Imperial College of Science, Technology and Medicine
Department of Physics

Optimisation of the SHiP Beam Dump Facility with generative surrogate models

Sergey Shirobokov

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Physics of Imperial College London, July 2021

Copyright Declaration

Statement of Originality

I state that the research presented in this thesis is my own work, and that all else is appropriately referenced.

Abstract

The SHiP experiment is a proposed fixed target experiment at the CERN SPS to search for new particles. To operate optimally, the experiment should feature a zero background environment. The residual muons flying from the target are one of the largest sources of the background. To remove them from the detector acceptance, a dedicated muon shield magnet is introduced in the experiment. The shield should be optimised to deliver the best physics performance at the lowest cost.

The optimisation procedure is very computationally costly and, thus, requires dedicated methods. This thesis comprises of a detailed description of a new machine learning method for the optimisation, comparisons to existing techniques, and the application of the method to optimising the muon shield magnet. In addition, the set of technological and simulation problems affecting the optimisation is discussed in details. Finally, the set of requirements for the muon shield prototype design and verification is presented.

Acknowledgements

I would like to thank my Higher School of Economics and Oxford colleagues Vladislav Belavin, Andrey Ustyuzhanin and Atılım Güneş Baydin for our collaborative efforts in producing the optimisation algorithm. Further, I would like to express my gratitude to Michael Kagan from SLAC, who was deeply involved in the optimisation algorithm project.

Furthermore, I would like to thank Fedor Ratnikov and Evgeny Kurbatov for our productive work on the muon shield prototype, and Pavel Dergachev for the provided CAD drawings.

I would like to extend my thanks to Thomas Ruf, Eric van Herwijnen, Richard Jacobsson and Oliver Lantwin for guiding my understanding of the SHiP experiment and, in particular the FairShip simulation package.

Additionally, I would like to thank many of my friends who provided valuable support and encouragement throughout my PhD experience and proofread my thesis. Furthermore, I would like to thank my family and especially my brother Alexey for their support, guidance, and encouragement throughout my time at university, particularly during my PhD.

Finally, I would like to thank my supervisor Andrey Golutvin for his guidance and support during my PhD.

Contents

Chapter 1

Introduction

With the discovery of the Higgs boson by the ATLAS [1] and CMS [2] collaborations, all particles predicted by the Standard Model (SM), with the exception of the anti-tau neutrino, have been observed [3]. Nevertheless, certain observed phenomena remain unexplained by the SM. It has been proven that neutrinos have non-zero mass as opposed to SM postulation [4]. The observed baryon asymmetry in the Universe is yet another problem not described by the SM [5]. If the amount of matter and anti-matter had been the same, the earlier-state Universe would have been preserved, and no large mass formations would have been possible [5]. One more issue is the presence of Dark Matter and Dark Energy in the Universe. From cosmological observations, it is estimated that ordinary matter accounts only for 5% of the total composition of the Universe. The remaining part is made of about 68% of Dark Energy and about 27% of Dark Matter [6, 7]. These observations suggest the possible presence of new physics beyond the SM (BSM physics), which is yet to be discovered. Given no new particles have been found, these new BSM particles are expected to either be very heavy, and thus not accessible by accelerators, or to interact weakly with ordinary matter, thus requiring high experiment intensity to be observed [8].

One possibility is that the hypothetical particles are heavy and require very high collision energy to be observed, the so-called "energy frontier" research. In the last decades, many major experiments, such as the Large Hadron Collider (LHC) at the European organisation for Nuclear Research (CERN) and Tevatron facility in the US, have followed this path. Future proposed experiments, such as the Future Circular Collider (FCC), are designed to pursue similar avenues. The second option is that new particles have very feeble interactions and, thus, stay unobserved. If this is true, the new experiments will need to cross the "intensity frontier" to detect new particles [8].

Hidden Sector (HS) models (or portals) are BSM models that have one or several long-lived neutral particles which interact with SM only via mediator particles. The idea of HS portals is very appealing since these models allow the construction of new complex forces that cause interactions between BSM and SM particles. The couplings of the mediator particles to SM particles are so weak that observed SM phenomena remain unaffected [9, 10, 11].

Three main possible portals will be described in more detail later in this thesis: Vector portal, Scalar portal and Neutrino portal. Another possible extension to the SM are non-renormalisable axion-like particles (ALP). Portals can provide a mediator between Dark Matter and SM particles, and additional annihilation channels relevant to the freeze-out of Dark Matter [8].

As mentioned above, new particles can be very light but still inaccessible at the LHC and other collider experiments because of their weak couplings to SM particles. In such experiments, any rare decay of a Hidden Sector particle may be obscured by a huge background. The searches via missing mass technique are also challenging due to the need to construct a highly granular detector that can cover 4π spherical angle and distinguish the constituents of the overlapping jets to determine the missing mass precisely [12].

The Search for Hidden Particles (SHiP) experiment is a recently proposed fixed target experiment at CERN designed to directly search for extremely weakly interacting new particles at low energies [8, 13, 14], hypothesised by Hidden Sector models. The SHiP experiment will operate at the "intensity frontier" of the searches.

At SHiP, protons at an energy of 400 GeV are injected by the CERN Super Proton Synchrotron (SPS) ring, and shot at a fixed target to produce potential BSM and SM particles via various processes. With such a high-intensity proton beam, the resulting rate of SM particles will be very high. Therefore, to fully exploit the high intensity of the experiment and to make the detection of BSM particles feasible, the experiment must be able to reduce the SM background rate to nearly zero. This poses an extremely challenging task. While hadrons produced in the target can be easily stopped by a dedicated hadron absorber, eliminating outgoing muons is one of the most important challenges of the experiment.

Thus, one of the most critical tasks of the experiment is to reduce the muon flux rate to a manageable level from the initial 10^{11} muons/spill. To solve this problem, an experiment subsystem, named the muon shield, was proposed. The crucial task is to perform an optimisation of the shape and magnetic fields of the shield to efficiently sweep out muons. Since the muon shield is a complex system of magnets

characterised by many parameters, its optimisation is an extremely daunting problem and requires the use of state-of-the-art optimisation algorithms and machine learning methods. Given the importance of the problem, several iterations of optimisation utilising machine learning methods have been done. The last iteration, described in this thesis, has resulted in a new optimisation algorithm called *local generative surrogate optimisation* (L-GSO).

Machine Learning methods are well-established methods in various computer science fields, such as computer vision, natural language processing [15] and reinforcement learning [16]. With their success in the original domain, the popularity and successful application of machine learning methods are extending to natural sciences domains, such as biology [17], chemistry [18], physics [19] and many more. Moreover, the field of high energy physics (HEP) is currently experiencing an exponential growth in the successful application of machine learning methods in a range of HEP topics such as: jet classification [20], particle identification (PID) [21], tracking [22], anomaly detection [23] and likelihood-free inference [24]. In addition to the problems above, machine learning methods such as Bayesian optimisation [25, 26] and evolutionary strategies [27] have been applied for the optimisation tasks in the domain of high energy physics [28, 29].

Many optimisation methods can be applied in the context of the SHiP muon shield optimisation task. They include Bayesian optimisation [25, 26], numerical differences [30], genetic algorithms [31, 32] and stochastic gradient estimators [33]. All the algorithms above share the property of being able to optimise noisy complex black-box functions where it is computationally expensive to evaluate a function. Interestingly, Monte Carlo simulators in the HEP domain can be treated as such functions, thus allowing the application of the aforementioned machine learning methods. However, each of these algorithms has its own drawbacks and benefits and might be nonoptimal for the particular task of shield optimisation. The reasons behind the performance of each algorithm will be discussed in detail in this thesis. In what follows, this thesis will be focused on the application of the machine learning techniques to the shield optimisation, and, in particular, the L-GSO algorithm, in the development of which the author of this thesis took a leading role.

Finally, after the last iteration of the shield optimisation it is clear that a dedicated muon shield prototype is needed to test all the assumptions included in the optimisation procedure. These assumptions include both physical and technological simplifications used during the optimisation, that require verification. The detailed description of the simplifications, potential solutions and the prototype test beam experiment are discussed later in the thesis.

Chapter 2 describes the theoretical motivation for the SHiP experiment. In particular, it gives a brief overview of the physics program that is possible to explore in the experiment and its importance as an extension to the Standard Model (SM).

Chapter 3 contains a detailed description of the SHiP experiment. It describes the overall experimental setup, the properties of the proposed experimental facility and the details of the implementation of the experiment subsystems. Additionally, it describes requirements imposed on the subsystems to meet planned physics performance and achieve zero background rates.

Chapter 4 provides a mathematical formulation of an optimisation problem and gives a broad overview of the algorithms used for optimisation in computer science and natural sciences. In particular, it provides an explicit description of the optimisation procedure in high energy physics experiments which utilise Monte Carlo methods to perform simulations. The chapter then dives into the details of various optimisation methods, their mathematical formulation and applicability in the SHiP shield optimisation.

Chapter 5 describes in detail the L-GSO algorithm, the motivation behind the algorithm, potential drawbacks and ways to mitigate them. It then demonstrates the performance of the algorithm on a set of toy mathematical and physics problems and performs a comparison of L-GSO to other available optimisation methods. The new algorithm has resulted in the publication in the machine learning venue [34]. Thus, some details of the algorithm in the chapter may appear similar to the publication.

Chapter 6 contains the application of the L-GSO algorithm for shield optimisation. It starts with an overview of previous findings, an explanation of the constructional limitations and the reasons for the re-optimisation. It then describes in detail a new strategy for the optimisation, a new objective function, and compares a newly found geometry with the previous baseline. Finally, simulation simplifications and their implication for the optimisation are discussed.

Chapter 7 concludes the thesis with the motivation behind the shield prototype, its design and a set of experimental measurements required to verify the design.

Chapter 2

Portal models

2.1 Experimental evidence for the Beyond Standard Model physics

The Standard Model (SM) is the cornerstone theory of modern particles physics. It is able to successfully describe and predict the majority of observed phenomena in particle physics, such as the W, Z and Higgs bosons, the gluon and the top quark. Additionally, the SM has been able to precisely predict not only the existence but also the properties of these particles. Overall, it successfully describes strong, weak and electromagnetic forces. However, there are challenges beyond the SM that still need to be solved to create a complete theory of the Universe. Furthermore, if in the case of the gravitational force, the general theory of relativity is a well established and experimentally checked theory which successfully describes all the gravitational interactions, the origin of baryon asymmetry, Dark Matter, and Dark Energy and neutrino masses are still open questions in science.

- **Dark Matter:** One of the most intriguing questions of the last decades is the nature of Dark Matter. It is known now that Dark Matter accounts for 27% of matter, whereas usual matter accounts only for 5% in the total composition of the Universe, with the rest being Dark Energy. Experimental proofs of the existence of Dark Matter stem largely from cosmological observations, such as galaxy rotation curves [35], gravitational lensing [36], and the cosmic microwave background [37] which lead to the creation of the Λ-CDM model. A few models are describing Dark Matter: cold, warm and hot Dark Matter. By cold Dark Matter, it is usually meant that the dark matter particle is non-relativistic [38].

This property allows Dark Matter to support modern galaxies formation, thus making the cold Dark Matter model very appealing for cosmology. Examples of such models include Axions and Weakly Interactive Massive particles (WIMPs). Warm Dark Matter consists of relativistic particles such as heavy neutral leptons (HNL) [39]. Finally, hot Dark Matter is an ultrarelativistic version of Dark Matter; for example, a neutrino particle can be classified as hot Dark Matter [39].

- **Baryon asymmetry of the Universe (BAU):** According to a modern understanding of fundamental processes, matter and anti-matter were produced in equal amounts during the Big Bang. However, as observed today, there is an asymmetry of baryons over anti-baryons, which is described by $\eta = \frac{n_B - n_{\bar{B}}}{n_\gamma} = \frac{n_B}{n_\gamma} \approx 10^{-10}$, where n_B is number of baryons and n_γ is number of photons resulting from the annihilation processes. There is no explanation for the observed BAU in the SM. Although there is evidence of CP-symmetry violation in the CKM matrix, the magnitude of the effect is not large enough to explain the observed baryon asymmetry. However, CP violation provides evidence that an asymmetry could have been created dynamically during baryogenesis. Recently, CP-asymmetry has also been observed in the T2K experiment through the measurement of non-zero δ_{CP} in neutrino mixing[40], thus making a new step to test the leptogenesis models [41].

- **Neutrino masses:** In the SM, neutrinos are massless; however, it is known from neutrino oscillation phenomena that they must have non-zero mass [4]. Since we only observe left-handed neutrinos, one can not utilise the Yukawa Coupling mechanism to generate their mass. As such, an extension of the SM is needed, which introduces right-handed neutrinos and thus naturally generates their mass. A new extension of the SM, however, must take into account constraints on the neutrino mass splitting from oscillation experiments and an upper limit on the sum of neutrino masses from cosmological observations of the baryonic and Dark Matter densities and neutrino freeze-out effect [42].

All of the above observations suggest the physics beyond the SM model. As was mentioned, one way to create an extension of the SM is via Hidden Sector portals. In what follows, the models that can be probed in a beam dump experiment are described.

2.2 Vector Portal

In this model, a new $U(1)$ gauge symmetry is introduced, under which only the Hidden Sector particles receive an associated charge, while the SM particles interact with the Hidden Sector via a kinetic mixing mechanism [43]. By analogy with the SM QED Lagrangian, a new vector particle charged under this symmetry is called a Dark Photon $A^{'}$, and the Lagrangian has the following form:

$$\mathcal{L} = \mathcal{L}_{SM} - \frac{1}{4}F^{'}_{\mu\nu}F^{'\mu\nu} - \frac{\epsilon}{2}F_{\mu\nu}F^{'\mu\nu} + \frac{1}{2}m^2_{A'}A^{'}_{\mu}A^{'\mu} \tag{2.1}$$

where $F^{'}_{\mu\nu} = \partial_\mu A^{'}_{\nu} - \partial_\nu A^{'}_{\mu}$ is the strength tensor of the "dark" field. The term $\frac{\epsilon}{2}F_{\mu\nu}F^{'\mu\nu}$ represents a kinetic mixing mechanism and provides interaction between Dark Photon and the SM photon. Then the simplest extension introduces the Dark Matter field χ by adding the following term $\mathcal{L}_\chi = \bar{\chi}[\gamma_\mu(i\partial_\mu - g^{'}A^{'}_{\mu}) - m_\chi]\chi$ to the Lagrangian above, and $g^{'}$ is the gauge coupling. The model can potentially help to explain a couple of experimental observations.

The $g-2$ anomaly refers to the discrepancy between the measured muon magnetic moment and the SM prediction [44]. Existing theories do not explain this deficiency of theory with respect to experimental observations. Although the anomaly cannot be explained by the Dark Photon alone, there is still a potential region of parameters where the Dark Matter-Dark Photon model can provide a positive correction to the theory required to match experimental data [45].

Utilising Dark Photon mediators, one can solve the problem of Dark Matter overproduction in the early Universe. To do so, a Dark Matter with mass m_χ is introduced and annihilates via the process shown in Figure 2.1, where $\alpha_D = (g^{'})^2/4\pi$. Thus, one can find the annihilation rate:

$$\sigma v \sim \frac{\alpha\alpha_D\epsilon^2 m^2_\chi}{m^4_{A'}}, \tag{2.2}$$

where v is the relative velocity of colliding Dark Matter particles. Extensive searches for Dark Matter have been conducted in the mass range of $O(10-10^4)$ GeV, usually referred to as the WIMP region. WIMPs have been extensively searched for by direct detection experiments such as XENON100 [46] and XENON1T [47], as well as presently operating experiments such as LUX-Zeplin [48]. However, up to now, no WIMPs have been observed. In the region below $O(10)$ GeV such experiments become

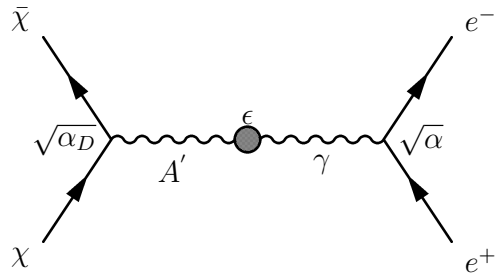


Figure 2.1: Feynman diagram of Light Dark Matter annihilation via Dark Photon with subsequent decay to two leptons.

insensitive to nuclear recoils and are overwhelmed by the neutrino background. The existing and projected limits for such experiments are presented in Figure 2.2.

Introducing Light Dark Matter (LDM) via vector portal has the advantage of setting α_D to an arbitrary value, with the only requirement of matching current DM abundance in the Universe. This mechanism allows to introduce LDM with mass $m_\chi \approx O(\text{MeV} - 10\text{ GeV})$ which can be explored in the accelerator-based experiments.

From Figure 2.1, it can be seen that Dark Photon can decay to Dark Matter in the case $2m_\chi < m_{A'}$. The ratio of the decay rates of A' to a visible mode and to Dark Matter scales as $\alpha\epsilon^2/\alpha_D$. In the case of $\alpha\epsilon^2 \gg \alpha_D$, Dark Photon can fly macroscopic distances and be detected in the downstream detectors via its decay to visible particles (leptons and hadrons). If $\alpha\epsilon^2 \ll \alpha_D$ holds, Dark Photon immediately decays to Dark Matter, which can be detected in the dense material of the dedicated detectors via Dark Matter scattering off electrons or nucleons, and this is discussed in details in section 2.6.

The main production mechanisms of Dark Photon in accelerator-based experiments are:

- Meson decays: Mesons are produced in large abundance at the beam dump of an experiment. Decays of π^0, η, ω contribute the most to Dark Photon production up to $m_{A'} \sim 0.9\,\text{GeV}$.
- Bremsstrahlung: Relevant for electron and proton beams, where either electron or proton scattering on a fixed target with a charge Z results in a radiated Dark Photon, $eZ \to eZA'$ or $pZ \to pZA'$.
- QCD production: Dark Photon is produced in the $q\bar{q}$ annihilation process, which can happen at a hadron collider or at a proton beam dump experiment.
- e^+e^- annihilation at e^+e^- colliders.

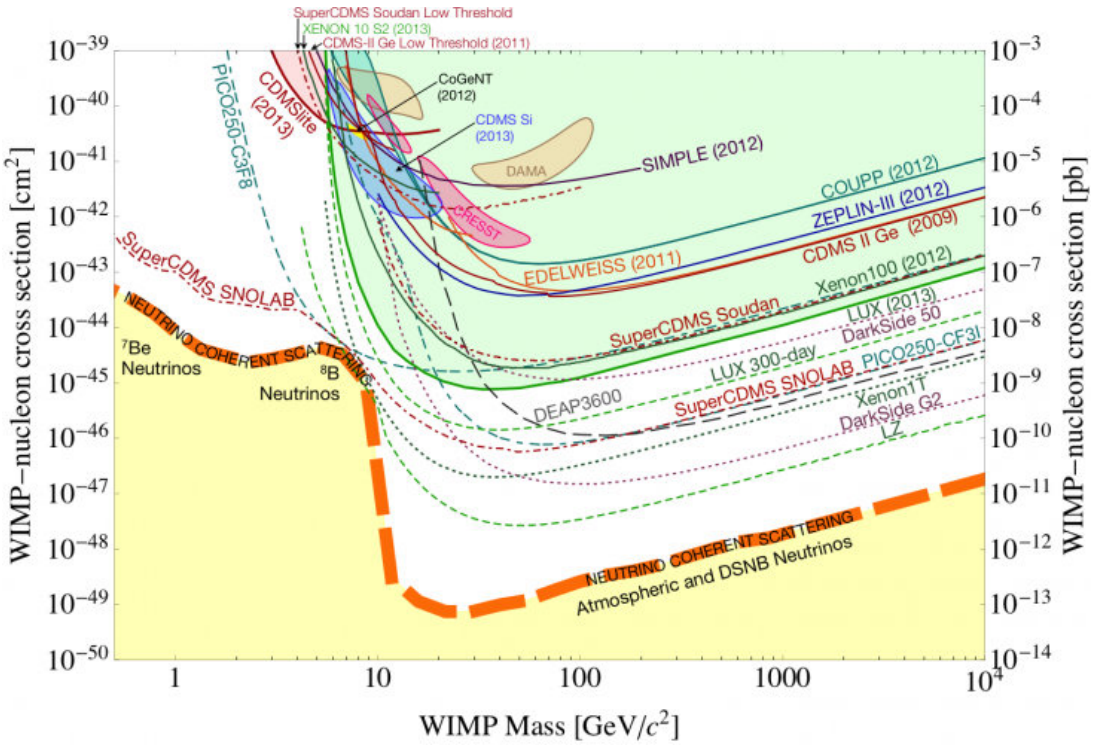


Figure 2.2: Existing and projected limits of the WIMP-nucleon scattering. [49]

The relative contribution of each mechanism depends on the Dark Photon mass range and the energy of colliding particles. Additionally, Dark Photons can be produced in an electromagnetic shower with the cascade effect, which can significantly contribute to the production rates.

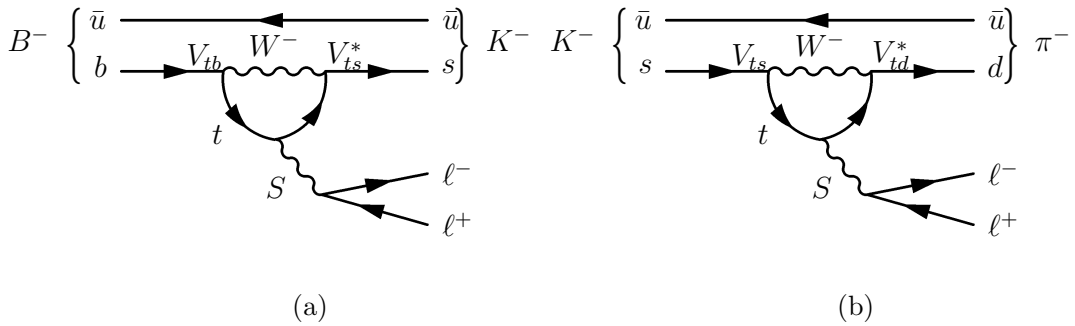
2.3 Scalar Portal

In this model, a scalar singlet S is coupled to the gauge invariant combination $H^\dagger H$ of the Standard Model Higgs doublet [8]:

$$\mathcal{L} = \mathcal{L}_{SM} + \frac{1}{2}(\partial_\mu S)^2 - \frac{M_S^2}{2}S^2 + gSH^\dagger H + \mathcal{L}_{self}, \tag{2.3}$$

where g is the coupling constant and $\mathcal{L}_{self}$ is a self-interaction term. After symmetry breaking, one obtains:

$$\mathcal{L}_{S,int} = \theta S \left[\sum_f m_f \bar{f} f + M_W W_\mu^+ W^{-\mu} + \dots \right] \tag{2.4}$$



(a) (b)

Figure 2.3: Feynman diagram of Dark Scalar S production from (a) a B decay and (b) a K decay and the subsequent decay of S to two leptons.

where $\theta = gv/m_H$ and $v = 246\,\text{GeV}$ is the Higgs boson vacuum expected value (VEV), and m_H is the Higgs boson mass. The sum goes over all massive fermions (leptons and quarks). Dark Scalars may be produced in the decays of $B^\pm$, $K^\pm$ and $D^\pm$ mesons. In fact, only the contributions from B and K are relevant since their decay width is proportional to the top-quark mass as m_t^4, whereas the decay width of the D meson scales only as m_b^4. Thus, the contribution of the D meson to the total flux of Dark Scalars is suppressed by $O(10^6)$. The decay modes of Dark Scalar consist of $e^+e^-, \mu^+\mu^-, \pi^+\pi^-$ pairs for masses up to 1 GeV, and from decays to kaons and tau-leptons after 1 GeV [50]. The relevant Feynman diagrams are shown in Figure 2.3.

A Light Dark Scalar can also be a mediator portal for Dark Matter. Dark Matter pairs can directly annihilate into SM particles via an interaction with a Dark Scalar. As long as the value of θ agrees with the SM processes rate, one can select a coupling of DS to DM to produce the correct Dark Matter density. In such a model, it is possible to search for Dark Matter via its scattering on nuclei in the same way as it is done for Dark Photons. In that case, a beam dump experiment can probe Dark Matter masses in the range of $1 - 10\,\text{GeV}$ [8].

2.4 Neutrino portal

It is possible to extend the Standard Model with right-handed neutrinos without violating any other parts of the theory. If the newly added neutrinos have the Dirac mass term, they are not capable of explaining BAU. However, since neutrinos are uncharged, one can substitute the Dirac mass term with the Majorana term, thus introducing lepton-number violation and a potential solution to BAU [51]. Finally,

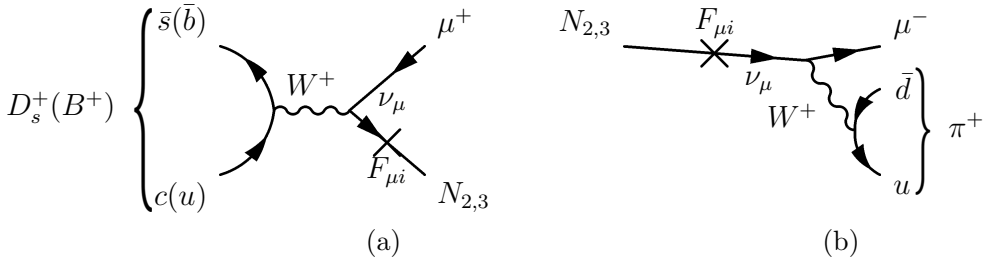


Figure 2.4: Feynman diagrams of possible NHL a) production and b) decay modes.

both Dirac and Majorana mass terms can be added to the model. In that case, after diagonalising the mass matrix, one would obtain three light active neutrinos and $\mathcal{N}$ sterile neutrinos or Heavy Neutral Leptons [52]. The Lagrangian for this case is presented below:

$$\mathcal{L} = \mathcal{L}_{SM} + \bar{N}_i i\partial\!\!\!/ N_i + F_{\alpha i}(\bar{L}_\alpha \Phi) N_i + M_i \bar{N}_i N_i + \text{h.c.}, \tag{2.5}$$

where L_α is the left lepton doublet, $\alpha \in \{e, \mu, \tau\}$ is a flavour index, $F_{\alpha i}$ is a dimensionless Yukawa coupling and Φ is the Higgs doublet.

Theoretically, it is possible to have an arbitrary number of HNLs with masses ranging from keV to $\sim 10^{18}$ GeV. However, it was shown in the νMSM [52] extension of the SM that by introducing three generations of HNLs, all three problems of the SM can be solved simultaneously. To do so, the see-saw mechanism is used [53], with at least one generation of the HNL. N_1 should have a mass of an order of 10 keV, and can be a candidate for warm Dark Matter [52]. To correctly determine active neutrino masses, one can further introduce $N_{2,3}$ that may have arbitrary large masses. The presence of three generations of right-handed neutrinos itself determines correct active neutrino masses [52]. In addition to it, it was shown that by restricting masses of $N_{2,3}$ to the order of the MeV - GeV scale, BAU could be introduced via leptogenesis [51, 54]. The ability to solve three major challenges of the Standard Model while introducing particles in the mass range within the reach of modern accelerators makes the νMSM theory very appealing to probe experimentally.

In beam dump experiments, HNLs can be produced in the decays of heavy flavour particles. Mostly B and D mesons contribute to the production of the HNL, as presented in Figure 2.4. The subsequent decay of the HNL occurs via mixing with an active neutrino and its subsequent decay to one charged lepton and a W or Z boson. Thus, in the final state, a HNL decay can produce two or three outgoing particles, e.g.:

$$N_i \to l^+ l^- \nu; \quad N_i \to l\pi, lK \tag{2.6}$$

2.5 Axion-Like Particles

One can add non-renormalisable theories to the Standard Model Lagrangian to introduce new particles. In that case, Pseudo Nambu-Goldstone bosons will originate from a spontaneously broken symmetry of the theory. Originally, such a mechanism was introduced by Peccei-Quinn [55] to solve the strong CP problem in QCD. Those new particles have two crucial properties. Firstly, their interactions are suppressed by the scale of the symmetry breaking f_a. Secondly, their masses are suppressed by the scale of symmetry breaking $m_A \sim 1/f_a$. The mass of the originally proposed axion is around 10^{-5} eV and the beam dump experiments are not sensitive to it [50].

However, one can introduce heavier particles to the theory, utilising the same mechanism as that used to generate QCD axions. Such particles are called axion-like particles (ALP). In a similar fashion to Dark Photons, ALPs can be mediators for Dark Matter, providing a mechanism for Dark Matter annihilation and its subsequent decay to Standard Model particles [8, 56]. This could provide the correct Dark Matter abundance [56].

ALPs can couple to SM fermions, photons and gluons. In the case of their coupling to SM fermions, their production and decay channels are similar to those of Dark Scalar: they are produced through B and K decays and their subsequent decay to leptons and hadrons, but with different couplings [8]. The Lagrangian in that case is as follows:

$$\mathcal{L} = \frac{\partial_\mu \phi}{f_a} \overline{\psi} \gamma^\mu \gamma^5 \psi \tag{2.7}$$

When ALPs couple to photons, they are produced via the Primakoff effect [57] and decay to two photons, with the Lagrangian being:

$$\mathcal{L} = \frac{\alpha}{4\pi f_a} \phi F_{\mu\nu} F^{\mu\nu} \tag{2.8}$$

2.6 Light Dark Matter

Different techniques exist to search for Light Dark Matter particles at accelerators:

- LDM is produced by the intensive beam of electrons or hadrons, and subsequently scatters in the detector downstream of the production point. Examples of such a setup are SHiP [14], LSND [58] and MiniBoone [59].
- Missing mass experiments, such as BaBaR [60] and Belle II [61], aim to reconstruct the energy and momentum of the outgoing photon in e^+e^- collisions $e^+e^- \to \gamma A'$. Such experiments require a mono-photon trigger and low background to achieve this task.
- Missing energy experiments, such as LDMX [62] and NA64 [63], reconstruct the recoil energy and momentum of an electron in the scattering: $eZ \to eZA'$. Such experiments have to suppress the background from neutrino scattering events.

All three techniques allow to search for LDM that can be produced via a Dark Photon mediator as discussed in the Section 2.2. However, it has to be noted, that the number of events in the first approach is proportional to ϵ^4, whereas in second and third to ϵ^2, which makes the direct Dark Matter scattering experiments challenging. Nevertheless, the benefit of the first approach is that it allows to probe LDM models with mediators other than Dark Photon, for instance, Dark Scalar from Section 2.3.

In its general form, the Lagrangian for LDM is as follows:

$$\mathcal{L}_{LDM} = \mathcal{L}_{dark}[\psi_{LDM}, \psi_{mediator}] + \mathcal{L}_{portal}[\psi_{mediator}, \psi_{SM}] + \mathcal{L}_{SM}[\psi_{SM}], \tag{2.9}$$

where $\mathcal{L}_{portal}$ is different for different portal models. For example, in the aforementioned case of Dark Photon, $\mathcal{L}_{portal} = -\frac{\epsilon}{2} F_{\mu\nu} F'^{\mu\nu}$, in the case of scalar portal $\mathcal{L}_{portal} = gSH^\dagger H$, and for the leptophobic mediators $\mathcal{L}_{portal} = \epsilon V_\mu J_\mu^{baryon}$, where J_μ^{baryon} is a baryon current.

In beam dump experiments, there are three main production mechanisms for LDM that coincide with Dark Photon production:

- Meson decays: $\pi^0/\eta \to \gamma + A' \to \gamma + \chi^\dagger + \chi$

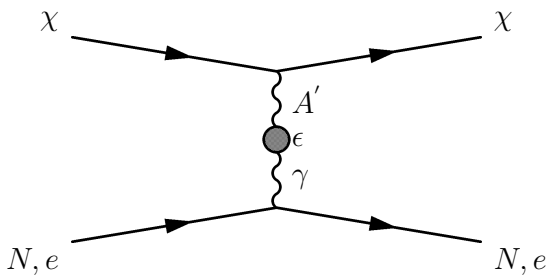


Figure 2.5: Feynman diagram of Light Dark Matter scattering on an electron or a nucleon via Dark Photon (Dark Scalar) mediator.

- Beam (proton or electron) bremsstrahlung: $p(e) + N \rightarrow p(e) + N + A' \rightarrow p(e) + N + \chi^{\dagger} + \chi$
- QCD production: $q + \bar{q} \rightarrow A' \rightarrow \chi^{\dagger} + \chi$

In the missing energy experiments only the second production mechanism is feasible to pursue. In such experiments, an intense beam of electrons impinges on the target and the recoil energy or momentum of the electron is measured. It is only possible to perform such experiments with an electron beam, since the bremsstrahlung probability is suppressed for a proton beam [62]. The main benefit of such an approach is that the number of observed events is only suppressed by a factor of ϵ^2, since the approach requires the observation of LDM production only, without scattering.

In contrast, beam dump scattering experiments are able to detect LDM produced in all of the above channels. The main drawback of this approach is that the sensitivity is suppressed by ϵ^4, first LDM must be produced (ϵ^2) and then scatter (ϵ^2). Nevertheless, it was shown [58] that even in this scenario it is possible to impose new constraints on the LDM parameter space by considering LDM elastic electron or nucleon scattering, inelastic π^0-like scattering and deep inelastic scattering. To detect each of these processes and minimise the background, the experiment must be equipped with a section of dense material, surrounded by appropriate detectors to observe the scattering, the electron scattering signature being the easiest to identify.

Chapter 3

The SHiP experiment

3.1 Experiment performance requirements

As it was discussed in Chapter 2, new hypothetical particles can be produced in heavy flavour decay. To achieve high production rates of heavy flavour particles, it would be advantageous to have high energy proton beam, as can be seen from Figure 3.1. The highest energy beam currently available for a beam dump experiment is located at CERN SPS and provides a proton energy of 400 GeV.

There are four main extensions of the SM Lagrangian that can introduce new particles that can be probed in beam dump experiments: Vector, Scalar, Neutrino portals and axion-like particles. A summary of the detection signatures for all the models is presented in Table 3.1. Given the final states, the new experiment should be able to provide particle identification (PID) in order to distinguish between different portal models. In addition, some SM particles originating from neutrino scattering (such as K_L^0, K_S^0), can decay into the same final states as in Table 3.1, thus creating a background for the portal particle decays. The background can also originate from muons that are copiously produced in the beam dump. Thus, the experiment must be able to veto such types of events using a dedicated set of detectors.

Finally, the presence of the Dark Photon portal allows LDM to be probed via its scattering. To detect LDM, an experiment should include a detector with a dense material to induce Dark Matter scattering, and tracking detectors with a high spatial and angular resolution to identify emerging electromagnetic showers and nucleus recoils, and distinguish them from neutrino scattering events.

The requirements outlined above lead to the experimental layout of the proposed

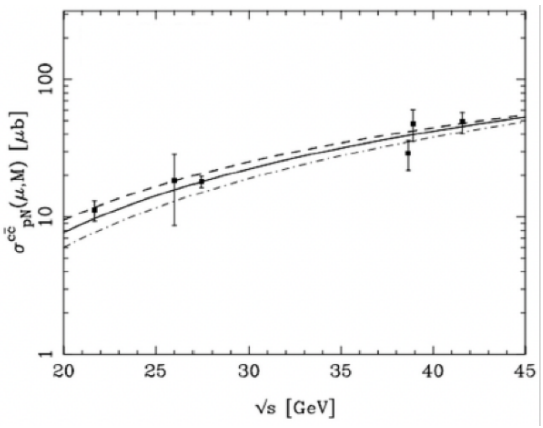


Figure 3.1: Charm production cross-section in a proton-isoscalar target collision as a function of proton beam energy in the centre of mass frame [14]. The data points shown are from a compilation of fixed target results [64].

Portal	Final states
Vector	l^+l^-, h^+h^-
Scalar	l^+l^-, h^+h^-
Neutrino(2-body)	$l^\pm K^\mp, l^\pm \pi^\mp$
Neutrino(3-body)	$l^\pm l^\mp \nu_l$
ALPs (photon)	$\gamma\gamma$
ALPs (fermion)	l^+l^-, h^+h^-

Table 3.1: Portal detection signatures. l and h denote a lepton and a hadron respectively.

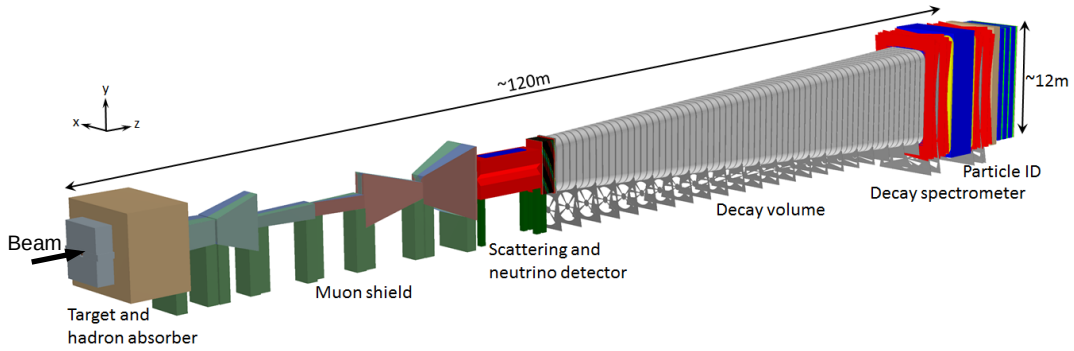


Figure 3.2: Current layout of the SHiP experiment. [50]

Search for **Hi**dden **P**articles experiment - **SHiP**.

3.2 Experimental layout

The general design of the SHiP experiment is presented in Figure 3.2. It starts with a heavy target that maximises heavy flavour production, followed by a hadron stopper. A dedicated active muon shield is located after it to sweep out muon particles produced in the target.

The scattering and neutrino detector (SND) is located downstream of the muon shield. Its main goal is to detect scattering signatures of LDM and τ neutrino physics. It is equipped with an emulsion-based spectrometer and target trackers, enclosed inside a magnetised volume.

The main detector, called the Hidden Sector (HS) detector, is located downstream of the decay volume and aims to measure the decays of HS particles to partially and fully reconstructible final states, originating in the 50 m long decay volume. To eliminate the neutrino-induced background while maximising detector acceptance, the decay volume has a pyramidal frustum shape, has a pressure of $< 10^{-2}$ bar, and is surrounded by liquid scintillator background taggers (SBT). It is followed by the straw tracker spectrometer and a magnet. The straw tracker can accurately reconstruct momenta, decay vertex, mass and impact parameter of the hidden particle at the proton target. Further downstream, the timing detector (TD), the ECAL and the muon system are located.

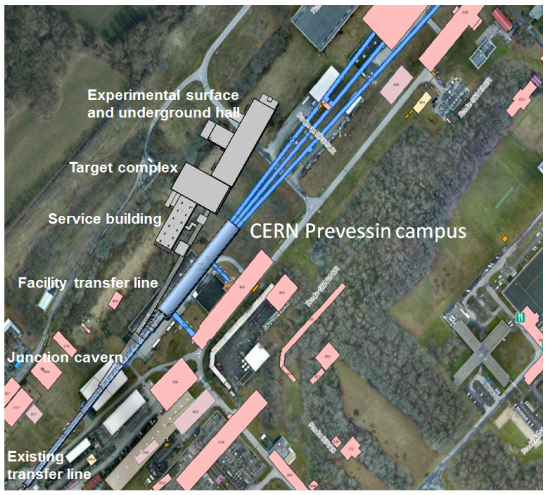


Figure 3.3: Overview of the CERN North Area including BDF. [50]

3.3 Beam Dump Facility

3.3.1 Beam line

The Beam Dump Facility (BDF) comprises the target, the hadron stopper and the muon shield and utilises the 400 GeV proton beam from the CERN SPS. A nominal intensity of 4×10^{13} protons on target (POT) per spill is projected for the experiment. This will result in 4×10^{19} protons per year for the SHiP experiment while respecting other experiments and the HL-LHC requirements. The projected POT for five years of operation is thus 2×10^{20} and is taken as a benchmark for evaluating SHiP physics sensitivities.

Because SHiP is designed to be a zero-background experiment, the control over beam extraction procedure and its uniformity is crucial, as the rate of the muon combinatorial background depends significantly on these factors. A slow extraction is also required to dilute the power of the beam on the target. A dedicated study of measuring proton rate per 400 ns and 100 ns in 1 s SHiP spill has been done [65]. The study shows that it is possible to achieve an average proton rate of just a factor of two to three larger than in a perfectly uniform extraction [50]. Dedicated studies have also been performed to minimise beam loss during slow extractions [50].

