UNIVERSITÀ DEGLI STUDI DI PADOVA

PhD School of Physics - XXXVII series
École Doctorale de Sciences Fondamentales - Particules, Interactions, Univers

Towards End-to-end optimization of experimental design with automatic differentiation

Thesis written in cotutelle between Università degli Studi di Padova and Université Clermont-Auvergne

Supervisors
Prof. Julien Donini
Prof. Tommaso Dorigo
Referees
Dr. Sascha Caron
Prof. Michele Gallinaro
Examiners
Dr. David Rousseau
Dr. Sofia Vallecorsa

Ph. D. Student:
Federico Nardi

Moi, je veux mourir sur scène

Abstract

In high-energy physics, experiment design is a foundational step, as decisions made early on can influence — and limit — the direction of scientific inquiry for decades to come. Our research, embedded in the MODE Collaboration, aims to generalize Machine Learning tools for creating a differentiable pipeline capable of suggesting optimal configurations for the Muon Collider Electromagnetic Calorimeter geometry. After introducing the state-of-the art of automatic differentiation applications in experimental setups, and introducing the optimization of the SWGO telescope array as introductory example, in this work we outline the structure of our pipeline, emphasizing the methods employed to ensure full code differentiability. Our primary focus lies in maximizing the reconstruction efficiency of photons amidst Beam-Induced background from muon decays. The approach relies on three core blocks: (I) Signal Event Generator: Responsible for generating signal events; (II) Background Generator: Focused on simulating background events; (III) Reconstruction Algorithm: Adapting the DeepJetCore Object Condensation framework. The thesis includes a showcase of performance tests for each core block, shedding light on their efficacy. Additionally, we provide a proof of concept derived from a first implementation of the developed modules into a full pipeline.

Contents

List of Figures

Chapter 1

Introduction

The strikingly rapid progression of fundamental physics over the past century has driven the development of increasingly sophisticated detectors, with performance requirements rising in parallel to the scale and ambition of modern experiments. In the last decade, one of the most transformative advances has been the emergence of machine learning as a powerful and reliable tool—not only for data analysis and reconstruction, but also for guiding the design of experimental apparatuses themselves. The interconnected nature of detector subsystems makes a compelling case for global rather than component-wise optimization: by adopting a holistic view, one can define utility functions that more faithfully reflect the ultimate scientific goals of the experiment, instead of relying on narrow, task-specific figures of merit.

At the heart of this paradigm lies differentiable programming—the application of automatic differentiation techniques to compute gradients throughout complex software pipelines, enabling principled optimization via gradient descent. Yet, a key bottleneck in this approach is the inherently stochastic nature of data generation. Many physical processes involved in particle detection, particularly those governed by quantum interactions, are not easily expressed in a differentiable form. This necessitates the development of surrogate models or approximate differentiable representations to unlock the full potential of gradient-based methods.

The present thesis is situated within the context of the MODE collaboration (Machine-learning Optimized Design of Experiments) [1], which brings together physicists and computer scientists to explore precisely this class of problems. The collaboration aims to establish a new methodology for experimental design—one that tightly couples simulation, reconstruction, and optimization within a unified computational framework.

1.1 Automatic Differentiation in experimental physics

The design of a measuring instrument—spanning its layout, material choices, and data extraction procedures—constitutes a high-dimensional optimization problem. The space of parameters is vast and often shaped by implicit, non-linear correlations, which renders the task loosely constrained and inherently complex. In favorable cases, tractable approximations are possible: if a parametric model of the system can be defined, one may construct a likelihood function $\mathcal{L} = p(x|\theta)$ for simulated observations $\mathbf{x}$, and identify optimal configurations via minimization of the negative log-likelihood, $-\log \mathcal{L}$, with respect to the model parameters θ. This approach, however, breaks down in scenarios where the instrument's response is governed by quantum-level interactions—such as radiation-matter processes—where no closed-form expression for $p(x|\theta)$ exists. In such cases, one must rely on the generative model of the detector, accessible only via forward simulation. The resulting inference task, where only synthetic data can be generated without an explicit likelihood, falls into the category of likelihood-free or simulation-based inference [2].

Despite the intrinsic intractability of most detector design problems in particle physics, the past eighty years have seen the successful conception and deployment of increasingly sophisticated instruments. This progress was sustained by a pragmatic and iterative strategy: while consistently incorporating technological advances in materials and electronics, designers relied heavily on paradigms validated by accumulated experimental experience. In collider experiments, for instance, a canonical design principle is the early measurement of charged particle momentum via magnetic deflection in low-density media, followed by energy measurement through the development of electromagnetic and hadronic showers in dense absorbers. Redundancy across detection layers—crucial for cross-calibration and fault tolerance—has long been a cornerstone of detector architecture. Likewise, symmetric layouts, such as regularly interleaved active and passive layers in calorimeters, have become the default. While these heuristics are grounded in historical successes, their relevance to future detector development—particularly in light of emerging optimization techniques—deserves critical re-evaluation.

The rapid evolution of computer science over the past two decades—particularly the advent of deep neural networks and the maturation of differentiable programming frameworks—has opened the door to a radical rethinking of detector design. These tools offer the unprecedented possibility to explore and optimize complex design spaces far beyond the reach of manual or heuristic approaches. The challenges we face typically involve the simultaneous tuning of hundreds or thousands of parameters: from the spatial arrangement and geometry of materials to the specification of detector layers,

their physical performance, and associated costs. Exhaustively probing such a high-dimensional landscape is beyond human capability. To chart a viable path through this complexity, we must now turn to differentiable programming as a means to unlock principled, scalable optimization.

It is important to recognize that the dimensionality of the design space itself has expanded alongside technological progress. Additive manufacturing techniques now allow the 3D printing of scintillating detectors [3], while advanced fabrication methods enable the integration of complex elements such as thin, AC-coupled resistive silicon sensors [4]. Fully capitalizing on these capabilities requires more than domain expertise—it demands the ability to continuously scan and optimize over the geometric configurations of the detectors we seek to build. This is precisely the role of differentiable programming pipelines: to turn design into a navigable, learnable landscape.

A further motivation for rethinking detector design in light of recent computational advances lies in the evolution of our pattern recognition and inference strategies. The complexity of modern experiments continues to rise—whether due to the approach of the High-Luminosity phase of the LHC or the scale-up of detection volumes in cosmic ray and neutrino observatories. At the HL-LHC, ATLAS and CMS will soon be tasked with reconstructing events occurring amidst $O(200)$ simultaneous proton-proton interactions per bunch crossing. In such high-pileup environments, traditional reconstruction techniques for charged tracks suffer severe degradation due to the explosive growth of combinatorial ambiguities. Deep learning methods —such as those explored in Refs. [5]–[25]— offer a promising path forward. Yet this raises a critical question: are the detectors we are building today truly optimal for the AI-driven reconstruction techniques of tomorrow? A mismatch between design choices and downstream inference capabilities could severely limit the effectiveness of both.

This concern becomes even more pressing when looking ahead to future facilities such as the proposed Future Circular Collider (FCC)[26], where center-of-mass energies will significantly exceed those of current machines. Given the steep and ongoing progress in artificial intelligence techniques[27], it is increasingly clear that detector design must not be divorced from the capabilities of the reconstruction algorithms that will ultimately interpret the data. To ensure maximal utility, the optimization of future detectors must integrate, from the outset, a model of the inference procedures available at operation time—however speculative their precise form may be today.

These considerations motivate a broad research agenda aimed at equipping both ourselves and the wider community with the tools and insights necessary to integrate all facets of the detector design problem into a unified optimization framework. This in-

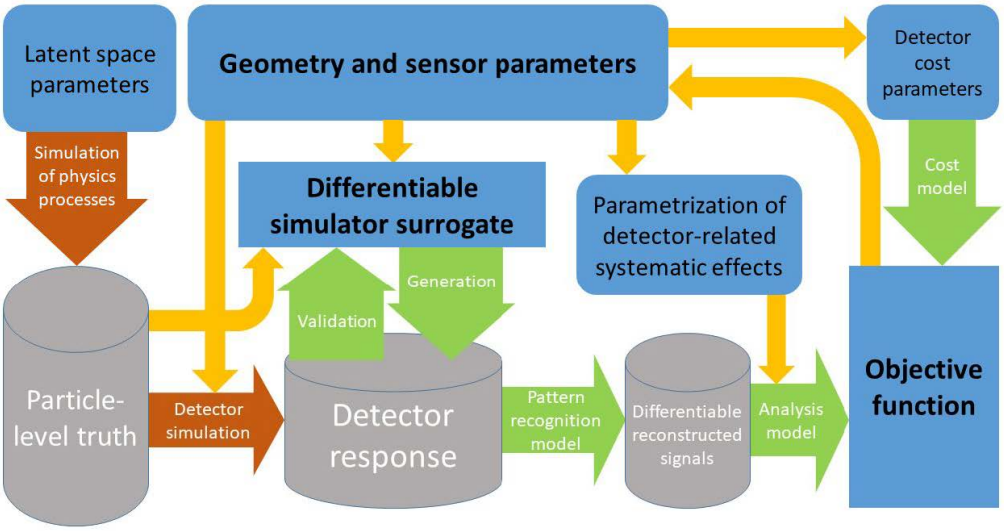


Figure 1.1: Schematic representation of the end-to-end optimization workflow for a generic detector. Adapted from Ref. [28].

cludes the modeling of quantum stochastic processes, the geometric and material layout of the detector, the performance characteristics of its components, the structure of pattern recognition and inference pipelines, and the interaction between design choices and both systematic uncertainties and resource constraints. As outlined schematically in Figure 1.1, these elements are deeply interconnected. Simulated data—depicted as containers labeled Particle-level truth and Detector response—are used to train and validate a differentiable surrogate model of the underlying physical processes. This surrogate, in turn, interfaces with modules responsible for pattern recognition, inference extraction, cost evaluation, and the definition of a global loss function. Crucially, all these components are expressed as functions of the detector's geometric and structural parameters. The loss gradients can then be propagated backward through the entire system, enabling joint optimization of design and reconstruction. Our central hypothesis is that implementing the full pipeline within a differentiable programming framework—thus enabling the computation of gradients with respect to any design parameter—will be essential for systematically exploring the vast configuration space and for uncovering novel, high-performing solutions.

Tackling the full design optimization of a detector on the scale of ATLAS or CMS is, at present, an overwhelmingly ambitious goal. The sheer complexity of such systems, coupled with current limitations in available resources and expertise, makes a direct approach unfeasible. Instead, a pragmatic strategy is required—one that builds capacity incrementally by focusing on a series of well-defined, tractable problems. In

this work, we outline a sequence of such tasks, each of intrinsic interest, whose resolution within the proposed framework serves a dual purpose: advancing specific design studies while contributing to a broader foundation of reusable methods and software tools.

Although the surrogate models used to describe the interaction of particles with detector components tend to be highly specific to the device under study—limiting their direct reusability—there are other parts of the optimization pipeline that offer broader applicability. In particular, recent deep learning–based reconstruction algorithms exhibit a remarkable degree of device independence [29]. Similarly, models describing component costs, the relationship between geometry and systematic uncertainties, and inference extraction strategies are amenable to reuse across detector concepts. By modularizing the design problem in this way, we aim to assemble a toolbox that can eventually scale to the full complexity of next-generation experiments.

