute error on fill-level prediction of 16 cm.
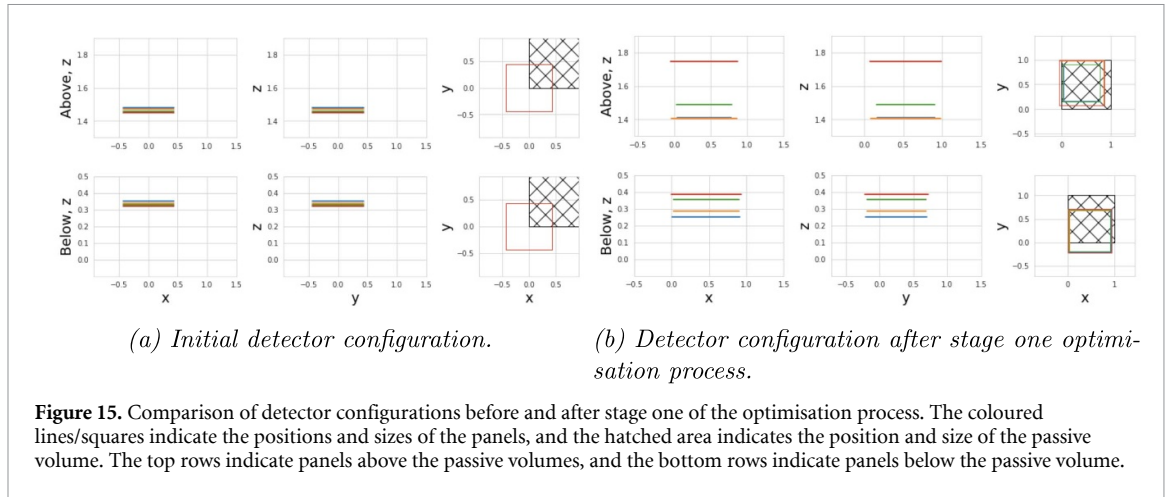
Our intuitive expectation is that from this position, the panels should move to cover the centre of the passive volume, and increase in z-separation. There is, however, a trade-off in increasing the z-separation: whilst it generally increases the PoCA resolution, it also decreases the hit-reconstruction for muons entering or leaving the passive volume at large angles. We therefore expect the optimisation to find a natural point at which the improved resolution is balanced by the decreased high-angle-muon flux. Additionally, we have no confident intuition over how the panels should be resized to best improve performance given a constant budget.

3.4.3. Optimisation callbacks

As mentioned in section 2.6, TomOpt has a callback system that allows the user to augment the basic optimisation loop to include different functionality, and to help stabilise the optimisation process. For this task, the following callbacks are used:

- MuonResampler—When muons are generated, they may not ever interact with the passive volume, and so form a source of *background*, which would ideally be removed by some filtering process. In the interests of simplicity, we are not considering background processes here[21]. The MuonResampler callback acts to regenerate muons until all muons in the batch will interact with the passive volume at some point. This therefore improves the generation efficiency.

- LinearCorrection—The de-biasing process mentioned in section 3.3.2 is implemented as a callback that can modify in-place the outputs of the inference module, collect raw predictions and targets during the volume batch, and then re-fit its correction after each detector update.

- SpreadRangeLoss—By default, TomOpt computes the loss per volume and tracks a running sum of losses over the volume batch. However the 'spread over range' loss can only be computed for a population of volume predictions. This callback collects the predictions over the course of the volume batch and then sets the value of the loss prior to back-propagation.

- NoMoreNaNs—Optimisations can take several hours, and whilst TomOpt is fairly resilient, given the number of muons being propagated and the complexity of the reconstruction and inference process, it is inevitable that invalid (*NaN*) gradients are occasionally produced. These would otherwise kill the optimisation process, however after back-propagation but before the parameter updates, this callback goes through the parameters and sets any *NaN* gradients to be zero.

- OptConfig—The learning rate of optimisers in gradient descent is one of their most important parameters, however it can be difficult to set suitably with limited *a priori* knowledge. Whilst techniques such as the 'learning rate range finder test' from [44] have been developed, they are designed for DNNs, where the learnable parameters lack a physical meaning. Here, our parameters do have physical meaning (e.g. the detector position is a learnable parameter measured in metres). We are thus able to do something better: we can specify the desired change per update of parameters in their physical units, and over the course of a *warm-up phase* (in which all parameters are fixed) compute the average gradient received by each parameter. From these averages, we can then compute the learning rate that is likely to produce the desired update rate. As an example we can say that we want the panel *xy* positions to update at a rate of 1 cm per iteration. We can then monitor the median gradients for a few iterations and then set an appropriate learning rate that results in a change of about 1 cm per iteration according to $\gamma = 1\,\mathrm{cm}/\bar{\nabla}_{xy-\mathrm{pos.}}\mathcal{L}$.

- PanelCentreing—In order to help stabilise the optimisation in TomOpt, we use this callback: after panel *xy* positions have been updated, it goes through and for panels above and below the passive volume separately, it computes the mean position in $x - y$ of panels and then sets panels to that position. Panels below the passive volume can still have different positions to those above, but all panels above and below each have a common $x - y$ position. Whilst this does limit the originality of detector configuration that TomOpt can explore, it is currently necessary to avoid instabilities and divergences in the optimisation.

- PanelMetricLogger—This provides real-time feedback to the user, and displays the current layout of the detector, the history of the loss values, and other diagnostic information.

- EpochSave—This automatically saves the state of the detector after every update, allowing the user to manually pick the best performing detector in case of performance divergence, or to recover in case of errors

- OneCycle—This is a learning rate scheduler that adapts the learning rate (and optionally momentum/β_1 coefficients) or the optimisers according to the 1-cycle schedule described in [45, 46]. This is mainly used to allow the learning rate to gradually decrease over the course of the optimisation, allowing the detector configuration to better converge, whilst allowing larger updates near the beginning to training to move quickly from sub-optimal to near-optimal positions.

- SigmoidPanelSmoothnessSchedule—as mentioned in section 2.3.1, the physical panels are approximated differentiably during the *training epochs* using a sigmoid-based model. It was also suggested that different values of the *smoothness* of the model allows different parameters of the detector panels mode or less of an effect on the loss. This callback allows the smoothness to be decreased across the optimisation process, meaning that close to the end of the training, the panels better approximate physical detectors, and the loss is more sensitive to the exact *z* position of the panels.