The above conditions of the accelerator operation would utilise the full power of the CERN SPS beam. Currently, there is no high-intensity facility at CERN that is compatible with these requirements. Thus, SHiP has proposed to create a new facility at the CERN North area based on minimal modifications to the existing facilities, as presented in Figure 3.3 [66].

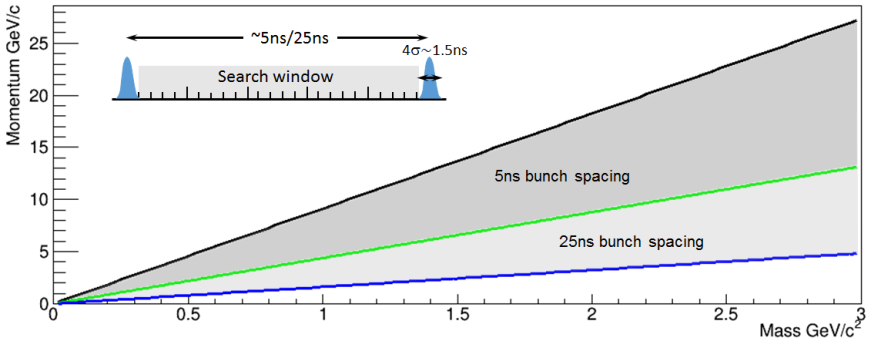


Figure 3.4: Mass-momentum areas in which it is possible to separate LDM from the neutrino by measuring the time-of-flight. Regions corresponsing to 5 ns and 25 ns in time between bunches are shown. A 40 m distance between the target and the scattering detector is assumed. [50]

Finally, it is worth noting that the uniform beam extraction has a negative effect in the case of LDM searches and neutrino physics. A bunched beam could provide evidence for LDM through the use of time-of-flight measurements. Should the observation require confirmation, SHiP should switch to a bunched beam to increase the discrimination power between Light Dark Matter and background. Figure 3.4 shows possible search window regions, assuming 4σ width of the SPS bunch to be 1.5 ns, and a 40 m distance between the target and the scattering detector.

3.3.2 Target system

As it was mentioned in Chapter 2, SHiP requires a target that maximises the production of charm and beauty mesons as well as photons. Alongside signal-producing particles, a large number of hadrons, especially pions and kaons, as well as short-lived resonances, are produced. A hadron stopper of a few meters, located after the target (see Figure 3.5), is sufficient to absorb most of the hadrons and electromagnetic radiation. However, it does not protect the experiment from the decay of the short-lived resonances, pions and kaons, resulting in a large flux of muons and neutrinos. Thus pions and kaons must be absorbed before their decay. To achieve this goal, the target is made from material with short interaction length λ_{int} and has a total length of twelve λ_{int}. It starts with TZM alloy blocks, followed by the blocks of pure tungsten. The schematic view of the target is presented in Figure 3.5. With such a configuration, SHiP is expected to produce $10^{18} D, 10^{14} B$ mesons and $10^{16} \tau$ leptons for five years of operation [67].

A smaller prototype of the target was tested during the dedicated beam measurements in 2018. The aim was to check the manufacturing process and mechanical properties

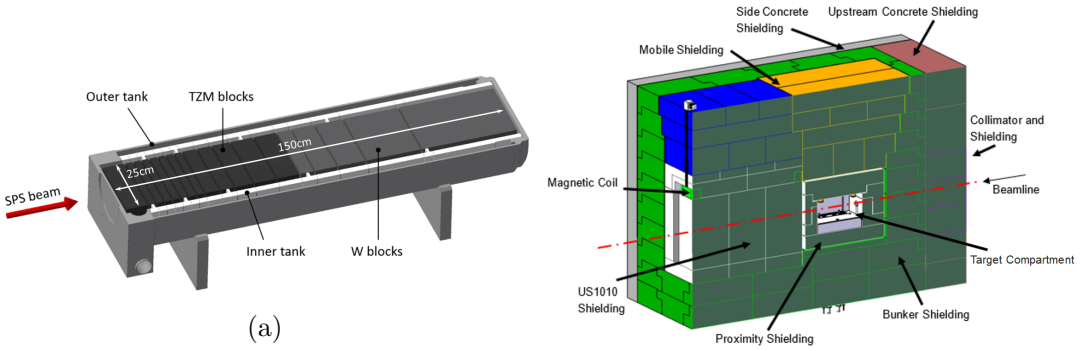


Figure 3.5: (a) Layout of the SHiP target. (b) Cross-section of the target station. The location of the target, shielding, hadron stopper and magnetic coil can be seen. [50]

of the target during the exposure to the beam and the flux of outgoing muons. The experimental setup of the traget area was similar to that of SHiP: a cylindrical SHiP-like target with small λ_{int}, followed by an iron hadron stopper and surrounded by shielding blocks of iron and concrete. 400 GeV protons were delivered by H4 SPS beamline in 4.8 s spills. Roughly 3.5×10^{11} POT were recorded for the analysis. The P and P_T distributions as well as the total yield of muons in various energy ranges were compared to the SHiP simulation. The muons originated from the production and decay of charm particles, EM resonances and from the decay of non-interacting pions and kaons were taken into account. The results demonstrated good agreement with the simulation, with a difference of only up to 20% in the yields [68].

The five-meter-long hadron stopper is located after the target. Ideally, one would like to make the hadron stopper as short as possible to increase the geometrical acceptance of the downstream detectors for the hidden particles that have a relatively large transverse momentum. Thus the minimal distance for the hadron stopper is dictated by its ability to absorb electromagnetic radiation and hadrons emerging from the target.

Since SHiP is designed as a zero-background experiment, it is favourable to optimise the hadron stopper and the target to act as a first section of the muon magnetic shield to separate positively and negatively charged muons. This allows muons to be swept out as early as possible, thus decreasing the shield's overall length and improving the geometrical acceptance. The hadron stopper is magnetised with the field of $\sim$1.6 T through the use of a coil located 1.3 m above the beam axis. Radiation exposure and heat extraction dictate the location and size of the coil. With the coil in place, the length of the hadron stopper is also limited by radiological requirements. The possibility of target magnetisation is currently being studied with the requirements

on the cavern's radiation exposure being challenging to satisfy.

3.3.3 Muon magnetic shield

The muon shield is one of the central parts of the SHiP experiment. SHiP is expected to operate at a rate of $O(10^{11})$ muons per spill, where the spill duration is about 1 s. Since SHiP is projected as a zero-background experiment, it is estimated that the muon flux needs to be suppressed by at least six orders of magnitudes to achieve the acceptable rate of $O(10^5 - 10^6)$ muons/s. The passive shield option was not considered by the collaboration, since at the SPS energies one would require an enormous amount of high-density material along the beam axis, thus drastically reducing the geometrical acceptance of the experiment. The choice was made to utilise classical ("warm") dipole electromagnets to sweep out muons.

Prior to the work of this thesis, several iterations of the optimisation were performed. The first version consisted of two parts with a 1.8 T magnetic field. The second option was constructed from six sections of magnets and a magnetised hadron stopper. Afterwards, a third optimisation was done using machine learning techniques and resulted in a lighter shield, with a 1.7 T field in the magnet and a 1.6 T field in the hadron stopper. This option is currently used as a baseline in the experiment simulation.

A detailed engineering drawing of the magnet is presented in Figure 3.6. It consists of six frustum sections, each characterised by six parameters: length plus height, width and gap size at the beginning and the end. The gap is needed to simulate the space required for coils to be inserted inside the magnet. Notice that the coils are located inside for the first half of the magnet and outside for the second half. This ensures uniformity of the magnetic field in the most active zones of the shield. The shield should be constructed from rectangular blocks due to production constraints, described later in the thesis. The total length of the current shield is about 35.5 m and the weight is ~ 1.3 kt.

To achieve a high value of the magnetic field, it was decided to use Grain-Oriented (GO) steel for the magnet material. It allows to achieve fields of 1.7-1.8 T with limited current and light coils which do not require cooling. The GO steel, however, comes in 0.3-0.5 mm thick sheets that are extremely sensitive to mechanical stress and high temperatures. This poses several technological challenges of cutting and assembling GO sheets and connecting GO sheets to create a full magnet. Given the challenges, it was estimated that an average field of 1.7 T is more realistic than 1.8 T

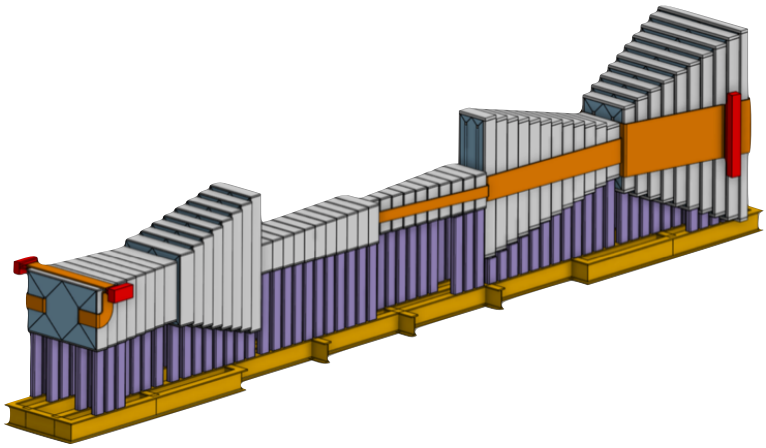


Figure 3.6: A CAD engineering model of the current SHiP magnet design.

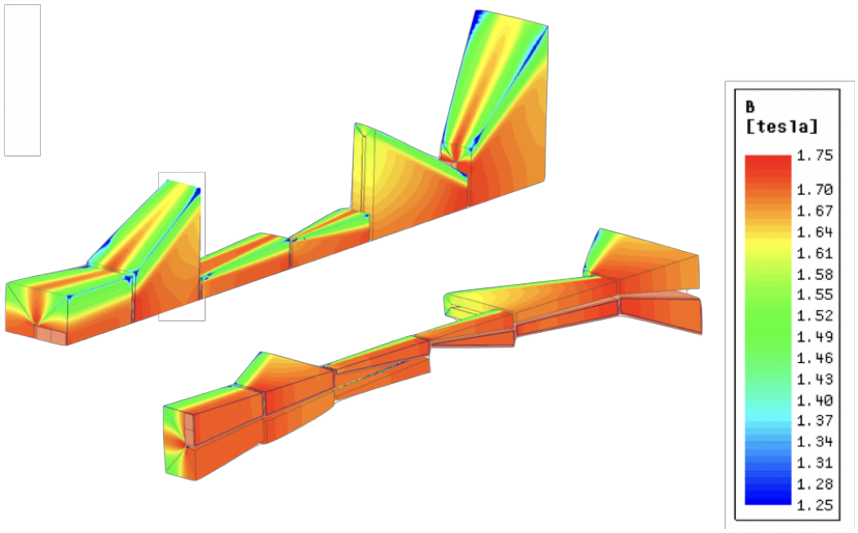


Figure 3.7: Distribution of the magnetic field in the magnet. A half of the shield is shown in Y and X cuts of the shield. [50]

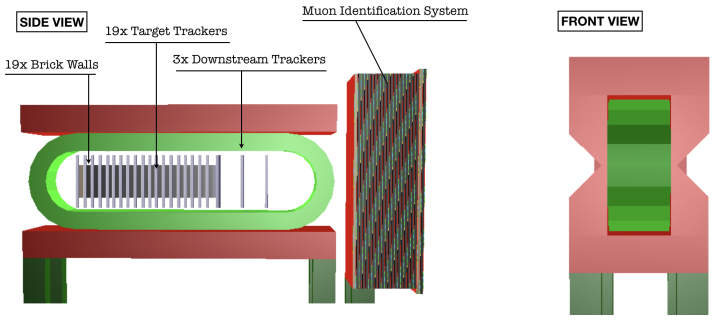


Figure 3.8: A schematic view of the SND detector. [50]

for the entire magnet. The simulated field map is presented in Figure 3.7.

After the current status of the optimisation was revisited, it became clear that a new iteration of the optimisation and a dedicated test-beam experiment are needed to validate the simulation and engineering grounds of the muon shield.

In particular, construction limitations, the realisability of the field and the efficiency of the design have to be taken into account in the new optimisation. The details of these concerns will be covered later in the manuscript.

Finally, it was understood that the effect of large-energy loss and large-angle scattering might affect the result significantly, since there is no experimental data available for such rare processes at the energies of muons above 10 GeV. Validation of the simulation of these rare processes motivates the construction of a dedicated test-beam experiment.

3.4 The SHiP detector

3.4.1 SND detector

After the muon shield, two special purpose detectors are located. The first one is the Scattering and Neutrino Detector(SND). The main goal of this detector is to detect LDM via its scattering on the dense material of the detector and study tau-neutrino physics. The schematic view of the detector can be seen in Figure 3.8.

The main part of the detector is a 1.2 T magnet that hosts emulsion target bricks

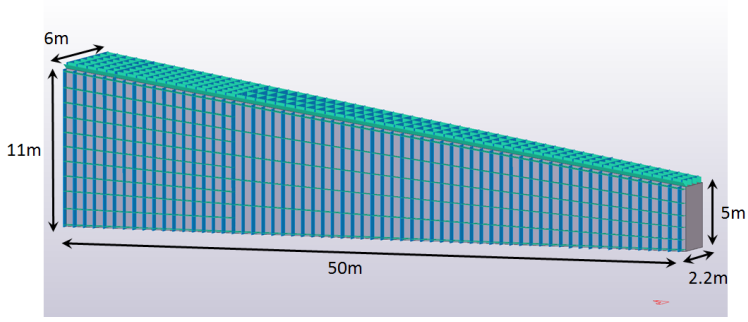


Figure 3.9: View of the SHiP decay volume. [50]

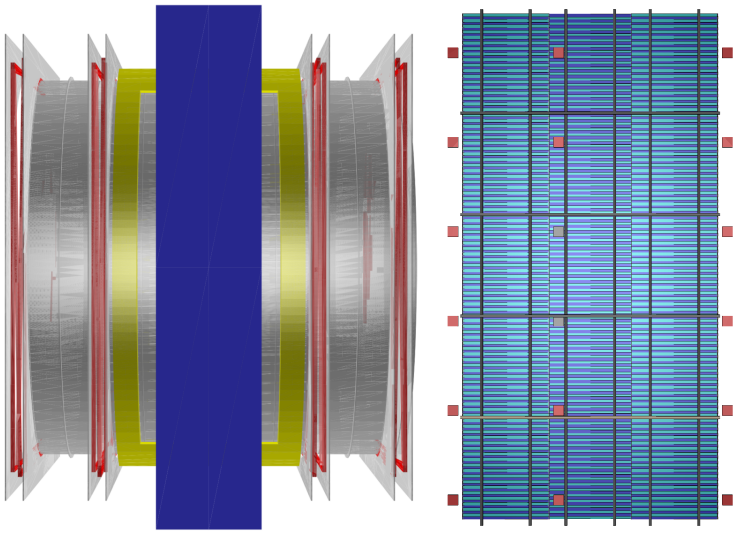
alternating with Target Trackers (TT) and enclosed with a downstream tracker. Each emulsion target brick consists of an Emulsion Cloud Chamber (ECC) made of 1 cm thick lead planes interleaved with an emulsion, and enclosed by a Compact Emulsion Spectrometer (CES). The ECC acts as a calorimeter and tracker simultaneously: with the emulsion in place, it is possible to reconstruct individual tracks with μm accuracy and the energy of electromagnetic showers with a resolution of about 15%. The main purpose of the CES is to reconstruct the charge and momentum of outgoing muons and pions. This is especially crucial for tau-neutrino physics. The transverse size of the ECC and CES is $40 \times 40\,\mathrm{cm}^2$ and there are four ECCs located in a XY-plane.

The main purpose of the Target Trackers is to provide a timestamp for the events in the emulsion films and link tracks in the emulsion to those reconstructed in the downstream tracker and muon identification system. The Target Tracker planes are located after each ECC brick and have the size of $80 \times 120\,\mathrm{cm}^2$. Three TT planes located after the ECC bricks constitute a downstream tracker and are needed to measure the charge and momentum of high-energy muons. The TT is based on scintillating fibre technology.

Finally, the muon identification system is located outside the magnet. Its main goal is to identify muons originating inside ECC bricks and those penetrating the muon shield. Thus, the muon identification system also acts as a veto detector in front of the decay volume and covers the whole entrance window of size $2 \times 5\,\mathrm{m}^2$. The muon identification system consists of 12 RPC chambers interleaved with 13 iron slabs.

3.4.2 Decay Volume

Since SHiP is a designed zero-background experiment, it requires a large clean area where the decays of hidden particles may be expected. To achieve this, the 50 m long decay volume (DV) is located after the SND detector and its view is presented in



(a) (b)

Figure 3.10: Layout of the (a) magnetic spectrometer, (b) timing detector. [67]

Figure 3.9. To minimise the neutrino-induced interactions inside the decay volume, it is kept under a pressure of 1 mbar. With such a low pressure, the neutrino interactions mainly occur in the decay volume walls. Deep inelastic scattering of muons, as well as products of neutrino interactions, can be efficiently suppressed by a cut on the impact parameter that does not point to the target. However, to ensure zero background, the decay volume is surrounded by the liquid scintillator system for the detection of associated activity in the proximity of the decay volume walls. On one side, the shape of the decay volume is dictated by the geometrical area that is cleaned from muons by the muon shield. On the other, the length of the DV is restricted by the decay probability of HS particles and the DV cross-section is motivated by the HS P_T spectrum.

3.4.3 Hidden Sector detector

The Hidden Sector detector is located after the decay volume and consists of the spectrometer, the timing detector, the ECAL and the muon identification system.

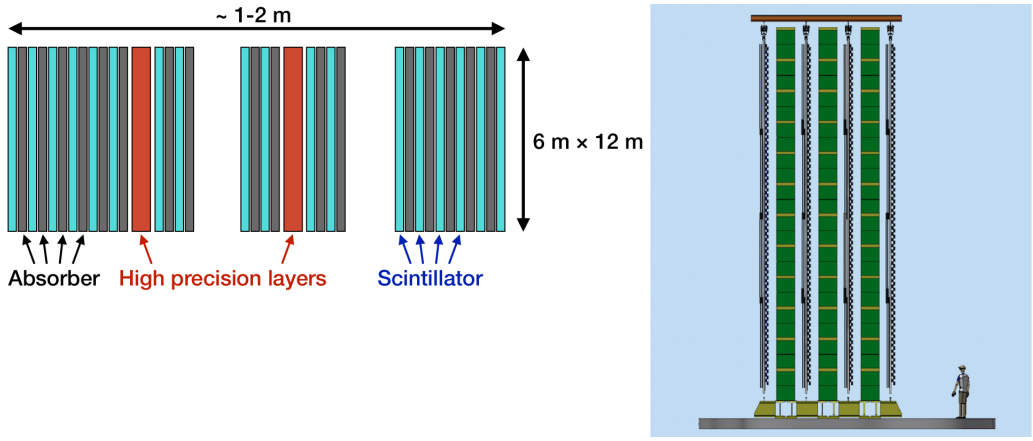
The magnetic spectrometer is located right after the decay volume. The main goal of the spectrometer is to provide track reconstruction and momentum measurement of charged particles. The reconstruction quality of the spectrometer must be high enough to allow the location of the decay vertices of hidden sector particles and to match the tracks with the segmentation of the timing detector to suppress the

background efficiently. The tracker consists of four tracking stations and a dipole magnet, located after the second station. The geometrical acceptance of the tracker is 5 m in X and 10 m in Y and the straw drift tubes, made of very light material, are used as a technology. Each tracker has the tubes tilted at different angles, thus allowing the identification of X and Y coordinates of the track. The magnetic field is oriented along the X-axis and is about 0.14 T in strength. The resulting configuration has a resolution of $O(1-2)$ mm in the XY plane and $O(10)$ cm in the Z-axis. The resulting momentum resolution is $(\sigma_p/p)^2 \approx (0.49\%)^2 + (0.022\%/[GeV])^2 p^2$. The main technological challenge is the elongation and sagging of the straw tubes. The layout of the detector is presented in Figure 3.10a.

The timing detector is located right after the spectrometer and covers an area of $5 \times 10\,\mathrm{m}^2$. The main goal of the timing detector is to provide time of coincidence resolution for the track candidates as well as start times for the spectrometer drift time measurement. To suppress the residual muon background to a manageable level, the timing resolution of the detector must be less than 100 ps. Two technologies are currently being considered for the timing detector: plastic scintillator with SiPM readout and time measuring resistive plate chambers. The SiPM option detector layout is presented in Figure 3.10b. It consists of three rows of scintillator bars being read out on both sides. Such a configuration allows a timing resolution of $O(85)$ ps to be achieved, thus meeting the requirements for the timing detector [50].

In the SHiP experiment the calorimeter system is composed of only one detector, called SplitCal. The main goal of the SplitCal is to identify electrons, photons and pions as well as to measure electron and photon energy in the $1-100$ GeV range. It must also provide direction information for photons. The SplitCal is a longitudinally segmented electromagnetic calorimeter with the capability of reconstructing trajectories of photons with a precision of a few mrad. This is especially important to identify the decay $ALP \to \gamma\gamma$. The layout of the calorimeter is presented in Figure 3.11a. It has a $0.5X_0$ thick lead absorber interleaved with a scintillator, and three high-resolution (200 μm) gas layers located in a way that maximises the angular resolution of low and high energy photons. The layers can precisely reconstruct the barycentre of the shower transverse profile and thus measure the angle of the photon with mrad precision. An energy resolution $\sigma(E)/E$ of about $12\%/\sqrt{E}$ is achieved.

Finally, the muon system encloses the HS detector. The layout is presented in Figure 3.11b. The main goal of the system is to identify muons with efficiency greater than 95% and in a wide range of momentum spanning from 5 to 100 GeV. It also contributes to background rejection by having a ~ 350 ps resolution. With dimensions of 6 by 12 m, the muon system consists of four active stations, made



(a) (b)

Figure 3.11: Layout of the (a) SplitCal, (b) Muon detector. [67]

from scintillating tiles of size $10 \times 20\,\mathrm{cm}^2$ and iron absorber walls, $3.4\lambda_{int}$ each. A muon has to have an energy of at least 2.6 GeV to reach the first station and at least 5.3 GeV to reach the last station. The granularity of the system is driven by the multiple scattering of muons inside the detector. A timing resolution of $O(400)$ ps was measured on the small prototype [50].

The hit pattern in the muon system and energy deposition in the SplitCal defines the identification of the track as a muon or a pion. Tracks that do not reach the muon system are tagged as pions. For tracks with momentum higher than 3 GeV, the pion misidentification rate is less than 0.1% with a muon misidentification efficiency of 99%.

3.5 Background rates

There are three main categories of background in SHiP: neutrino-induced background, muon deep inelastic scattering (DIS) and muon combinatorial background. The categories are schematically presented in Figure 3.12. The background from cosmic muons was found to be negligible [50].

The neutrino-induced background consists of long-lived neutral SM particles, such as K_L^0, that are produced in neutrino inelastic interactions in the material of the detector and the cavern walls. These neutral particles can decay inside the decay volume to final states similar to HS final states. The estimated flux of neutrinos from the target is $\sim 2 \times 10^{18}$ particles over five years of SHiP operation. This will result in $\sim 10^5$ neutrino interactions nearby the decay volume, the major fraction of

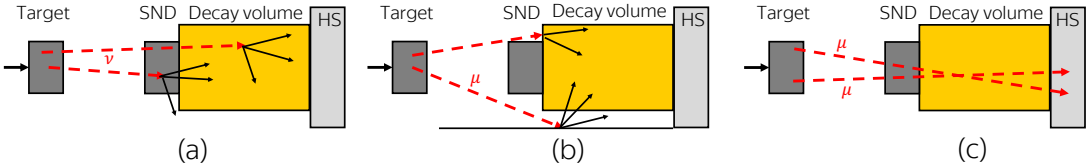


Figure 3.12: Background types in SHiP: (a) Neutrino-induced, (b) muon DIS, (c) muon combinatorial.

Table 3.2: Common selection criteria used for background rejection

Cut	Value
Track momentum	$> 1.0\,\text{GeV}$
Children distance of closest approach	$< 1\,\text{cm}$
Decay vertex position	$> 5\,\text{cm}$ from DV wall
IP w.r.t. to target (fully reconstructed)	$< 10\,\text{cm}$
IP w.r.t. to target (partially reconstructed)	$< 250\,\text{cm}$

which will happen in the SND detector and decay volume walls. By applying the common selection cirteria, presented in Table 3.2, and the SBT veto, it is possible to reduce the neutrino-material interactions to $0.1 - 0.3$ events in five years [50]. The estimated number of neutrino-air interactions inside the decay volume is 10^{-2} due to the low (1 mbar) pressure inside it.

Similarly to the neutrino-induced background, the muon DIS interactions will produce long-lived particles in the DIS of muons at the cavern walls, the material of the decay volume and the SND detector, resulting in about 2×10^8 interactions. However, after applying the selection criteria from Table 3.2 and the SBT veto, the background level from DIS is estimated to be less than 6×10^{-4} events in five years.

The muon combinatorial background results in two reconstructed charged tracks in the HS spectrometer, which mimic the signal final states with two leptons. If these tracks coincide in time and position, they might be misidentified with the decay products of a hidden sector particle. The current design of the muon shield reduces the flux of muons to the rate of 50 kHz. Assuming a spill duration of 1 s and two spills per minute, this rate will result in $\sim 10^{16}$ pairs of muons in the acceptance of the detector over five years of operation. Applying the selection criteria from Table 3.2 this number is further reduced to 10^9 events in five years. To further suppress the background, the SHIP's timing detector, with a resolution $\sigma_t \approx 100\,\text{ps}$, is used. The probability of obtaining two muons within the timing window of 340 ps (that is three times the resolution of the timing detector) is about 0.5×10^{-10}, where Poisson statistics was used to compute the probability. Thus, the number of di-muon events in five years, where both tracks lie in the same time slot of 340 ps, is $\approx 5 \times 10^{-2} - 10^{-1}$.

This number can be further reduced by a factor of 10^{-4} by utilising the SBT veto criteria.

Finally, there is a different type of background for the LDM search in the SND detector via LDM scattering off electrons. After applying geometric, kinematic and topological cuts, two sources of the background dominate. Topologically indistinguishable neutrino elastic scattering on electrons or protons constitutes the main source of the background, resulting in 207 events in five years. Quasi-elastic scattering on an electron, with an unidentified soft outgoing proton is the second most common background and adds 18 events in five years [69].

3.6 Sensitivity to Hidden Sector particles and Light Dark Matter

In experiments searching for visible decays of the HS particles, the sensitivity to individual HS particles may be determined by computing the estimated number of events:

$$N_{events} = N_{prod} \times \mathcal{P}_{decay} \times A_{tot}, \tag{3.1}$$

where N_{prod} accounts for the number of HS particles produced in the experiment. This term includes the number of protons on target, the cross-section of heavy flavour production and the probability of their subsequent decay into HS particles. The $\mathcal{P}_{decay}$ term resembles the probability of the HS particle to decay inside the dedicated decay volume. Finally, A_{tot} accounts for the geometrical acceptance of the experiment and efficiency of the detectors.

In Figure 3.13, 3.14a, the sensitivities to HNLs, Dark Scalars, ALPs and Dark Photons are shown. Solid regions correspond to the areas either excluded by previous experiments or inaccessible from a theoretical perspective. Dashed lines correspond to the proposed experiments. As can be seen, the SHiP experiment outperforms all proposed experiments in all portals, save one.

The number of events of Light Dark Matter scattering can be estimated as:

$$N_{events} = \sigma(\chi e(N) \to \chi e(N)) \times \frac{\phi}{S_{det}} \times N_{e(N)} \tag{3.2}$$

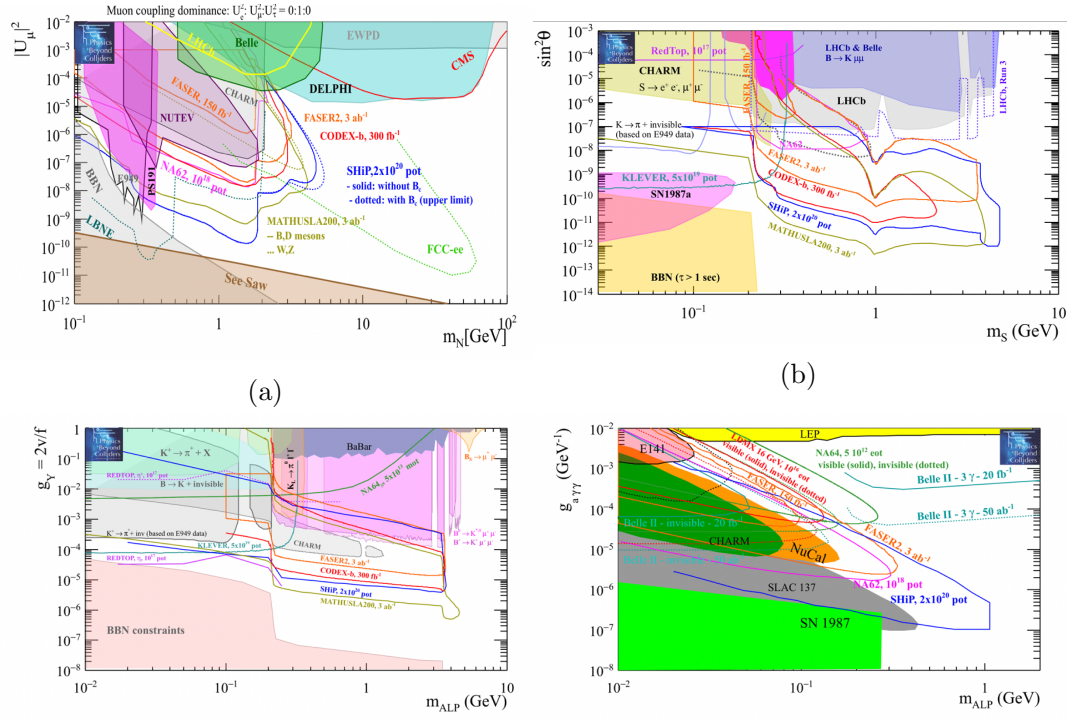


Figure 3.13: Existing limits and limits from the proposed experiments for Hidden Sector models: a) Heavy Neutral Leptons b) Dark Scalar c) ALP couples to SM fermions d) ALP couples to SM photons. [12]

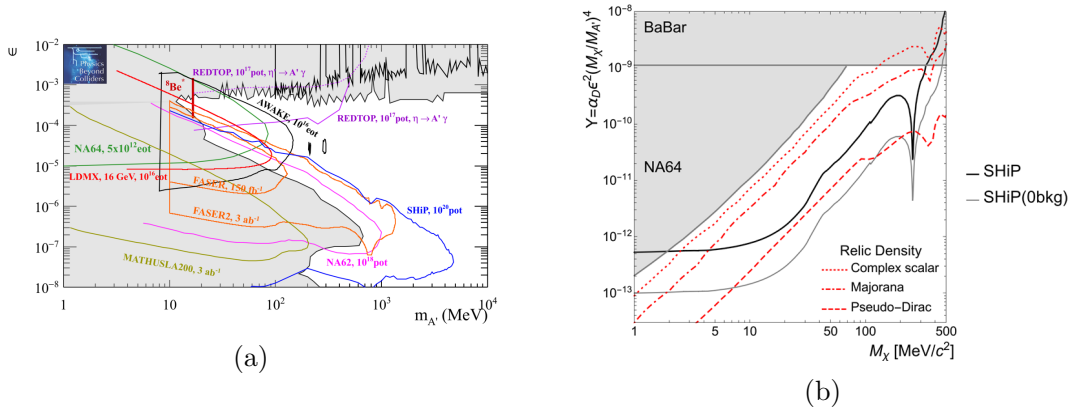


Figure 3.14: Existing limits and limits from the proposed experiments to a) Dark Photon b) Light Dark Matter. [12]

where $\sigma(\chi e(N) \to \chi e(N))$ is the model-dependent cross-section of LDM scattering on electrons (nucleons), ϕ is the flux of the LDM, S_{det} is the transverse (w.r.t to the flux) area of the detector and $N_{e(N)}$ is the number of scattering centres - electrons or nucleons. The flux ϕ is given by the $\phi \sim N_{POT} \times r_{parent} \times \epsilon^2 \times A_{geo}$, where N_{POT} is the number of protons on target, r_{parent} is the contribution from the production mechanisms, discussed in sub-section 2.6, ϵ is coupling constant and A_{geo} is the geometrical acceptance. The resulting sensitivity, calculated for the SHiP experiment, is presented in Figure 3.14b. It can be seen that the experiment improves the existing experimental constraints by several orders of magnitude.

Chapter 4

Algorithms for Optimisation

Machine Learning (ML) methods are applied to a wide range of High Energy Physics (HEP) tasks. Such tasks include particle identification [21], classification of events type [20], anomaly detection in the analysis [23] and tracking [22]. More recently, ML methods have been successfully used for matrix element computations and real-time analysis [70].

Monte Carlo simulators, such as GEANT4 [71] are a crucial part of the HEP domain and are used extensively since the creation of computers. However, HEP simulations are very computationally expensive. For example, an order of a trillion simulated collisions is required to achieve statistical accuracy for the hypothesis testing in the HL-LHC environment. Simulating the detector response of proton-proton collisions can take up to several minutes. In SHiP, the simulation of proton interactions with the beam target and the subsequent muon propagation can take days of computing time, even when using CERN High Throughput Computing (HPC) system. Thus, many efforts have been made to speed up the simulation in HEP using modern ML techniques. For example, some of the earliest work was done to speed up the simulation of particle showers in calorimeters utilising generative neural networks [72]. In SHiP generative networks were used to speed up the proton-target collisions [73].

Finally, modern simulators have a large number of parameters that affect the outputted events and their distributions. Tuning the high dimensional parameter space of a simulator to achieve the desired output can often be computationally expensive. Machine learning methods such as Bayesian optimisation [25, 26] and evolutionary strategies [27] have been applied for the optimisation tasks in high energy physics domain [29, 74]. For example, Ref. [74] utilised Bayesian optimisation to fine-tune Pythia parameters to reproduce data in e^+e^- collisions.

While many of the problems above are relevant to the SHiP experiment, this thesis mainly explores the problem of optimising the simulator's parameters. Thus, the main definitions, formulas and concepts relevant to the problem of simulator optimisation will be described below. The mathematical notation used throughout the thesis is summarised in Table 4.1.

Table 4.1: Mathematical notation

x	Scalar variable x
$\boldsymbol{x}$	Vector variable $\boldsymbol{x}$
$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$	n-dimensional vector $\boldsymbol{x}$ with components $x_1, x_2, \ldots, x_n$
$\mathbb{R}^n$	n-dimensional space of real-valued numbers
$\|\|\boldsymbol{\psi}\|\|_{L_2}$	Norm of n-dimensional vector $\boldsymbol{\psi}$: $\|\|\boldsymbol{\psi}\|\|_{L_2} = \sqrt{\psi_1^2 + \psi_2^2 + \cdots + \psi_n^2}$
$\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$	Nabla (gradient) operator acting on variable $\boldsymbol{x}$ of f
$\sigma(x)$	Sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$
$\mathbb{E}[x]$	Expected value of a random variable x
$\arg\min_x$	Arguments of the minima over variable x
$p(x), q(x)$	Probability distribution over x
$p(x\|\psi)$	Conditional probability distribution over x given ψ
$\mathbb{E}_{p(x\|\psi)}[R(x)]$	Expected value of a function $R(x)$ w.r.t. to the $p(x)$ distribution: $\mathbb{E}_{p(x\|\psi)}[R(x)] = \int R(x)p(x\|\psi)dx$
$\mathcal{N}(\mu, \sigma)$	Normal distribution with mean μ and standard deviation σ
$\mathrm{U}[L, R]$	Uniform distribution over interval [L,R]
$\mathcal{R}(x)$	Objective/loss function over x

4.1 Optimisation problem

To formulate an optimisation problem, one starts with the definition of the function to be minimised. Throughout this thesis the names **objective** and **loss** function

are used interchangeably. We define the objective function $\mathcal{R}(\boldsymbol{\psi})$ as a function that depends on the parameter vector $\boldsymbol{\psi}$. This function returns a number that in some way characterises the performance of the optimised object (e.g. the total flux of particles through the detector). Thus, the goal of the optimisation is to find a $\boldsymbol{\psi}^*$ that minimises the objective function:

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \mathcal{R}(\boldsymbol{\psi}) \tag{4.1}$$

The optimisation may be subject to some constraints, for example, the constraint on the length of the muon shield in the SHiP experiment.

As an example, let the objective function be the 1-dimensional Higgs potential:

$$\mathcal{R}(\psi) = -5\psi^2 + \psi^4 \tag{4.2}$$

In this simple scenario, one can find the minimum analytically by computing first and second-order derivatives of $\mathcal{R}$ and requiring $\mathcal{R}^{'}(\psi) = 0, \mathcal{R}^{''}(\psi) > 0$. There are two degenerate minima at $\psi^* = \pm\sqrt{5/2}$. However, such a simple function does not resemble what one actually observes in a simulation or experiment. For example, in an experiment there will always be some noise observed. In a simulator, one often does not even know the exact dependency of the loss on the inputs.

For instance, the parameters ψ may be treated as unknown, yet affect the value of the observables y. The observable y may be a random variable, sampled from a normal distribution with mean equal to ψ and fixed standard deviation σ: $y \sim \mathcal{N}(y; \psi, \sigma)$. Now, the objective function $\mathcal{R}$ depends on ψ implicitly: the only variable that we actually observe is y:

$$\mathcal{R}(y) = -5y^2 + y^4, \quad \text{s.t.} \;\; y \sim \mathcal{N}(y; \psi, \sigma) \tag{4.3}$$

Such formulation results in a distribution over values of $\mathcal{R}$ as a function of ψ, rather than an exact value, as in the initial example. The corresponding plot of the mean and the standard deviation of $\mathcal{R}(\psi)$ is shown in Figure 4.1.

Since $\mathcal{R}$ is a random variable now, it is natural to reformulate the optimisation problem in terms of the expected value of the function:

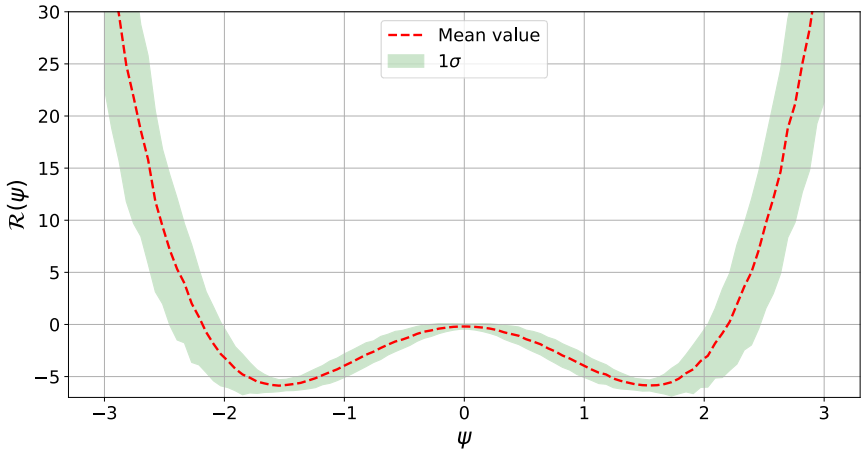


Figure 4.1: Mean (red line) and one standard deviation (green area) of $\mathcal{R}(\psi)$ as a function of ψ.

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \mathbb{E}_{p(y;\psi,\sigma)}[\mathcal{R}(y)] \tag{4.4}$$

In the Higgs example above that corresponds to the minimum of the red curve.

With the implicit dependency on ψ, it is impossible to solve the problem analytically anymore since there is no closed-form solution due to the presence of the expectation value $\mathbb{E}$. Although an analytical solution is not available, various methods exist to solve such types of problems. To simplify the notation for further examples, we define $f(\psi) = \mathbb{E}_{p(y;\psi,\sigma)}[\mathcal{R}(y)]$.

4.2 Existing optimisation algorithms

When the exact analytical solution can not be computed, one of two classes of methods is usually applied. Methods from the first class try to approximate the derivative $\nabla f(\psi)$ in various ways and use the *gradient descent* technique to find the minimum of the function. Examples of such methods include numerical differentiation [30] or REINFORCE gradient estimator [33]. A well-known example of gradient-based optimisation in HEP is the MINUIT package [75].