At the heart of any optimization process lies the definition of an objective function—an explicit mathematical expression of the goals we seek to achieve through the design. For large-scale scientific instruments, this task may appear daunting. Detectors often serve multiple purposes, enabling a wide range of physics analyses with differing scientific priorities and evaluation criteria. Nonetheless, we contend that attempting to quantify these goals—however approximate or subjective—can be both feasible and instructive. A well-motivated objective function, or even a family of such functions reflecting diverse use cases, provides the structure necessary to navigate the design space productively. This challenge bears a meaningful analogy to the definition of a trigger menu in collider experiments: a carefully balanced list of event selection strategies, bandwidth allocations, and prescale factors that encode complex scientific tradeoffs. Similarly, the aim of our optimization study is not to deliver a single definitive answer, but to illuminate regions of the design space that offer advantageous tradeoffs—robust configurations that perform well across multiple metrics and constraints. In this sense, the optimization process becomes a tool for informed decision-making, rather than an oracle for final design.

1.2 Methods and tools in Automatic Differentiation

Over the past several decades, scientists and engineers have consistently harnessed the growth of computational power to refine and extend the reach of numerical simulations. In many technical domains—particularly those where traditional theoretical methods fall short—simulations have become indispensable tools for answering the

question: what happens if we implement a given design choice? The increasingly close agreement between simulated and experimental results has, in turn, opened the door to a more ambitious question: which design leads to the optimal outcome? In this context, simulations are no longer just predictive tools but active components in the optimization process itself.

Among the technical disciplines that have most effectively embraced simulation-based optimization, computational fluid dynamics (CFD) stands out as a model example. Early efforts to improve airfoil shapes numerically [30], [31] relied on finite-difference schemes to approximate sensitivities. However, as the dimensionality of the design space grew, more efficient approaches emerged—most notably the analytic derivation of sensitivity equations through adjoint methods [32]–[34]. Over the past two decades, automatic differentiation (also referred to as algorithmic differentiation or simply autodiff) has been increasingly integrated into CFD codes [35]–[37], enabling a new generation of high-fidelity design optimization studies [37]–[40].

The notion of optimality extends well beyond technical performance metrics; in a broader context, it underpins the fitting of models and parameters to maximize predictive accuracy given empirical data. Automatic differentiation has played a central role in this domain, powering parameter estimation across a wide range of fields—from ice sheet modeling [41] and optimal control [42] to applications in quantitative finance [43]. Deep learning [44], [45] represents a particularly prominent instance of parameter optimization, where the focus shifts toward learning expressive representations through structured model architectures. This paradigm has led to breakthroughs in computer vision [46], natural language processing [47], and reinforcement learning [48]. Crucially, the widespread availability of automatic differentiation in frameworks such as PyTorch [49] and TensorFlow [50] forms the computational backbone of many of these advances [51].

The term differentiable programming—relatively recent in origin—has emerged to encapsulate the core philosophy behind much of modern deep learning. Initially popularized by Christopher Olah [52], David Dalrymple [53], and Yann LeCun [54], the term refers to the construction of differentiable computer programs using automatic differentiation, which are then optimized via gradient-based techniques with respect to a defined objective on training data. What distinguishes this paradigm from traditional neural network design is its emphasis on general-purpose programming: differentiable programs can include control flow, loops, and recursion, extending well beyond feed-forward architectures. In this broader view, neural networks are simply one class of differentiable programs—composable, trainable blocks within a much larger space of

algorithms. Crucially, this expanded formulation brings into scope differentiable numerical simulators across scientific and engineering disciplines, bridging machine learning and physical modeling in a unified computational framework.

In particle physics, measuring instruments function by extracting inference from the data they collect—thus inherently depending on the models and parameter estimation procedures discussed above. As a result, the optimization of such instruments is tightly coupled to both the precision of the underlying models and the effectiveness of the inference mechanisms. However, particle detectors introduce an additional layer of complexity: their operation is fundamentally rooted in the interaction of radiation with matter, a process governed by quantum mechanics and characterized by intrinsic stochasticity. This non-determinism complicates the formulation of fully differentiable models, often requiring specialized solutions. For this reason, the end-to-end optimization of detectors and accelerators remains a challenging frontier, explored only in a limited number of recent studies [28], [55]–[61].

There are two principal strategies for making a simulation differentiable. The first involves applying automatic differentiation (AD) directly within the simulation code itself, leveraging AD frameworks that operate through operator overloading or source code transformation, depending on the programming language used. As discussed in Sections 1.2.6 and 1.2.7, the feasibility of this approach varies significantly across simulators, and in many cases the technical complexity renders it impractical. An alternative strategy is to construct a differentiable surrogate of the simulator using deep learning. In this case, a neural network is trained in a supervised fashion on data sampled from the original simulation. The advantages of this surrogate-based approach are examined in detail in Section 1.2.8. The tools description presented here follows closely the discussion in [62].

1.2.1 Optimization

The most general formulation of a mathematical optimization problem takes the form

$$\min_{x \in \mathbb{X}} f(X),$$

where $\mathbb{X}$ denotes the space of admissible choices -typically a subset of $\mathbb{R}^d$- and $f : \mathbb{X} \to \mathbb{R}$ is an objective, cost or loss function that quantifies the quality of a given choice. While the terms *loss* and *cost* often refer specifically to functions measuring the discrepancy between predicted and target values at the level of individual data points, the broader term “objective” also encompasses aggregate quantities, regular-

```python
import torch
import scipy.optimize

def f(x): # returns a tuple (value, gradient)
  x_ad = torch.tensor(x, requires_grad=True)
  f_ad = x_ad[1]*x_ad[1]+torch.sin(x_ad[0])
  f_ad.backward()
  return ( f_ad.item(), [x_ad.grad[0], x_ad.grad[1]] )

x0 = (2,-3)
res = scipy.optimize.minimize(f, x0, method='BFGS', jac=True)
print("Minimum: ", res.x[0], res.x[1])
```

(a) Using SciPy [63] for optimization. See the other subfigures, especially 1.2e, for how the gradient is obtained.

```python
import math

def func(x, y):
  return y*y + math.sin(x)

x = 0.
y = 5.
z = func(x,y)


print("z =", z)
```

(b) *Primal*, i. e. undifferentiated, program. It computes the value $f(0{,}5) = 25.0$.

```python
import ad_tool

def func(x, y):
  return y*y + ad_tool.sin(x)

x = ad_tool.DualNumber(0.,1.)
y = ad_tool.DualNumber(5.)
z = func(x,y)

print("z =", z.primal)
print("dz/dx =", z.tangent)
```

(c) Program differentiated in *forward mode* AD using the ad-hoc tool in Fig. 1.2d. It computes $\frac{\partial f}{\partial x} = 1.0$.

```python
import math

class DualNumber:
  def __init__(self, primal, tangent=0):
    self.primal = primal
    self.tangent = tangent

  def __add__(self, other):
    return DualNumber( \
        self.primal+other.primal, \
        self.tangent + other.tangent)
  def __mul__(self, other):
    return DualNumber( \
      self.primal*other.primal, \
      self.primal*other.tangent + \
        self.tangent*other.primal)
  # ... define further operations

def sin(val):
  return DualNumber( \
      math.sin(val.primal), \
      math.cos(val.primal)*val.tangent)
# ... define further operations
```

(d) Ad-hoc tool for forward AD.

```python
import torch

def func(x, y):
  return y*y + torch.sin(x)

x = torch.tensor(0.,
                 requires_grad=True)
y = torch.tensor(5.,
                 requires_grad=True)
z = func(x, y)

z.backward()
print("z    =", z.item())
print("dz/dx=", x.grad)
print("dz/dy=", y.grad)
# alternative:
# torch.autograd.grad(z, [x,y])
```

(e) Program differentiated in *reverse mode* AD with PyTorch [49]. It computes all partial derivatives of the single output variable in one turn, $\frac{\partial f}{\partial x} = 1.0$, $\frac{\partial f}{\partial y} = 10.0$.

Figure 1.2: Applying AD and optimization to a simple function $f(x) = y^2 + \sin(x)$. See the text for more detail. Adapted from [62]

ization terms, and constraints. The precise definition of both $\mathbb{X}$ and f is inherently problem-specific and requires input from domain experts.

In the context of end-to-end detector optimization, the parameter space $\curvearrowleft$ typically spans all tunable aspects of the detector's hardware and software design—from geometric layout and material properties to reconstruction algorithm hyperparameters. The objective function f encodes a complex interplay of physical accuracy, technical feasibility, and economic cost. Evaluating it generally involves running a complete simulation pipeline, which includes modeling of particle interactions, detector response, and downstream data processing.

1.2.2 Gradient-Based optimization

Many of the design parameters x_i in detector optimization problems can be treated as continuous variables that influence the objective function f in a smooth and differentiable manner. This makes them natural candidates for gradient-based optimization within a differentiable programming framework. Gradient descent algorithms are a foundational tool in this setting. Starting from an initial guess $x^{(0)}$, they generate a sequence of parameter updates $x^{(1)}, \ldots, x^{(n)}$ using the rule

$$x^{(k+1)} = x^{(k)} - \eta_k \, \nabla_x f(x^{(k)}), \tag{1.1}$$

where $\eta_k > 0$ is the step size at iteration k, and $\nabla_x f(x^{(k)})$ indicates the direction of steepest descent. Careful tuning of η_k is required to ensure convergence without overshooting the minimum.

When the objective function f is twice differentiable, second-order methods such as (damped) Newton's method can be employed. These incorporate the curvature information encoded in the Hessian matrix $H_x f$, using the update rule

$$x^{(k+1)} = x^{(k)} - \eta_k \, H_x f(x^{(k)})^{-1} \, \nabla_x f(x^{(k)}). \tag{1.2}$$

Quasi-Newton methods such as BFGS [64]–[66] and L-BFGS-B [67], [68] bypass the need for an explicit Hessian by internally constructing an approximation based on gradient evaluations.

These and other optimization techniques are readily accessible through open-source libraries such as SciPy [63]. Figure 1.2a illustrates an example in which a SciPy optimizer is used in conjunction with PyTorch's automatic differentiation to supply gradients. In practice, the user provides an initial parameter configuration and the code

required to evaluate both the objective function and its derivatives. The remainder of this section outlines three principal strategies for obtaining these derivatives in the context of differentiable programming.

1.2.3 Computing derivatives

There are several approaches to computing derivatives of a function f, each with distinct trade-offs in terms of accuracy, efficiency, and applicability.

Numerical differentiation approximates individual gradient components $\frac{\partial f(\mathbf{x})}{\partial x_i}$ by evaluating the function at nearby points. The most common schemes are the forward and central finite-difference quotients

$$\frac{f(\mathbf{x} + h\,\mathbf{e}^{(i)}) - f(\mathbf{x})}{h} \quad \text{and} \quad \frac{f(\mathbf{x} + h\,\mathbf{e}^{(i)}) - f(\mathbf{x} - h\,\mathbf{e}^{(i)})}{2h}$$

where $\mathbf{e}^{(i)}$ is the ith is the unit basis vector and $h > 0$ a small step size. As $h \to 0$, these expressions converge to the true derivative, but in practice, round-off and truncation errors limit how small h can be chosen [51], [69]. While straightforward to implement, numerical differentiation scales linearly in time with the number of input variables and may become inefficient in high-dimensional settings.

Analytic differentiation —performed by hand— or *symbolic differentiation* using computer algebra systems such as Mathematica [70] yields exact, closed-form derivatives. However, these approaches are typically limited to conventional mathematical expressions and do not generalize well to algorithmic constructs like loops, branching, or recursion. Symbolic methods also suffer from expression swell, where the derivative expressions become disproportionately complex and expensive to evaluate.