---

[21] Given the small number of muons used, relative to the incoming flux and acquisition time, a highly pure sample of muons is expected to be acquirable by even a low-efficiency trigger/filter.

(a) *Initial detector configuration.*   (b) *Detector configuration after stage one optimisation process.*

Figure 15. Comparison of detector configurations before and after stage one of the optimisation process. The coloured lines/squares indicate the positions and sizes of the panels, and the hatched area indicates the position and size of the passive volume. The top rows indicate panels above the passive volumes, and the bottom rows indicate panels below the passive volume.

3.4.4. Optimisation process

Five parameter families are considered: xy positions of the detector panels and the z positions of the panels are considered separately for panels above and below the volume (four parameter families); and the budget assignment per panel (one parameter family). Over the course of five epochs, we monitor the gradients received by the trainable parameters in these families and compute nominal learning rates that correspond to position updates at a rate of 1 cm per epoch, and budget weights at a rate of 0.1 (dimensionless).

3.4.4.1. Stage one

Over the course of 10 epochs, the detector is updated, using 1-cycle learning rate annealing to scale the nominal learning rates over the course of the optimisation: initially they increase to allow large updates of the detector, and towards the end the decrease to much smaller values, allowing the panels to converge and stabilise. Additionally, the sigmoid detector model is annealed from an initially smooth configuration to a sharper, more physical distribution of efficiency and resolution. On a 2018 Macbook Pro with an Intel i7, this takes about 30 min to run (no GPU is used).

Figure 15 illustrates the detector configuration before and after this stage of the optimisation process. From this we can see that the panels have indeed moved to be more central in $x - y$, and have expanded in z.

Figure 16 compares the performance of the optimised detector to the initial detector, and indicates a significant improvement in both the precision and bias: now predictions are clearly distinguishable by true fill-height. The optimised detector has a mean MSE of 0.0014 m^2, and an average absolute error on fill-level prediction of 2.8 cm.

3.4.4.2. Stage two

The previous optimisation process showed that TOMOPT can take a poorly designed detector and quickly shift it into a configuration that provides much better performance. In this second stage we will demonstrate that TOMOPT can be used to refine detectors that are already thought to be well performing, by altering only subsets of the total available parameters. To do so, we will use our detector from stage one.

During stage one, we saw that the panels moved to become central in $x - y$. We will take this experimental evidence, along with our domain knowledge of the problem, to assume that the ideal position for the detectors is directly above and below the passive volume. We will manually shift the panel to the centre and fix them there, allowing only the z and budget assignments to slowly change.

Once again, we compute suitable nominal learning rates for our three remaining parameter families by monitoring the gradients they receive for five epochs. Over another 10 epochs we allow the detector to further be refined, this time with smaller update steps. Figure 17 illustrates the detector after this refinement stage, and figure 18 its performance. Mainly we can see that the panels below the passive volume have moved to cover a larger z range. While this could come at the expense of missing some high-angle muons leaving the passive volume, the panels have slightly increased in size by decreasing the intermediate panels above the passive volume. The refined detector has a mean MSE of 0.0011 m^2, and an average absolute error on fill-level prediction of 2.5 cm.
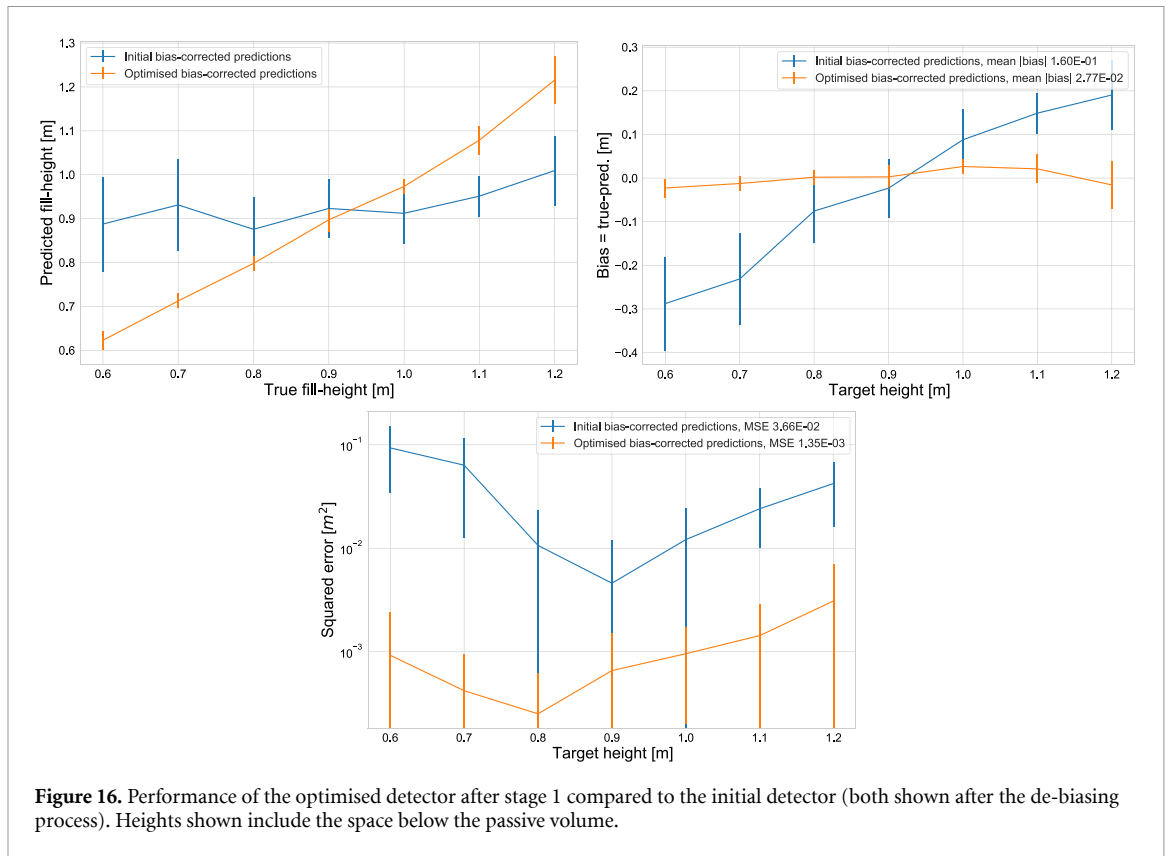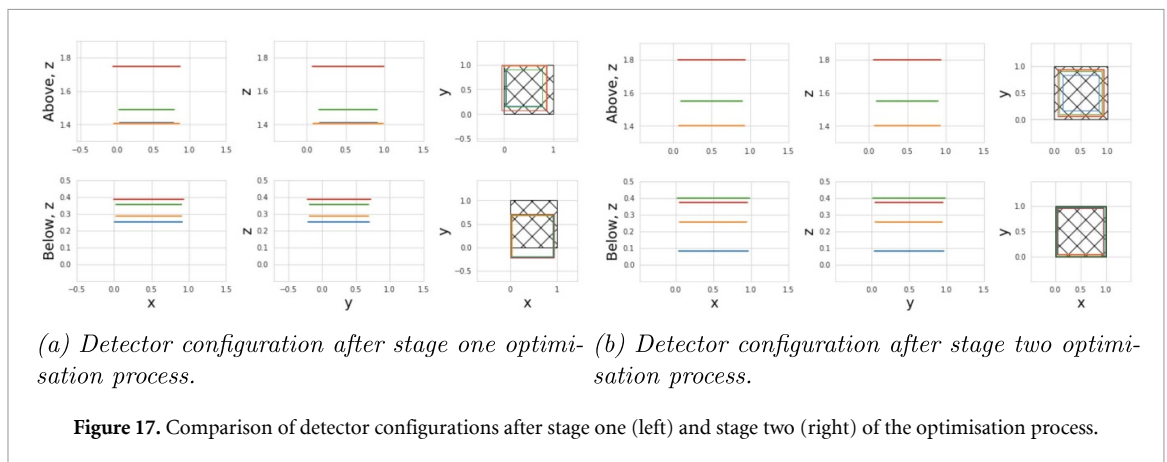
Figure 16. Performance of the optimised detector after stage 1 compared to the initial detector (both shown after the de-biasing process). Heights shown include the space below the passive volume.