Another set of methods is called derivative-free and includes many methods that perform optimisation without estimating or using available information about a derivative. The most common ones are Bayesian optimisation (BO) [28] and evolutionary algorithms [29], which are a large collection of generic population-based

optimisation methods.

4.2.1 Gradient descent optimisation

Gradient descent (GD) is one of the oldest optimisation techniques used in science. The idea behind the algorithm is to iteratively move in the direction opposite to the gradient $\nabla_\psi f(\psi)$ starting from some initial guess ψ_0. At each iteration the gradient at point ψ_i is estimated and the new point is computed as $\psi_{i+1} = \psi_i - \alpha \nabla_\psi f(\psi)$. The procedure is repeated until some convergence criterion is reached. The algorithm is summarised in Algorithm 1. The parameter α is called learning rate and controls the size of the update. If it is too large, the algorithm can diverge, whereas if it is too small, the optimisation might get stuck at some point.

Algorithm 1 Gradient descent

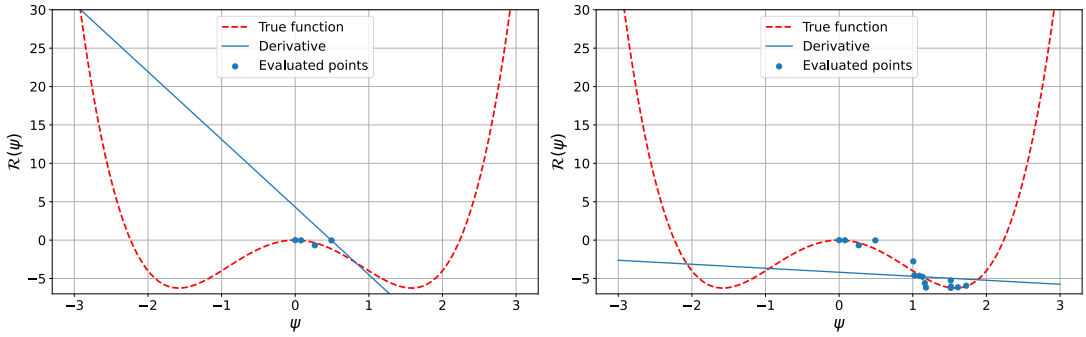
Require: Initial point ψ_0, learning rate α
1: **while** ψ has not converged **do**
2: $\quad \psi_{i+1} = \psi_i - \alpha \nabla_\psi f(\psi)$
3: **end while**

If the function f is differentiable, there are theoretical guarantees on the convergence rate and an optimal value of α. However, in the formulation of the problem above the analytical gradient $\nabla_\psi f(\psi)$ is not known and needs to be approximated.

The simplest way to approximate the gradient is to use numerical differentiation. Finite difference methods are simple yet effective and provide theoretical guarantees on the convergence rate. To estimate the gradient, it is enough to compute the function value at some point $\psi + h$, where h is called step size and is usually small. The so-called central derivative, which uses both $f(\psi + h)$ and $f(\psi - h)$ has a smaller error and is often preferred:

$$\nabla_\psi f(\psi) \approx \frac{f(\psi + h) - f(\psi - h)}{2h} \tag{4.5}$$

Once the derivative has been approximated, it can be used to perform gradient descent optimisation using Algorithm 1. Figure 4.2 shows points evaluated during GD optimisation using the above approximation of the gradient applied to the Higgs potential example starting at point $\psi_0 = 0$. As it can be seen, the evaluated points and the gradient direction do not match perfectly with the true mean value of the function (red line) due to the noise in the observations y. Nevertheless, the GD was able to converge to one of the function's minima.



(a) (b)

Figure 4.2: True function value (red) and evaluated points (blue) during GS optimisation with numerical derivatives after (a) 4 iterations, (b) 15 iterations. Blue line shows gradient direction.

In this simple example, the numerical estimation of the derivative works well and is enough to find a minimum of the function. However, in more challenging problems, the drawbacks of the method start to show up. Specifically, the numerical derivatives are mostly used for noiseless functions, and may struggle to accurately estimate the derivative for a very noisy or fast-changing function. The latter comes from the fact that numerical derivatives linearly interpolate the function and may estimate the function inaccurately if the step size h is too large. However, if h is too small, the noise in the function evaluations will dominate, and the numerical estimation will be inaccurate again. In addition to that, only two evaluations of the function $f(\psi)$ per iteration were required in the simple example above. In general, numerical derivatives require $O(2d)$ function evaluations, where d is the dimensionality of the optimised parameter ψ. In the case of high-dimensional parameters, and when the evaluation of the function f is computationally expensive, this might induce a high computation cost for the method, and, in turn, slow optimisation.

Other, more recent methods, such as REINFORCE [33] and the reparametrisation [76] exist, that allow approximating the gradients in Equation 4.4. The details of the algorithms are not covered in the thesis for brevity. However, the latter algorithm is applicable only in a minimal number of problems, when it is possible to reparametrise the distribution $p(y; \psi, \sigma)$. The former algorithm found application in many domains of machine learning and only requires the existence of $\nabla_\psi \log p(y; \psi, \sigma)$. The main disadvantage of the algorithm is a very high variance of the estimation of the gradient, meaning that estimations of the gradient of Equation 4.4 done several times at the same point ψ, would differ from each other significantly. This flaw slows down the convergence of the algorithm or requires a large number of function $f(\psi)$ evaluations, which again is computationally costly.

4.2.2 Bayesian optimisation

Bayesian optimisation (BO) is the most common approach to optimise non-differentiable, noisy and computationally expensive functions. The algorithm builds a probabilistic *surrogate* model of the function $f(\psi)$ in the whole search space of parameters based on the initial observations. It starts with sampling points, usually at random, to create an initial dataset $D = \{\psi_t, f(\psi_t)\}_{t=1}^{T}$. At each iteration t, BO tries to approximate the behaviour of the function $f(\psi)$ with a surrogate model, using the data available in D. It then uses an acquisition function β that estimates the most promising point ψ_{t+1} to evaluate the next. The point $(\psi_{t+1}, f(\psi_{t+1}))$ is added to the dataset D and the process is repeated until the optimisation budget is exhausted. A summary of the algorithm is presented in Algorithm 2.

Algorithm 2 Bayesian optimisation

Require: Search space U, surrogate model $\mathcal{M}$, acquisition function β
1: Sample initial dataset $D = \{\psi_t, f(\psi_t)\}$ from U
2: **while** computationally feasible **do**
3: Define mean $\mu(\psi|D)$ and variance $\sigma(\psi|D)$ using surrogate $\mathcal{M}$
4: Find $\psi^* = \arg\max_{\psi} \beta(\mu(\psi|D)), \sigma(\psi|D))$
5: Add new point $(\psi^*, f(\psi^*))$ to the dataset D
6: **end while**

There are a couple of design choices in Bayesian optimisation. The first one is the surrogate model $\mathcal{M}$, which should provide an approximation of f at any point ψ by estimating its mean $\mu(\psi|D)$. This corresponds to the red curve in the Higgs example above. The surrogate should also provide an uncertainty estimation that represents the level of the surrogate confidence in its prediction. Many machine learning models, such as random forest, are suitable for the role of the surrogate [77], but the most common choice is Gaussian Processes (GP) [78].

Secondly, the acquisition function must be specified. Different variants of the functions exist, but all of them return high scores for regions that have either high variance (meaning high, but uncertain reward) or low mean value (meaning modest, but certain reward). The choice of the function and its parameters controls the *exploration/exploitation trade-off*. This means that, depending on the chosen function and its parameters, the BO would either explore the search space (giving more weight to high uncertainty regions) or exploit gathered information (giving more weight to the value of the surrogate mean). One of the most commonly used acquisition functions is the Expected Improvement (EI):

$$\beta = \mathrm{EI}(\psi) = max(f(\psi^+) - \hat{f}(\psi), 0), \tag{4.6}$$

where ψ^+ is the best point so far ($f(\psi^+)$ is the smallest) and $\hat{f}(\psi)$ is the value obtained from the surrogate model.

Thirdly, GP has a parameter called the kernel function (or covariance function) that significantly affects how the mean $\mu(\psi|D)$ and variance $\sigma(\psi|D)$ are modelled. Most importantly, for any two points ψ and $\psi^{'}$ the covariance function often depends uniquely on the distance $d(\psi, \psi^{'})$ between the points. The optimisation task usually determines the choice of the kernel.

Coming back to the example with the Higgs potential, the Bayesian optimisation with GP and EI as an acquisition function was applied to the problem. The resulted mean surrogate function $\mu(\psi)$ and standard deviation $\sigma(\psi)$ after 1, 5 and 15 iterations respectively are presented in Figure 4.3. The plots on the right represent the EI function and the point that maximises it. The BO starts by exploring the space first and by the fifth iteration it already adequately represents the function in the interval from -1 to 1. As expected, the uncertainty near the evaluated points (red dots) is small. By the fifth iteration, the [-1, 1] region is well explored, and the EI is maximised in the areas with high uncertainty, as seen in the right plot of Figure 4.3b. Finally, by the end of the optimisation, the EI function is almost zero everywhere except at the two minima. We notice that BO was able to find both of the degenerate minima in this example, unlike the gradient-based method, which is always capable of finding only one local minimum.

Again, for such a simple example BO outperforms GD and is capable of finding both minima. However, this is not generally the case, and although BO is considered to be a global optimisation algorithm, it is almost impossible to find a real global minimum in high-dimensional spaces.

Speaking about drawbacks, there are two main reasons why applying BO is challenging. The first, and most commonly referenced one, is that the estimation of $\mu(\psi)$ and $\sigma(\psi)$ requires inverting the covariance matrix, a process which scales as $O(n^3)$, where n is the number of training points in the dataset D. Since the algorithm constantly adds points to the dataset D, this inversion operation quickly becomes expensive. Typically, when the dataset size reaches $O(10^3)$ samples the GP-based BO becomes too costly to compute. Usually, at this point, the time required for one iteration of BO exceeds the time needed for the evaluation of the function $f(\psi)$.

The second problem that is often encountered is the "curse of dimensionality" in high-

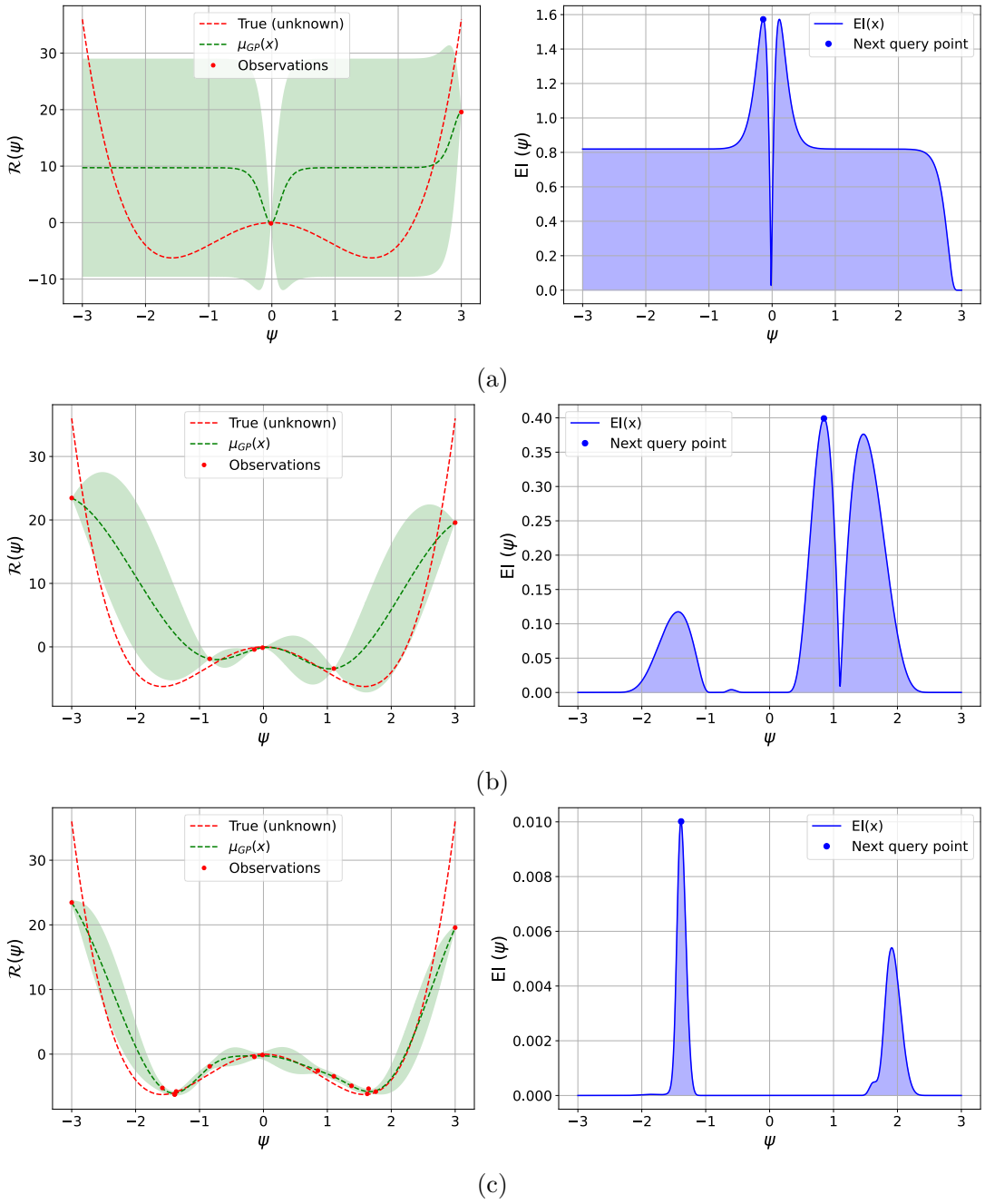


Figure 4.3: Left: True function value (red), $\mu(\psi)$ (green) and standard deviation $\sigma(\psi)$ (shaded green) obtained from GP surrogate model. Right: Expected improvement. After iteration (a) 1, (b) 5, (c) 15 of the Bayesian optimisation procedure.

dimensional spaces. The "curse of dimensionality" states, that in high-dimensional spaces most of the volume is concentrated on the boundary of the space. This is illustrated by the ratio of volumes V of two concentric D-dimensional balls with radius R and $R - \Delta R$.

$$\frac{V(R - \Delta R)}{V(R)} = o((1 - \Delta)^D) \tag{4.7}$$

Since $\Delta < 1$ this ratio goes to zero as D grows, meaning that the volume is concentrated near the boundaries of the space.

By construction, BO assumes that the solution lies close to the origin of the search space. Due to the "curse of dimensionality" most points would be located on the boundaries and far away from each other. This will induce the covariance function, which depends on the distance between points, to be large for the boundary region and result in a large variance $\sigma(\psi)$. Thus, BO would prefer to evaluate points far away from the centre, which contradicts the initial assumption. In practice, this leads to the stagnation of BO. This was observed in the previous iteration of the muon shield optimisation [79].

4.2.3 Evolutionary algorithms

Evolutionary algorithms describe many types of genetic-inspired optimisation methods. The main idea is to sample K points at random, evaluate the function $f(\psi)$ at those points and then select K best points that correspond to K minimal values of f. Afterwards, the new points are constructed by crossover and mutation of the K best samples. The newly obtained points then replace the worst-performing samples in the initial dataset, and the process repeats. The algorithm is summarised in Algorithm 3.

Algorithm 3 General purpose evolutionary algorithm

Require: Number K best samples
1: Generate initial population $D = \{\psi_t\}$ randomly
2: **while** computationally feasible **do**
3: Compute $f(\psi)$ for each point in D
4: Select K best points ψ corresponding to K minimal values of $f(\psi)$
5: Breed the K best selected points
6: Replace the least fit samples from D with K breed points
7: **end while**

By the crossover and mutation operations one often means some mathematical operations over vectors. For example, for two vectors $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ a crossover operation will randomly substitute some of the $\boldsymbol{\psi}_1$ components with components from $\boldsymbol{\psi}_2$ and vice versa. The mutation will change the components of the vector randomly or following some specific distributions. It is thanks to this idea of finding the best performing sample by following biological evolutionary principles that this optimisation method is called evolutionary.

The strongest advantage of this algorithm is, at the same time, its biggest flaw. On one side, the algorithm does not make any structural assumptions about the function $f(\psi)$ nor creates any function surrogate. It directly uses the function to evaluate new samples, thus, avoiding any bias in the estimation of the function value. On the other hand, because the function is evaluated for all samples and the samples are obtained at random, the algorithm usually requires a large number of function calls to find the best performing sample. For computationally expensive functions such an approach is ineffective and results in wasteful usage of the resources.

4.2.4 Motivation for a new algorithm

In HEP, one often encounters computationally expensive simulations. Typically, the objective of the simulation is to prove or optimise the design of the detector by using, for example, GEANT4 [71]. Such simulations can be treated as *black-box* simulations, since a physicist does not know the *exact* relation between the inputs ψ and the outputs y in the simulator, as in our toy Higgs example before. We can think of the simulator as a black-box function F, that takes as input some physical parameters of the simulation $\boldsymbol{x}$ and parameters of the detector $\boldsymbol{\psi}$, and returns observable variables $\boldsymbol{y}$: $\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{\psi})$. For example, $\boldsymbol{x}$ can be a vector of energies of colliding particles, $\boldsymbol{\psi}$ can characterise the location or shape of the detectors, and the observables $\boldsymbol{y}$ can be a detector response, such as the particle's energy or position.

However, there is one important part missing from such a formulation: the stochastic nature of all HEP simulations. Any process in a simulation, whether a matrix element computation, a simulation of an electromagnetic shower, or simply a particle trajectory prediction, is non-deterministic. Every simulator has a random seed, which we denote by z, that introduces randomness to the simulation by trying to mimic the stochasticity of physics processes. Thus, in reality $\boldsymbol{y}$ is deterministic *only* if all parameters, including the random seed, are fixed: $\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{\psi}, z)$. Usually, one does not want to fix the random seed z to ensure the stochasticity is properly taken into account in the simulation. In such cases, the outcome of the simulation $\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{\psi})$

becomes a random variable sampled from some *unknown* distribution $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$. This distribution is often called *intractable*, because we can obtain samples from it, by running a simulator, but the exact analytical form of the distribution is not known.

Summarising the problem, one has a black-box simulator F that produces a random variable observable $\boldsymbol{y}$ sampled from $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$: $\boldsymbol{y} = F(\boldsymbol{x},\boldsymbol{\psi}) \sim p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$. The exact relation between the observable and the parameters is not known. The physicist wants to minimise some objective (loss) function $\mathcal{R}$ of the observables:

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} f(\boldsymbol{\psi}) = \arg\min_{\boldsymbol{\psi}} \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})}[\mathcal{R}(\boldsymbol{y})] \tag{4.8}$$

This optimisation problem looks exactly like the toy example in Equation 4.4, meaning that we can use all the methods above to solve the problem. In fact, Bayesian optimisation was used in the previous iteration of the optimisation [79]. The question is, then, why a new algorithm is needed?

As outlined earlier, the aforementioned algorithms have some drawbacks. For example, numerical derivatives and evolutionary algorithms do not take into account that it is costly to evaluate function f. Moreover, the numerical derivatives provide only a linear approximation of the objective function's surface. The REINFORCE-based algorithm may suffer from high variance and have slow convergence rate, resulting in many evaluations of the function. Finally, Bayesian optimisation scales as the cube of the number of training samples, so it quickly becomes computationally infeasible. In addition, it is susceptible to the "curse of dimensionality" and thus may waste computational resources by evaluating irrelevant (close to the boundary of the search space) points. Leveraging the recent progress in the field of machine learning, we have tried to develop a new optimisation method that might be more suitable for the optimisation problem defined above. As it is shown in Chapter 6, this formulation of the problem is relevant to the muon shield optimisation in the SHiP experiment. Before describing the details of the algorithm, a short introduction into neural networks is provided in Section 4.3.1.

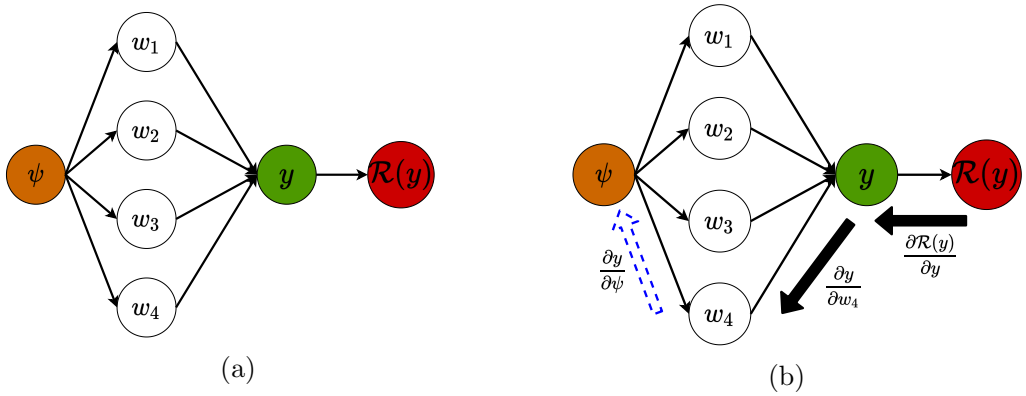


Figure 4.4: Example of a neural network. (a) Forward connections are shown. (b) Example of backward pass is added. Black solid arrows correspond to backpropagation over weights, and blue dashed arrows show backpropagation over inputs.

4.3 Neural Networks fundamentals

4.3.1 Training a neural network

Neural networks (NN) have been used in many applications of HEP long before their recent popularity. However, the computational expenses have been a limiting factor in the applicability of NNs to practical problems until the beginning of the previous decade. From that point onwards, the availability of computational resources has allowed for faster training of deep neural networks and resulted in a new discipline called Deep Learning (DL). Since then, DL led to major breakthroughs in computer science and found its application in natural sciences.

The simplest neural network consists of inputs $\boldsymbol{\psi}$, a *hidden layer* and outputs $\boldsymbol{y}$, as shown in Figure 4.4. The hidden layer is characterised by a learnable weight matrix $W = (w_1, w_2, \ldots, w_n)$. We set a d-dimensional vector $\boldsymbol{\psi} \in \mathbb{R}^d$ as input, an n-dimensional vector $\boldsymbol{y} \in \mathbb{R}^n$ as output and a weight matrix of shape $n \times d$: $W \in \mathbb{R}^{n \times d}$. In the example of Figure 4.4 the $\boldsymbol{\psi}$ and $\boldsymbol{y}$ are 1-dimensional, and the matrix W is a 4-dimensional vector. The relationship between the input and the output is given by matrix multiplication $\boldsymbol{y} = W\boldsymbol{\psi}$. Given the predictions of the NN, one wants to optimise the weights W such that some loss function $\mathcal{R}$ is minimised. As an example, the objective function can be a mean squared error (MSE) between NN predictions $\boldsymbol{y}$ and known answers $\boldsymbol{y}_{train}$.

The idea behind training a NN is the same as that of the gradient descent. One would like to update the NN weights W until convergence using the following rule

$$W_{t+1} = W_t - \alpha \nabla_W \mathcal{R}(\boldsymbol{y}) \tag{4.9}$$

The main question here is how does one compute $\nabla_W \mathcal{R}(\boldsymbol{y})$? This is where the *backpropagation* algorithm comes in; it is a fundamental concept in training a NN. Basically, the backrpopagation uses the chain rule of differentiation to compute the required gradient. In the example above, one can compute the gradient as follows:

$$\nabla_W \mathcal{R}(\boldsymbol{y}) = \frac{\partial \mathcal{R}(\boldsymbol{y})}{\partial W} = \frac{\partial \mathcal{R}(\boldsymbol{y})}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial W} = \frac{\partial \mathcal{R}(\boldsymbol{y})}{\partial \boldsymbol{y}} \boldsymbol{\psi} \tag{4.10}$$

where the last equality follows from $\partial_W \boldsymbol{y} = \partial_W W \boldsymbol{\psi} = \boldsymbol{\psi}$. The procedure is summarised in Figure 4.4b. For the MSE objective function $\mathcal{R} = ||\boldsymbol{y} - \boldsymbol{y}_{train}||^2_{L_2}$ the derivative is $(\boldsymbol{y} - \boldsymbol{y}_{train})$, so the equation above becomes:

$$\nabla_W \mathcal{R}(\boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{y}_{train}) \boldsymbol{\psi} \tag{4.11}$$

Knowing this derivative allows for the updating of the weights according to Equation 4.9 and perform this process until convergence.

The procedure described above is how all modern NN are trained. However, the same procedure allows to optimise the inputs $\boldsymbol{\psi}$, since one can compute the derivative of the objective function over its inputs in the same way we did for its weights:

$$\nabla_{\boldsymbol{\psi}} \mathcal{R}(\boldsymbol{y}) = \frac{\partial \mathcal{R}(\boldsymbol{y})}{\partial \boldsymbol{\psi}} = \frac{\partial \mathcal{R}(\boldsymbol{y})}{\partial \boldsymbol{y}} W \tag{4.12}$$

The obtained derivative can be used to minimise any differentiable objective function $\mathcal{R}(\boldsymbol{y})$ over $\boldsymbol{\psi}$, which is exactly the optimisation problem in Equation 4.8! So on one side, a NN can be used to predict some outputs $\boldsymbol{y}$, but on the other hand, it can be used to perform optimisation over the inputs. This is the key concept in the optimisation algorithm described in Chapter 5. The backpropagation over inputs is indicated by a blue arrow in Figure 4.4b.

In this simple example, we only have a single hidden layer of size four: the weight matrix W is just a 4-dimensional vector. In real-life, NNs have many hidden layers of different sizes, and some non-linear function of outputs of one hidden layer is the input for the next one. We note that it is necessary to apply some non-linear functions between hidden layers because without them any consecutive combination

of linear operations can be expressed as one linear operation. Without such functions even the deepest network would act as a simple NN with one hidden layer.

The same principle of backpropagation (chain rule) applies to deep neural networks with many layers. For example, for a NN with three hidden layers, the derivative of objective function with respect to the weights of the first layer is:

$$\nabla_{W_1}\mathcal{R}(\boldsymbol{y}) = \frac{\partial\mathcal{R}(\boldsymbol{y})}{\partial\boldsymbol{y}}\frac{\partial\boldsymbol{y}}{\partial i_3}\frac{\partial i_3}{\partial i_2}\frac{\partial i_2}{\partial W_1}, \tag{4.13}$$

where

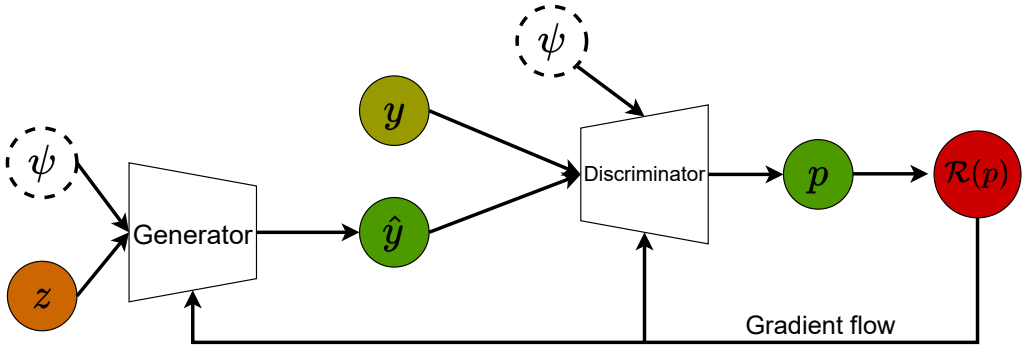
$$\boldsymbol{y} = \gamma(W_3 i_3) \tag{4.14}$$
$$i_3 = \gamma(W_2 i_2) \tag{4.15}$$
$$i_2 = \gamma(W_1 \boldsymbol{\psi}) \tag{4.16}$$

and γ is some non-linear function. Modern Deep Learning frameworks allow to compute such derivatives automatically by specifying only the inputs and the architecture of the NN. Thus, the procedure of training a neural network or using it for optimisation over its inputs is straightforward. However, since the state-of-the-art neural networks have millions of parameters w they require extremely extensive computation resources.

To conclude, *by construction* any DL model allows to compute gradients over model parameters or inputs. Those gradients can successively be used for the NN training or optimisation of some objective function over inputs. However, the only question left is how to make the gradients $\nabla_{\boldsymbol{\psi}}\mathcal{R}(\boldsymbol{y})$ meaningful, since an untrained NN would produce random outputs $\boldsymbol{y}$ and thus random gradients estimation.

4.3.2 Generative neural networks

In the modern DL era, generative networks were first introduced in Ref. [80] and then quickly gained popularity. There are various types of generative networks currently available, but for all of them the main task is to sample from some learned distribution. For example, the generative network can learn the phase space distribution of particles and then sample from it, thus, speeding up a simulation. One



(a)

Figure 4.5: Diagram of a Generative Adversarial Network. Dashed nodes represent additional inputs to a conditional GAN.

particular type of the generative model is called a Generative Adversarial Network (GAN) and is schematically presented in Figure 4.5.

A GAN consists of two NN called a generator (G) and a discriminator (D). The task of the generator is to produce outputs $\hat{\boldsymbol{y}}$ that are as close as possible to the real examples $\boldsymbol{y}$ from some distribution $p(\boldsymbol{y})$. The task of the discriminator is to distinguish between real and generated outputs. By default, a generator network takes a random noise $\boldsymbol{z}$, sampled from a pre-defined distribution $q(\boldsymbol{z})$, as an input and returns the generated output $\hat{\boldsymbol{y}}$. The discriminator takes as input either a real sample or a generated one and outputs the probability p of the sample being a real one. Afterwards, a loss function $\mathcal{R}(p)$ is computed, and the weights of D and G are updated using backpropagation and gradient descent, as described in Section 4.3.1.

The loss function for the GAN is interesting in the way that the generator tries to minimise it and the discriminator tries to maximise it simultaneously:

$$\min_G \max_D \mathcal{R}(D, G) = \mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y})}[\log(D(\boldsymbol{y}))] + \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{4.17}$$

The discriminator should output $p = D(\boldsymbol{y})$ equal to one for real samples and to zero for fake ones. Thus, it tries to maximise the loss above and to make it zero. The generator, on the contrary, tries to minimise the loss, in particular, it tries to force $D(G(\boldsymbol{z}))$ to be one, thus making the whole term negative. Eventually, in the ideal scenario, the equilibrium state is reached when neither the generator nor the discriminator can improve any further. In reality, the GAN training is a very subtle procedure, and the most common problems that arise during GAN training are mode-collapse, vanishing of gradients, and failure of convergence of the training

procedure.

The mode-collapse happens when the generator finds some perfect output $\hat{\boldsymbol{y}}$ that fools the discriminator ($D(\hat{\boldsymbol{y}}) = 1$). In that case, there is no reason for the generator to produce any output other than $\hat{\boldsymbol{y}}$, meaning that it will not sample diverse examples.

The vanishing gradients problem happens when the discriminator is much better than the generator. Then $D(\hat{\boldsymbol{y}}) = 0$ and the last term in Equation 4.17 is zero, resulting in zero gradients for the generator. In such a scenario, the generator stops improving, even though the quality of its samples is poor.

Finally, because the training objective is a non-convex optimisation problem, the convergence of the procedure is not guaranteed. Depending on the hyper-parameters of D and G, as well as the parameters of the optimisation algorithm, the whole training procedure might diverge.

Many variations of the original approach have been created to make it more robust while trying to solve some of the above problems [81, 82, 83].

After the training is finished, the generator's weights are fixed, and one can start generating new samples $\hat{\boldsymbol{y}}$ from it. Ideally, the samples from the generator would be close to the real ones: $G(\boldsymbol{z}) \approx p(\boldsymbol{y})$.

One can also modify the training procedure to learn the *conditional* distribution $p(\boldsymbol{y}|\boldsymbol{\psi})$ by providing $\boldsymbol{\psi}$ as an additional input to the generator and the discriminator, as shown by the dashed nodes in Figure 4.5. The training procedure remains the same, and the resulted generator can now sample from $p(\boldsymbol{y}|\boldsymbol{\psi})$.

The GAN provides the last piece on the way to understanding the idea behind the new optimisation algorithm. Since it can sample from a conditional distribution, it might be capable of mimicking the distribution $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$, induced by a simulator, as discussed in Section 4.2.4. Thus, the conditional samples $\hat{\boldsymbol{y}}$ from the GAN should resemble the laws of physics, inbuilt in the simulator. Now, by removing the discriminator and defining some physics-motivated objective function $\mathcal{R}(\hat{\boldsymbol{y}})$ based on the generator's outputs, it is possible to obtain gradients $\nabla_{\boldsymbol{\psi}}\mathcal{R}(\hat{\boldsymbol{y}})$ with respect to the generator's inputs $\boldsymbol{\psi}$, as discussed in Section 4.3.1. By knowing the gradient, one can apply gradient descent to optimise the objective function. This is the key idea in the optimisation algorithm that will be described in the next chapter.

Chapter 5

Local Generative Surrogate Optimisation

As discussed in the previous chapter, a simulator is a compelling method for modelling complex real-world systems. It is hard to imagine High Energy Physics domain without using dedicated simulators, such as GEANT [71] or Pythia [84].

In the SHiP experiment, one needs to find optimal parameters that produce optimal data. To formulate the optimisation problem, we introduce an objective function $\mathcal{R}$, a simulator F, some optimised parameters $\boldsymbol{\psi}$, together with the inputs $\boldsymbol{x}$ and outputs $\boldsymbol{y}$. The objective function $\mathcal{R}$ is a physics-dictated objective function, defined by the user, that one wants to minimise. For example, in the SHiP experiment, it may be the number of muon hits in the detector plane. The parameters $\boldsymbol{\psi} \in \mathbb{R}^d$ correspond to the physical properties of the optimised object, such as the magnet length or width or the strength of a magnetic field. The inputs $\boldsymbol{x}$ correspond to the input data provided to the simulator, such as muon momentum and production coordinates. The outputs $\boldsymbol{y}$ correspond to the observable quantities, such as the detector hit coordinates of the muons. Finally, F corresponds to the GEANT4 simulator, that given some inputs $\boldsymbol{x}$ and some characteristics of the experiment setup $\boldsymbol{\psi}$ will output observable variables $\boldsymbol{y}$: $\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{\psi})$.

As was discussed in Chapter 4, given the stochastic nature of the problem in HEP, one may treat $\boldsymbol{y}$ as a random variable as well, such that $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\psi})$, and the inputs $\boldsymbol{x}$ are sampled from the distribution $q(\boldsymbol{x})$, which resembles the known distribution over the inputs. For instance, the distribution of the muon momenta and production coordinates. Thus, the objective becomes a stochastic function and is expressed as an expectation $\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})}[\mathcal{R}(\boldsymbol{y})]$. The minimisation problem is thus the same as in

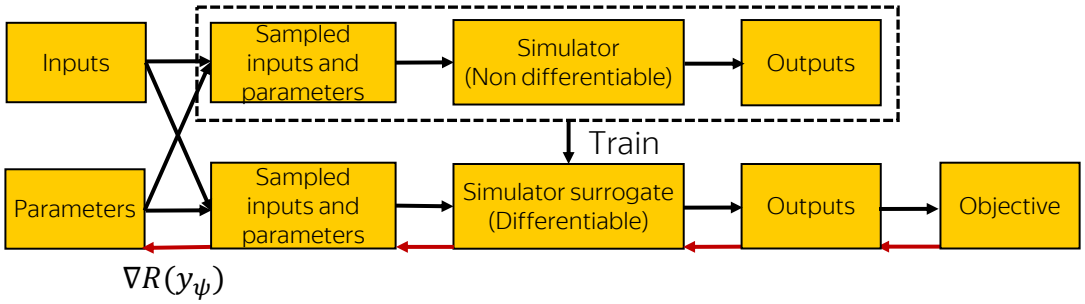


Figure 5.1: Overview of the simulation process and surrogate training.

Equation 4.8:

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})}[\mathcal{R}(\boldsymbol{y})] \tag{5.1}$$

Usually in the HEP domain one needs to obtain large population statistics in order to estimate the optimised parameters. Given that the computational cost of running a simulator such as GEANT4 is very high, it is desirable to devise an optimisation method that would minimise the usage of CPU time during optimisation. Thus the algorithm should infer as much information from the simulator as possible during one run of the simulator and maximise the re-usage of the datapoints obtained from previous runs.

In case the gradient of the objective function is available, one can use various methods for optimisation, such as Minuit [75]. However, in many scientific domains, including HEP, non-differentiable functions occur frequently, as in the problem formulation above, where a Monte Carlo simulator is used to produce data samples from an intractable probability distribution. In such cases, the whole spectrum of the algorithms, such as genetic algorithms, Bayesian optimisation, numerical differentiation or REINFORCE-based estimators, may be employed to estimate the gradients of non-differentiable functions (see Chapter 4 for details on the algorithms).

To utilise the strengths of gradient-based optimisation while avoiding the drawbacks of the other algorithms (see Chapter 4), our approach utilises deep generative models (see Chapter 4) as differentiable surrogate models (sometimes referred to as the surrogate) to approximate non-differentiable simulators, as described in Figure 5.1. We have shown that the surrogate model is able to approximate the stochastic behaviour of the GEANT4 simulator, thus, enabling the gradient-based optimisation of the objective function. We have first developed Generative Surrogate Optimisation, described below, and then modified it to account for some problems, frequently arising

in a high-dimensional optimisation scenario. The modified algorithm is named Local Generative Surrogate Optimisation or, in short, L-GSO.

5.1 Generative Surrogate Optimisation (GSO)

To further set up the mathematical formulation of the problem, we note one more important property: the simulator F is a black-box function for the user. This means that the simulator can only draw samples from the distribution $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$, but the exact analytical form of this distribution is inaccessible to the user. This is precisely the case of the GEANT4 simulator, for which the user can not infer any information about the analytical form of the distribution without an additional complicated routine tailored to a specific optimisation task [85].

Let a simulator be characterised by the parameters $\boldsymbol{\psi}$, consume stochastic inputs $\boldsymbol{x} \sim q(\boldsymbol{x})$ and produce outputs (observations) $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$. The objective function $\mathcal{R}(\boldsymbol{y})$ is defined by the physicist and encodes the task-specific relationship between the outputs $\boldsymbol{y}$ and the penalty values. The function should be minimised over the parameters $\boldsymbol{\psi}$:

$$\begin{aligned} \boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(\boldsymbol{y})] &= \arg\min_{\boldsymbol{\psi}} \int \mathcal{R}(\boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi}) q(\boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y} \\ &\approx \arg\min_{\boldsymbol{\psi}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{R}(F(\boldsymbol{x}_i;\boldsymbol{\psi})) \end{aligned} \tag{5.2}$$

where $\boldsymbol{y}_i = F(\boldsymbol{x}_i;\boldsymbol{\psi}) \sim p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi}), \quad x_i \sim q(\boldsymbol{x})$, and a Monte Carlo approximation of the expected value of the objective function is computed using samples drawn from the simulator.

Given such an objective function, one would like to compute its gradient to perform optimisation:

$$\nabla_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(\boldsymbol{y})] \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\psi}} \mathcal{R}(F(\boldsymbol{x}_i;\boldsymbol{\psi})) \; . \tag{5.3}$$

However, since the simulator F is non-differentiable, it is not possible to compute the gradient in Equation 5.3

The GSO algorithm approximates F with a generative neural network (a surrogate)

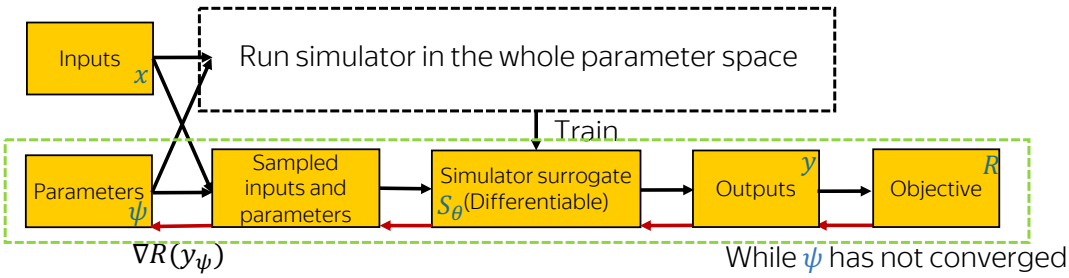


Figure 5.2: Block diagram of the Generative Surrogate Optimisation algorithm.