Automatic differentiation (AD), sometimes called algorithmic differentiation, provides a powerful alternative. AD augments the original code that computes f with additional operations that propagate derivative information through all intermediate computations. The result is exact (up to floating-point precision) and can be computed efficiently, particularly in reverse mode, which is well-suited for functions with many inputs and a single output. Unlike symbolic differentiation, AD handles all programming constructs and is fully compatible with standard software development practices. Together, AD and gradient-based optimization form the foundation of differentiable programming: a paradigm in which optimization problems are expressed as executable code, differentiated automatically, and solved using gradient-based methods. The two primary

modes of AD—forward and reverse—are reviewed in Sections 1.2.4 and 1.2.5, while implementation strategies are discussed in Section 1.2.6. For a broader introduction, see the textbook by Griewank and Walther [71] for a numerical analysis perspective, or the survey by Baydin et al. [51] for a machine learning–oriented treatment.

1.2.4 The AD forward mode

In forward-mode automatic differentiation [72], each variable a is augmented with a corresponding *tangent* variable $\dot{a}$, representing its partial derivative along a given direction. This is typically implemented by introducing a new dual-number data type, where each value stores both its *primal* component (the value of a) and its tangent component (the value of $\dot{a}$). An example of such an implementation is shown in Figure 1.2d, where a custom class defines the dual structure, and overloads standard arithmetic operations—such as multiplication (`mul`)—to propagate derivatives according to the rules of calculus. For instance, evaluating $a = b_1 \cdot b_2$ automatically computes $\dot{a} = \dot{b}_1 \cdot b_2 + b_1 \cdot \dot{b}_2$.

As illustrated in Figure 1.2c, the partial derivative $\frac{\partial f}{\partial x_i}(\mathbf{x})$ is computed by initializing all tangents $\dot{x}_j$ to zero, except for $\dot{x}_i = 1$, and then evaluating the function $y = f(\mathbf{x})$ using the extended dual arithmetic. The resulting derivative can be read directly from the tangent component $\dot{y}$. While conceptually straightforward and easy to implement, forward mode incurs a time complexity proportional to the cost of evaluating f multiplied by the number of input variables. However, it is highly efficient for computing directional derivatives of the form $\nabla_x f(\mathbf{x})^T \cdot v$, which can be obtained in a single pass by initializing the tangents $\dot{x}_j = v_j \forall j$.

1.2.5 The AD reverse mode

Reverse-mode automatic differentiation [73]–[75] operates in two distinct phases. During the first —or *primal*— phase, the function $f(x)$ is evaluated normally, while each computational step is recorded in a data structure such as a computational graph or a tape (alternative representations exist, but the tape abstraction is widely used). n the second —or *reverse*— phase, the algorithm traverses this record backward, from the output toward the inputs, computing derivatives by propagating adjoint variables.

For each variable a encountered during the primal computation, an associated *adjoint* variable $\bar{a} := \frac{\partial f}{\partial a}$ is introduced. These adjoints represent the sensitivity of the final output with respect to each intermediate variable. For example, if the primal computa-

tion contains a step $a = b_1 \cdot b_2$, the reverse pass will apply the updates

$$\bar{b}_1 += \bar{a} \cdot b_2 \qquad \bar{b}_2 += \bar{a} \cdot b_1. \tag{1.3}$$

These equations reflect the application of the chain rule, where the adjoint $\bar{a}$ carries the accumulated gradient information from downstream in the computation. This backward propagation continues recursively, with each operation contributing partial derivatives to its input adjoints based on its role in the overall function.

To initialize the process, all adjoints are set to zero except for the output $y = f(x)$ that receives $\bar{y} = 1$. After the reverse pass, the adjoints $\bar{x}_i$ hold the gradient components $\frac{\partial f}{\partial x_i}$ with respect to each input variable.

A key advantage of reverse-mode AD is its time complexity: it computes the full gradient of a scalar-valued function with a cost that is asymptotically constant with respect to the number of inputs. This makes it particularly well-suited for high-dimensional optimization problems, as commonly encountered in machine learning and detector design. However, this efficiency comes at the cost of memory: the entire computational graph must be retained, making memory management a central concern. Furthermore, reverse-mode AD is substantially more involved to implement than its forward counterpart.

An example of reverse-mode differentiation using PyTorch [49] is shown in Figure 1.2e, where a simple function is differentiated automatically by tracing the backward graph built during the forward computation.

1.2.6 Implementation aspects of AD

The integration of automatic differentiation into a simulation program depends significantly on the capabilities of the programming language in use. One of the most accessible strategies is to replace all arithmetic operations with calls to an AD library that transparently implements the required differentiation logic. This is typically achieved through operator overloading—a feature supported by many modern languages—where the standard arithmetic operators are redefined for custom data types.

As illustrated by the forward-mode example in Figure 1.2d, many AD frameworks rely on polymorphism and operator overloading to extend the behavior of numeric types. In this approach, substituting the native floating-point type (e.g., `float` or `double`) with a custom type provided by the AD library is often sufficient to enable gradient computation. The underlying AD logic is then triggered automatically whenever arithmetic operations are performed on those types. This strategy, adopted by several widely

used AD tools [76]–[79], offers a straightforward path to incorporating AD into existing codebases with minimal changes to program structure.

In Ref.[80], automatic differentiation was incorporated into the general-purpose matrix element generator MadGraph [81] by leveraging the `JAX` framework [82]. While this integration enabled differentiable computations over complex physics processes, it also required non-trivial modifications to the original codebase. This is a common scenario when working with large, mature simulation frameworks: retrofitting AD capabilities often necessitates structural adjustments to ensure compatibility with the AD tool's execution model. We outline typical challenges and required adaptations in more general terms in the following subsections.

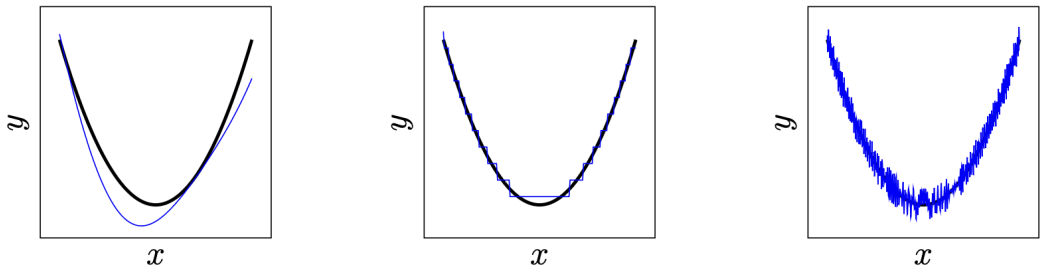
1.2.7 Adaptations to the primal program

A central objective in the development of automatic differentiation tools is to enable seamless integration with existing codebases. Ideally, one would like to differentiate a given program with minimal or no changes to its structure. In practice, however, problem-specific adaptations are often necessary to accommodate the AD workflow. These adjustments may arise from several typical challenges:

- Initialization and output of derivatives: The primal program must often be extended to support the initialization and storage of tangent or adjoint variables, depending on whether forward or reverse mode is used.

- External function calls: Many simulation codes invoke external numerical libraries that are only available in compiled form, without accessible source code. In such cases, derivatives of these functions must be supplied manually -either analytically, numerically, or via surrogate models.

- Memory constraints in reverse mode: Recording every operation during the primal run may exceed memory limits. Techniques such as checkpointing [83] allow specific regions of the code to be re-executed on demand, reducing tape size. Alternatively, preaccumulation can be used to compute derivatives of small subroutines with limited input/output dimensions on the fly, avoiding their explicit recording. This is particularly effective for numerical algorithms that involve many operations to solve conceptually simple tasks and for which analytic derivatives can be substituted. Shared-memory parallelism: In reverse mode, concurrent reads in the forward pass may translate to concurrent writes during adjoint updates (as per Equation 1.3). While AD frameworks can enforce thread safety



(a) Good.

(b) Bad: The derivative is 0, if it exists at all.

(c) Bad: The derivative is dominated by high-frequency noise.

Figure 1.3: Value-wise good approximations can, but do not need to be, good derivative-wise as well. Reproduced from [85].

through atomic operations, these can introduce performance overhead. In some cases, the programmer may opt to disable atomicity manually when the data access patterns are well understood and race conditions can be safely avoided [84].

- Accuracy of the objective approximation: If the primal code only approximates the true objective function, there is no guarantee that its computed derivatives faithfully represent those of the underlying quantity of interest. This issue is illustrated in Figure 1.3. Manual intervention may be required to ensure that the approximation and its derivative are aligned, corresponding to the well-behaved scenario of Figure 1.3a.

These considerations highlight the importance of domain knowledge and careful implementation when adapting complex scientific software for use with AD.

1.2.8 Surrogate models

An alternative strategy for enabling differentiable optimization is to apply algorithmic differentiation not to the original objective function directly, but to a surrogate model trained to approximate it [55]. While the original objective may involve complex, domain-specific simulation code that is not readily differentiable, surrogate models are typically drawn from simpler, general-purpose function classes—most commonly neural network architectures from the deep learning literature. These models are trained to mimic the behavior of the original function by fitting their parameters (or weights) on a representative dataset of input-output pairs.

When the surrogate is implemented using a deep learning framework, automatic differentiation comes for free: both the model and its derivatives can be evaluated ef-

ficiently within the same computational environment used for training. This makes the surrogate differentiable even when the original function is not. Moreover, inference and gradient computation with such models are often orders of magnitude faster than evaluating the full simulation [86], thanks to optimized implementations that exploit hardware acceleration through GPUs or TPUs.

However, these advantages come with trade-offs. Training a surrogate requires a potentially large number of evaluations of the original objective function—at least scaling with the dimensionality of the design space. Furthermore, any mismatch between the surrogate and the true function may introduce bias into the optimization process, particularly if the surrogate fails to capture key features of the underlying model. For this reason, careful training and validation are essential. The development of surrogate models takes a relevant role in this thesis, as we choose this strategy to approach the optimization study of the Muon Collider (MuC-MuColl) Electromagnetic Calorimeter (ECal). In particular Chapter 5 showcases examples of such methods and discusses in detail their implementation.

Chapter 2

New Challenges in Experimental Physics: The SWGO Case

The design of particle detectors has traditionally relied on expert intuition, performance scans, and subsystem-level studies conducted in isolation. While effective in many cases, this approach makes it difficult to globally optimize experimental performance, especially when physics goals require joint consideration of geometry, reconstruction, and signal-background discrimination.

In recent years, differentiable programming has emerged as a powerful strategy for end-to-end optimization of scientific instruments. By making each stage of the simulation and reconstruction process differentiable with respect to design parameters, one can propagate gradients from final physics objectives back to detector layouts, enabling iterative improvements in a principled, automated fashion.

This chapter explores the application of this methodology to the Southern Wide-field Gamma-ray Observatory (SWGO), a proposed high-altitude array of water Cherenkov detectors (WCDs) designed to observe extensive air showers from cosmic gamma rays. SWGO presents a compelling test case: its layout must balance competing requirements for sensitivity to low-energy events, angular resolution for source localization, and background rejection of hadronic showers. These objectives, while physically meaningful, translate into complex trade-offs in the spatial arrangement of detectors.