(a) Detector configuration after stage one optimisation process.

(b) Detector configuration after stage two optimisation process.

Figure 17. Comparison of detector configurations after stage one (left) and stage two (right) of the optimisation process.

3.5. Comparison to human baselines

Having demonstrated that TOMOPT can successfully optimise detectors, we now want to see whether the resulting detector is good compared to a human-designed detector. For this we will take two baseline configurations:

- Baseline 1: detector panels are placed in pairs with a 5 cm separation between panels, and a 25 cm separate between pairs above and below the passive volume. This configuration aims to maximise the precision of the muon trajectory reconstruction by allowing precise measurements of the muon positions over a long baseline.
- Baseline 2: detector panels are placed with an even 10 cm spacing. Baseline 1 potentially misses some high-angle muons due to the large distance to the second pair panels. This baseline attempt to strike a balance between precision and exposure.
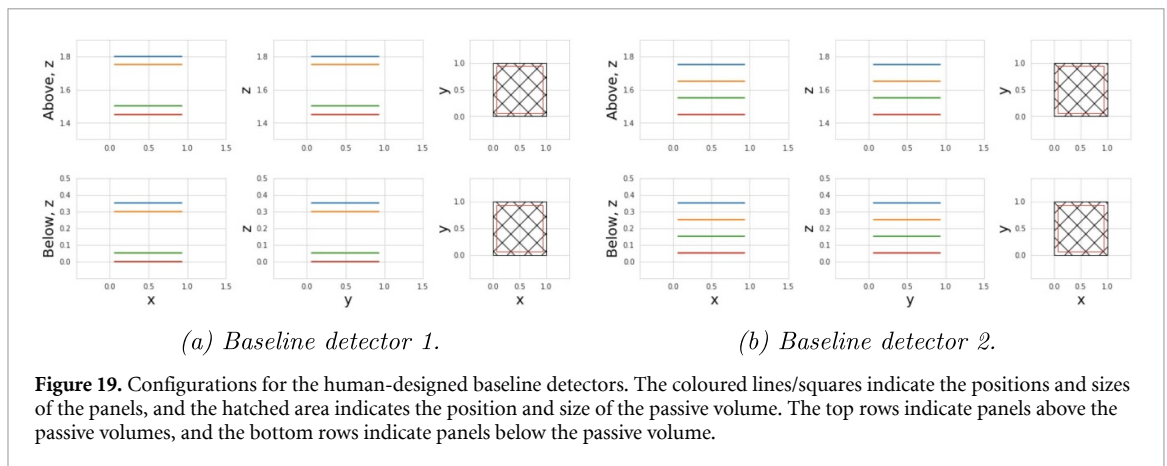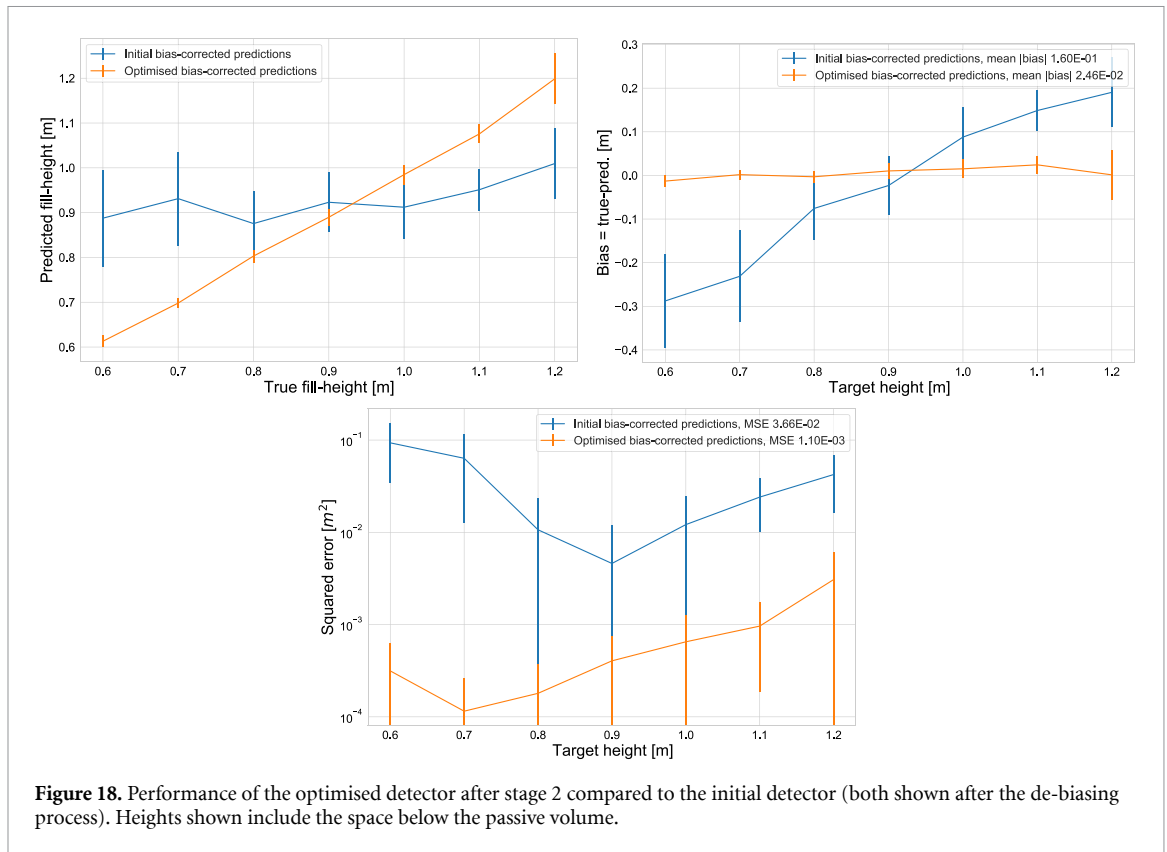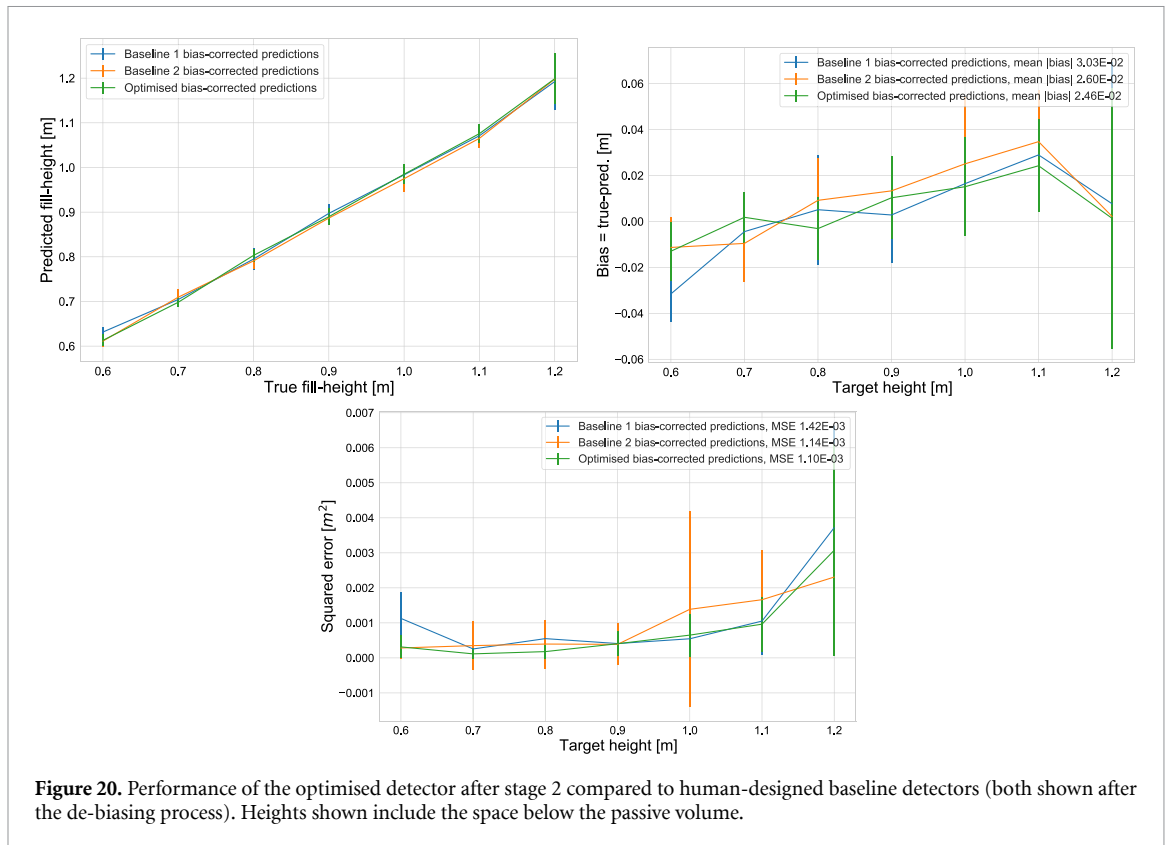
These are illustrated in figure 19.

Figure 18. Performance of the optimised detector after stage 2 compared to the initial detector (both shown after the de-biasing process). Heights shown include the space below the passive volume.

(a) Baseline detector 1.   *(b) Baseline detector 2.*

Figure 19. Configurations for the human-designed baseline detectors. The coloured lines/squares indicate the positions and sizes of the panels, and the hatched area indicates the position and size of the passive volume. The top rows indicate panels above the passive volumes, and the bottom rows indicate panels below the passive volume.

Figure 20 compares the performances of the stage-two optimised detector against the two baseline detectors. Whilst all three detectors offer similar levels of performance, our optimised detector is able to not only match human performance, but slightly and consistently outperforms it.

4. Conclusions

In this document, we have described TomOpt; a system that allows muon-tomography detector designers to specify their end goal numerically as a task-specific loss function, and through the use of automatic differentiation optimise their designs to provide the best possible performance. Since the updates to the detector are made in the presence of both the inference algorithm and end-goal, the designers can be sure that changes made have a genuine impact on the final performance, rather than focussing on optimising task-agnostic proxy metrics (such as track resolution), and hoping that these eventually translate to better actual performance for the task at hand.