S_θ, where θ corresponds to the weights of the neural network. After training, the surrogate outputs $\bar{\boldsymbol{y}} = S_\theta(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\psi})$ approximate the simulator's outputs $F(\boldsymbol{x}; \boldsymbol{\psi})$, where $\boldsymbol{z} \sim p(\boldsymbol{z})$ is the noise variable accounting for the stochastic nature of the simulator (see Chapter 4 for more details). Since the samples $\bar{\boldsymbol{y}}$ are differentiable with respect to $\boldsymbol{\psi}$, the computation of Equation 5.3 becomes feasible by swapping the simulator F with the surrogate S:

$$\nabla_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(\boldsymbol{y})] \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\psi}} \mathcal{R}(S_\theta(\boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\psi})) \,. \tag{5.4}$$

The optimisation procedure is described in Algorithm 4 and is schematically depicted in Figure 5.2. A set of training data for the surrogate is created by sampling a set of values for the parameters $\boldsymbol{\psi}$ and the inputs $\boldsymbol{x}$, and then computing the simulator's output $F(\boldsymbol{x}, \boldsymbol{\psi})$ (step 3). Then, the surrogate is trained using the standard neural network training procedure, as described in Chapter 4 (step 4). Subsequently, the gradient of the objective function is estimated using the trained surrogate model by iteratively sampling the inputs $\boldsymbol{x}$ and the surrogate outputs $\bar{\boldsymbol{y}}$ (step 7) and then computing the gradient estimation (step 8). Finally, $\boldsymbol{\psi}$ is updated with the stochastic gradient descent (SGD) procedure, described in Chapter 4.

Here we used the crucial property of any deep neural network model: the ability to produce differentiable samples $\bar{\boldsymbol{y}}$. It is exactly this property that allows us to apply the chain rule of differentiation and obtain the estimation of the gradient in step 8 of the algorithm:

$$\nabla_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{y}}[\mathcal{R}(\boldsymbol{y})] \sim \sum \frac{\partial \mathcal{R}(\boldsymbol{y})}{\partial \boldsymbol{y}_i} \frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{\psi}} = \sum \frac{\partial \mathcal{R}(\boldsymbol{y})}{\partial \boldsymbol{y}_i} \frac{\partial S_\theta}{\partial \boldsymbol{\psi}} \tag{5.5}$$

Before applying the algorithm to the full-scale SHiP optimisation, several toy experiments have been performed to assess the desired properties of the algorithm.

Algorithm 4 Generative surrogate optimisation (GSO)

Require: number N of $\boldsymbol{\psi}$, number M of $\boldsymbol{x}$ for surrogate training, number K of $\boldsymbol{x}$ for $\boldsymbol{\psi}$ optimisation step

1: Choose parameters $\{\boldsymbol{\psi}'_i\}_{i=1}^N$
2: For each $\boldsymbol{\psi}'_i$, sample inputs $\{\boldsymbol{x}^i_j\}_{j=1}^M \sim q(\boldsymbol{x})$
3: Sample $M \times N$ training examples from simulator $\boldsymbol{y}_{ij} = F(\boldsymbol{x}^i_j; \boldsymbol{\psi}'_i)$
4: Train generative surrogate model $S_\theta(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\psi}')$, where $\boldsymbol{z} \sim \mathcal{N}(0, 1)$
5: Fix weights of the surrogate model θ
6: **while** $\boldsymbol{\psi}$ has not converged **do**
7: Sample $\bar{\boldsymbol{y}}_k = S_\theta(\boldsymbol{z}_k, \boldsymbol{x}_k; \boldsymbol{\psi}),\ \boldsymbol{z}_k \sim \mathcal{N}(0, 1),$ $\boldsymbol{x}_k \sim q(\boldsymbol{x}),\ k = 1, \ldots, K$
8: $\nabla_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(\bar{\boldsymbol{y}})] \leftarrow \frac{1}{K} \sum_{k=1}^{K} \frac{\partial \mathcal{R}}{\partial \bar{\boldsymbol{y}}_k} \frac{\partial S_\theta(\boldsymbol{z}_k, \boldsymbol{x}_k; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$
9: $\boldsymbol{\psi} \leftarrow \mathrm{SGD}(\boldsymbol{\psi}, \nabla_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(\bar{\boldsymbol{y}})])$
10: **end while**

5.1.1 Experiments and results

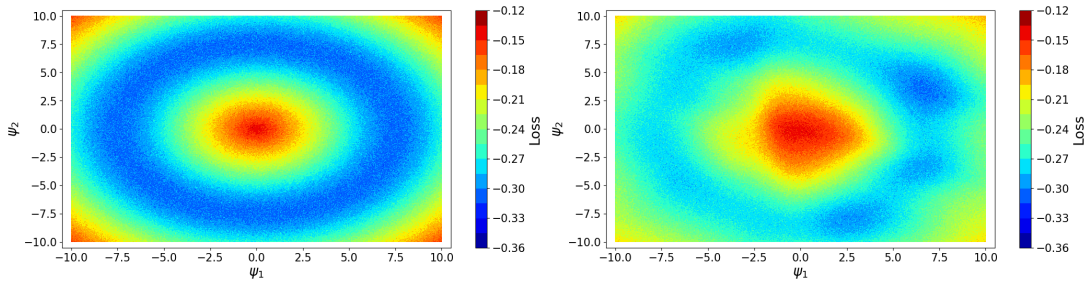
To be in complete control of the environment and demonstrate the capabilities of the new algorithm, we have solved a simple mathematical optimisation problem and a straightforward physics optimisation problem, where the solutions are known in both cases.

Toy mathematical problem

We start with a simple mathematical model that is defined as follows:

$$\begin{gathered} \boldsymbol{x} \sim \mathrm{U}[L, R]: \quad L = -10,\ R = 10,\ \boldsymbol{x} \in \mathbb{R}^1 \\ X \sim \mathcal{N}(\boldsymbol{x}, 1) \\ \psi = \sqrt{\boldsymbol{\psi}_1^2 + \boldsymbol{\psi}_2^2} \\ \boldsymbol{y} \sim \mathcal{N}\left(\psi + X, 0.1 + \frac{1}{2}|X|\right) \end{gathered} \tag{5.6}$$

The variable $\boldsymbol{x}$ is first sampled from the uniform distribution and then transformed via a normal distribution to arrive at X. With this sampling procedure, we have tried to mimic a stochastic transformation over the inputs $\boldsymbol{x}$ that might actually happen in the GEANT4 simulator. Then, we sample the observable outputs $\boldsymbol{y}$ that depend both on X and $\boldsymbol{\psi}$, resembling the dependence of the outputs on the parameters of the simulator (for example, the shape of the magnet). As discussed above, the



(a) (b)

Figure 5.3: Objective function value as a function of $\boldsymbol{\psi}$ vector components. Objective function value obtained from (a) the mathematical model, (b) the surrogate S_θ.

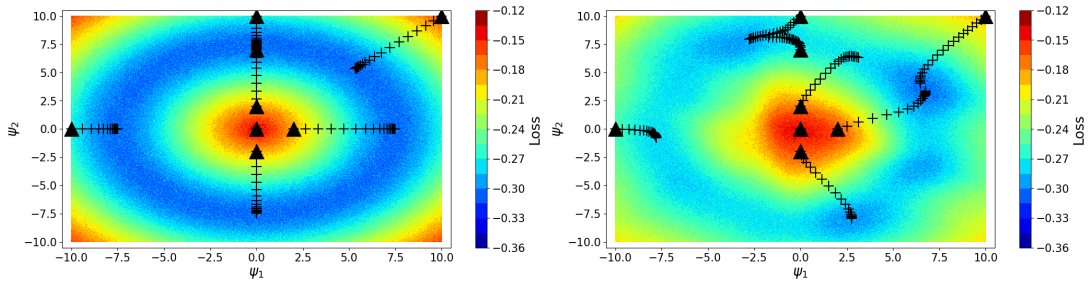
algorithm's goal is to learn the surrogate model $S_\theta(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\psi})$, such that the samples from it are close to $\boldsymbol{y}$.

On top of the mathematical model, we impose an objective function that we want to minimise. The choice of the objective function for this particular example is dictated by the requirements of being differentiable with respect to the outputs $\boldsymbol{y}$ and being sufficiently varying in the region of parameters $\boldsymbol{\psi}$. Thus, the function is defined as follows:

$$\mathcal{R}(y; L, R) = -\frac{1}{N} \sum_{i=0}^{N} \sigma(\boldsymbol{y}_i - L) - \sigma(\boldsymbol{y}_i - R) \tag{5.7}$$

where $\boldsymbol{y}_i$ is the sample from either the mathematical model defined above or from the trained surrogate model S_θ, and L, R denote the left and the right boundaries of the $\boldsymbol{\psi}$ space. The function σ denotes the sigmoid function: $\sigma(x) = 1/(1 + e^{-x})$.

To evaluate the proposed algorithm, we would like to compare the landscapes of the objective function (in the $\boldsymbol{\psi}$ space), as this would define the quality of the gradient estimation and, in turn, the convergence of the optimisation. The comparison of the loss surfaces obtained from the mathematical model and the surrogate is presented in Figure 5.3. The left figure, obtained from the model, has many degenerate minima due to the rotational symmetry of the problem. The goal is to validate that the optimisation algorithm can find any of them. Examining the right figure, where the loss is calculated using the surrogate model, the deviation from the actual loss landscape is clearly seen. At first glance, one might expect the optimisation to diverge due to the inaccurate estimation of the objective function by the surrogate model. To study how the reconstruction quality affects actual optimisation, we have run the optimisation procedure from multiple initial points, using the mathematical model



(a) (b)

Figure 5.4: Optimisation trajectories from different starting points overlaid on the loss surface. The triangles depict the different starting points. Trajectories obtained from (a) the mathematical model, (b) the surrogate model.

and the surrogate. We note that using the mathematical model for gradient-based optimisation is only possible in toy problems, where the gradient information is available. In our implementation, we have used the Pyro package [86] to compute the information needed for the optimisation. The result of such optimisation is presented in Figure 5.4.

As observed in Figure 5.4a, all optimisation paths lie along the radius of the loss landscape. This happens because the gradient information is extracted from the mathematical model. In such a case, the gradient descent will be performed in the most optimal way, and the desired minimum achieved in the smallest number of model calls. When using the gradient estimation obtained from the surrogate model, the optimisation also works and, moreover, attains one of the possible minima, regardless of the starting point. However, because the surrogate does not model the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$ perfectly, the optimisation trajectories follow the surrogate loss landscape and do not any longer lie along the radius, as seen in Figure 5.4b. This observation raises questions regarding the quality of the surrogate model as a function of the number of training samples.

Different instances of the surrogate model have been trained on 10, 100, 1000 and 10000 training points to address this concern. The values of $\boldsymbol{\psi}$ were randomly selected on an integer grid, and a value of $\boldsymbol{x}$ sampled from the uniform distribution was assigned to each $\boldsymbol{\psi}$. The procedure was repeated until the desired number of training samples was reached. The distributions of the samples for different dataset sizes are presented in Figure 5.5. The grid step size was set to one, meaning 20 points were required just to cover one dimension of $\boldsymbol{\psi}$. To fill the whole grid with at least one sample, 20^2 points were required.

The contours of the objective function obtained from the surrogate, corresponding

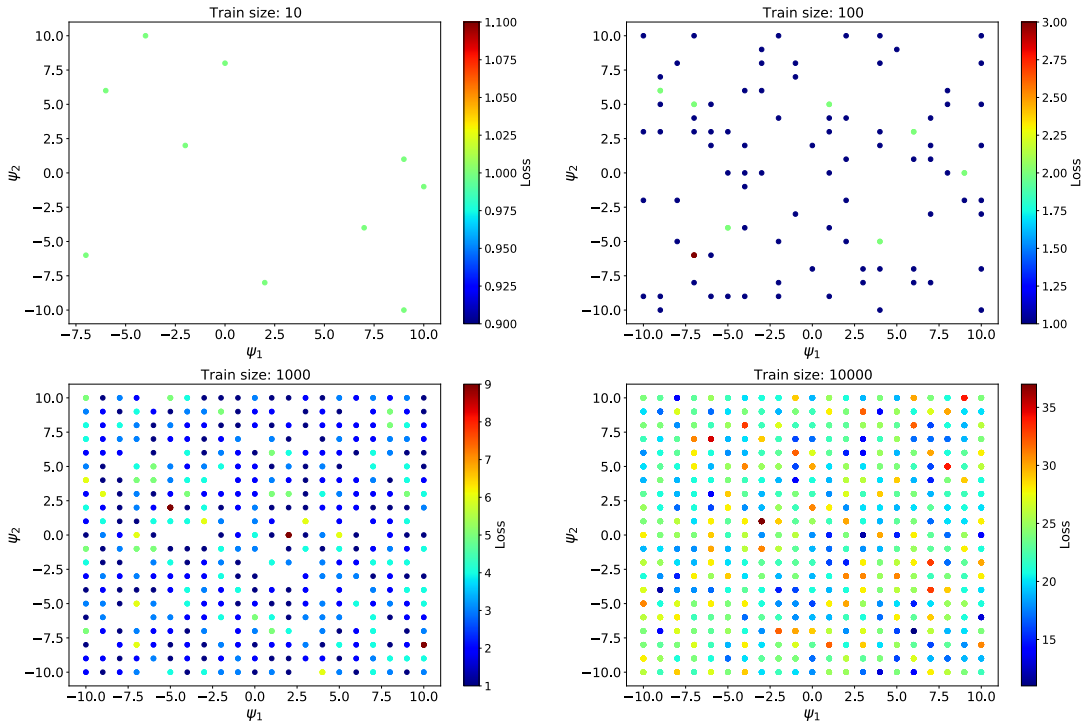


Figure 5.5: Distribution of $\boldsymbol{\psi}$ samples for the training set of size 10, 100, 1000 and 10000 points. Colour represents the amount of points with a particular $\boldsymbol{\psi}$ value in the training set.

to different dataset sizes, are presented in Figure 5.6. Examining the results and comparing them to Figure 5.3a one notices, that in the cases of 10 and 100 training samples, the loss surface does not resemble the actual landscape. Even the objective function surface obtained with 1000 training points seems far away from the true landscape.

The plots of the samples density in $\boldsymbol{\psi}$ space in Figure 5.5 suggest that, even though the $\boldsymbol{\psi}$ space is covered uniformly in the example of 1000 training points, this is still not enough to learn a good surrogate model. Only starting from the training set of size 10000 the loss surface looks like the one in Figure 5.3a. It is worth noting that the above discrepancies already arise in a simple two-dimensional optimisation problem. We note, that there are no theoretical guarantees on the optimal number of points needed to achieve a good agreement between the loss landscapes obtained from the surrogate and the model. However, as discussed later in this section, there is an exponential relationship between the dimensionality of the space and the number of points needed to train a surrogate model. Thus, we can be sure that the discrepancy observed in this toy example will be even more pronounced in high-dimensional problems.

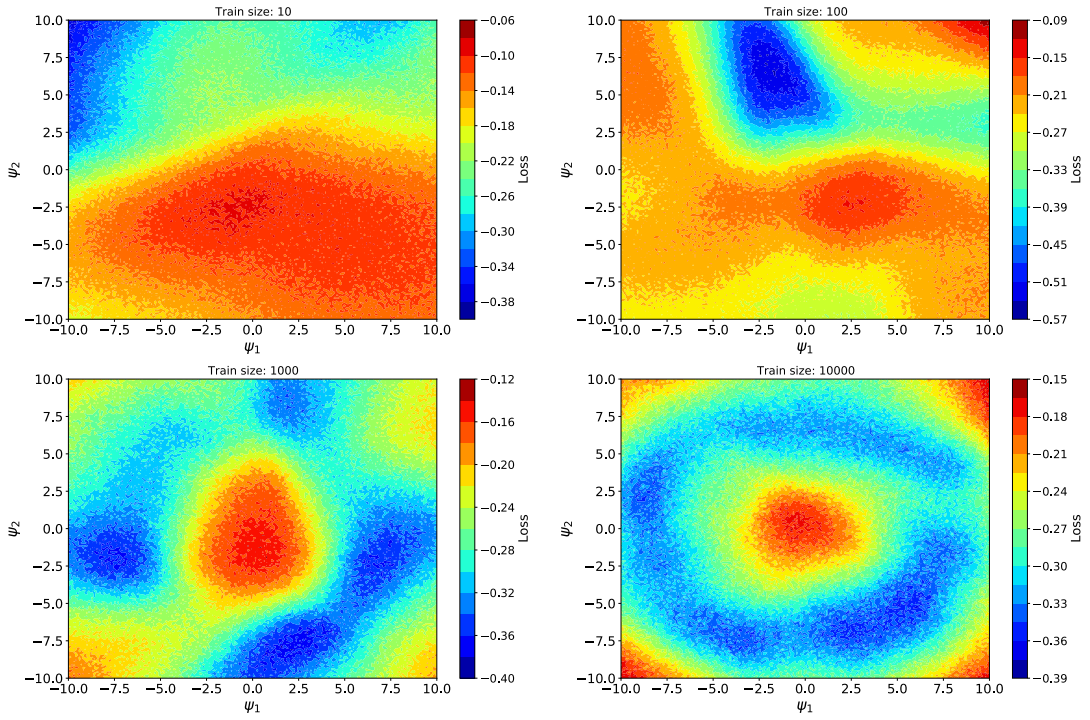


Figure 5.6: Contour levels of the objective function obtained from the surrogate, trained with different dataset sizes as a function of ψ.

The conclusion is that the parameter space coverage and the number of samples per point are essential for the surrogate to be correctly trained. Moreover, in high-dimensional spaces the number of points needed to train the surrogate model accurately grows exponentially with the dimensionality of the space. However, by decreasing the search size (not dimensionality!), the amount of samples needed is significantly reduced. This is a crucial observation, and we will use it to devise a local surrogate model in Section 5.2.

Toy physics problem

To make sure the algorithm works not only in the previous simple artificial setting, but can also be applied to a more realistic physics problem, a simple one-dimensional optimisation was performed. We simulated the propagation of muons through the magnetised vacuum volume of a rectangular shape using GEANT4, and recorded muons coordinates on the sensitive plane located after the magnetised area. The muons were shot using a particle gun with a uniform distribution over the muon momentum range, and polar and azimuthal angles. The muon charge was selected randomly for each event. Thus, the momentum P, polar angle ϕ and azimuthal angle θ as well as the electric charge Q of a particle constitute the random variable

$\boldsymbol{x}$, which we set as an input to our algorithm. The muon particle coordinates $\boldsymbol{y} \in \mathbb{R}^2$ on the sensitive plane are observed as an output. We keep the width and height of the magnetised volume fixed and vary only its length. Thus, the parameter $\boldsymbol{\psi} \in \mathbb{R}$ corresponds to the length of the volume.

The mathematical formulation of this problem is as follows:

$$\begin{gathered} \boldsymbol{x} = (P, \phi, \theta, Q), \boldsymbol{x} \in \mathbb{R}^4, \\ P \sim \mathrm{U}[5, 10] \text{ GeV}, \ \phi \sim \mathrm{U}[-\pi, \pi], \\ \theta \sim \mathrm{U}\left[0, \frac{5}{180}\pi\right], \ Q \in \{+1, -1\} \end{gathered} \tag{5.8}$$

Before running the optimisation algorithm, we can visually examine how well the surrogate approximates the distribution of the muon hits on the sensitive plane for different lengths of the magnet. To do that, we have generated events using GEANT4 and trained a GAN surrogate model on the obtained data. Afterwards, new events were sampled from GAN and compared to the data from GEANT4. Unsurprisingly, GAN was able to reproduce the distribution of the hits very well, as demonstrated in Figure 5.7.

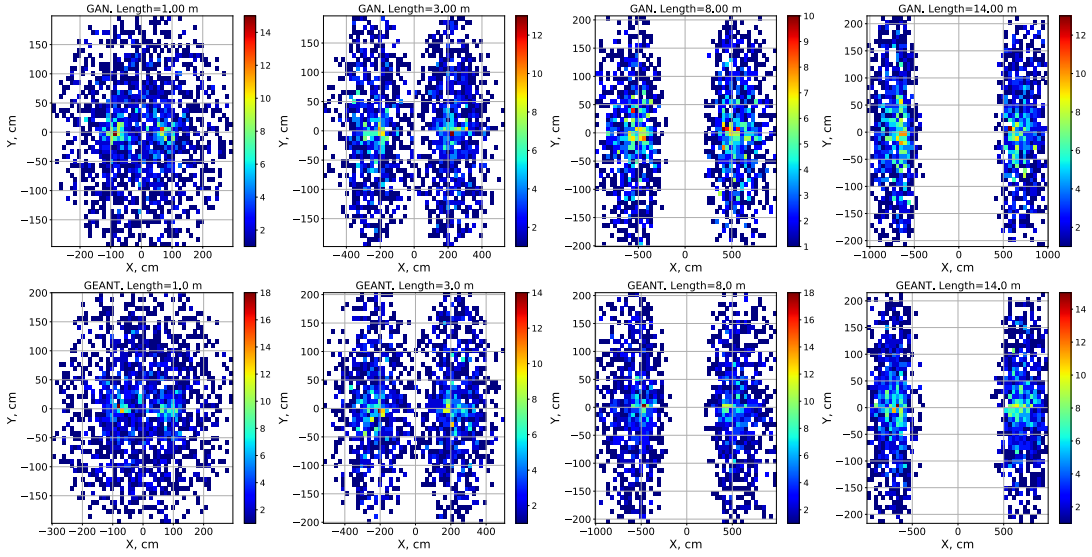


Figure 5.7: Distribution of the muons hits on the sensitive plane. The first row corresponds to the samples from GAN, the second row represents the output of the simulator. Columns indicate different magnet length: 1, 3, 8 and 14 meters correspondingly.

After making sure that the GAN could capture the behaviour of the muon distribution as a function of the magnet length, the next step was to perform an actual

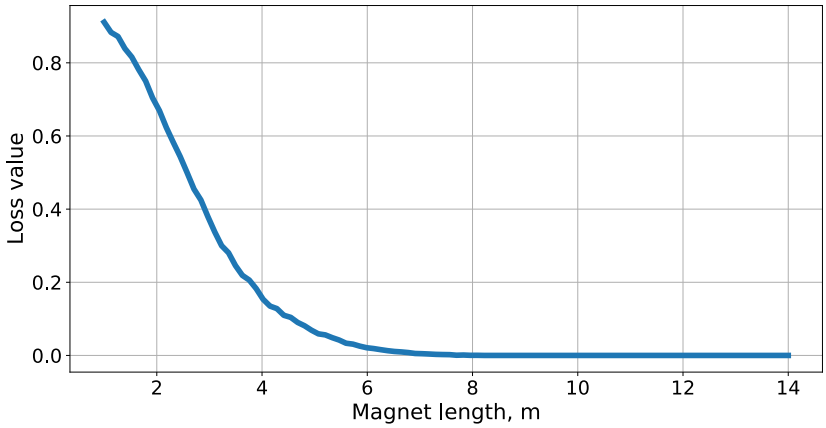


Figure 5.8: True value of the loss as a function of the magnet length, obtained from GEANT4.

optimisation.

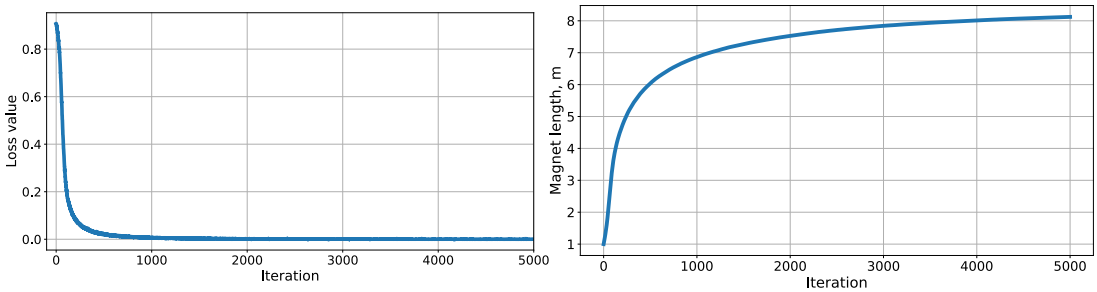
The objective function to minimise was defined as:

$$\begin{aligned}\mathcal{R}(\boldsymbol{y}; -200, 200) = -\frac{1}{N}\sum_{i=0}^{N} \sigma(\boldsymbol{y}_i^0 + 200) * \sigma(\boldsymbol{y}_i^1 + 200) \\ -\sigma(\boldsymbol{y}_i^0 - 200) * \sigma(\boldsymbol{y}_i^1 - 200)\end{aligned} \tag{5.9}$$

where the index i runs over all N muons in the sample, $\boldsymbol{y}^0$ is the x-coordinate of the hit and $\boldsymbol{y}^1$ is the y-coordinate. The value 200 denotes the half-width and the half-height of the restricted area that we want to clean up from the muons. The objective function equally penalises all muons that have hits inside the 200×200 square of the sensitive plane.

The idea behind such a simple experiment is that we know *a priori* the correct answer: the algorithm must increase the length of the magnet until there are no muons left inside the restricted area. The loss value as a function of the magnet length, obtained using the GEANT4 simulator, is presented in Figure 5.8. Examining the plot, one would expect that the optimisation should rapidly increase the magnet length to 6-7 meters.

In Figure 5.9a the loss function value during the optimisation is presented. As expected, the optimisation algorithm was able to minimise the loss function rapidly. In Figure 5.9b the magnet's length as a function of the optimisation iteration is presented. The magnet's length is rapidly increasing until it reaches seven meters and then continues to increase, but at a slower pace. This is what we expect: once the loss reaches zero, the gradients provided by the surrogate become very noisy,



(a) (b)

Figure 5.9: (a) Attained loss value as a function of the iteration step, during optimisation. (b) Magnet length as a function of the iteration step.

and the optimisation slows down. The algorithm, however, continues to increase the magnet length to reduce the muon rate to an absolute zero.

5.1.2 Summary

We have shown that the newly devised GSO algorithm can perform simple one- and two-dimensional optimisation by applying it to the toy mathematical and physics problems. This suggests that the idea of generative surrogates is worth pursuing. However, we have already seen in Figure 5.4 that the GSO can not capture the distribution $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$ perfectly even in the two-dimensional case, given the limited amount of data.

In general, training the surrogate model in the whole parameter space is infeasible. First of all, the main problem is that the surrogate model needs to be trained on a grid of parameters $\boldsymbol{\psi}$. To fill in this grid uniformly, one would require $O\left((L/\Delta)^D\right)$ samples, where L is the size of the space, Δ is the discretisation step and D is the dimensionality of $\boldsymbol{\psi}$. For instance, in the toy mathematical example above $\boldsymbol{\psi}$ varies from -10 to 10 cm, $L = 20$ cm and $\boldsymbol{\psi}$ is a two-dimensional vector. By setting $\Delta = 1$ cm the points will be sampled uniformly in the $\boldsymbol{\psi}$ space every cm. With these settings, 400 points would be sampled to train the model and, as as it was shown, even this might not be enough. If the parameter space is 10-dimensional, the number of samples required to fill the grid is $O(20^{10})$, which is computationally infeasible! This is an example of the exponential growth of the number of samples with the dimensionality D of the parameter space.

Secondly, when one samples points in a high-dimensional setting, the "curse of dimensionality" problem arises (see Chapter 4). This problem states that as the dimensionality of space increases, the amount of datapoints lying on the boundary

of the space is infinitely larger than the number of points inside the space volume. Thus, even if one would randomly sample points, instead of using the grid, it would be impossible to fill in the volume uniformly since most of the sampled points would lie on the boundary.

The above observations mean that the GSO will not work in high-dimensional spaces. Indeed, we performed some basic toy mathematical and physics experiments and observed that the GSO could not optimise even a five-dimensional physics model. Therefore, to make GSO work in a high dimensional scenario, we have developed the *Local Generative Surrogate Optimisation*: L-GSO.

5.2 Local Generative Surrogate Optimisation (L-GSO)

As concluded in the previous section, when the parameters $\boldsymbol{\psi}$ are high-dimensional, a large training set is required to train a surrogate model in the whole parameter space. In the toy mathematical example above, we have seen that if this requirement is not satisfied, the surrogate model does not estimate the gradients well, and the optimisation might diverge. However, it is computationally unfeasible to train the surrogate in the whole parameter space. To solve this problem, the GSO algorithm was modified so that the surrogate model is trained locally in a small region around the current value of the parameter $\boldsymbol{\psi}$.

In this approach, a new parameter value $\boldsymbol{\psi}'$ is sampled within the multidimensional cube (hypercube) with the current point $\boldsymbol{\psi}$ being in the centre of the hypercube: $U_\epsilon^\psi = \{\boldsymbol{\psi}' : |\boldsymbol{\psi}'_i - \boldsymbol{\psi}_i| \leq \epsilon, \forall i \in \{1, \ldots, D\}\}$, where D is the dimensionality of $\boldsymbol{\psi}$. Afterwards, the surrogate model is trained locally using only the points from the hypercube as the training data. At the next step, the gradient at the current point $\boldsymbol{\psi}$ is estimated using the trained model, and a step of the SGD is performed. A new surrogate model is then trained in the hypercube around the new point. The algorithm is graphically summarised in Figure 5.10 and in Algorithm 5.

There are several hyperparameters in the L-GSO algorithm that require tuning before the optimisation. First of all, a dedicated sampling algorithm for the hypercube U_ϵ^ψ must be defined. Since simple uniform sampling is inefficient due to the "curse of dimensionality" in high-dimensional spaces, a dedicated sampling technique called Latin Hypercubes was used [87] in step 3 of Algorithm 5.

Another crucial hyperparameter is the size ϵ of the hypercube U_ϵ^ψ. Since this

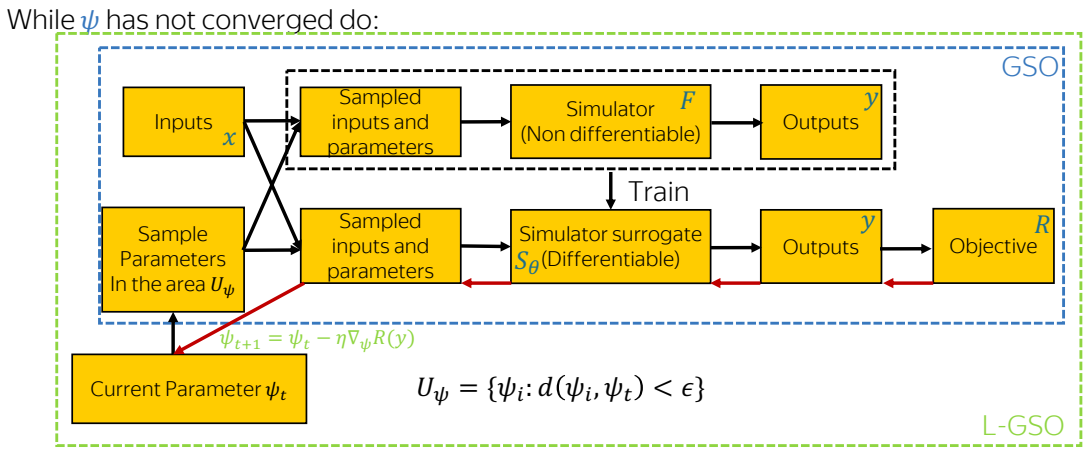


Figure 5.10: Block diagram of the Local Generative Surrogate Optimisation algorithm.

parameter controls the size of the volume in which the surrogate is trained, it should be carefully selected. If it is too large, the surrogate model might provide an inaccurate estimation of the gradient due to the insufficient density of the training points in the region, and the optimisation might diverge. If it is too small, the hypercube volume might be dominated by the noise, and the surrogate model might not be able to extract any meaningful information, resulting in bad gradient estimation and divergence of the optimisation.

Finally, the number of points $\boldsymbol{\psi}$ sampled in the hypercube U_ϵ^ψ is another essential hyperparameter. It is related to the size of the hypercube since the two parameters control the density of the samples inside the volume. Empirically, in the experiments, we have observed that approximately $O(D)$ samples in the hypercube are required to accurately estimate the gradient in the area, where D is the dimensionality of the $\boldsymbol{\psi}$ parameter space. However, if the components of the vector $\boldsymbol{\psi}$ relevant to the optimisation lie in some subspace of the dimension d, where $d < D$, L-GSO requires $O(d)$ samples for producing a reasonable gradient estimation, resulting in faster convergence of the algorithm in comparison to other methods. An example of such subspace could be a fixed radius circle in the two-dimensional space: in Cartesian coordinates a point on the circle is described by x and y coordinates. In polar coordinates however, only an angle is needed to set the position of the point. The same idea can be applied to high-dimensional spaces.

In addition to the above, the algorithm stores previously sampled points in the history bank H in step 6 of Algorithm 5. This allows reusing the previous points if they are located in the current hypercube (step 7 of the algorithm). The ability to reuse previously sampled points is an essential property of the algorithm: it results

Algorithm 5 Local Generative Surrogate Optimisation (L-GSO) procedure

Require: number N of $\boldsymbol{\psi}$, number M of $\boldsymbol{x}$ for surrogate training, number K of $\boldsymbol{x}$ for $\boldsymbol{\psi}$ optimisation step, region U_ϵ, size of the neighbourhood ϵ, Euclidean distance d

Choose initial parameter $\boldsymbol{\psi}$
while $\boldsymbol{\psi}$ has not converged **do**
  Sample $\boldsymbol{\psi}'_i$ in the region U^{ψ}_{ϵ}, $i = 1, \ldots, N$
  For each $\boldsymbol{\psi}'_i$, sample inputs $\{\boldsymbol{x}^i_j\}_{j=1}^{M} \sim q(\boldsymbol{x})$
  Sample $M \times N$ training examples from simulator $\boldsymbol{y}_{ij} = F(\boldsymbol{x}^i_j; \boldsymbol{\psi}'_i)$
  Store $\boldsymbol{y}_{ij}, \boldsymbol{x}^i_j, \boldsymbol{\psi}'_i$ in history H $i = 1, \ldots, N; j = 1, \ldots, M$
  Extract all $\boldsymbol{y}_l, \boldsymbol{x}_l, \boldsymbol{\psi}'_l$ from history H, iff $d(\boldsymbol{\psi}, \boldsymbol{\psi}'_l) < \epsilon$
  Train generative surrogate model $S_\theta(\boldsymbol{z}_l, \boldsymbol{x}_l; \boldsymbol{\psi}'_l)$, where $\boldsymbol{z}_l \sim \mathcal{N}(0, 1)$
  Fix weights of the surrogate model θ
  Sample $\bar{\boldsymbol{y}}_k = S_\theta(\boldsymbol{z}_k, \boldsymbol{x}_k; \boldsymbol{\psi})$, $\boldsymbol{z}_k \sim \mathcal{N}(0, 1)$, $\boldsymbol{x}_k \sim q(\boldsymbol{x})$, $k = 1, \ldots, K$
  $\nabla_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(\bar{\boldsymbol{y}})] \leftarrow \frac{1}{K} \sum_{k=1}^{K} \frac{\partial \mathcal{R}}{\partial \bar{\boldsymbol{y}}_k} \frac{\partial S_\theta(\boldsymbol{z}_k, \boldsymbol{x}_k; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$
  $\boldsymbol{\psi} \leftarrow \mathrm{SGD}(\psi, \nabla_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(\bar{\boldsymbol{y}})])$
end while

in a better surrogate model at almost no additional computational cost. We note that this property is crucial to reduce the number of expensive simulator runs.

An illustration of the L-GSO procedure is presented in Figure 5.11. The left plot shows the loss landscape for the "three hump problem" defined below. The goal of the optimisation was to converge to one of the red stars by starting the procedure from the black star. The black rectangle shows the area U^{ψ}_{ϵ} around the current point $\boldsymbol{\psi}$, indicated by a yellow dot. The right plot features the gradient vector fields estimations from the toy model and the GAN surrogate model. The key observation is that the fields closely match each other inside the training area U^{ψ}_{ϵ}.

5.2.1 Experiments and results

For the L-GSO algorithm an extensive study of the performance was done by comparing it with a wide selection of other methods (from Chapter 4) and estimating the algorithm's properties. In addition, the algorithm was applied to the toy five-dimensional physics optimisation problem before being applied to the full-scale magnet optimisation.

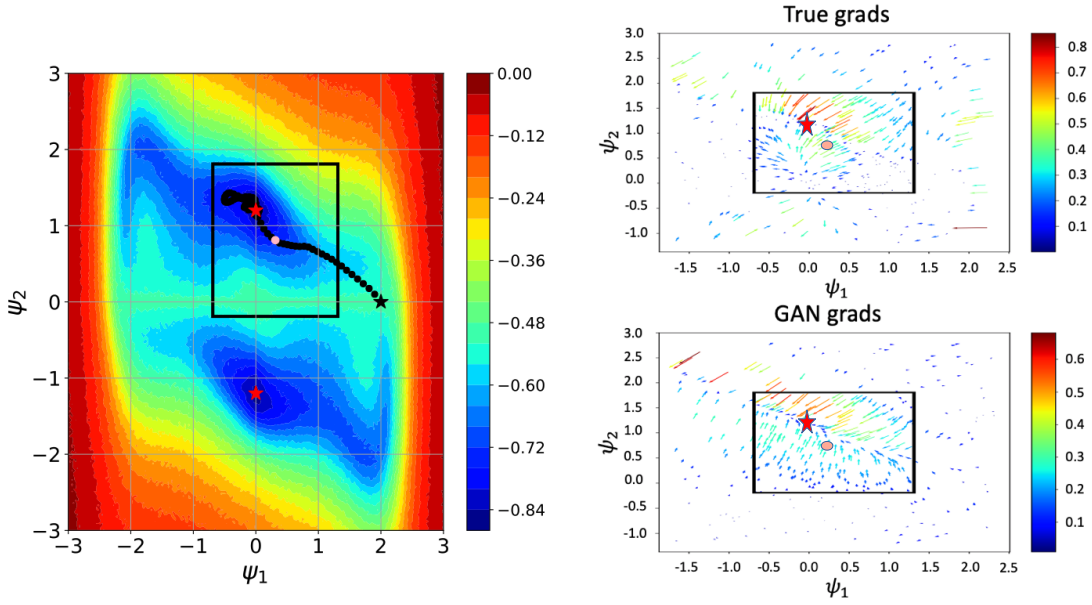


Figure 5.11: (Left) objective function landscape of the "three hump problem" and the optimisation trajectory. The black star is the initial point. The red stars represent the optimal values of the function. The yellow point corresponds to the current point $\boldsymbol{\psi}$. The black rectangle is the hypercube around the current point. [34]

5.2.2 Toy mathematical problems

We have based the evaluation and the comparison of the algorithms on the speed of convergence and the value of the attained optima. The speed of convergence was measured in terms of the number of simulator calls, as this typically is the most time-consuming part of the optimisation.

L-GSO was compared to two types of the Bayesian optimisation with Gaussian processes, with different kernel functions: the cylindrical kernels [88], denoted as BOCK, and RBF kernels [25], denoted as BO-RBF. Furthermore, the algorithm was compared to numerical differentiation, denoted as numerical optimisation, various REINFORCE-based algorithms, denoted as LAX [89] and LTS [90], and, finally, to the evolutionary algorithm [27], denoted CMS-ES. In some examples, the information about the true gradients was available and presented in the comparison.