The work presented here builds on two key contributions:

1. A surrogate-based optimization framework that computes utility gradients analytically from a parametrized model of shower signals and reconstruction likelihoods;

2. A theoretical treatment of utility functions as representations of experimental goals, grounded in information theory and performance-based inference.

We show that this pipeline can be used to improve sensitivity and resolution relative to benchmark SWGO layouts, and that it naturally produces geometric configurations that encode physical priorities—e.g., elliptical elongation for improved triangulation, or compact cores for low-energy triggering.

The remainder of this chapter is structured as follows:

- Section 2.1 provides an overview of the SWGO experiment and its goals.
- Section 2.2 introduces the physical and operational features of the SWGO detector units and simulation model;
- Section 2.3 details the gradient-based optimization pipeline and algorithm;
- Section 2.4 defines the utility components and interprets them in both physical and information-theoretic terms;
- Section 2.5 presents optimization outcomes, performance gains, and layout evolution under varying spectral assumptions;
- Section 2.6 summarizes key findings and discusses the implications for future detector optimization efforts.

Throughout, we emphasize clarity, reproducibility, and physical motivation, aiming to make this case study a benchmark for differentiable design in particle astrophysics.

2.1 Context

The Southern Wide-field Gamma-ray Observatory (SWGO) is a proposed experiment designed to study Very-High-Energy (VHE) gamma rays, i.e. in the Tev-PeV range, which range from sources in the southern sky. Its scientific objectives include the study of Galactic accelerators, transients, and diffuse emissions, as well as the search for new signatures of physics beyond the Standard Model [87]. Like its northern counterparts HAWC[88] in Mexico and LHAASO[89] in China, SWGO is conceived as an array of water Cherenkov tanks deployed at high altitude to sample the secondary particles produced by extensive air showers (EAS). When a primary gamma ray interacts with atmospheric nuclei, it initiates an electromagnetic shower whose footprint can extend over hundreds -if not thousands- of meters. Hadronic primaries, such as protons, generate more complex showers with a richer muonic component and higher stochasticity. At ground level, these differences can be exploited to reconstruct the properties of the

primary and to discriminate signal from background. The different shower development and topology is illustrated in figure 2.1, comparing lateral distributions for photon- and proton-induced events.

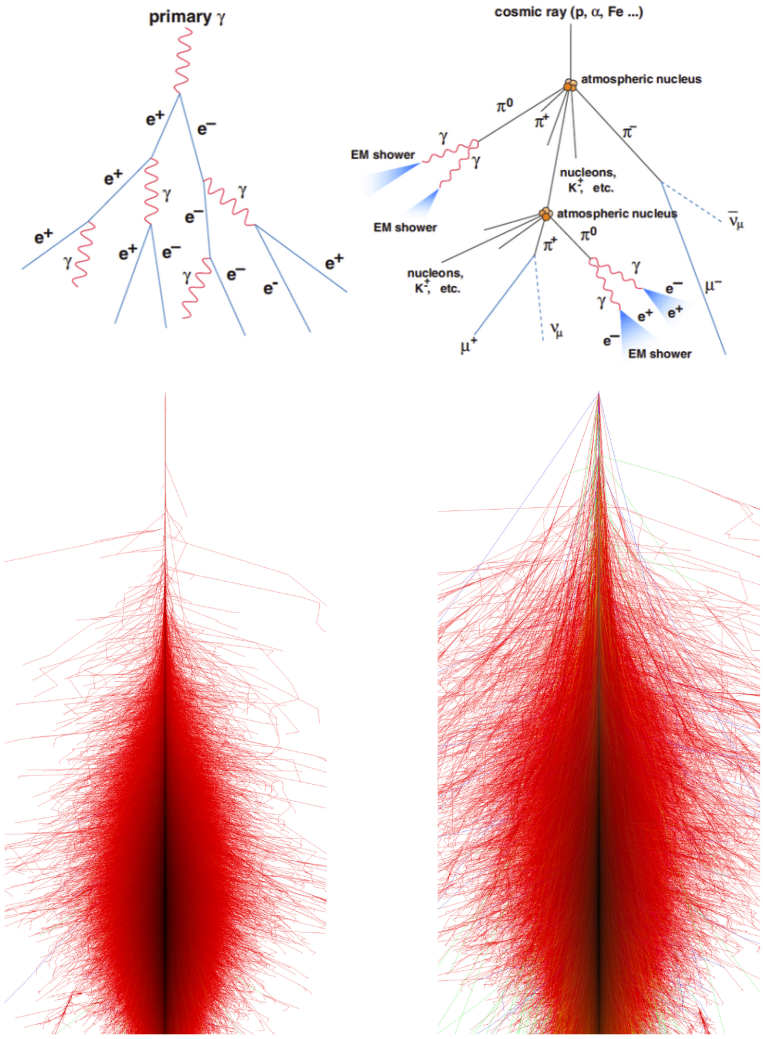


Figure 2.1: Top: sketch of an electromagnetic and hadronic shower in the atmosphere; Bottom: Lateral development of simulated 10TeV showers originated by primary photon (left) and proton (right). Source: CORSIKA web page [90]. Adapted from [91]

The layout of the SWGO array -namely the number, type and position of detector units- has a critical impact on the instrument's performance across energy regimes. Low-energy (O(10 GeV)) events are abundant but produce compact showers that require a densely instrumented core for efficient triggering and accurate reconstruction. Conversely, high-energy (PeV-scale) photons are extremely rare and demand larger instrumented areas to achieve acceptable exposures. The result is a design tension between competing objectives: a dense inner array to ensure high sensitivity at low

energies, and a sparse periphery to maximise effective area at higher energies. This tension is evident in the family of benchmark layouts (A1-F1) in Figure 2.2 proposed by the SWGO collaboration[87], which exhibit concentric regions with decreasing fill factor toward larger radii.

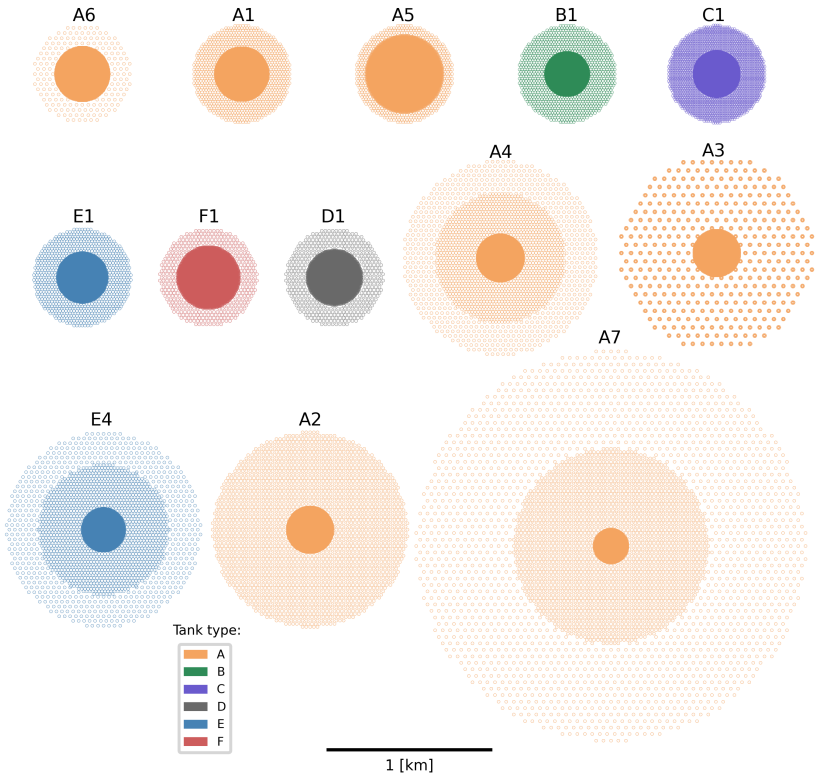


Figure 2.2: Preliminary arrays considered by the SWGO collaboration. Different tank types and filling fractions are considered. Adapted from [91].

Historically, optimization studies in experimental astroparticle physics have relied in discrete comparisons of hand-crafted layouts, evaluated via full simulations using tools like CORSIKA[90]. Such simulations are computationally expensive, making the number of different configurations that can realistically be compared very limited. Moreover, the optimization process is seldom gradient-driven and typically lacks continuity in the parameter space. This limits the ability to uncover subtle structures or correlations in the design landscape. The present study aims to overcome these limitations by introducing an end-to-end optimization pipeline that treats the detector layout as a continuously parametrized object. A differentiable surrogate model captures the particle flux from gamma and proton showers as a function of energy, zenith angle, and ground position. These parameterizations are then linked to a statistical inference procedure that reconstructs primary particle properties from synthetic detector responses. Finally, a scientifically motivated utility function quantifies the array performance, and

its derivatives with respect to the detector positions, computed analytically thanks to the closed form of all its ingredients, is used to steer a gradient-based optimization loop.

This pipeline is modular and interpretable, designed to trace analytical gradients through each step of the simulation-to-utility chain. It avoids reliance on black-box heuristics and instead foregrounds physical intuition and control over modeling assumptions. In doing so, it embodies the methodological shift advocated by the MODE collaboration: to reformulate experiment design problems in terms of differentiable programming frameworks and continuous optimization [62]. A beta release of the code (15,000 lines in C++) with commentary and model dataset files is available on github at >SWGOLO REPO<

The SWGO case serves not only as a practical optimization target, but also as a pedagogical prototype. It allows us to illustrate, in a concrete and tractable setting, the core ideas that will later be generalized to more complex scenarios, such as the TensorFlow-based optimization of the Muon Collider calorimeter discussed in the later chapters. In both cases, the focus is on constructing a pipeline where each component (simulation, reconstruction, evaluation) is differentiable and contributes meaningfully to the global objective. SWGO however, with its comparatively low-dimensional design space and well-understood physics, provides and ideal first case study.

2.2 SWGO detector and design landscape

The core detection principle of the SWGO detector array involves the sampling of extensive air showers (EAS) generated when a high-energy photon or hadron interacts with atmospheric nuclei. Secondary particles—mainly electrons, positrons, muons, and photons—reach ground level within a few tens of nanoseconds, forming a spatial footprint that extends from hundreds of meters (for sub-TeV primaries) to over a kilometer at PeV energies. Placing the array at a high-altitude site (above 4400 m a.s.l.) maximizes the number of particles reaching the detectors before shower attenuation becomes dominant. Figure 2.3 illustrates the footprint at ground level arising for the two primaries considered for this study, protons and gamma rays, highlighting their respective constituents. The muonic component, arising from the decay of K and π mesons, is a relevant signature of hadronic showers.