We demonstrated the capabilities of TomOpt in a simulation of one of the typical industrial use-cases—that of estimating the fill-level of furnace ladles at a metal refinery. Using task-specific inference and losses, we were able to optimise an initial, poorly designed detector into one that was able to not-only

Figure 20. Performance of the optimised detector after stage 2 compared to human-designed baseline detectors (both shown after the de-biasing process). Heights shown include the space below the passive volume.

match, but slightly outperform two human-designed baseline detectors in terms of fill-height prediction precision. In this example we performed the optimisation such that the detector always conformed to a fixed cost, however alternatively the optimisation could be performed to minimise the detector cost and not exceed a specified budget.

We believe that TOMOPT provides the first demonstration of end-to-end-differentiable and inference-aware optimisation of particle physics instruments, and represents an important step in changing the way we design experimental apparatus. Given the number of free parameters considered here, it is worth mentioning that non-differentiable, black-box optimisers such as Bayesian optimisation (BO) could be well applicable. However given the parameter spaces of larger detectors, such as those at the CERN LHC, where perhaps more than 1000 parameters might be available for control, we believe that differentiable optimisation will be required. With that in mind, our primary of this paper was to demonstrate the viability of measurement-aware detector design in a differentiable pipeline.

In practical application on larger detectors, BO could potentially be used on a reduced parameter set to act as an initial global search to find a decent initialisation, after which full differentiable optimisation can be used on the larger parameter space to refine the detector. Given BO's capability to more easily optimise categorical parameters, this initial search may be additionally beneficial in fixing such parameters prior to differentiable optimisation.

Looking to the immediate future of this work, there are many possible extensions. So far, we have relied on 'classical' inference algorithms, but inference via deep-neural networks are naturally compatible with TOMOPT's requirement on differentiable processes. We will be exploring their use in future publications, as these can help to ease the burden of designing task-specific differentiable inference algorithms. Additionally, as a quick simulator of the physics, TOMOPT can be exploited for developing and testing advanced (non-differentiable) inference algorithms for muon-tomography in general. The available degrees of freedom of detectors, in terms of placement, technology, and modelling needs to be expanded, to allow for e.g. detectors to be placed on the side of the passive volume, or otherwise rotated. Finally, we will be looking to expand the repertoire of benchmark scenarios used for demonstration and testing.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

Appendix

Table 2. Description of variables and learnable parameters from figures 21–23.