The toy experiments were chosen to explore low- and high-dimensional optimisation problems. The *probabilistic three hump problem* [91] and the *Rosenbrock problem* [91] are two and ten-dimensional problems corresponding to low and medium dimensional problems respectively. The probabilistic three hump problem is defined as

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(y)] = \mathbb{E}[\sigma(y-10) - \sigma(y)], \text{ s.t.}$$
$$y \sim \mathcal{N}\left(y; \mu_i, 1\right), i \in \{1,2\}, \quad \mu_i \sim \mathcal{N}(x_i h(\boldsymbol{\psi}), 1), \quad x_1 \sim \mathrm{U}[-2,0], \quad x_2 \sim \mathrm{U}[2,5]$$
$$\mathrm{P}\left(i=1\right) = \frac{\psi_1}{\sqrt{\psi_1^2+\psi_2^2}} = 1 - \mathrm{P}\left(i=2\right), \quad h(\boldsymbol{\psi}) = 2\psi_1^2 - 1.05\psi_1^4 + \psi_1^6/6 + \psi_1\psi_2 + \psi_2^2 \tag{5.10}$$

and the Rosenbrock problem states that:

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(y)] = \arg\min_{\boldsymbol{\psi}} \mathbb{E}[y], \text{ s.t.}$$
$$y \sim \mathcal{N}\left(y; \sum_{i=1}^{n-1}\left[(\psi_i - \psi_{i+1})^2 + (1-\psi_i)^2\right] + x, 1\right), \tag{5.11}$$
$$x \sim \mathcal{N}(x; \mu, 1), \quad \mu \sim \mathrm{U}[-10,10]$$

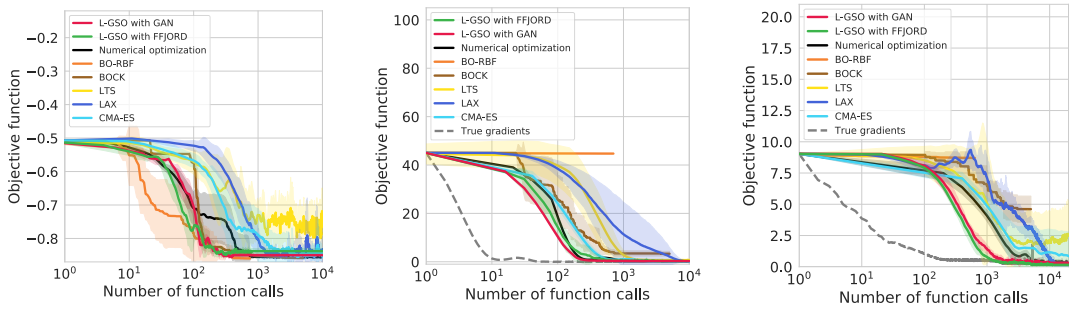
The *Submanifold Rosenbrock problem* is a 100-dimensional problem where only a 10-dimensional subspace is relevant. For such types of problems, we would expect L-GSO to outperform other methods. To complicate the relation between inputs and outputs even more, the *Nonlinear Three Hump Problem* was introduced. The initial dimension of the optimised space is 40 and the subspace is only two-dimensional. A series of non-linear transformations produce the subspace.

The Submanifold Rosenbrock problem is defined as follows:

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(y)] = \arg\min_{\boldsymbol{\psi}} \mathbb{E}[y], \text{ s.t.} \tag{5.12}$$
$$y \sim \mathcal{N}\left(y; \sum_{i=1}^{n-1}\left[(\psi'_i - \psi'_{i+1})^2 + (1-\psi'_i)^2\right] + x, 1\right)$$
$$\boldsymbol{\psi}' = A\boldsymbol{\psi}, \quad x \sim \mathcal{N}(x; \mu, 1), \quad \mu \sim \mathrm{U}[-10,10] \tag{5.13}$$

where A is a matrix projecting the 100-dimensional vector $\boldsymbol{\psi}$ to a 10-dimensional vector $\boldsymbol{\psi}'$. The Nonlinear Three Hump Problem is defined exactly like the problem above, but defines $\boldsymbol{\psi}$ as $\hat{\boldsymbol{\psi}} = B\tanh(A\boldsymbol{\psi})$, where $\boldsymbol{\psi} \in \mathbb{R}^{40}$, $A \in \mathbb{R}^{16\times 40}$, $B \in \mathbb{R}^{2\times 16}$, with matrices A and B generated as in the Submanifold Rosenbrock problem.

Finally, our method was applied to optimise the weights of the neural network:



(a) (b) (c)

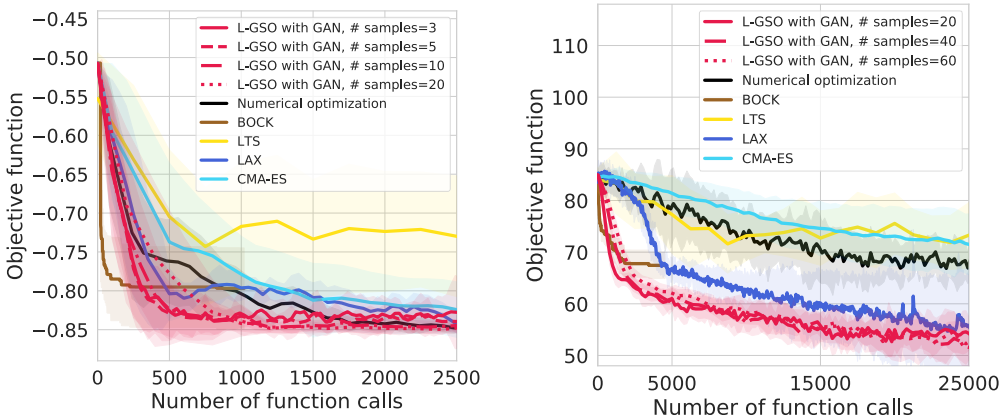
Figure 5.12: The objective function value for the toy problems obtained from L-GSO and the baseline methods. (a) Three hump problem, (b) Rosenbrock problem in 10 dimensions, initial point is $\vec{2} \in \mathbb{R}^{10}$, (c) Submanifold Rosenbrock Problem in 100 dimensions, initial point is $\vec{2} \in \mathbb{R}^{100}$. True gradients are shown in grey dashed curves when available. The shaded region corresponds to 1σ confidence intervals. [34]

instead of using a conventional SGD procedure, the L-GSO was used to find an optimal set of weights. This problem is referred to as the *Neural Network Weights Optimisation Problem* [92] and is a 91-dimensional optimisation problem.

The value of the objective function versus the number of the simulator calls for the first three toy problems is shown in Figure 5.12. In all problems, except the first one, L-GSO performs similar or better than other methods. We note, that BO converges faster than other methods in the first problem due to the low dimensionality of the parameter space $\boldsymbol{\psi}$, but struggles to attain optima in the second and especially the third problems. This observation demonstrates the drawbacks of applying BO to high-dimensional problems, discussed in Chapter 4. Notably, even in low dimensional problems, L-GSO performs on par with numerical differentiation.

The actual benefits of the L-GSO algorithm are noticeable in high-dimensional problems. The loss value as the number of the simulator calls for the Submanifold Rosenbrock, Nonlinear Submanifold Hump Problem and Neural Network Weights Optimisation problems is shown in Figure 5.12c, Figure 5.13a and Figure 5.12b respectively. In all three problems L-GSO outperforms all baseline algorithms by a large margin. We note that no prior knowledge about the subspace structure was used by L-GSO. Figure 5.13 further shows the challenging task of the BO application for high-dimensional setting, as BO struggles to converge.

Finally, any optimisation algorithm should be unbiased and have as small variance as possible. When a theoretical derivation is possible, those properties should be proved and some quantitative guarantees obtained. Since we use deep learning



(a) (b)

Figure 5.13: The objective function value as a function of the accumulated number of simulator calls for (a) Nonlinear Submanifold Three Hump problem, $\boldsymbol{\psi} \in \mathbb{R}^{40}$, (b) Neural Network Weights Optimisation problem , $\boldsymbol{\psi} \in \mathbb{R}^{91}$. The shaded region corresponds to 1σ confidence intervals. [34]

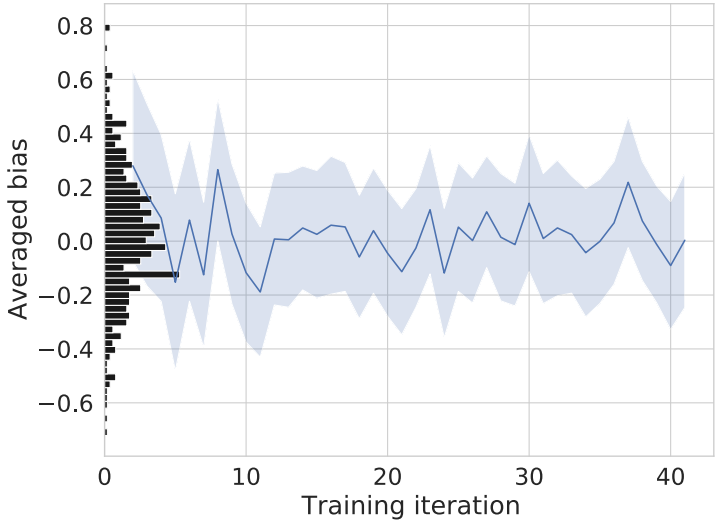


Figure 5.14: The bias (solid line) and one standard deviation (shaded region) of the GAN based L-GSO gradient averaged over all $\boldsymbol{\psi}$ dimensions in the 10D Rosenbrock problem versus training step. The grey histogram shows the empirical bias distribution over all training iterations. [34]

models as our surrogate, it is hard to obtain any mathematical guarantees on the convergence. This is only possible for a very restricted class of deep learning models, where the exact mathematical formulation of the model and the backpropagation can be obtained. However, one would like to have some convergence guarantees for the optimisation procedure, at least empirically. Thus, the bias and variance of L-GSO were empirically estimated on the example of the ten-dimensional Rosenbrock problem. The values of bias and variance at each iteration of the optimisation can be seen in Figure 5.14. The bias is computed as the difference between the true gradient and the average gradient estimation obtained from the surrogate. The averaging is performed over different sets of training data and weights initialisation of the surrogate neural network. Mathematically, the bias at iteration t of the optimisation is defined as

$$\text{Bias}_t = \nabla_{\boldsymbol{\psi}|\boldsymbol{\psi}_t} \mathcal{R}(\boldsymbol{y}_{\boldsymbol{\psi}}) - \nabla_{\boldsymbol{\psi}|\boldsymbol{\psi}_t} \mathbb{E}[\mathcal{R}(\bar{\boldsymbol{y}}_{\boldsymbol{\psi}})] \tag{5.14}$$

where $\bar{\boldsymbol{y}}_{\boldsymbol{\psi}}$ denotes samples obtained from the surrogate model and the expectation is taken over training data(samples of $\{\psi'\}$) and surrogate initialisation. The bias stays close to zero throughout the entire optimisation procedure, as seen in Figure 5.14.

5.2.3 Toy physics problem and results

The last check is to apply the L-GSO algorithm to optimise the toy magnet design, but now with a realistic setup of the simulator and objective function. The optimisation is applied to one section of a SHiP muon shield-frustum, characterised by five parameters: height and width at the beginning and the end plus the frustum length. The sensitive plane that resembles the tracking station is located after the magnet. The event display of the setup is presented in Figure 5.15. The muons were shot using the particle gun, with the same parameters, as in Equation 5.8. The main task was the same: find an optimal shape of the magnet, which sweeps muons away from the restricted area. Thus the objective function was defined as:

$$\mathcal{R}(\mathbf{y}; \boldsymbol{\alpha}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{Q_i=-1} \sqrt{(\alpha_1 - (\boldsymbol{y}_i + \alpha_2))/\alpha_1} + \mathbb{1}_{Q_i=1} \sqrt{(\alpha_1 + (\boldsymbol{y}_i - \alpha_2))/\alpha_1} \right) ,$$

where $\alpha_{1,2}$ defines the width of the restricted area, Q is the electric charge of a muon, $\boldsymbol{y}$ is the x-coordinate of a hit in the sensitive plane, and N is the total number of

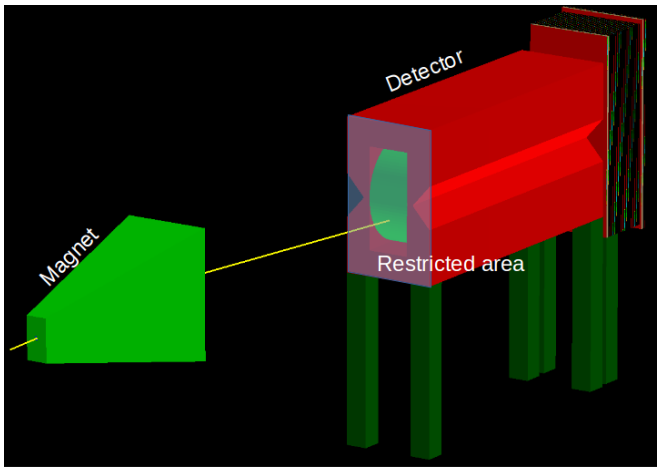
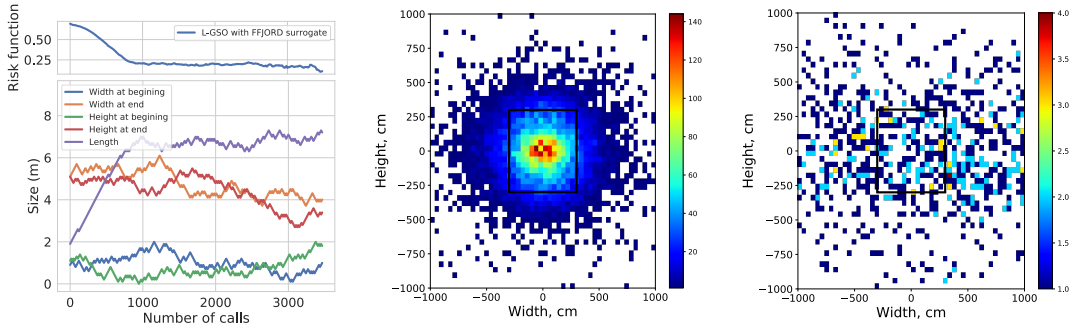


Figure 5.15: Toy physical detector setup. Green: optimised magnet; red: sensitive detector; thin blue area: restricted area, which we want to keep clear of the muons. The yellow line represents the trajectory of the particle. The screenshot is from FairShip software.

muons reaching the sensitive plane. The motivation behind this choice of the loss function will be discussed in the next chapter.

The results of the optimisation are presented in Figure 5.16. In Figure 5.16a the value of the loss as a function of the iteration is presented at the top and the dynamics of frustum parameters at the bottom. We see that the loss is rapidly decreasing, as the length of the magnet increases, as expected. However, afterwards, the loss reaches the plateau, and the frustum parameters oscillate without significant change. At first, this might seem strange because we would expect the length of the magnet to increase constantly until no more muons reach the sensitive plane. However, after deeper inspection of this phenomenon, we observed that the stagnation of the optimisation was due to the "wrong polarity" muons. These are muons that enter the magnet on the wrong side and are bent inwards instead of outwards. Such particles can traverse the whole magnet diagonally and exit from it on the other side, while still having hits in the acceptance of the sensitive plane. As we will see in Chapter 6, these types of events may constitute an irreducible background for the SHiP experiment and, thus, require thorough examination. If such muons enters the magnet somewhere in the middle of the frustum, it can easily hit the sensitive plane. Thus, there will be a saddle point for the optimisation when the number of muons bent out by increasing the length is equal to the amount of muons bent in by the "wrong polarity" and long magnet. The optimisation might get stuck at such a point for a long time or even forever. This is precisely what is observed in the current optimisation. However, the algorithm was able to escape the saddle point, as seen in the loss plot. The loss starts to decrease again after iteration 2500. At the same time, the transverse



(a) (b) (c)

Figure 5.16: (a) Magnet objective function (top) and five $\boldsymbol{\psi}$ parameters (bottom) during optimisation with L-GSO. (b) Histogram of the muons hits distribution in the detection apparatus (depicted as black contour) before (b) and after (c) optimisation of the magnet. Colour represents the number of hits in a bin.

shape of the frustum at the end is shirking and the length is growing. This is also well-motivated from a physics point of view: the optimisation tries to sweep away those "wrong polarity" muons by reducing the time they spend inside the magnet or not letting them touch the magnet at all, thus, potentially missing the restricted area, but at the same time increasing the length to sweep out "correct polarity" muons. The same effect will be observed in the full-scale SHiP shield optimisation. The muon hits distribution before and after optimisation is presented in Figure 5.16b and Figure 5.16c. As one can see, the density of muons has dropped by two orders of magnitude.

Examining the results for the toy physics model, we can conclude that the L-GSO algorithm was able to perform the optimisation of the five-dimensional frustum successfully. Thus, it is worth applying the algorithm to the full-scale muon shield optimisation and improving the shield performance.

5.3 Summary

In this chapter, a novel optimisation algorithm was presented. It is tailored to the optimisation of black-box stochastic non-differentiable functions, such as the GEANT4 simulator. The core idea and the drawbacks of a generative surrogate optimisation (GSO) algorithm were discussed. A range of toy experiments proved the feasibility of the idea. To solve the drawbacks of the GSO, a local generative surrogate optimisation was introduced (L-GSO). It was extensively tested against

other existing optimisation algorithms on a wide range of toy problems and then successfully applied to the optimisation of the frustum shape magnet.

The L-GSO algorithm was compared to various optimisation algorithms, from different optimisation approaches. In all the problems L-GSO is comparable or outperforms the baseline methods in terms of speed of convergence. L-GSO especially excels in high-dimensional problems, where it outperforms other methods by a large margin. Although, theoretical guarantees of the L-GSO convergence do not exist, empirically, it is unbiased and has low variance.

There are two main directions for further improvement of the algorithm. As discussed, currently, the hypercube size ϵ is fixed before the optimisation, which might not work in the case of very noisy or complex functions. In addition to that, the user must perform some hyperparameter search before the optimisation to find an optimal value of ϵ for a particular task. Thus, the first direction of work would be to make an adaptive computation of ϵ, such that the size of the hypercube would change according to the current landscape of the objective function. The second direction of work would be to combine L-GSO optimisation with more exploratory algorithms, such as Bayesian optimisation or genetic algorithms. This might solve the problem of getting stuck at a local optimum and provide a better strategy for selecting a starting point for the optimisation.

Chapter 6

SHiP shield optimisation

The problem of optimising the SHiP muon shield has been a central topic since the proposal of the experiment. No existing beam dump experiments feature the same proton beam energy and intensity and, thus, the same muon rate. In the SHiP experiment, a rate of 10^{11} muons/s with energies up to 350-400 GeV is expected. While, in theory, it is possible to absorb muons using passive shielding, it can not be done in the case of SHiP. At SHiP's energies, one would require hundreds of meters of concrete as an absorber, which would dramatically limit the sensitivity to the Hidden sector particles due to their large transverse momentum. Thus an active shield constructed from "warm" magnets was proposed. The shield must have a high efficiency of background rejection while being as short as possible to maximise signal acceptance. In addition, the construction of the shield must be feasible in terms of the constraints posed by the available technologies and funds. As such, several iterations of optimisation have been done, and the optimisation is still ongoing.

6.1 Optimisation problem statement

For the experiment to operate, the muon flux rate must be reduced by at least six orders of magnitude at the tracker station. The rate at the SND detector and the decay volume walls must be manageable as well. For example, the rate at the SND's emulsion films must be lower than 10^3 muons/mm^2 before emulsion replacement, otherwise, it would not be possible to reconstruct signal tracks. For the decay volume's SBT, the rate is solely dictated by the limitations of the electronics signal readout.

As in the previous optimisation iterations, only the rate at the tracker station is

considered. Although it is possible to include the rate at the SND and the SBT in the optimisation, it has been left for future work. Some preliminary considerations about how that could be accomplished are discussed later in the chapter.

The formulation of the optimisation problem is the same as for the toy physics problem of Chapter 5: we want to find the optimal parameters of the magnets that minimise the rate of muons in the tracker. To simplify the optimisation and remove other factors that introduce randomness into the optimisation, all sub-systems of the experiment were removed, and only the proton target, hadron stopper and muon shield were left in place. The tracker stations have been substituted with a sensitive plane located in the place of the first station of the tracker system. The SND and decay volume were removed because they were not included in the optimisation loss but would induce muon scattering and thus complicate the optimisation. Including them might have been beneficial if a full muon sample could have been used, but this is computationally infeasible, as discussed later.

The hadron absorber geometry is fixed and is excluded from the optimisation, the same as was done in the previous optimisation iteration [79]. The shape of the hadron stopper was excluded from the optimisation due to the radiological constraints imposed by the civil protection requirements.

With all the above simplifications, the optimisation problem is formulated as follows:

$$\begin{gathered} \boldsymbol{x} = (P_x, P_y, P_z, PID, x, y, z), \boldsymbol{x} \in \mathbb{R}^7, \\ (P_x, P_y, P_z, PID, x, y, z) \sim \text{available data samples} \\ \boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \mathbb{E}[\mathcal{R}(y)] \end{gathered} \tag{6.1}$$

where $\mathcal{R}(y)$ is a user-defined objective function that resembles the penalty for a muon for reaching the tracker, and $\boldsymbol{\psi}$ are the parameters of the shield. The input vector $\boldsymbol{x}$ characterises the kinematics properties of the muon: P_x, P_y, P_z are the components of the momentum, x, y, z are the coordinates of the muon production point inside the target, and PID is the electric charge of the muon. These quantities are sampled from a pre-production distribution, obtained from the simulation of a proton-target collision, and are described in the next section.

Currently, the magnet consists of six parts included in the optimisation. Note that everywhere throughout the thesis the numbering of the magnet sections starts from zero. This is because the zeroth section is now a part of the hadron absorber and excluded from the optimisation but presents in the figures for completeness. For example, in Figure 6.12, the initial rectangular part is the zeroth section.

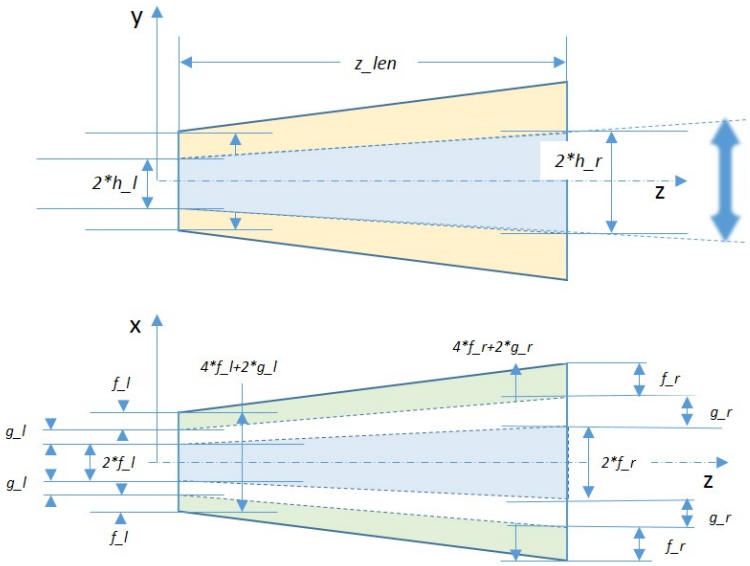


Figure 6.1: A schematic drawing of YZ and XZ projections of the frustum with the parameters caption. [93]

The YZ and XZ projections of the frustum with the labelled parameters are presented in Figure 6.1. Each part is a frustum characterised by seven parameters:

1. f_l, f_r correspond to the X-half width of the shield central limb at the beginning and the end of the frustum.
2. h_l, h_r correspond to the Y-half height of the coil gaps at the beginning and the end of the frustum.
3. g_l, g_r correspond to the X-width of the gaps at the beginning and the end of the frustum.
4. z_len correspond to the half-length of the frustum.

The magnetic field is fixed to 1.7 T throughout the optimisation and is directed upward (in the Y direction) in the central limb of the magnet for the first three magnets. In the last three magnets the direction of the field is flipped. This is the critical aspect of the design, as many muons are bent back into the detector acceptance in the first three sections of the magnet and must be removed in the last three sections. Furthermore, the field integral provided by the first sections and by the return field in the last sections is sufficient to deflect high energy muons. Notably, the field strength of 1.7 T is a conservative estimation of the achievable magnetic field at low coil current for grain-oriented (GO) steel.

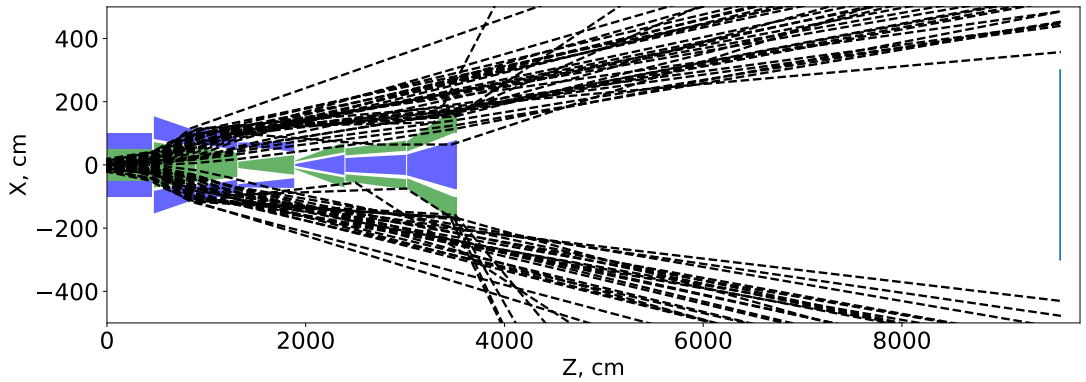


Figure 6.2: A schematic drawing of the muon shield optimisation setup in the XZ projection. The colour corresponds to the direction of the magnetic field.

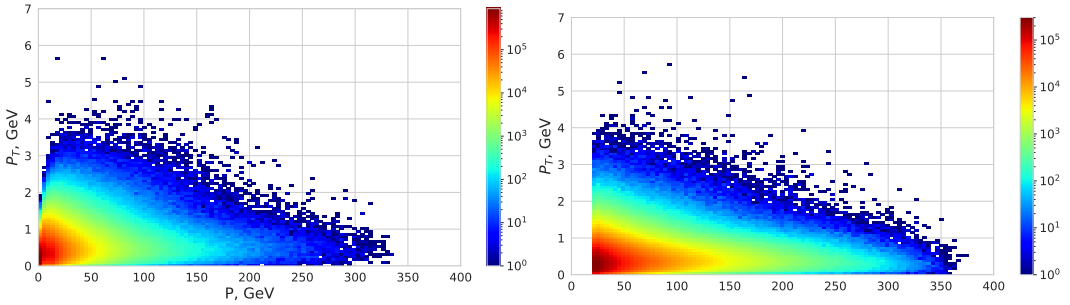
The schematic drawing of the optimisation setup is shown in Figure 6.2. The colour represents the magnetic field direction, with green indicating the field direction in the positive Y-axis. Black dashed lines are the trajectories of the muons deviated by the magnetic field. The blue line at $X \approx 9000\,\text{cm}$ is the sensitive plane.

Given that both the strength and the direction of the magnetic field as well as the number of magnets are fixed, the total number of optimised parameters is 42. This results in a high-dimensional optimisation problem, where it is costly to run a simulator — an ideal application for the L-GSO algorithm.

6.1.1 The muon sample

The $\boldsymbol{x}$ input in the mathematical formulation above describes all kinematic properties of the incoming muons. Unlike in the toy example in Chapter 5, where the distributions from which those properties are sampled are known, in real optimisation the samples are obtained from the simulation.

A dedicated pre-production stage using Pythia and GEANT4 was done, where 400 GeV protons impinged on the target and the muons surviving the hadron stopper were stored. The magnetic field in the hadron stopper was removed for the pre-production. In addition, the fraction of rare muon production processes, such as resonant EM and charm decays, was enhanced and stored with the appropriate event weights. After the simulation, a normalisation of the spectrum was performed by assigning the weights to the events, such that the total yield corresponded to one spill of the SHiP experiment. The weighting was required to appropriately scale the rare production channels and perform estimations of the background with the limited computing power.



(a) (b)

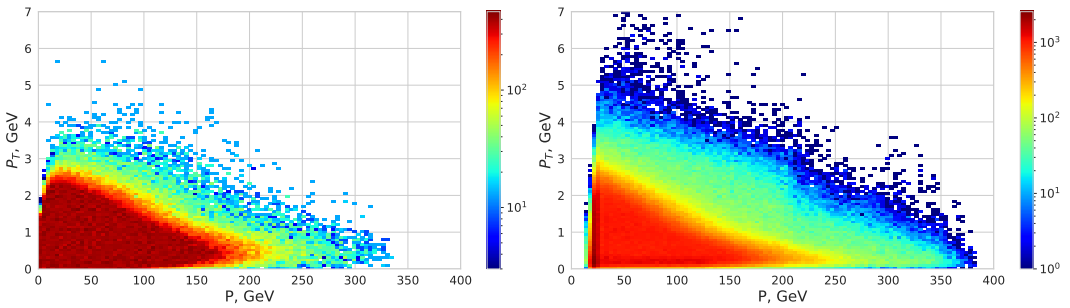
Figure 6.3: $P - P_T$ distribution for muons from (a) the small muon pre-production sample, (b) large muon pre-production sample (the histogram is based on 3% of the sample).

At the time of this thesis is written, two large samples of muons are available. The first one contains $\sim 1.7 \cdot 10^7$ muons, whereas the second one has $4.96 \cdot 10^8$ muons. The $P - P_T$ spectra corresponding to these pre-production samples are presented in Figure 6.3.

The key difference between those distributions is that the large sample has a 20 GeV cut on the muon energy. It was estimated that the muons with energy less than 10 GeV would not survive the hadron stopper and would waste the simulation CPU time [94]. Thus, muons with initial energy less than 10 GeV are not propagated through GEANT4. Furthermore, muons are only saved if the muon energy is larger than 10 GeV after the hadron stopper. Such a strategy results in a 20 GeV energy cutoff. In addition, the new sample was enriched with the muons from the EM resonance decays that typically result in high momentum muons. This explains the heavier tail of the large sample.

However, it is computationally impossible to use those samples for optimisation. For example, the large sample takes approximately one day to run on the CERN grid, which is impractical as the optimisation requires around $O(100)$ steps to converge. Thus during previous optimisation a "resampled" sample was devised [79]. It was obtained by resampling the small sample in such a way that each $P - P_T$ bin carries at least ten but at most 100 muons, where the azimuthal angle ϕ was sampled from a uniform distribution. This resulted in a sample of the size $4.75 \cdot 10^5$ muons, which was feasible to run through SHiP simulation software in around 30 minutes. This distribution also added heavier tails to the original one, as it was estimated that high energy muons are the most problematic for the muon shield. The "resampled" distribution is shown in Figure 6.4a, and further details can be found in [79].

When using the "resampled" distribution, the amount of muons that passes the shield



(a) (b)

Figure 6.4: $P - P_T$ distribution for muons from (a) "resampled" sample, (b) muGAN generated sample.

reduces to $O(10)$ events at the end of the optimisation. This results in a large amount of noise in the optimisation, preventing it from obtaining a better shield design. To solve this problem, the optimisation was performed using GAN sampled muons. A dedicated Generative Adversarial Network (GAN) was trained [73] to mimic the pre-production distribution in Figure 6.3b. It was further modified with auxiliary variables that allowed tweaking the output distribution and manually setting the preference for heavier tails. Those variables were adjusted to make the output distribution of the GAN close to the "resampled" distribution but with heavier tails to represent the large pre-production sample. The main benefit of the GAN sampling approach, is that it allows to sample arbitrary large quantities of muons and adjust the number of samples dynamically during the optimisation. For example, when the very few muons reach the sensitive plane, the optimisation becomes noisy and may stall due to a very small gradient. With the GAN generated muons, one can increase the number of input particles by the desired quantity, potentially increasing the number of muons reaching the sensitive plane. In return, larger statistics will result in more robust optimisation. However, this comes at the cost of an additional simulation time. Thus, the amount of the sampled muons was constrained to 10^6 in all the optimisation runs.

The last component of the optimisation problem formulation is the loss function.

6.1.2 Objective function

The objective (or loss) function should, on one side, penalise the rate of muon hits in the tracker stations and the length and/or weight of the magnet on the other. The latter is needed to minimise the shield length as well as construction and production cost. The minimisation of the shield length is crucial for the performance of the

whole experiment.

The first part of the loss that quantifies the rate of muons in the tracker may be defined in different ways. The following loss function was used in the previous iterations of the optimisation:

$$\mathcal{R}_{muon}(\mathbf{x}, \mathbf{y}; L, R) = \sum_{i=1}^{N} \Big(\mathbb{1}_{Q_i=-1} \sqrt{[(L+R) - (\mathbf{x}_i + R)]/(L+R)} + \mathbb{1}_{Q_i=1} \sqrt{[(L+R) + (\mathbf{x}_i - R)]/(L+R)} \Big) \ , \quad (6.2)$$

where $\mathbb{1}$ denotes the indicator function, Q is the muon charge, $\mathbf{x}$ is the x-coordinate of a muon hit in the sensitive plane and L, R are the boundaries that resemble the detector shape. The parameter L corresponds to a half-width of the detector and controls the penalty margin for muons that are bent in the correct direction. The parameter R controls the penalty margin for the "wrong" polarity muons. By varying those parameters one can control the effective size of the detector. For example, if the muons are swept out too close to the detector's edges, the value of the parameter L can be increased to shift muons away from the edges. The loss is only applied to the muon hits inside the acceptance of the tracker: $|\mathbf{y}_i| < 500\,\text{cm}$. In the default definition of the loss function $L = 260\,\text{cm}$ and $R = 300\,\text{cm}$ was used. The loss function is shown in Figure 6.5. The choice of this particular loss function is motivated by the idea of penalising muons for being on the "wrong" side of the magnet. For example, μ^- must be deflected to the positive side of the X-axis. If, on the contrary, the particle is deflected to the negative side, it is treated as a "wrong" polarity muon. The "wrong" polarity particles can not be conceptually deflected by the muon shield and constitute the most challenging background source. Thus, with this choice of the objective function, we want to penalise the shield for sweeping away muons in the wrong direction, even if they miss the detector.

The weight term of the loss function is defined as follows:

$$\mathcal{R}_W(W) = 1 + e^{10 \cdot (W - W_0)/W_0}, \quad (6.3)$$

where W is the total weight of the shield and W_0 is the baseline configuration weight. The total loss function used in the previous optimisation looked as follows [79]:

$$\mathcal{R} = \begin{cases} 10^8, & \text{if } W > 3\text{kt} \\ \mathcal{R}_W(W) \cdot \mathcal{R}_{muon}(\mathbf{x}, \mathbf{y}; L, R), & \text{otherwise} \end{cases} \tag{6.4}$$

However, this objective function is not suitable for any gradient-based optimisation. First of all, there is a non-differentiable part of the loss at the point $W = 3\,\text{kt}$, when the function switches regimes. In addition, the branch of the function corresponding to $W > 3\,\text{kt}$ does not depend on the parameters and will result in zero gradients. Secondly, the term $\mathcal{R}_W(W)$ results in gradient values that are similar to the gradients from $\mathcal{R}_{muon}(\mathbf{x}, \mathbf{y}; L, R)$ even for a low number of penetrating muons. Such function dependency assigns too much weight on the term $\mathcal{R}_W(W)$, which will result in a small, but an inefficient shield. Thus, one should either modify the coefficients of the loss or rewrite it in a more suitable way. During the initial runs of the optimisation it was observed that the algorithm was more stable without the weight term. Moreover, configurations produced by L-GSO were lighter than those obtained with BO. Thus, it was decided to define the objective as $\mathcal{R} = \mathcal{R}_{muon}(\mathbf{x}, \mathbf{y}; L, R)$ without the mass term.

Another observation that became available by the time of this work is that the cost of the shield is not strongly correlated with the weight. The most significant part of the cost comes from the manufacturing process, rather than the amount of material required. Thus, the weight term completely lost its relevance for the optimisation. Manufacturing and production processes are the main drivers of the cost but are extremely hard to quantify in terms of loss value. Thus, this part was left for future work.

Finally, as discussed before, no other sub-systems of the experiment were included in the optimisation. By the time of this work preliminary considerations suggested that the SND and SBT should be included in the optimisation, but the strict requirements for the sub-systems have not yet been deduced. For example, it is known that the SBT could not tolerate high EM background rates and, thus, this constraint should be included in the optimisation. It is, however, quite challenging to do, since by default, only muons are being simulated to estimate the muon background and adding an EM component would significantly increase simulation and optimisation complexity. For the SND, the background rate through an emulsion film should not exceed 10^3 tracks/mm^2 for the detector to operate. It is easy to shield from an EM background by placing a section of lead or concrete between the muon shield and the SND. On the contrary, one can not easily shield from the muon flux, and thus it has to be included in the optimisation procedure.

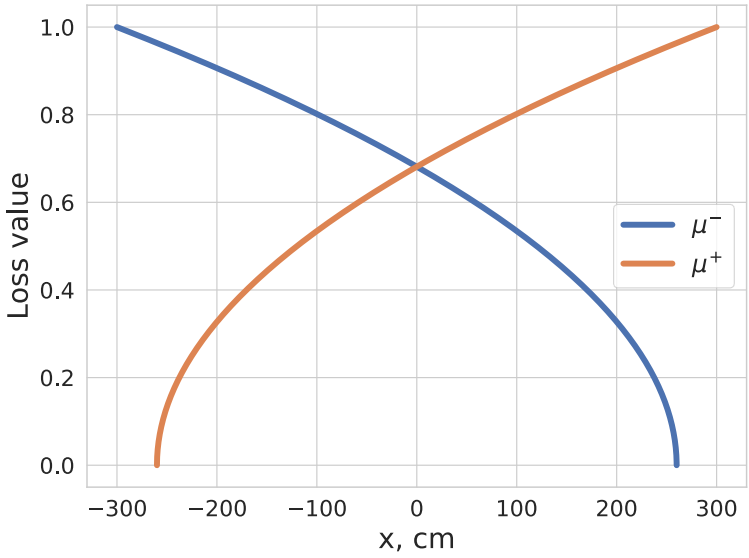
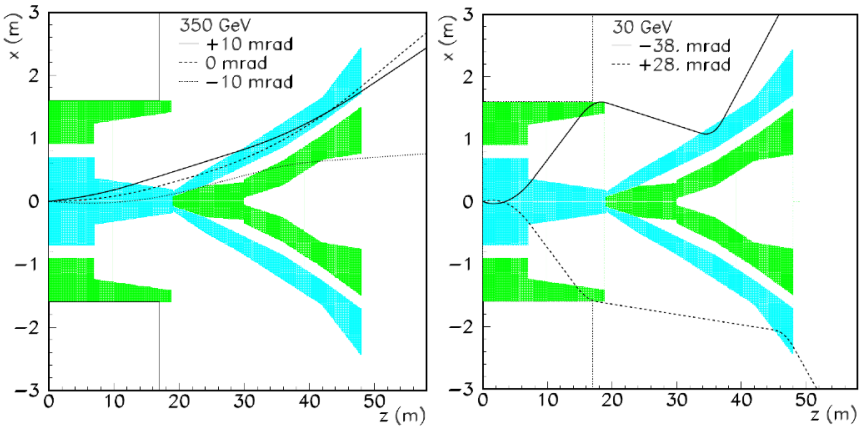


Figure 6.5: The objective function $\mathcal{R}$ to penalise muon hits location on the sensitive plane. The plot corresponds to parameters $L = 260, R = 300$.

6.2 Previous studies

Given the complexity of the muon shield and its crucial role in reducing the combinatorial background, several iterations of the shield optimisation have been done. In the Technical Proposal (TP) phase, the first design of the magnet was introduced. It consisted of two parts, both magnetised to 1.8 T, where the magnetic field's polarity is flipped in the second part with respect to the first part. The schematic view of such a shield can be seen in Figure 6.6. Such a configuration of the shield was optimised for the cylindrical shape of the decay volume and was very impractical: its weight was thousands of tons of iron and its length was around 50 m. However, it provided a key intuition to the construction of the magnet system that one has to switch the polarity of the magnets to sweep out slow muons that are turned back into the detector acceptance by the return field. This is well illustrated in Figures 6.6a and 6.6b, where in the first scenario, a 350 GeV muon is constantly located in the correct field and swept outside. On the contrary, in the second example, the 30 GeV muon is being swept out immediately after entering the shield and re-scatters back into the acceptance in the return field of the first magnet. As can be seen, if the polarity of the second magnet had not been flipped, the muon would have been focused back into the detector acceptance. Due to the reversed polarity, it is swept outside and thus misses the downstream detector.

The TP configuration was further optimised [95] using MINUIT [75]. The main benefits after this optimisation were the reduction of the length to 35 m and a weight



(a) (b)

Figure 6.6: Schematic view of the SHiP shield in XZ plane as proposed in the TP. Potential trajectories of (a) 350 GeV muon and (b) 30 GeV muon are shown. Colour represents the direction of the magnetic field. [79]

reduction in half. This was possible due to the magnetisation of the hadron stopper with a field of 1.8 T and the construction of the shield with six magnets instead of two. An additional iron plate was introduced after the last magnet to prevent muons from scattering on the edges. This design, however, was relying on fast simulation and unrealistic assumptions on the magnetic field strength and thus, a third optimisation was performed.

A third iteration of the optimisation was performed utilising the GEANT4 simulator [71] in combination with machine learning techniques [96]. Before this optimisation, a realistic design and magnetic field of the hadron stopper had been devised and thus, it was removed from the optimisation. The field inside the hadron stopper was set to 1.6 T and inside the shield to 1.7 T. Lower field values provide a more robust design from the engineering perspective and are easier to implement in practice. The resulting design works as efficiently as the previous optimum with the muon rate of ~ 50 kHz but is $\sim 25\%$ lighter. A detailed engineering drawing of the magnet is presented in Figure 6.7. Notice that the coils, shown in orange, are located inside the magnet for the first half of the magnet and outside for the second half. This ensures uniformity of the magnetic field in the most active zones of the shield.