To detect the shower footprint, SWGO proposes an array of several thousand modular detector units, primarily water Cherenkov tanks of cylindrical shape, each equipped with one or more photomultiplier tubes (PMTs). These units are designed to distinguish between electromagnetic and muonic components of the shower, for example

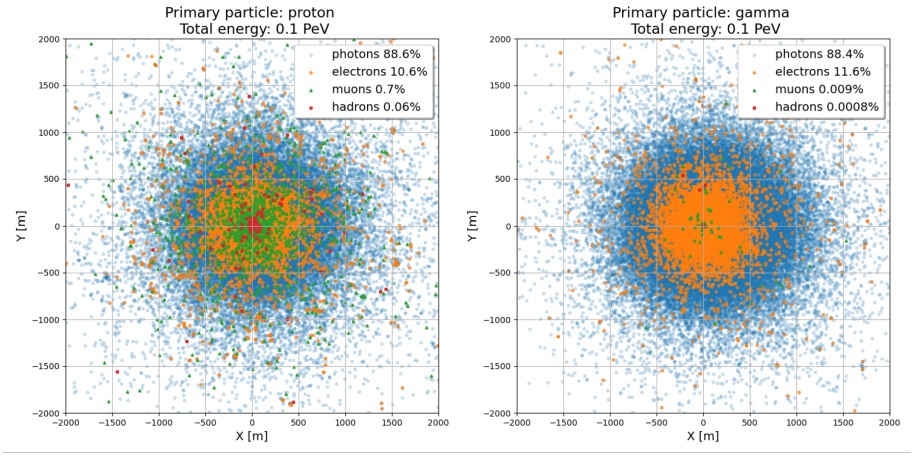


Figure 2.3: Radial distribution of secondary particles at ground level for a 100TeV proton (left) and photon (right) primary. From [91]

by employing vertical segmentation to isolate the short-range electrons and photons in an upper volume, while allowing muons to penetrate to a lower chamber. A selection of candidate tank designs is shown in Figure 2.4. For the purpose of this optimization study, a single reference design—tank A, a double-layer unit 3 meters tall and 3.82 meters in diameter—was adopted to reduce model complexity and isolate the geometric effects of layout choices.

In the absence of a finalized detector concept, we adopt a simplifying assumption of perfect detection efficiency and species identification, thereby decoupling layout optimization from the detailed design of individual units. While clearly not realistic, this idealization allows us to isolate the contribution of spatial geometry to the scientific performance of the array and will be relaxed in future work.

A central design tension arises from the conflicting requirements of low-energy triggering and high-energy coverage. The detection of 100 GeV–scale showers demands high sampling density to ensure sufficient particle counts in the core, while the detection of PeV photons requires a large total footprint to compensate for the extremely low flux. To address this, the SWGO collaboration defined a set of benchmark layouts, each structured into concentric zones with decreasing fill factors toward the periphery. These designs, articulated in different Zones, typically include a dense core (with fill factor above 80%) and outer regions with fill factors (FF) down to 1% or less. Figure 2.2 provides a bird's-eye view of several proposed configurations, while Table 2.1 summarizes their geometric parameters, including tank counts and radial extent. The so-called "A1" layout serves as a reference configuration, balancing a compact inner array with a modestly instrumented outer zone.

The scientific viability of each layout is typically assessed through Monte Carlo sim-

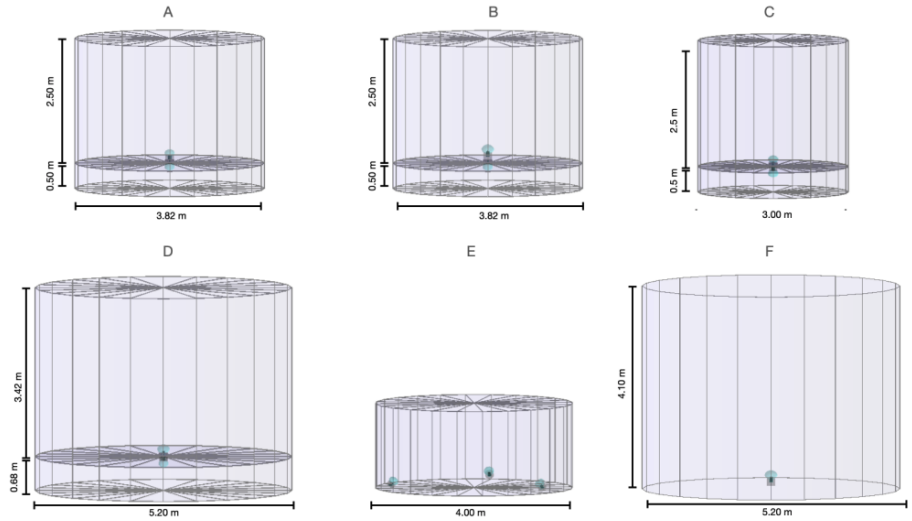


Figure 2.4: Schematic design of different SWGO tank designs. Tank design A is called the reference one; it consists in a double layer unit. Each layer holds a photomultiplier tube, both of equal size. With same dimensions tank B replaces a photomultiplier tube by a larger one. Tank design C (D) is a smaller (larger) version of tank A (B). Tank designs E and F are single-layer units with three (one) photomultiplier tubes of equal size, respectively. From [87]

ID	N_{det}	Zone 1			Zone 2			Zone 3		
		FF(%)	Radius (m)	Units	FF(%)	Radius (m)	Units	FF(%)	Radius (m)	Units
A1	6589	80	160	5731	5	300	858			
A2	6631	80	138	4303	2.5	600	2328			
A3	6823	80	138	4303	2.5	600	2520			
A4	6625	80	140	4429	4.0	400	1518	1.25	600	678
A5	6541	40	234	6109	5.0	300	432			
A6	6637	88	162	6469	1.0	300	168			
A7	6571	80	101	2335	2.5	600	2394	0.63	1200	1842
B1	4849	80	131	3865	5	300	984			
C1	8371	80	137	6829	5	300	1542			
D1	3805	80	166	3367	5	300	438			
E1	5461	80	150	4639	5	300	822			
E4	5455	80	140	3403	4.0	400	1428	1.25	600	624
F1	4681	80	188	4303	5	300	378			

Table 2.1: Characteristics of the benchmark layouts considered for the SWGO Cherenkov tanks. See Figure 2.2 for an illustration of the plans of the layouts.

ulations of atmospheric showers. For this study, which decided to focus on the very high-energy end of the spectrum because of its expected higher complexity and dependence on the pattern of the layout at very low fill factors required for very extended showers, a comprehensive set of photon and proton showers was generated using the CORSIKA package, which provides a high-fidelity model of secondary particle pro-

duction and propagation in the atmosphere. Showers were generated with energies spanning from 0.1 PeV to 10 PeV and zenith angles up to 65°, matching the energy range of primary interest for high-impact astrophysical sources. Table 2.2 details the main simulation campaigns, including their primary type, energy range, and angular coverage. These datasets provide the empirical basis for the construction of a continuous, differentiable model of the secondary flux at ground level, which will be used in the next section to drive the optimization pipeline.

	ID	Energy	Angle	Events
1-8	γ	0.1/0.3/1/3 PeV	$20°/40°$	10,000
9	γ	0.1-10 PeV	$0°$-$65°$	100,000
10-11	p	0.1-10 PeV	$20°/40°$	200,000
12	p	1.0-10 PeV	$0°$-$65°$	200,000

Table 2.2: Summary of dataset used for the study.

2.3 Surrogate models and pipeline

The central methodological contribution of this study is the construction of a fully differentiable pipeline that maps a detector layout configuration into a set of physics-informed observables, from which a scientific utility function is derived. The pipeline is built from analytically tractable components, enabling the use of gradient-based optimization techniques to update the positions of detector units with respect to the utility function. The architecture is modular, interpretable, and designed to allow backpropagation of gradients through every computational block. A schematic representation is shown in Figure 5. Each stage of the pipeline performs a distinct operation: from simulating the particle flux on the ground, to reconstructing the shower parameters from synthetic detector hits, to computing a statistical discriminator for background rejection. These operations, described in detail below, are implemented under simplifying assumptions that maintain analytical accessibility while retaining physical relevance.

Modeling of shower fluxes

The pipeline begins with a continuous model of the flux of secondary particles arriving at ground level from air showers. This model is extracted from high-statistics CORSIKA simulations of gamma and proton primaries, spanning the energy range 0.1–10 PeV and zenith angles up to 65° as summarized in Table 2.2. The spatial distribution

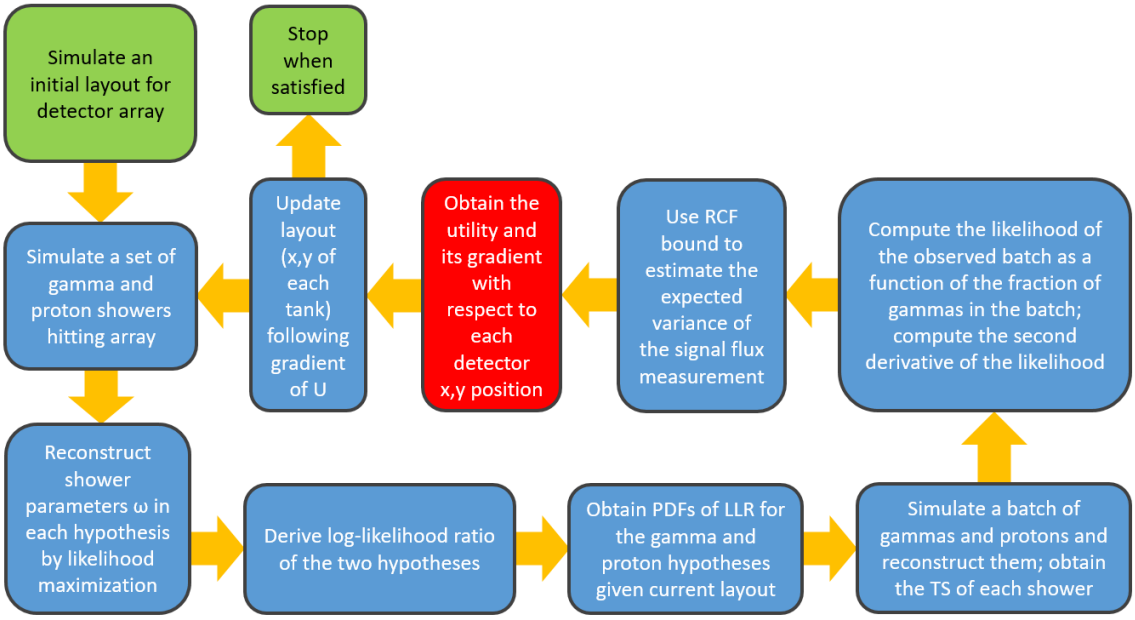


Figure 2.5: Illustrative pipeline diagram for the SWGO detector array optimization.

of secondary electrons, photons, and muons is parametrized as a function of radial distance from the shower axis, primary energy E, and polar angle θ. To first order, the flux is assumed to be azimuthally symmetric and depends only on the lateral distance r from the core.

For each particle type $i \in \{e, \mu\}$ and primary hypothesis $k \in \gamma, p$, we fit the radial distribution of ground-level flux $\lambda_i^{(k)}(r; E, \theta)$ using a closed-form expression. The chosen functional forms are empirical fits designed to be differentiable, monotonic, and compactly supported, typically of the form:

$$\lambda_i^k(r; E,\theta) = A(E,\theta)\left(1 + \frac{r}{r_0(E,\theta)}\right)^{-\beta(E,\theta)}$$

where A, r_0 and B are energy- and angle-dependent parameters extracted from the simulation grid. These fits are shown to accurately reproduce the mean lateral profiles of both gammas and protons, including the sharper falloff of electromagnetic components and the broader tails of muons. Full parameterizations and fit diagnostics are reported in Appendix A of [91].

Modeling of detector response

To compute the expected response of the detector, we assume that each unit is a perfect circular cylinder with radius R centered at position $(x,\, y)$. Letting r_j denote the lateral distance between the detector and the shower core, the expected number of particles of type i that can be detected is:

$$\Lambda_{i,j}^{(k)} = \int_{\text{tank } j} \lambda_i^k(r; E,\theta)\, d\Sigma \;\approx\; \lambda_i^{(k)}(r_j; E, \theta)\; \Sigma$$

where $\Sigma = \pi R^2$ is the cross-sectional detector area. This approximation holds in the regime where fluxes vary slowly over the scale of a unit's size.