| Symbol	Definition	Unit
θ	Muon zenith angle	rad, °
ϕ	Muon azimuthal angle	rad, °
X_0	Radiation length	m
S_{gen}	Muon generation surface	
$\mathrm{d}x, \mathrm{d}y$	Generation surface span in x, y	m
E_{range}	Muon generation energy range	GeV
θ_{range}	Muon generation zenith angle range	GeV
N_{muons}	Number of generated muons	
ϵ_{max}	Panel maximum efficiency	
σ_{max}	Panel maximum spatial resolution in x and y	m
$x_{\text{span},j}, y_{\text{span},j}$	Panel j span in x and y as a learnable parameter	m
x_j, y_j, z_j	Panel j position in x, y, z as a learnable parameter	m
ϵ_{eff}	Reconstructed hit effective uncertainty	
$x_{\text{rec}}, y_{\text{rec}}$	Reconstructed muon hit	m
ϵ_μ	Per-muon efficiency	
ϵ_{batch}	Muon batch efficiency, sum of the per-muon efficiencies	
$\mathcal{L}$	Total loss function	
$\mathcal{L}_{\text{cost}}$	Detector cost loss function	
$\mathcal{L}_{\text{error}}$	Error loss function	
β	Loss function cost weight	
$\mathcal{L}_{\text{volume}}$	Volume loss function	
X^k	Set of all learnable parameters $\{\ldots, (x_j, y_j, z_j, x_{\text{span},j}, y_{\text{span},j}), \ldots\}$ at iteration k	
∇_X	Loss function gradients w.r.t learnable parameters X	
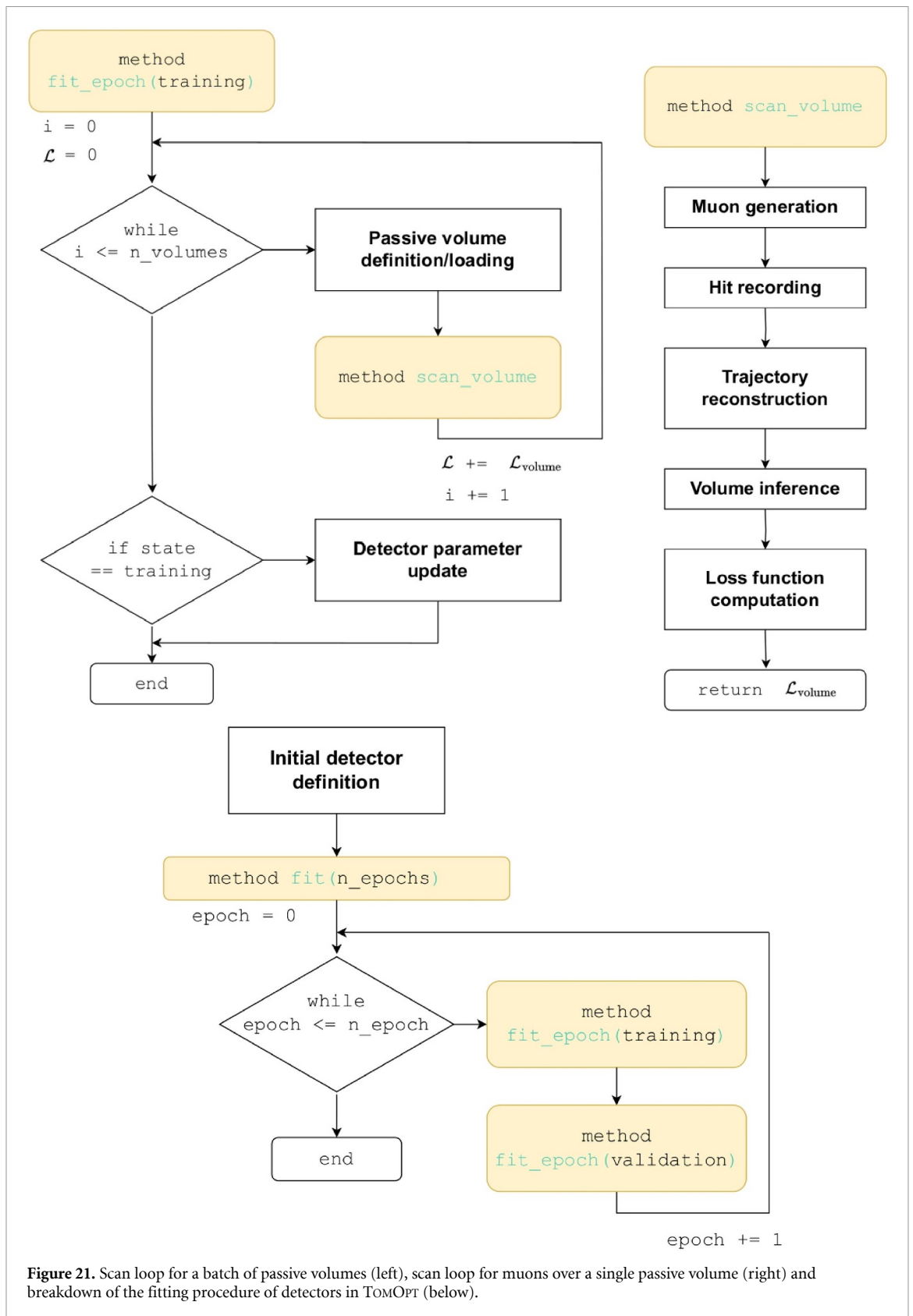
Figure 21. Scan loop for a batch of passive volumes (left), scan loop for muons over a single passive volume (right) and breakdown of the fitting procedure of detectors in TomOpt (below).
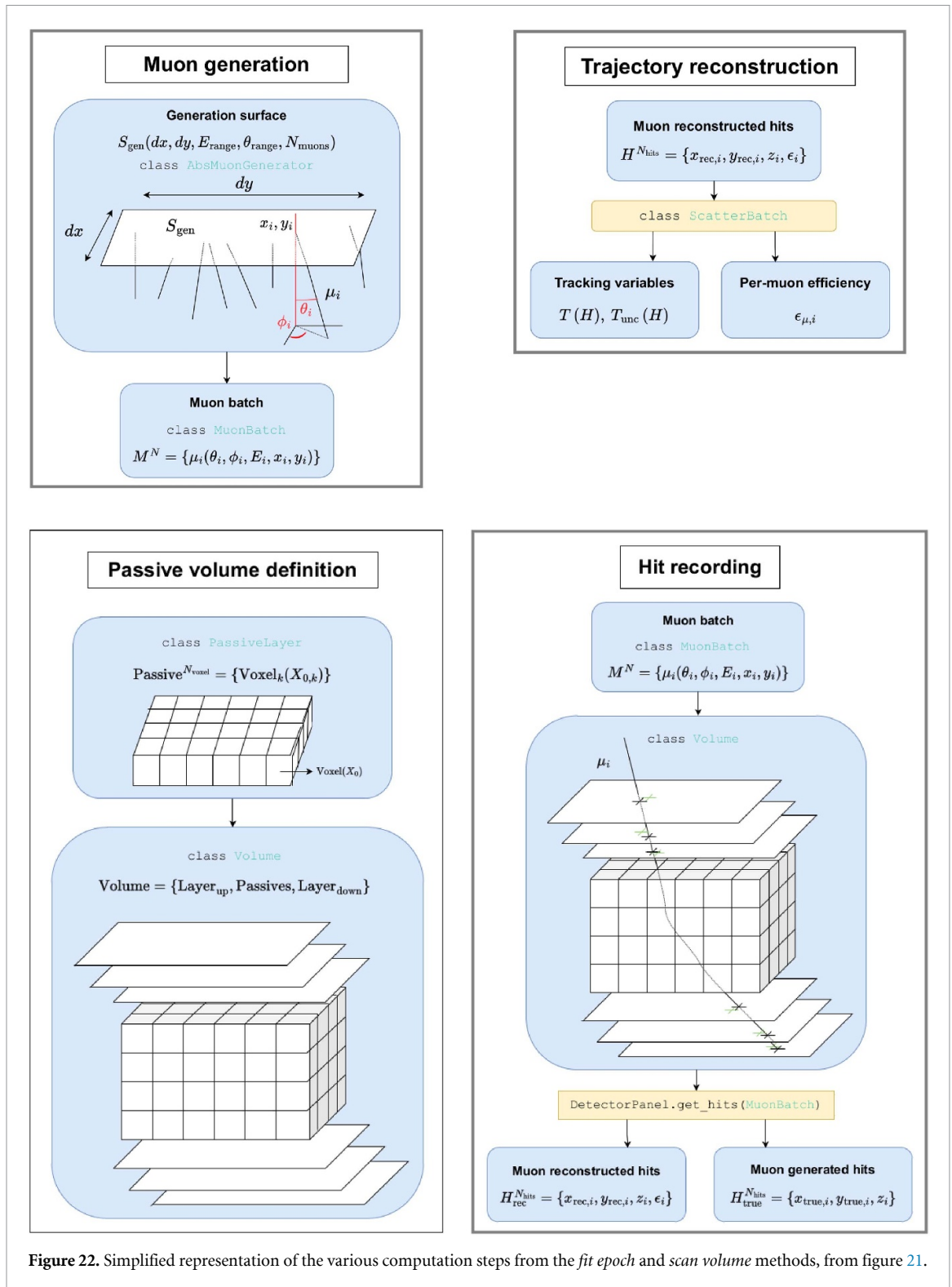
Figure 22. Simplified representation of the various computation steps from the *fit epoch* and *scan volume* methods, from figure 21.
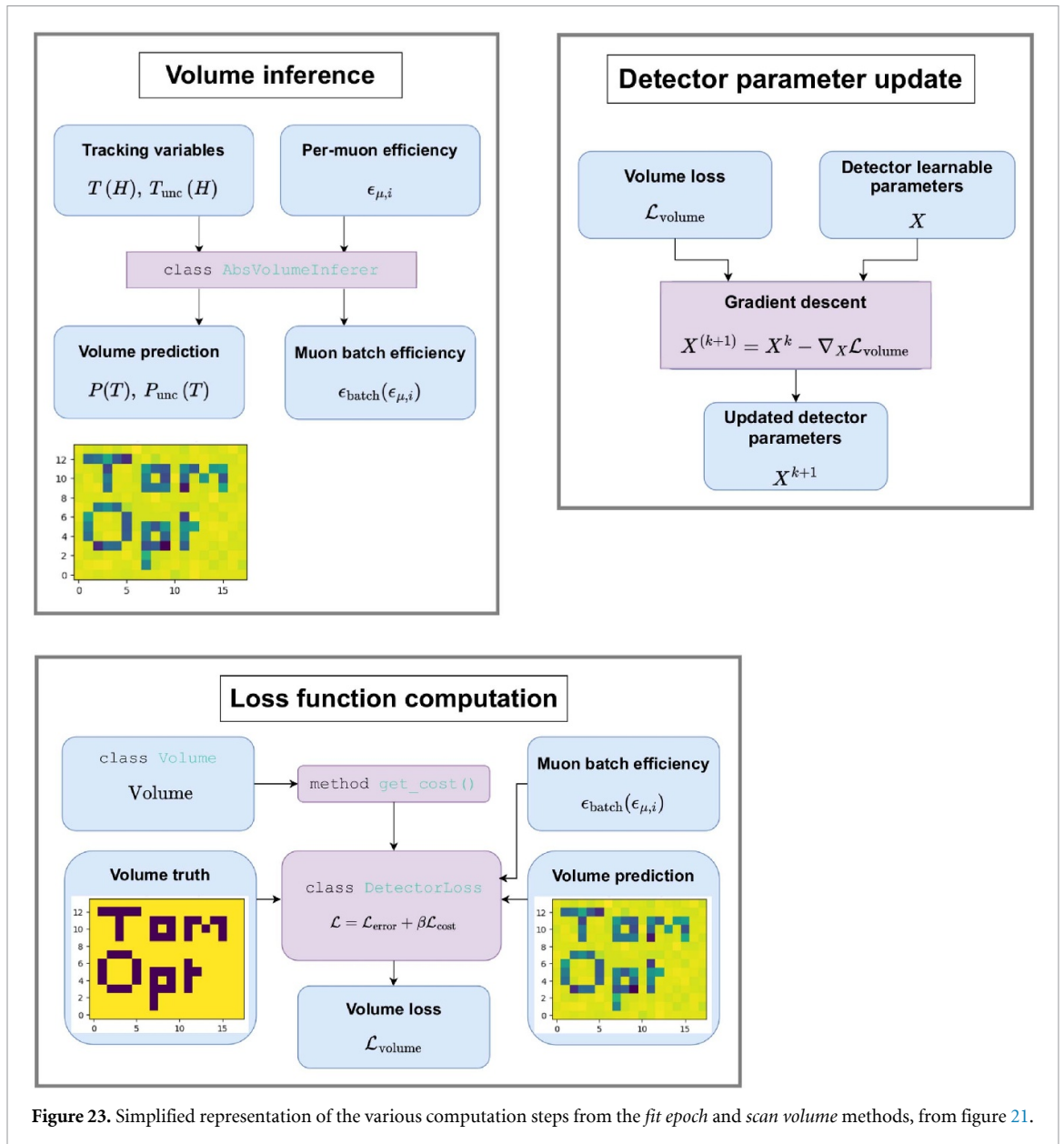
Figure 23. Simplified representation of the various computation steps from the *fit epoch* and *scan volume* methods, from figure 21.