Examining the distribution of the field closely in Figure 6.8, it can be seen that the distribution of the field is uniform in the central part of the shield. However, the field drops dramatically on the sides and edges, which can negatively affect the performance of the shield. Even if the fraction of muons affected by the field non-uniformity is small, at the rate of 10^{11} muons/s this number can become significant and requires further investigation.

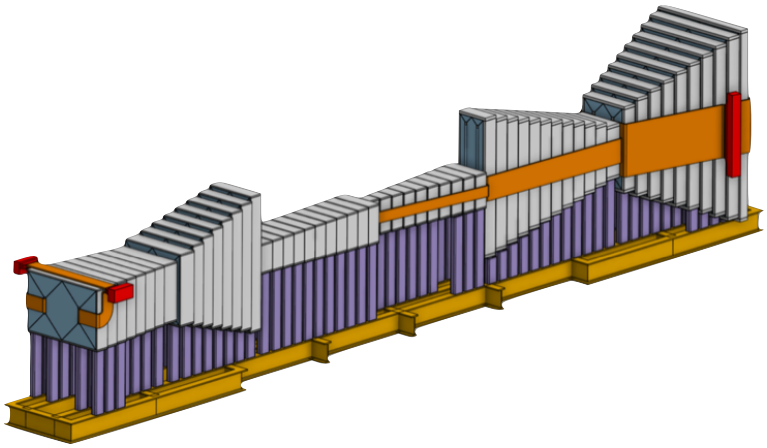


Figure 6.7: A CAD engineering model of the current SHiP magnet design. The magnetic coils are shown in orange.

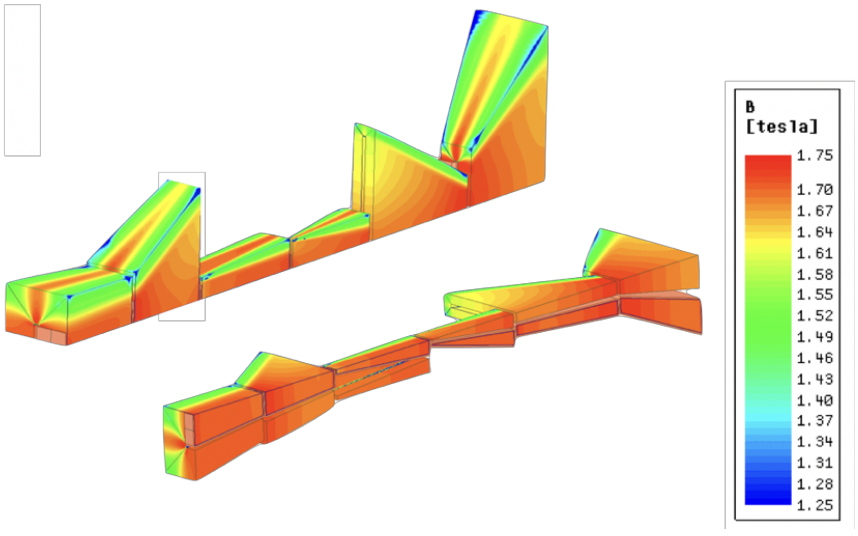


Figure 6.8: Distribution of the magnetic field in the magnet. A quarter of the shield is shown in Y-cut and Y-cut rotated by 90 degrees. [50]

After revisiting the current status of the optimisation, it has become clear that a new iteration of the optimisation and a dedicated test-beam experiment is needed to validate the simulation and engineering grounds of the muon shield.

In particular, the previous iteration of the optimisation did not take into account that each section of the shield can only be constructed from the rectangular blocks and could not be a perfect frustum (see Figure 6.7). This may affect muons scattering and thus can lead to an increase in the penetrating rate. Secondly, and most crucially, throughout the optimisation process the field in the shield was assumed to be uniform in the whole magnet. Dedicated studies were done to check how the above imperfections affect the results and are discussed further in this chapter.

In addition, some concerns were raised whether the current design found by the BO is optimal. As shown in Chapter 4, theoretically, BO can find a global minimum of a function. In practice, it is rarely possible for complex problems, such as shield optimisation. Thus, the longstanding question was whether it is possible to improve the current design of the shield, without sacrificing the length constraint. With the local optimisation, such as L-GSO, it is quite easy to answer this question.

Finally, it was understood that the effect of large-energy loss and large-angle scattering might significantly affect the result since no experimental data is available for such rare processes at the energies of muons above 10 GeV. Validation of such rare processes and their comparison to the description in GEANT4 are another motivation to perform a dedicated test-beam experiment, which is described in detail in Chapter 7.

6.3 Muon shield re-optimisation

With all the ingredients in place, we first start by applying the L-GSO algorithm to the shield optimisation. The optimisation has been performed from a couple of random starting points around the current optimum as well as from the optimum itself. This strategy allows to answer the longstanding question whether the current configuration indeed corresponds to the optimum. Also, it allows to understand whether this optimum is global or local.

6.3.1 Application of the L-GSO

The L-GSO algorithm was described in detail in Chapter 4. However, to satisfy the physical constraints of the shield, one must account for the minimal size of the coil gaps

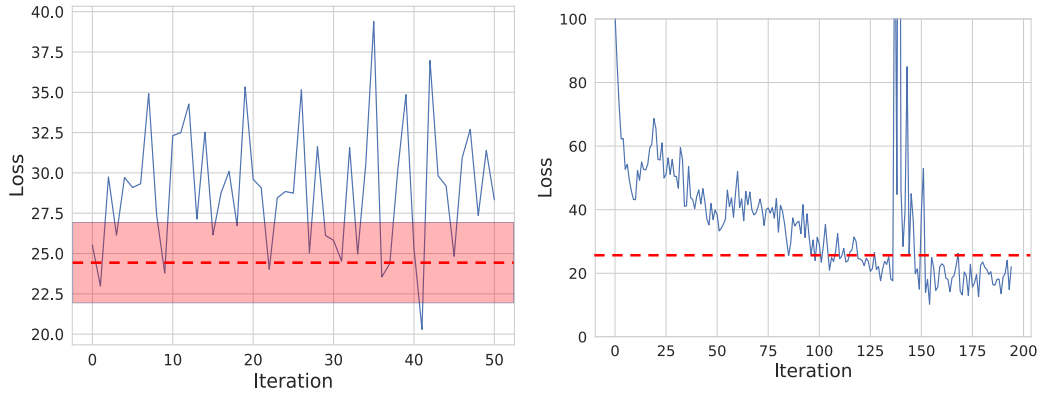
inside the magnet. It is achieved by limiting the corresponding magnet parameters in the GEANT4 simulation. In addition, as the shield cannot be constructed with infinite precision, the shield parameters should change discretely throughout the optimisation. Since the simulation was found to be insensitive to parameters changes smaller than 0.1 cm, it was decided to set this value as a minimal allowed change in the parameters.

The performance of the algorithm is compared using the loss function with the original parameters $L = 260\,\text{cm}$ and $R = 300\,\text{cm}$. This ensures that the same quantities are being compared since increasing or decreasing L, R would increase or decrease the loss correspondingly. Thus, all values of the objective function presented in this chapter are computed for the above values.

The computing power for the optimisation was provided by the Yandex Cluster and Higher School of Economics (HSE) GPUs. Due to the usage of neural networks in the L-GSO on one side and CPU extensive usage by the FairShip on the other, each iteration of the optimisation was performed on separate machines. At each iteration $O(80)$ $\boldsymbol{\psi}$ magnet configurations were generated using Latin Hypercube sampling. For each of those configurations 500000 muons, generated using muGAN, were simulated using FairShip. Each such configuration utilised 16 Cores of the Yandex Cluster, totalling in ~ 1300 simultaneous CPU cores utilisation. The distribution of muon hits $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\psi})$ on the sensitive plane for each magnet configuration $\boldsymbol{\psi}$ and muon parameters x was saved into the history replay as in step 5 of Algorithm 5. This step ensured the re-usage of the data for the surrogate training procedure. After this step, the data was transferred to the HSE GPU machine, where the surrogate was trained on a single Tesla K80 GPU. Then the gradient step was done, and the process repeated for the updated configuration of the magnet.

First, the value of the loss is assessed using the "resampled" muon sample to make a fair comparison with the results obtained in the previous iteration of the shield optimisation. Afterwards, the performance is evaluated using the number of muons hits in the first Target Tracker (TT) station, corresponding to the number of muons per spill after normalising using the weights. This is done using large pre-production sample but simplified experiment geometry. Finally, the same pre-production sample is run for the complete geometry of the SHiP experiment to account for SND and TT magnetic fields and scattering in the material. The rates of muons between baseline and new shield configurations in the first TT are then compared.

To test the L-GSO algorithm, the optimisation was repeated starting from the current configuration. If the configuration corresponds to the local optimum, then

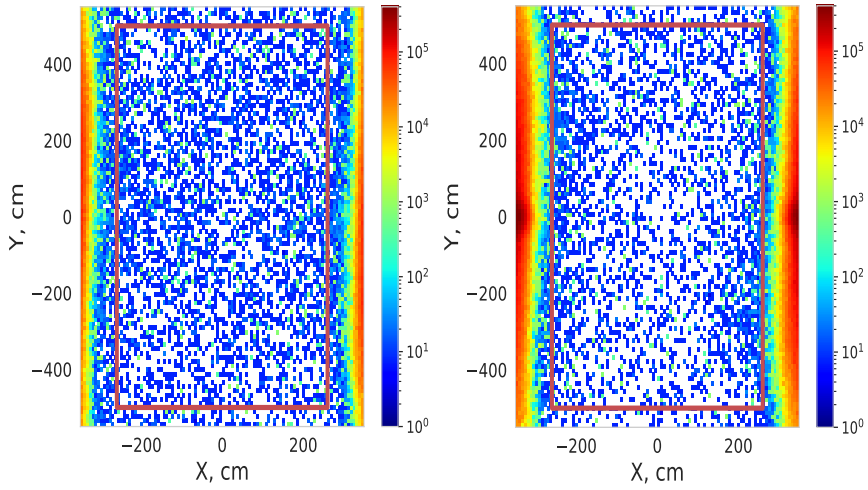


(a) (b)

Figure 6.9: (a) Objective function value during optimisation (a) started from the current configuration, (b) started from the perturbed configuration. The red line corresponds to the loss value as in [79]. The shaded region corresponds to the estimated uncertainty.

the L-GSO algorithm would oscillate around the current value without any significant decrease of the objective function value. It was found to be the case, as can be seen in Figure 6.9a. The optimisation was run for 50 iterations and showed oscillating behaviour around the initial value of the objective function ~ 25, shown in red. This proved that the current shield configuration was indeed the local optimum.

Since L-GSO is a gradient-based procedure and is much more sensitive to small changes in the muon flux rate, the objective function parameter L was changed from 260 cm to 290 cm during the optimisation. The increase of the parameter value shifted the boundary point at which the objective function became zero, improving the optimisation robustness. Afterwards, the optimisation was run from a new magnet configuration, which was a small perturbation of the current optimum. If the current optimum is a local one, it would be possible to find a better value of the objective function by starting L-GSO from a perturbed geometry. Figure 6.9b shows the value of the objective function when the optimisation was run from such a perturbed initial point. One can see, that even though the optimisation starts from a much higher loss value, it is able to converge to a loss value below 20, improving the previous optimum. The reason behind the large variance in the loss plot around iteration 130 is the stochastic nature of the problem. Due to the stochastisity of the optimisation it can sometimes make a step in an suboptimal direction. Furthermore, since the shield configuration is sensitive to small changes, the suboptimal step can lead to large values of the loss function. However, the optimisation is capable of escaping such regions of high uncertainty and successfully converging, as seen in Figure 6.9b.



(a) (b)

Figure 6.10: Weighted distribution of the muon hits in the sensitive plane obtained from (a) BO, (b) L-GSO. Red contour corresponds to the detector boundaries.

6.3.2 Characteristics of the configuration obtained with L-GSO

The obtained configuration and the BO baseline were run on the largest available muon sample of $4.96 \cdot 10^8$ muons to validate the results. The weighted distribution of muon hits, corresponding to one spill, is presented in Figure 6.10. Even by eye, it is possible to notice that the density of the hits in the middle of the detector acceptance is more sparse for the L-GSO configuration. The loss function value for the new optimum is, indeed, smaller: 2200, versus 3000 for the baseline. Additionally, the muon shield corresponding to the new optimum is lighter and shorter. These results are summarised in Table 6.1.

Another notable aspect of the optimisation is that the high density of the muon hits lies much closer to the detector boundaries in the case of using L-GSO, as can be seen in Figure 6.10b, but drops fast as one moves close to the centre of the detector. That can be further seen in Figure 6.11. In the optimum, found by the L-GSO, the number of hits drops polynomially as the detector size shrinks. This suggests that the L-GSO optimisation is more precise than the Bayesian optimisation and allows accurate fine-tuning of the shield parameters, if needed.

The resulting shape is presented in Figure 6.12b. One can notice that the optimisation has shortened the fifth section of the shield and widened the return yokes of all magnets. This change, in turn, reduced the amount of low and high energy muons penetrating the shield. The $P - P_T$ distribution of initial muon momentum plotted

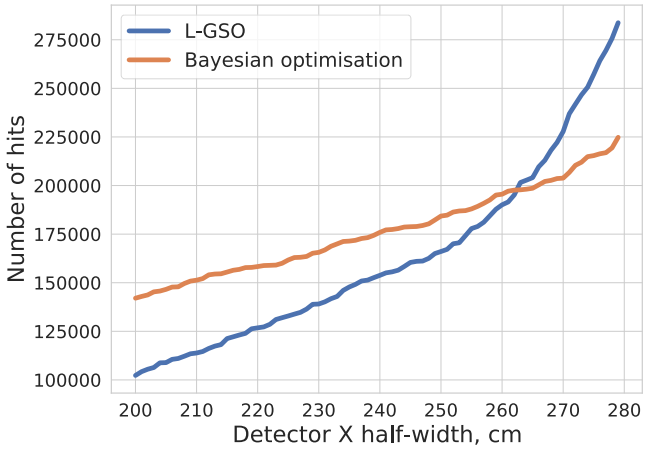


Figure 6.11: Number of the muon hits per one spill in the sensitive plane as a function of the plane X half-width.

Table 6.1: Comparison of the optimisation results for Bayesian optimisation and L-GSO.

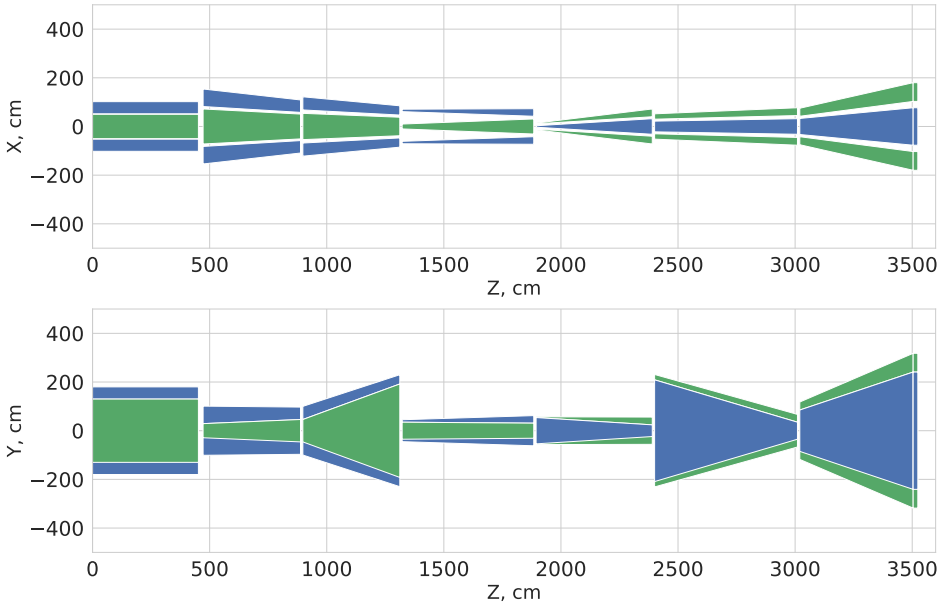
Method	Loss	Shield length (m)	Magnet weight (kt)
Bayesian opt.	3000	35.5	1.27
L-GSO	2200	34.4	1.05

for muons penetrating the shield is shown in Figure 6.13.

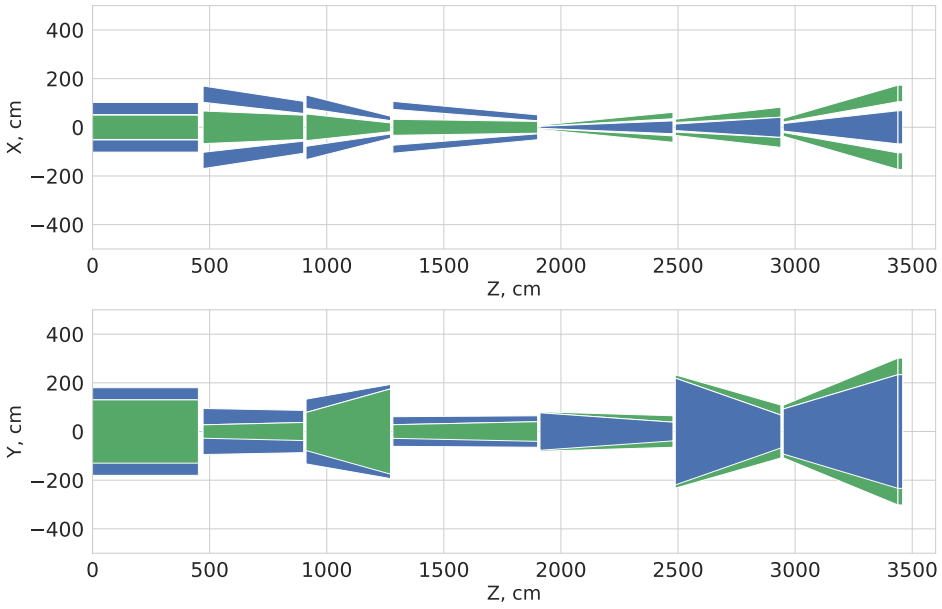
It is clearly seen that in the case of the L-GSO design, there are no muons with high momentum and much fewer muons with high P_T for energies above 200 GeV as well as less $30 - 40$ GeV muons. However, there are more $50 - 100$ GeV muons present in the L-GSO spectrum. The widening of the spectrum suggests that the new optimum has a less pronounced focusing effect than the BO configuration.

The observation that the new optimum can remove all high energy muons, while having a shorter length, is quite peculiar. As it turns out, the ability to remove high energy muons comes from the last three sections of the magnet, where now each next magnet is "nested" in the previous and not a "continuation" of the previous magnet. In such a way, even if high energy muons are flying inside the coil gap in one magnet, they eventually face the return field in the next magnet, thus being constantly swept out. In the baseline configuration, they might fly along the coil gap until they exit the shield. This also explains why the L-GSO result has lower loss and fewer muons in the centre since high energy muons contribute the most to the central hits.

After the optimum performance was assessed on a simplified geometry and the large pre-production muon sample, the performance was ultimately tested with

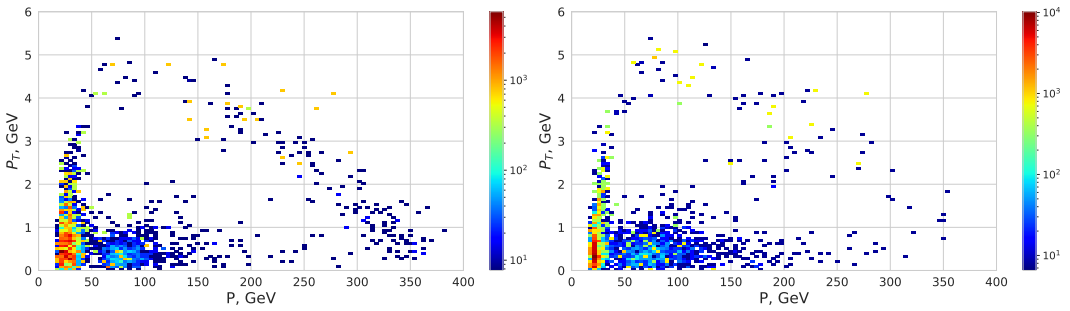


(a)



(b)

Figure 6.12: Muon shield XZ and YZ projections from (a) the BO optimum, (b) the L-GSO optimum.



(a) (b)

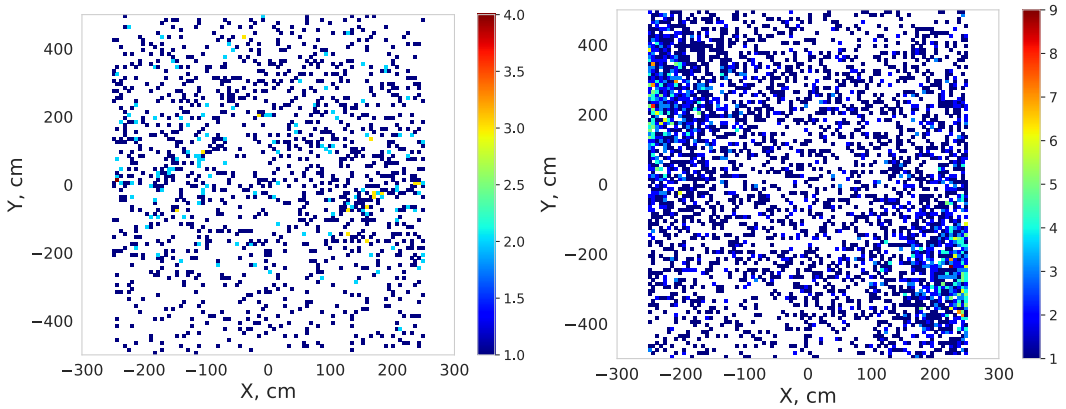
Figure 6.13: Initial weighted $P - P_T$ distribution for muons reaching the sensitive plane. (a) BO configuration, (b) L-GSO configuration.

the complete geometry. The large sample of muons was run through the FairShip simulation with all the sub-detectors in place. The resulting distribution of the hits in the first TT for the BO configuration and the L-GSO configuration is presented in Figure 6.14.

The weighted amount of muons with the hits in the TT is $\sim 5.5 \cdot 10^4$ for the BO configuration and $24.5 \cdot 10^4$ for the L-GSO. This means the L-GSO optimum has four times more muons when simulated with complete geometry instead of a 30% decrease in the loss function value and a 15% decrease in the amount of muons for the simplified geometry. One can also see an asymmetry in the hits distribution, which can not be explained by the magnetic field, since the first TT is located before the spectrometer magnet. A dedicated study showed that the cause of the hits asymmetry is the distribution skewness in the pre-production files, further enhanced when the muon propagates through the shield and the decay volume. However, the asymmetry does not affect the estimation of the muon flux.

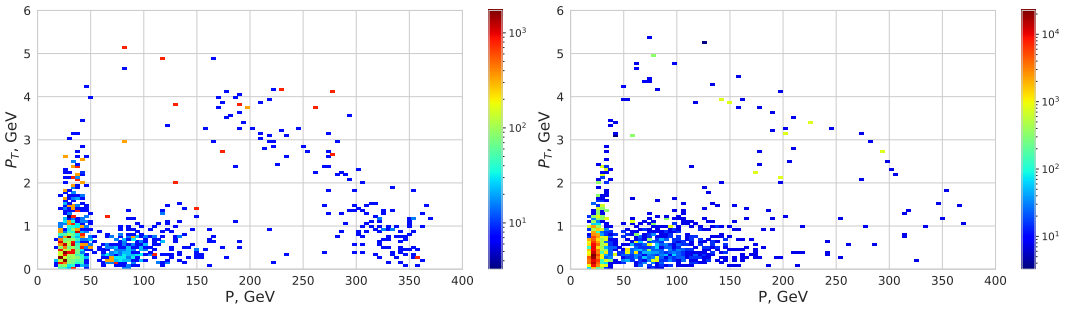
To understand the cause of the rate increase, we first compare the muon $P - P_T$ distributions for the complete geometry in Figure 6.15. The distributions look similar to those in Figure 6.13, except that the BO configuration has fewer low-P muons. This suggests that the effect is twofold. First, low energy muons are being stopped and/or swept out by the SND and the muon filter in the complete geometry. This effect is the same for the BO and L-GSO optimum. Second, due to the precise optimisation provided by the L-GSO, all muons that were swept just enough to miss the detector in the simple optimisation were then scattered back by the decay volume walls.

As it turns out, if we plot the distribution of the hits in the decay volume walls, that corresponds to passed muons and presented in Figure 6.16, we see that most of them



(a) (b)

Figure 6.14: Muon hits distribution (unweighted) in the first Target Tracker for large sample of muons and complete geometry. (a) BO shield configuration, (b) L-GSO shield configuration.



(a) (b)

Figure 6.15: Initial weighted $P - P_T$ distribution for muons reaching the first Target Tracker in complete geometry. (a) BO configuration, (b) L-GSO configuration.

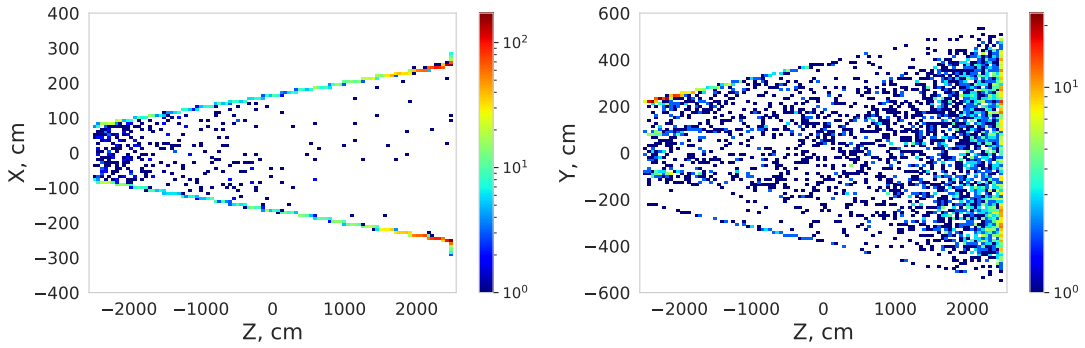


Figure 6.16: Hitmap of the muons in the decay volume walls in XZ and YZ projections.

have hits just at the end of the decay volume. Thus, it is precisely these muons that provide the hotspots in the distribution in Figure 6.14b. To remove the hotspots, a couple of approaches might be taken. In the first, one can reduce the transverse size of the decay volume such that its edges are entirely located in the region free from high muon flux. In the second one, the boundary conditions L and R in the Equation 6.2 can be increased, effectively forcing the optimisation to clean a wider area of the detector surface. This will shift high flux regions of muons further away in the transverse direction, where no decay volume walls are present.

The flux of penetrating muons can be further suppressed by requiring the muon momentum at the TT to be greater than 1 GeV. As discussed in Chapter 3, the common signal selection criterion is that track momentum is larger than 1 GeV. Since $\sim 40\%$ of muons have a momentum smaller than 1 GeV, a significant fraction of the scattered muons will be automatically discarded. The background can be further reduced by assuming that the muon hits in the TT can be matched with the hits in the SBT, since only 4% of muons do not have hits in the SBT.

Another peculiar observation is that within the muons with momentum larger than 1 GeV at the TT station, a significant fraction of them exhibit large energy loss during the propagation through the shield. In Figure 6.17, the ratio of energy loss to that predicted by the Bethe-Bloch formula is shown as a function of initial muon energy. Quantitatively, around 20% of such muons exhibit energy loss at least twice as large as the one predicted by the Bethe-Bloch formula. It can be seen that most of these muons lie in the energy range of $50 - 100$ GeV. Thus, such types of muons can potentially be an irreducible background for the experiment due to the large energy losses, which are rare and hard to simulate. A significant energy loss of these medium-low energy muons accompanied with large-angle scattering will significantly affect the trajectory of the muons in the magnetic field. For example, after exhibiting large energy loss, the muon can be trapped between return yokes

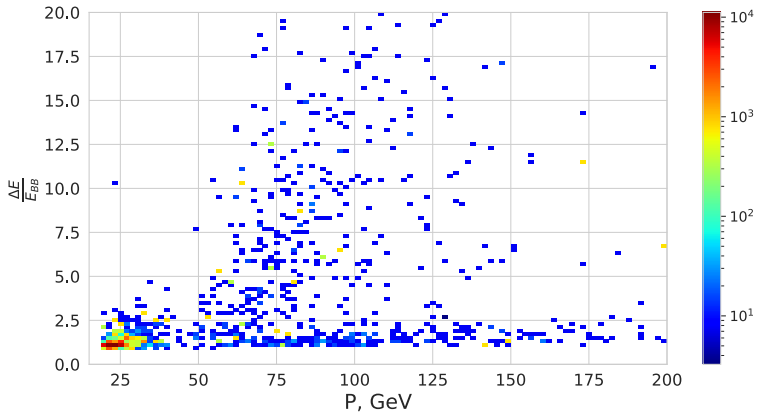


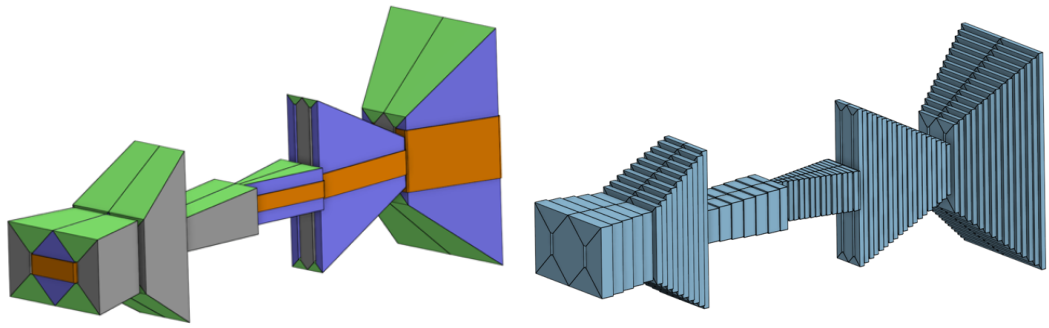
Figure 6.17: Ratio of muon energy loss to Bethe-Bloch theory prediction as a function of the muon momentum.

of the magnet, penetrate the shield and end up in the acceptance of the detector. Similarly, a significant angle change might affect the trajectory of the muons in the magnetic field. Due to low simulation statistics for such types of muons it might be challenging to properly account for such events in the optimisation. On the scale of 10^{11} muons/spill, these numbers can become significant and result in a large uncertainty on the muon rate. This and other factors discussed below, provide yet another motivation to perform a dedicated muon shield prototype beam testing.

After it was made sure that the newly devised L-GSO algorithm can perform a full-scale magnet optimisation, it was subsequently applied to solve some technological challenges of the muon shield.

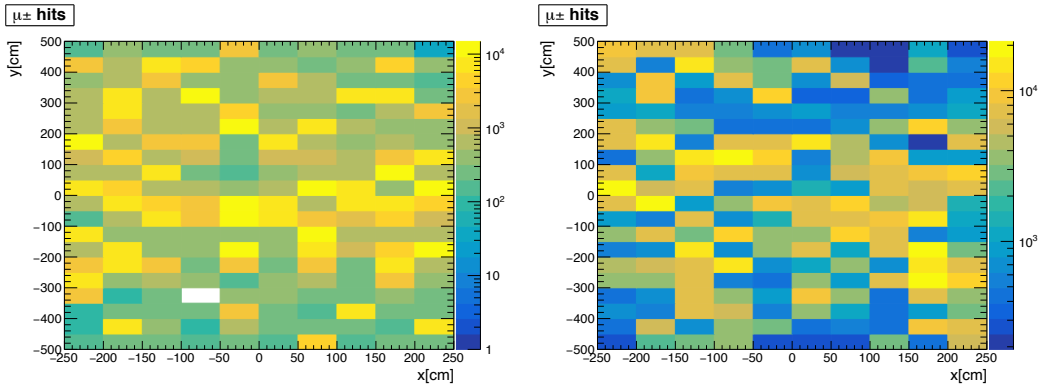
6.3.3 Optimisation of a feasible geometry

During the R&D studies of the shield, it became clear that GO steel must be used, as it was mentioned in Chapter 3. This allows achieving a high magnetic field inside the magnet with a small current inside the coils, all without the need of cooling the magnets. However, such technology brings another limitation: the GO steel can only be produced in sheets of $300 - 500\,\mu$m thickness. The sheets require delicate welding with electron-welding or laser beam technologies, limiting the maximum amount of sheets that can be welded together to ≈ 150, forming a pack of 50 mm thickness. Those packs are then bolted together into rectangular modules of 0.5 m thickness and the magnet is constructed from such modules. In Figure 6.18 the schematic view of the idealised magnet (left) and a magnet constructed from rectangular modules (right) is presented.



(a) (b)

Figure 6.18: View of the muon shield assembled from (a) ideal frustums, (b) module-based frustums. [50]

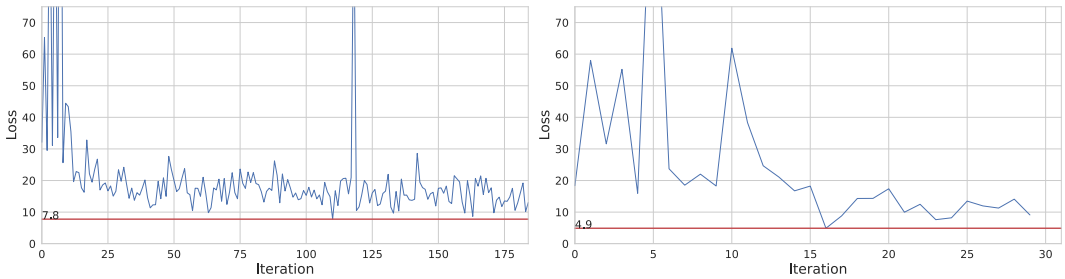


(a) (b)

Figure 6.19: Muon hits distribution in the first TT station. Obtained for (a) baseline muon shield, (b) step muon shield.

Such construction slightly changes the overall profile of the shield and introduces a lot of sharp corners, which can lead to significant degradation of the shield performance. To test this hypothesis, a large pre-production sample was run for the baseline shield configuration and the baseline configuration, constructed from the rectangular blocks (to which we refer as the step configuration). The distribution of muon hits in the first station of the Target Tracker for the baseline and the step option is shown in Figure 6.19. Although the distributions look similar, the total amount of muons crossing the TT increased by a factor of two by introducing the step modules. This is another hint that the optimum, found by BO, is not stable and is sensitive to small changes in the geometry of the shield. If it is indeed so, then the step modules act as a perturbation on the shield configuration, and it is possible to perform shield fine-tuning using L-GSO.

The L-GSO algorithm was applied to the baseline configuration with steps. Two par-



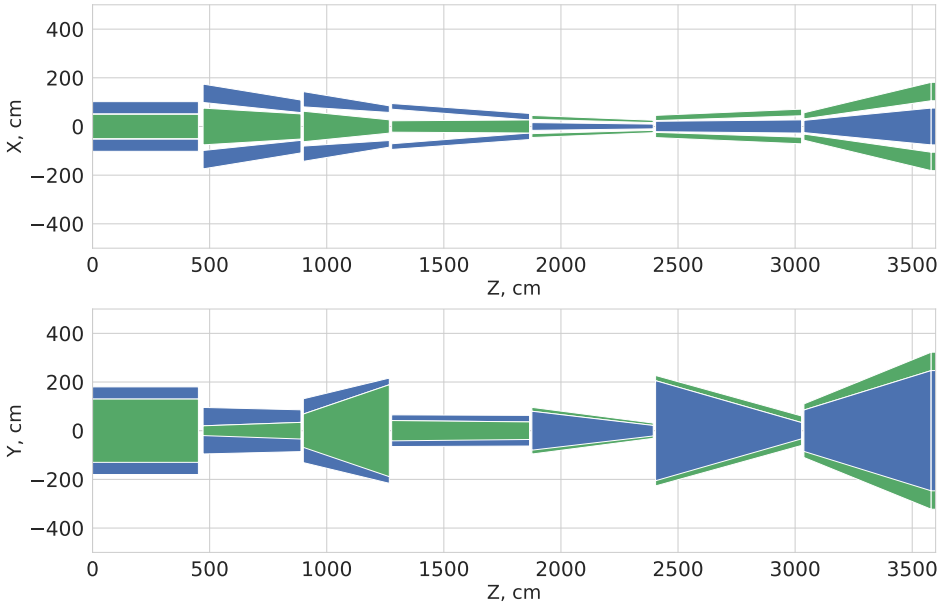
(a) (b)

Figure 6.20: Objective function value during optimisation of the shield made from rectangles using (a) resampled muons, (b) muGAN muons. Initial point is the BO optimum. Red line correspond to the minimal value attained.

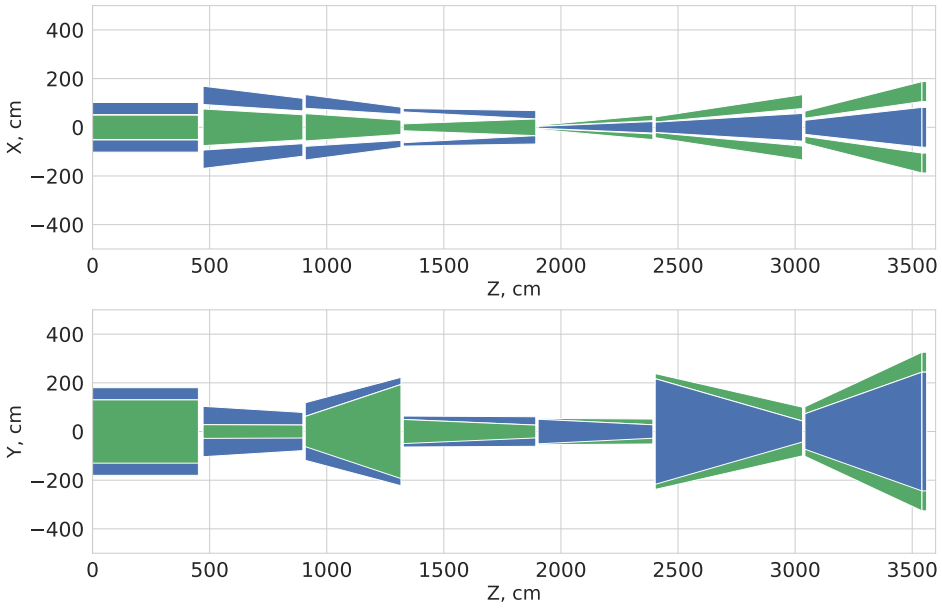
allel optimisations were performed: using the reduced muon sample from Figure 6.4a and the GAN generated sample from Figure 6.4b. The variation of the objective function during the optimisation is shown in Figure 6.20. The initial loss value is larger than in the original optimisation since, as mentioned, the step configuration of the shield has worse performance than its ideal counterpart. However, L-GSO successfully performed optimisation and reduced the loss value to the level of the "ideal" shield optimum or even lower. The optimisation using an extended muGAN sample proved to work well too, and improved the convergence of the algorithm. In Figure 6.20b the convergence of the algorithm occured already after the 15th iteration as opposed to 30th in Figure 6.20a.

Examining the resulted magnet shapes in Figure 6.21, we notice that the algorithm did not perform any drastic changes. For instance, the configuration in Figure 6.21b is very close to the baseline configuration yet features a much lower muon rate. On the contrary, Figure 6.21a features a fourth magnet that is completely different from any other designs, but has muon rate similar to configuration in Figure 6.21b. This suggests that there numerous optimal shield configurations that differ from each other by small adjustments.

The noticeable variation of the loss function for close design configurations suggests that the resulting muon rate is susceptible to small changes in the shield's geometry. If one thinks about the step shield design, it is only the edges of the shield that were affected by the steps, but even such a small variation of geometry increased the muon rate by a factor of two. The prototype test beam should help in understanding the magnitude of such effects by comparing the measured and simulated fluxes.



(a)



(b)

Figure 6.21: Muon shield XZ and YZ projections for the step optimisation for (a) "resampled" muon sample, (b) muGAN muon sample.