Given this expectation, the actual number of detected particles $N_{i,j}$ is sampled from a Poisson distribution, with optionally some Gaussian smearing to model the detector resolution:

$$N_{i,j} \sim \text{Poisson}\big[\mathcal{N}(\Lambda_{i,j},\; \sigma_{rel}\Lambda_{i,j})\big].$$

Here σ_{rel} denotes a relative resolution term, set to 5% throughout the simulation. The idealized assumption of perfect detection and species identification removes the need for detailed detector response simulations, enabling direct gradient computation with respect to layout parameters.

Shower parameters reconstruction

Given the set of detector responses $\{N_{\mu,i}, N_{e,j}\}$, the reconstruction task is to infer the most probable shower parameters under two competing hypotheses: that the primary particle was a gamma ray, or that it was a proton. The parameter vector is defined as:

$$\Theta = (E, \theta, \phi, X_0, Y_0).$$

For a given hypothesis $k \in \{\gamma, p\}$, the expected number of muonic and electromagnetic particles at detector unit j is denoted respectively by $\Lambda_{\mu,j}^{(k)}(\Theta)$ and $\Lambda_{e,j}^{(k)}(\Theta)$. These quantities are obtained from the surrogate flux models introduced earlier.

Assuming statistical independence across detectors and particle types, the likelihood function under hypothesis k is expressed as:

$$\mathcal{L}_k(\Theta) = \prod_j P(N_{\mu,j}|\Lambda_{\mu,j}^k(\Theta))\; P(N_{e,j}|\Lambda_{e,j}^k(\Theta))$$

where $P(N|\Lambda)$ denotes the Poisson probability mass function with expectation Λ. In

each case the best-fit parameters $\hat{\Theta}$ are found as:

$$\hat{\Theta}_k := \arg\max_{\Theta} \log \mathcal{L}_k(\Theta), \qquad k = \gamma,\ p.$$

The resulting log-likelihoods $\log \mathcal{L}_\gamma$ and $\log \mathcal{L}_p$ are then used to compute the test statistic

$$T := \log \mathcal{L}_\gamma(\hat{\Theta}_\gamma) - \log \mathcal{L}_p(\hat{\Theta}_p). \tag{2.1}$$

The choice the such test statistic is inspired by the Neyman-Pearson lemma [92]. Although the dependence on five parameters sets ourselves outside the simple-versus-simple hypothesis scenario wherein the likelihood ratio is the most powerful test statistic for discrimination of the two components, such quantity still encodes excellent discrimination capabilities in the considered parameter space. This scalar variable provides a measure of classification power between gamma and proton events. Additionally, the fitted parameters $\hat{\Theta}_k$ carry information about the expected measurement precision on energy, angle, and core location.

Utility evaluation and optimization loop

Once a batch of simulated events has been reconstructed, the resulting shower parameters and test statistics are used to evaluate the scientific performance of the current detector layout. This is done through a scalar utility function U, which quantifies how well the layout supports key experimental objectives such as flux sensitivity, angular and energy resolution, and gamma/hadron discrimination.

Although the explicit structure of U is deferred to Section 2.4, it is important to emphasize that the utility is constructed exclusively from quantities produced by the reconstruction pipeline: the best-fit parameters $\hat{\theta}_\gamma$, the likelihood scores $\mathcal{L}_l$, and the derived test statistic T. These outputs are aggregated across a batch of simulated showers to form statistical summaries (e.g., variances, class separations), which are then mapped into utility components and ultimately combined into a scalar score.

With the utility function providing a scalar performance measure that is fully differentiable with respect to the detector layout, the entire pipeline can be embedded in an optimization loop. Each iteration samples new events, evaluates the expected response of the array, reconstructs shower properties, computes the utility, and updates detector positions via analytic gradients. This procedure is formalized in Algorithm 1, which defines the core of the optimization strategy explored in the remainder of this chapter.

Critically, the entire chain of transformations — from detector layout to reconstructed observables and utility — is differentiable with respect to the detector positions $\{(x_i, y_i)\}_i$. These gradients $\nabla_{(x_i,y_i)}U$ are then used to update the layout parameters at each iteration of the optimization cycle. Derivatives are computed analytically by chaining the differentiable expressions across the flux, response, reconstruction, and utility components. While this work relies on explicit analytical gradients, the modular structure of the pipeline ensures that autodiff-compatible code structures may be used to propagate gradients efficiently in future extensions.

Algorithm 1 Differentiable Layout Optimization Loop

Require: Initial layout $\mathcal{L}_0 = \{(x_i, y_i)\}$, learning rate η, surrogate model, spectrum $f(E)$

for each optimization epoch $e = 1, \ldots, E$ **do**
    Sample batch of gamma and proton showers from spectrum $f(E)$
    Compute expected detector responses using surrogate model
    Reconstruct shower observables under both hypotheses
    Compute test statistic $T = \log \mathcal{L}_\gamma - \log \mathcal{L}_p$ and its uncertainty σ_T for each sampled event
    Substitute T_p values of proton showers in batch with $G(T_p, \sigma_{T_p})$ distributions to obtain continuous model of T_p distribution for protons and photons
    Do same for gamma showers in batch
    Sample new batch of protons and gamma showers, reconstruct, evaluate T values for each
    Derive estimate of gamma flux and uncertainty, energy and angle resolutions
    Evaluate classification and reconstruction utility components
    Aggregate scalar utility $U = \sum_k \lambda_k U_k$
    Compute analytic gradients $\nabla_{x_i} U$, $\nabla_{y_i} U$
    Update layout: $(x_i, y_i) \leftarrow (x_i, y_i) + \eta \nabla_{(x_i,y_i)} U$
end for

Note on triggering

One important parameter influencing the optimal configuration of the array is the minimum number of detection units required to register a signal —either from muons or electromagnetic particles— for an extensive air shower (EAS) to be considered in the analysis. If too few units are triggered, the resulting reconstruction of the shower's origin and properties becomes unreliable, motivating the application of a lower threshold on the number of coincident detectors. The SWGO collaboration currently applies such a threshold in its official reconstruction pipeline, typically requiring between 30 and 50 triggered units [93].

In the present study, which focuses on the upper end of the primary gamma-ray

energy spectrum of interest to SWGO, we adopt a conservative selection criterion and require $N_{trigger} \geq 50$ units for an event to be included in the calculation of the gamma-ray flux and the utility function. This choice ensures that only well-reconstructed showers contribute to the optimization process. Consequently, any gradient-based update of the array configuration must also account for the probability that a shower falls below this trigger threshold if a unit is moved. This probabilistic consideration becomes an integral component of the utility gradient, ensuring that layout modifications do not inadvertently suppress the effective acceptance of high-energy events.

To correctly inform the gradients during optimization, we compute the probability that a given detector unit registers at least one secondary particle —whether electromagnetic or muonic— by evaluating the expected detection probability across the array. For each unit, the probability of detecting at least one particle is given by

$$S(N_{\text{obs}} \geq 1) = \sum_{i=1}^{N_{\text{det}}} \left(1 - \exp(-\hat{N}_{\mu,i} - \hat{N}_{e,i})\right) \tag{2.2}$$

where $\hat{N}_{\mu,i}$ and $\hat{N}_{e,i}$ denote the expected number of muons and electromagnetic particles at detector i, respectively. This expression aggregates the individual probabilities over all detectors, yielding the expected number of active units for a given shower. To enforce the event selection threshold $N_{trigger}$, we then estimate the probability that at least $N_{trigger}$ detectors are active using the cumulative distribution of the Poisson law:

$$p(N_{\text{active}} \geq N_{\text{trigger}}) = 1 - \sum_{k=0}^{N_{\text{trigger-1}}} [\text{Poisson}(k,S)] \tag{2.3}$$

where S is the aggregate expected number of active units computed above. This probability enters directly into the utility function, and hence its gradient, allowing the optimization process to account for changes in detection efficiency due to layout modifications.

In essence, we approximate the combinatorial probability that at least $N_{trigger}$ out of N_{det} detectors register a signal S as the expected value of a Poisson process. This yields a tractable expression for the cumulative probability using only the aggregate detection probability, avoiding the need for explicit enumeration of detector-level outcomes. While this simplification neglects correlations between units, it proves sufficiently accurate for our purposes. Specifically, we verify through dedicated toy simulations that the approximation reproduces the true activation probability with a typical deviation below 1.5% in scenarios relevant to our study—namely, when the number of detectors triggered is near the threshold $N_{trigger}$, and the probability is therefore most

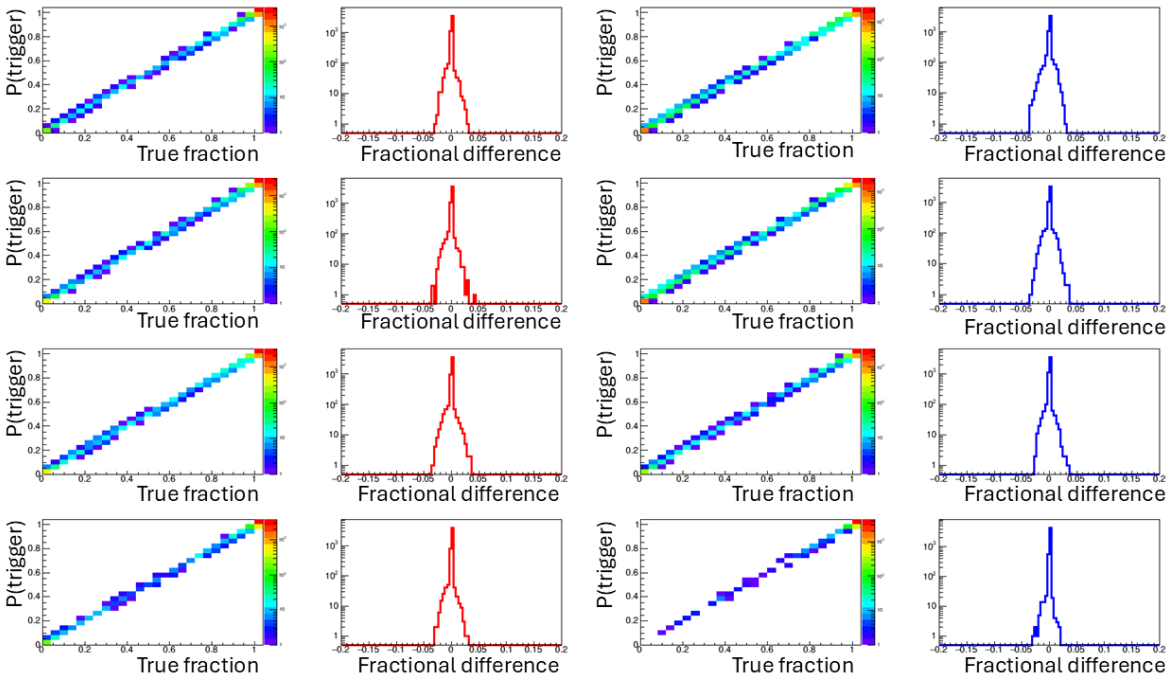


Figure 2.6: Top: Log-mean relative angular (left) and energy (right) error for gamma showers reconstructed with the A5 layout (6288 detectors). Bottom: Mean energy and pointing error vs. energy and core distance for a dense 361-detector array, representing the proposed SWGO first-phase configuration.

sensitive to layout changes. The result of one such validation test is shown in Figure 2.6

The above expressions have derivatives with respect to the detector positions (x_i, y_i) that can be determined analytically via chain rule (see [87] for the derivation).