ORCID iDs

Giles C Strong ● https://orcid.org/0000-0002-4640-6108
Maxime Lagrange ● https://orcid.org/0000-0002-9332-8210
Aitor Orio ● https://orcid.org/0000-0003-3150-6488
Florian Bury ● https://orcid.org/0000-0002-3077-2090
Tommaso Dorigo ● https://orcid.org/0000-0002-1659-8727
Andrea Giammanco ● https://orcid.org/0000-0001-9640-8294
Jan Kieseler ● https://orcid.org/0000-0003-1644-7678
Pablo Martínez Ruíz del Árbol ● https://orcid.org/0000-0002-7737-5121
Federico Nardi ● https://orcid.org/0000-0003-4324-2811
Pietro Vischia ● https://orcid.org/0000-0002-7088-8557
Haitham Zaraket ● https://orcid.org/0000-0003-3929-0431

References

[1] George E P 1955 Cosmic rays measure overburden of tunnel *Commonwealth Eng.* **1** 455
[2] Alvarez L W *et al* 1970 Search for hidden chambers in the pyramids *Science* **167** 832
[3] Morishima K *et al* 2017 Discovery of a big void in Khufu's Pyramid by observation of cosmic-ray muons *Nature* **552** 386

[4] Procureur S *et al* 2023 Precise characterization of a corridor-shaped structure in Khufu's Pyramid by observation of cosmic-ray muons *Nat. Commun.* **14** 1144

[5] Oláh L, Tanaka H K and Varga D (eds) 2022 *Muography: Exploring Earth's Subsurface With Elementary Particles* (Wiley)

[6] Rutherford E 1911 The scattering of α and β particles by matter and the structure of the atom *London, Edinburgh Dublin Phil. Mag. J. Sci.* **21** 669

[7] Lynch G R and Dahl O I 1991 Approximations to multiple Coulomb scattering *Nucl. Instrum. Methods Phys. Res.* B **58** 6

[8] Borozdin K N *et al* 2003 Radiographic imaging with cosmic-ray muons *Nature* **422** 277

[9] Barnes S, Georgadze A, Giammanco A, Kiisk M, Kudryavtsev V A, Lagrange M and Pinto O L 2023 Cosmic-ray tomography for border security *Instruments* **7** 13

[10] Weekes M, Alrheli A, Barker D, Kikoła D, Kopp A K, Mhaidra M, Stowell J, Thompson L and Velthuis J J 2021 Material identification in nuclear waste drums using muon scattering tomography and multivariate analysis *J. Instrum.* **16** 05007