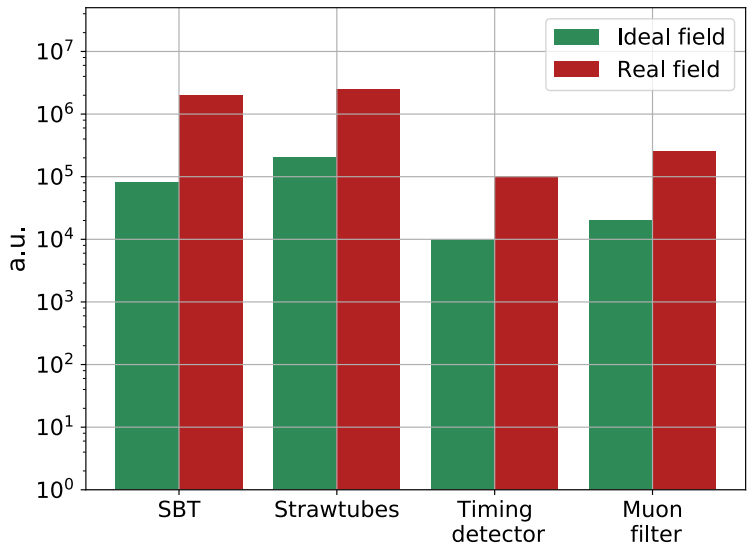


Figure 6.22: Number of muon hits per one spill in the experiment sub-systems. Blue is an ideal field map, red is a realistic map.

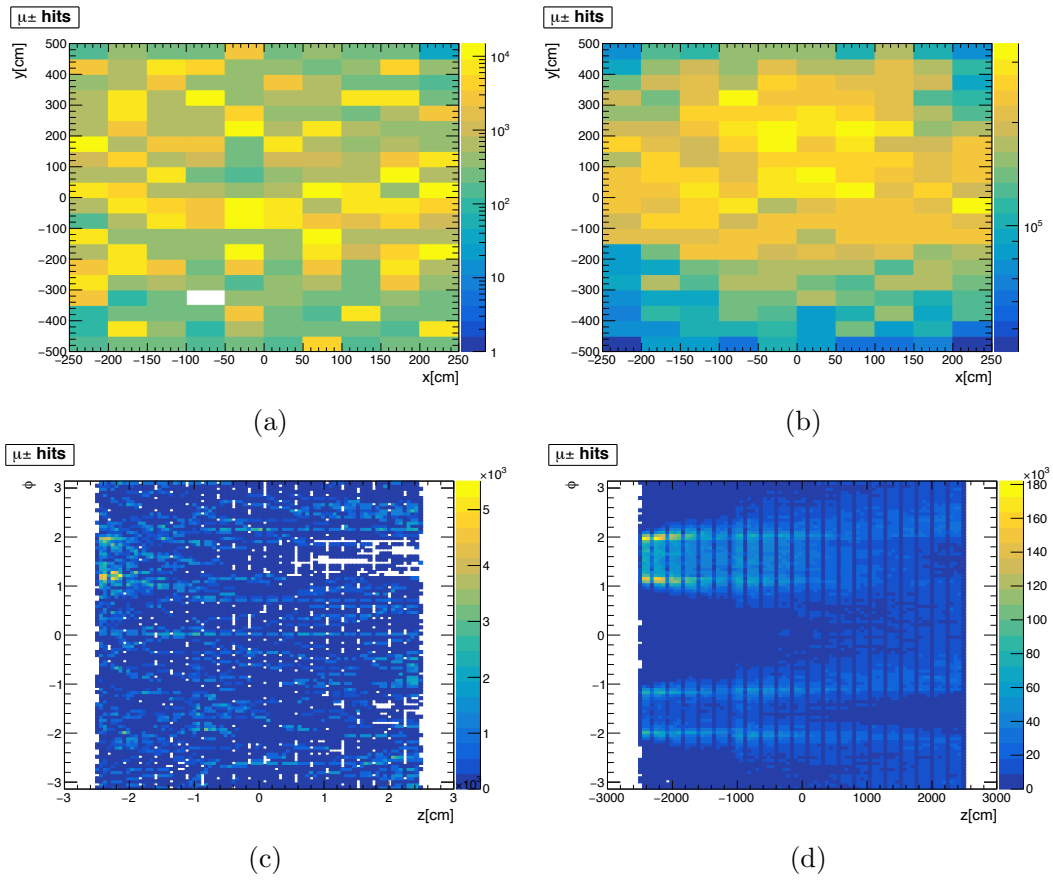


Figure 6.23: Muon hits distribution in (a) first TT station, (c) SBT for the ideal field and (b) first TT station, (d) SBT for the realistic field.

6.3.4 Ideal versus real magnetic fields

To bring the muon shield even closer to the realistic design a dedicated calculation was done to simulate the magnetic field inside the shield. The distribution of the field is presented in Figure 6.8. The field has a uniform value of $\sim 1.7\,\mathrm{T}$ in the "working" area of the shield: the central parts of the yoke and a significant decrease of the field to $\sim 1.5\,\mathrm{T}$ in the corners.

To check the effect of the real field, two simulations were done: one with the BO configuration and ideal field and the other with the BO configuration and realistic field. In Figure 6.22 the rates of muons reaching various sub-systems of the SHiP are presented. Unfortunately, an order of magnitude change in the muon rate is observed in all sub-detectors. The distributions of the hits in the TT and SBT for the ideal and real field are presented in Figure 6.23. Hotspots in the distributions in both detectors are seen.

There are two possible explanations why such a degradation in the performance might be observed. First of all, the realistic field is calculated on the 2.5 cm grid and then overlaid on the muon shield geometry during a simulation. This results in the discretisation of the field value inside the shield and on the boundaries. Having a discrete field inside the shield might not be that important, but having the discrete field on the boundaries would result in a "stray" field outside of the magnet as well as an incorrect field interpolation, which are not present in the ideal field scenario.

To test this hypothesis, the ideal field was discretised in the same manner as a realistic one with a 2.5 cm step. The simulations with the ideal and discretised ideal field were compared. It was observed that the rate had increased twice for the discretised field in comparison to the ideal. This suggests that the discretisation does indeed introduce side effects that make the current configuration non-optimal. However, this rate increase could not explain an order of magnitude change in the case of the realistic field.

Another option is that additional muons are caused by the weakened return field of the shield, as seen in Figure 6.8. A weaker return field means that low and mid energy muons do not experience enough bending and scatter back to the acceptance on the edges of a magnet. If it is indeed the case, the optimum found by the optimisation is sensitive to the non-uniformities in the magnetic field on the magnet's edges. It is an additional uncertainty of the simulation that requires verification at the test beam.

The simplest way to solve this problem would be to include the real magnetic field into the optimisation procedure, such that at each iteration the realistic field is generated

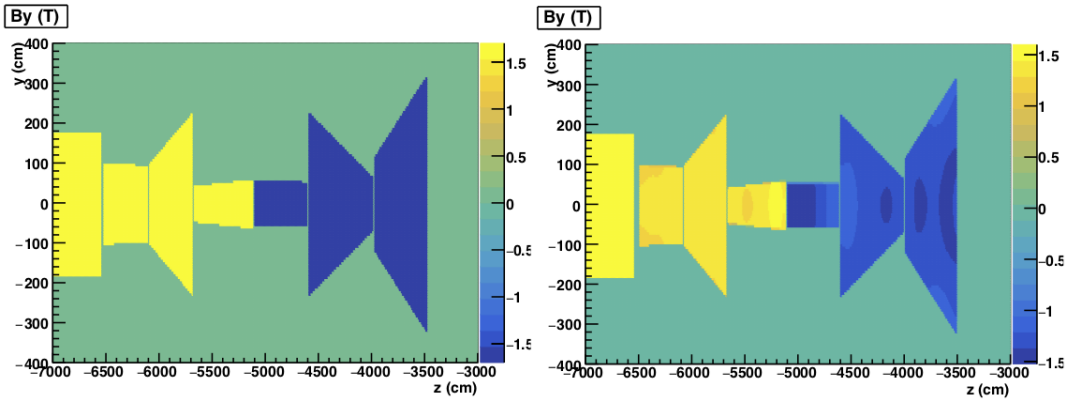
for the current geometry configuration. The problem here is that the computation of the field is not an automated procedure and requires a human expert to simulate the field. The second problem is that roughly one day is required to compute the field map for any specific geometry. At the scale of 80 parallel evaluations of different configurations (as discussed in Chapter 5), each of which takes 30 minutes, such an approach is infeasible.

A couple of solutions is possible to solve this problem. The best solution would be to devise a surrogate model for fast estimation of the magnetic field. This approach would allow approximating the field as close as possible to the real one, while maintaining low computation cost. For example, one can use generative networks or regression models to predict the field grid based on the magnet shape and the field strength. However, such an approach requires a dedicated long-term study to set up the problem, generate the data and train a model.

Another approach that is easy to implement is to add some noise to the ideal magnetic field. The added noise can be distributed uniformly in the whole volume of the shield or only on the edges, where the effect of the field degradation is the most pronounced. By introducing a new noise location and strength every optimisation iteration, one can regularise the optimisation process by preventing the algorithm from converging to an optimal but less stable solution.

An example of such a noise adding procedure is presented in Figure 6.24. Figure 6.24b is obtained by taking the field from Figure 6.24a and randomly adding 1000 Gaussian noise cores to it. As a result, patches of the non-uniform field are produced. The number of cores and the strength of the noise were calibrated so that the amount of muons passing the ideal noisy field is the same as the number of muons passing the realistic field configuration. This procedure was implemented in the optimisation process as it adds almost no additional computation time. This approach, however, was not able to steer the procedure in the correct direction. All optimisation iterations with noisy field map diverged. The reason for that is yet not completely understood, but it appears that the same noise strength affects the muon rate by orders of magnitude depending on the shield geometry. Thus, the noise parameters should be adjusted separately for each particular configuration of the shield, which is a computationally ineffective procedure. It was thus decided to switch to surrogate field modelling, which is currently under development.

The importance of the field effect and its influence on the muon rate has further proved the significance of the dedicated test beam. During the test beam, it will be possible to measure the impact of edges and welding on the magnetic field distribution



(a) (b)

Figure 6.24: Distribution of the magnetic field for the shield in YZ projection. (a) Ideal field, (b) Ideal field with noise.

and how those affect muon propagation and the final rate.

6.3.5 Summary

All the studies presented above conclude that the shield's design is highly fine-tuned so that small changes in the geometry or magnetic field affect the muon flux. Moreover, some critically important factors, such as magnetic field, are not properly modelled in the optimisation procedure. To better understand how the above aspects can be accounted for in the next iteration of shield optimisation, the muon shield prototype test beam is needed.

In addition, there might be another reason for the test beam. As discussed earlier, the shield is sometimes penetrated by wrong polarity muons. It is hard to predict under which circumstances a muon will become a wrong polarity particle and inevitably pass through the shield. To understand how many of such muons penetrate the shield and what are the characteristics of those, a dedicated muon classification was done.

6.4 Muon classification

Although there are only 1100 out of $O(10^7)$ muons from the sample that pass the shield, which corresponds to a rate of 50 kHz, it is important to understand the properties of those muons.

A visual examination of the passed muons provides a good initial separation of the particles. Overall, muons fall into four categories. The first corresponds to muons that exit the shield in the first or second magnet and do not have any hits in magnets 3-5. Instead, all such muons scatter back to the detector acceptance in the last sixth magnet. The mean energy of such muons is 30 GeV and does not exceed 70 GeV. There are 319 of particles of this category. Interestingly, such muons hit the last magnet with the correct field direction but are not swept outside due to the minimal distance they travel inside and the scattering effect.

The second category corresponds to low energy muons, with a mean (maximum) energy of 35 GeV (60 GeV), that exit the shield in the first two magnets, but then fly above or below magnets 3-4 and thus enter the shield on the wrong side. Because such particles appear on the wrong side, they are being swept back into the detector. There are 412 of such muons.

Muons with energies ranging from 60 to 125 GeV constitute the third category. Particles from this category usually exit the shield in the third magnet at a large angle and are not bent out by the following magnets. Muons traverse the magnets and again appear on the wrong side of the shield, where they are now swept in the detector. This category has 204 particles.

Finally, the fourth category contains high energy muons with momentum larger than 125 GeV. Such muons usually traverse the first three shield sections properly and then either end up in the wrong field or in the coil gaps. Thus, they are not bent enough to miss the detector. 184 particles fall in this category. However, this category can be removed via shield fine-tuning, as shown in the subsection 6.3.2 of this chapter.

The examples of trajectories for all categories are presented in Figure 6.25.

Apart from the difference in energy, it is interesting to further understand other properties of the categories. The initial hypothesis was that the penetrated muons might suffer from anomalous energy losses or large-angle scatterings while traversing the shield. These effects, in turn, may result in non-standard trajectories and, thus, shield penetration. To check this hypothesis, the following calculations have been done. First of all, the muon's energy loss ΔE in all traversed magnet sections is calculated. Then, the difference between the XZ-plane angle at the entrance and the exit of the section, θ_{diff} is calculated. This angle, however, can not be a measure of scattering since it reflects the effect of the magnetic field. To adjust for the particle bending, the change in the angle θ_{field} due to the magnetic field is calculated as follows:

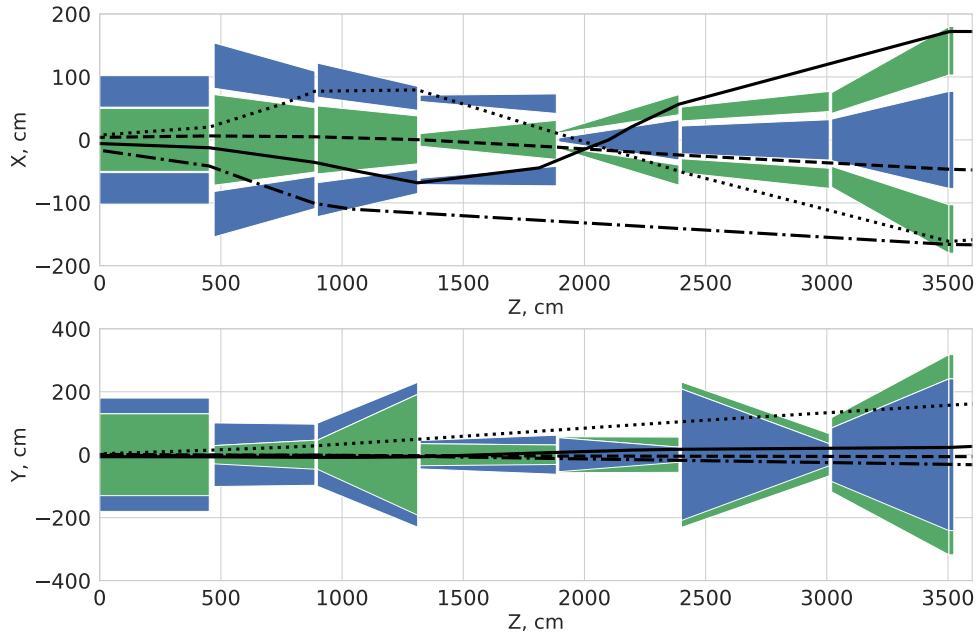


Figure 6.25: Examples of penetrating muons. Categories: dashed-dot:1, dotted:2, solid:3, dashed:4.

$$\theta_{field} = \frac{cBL}{E\beta^2}, \tag{6.5}$$

where c is the speed of light, B is the magnetic field strength in T, L is the distance traversed in the XZ-plane in m, E is particle energy in GeV and β is relative to the speed of light particle velocity. The absolute angle difference between the muon's entrance and exit direction due to the scattering can be then estimated as $\theta_{adjusted} = ||\theta_{diff}| - \theta_{field}|$. To understand if the scattering angle is large, it can be compared to a multiple scattering theory.

In multiple scattering, a particle traversing the medium is deflected by many small-angle scatters. Most deflections come from many consecutive scatterings on a nuclei described by the Rutherford scattering and often referred to as a multiple Coulomb scattering (MS). The multiple scattering is described by a Gaussian distribution for small angles. However, occasionally a single large-angle scattering could happen, which will result in a significant change of the particle trajectory. The Gaussian distribution for MS is characterised by zero mean and standard deviation σ_{MS} that is proportional to $\sigma_{MS} \sim 1/p\sqrt{L/X_0}$ [97], where p is the momentum of the particle, L is the distance travelled in the medium and X_0 is the radiation length of the material. Usually, it is considered that a Gaussian approximation works well for angles up to a few σ_{MS} and is accurate to 15% or better for $L/X_0 < 200$ [97].

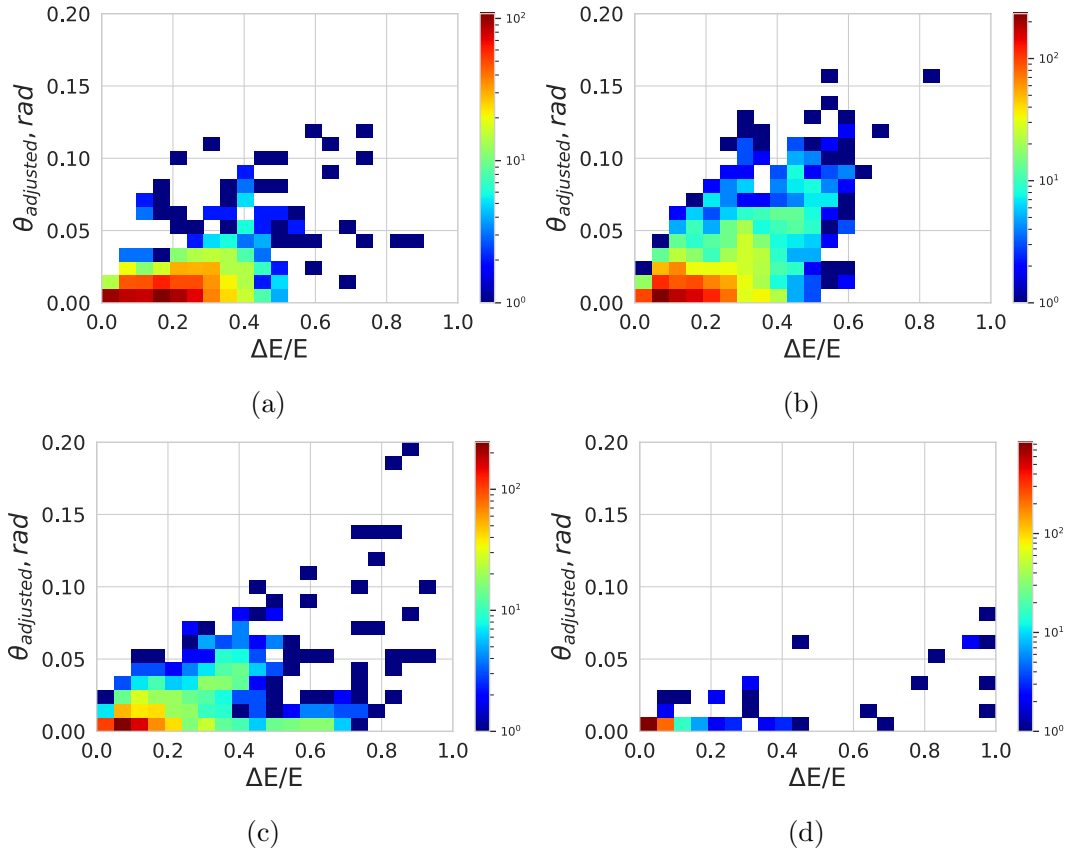


Figure 6.26: Histogram of correlation between scattering angle $\theta_{adjusted}$ and the relative muon energy loss. Histograms plotted separately for different muon categories (a) one, (b) two, (c) three, (d) four.

The mean distance L travelled by the muon inside one magnet section is around 300 cm and the radiation length of iron is 1.757 cm result in $L/X_0 \approx 170$. Thus, it is possible to apply multiple scattering theory in the first approximation. As a result, one would expect that $\theta_{adjusted}$ would correspond to an angle change induced by multiple or single scattering in the material.

Each muon, when flying through the shield, would traverse a couple of magnets. The relative energy loss $\Delta E/E$ and the angle $\theta_{adjusted}$ is computed for each magnet, traversed by the muon, resulting in $5-10$ data points per muon. The 2D histograms of $\Delta E/E - \theta_{adjusted}$ are presented in Figure 6.26.

Although all the computed quantities are simple approximations, there are two notable properties. First, muons in categories two and three have a correlation between large-angle scattering and large energy loss. Interestingly, these categories correspond to muons of the wrong polarity, as explained above. The second observation is that categories one to three have a significant amount of events with large-angle scattering.

Figure 6.26 shows that muons in categories 1-3 have a noticeable fraction of them with angles larger than 50 mrad. There are approximately 14 %, 59 %, 32 %, 3 % of muons with an angle larger than 50 mrad in categories 1-4 correspondingly. In total, this adds up to about 30 % of penetrated muons that have undergone scattering at an angle larger than 50 mrad. Thus, it is vital to understand how well large-angle scattering is modelled in GEANT4 to avoid surprises when the experiment becomes live.

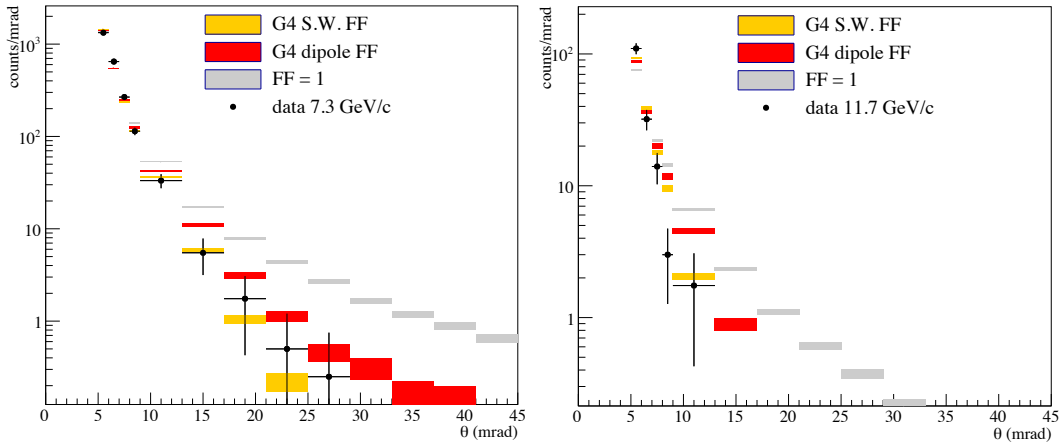
To answer this question the literature on the GEANT4 simulation of multiple scattering was studied [98, 99, 100]. It turns out that there are multiple models available in GEANT4 for multiple scattering description and that the result depends significantly on the step size. The step size describes how often GEANT4 recomputes the propagation of the particle, its energy and direction. A smaller step size provides more precise results but increases the usage of the CPU. The new combined single and multiple scattering model was introduced in [99]. This model automatically chooses which type of scattering to use and is currently the default option in GEANT4 [99].

However, examining the comparison of the different models, it is clear that all the models tend to agree for scattering angles smaller than $20 - 40$ mrad and show growing discrepancies for larger angles [99]. It is also important to note that the results provided in the paper are for low energy muons of 200 MeV. Therefore, the results are not applicable to the muon shield, where the minimum muon energy is $\sim 15 - 20$ GeV.

The latest results for GEANT4 comparison with the data comes from [100], where large-angle muon scattering in lead was compared to the simulation. Muons with energy up to 12 GeV were used in the comparison. Although having low statistics and still being below SHIP's energy range, this paper has the closest energies for such type of scattering experiments. At the moment of this work, no publications on the subject with higher energies were found.

First of all, the results of the comparison GEANT4 Monte Carlo simulation with theoretical models of multiple scattering show a diverging behaviour starting from a certain angle θ. Moreover, the value of θ decreases with muon energy, as seen in Figures 3-5 of the paper. This suggests that using simple multiple and single scattering theoretical models for muons with energies more than $O(10)$ GeV is incorrect.

Secondly, and most importantly, the GEANT4 simulation does not agree well with the experimental data. The distributions of the scattering angle from simulations and experimental data are presented in Figure 6.27. Figure 6.27a corresponds to 7.3 GeV



(a) (b)

Figure 6.27: Scattering angle distribution from GEANT4 simulations (coloured) and experimental data (black dots). For (a) 7.3 GeV muons, (b) 11.7 GeV muons.[100]

muons and Figure 6.27b to 11.7 GeV muons. As one can see, for 7.3 GeV muon the discrepancy starting to appear at $\theta \approx 15 - 20$ mrad and for 11.7 GeV muons at around $5 - 10$ mrad. Note, that the size of the data sample in the regions of interest is just a few events per bin. The coloured histograms correspond to various form-factor models of the nucleus used by GEANT4, and the black dots are the experimental data. The "FF=1" corresponds to the point-like nucleus with form-factor equal to one. "G4 dipole FF" is a dipole form-factor model that assumes an exponential decrease of the nuclear charge density and is a default GEANT4 option. This model might be inaccurate for the charge distribution of large nuclei [100]. The authors propose a "G4 S.W.FF" model, which adds a more realistic Saxon-Woods nuclear charge density and introduces form-factor models of a proton. The authors show that the inclusion of the proton form-factor significantly increases the probability of large-angle scattering [100]. Notably, none of the single form-factor models describes the data fully in Figure 6.27. Thus, the GEANT4 simulation results will depend on the region of muon momenta to which the simulator was tuned to. This might lead to an inaccurate prediction for the SHiP's energy range.

As stated above, 30% of muons underwent large-angle scattering above 50 mrad, which affected the muon trajectory. With this estimate being very conservative, there are no data available for such angles in Figure 6.27. This observation suggests that a significant fraction of penetrating muons might not be well-simulated by GEANT4. In turn, it may result in an inaccurate simulation of the muon shield and, as a consequence, under- or over-estimation of the flux. To ensure that the GEANT4 computations agree with the experiment, dedicated measurements of large-angle

scatterings should be done at the test beam.

6.5 Conclusion and future work

In this chapter the motivation behind the re-optimisation, the application of L-GSO and its application to technological challenges were discussed.

The formal problem statement and the definition of the loss function were stated. In addition to that, the pre-production samples used were discussed in detail. The L-GSO algorithm was successfully applied to the problem and showed that the optimum found by BO is indeed a local optimum, unstable to perturbations. L-GSO was able to fine-tune the BO configuration and achieve a better result for the simplified geometry, however, it showed a worse result in the case of the complete geometry. The main reason for such a discrepancy was shown to be an exact configuration provided by L-GSO, that did not account for muon scattering on the decay volume walls, since the latter was excluded in the simplified geometry.

The technological challenge of constructing the shield from the GO steel sheets was considered. It was shown that it is possible to optimise the shield constructed from steps to reduce the muon rate to that of the ideal shield.

Finally, simulation simplification of the magnetic field was considered. The simulation shows that the muon flux increased by order of magnitude in the case of the realistic magnetic field. This is due to the significant degradation of the field on the edges and corners of the magnet. An attempt to perform an optimisation with the noisy field was discussed.

Currently, there is an effort in reconsidering all sub-systems of the SHiP experiment. With all the optimisation techniques and results at hand, it is possible to perform new optimisation with all sub-systems included. Ideally, the new objective function should look as follows:

$$\mathcal{R} = \alpha\text{SND} + \beta\text{HS} + \gamma\text{DV} + \delta\text{Shield}, \tag{6.6}$$

where each term in the sum corresponds to the loss of the particular sub-detector. For example, it is now clear that the weight of the shield does not affect the cost. Thus, the shield term should include the length of the shield and any other shield characteristic that corresponds to its price, for example, the price of the sheets welding

per unit length. The HS term might be reconsidered to include any information used in the signal selection requirements. The decay volume term should include the rate of the EM background through the SBT. Finally, the requirement for the SND was recently devised: the density of the muon tracks should be less than $10^3/\text{mm}^2$. The author of the thesis has already done a preliminary estimation of the muon flux for the SND detector and proposed a loss term for the SND.

It is now clear that Bayesian optimisation or evolutionary algorithms provide much better exploration capabilities than L-GSO, but L-GSO is able to perform the exploitation step much better. Thus, an exciting idea is to combine those algorithms: for example, to use BO for probing new configurations and then further optimise them with L-GSO.

However, given the technological challenges described above, a dedicated test beam must be performed before progressing further with an optimisation. The test beam will allow the collaboration to estimate the cost of the prototype and thus the cost of the whole muon shield, the effect of the steps and magnetic field imperfections. Furthermore, the test beam will help in understanding the effect of large-angle scattering and its impact on the wrong polarity muon's trajectories. The resulting measurements can be compared to the GEANT4 simulations.

Chapter 7

Validation of the muon shield design

As was pointed in previous chapters, many findings that emerged from simulations and R&D studies of the muon shield show the necessity of a dedicated muon shield prototype construction and testing. Although successful, the muon shield optimisation relied heavily on simulation assumptions that do not hold in reality. These assumptions include the ideal frustum-like shape of the magnets, precise knowledge of the magnetic field inside the magnet's body and precise knowledge of muon track propagation inside the material. As a result, the design of the shield was fine-tuned and hard to construct in reality. Notably, because of the fine-tuning effect, the current designs of the shield, whether obtained by BO or L-GSO, are susceptible to the change of underlying assumptions.

As shown in the previous chapter, switching from ideal frustums to a realistic rectangular-based design of the magnets resulted in a twofold increase in the rate of muons. The substitution of the ideal field with the realistic simulated field increased the muon rate by an order of magnitude.

The effect of large energy-angle scattering contributes to 30% of the muons penetrating the shield, and its description in the GEANT4 simulator is not accurately modelled, as was also shown in Chapter 6. Thus, it is essential to understand the mechanism behind large-angle scattering and large-energy loss, because such events may result in "wrong" polarity muons, which can be an irreducible background for the experiment.

Finally, the technological processes of the shield production, welding and assembling have never been tested on a large scale prototype. Some technological aspects that require verification may have a significant impact on the magnetic field. For instance,

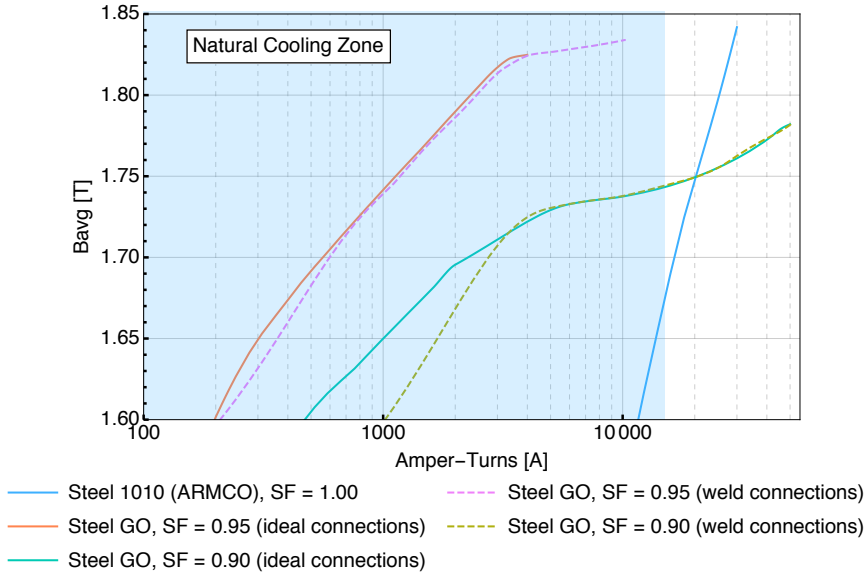


Figure 7.1: Magnetic field strength as a function of a coil current for various types of steel and stacking factor(SF) values. Courtesy of Fedor Ratnikov.

as discussed in Chapter 6, the shield is constructed from $0.3 - 0.5\,\text{mm}$ thick sheets that are welded together in units. It was measured that the field strength might drop by $\sim 20\%$ in the vicinity of the welding seams, thus adding even more distortions to an ideal field used in the optimisation. Another unaccounted effect is the production stacking factor (SF). The stacking factor is a measure of how tight steel sheets are interconnected. For example, a SF value of one represents an ideal connection and a SF value of 0.95 means that the material density has dropped by 5%. This will result in a proportional degradation of the magnetic field strength and a deviation of the actual shield performance from the simulation. As seen in Figure 7.1, the difference in the magnetic field strength for various SF can result in a $\sim 5\%$ field degradation.

With the above-mentioned concerns in mind, it was concluded that a dedicated muon shield test beam experiment is needed for the final iteration of the experiment optimisation. The test beam will verify the technological processes as well as allow for measurements of the magnetic field and studies of muon propagation in the prototype.

7.1 Prototype requirements

A dedicated muon flux experiment was conducted in 2018, where a SHiP target replica was used to produce the particles. As discussed in Chapter 6, the flux was

found to be in agreement with the SHiP simulation, with regards to both the total muon yield and momentum spectra [68]. Thus, to simplify the prototype testing, it was decided to use the muon beam available at the SPS. With such an approach, the experiment does not require a hadron stopper, a target and, potentially, a spectrometer.

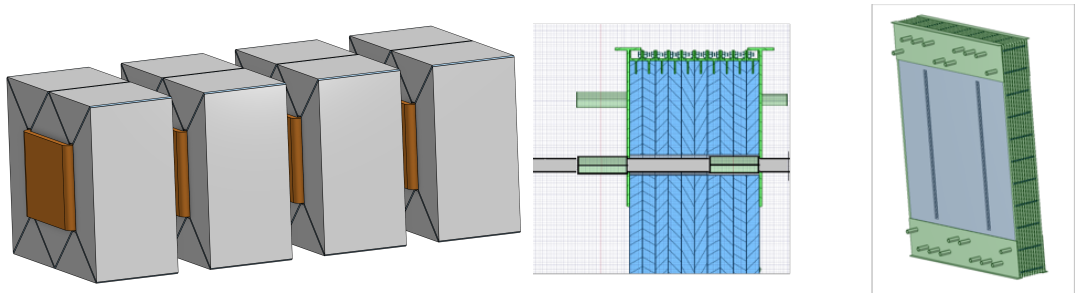
Presently, the CERN SPS North area H2, H4 and H8 beams are considered. Each of them can deliver $10^3 - 10^4$ muons per spill with a $5 - 7\%$ momentum spread, a $\sim 10\,\text{cm}$ spatial beam spread and an energy range of 10-250 GeV. Additionally, the North area has all the required infrastructure to quickly install the experimental equipment.

Given the muon beamline as a potential area for the experiment, a set of requirements should be devised for the prototype itself. The shield prototype should be, on one hand, large enough to put to test the outlined optimisation problems and, on the other, small enough to minimise the production and installation cost. The prototype should be large enough to maximise the amount of large-angle single scatters and large-energy losses and observe the muon trajectories within the shield. To detect such events, the shield should be instrumented with active detectors to track the particle trajectory inside the prototype. In addition, an EM calorimeter should be located downstream the prototype to detect emerging EM showers and measure their position and energy. This study will help to devise an optimal strategy of shielding SND from the EM background emerging from the muon shield. The prototype should have large enough cross-section, so that the field inside is not uniform, and resembles the real magnetic field distribution. Finally, the prototype should be large enough so that the manufacturing and assembling procedures mimic those of the real shield, thus allowing to test them.

The requirements above result in a prototype constructed from a couple of modules, interleaved with active detectors. The CAD drawing of the potential prototype is presented in Figure 7.2a. It consists of four identical modules of size $1 \times 1 \times 0.5\,\text{m}$ with the coil inserted in the middle, as it would be for the full-scale muon shield. Each module is constructed from 5 cm thick units of GO steel sheets welded together. The units are then bolted together to construct a module, as shown in Figure 7.2b.

7.2 Magnetic field measurements

As discussed above, the discrepancy between the ideal and the simulated fields raises concerns about the robustness of the shield performance. Switching from the ideal to



(a) (b)

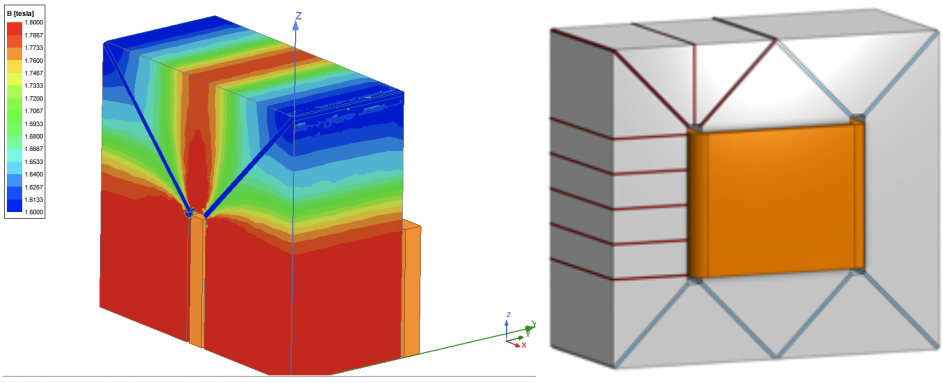
Figure 7.2: (a) The muon shield prototype engineering drawing. (b) Assembly procedure for a module [50].

a realistic field increases the flux of the penetrating muons by an order of magnitude. The magnetic field measurement will help in understanding the limitations of the simulated magnetic field and how those affect the muon flux. By comparing simulated and measured muon fluxes, it will be possible to modify the simulation to match the experimental results.

The simulated distribution of the magnetic field is presented in Figure 7.3a. The red colour corresponds to a field strength of 1.8 T, while the blue colour represents a field strength of 1.6 T. There is a noticeable degradation of the field at the edges and welding joints. Even though the field in the primary deflection area and the return field are close to the nominal value of 1.8 T, the integral effects of overall field non-uniformity can significantly affect the muon flux, as was shown previously. Thus, the primary deflection region, the return yoke, welding joints, coil gaps and edges of the magnet are required to be probed experimentally.

In addition, because the magnet is constructed from 5 cm units of sheets that are bolted together, the sheets may not fill the whole volume with steel. The stacking factor that describes this effect, and is discussed above, affects the magnetic field strength significantly, as seen in Figure 7.1. The SF is embedded in the magnetic field integral over the prototype and can not be simulated separately. The effect of SF is not shown in the field distribution above but can result in the incorrect estimation of the flux.

The simplest way to estimate the field is to measure it with measuring coils: a set of thin coils may be introduced into the prototype, as shown in Figure 7.3b. By placing the measuring coils in various locations of the prototype, such as welding joints, it is possible to precisely measure the field through the particular cross-section of the magnet. By successively measuring the field in the 5 cm units and then in 0.5 m modules and comparing it to the simulation, all the technological effects impacting



(a) (b)

Figure 7.3: (a) The simulated distribution of the magnetic field in the prototype module. (b) The measuring coils (red stripes) placed around a prototype module. The orange area is the main magnetic coil.

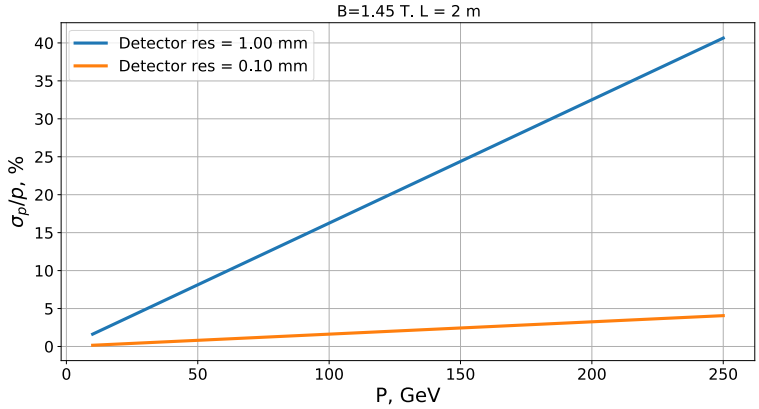


Figure 7.4: Momentum resolution as a function of momentum. Blue line corresponds to a detector resolution of 1 mm, yellow - 0.1 mm.

the field can be quantified. After that, the measured field map can be uploaded to the simulation, and a comparison of the simulated and measured muon flux performed. To emulate the scenario of the SHiP shield magnet construction it was decided to use four modules in the prototype. This way it is possible to test the shield assembling technology while instrumenting the prototype with active detectors.

Since the field measurements are done with the coils, precise knowledge of muon momentum is not required. As it was mentioned, the beam at the North area has a $5-7\%$ momentum spread, which should be sufficient for the comparison of simulation and experimental flux measurements. However, if more precise information about the muon momentum is required, the GOLIATH magnet [101] and a set of detectors might be used together as a spectrometer.