2.4 Utility function definition

2.4.1 Signal fraction uncertainty from shower batches

A measurement of the gamma-ray flux requires, in practice, two key ingredients: the reconstruction of all observed showers within a given time window, and an estimation of the fraction of those showers that originate from genuine gamma primaries. As discussed in the preceding sections, we adopt a likelihood ratio–based test statistic to quantify the separation power between gamma- and proton-induced showers, given a specific detector layout. For a batch of simulated showers—generated under both gamma and proton primaries and corresponding to a fixed data-taking interval—we reconstruct each event under both hypotheses. From these reconstructions, we compute a test statistic value T for every shower. Importantly, rather than applying a hard

threshold on T to isolate a signal-enriched subset, we take a more flexible and differentiable approach: we directly fit for the underlying gamma-ray fraction by maximizing the likelihood of the observed $T-$distribution under a parametric mixture model.

This choice has two main advantages. First, it circumvents the non-differentiable nature of threshold-based selections, which would otherwise obstruct gradient-based optimization. Second, it retains full information from the entire batch of showers, including those with intermediate T-values that would be discarded by a binary cut but still contribute to the overall statistical inference. This continuous, differentiable formulation is essential for embedding signal extraction within a broader end-to-end optimization framework, where the gradients of the objective function must be smoothly propagated with respect to detector positions and other layout parameters.

Given a batch of N_{batch} reconstructed showers, each associated with a value of the test statistic $\mathcal{T}$, we estimate the fraction i of photon-induced events, f_γ, by fitting a two-component mixture model. Specifically, we maximize the likelihood function

$$\mathcal{L}(f_\gamma) = \prod_{i=1}^{N_{\text{batch}}} \left[f_\gamma P_\gamma(\mathcal{T}_i) + (1 - f_\gamma) P_p(\mathcal{T}_i) \right], \tag{2.4}$$

where $P_\gamma(\mathcal{T})$ and $P_p(\mathcal{T})$ denote the probability density functions of the test statistic under the gamma and proton hypotheses, respectively. These densities are derived from large samples of simulated events and remain fixed during the fitting procedure. The recipe for deriving the two distributions $P_\gamma(\mathcal{T})$ and $P_p(\mathcal{T})$ is detailed on Section 5.1 of [87].

Determining those distributions is needed to define our utility function. This will not depend on the gamma fraction itself f_γ -which is a fixed value during event generation and dependent on the triggering criterion- but rather on how precisely our experiment is able to measure it. In practice what we are interested in is the variance of f_γ, which we can estimate through the RCF bound [94] by plugging in the likelihood of equation 2.4:

$$\sigma^2[f_\gamma] = \left[- \frac{d^2 \log \mathcal{L}}{df_\gamma^2} \right]^{-1}. \tag{2.5}$$

2.4.2 Calculation of the utility function

The recent discovery of twelve PeV gamma-ray emitters in the northern sky by the LHAASO collaboration [89] has generated renewed interest in the search for ultra-high-energy gamma sources. These detections were surprising in both number and intensity, given prior expectations based on known astrophysical accelerators. This result has

significantly raised the stakes for a complementary southern-hemisphere instrument. In this context, SWGO is uniquely positioned to explore the PeV gamma-ray sky from the southern hemisphere, with the potential to uncover new classes of sources or reveal anisotropies in source populations

As such, the measurement of the gamma-ray flux in the 100 TeV to 10 PeV range is adopted here as a baseline scientific objective for the optimization study. To quantify how well a given array layout supports this goal, we now define a utility function that encapsulates both sensitivity and statistical uncertainty in a form compatible with gradient-based optimization.

$$U_{GF} = \frac{\Phi_\gamma}{\sigma_\Phi \sqrt{\rho}}, \tag{2.6}$$

In the definition of the utility function, Φ_γ denotes the gamma-ray flux, σ_Φ its associated uncertainty, and ρ a shower density factor, which will be discussed in detail below. This utility function provides an integrated measure of the expected precision in determining Φ_γ across the 0.1–10 PeV energy range. While it is possible to subdivide this objective into energy-dependent components—assigning different weights to distinct spectral intervals—we do not implement such fine-grained structure here. The formulation can be easily extended in that direction if needed.

Nonetheless, we note that the choice of spectral index used when generating the simulated events implicitly acts as a weighting factor. A steeply falling spectrum emphasizes lower-energy showers, whereas a harder (or even rising) spectrum gives more weight to higher energies. To avoid introducing such biases into the optimization procedure, we generate showers uniformly in energy across the target range in most of our studies. This design choice ensures that the resulting detector configurations are not driven by a specific energy preference unless explicitly modeled—and must be borne in mind when interpreting the optimized layouts.

The density factor ρ in the utility function accounts for the number of showers generated during each step of the optimization process. As the array layout changes, the effective ground area over which shower cores are sampled may vary—particularly if detectors move closer together or spread out. If the number of simulated showers per optimization epoch is held fixed, then ρ serves as a corrective factor, capturing variations in sampling density that would otherwise bias the utility estimation.

Although the scaling of the flux uncertainty σ_Φ with $1/\sqrt{N_{batch}}$ is a reasonable approximation under ideal conditions, it introduces an implicit dependence on batch size. A more robust strategy would involve generating a fixed density of showers over a sufficiently large area encompassing all plausible array configurations. This would de-

couple the statistical precision from the batch size entirely, at the cost of increased computational expense.

For the present work, we adopt the simpler approach and rely on the scaling $1/\sqrt{\rho}$, where ρ is computed dynamically from the number of events in each batch. While this choice enables faster iterations and is sufficient for exploratory optimization, we anticipate future runs of the pipeline will adopt the more conservative fixed-area sampling scheme to ensure maximal robustness.

To adapt the simulation volume dynamically during optimization, we update the region over which shower cores are generated at each training epoch according to the current configuration of the array. This ensures that all detector units remain adequately illuminated by the incoming flux. The procedure is as follows:

1. Compute array scale — The average radial extension $\bar{R}$ and the RMS spread σ_R of the detector positions are computed relative to the array center, defined as the origin $(0,0)$ in the detector plane. While cylindrical symmetry is not assumed, this radial metric serves as a reliable proxy for the array's overall spatial extent during optimization.

2. Add a slack margin — A fixed slack term R_{slack}, set by default to 2km, is added to ensure coverage of outlier units and allow for layout evolution. This margin avoids the need to track the outermost detector (which would lead to a non-differentiable condition) and instead offers a smooth measure of spatial coverage. In later optimization stages, some distant units may fall outside the core generation area, but this is acceptable for exploratory studies and implicitly constrains the array's radial expansion.

3. Define the core generation region — Shower cores are sampled uniformly in x and y within a circular region of total radius

$$R_{\text{tot}} = \overline{R} + 2 \times \sigma_R + R_{\text{slack}}$$

   centered at the origin of the detector plane.

This approach ensures that the shower sampling region evolves naturally with the array geometry, maintains coverage for all active units, and supports differentiable optimization of the utility.

To ensure uniform and relevant coverage of the instrumented region, shower cores are only retained if they fall within a distance R_{slack} from at least one detector unit. Core positions lying outside this range are discarded and replaced by new samples. This

process continues until exactly N_{batch} accepted core positions are obtained. The total number of generated positions, including discarded trials, is recorded as N_{trials}. For each accepted shower, energy, polar and azimuthal angle of incidence, and particle identity are drawn at random from the specified priors.

This sampling strategy is repeated at every epoch of the optimization, dynamically adapting to the evolving array configuration. The effective area Σ covered by the accepted showers is computed as

$$\Sigma = \pi R_{\text{tot}}^2 \frac{N_{\text{batch}}}{N_{\text{trials}}}. \tag{2.7}$$

where R_{tot} is the total sampling radius defined previously. Consequently, the shower density per unit area —relevant for normalizing the utility function— scales with

$$\rho \propto \frac{N_{\text{trials}}}{R_{\text{tot}}^2}. \tag{2.8}$$

While the normalization constant (including π) is not strictly required, the proportionality is used to compute relative densities across epochs.

A further consideration in the formulation of the utility function is the precision with which key physical parameters of gamma-originated showers can be reconstructed. In particular, the energy of the primary and its polar and azimuthal angles of incidence are of scientific interest. By contrast, the ground-level core coordinates

$$(X_0, Y_0)$$

do not encode information about the primary flux and are thus treated as nuisance parameters in the inference framework. Their uncertainty does not enter the utility directly but influences the reconstruction indirectly through geometric effects.

To assess the overall energy resolution achievable by a given detector layout, we define a utility function U_{IR}(Integrated Resolution Utility), which captures how accurately the energies of gamma-ray showers are reconstructed across a batch of simulated events. The function is given by

$$U_{IR} = \frac{\sum_{k=1}^{N_{\text{batch}}} P_{\text{tr}}(k) w(k)}{\sum_{k=1}^{N_{\text{batch}}} P_{\text{tr}}(k) w(k) \frac{\hat{\sigma}_{E_k}}{E_{k,t}}}. \tag{2.9}$$

where $P_{tr}(k)$ is a binary indicator that is nonzero only for gamma-ray showers passing the trigger condition, σ_{E_k} is the estimated energy uncertainty for shower k, and $E_{k,t}$ is the corresponding true energy. The ratio expresses a weighted inverse average of the

relative energy uncertainty, such that the utility increases when the overall resolution improves. Notably, the denominator uses the relative uncertainty $\hat{\sigma}_{E_k}/E_{k,t}$, meaning that low- and high-energy showers are treated on equal footing by default. This design choice has a significant influence on the optimization outcome, as it avoids biasing the layout toward any specific energy scale unless explicitly introduced. Such biasing, if desired, can be introduced through the weighting factor $w(k)$ described in [87], but for this study we consider flat weighting across energies $w = 1$.

third component may be incorporated into the global utility function to capture the pointing resolution (PR) of the array—that is, the precision with which the incident direction of primary gamma rays can be reconstructed. This is particularly relevant for source localization and the detection of astrophysical anisotropies. We define the corresponding utility term U_{PR} based on the residuals between the true and reconstructed arrival angles (θ_k, ϕ_k) for each gamma-ray shower as:

$$U_{PR} = \frac{\sum_{k=1}^{N_{\text{batch}}} P_{\text{tr}}(k) w(k) \frac{\delta_{min}}{\Delta R}}{\sum_{k=1}^{N_{\text{batch}}} P_{\text{tr}}(k) w(k)} \tag{2.10}$$

where δR_k is the angular deviation of shower k, defined by

$$\Delta R = \sqrt{(\theta_t - \theta_m)^2 + (\phi_t - \phi_m)^2 + {\delta_{min}}^2}. \tag{2.11}$$

Here, $(\theta_{t,k}, \phi_{t,k})$ are the true polar and azimuthal angles of the incoming primary, and $(\theta_{m,k}, \phi_{m,k})$ their reconstructed counterparts. A small regularization term δ_{min} (set to 0.001) is added in quadrature to prevent instabilities in the denominator, particularly for showers that are reconstructed with near-perfect accuracy.

This formulation avoids the need for explicit uncertainty estimation —unlike the energy resolution utility U_{IR}— and instead uses directly measurable angular residuals. We find this approach to be more robust and easier to implement within the optimization pipeline, especially in early exploratory studies. Periodicity effects in angular differences are handled explicitly to ensure consistency across the full angular domain.