[11] Morris C L, Bacon J, Borozdin K, Fabritius J, Miyadera H, Perry J and Sugita T 2014 Horizontal cosmic ray muon radiography for imaging nuclear threats *Nucl. Instrum. Methods Phys. Res.* B **330** 42–46

[12] International Atomic Energy Agency 2022 *Muon Imaging: Present Status and Emerging Applications Iaea Tecdoc 2012* (IAEA)

[13] Mrdja D, Bikit I, Bikit K, Slivka J, Hansman J, Oláh L and Varga D 2016 First cosmic-ray images of bone and soft tissue *Europhys. Lett.* **116** 48003

[14] Bikit I, Mrdja D, Bikit K, Slivka J, Jovancevic N, Oláh L, Hamar G and Varga D 2016 Novel approach to imaging by cosmic-ray muons *Europhys. Lett.* **113** 58001

[15] Yifan Z, Zhi Z, Ming Z, Xuewu W and Ziran Z 2018 Discrimination of drugs and explosives in cargo inspections by applying machine learning in muon tomography *High Power Laser Part. Beams* **30** 086002

[16] Xuan-Tao J, Si-Yuan L, Yu-He H, Kun Z, Jin Z, Xiao-Yu P, Min X and Xiao-Dong W 2022 A novel 4D resolution imaging method for low and medium atomic number objects at the centimeter scale by coincidence detection technique of cosmic-ray muon and its secondary particles *Nucl. Sci. Tech.* **33** 2210–3147

[17] Holma M, Joutsenvaara J and Kuusiniemi P 2022 Trends in publishing muography related research: the situation at the end of 2020 *J. Adv. Instrum. Sci.* **2022** 292

[18] Bonechi L, D'Alessandro R and Giammanco A 2020 Atmospheric muons as an imaging tool *Rev. Phys.* **5** 100038

[19] Gaisser T K 1990 *Cosmic Rays and Particle Physics* (Cambridge University Press)

[20] Guan M, Chu M -C, Cao J, Luk K-B and Yang C 2015 A parametrization of the cosmic-ray muon flux at sea-level

[21] Shukla P and Sankrith S 2018 Energy and angular distributions of atmospheric muons at the earth *Int. J. Mod. Phys.* A **33** 1850175

[22] Workman R L (Particle Data Group Collaboration) 2022 Review of particle physics *Prog. Theor. Exp. Phys.* **2022** 083C01

[23] Moliere G 1948 Theorie der streuung schneller geladener teilchen II mehrfach-und vielfachstreuung *Z. Nat.* A **3** 78–97

[24] Bethe H A 1953 Moliere's theory of multiple scattering *Phys. Rev.* **89** 1256

[25] Strong G C *et al* 2024 TomOpt: differential muon tomography optimisation (available at: https://github.com/GilesStrong/tomopt)

[26] Paszke A *et al* 2019 PyTorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F d Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.) pp 8024–35

[27] Baydin A G *et al* 2021 Toward machine learning optimization of experimental design *Nucl. Phys. News* **31** 25–28

[28] Dorigo T *et al* 2023 Toward the end-to-end optimization of particle physics instruments with differentiable programming *Rev. Phys.* **10** 100085

[29] Strong G and Dorigo T 2021 TOMOPT: differential muon tomography optimisation *1st MODE Workshop on Differentiable Programming*

[30] Strong G *et al* 2022 Data Challenge for the 2nd MODE *Workshop on Differentiable Programming*, 2nd MODE Workshop on Differentiable Programming (Repository)

[31] Strong G *et al* 2022 TomOpt: differential optimisation of muon-tomography detectors *2nd MODE Workshop on Differentiable Programming*

[32] Agostinelli S *et al* 2003 Geant4-a simulation toolkit *Nucl. Instrum. Methods Phys. Res.* A **506** 250–303

[33] Allison J *et al* 2006 Geant4 developments and applications *IEEE Trans. Nucl. Sci.* **53** 270–278

[34] Allison J *et al* 2016 Recent developments in Geant4 *Nucl. Instrum. Methods Phys. Res.* A **835** 186–225

[35] Schultz L J, Blanpied G S, Borozdin K N, Fraser A M, Hengartner N W, Klimenko A V, Morris C L, Orum C and Sossong M J 2007 Statistical reconstruction for cosmic ray muon tomography *IEEE Trans. Image Process.* **16** 1985–93

[36] Qasim S R, Kieseler J, Iiyama Y and Pierini M 2019 Learning representations of irregular particle-detector geometry with distance-weighted graph networks *Eur. Phys. J.* C **79** 608

[37] Strong G 2022 2-level graphs for muon-tomography *5th CERN Inter-Experiment Machine Learning Workshop*

[38] Linnainmaa S 1970 The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors *Master's Thesis* University of Helsinki

[39] Werbos P J 1981 Applications of advances in nonlinear sensitivity analysis *Proc. 10th IFIP Conf. 31.8–4.9* (NYC) pp 762–70

[40] Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating errors *Nature* **323** 533–6

[41] Hadamard J 1908 *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées* (*MéMoires présentés par divers savants à l'Académie des sciences de l'Institut de France: ÉXtrait*) (Imprimerie Nationale)

[42] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *3rd Int. Conf. on Learning Representations, ICLR 2015* (*Conf. Track Proc.*) ed Y Bengio and Y LeCun pp 7–9

[43] Martínez P, Díez C, Gómez P and Orio A 2022 Process monitoring: measurement of the metal-slag interface in furnace ladles, *IAEA TECDOC SERIES. Muon Imaging (Present Status and Emerging Applications)* vol 2012 pp 34–35

[44] Smith L N 2017 Cyclical learning rates for training neural networks *2017 IEEE Winter Conf. on Applications of Computer Vision (WACV)* pp 464–72

[45] Smith L N and Topin N 2019 Super-convergence: very fast training of neural networks using large learning rates *Proc. SPIE* **11006** 369–86

[46] Smith L N 2018 disciplined approach to neural network hyper-parameters: part 1 - learning rate, batch size, momentum, and weight decay (arXiv:1803.09820)