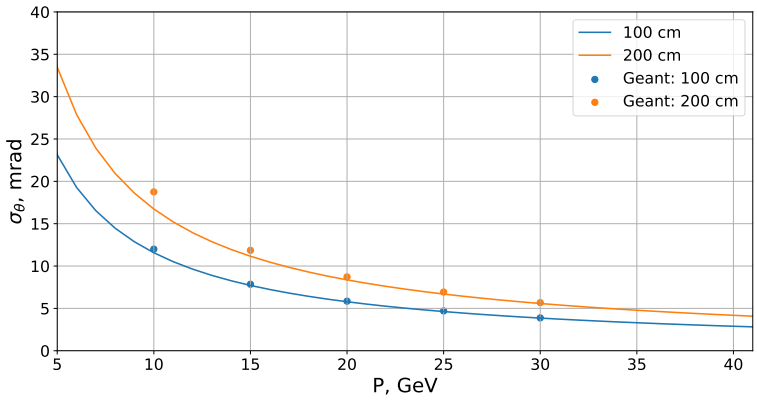


Figure 7.5: σ_{MS} as a function of muon momentum. Dots correspond to data from GEANT4, while lines correspond to the theoretical prediction. Two lengths of prototype are considered: orange - 100 cm, blue - 200 cm.

This magnet was already used during muon flux measurements in 2018 [68]. It provides a magnetic field of ~ 1.45 T over a 2 m distance. The momentum resolution as a function of the muon momentum is shown in Figure 7.4. It can be clearly seen that with the moderate detector resolution of > 1 mm, precise momentum identification can not be achieved. To measure the muon momentum more precisely than provided by the beam characteristics, it is necessary to install high-resolution detectors, such as SciFi tracking planes, used in the SHiP SND detector, that have spatial resolution of ~ 0.05 mm. The study of other available options for the spectrometer and the discussion of whether it is actually needed are currently ongoing.

7.3 Large-angle scattering

As presented in Figure 6.27, there is a notable discrepancy between data and various GEANT4 simulation models for scattering angles above $10 - 20$ mrad. Moreover, the experimental results have low statistics and are only available for muons with energy up to 12 GeV, which is well below SHiP energy range. As discussed in Chapter 6, approximately 30% of penetrating muons have large scattering angles. Categories corresponding to "wrong" polarity muons have an even larger fraction of such events of about 50%. The main concern about such types of events is that they may constitute an irreducible background for the experiment because they are not easily accounted for in the optimisation. Thus, it is crucial to verify that the simulation correctly estimates the number of such large-angle scatterings.

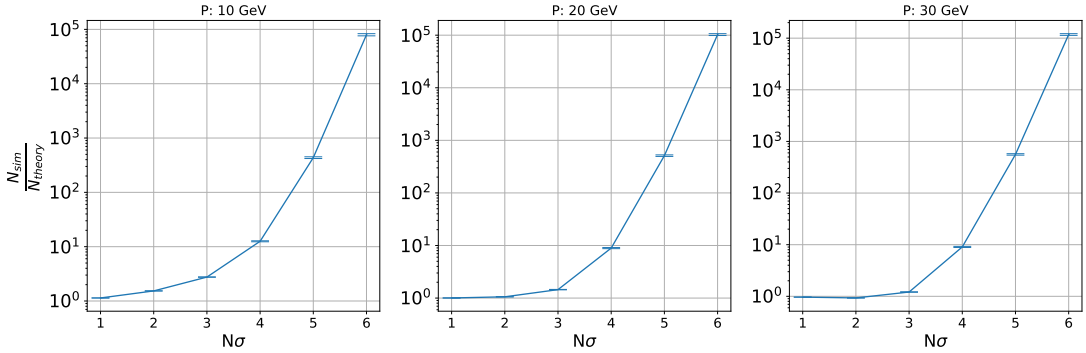
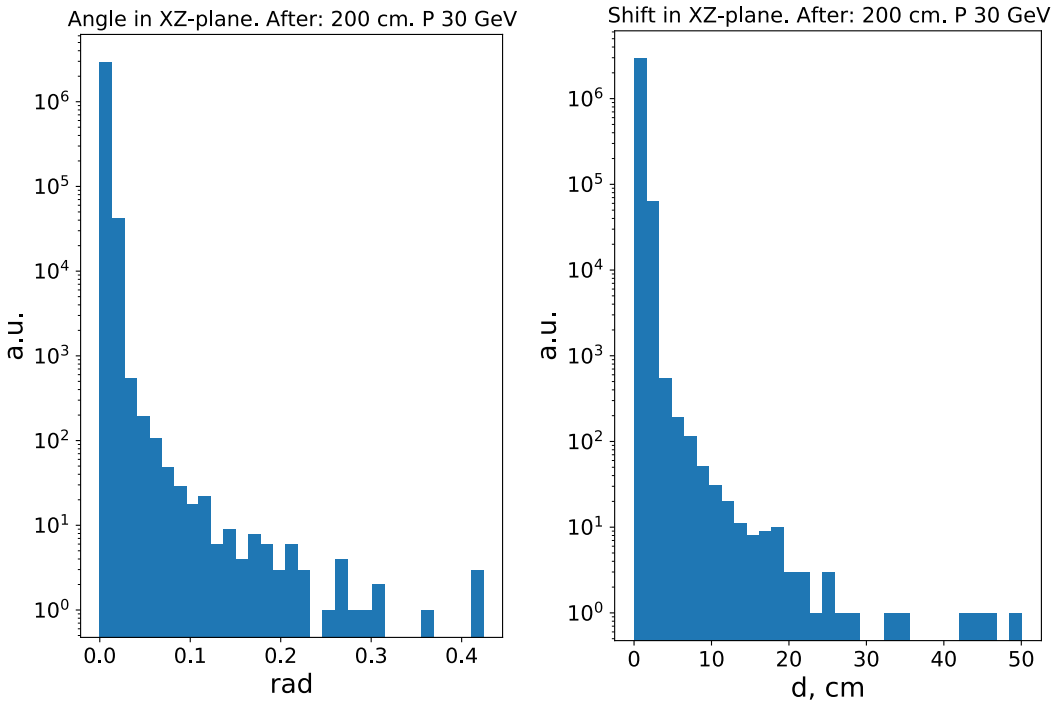


Figure 7.6: Ratio of muons in excess to theory as a function of the distribution quantile, for 10, 20, 30 GeV muons respectively.

The measurement of large-angle scattering, provided in the Figure 6.27, was done using a 1.44 mm copper plate and, thus, the angle distribution was mostly driven by single scattering rather than by multiple scattering (MS). In the case of the prototype, the probability of obtaining a large-angle scattering due to MS is non-negligible, as seen in Figure 7.5. This is because the effect of MS is proportional to $\sqrt{L}/P$ where L is the prototype length and P is the particle momentum. However, it is beneficial for the experiment to keep the long prototype, since the number of single scattering events is proportional to L. Thus, both single scattering and MS will populate the large-angle scattering region.

Since the comparison will be performed with the GEANT4 simulator, both effects will be automatically taken into account. However, to roughly estimate the number of single scattering events, one might still use simple MS theory [97]. Since the MS angle has a Gaussian distribution, one can compute the excess of events in the simulation with respect to a normal distribution. For example, for a thick layer of a material the number of muons with an angle larger than $3\sigma_{MS}$ should be just $\approx 0.3\%$ of the simulated events. The excess of events above this value would correspond to a single scattering. We define $N_{sim}(N) = \sum_i^{N_{total}} [\theta_i > N\sigma_{MS}]$ as the number of GEANT4 simulated events, where the scattering angle θ exceeds σ_{MS} N times. The $N_{theory}(N) = N_{total}(1 - \mathrm{erf}(\sigma_{MS}/\sqrt{2}))$ corresponds to the number of such events predicted by the theory. Thus, the quantity N_{sim}/N_{theory} corresponds to the excess of rare events in simulation in comparison to MS theory and is a first order measure of the number of single scattering events. In Figure 7.6, the ratio N_{sim}/N_{theory}, as a function of multiple scattering standard deviation σ_{MS} for 10, 20, and 30 GeV muons after travelling through 200 cm of iron is shown. It can be clearly seen that up to 3σ the excess of events in the simulation is small but rapidly increases starting from 4σ for all energies.



(a) (b)

Figure 7.7: (a) Distribution of (a) scattering angle, (b) scattering distance in X-axis for 30 GeV muons after passing through a 200 cm long prototype.

The $N_{total} = 3 \times 10^6$ events have been simulated for 10, 20 and 30 GeV muons. As an example, the distributions of the scattering angle and the coordinate shift in the XZ-plane for 30 GeV particles are shown in Figure 7.7. With a conservative estimation of the test beam intensity of 10^3 muons/spill and two spills per minute, the experiment is expected to have around 2000 events with angles larger than $4\sigma_{MS}$ (or ≈ 22 mrad) per day and ≈ 100 events per day for angles above 100 mrad, which corresponds to 20 σ_{MS}. Figure 7.7b also motivates the choice for the transverse size of the prototype's working area to be at least 20-30 cm not to lose any rare events.

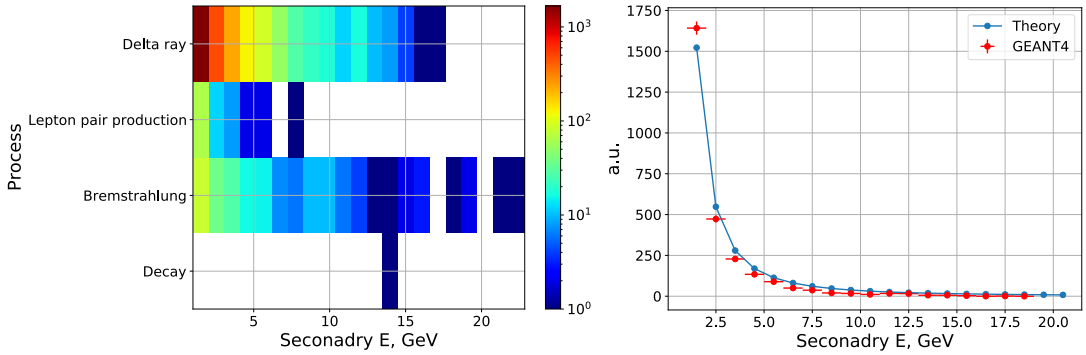
Thus, irrespective of the prototype technology and the magnetic field testing, the proposed test beam provides a standalone experiment for the first measurements of large-angle scattering for muons above 12 GeV. Given the discrepancy between previous experimental results [100] and the GEANT4 models, such a test beam experiment will allow the magnitude of the discrepancy to be understood further, so that the GEANT4 form factor models may be fine-tuned to the data in the energy range of $10 - 250$ GeV. With the conservatively estimated beam intensity, the experiment will be able to detect $O(100 - 1000)$ events per day.

7.4 Electromagnetic showers detection

Unlike large-angle scattering, large-energy loss for muons in dense materials has been well-studied [102]. However, large-energy loss is usually accompanied by an energetic electron or photon, which results in an electromagnetic (EM) shower. The EM shower can be detected in the calorimeter stations located in between prototype modules or after it. The reconstruction of the EM shower energy and position is an important step in understanding the electromagnetic background in the SND detector, testing a calorimeter prototype, and can serve as a supplementary measurement for a large-angle scattering study.

To estimate the number of detectable EM showers, 3×10^4 muons with energies of 25 GeV were simulated and EM showers with the energy of the initial particle above 1 GeV were counted. The distribution of a production process as a function of the energy of a secondary particle is shown in Figure 7.8a. There are only ≈ 3000 events with secondary particles of energies above 1 GeV, with ≈ 2500 of them being high energy delta-electrons. The distribution of delta-electron energy in comparison to simple theory prediction [97] is shown in Figure 7.8b. As expected, the distribution obtained from the GEANT4 simulator closely follows theoretical predictions.

Assuming that it is possible to reconstruct the EM shower energy if the shower



(a) (b)

Figure 7.8: (a) 2D distribution of EM production mechanism and secondary particle energy, (b) Comparison of number of delta electrons obtained from GEANT4 (red) and theory (blue).

maximum lies outside of the prototype module, there will be $\approx 15-150$ showers/spill with energy above 1 GeV, depending on the beam intensity. Thus, it will be possible to collect large statistics of muon large-energy loss events and the resulting EM showers.

It has to be noted, that a detailed simulation of the spectrometer and the particular type of detectors used should be done in the future. The simulation should include the realistic geometry of the spectrometer and the magnetic field (for example, such as GOLIATH [101]), and incorporate a reconstruction procedure for the detectors. Such a study is needed to pinpoint the required resolution of the detectors to accurately measure muon momentum and deviation in the shield prototype, as well as understand the requirements on the spectrometer. Finally, a future study must include a detailed simulation of EM showers and the calorimeter to understand the energy and position resolution of the experimental setup.

Chapter 8

Conclusion

To successfully exploit high-intensity proton beam and detect BSM particles, the SHiP experiment should operate at almost zero SM background level. At SHiP an initial muon flux of 10^{11} particles/spill should be reduced to a manageable level for the experiment to operate: at least five-six orders of magnitude. This poses a challenging optimisation problem of designing an optimal muon magnetic shield to effectively sweep out muons while minimising the experiment cost.

The L-GSO optimisation algorithm and its successful application to the SHiP BDF muon shield optimisation are the key outcomes of this thesis. The devised algorithm was specifically tailored to the needs of HEP optimisation tasks where it is computationally expensive to run a simulator. The full description of the GSO, its drawbacks, and how it may be applied in HEP tasks were presented. A modification of the algorithm, called L-GSO, was presented where the problem of scalability to high-dimensional problems was solved. It was shown that the algorithm outperforms other existing methods and that it works well for simple physics problems. In the toy setting, the algorithm was able to optimise one frustum section of the muon shield successfully.

Afterwards, the outcomes of the BDF magnetic shield re-optimisation were presented. It was shown that the current design of the shield is the fine-tuned local optimum and that it is possible to improve the current design. The L-GSO algorithm was successfully applied for the shield re-optimisation. However, it was discovered that it is essential to perform optimisation of the complete geometry of the experiment rather than its simplified version. Thus, it became clear that a new iteration of the optimisation must include the decay volume and the SND detector in the optimisation procedure. Furthermore, it was shown that small changes of the shield geometry,

such as the introduction of steps, can significantly affect the resulting muon flux. With the L-GSO method, it was possible to tune the modified geometry to bring the rates back to the baseline level. However, it was also discovered that it is crucial to use the real approximation of the magnetic field in the shield rather than an ideal default implementation in GEANT4.

In addition to the above, it was shown that large-angle and large-energy scattering of muons could affect their trajectories in the magnetic field. These low-medium energy muons can constitute up to 30% of the penetrating muons, but the description of such processes is not accurately modelled in the GEANT4 simulator. These effects may result in an inaccurate estimation of the muon flux in the detector acceptance. The above findings resulted in the conclusion that a dedicated muon shield prototype test beam experiment is needed to validate the above discrepancies.

Finally, a preliminary estimation of the prototype test beam requirements was done. It was shown that, if located at the CERN SPS muon beam, the test beam will allow all technological aspects of the shield construction to be tested and, measurements of the magnetic field and the muon yield to be performed and compared to the simulation. Ultimately, important standalone measurements of muon large-angle scattering and large-energy losses in the energy range of $10 - 250\,\text{GeV}$ can be performed. This will allow GEANT4 predictions to be compared with the new experimental data, and the simulator to be calibrated if needed.

A new global optimisation of the BDF facility will take place in the future. As stated in the SPS BDF Memorandum of Understanding, one of the key focuses will be the R&D development of the muon shield prototype and the related detector technologies followed by the re-optimisation of the shield. The simultaneous optimisation of the complete experiment geometry, including the muon shield, the SND detector and the vacuum vessel will take place afterwards. The joint optimisation will allow not only satisfy the ultimate constraint for the muon flux rate in the Hidden Sector detector but also ensure that other detectors, such as SND and SBT, are operating under acceptable muon and EM background rates. The combination of BO with the L-GSO algorithm, developed in this thesis, will form a new framework for optimisation during the R&D stage and after.

Bibliography

[1] The ATLAS collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020. URL https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[2] The CMS collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30 – 61, 2012. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2012.08.021. URL http://www.sciencedirect.com/science/article/pii/S0370269312008581.

[3] The DONUT collaboration. Observation of tau neutrino interactions. *Physics Letters B*, 504(3):218–224, Apr 2001. ISSN 0370-2693. doi: 10.1016/s0370-2693(01)00307-0. URL http://dx.doi.org/10.1016/S0370-2693(01)00307-0.

[4] R. Davis, D.S. Harmer and K.C. Hoffman. Search for Neutrinos from the Sun. *Phys. Rev. Lett.*, 20:1205–1209, May 1968. doi: 10.1103/PhysRevLett.20.1205. URL https://link.aps.org/doi/10.1103/PhysRevLett.20.1205.

[5] L. Canetti, Drewes M. and M. Shaposhnikov. Matter and antimatter in the universe. *New Journal of Physics*, 14(9):095012, sep 2012. doi: 10.1088/1367-2630/14/9/095012. URL https://doi.org/10.1088/1367-2630/14/9/095012.

[6] B. Gianfranco, D. Hooper and J. Silk. Particle dark matter: evidence, candidates and constraints. *Physics Reports*, 405(5):279–390, 2005. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2004.08.031. URL https://www.sciencedirect.com/science/article/pii/S0370157304003515.

[7] Planck satellite collaboration. Planck2013 results. i. overview of products and scientific results. *Astronomy & Astrophysics*, 571:A1, Oct 2014. ISSN 1432-0746. doi: 10.1051/0004-6361/201321529. URL http://dx.doi.org/10.1051/0004-6361/201321529.

[8] The SHiP collaboration. A facility to Search for Hidden Particles at the CERN SPS: the SHiP physics case. *Rept. Prog. Phys.*, 79(12):124201, 2016. doi: 10.1088/0034-4885/79/12/124201. URL https://arxiv.org/abs/1504.04855.

[9] R. Foot, H. Lew and R.R. Volkas. A model with fundamental improper spacetime symmetries. *Physics Letters B*, 272(1):67–70, 1991. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(91)91013-L. URL https://www.sciencedirect.com/science/article/pii/037026939191013L.

[10] B. Patt and F. Wilczek. Higgs-field portal into hidden sectors. 5 2006. URL https://arxiv.org/abs/hep-ph/0605188.

[11] B. Batell, M. Pospelov and A. Ritz. Exploring portals to a hidden sector through fixed targets. *Phys. Rev. D*, 80:095024, Nov 2009. doi: 10.1103/PhysRevD.80.095024. URL https://link.aps.org/doi/10.1103/PhysRevD.80.095024.

[12] J. Beacham, C. Burrage *et al.* Physics Beyond Colliders at CERN: Beyond the Standard Model Working Group Report. *J. Phys. G*, 47(1):010501, 2020. doi: 10.1088/1361-6471/ab4cd2. URL https://arxiv.org/abs/1901.09966.

[13] The SHiP collaboration. Proposal to Search for Heavy Neutral Leptons at the SPS, 2013. URL https://cds.cern.ch/record/1606085/plots?ln=en.

[14] The SHiP collaboration. A facility to Search for Hidden Particles (SHiP) at the CERN SPS. *arXiv preprint arXiv:1504.04956*, 2015. URL http://cds.cern.ch/record/2007512/files/SPSC-P-350.pdf.

[15] I. Goodfellow, M. Bengio and A. Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org.

[16] K. Arulkumaran, M. P. Deisenroth *et al.* Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017. URL https://ieeexplore.ieee.org/document/8103164.

[17] A. Tarca, C. Vincent *et al.* Machine Learning and Its Applications to Biology. *PLOS Computational Biology*, 3(6):1–11, 06 2007. doi: 10.1371/journal.pcbi.0030116. URL https://doi.org/10.1371/journal.pcbi.0030116.

[18] T. Cova and Alberto A. Pais. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. *Frontiers in Chemistry*, 7:809, 2019. ISSN 2296-2646. doi: 10.3389/fchem.2019.00809. URL https://www.frontiersin.org/article/10.3389/fchem.2019.00809.

[19] G. Carleo, I. Cirac *et al.* Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91:045002, Dec 2019. doi: 10.1103/RevModPhys.91.045002. URL https://link.aps.org/doi/10.1103/RevModPhys.91.045002.

[20] J. Cogan, M. Kagan *et al.* Jet-images: computer vision inspired techniques for jet tagging. *Journal of High Energy Physics*, 2015(2), Feb 2015. ISSN 1029-8479. doi: 10.1007/jhep02(2015)118. URL http://dx.doi.org/10.1007/JHEP02(2015)118.

[21] D. Belayneh, F. Carminati *et al.* Calorimetry with Deep Learning: Particle Simulation and Reconstruction for Collider Physics, 2019. URL https://arxiv.org/abs/1912.06794.

[22] S. Farrell, P. Calafiura *et al.* Novel deep learning methods for track reconstruction, 2018. URL https://arxiv.org/abs/1810.06111.

[23] R. D'Agnolo and A. Wulzer. Learning new physics from a machine. *Physical Review D*, 99(1), Jan 2019. ISSN 2470-0029. doi: 10.1103/physrevd.99.015014. URL http://dx.doi.org/10.1103/PhysRevD.99.015014.

[24] K. Cranmer, J. Brehmer and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, page 201912789, May 2020. ISSN 1091-6490. doi: 10.1073/pnas.1912789117. URL http://dx.doi.org/10.1073/pnas.1912789117.

[25] B. Shahriari, K. Swersky *et al.* Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015. URL https://ieeexplore.ieee.org/document/7352306.

[26] J. Snoek, H. Larochelle and R.P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012. URL https://papers.nips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

[27] N. Maheswaranathan, L. Metz *et al.* Guided evolutionary strategies: augmenting random search with surrogate gradients. In *ICML*, 2018. URL http://proceedings.mlr.press/v97/maheswaranathan19a.html.

[28] O. Alenkin, M. Hushchyn *et al.* Optimization of the SHiP Spectrometer Tracker geometry using the Bayesian Optimization with Gaussian Processes. *EPJ Web Conf.*, 214:06036, 2019. doi: 10.1051/epjconf/201921406036. URL https://doi.org/10.1051/epjconf/201921406036.

[29] L. Teodorescu. Evolutionary Computation in High Energy Physics. In *Inverted CERN School of Computing (iCSC2006)*, pages 35–46, 4 2008. URL https://arxiv.org/abs/0804.0369.

[30] K. Svanberg. The method of moving asymptotes—a new method for structural optimization. *International Journal for Numerical Methods in Engineering*, 24(2):359–373, 1987. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.1620240207.

[31] P. Y. Papalambros and D. J. Wilde. Principles of Optimal Design: Modeling and Computation. *The Aeronautical Journal (1968)*, 105(1050):458–459, 2001. doi: 10.1017/S0001924000012458.

[32] W. Banzhaf, P. Nordin *et al. Genetic Programming: An Introduction*. Morgan Kaufmann Series in Arti. Elsevier Science, 1998. ISBN 9781558605107. URL https://books.google.ch/books?id=1697qefFdtIC.

[33] R.J. Williams. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.*, 8(3-4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

[34] S. Shirobokov, V. Belavin *et al.* Black-box optimization with local generative surrogates. In *Advances in Neural Information Processing Systems*, volume 33, pages 14650–14662. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a878dbebc902328b41dbf02aa87abb58-Paper.pdf.

[35] E. Corbelli and P. Salucci. The extended rotation curve and the dark matter halo of M33. *Monthly Notices of the Royal Astronomical Society*, 311(2):441–447, 01 2000. ISSN 0035-8711. doi: 10.1046/j.1365-8711.2000.03075.x. URL https://doi.org/10.1046/j.1365-8711.2000.03075.x.

[36] N. Priyamvada, C. Urmila *et al.* Mapping substructure in the HST Frontier Fields cluster lenses and in cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 468(2):1962–1980, 02 2017. ISSN 0035-8711. doi: 10.1093/mnras/stw3385. URL https://doi.org/10.1093/mnras/stw3385.

[37] G. Hinshaw, J. L. Weiland *et al.* Five-year wilkinson microwave anisotropy probe observations: Data processing, sky maps, and basic results. *The Astrophysical Journal Supplement Series*, 180(2):225–245, feb 2009. doi: 10.1088/0067-0049/180/2/225. URL https://doi.org/10.1088%2F0067-0049%2F180%2F2%2F225.

[38] P. J. E. Peebles. Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations. *apjl*, 263:L1–L5, December 1982. doi: 10.1086/183911. URL https://inspirehep.net/literature/176406.

[39] A. Boyarsky, O. Ruchayskiy and M. Shaposhnikov. The Role of Sterile Neutrinos in Cosmology and Astrophysics. *Annual Review of Nuclear and Particle Science*, 59(1):191–214, Nov 2009. ISSN 1545-4134. doi: 10.1146/annurev.nucl.010909.083654. URL http://dx.doi.org/10.1146/annurev.nucl.010909.083654.

[40] The T2K collaboration. Constraint on the matter–antimatter symmetry-violating phase in neutrino oscillations. *Nature*, 580(7803):339–344, Apr 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2177-0. URL http://dx.doi.org/10.1038/s41586-020-2177-0.

[41] S. Davidson, E. Nardi and Y. Nir. Leptogenesis. *Physics Reports*, 466(4-5):105–177, Sep 2008. ISSN 0370-1573. doi: 10.1016/j.physrep.2008.06.002. URL http://dx.doi.org/10.1016/j.physrep.2008.06.002.

[42] M. C. Gonzalez-Garcia, M. Maltoni and T. Schwetz. Updated fit to three neutrino mixing: status of leptonic CP violation. *Journal of High Energy Physics*, 2014(11), Nov 2014. ISSN 1029-8479. doi: 10.1007/jhep11(2014)052. URL http://dx.doi.org/10.1007/JHEP11(2014)052.

[43] B. Holdom. Two U(1)'s and Epsilon Charge Shifts. *Phys. Lett. B*, 166:196–198, 1986. doi: 10.1016/0370-2693(86)91377-8. URL https://inspirehep.net/literature/219689.

[44] G. W. Bennett, B. Bousquet *et al.* Final report of the E821 muon anomalous magnetic moment measurement at BNL. *Physical Review D*, 73(7), Apr 2006. ISSN 1550-2368. doi: 10.1103/physrevd.73.072003. URL http://dx.doi.org/10.1103/PhysRevD.73.072003.

[45] G. Mohlabeng. Revisiting the dark photon explanation of the muon anomalous magnetic moment. *Phys. Rev. D*, 99(11):115001, 2019. doi: 10.1103/PhysRevD.99.115001. URL https://arxiv.org/abs/1902.05075.

[46] E. Aprile, J. Aalbers *et al.* XENON100 dark matter results from a combination of 477 live days. *Physical Review D*, 94(12), Dec 2016. ISSN 2470-0029. doi: 10.1103/physrevd.94.122001. URL http://dx.doi.org/10.1103/PhysRevD.94.122001.

[47] E. Aprile, J. Aalbers *et al.* Dark Matter Search Results from a One Ton-Year Exposure of XENON1T. *Physical Review Letters*, 121(11), Sep 2018. ISSN

1079-7114. doi: 10.1103/physrevlett.121.111302. URL http://dx.doi.org/10.1103/PhysRevLett.121.111302.

[48] The LZ Collaboration. LUX-ZEPLIN (LZ) Conceptual Design Report, 2015. URL https://arxiv.org/abs/1509.02910.

[49] P. Cushman, C. Galbiati *et al.* Working Group Report: WIMP Dark Matter Direct Detection. In *Community Summer Study 2013: Snowmass on the Mississippi*, 10 2013. URL https://www.osti.gov/biblio/1332188-working-group-report-wimp-dark-matter-direct-detection.

[50] The SHiP collaboration. SHiP Experiment - Progress Report. Technical Report CERN-SPSC-2019-010. SPSC-SR-248, CERN, Geneva, Jan 2019. URL https://cds.cern.ch/record/2654870.

[51] M. Fukugita and T. Yanagida. Baryogenesis Without Grand Unification. *Phys. Lett. B*, 174:45–47, 1986. doi: 10.1016/0370-2693(86)91126-3. URL https://lib-extopc.kek.jp/preprints/PDF/1986/8602/8602061.pdf.

[52] A. Takehiko, S. Blanchet and M. Shaposhnikov. The νMSM, dark matter and neutrino masses. *Physics Letters B*, 631(4):151 – 156, 2005. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2005.09.070. URL http://www.sciencedirect.com/science/article/pii/S0370269305013675.

[53] P. Minkowski. $\mu \to e\gamma$ at a rate of one out of 109 muon decays? *Physics Letters B*, 67(4):421 – 428, 1977. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(77)90435-X. URL http://www.sciencedirect.com/science/article/pii/037026937790435X.

[54] E. Kh. Akhmedov, V. A. Rubakov and A. Yu. Smirnov. Baryogenesis via Neutrino Oscillations. *Physical Review Letters*, 81(7):1359–1362, Aug 1998. ISSN 1079-7114. doi: 10.1103/physrevlett.81.1359. URL http://dx.doi.org/10.1103/PhysRevLett.81.1359.

[55] R. D. Peccei and Helen R. Quinn. CP Conservation in the Presence of Pseudoparticles. *Phys. Rev. Lett.*, 38:1440–1443, Jun 1977. doi: 10.1103/PhysRevLett.38.1440. URL https://link.aps.org/doi/10.1103/PhysRevLett.38.1440.

[56] M.J. Dolan, F. Kahlhoefer *et al.* A taste of dark matter: flavour constraints on pseudoscalar mediators. *Journal of High Energy Physics*, 2015(3), Mar 2015. ISSN 1029-8479. doi: 10.1007/jhep03(2015)171. URL http://dx.doi.org/10.1007/JHEP03(2015)171.

[57] A. Browman, J. DeWire *et al.* Decay Width of the Neutral π Meson. *Phys. Rev. Lett.*, 33:1400–1403, Dec 1974. doi: 10.1103/PhysRevLett.33.1400. URL https://link.aps.org/doi/10.1103/PhysRevLett.33.1400.

[58] P. deNiverville, C. Chen *et al.* Light dark matter in neutrino beams: Production modeling and scattering signatures at MiniBooNE, T2K, and SHiP. *Physical Review D*, 95(3), Feb 2017. ISSN 2470-0029. doi: 10.1103/physrevd.95.035006. URL http://dx.doi.org/10.1103/PhysRevD.95.035006.

[59] A.A. Aguilar-Arevalo, M. Backfish *et al.* Dark matter search in nucleon, pion, and electron channels from a proton beam dump with MiniBooNE. *Physical Review D*, 98(11), Dec 2018. ISSN 2470-0029. doi: 10.1103/physrevd.98.112004. URL http://dx.doi.org/10.1103/PhysRevD.98.112004.

[60] The BABAR collaboration. The BABAR detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 479(1):1 – 116, 2002. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(01)02012-5. URL http://www.sciencedirect.com/science/article/pii/S0168900201020125. Detectors for Asymmetric B-factories.

[61] J. Kahn. The Belle II Experiment. In *CERN-BINP Workshop for Young Scientists in e+e- Colliders*, 2017. doi: 10.23727/CERN-Proceedings-2017-001.45.

[62] T. Åkesson, A. Berlin *et al.* Light Dark Matter eXperiment (LDMX), 2018. URL https://arxiv.org/abs/1808.05219.

[63] S. N. Gninenko, D. V. Kirpichnikov *et al.* Combined search for light dark matter with electron and muon beams at NA64. *Phys. Lett. B*, 796:117–122, 2019. doi: 10.1016/j.physletb.2019.07.015. URL https://arxiv.org/abs/1903.07899.

[64] C. Lourenço and H.K. Wöhri. Heavy-flavour hadro-production from fixed-target to collider energies. *Physics Reports*, 433(3):127–180, Oct 2006. ISSN 0370-1573. doi: 10.1016/j.physrep.2006.05.005. URL http://dx.doi.org/10.1016/j.physrep.2006.05.005.

[65] F. Addesa, L. Burmistrov *et al.* The SE-CpFM Detector for the Crystal-Assisted Extraction at CERN-SPS. page WEPCF04. 4 p, 2018. doi: 10.18429/JACoW-IBIC2017-WEPCF04. URL https://cds.cern.ch/record/2661465.

[66] C.C. Ahdida, G. Arduini *et al.* *SPS Beam Dump Facility - Comprehensive Design Study*. CERN Yellow Reports: Monographs. Dec 2019. doi: 10.23731/CYRM-2020-002. URL https://cds.cern.ch/record/2703984.

[67] The SHiP collaboration. SHiP Experiment - Comprehensive Design Study report. Technical Report CERN-SPSC-2019-049. SPSC-SR-263, CERN, Geneva, Dec 2019. URL https://cds.cern.ch/record/2704147.

[68] The SHiP collaboration. Measurement of the muon flux from 400 GeV/c protons interacting in a thick molybdenum/tungsten target. *Eur. Phys. J. C*, 80(3):284. 11 p, 2020. doi: 10.1140/epjc/s10052-020-7788-y. URL http://cds.cern.ch/record/2715745.

[69] The SHiP collaboration. Sensitivity of the SHiP experiment to light dark matter. *JHEP*, 04:199, 2021. doi: 10.1007/JHEP04(2021)199. URL http://cds.cern.ch/record/2743145?ln=en.

[70] K. Albertsson, P. Altoe *et al.* Machine learning in high energy physics community white paper, 2019. URL https://arxiv.org/abs/1807.02876.

[71] The GEANT4 collaboration. GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.*, A506:250–303, 2003. doi: 10.1016/S0168-9002(03)01368-8. URL https://cds.cern.ch/record/602040?ln=en.

[72] M. Paganini, L. de Oliveira and B. Nachman. CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Phys. Rev. D*, 97(1):014021, 2018. doi: 10.1103/PhysRevD.97.014021. URL https://arxiv.org/abs/1712.10321.

[73] The SHiP collaboration. Fast simulation of muons produced at the SHiP experiment using Generative Adversarial Networks. *Journal of Instrumentation*, 14 (11):P11028, 2019. URL https://iopscience.iop.org/article/10.1088/1748-0221/14/11/P11028.

[74] P. Ilten, M. Williams and Y. Yang. Event generator tuning using Bayesian optimization. *JINST*, 12(04):P04028, 2017. doi: 10.1088/1748-0221/12/04/P04028. URL https://arxiv.org/abs/1610.08328.

[75] F. James and M. Roos. Minuit - a system for function minimization and analysis of the parameter errors and correlations. *Computer Physics Communications*, 10 (6):343–367, 1975. ISSN 0010-4655. doi: https://doi.org/10.1016/0010-4655(75)90039-9. URL https://www.sciencedirect.com/science/article/pii/0010465575900399.

[76] D.P. Kingma, T. Salimans and M. Welling. Variational Dropout and the Local Reparameterization Trick. In C. Cortes, N. Lawrence *et al.*, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/bc7316929fe1545bf0b98d114ee3ecb8-Paper.pdf.

[77] F. Hutter, H.H. Hoos and K. Leyton-Brown. Sequential Model-Based Optimization for General Algorithm Configuration. In Carlos A. Coello Coello, editor, *Learning and Intelligent Optimization*, pages 507–523, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. URL https://ml.informatik.uni-freiburg.de/papers/11-LION5-SMAC.pdf.

[78] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X. URL http://www.gaussianprocess.org/gpml/chapters/RW.pdf.

[79] O. Lantwin. Optimisation of the SHiP experimental design, 2019. URL https://cds.cern.ch/record/2693177.

[80] I. Goodfellow, J. Pouget-Abadie *et al.* Generative adversarial networks. *arXiv e-prints*, art. arXiv:1406.2661, Jun 2014. URL https://arxiv.org/abs/1406.2661.

[81] M. Arjovsky, S. Chintala and L. Bottou. Wasserstein GAN. *ArXiv*, abs/1701.07875, 2017. URL https://arxiv.org/abs/1701.07875.

[82] I. Gulrajani, F. Ahmed *et al.* Improved training of Wasserstein GANs. In *Advances in neural information processing systems*, pages 5767–5777, 2017. URL https://arxiv.org/abs/1704.00028.

[83] M.G. Bellemare, I. Danihelka *et al.* The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017. URL https://arxiv.org/abs/1705.10743.

[84] T. Sjöstrand, S. Mrenna and P. Skands. A brief introduction to PYTHIA 8.1. *Computer Physics Communications*, 178(11):852–867, 2008. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2008.01.036. URL https://www.sciencedirect.com/science/article/pii/S0010465508000441.

[85] A.G. Baydin, L. Shao *et al.* Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model. In H. Wallach, H. Larochelle *et al.*,

editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/6d19c113404cee55b4036fce1a37c058-Paper.pdf.

[86] E. Bingham, J.P. Chen *et al.* Pyro: Deep Universal Probabilistic Programming. *CoRR*, abs/1810.09538, 2018. URL http://arxiv.org/abs/1810.09538.

[87] R. Iman, J. Davenport and D. Zeigler. *Latin hypercube sampling (program user's guide). [LHC, in FORTRAN]*. 01 1980. ISBN SAND-79-1473 United StatesThu Feb 07 00:00:56 EST 2008Dep. NTIS, PC A05/MF A01.SNL; ERA-05-017293; EDB-80-043435English.

[88] C. Oh, E. Gavves and M. Welling. Bock : Bayesian optimization with cylindrical kernels. In *ICML*, 2018. URL https://arxiv.org/abs/1806.01619.

[89] W. Grathwohl, D. Choi *et al.* Backpropagation through the void: Optimizing control variates for black-box gradient estimation, 2018. URL https://arxiv.org/abs/1711.00123.

[90] N. Ruiz, S. Schulter and M. Chandraker. Learning to simulate, 2019. URL https://arxiv.org/abs/1810.02513.

[91] J. Momin and Y. Xin-She. A literature survey of benchmark functions for global optimization problems. *Int. J. of Mathematical Modelling and Numerical Optimisation*, 4, 08 2013. doi: 10.1504/IJMMNO.2013.055204. URL https://arxiv.org/abs/1308.4008.

[92] D. Harrison and D.L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1): 81–102, 1978. ISSN 0095-0696. doi: https://doi.org/10.1016/0095-0696(78)90006-2. URL https://www.sciencedirect.com/science/article/pii/0095069678900062.

[93] P. Gorbunov. On the field shape of the SHiP Active Muon Shield. *Memo PG-SHiP-05-20.04.2017*, 2017.

[94] T. Ruf and E. Van Herwijnen. Description of the SHiP FixedTargetGenerator. Aug 2017. URL https://cds.cern.ch/record/2280572.

[95] The SHiP collaboration. The active muon shield in the SHiP experiment. *Journal of Instrumentation*, 12(05):P05011–P05011, may 2017. doi: 10.1088/1748-0221/12/05/p05011. URL https://doi.org/10.1088/1748-0221/12/05/p05011.

[96] A. Baranov, E. Burnaev *et al.* Optimising the Active Muon Shield for the SHiP Experiment at CERN. *Journal of Physics: Conference Series*, 934:012050, dec 2017. doi: 10.1088/1742-6596/934/1/012050. URL https://doi.org/10.1088%2F1742-6596%2F934%2F1%2F012050.

[97] D.E. Groom and S.R. Klein. Passage of particles through matter. *Eur. Phys. J*, 15:163–173, 2000. URL https://doi.org/10.1007/BF02683419.

[98] L. Urban. Multiple scattering model in Geant4. *https://cdsweb.cern.ch/*, 2002. URL https://cdsweb.cern.ch/record/592633/files/open-2002-070.pdf.

[99] V.N. Ivanchenko, O. Kadri *et al.* Geant4 models for simulation of multiple scattering. *Journal of Physics: Conference Series*, 219(3):032045, apr 2010. doi: 10.1088/1742-6596/219/3/032045. URL https://doi.org/10.1088/1742-6596/219/3/032045.

[100] A. Longhin, A. Paoloni and F. Pupilli. Large-angle scattering of multi-GeV muons on thin Lead targets. *IEEE Trans. Nucl. Sci.*, 62:2216–2225. 10 p, Jun 2015. doi: 10.1109/TNS.2015.2473674. URL http://cds.cern.ch/record/2029819.

[101] M. Rosenthal, N. Charitonidis *et al.* Magnetic Field Measurements of the GOLIATH Magnet in EHN1. Mar 2018. URL https://cds.cern.ch/record/2310483.

[102] D.E. Groom, N.V. Mokhov and S.I. Striganov. Muon stopping power and range tables 10-MeV to 100-TeV. *Atom. Data Nucl. Data Tabl.*, 78:183–356, 2001. doi: 10.1006/adnd.2001.0861. URL https://pdg.lbl.gov/2020/AtomicNuclearProperties/adndt.pdf.

List of Tables

List of Figures