The three utility components just described can be employed individually or combined into a multi-objective utility function. In the latter case, suitable coefficients must be introduced to weigh their relative contributions according to the priorities of the experiment or the optimization goal. For the preliminary investigations presented here, which are primarily aimed at validating the software framework and exploring basic optimization behavior, we define a composite utility function as

$$U_1 = \eta_{GF} U_{GF} + \eta_{IR} U_{IR} + \eta_{PR} U_{PR}. \tag{2.12}$$

Without loss of generality, we set $\eta_{GF} = 1$, effectively normalizing the expression to the gamma flux utility. In practice, we found that setting $\eta_{IR} = 0.2$ and $\eta_{PR} = 0.0008$ yields a reasonable balance between the three components during optimization. These values were chosen heuristically to ensure that no single term dominates the others over the course of an optimization loop, allowing all three objectives to contribute meaningfully to the evolution of the layout.

Although this choice is arbitrary and not yet informed by a physics-driven weighting scheme, it serves the purpose of demonstrating the interplay between resolution and sensitivity within a unified optimization framework. Future studies may refine these weights based on specific science targets or operational constraints.

The derivative expressions for the utility terms are derived in Section 6.2 of [87].

An alternative utilty definition The definition of the composite utility function U_1 (Equation 2.6) relies on user-defined coefficients to specify the relative importance of its three components: gamma flux sensitivity, energy resolution, and pointing accuracy. While this flexibility allows the function to be tuned to different experimental priorities, it may also obscure a more unified perspective on array performance. For this reason, it is instructive to explore how a more integrated, experiment-wide utility function could be formulated—one that inherently balances these objectives without the need for arbitrary weight assignments. To assess the sensitivity of the array to isolated astrophysical gamma-ray sources, we define a point-source utility function U_{PS}, following the approach outlined in Ref. [87]. This function quantifies the statistical significance with which a gamma-ray excess may be detected within a signal window centered on the source's angular and energy coordinates, and is derived from a simplified on/off analysis with a surrounding sideband region used for background estimation. By enforcing a 5σ detection criterion, the utility is inversely proportional to the minimum number of source events required for discovery. U_{PS} thus serves as a proxy for directional sensitivity, incorporating per-shower angular and energy resolutions. Like U_{IR}, it is fully differentiable and contributes gradients with respect to detector positions by tracing how energy and direction estimation precision vary with geometry.

2.5 Optimization loop and main results

2.5.1 Optimization framework settings

The optimization framework developed in this study integrates algorithmic flexibility with physical realism to produce meaningful detector configurations under a variety of utility-driven objectives. A key feature of the implementation is the `CommonMode` parameter, which governs the collective motion of detector units during gradient descent. When `CommonMode=3`, triplets of detectors are constrained to move coherently while preserving a 120-degree angular symmetry about their centroid. This constraint ensures a triangular tiling symmetry and reduces the dimensionality of the optimization space, allowing more stable convergence in early tests. This mode is particularly well-suited for layouts built from macro-tanks—aggregates of 19 individual detectors arranged in a compact hexagonal pattern—as it allows the optimizer to operate in a reduced configuration space while preserving meaningful geometric flexibility. In this mode, the optimizer does not update individual unit positions arbitrarily but rather projects the gradient of each triplet onto radial and azimuthal directions, updating the positions according to the averaged directions within the triplet (see Figure 2.7). This averaging strategy has proven successful to reach faster convergence [87]. Because of this the results discussed in this Section are produced with `CommonMode=3` as default setting.

Configuration	$U_{GF,in}$	$U_{GF,fi}$
Wide random ball	927 ± 72	1297 ± 36
Tightly packed ball	727 ± 76	1235 ± 37
Two annuli	714 ± 69	1235 ± 66

Table 2.3: Values of the utility at the start of the optimization cycle $U_{GF,in}$ and at the end $U_{GF,fi}$, for the three configurations of Figure 2.8.

To regulate the learning dynamics, the framework employs a globally scheduled learning rate with exponential damping and periodic modulation, supplemented by a per-unit correction based on recent displacement directionality. Units that oscillate without making net progress are gradually damped, while those exhibiting coherent motion retain higher mobility. The displacement of each unit is capped to preserve smooth evolution and avoid spurious jumps driven by large gradients.

Physical realism is further enforced through post-update overlap resolution, which adjusts detector positions to respect finite sizes and mandatory spacing constraints. When macro-tank aggregates are used, as in most tests described here, the system accounts for the geometric footprint of 19 tightly packed units and guarantees minimal

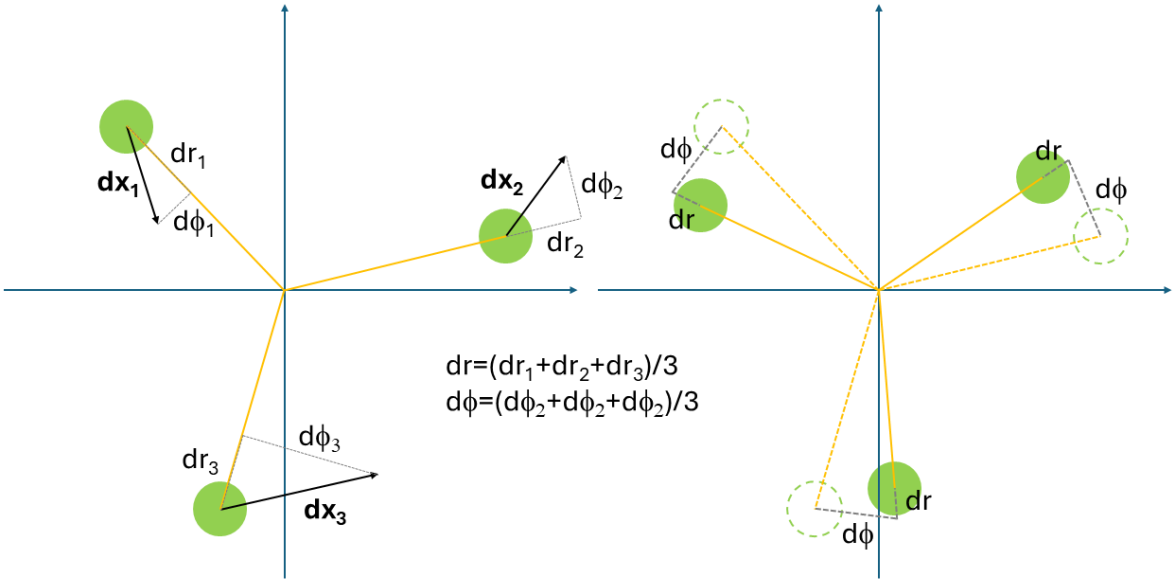


Figure 2.7: Illustration of the gradient update of a triplet of detectors when `CommonMode=3`. First the gradient is projected along the directions (left), then it is updated for the whole system averaging over the projections (right).

inter-aggregate spacing via analytic constraints. Boundary constraints are implemented for realistic site profiles, such as the Pampa La Bola plateau in the Chilean Andes, by generating repulsive forces near forbidden regions and projecting gradient updates accordingly (see [87] for more details).

2.5.2 Optimizer behavior and emergent patterns

Several qualitative behaviors are consistently observed across optimization runs:

Convergence toward symmetric configurations To verify the stability and convergence of the optimization process, we performed a controlled test in which the task was simplified by isolating the U_{GF} term in the utility function (see Equation 2.6), and generating showers orthogonally to the ground $(\theta = 0)$. This choice reduces the variability of shower patterns across a batch, effectively minimizing stochastic fluctuations in the gradient descent procedure. Energy was fixed at 1.0 PeV, and reconstruction was limited to core position inference, assuming perfect energy measurement. We initialized the layout with 36 macro-tanks (19 units each) in three distinct configurations: a wide random distribution, a tightly packed central ball, and a double annulus. The optimiza-

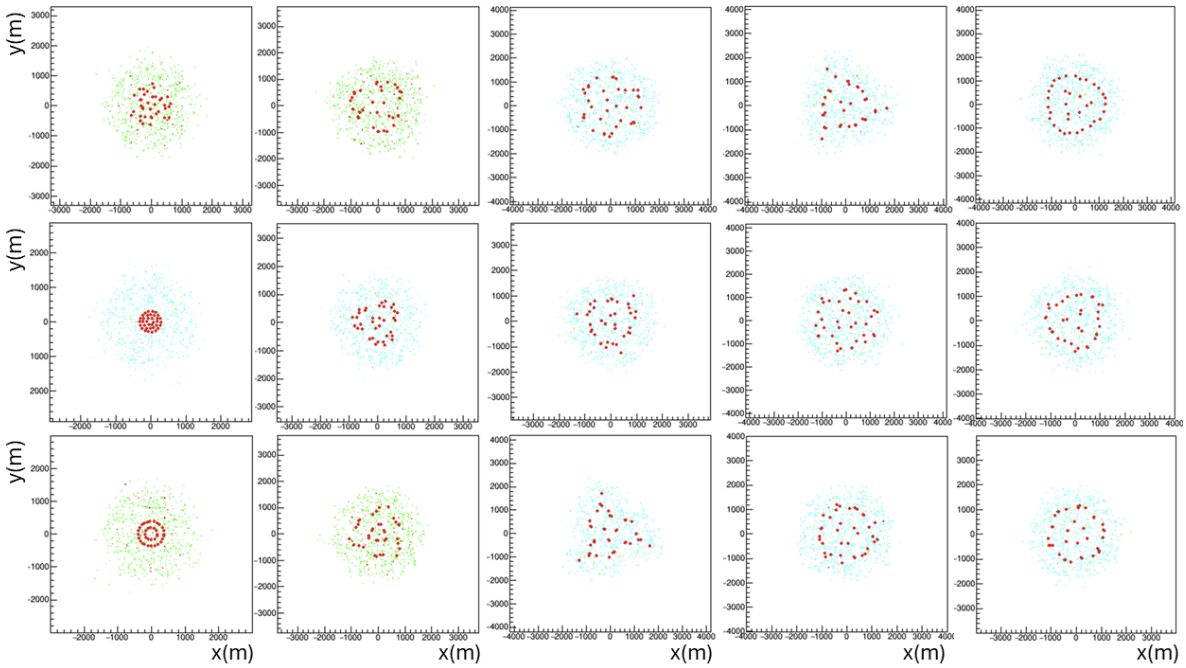


Figure 2.8: Caption

tion was run under CommonMode = 3, enforcing triangular symmetry across 12 triplets of tanks. With 3000 events per batch and 400 training epochs, each run required approximately one day on a single CPU. As shown in Figure **??**, all three initial layouts converge toward nearly identical triangle-symmetric configurations, confirming that the optimizer identifies consistent maxima of the utility landscape regardless of starting conditions. Final utility values also agree closely across the three runs (see Table 2.3), demonstrating both robustness and reproducibility of the method under these simplified conditions.

Expansion of the array : A striking feature observed in early optimization runs is the strong outward trend in the utility gradients when starting from a compact initial layout. In particular, the U_{GF} component of the utility function (Equation 2.6) consistently drives expansion, pushing detector units toward the periphery (see Figure 2.9, top). This behavior reflects the gain in sensitivity to gamma showers with cores falling outside the central region—events that still produce partial signals in the array and are valuable for distinguishing gamma from proton backgrounds. Notably, this outward pressure persists even in the absence of the $1/\rho$ factor, as wider coverage increases the likelihood of observing showers with larger lever arms, thereby enhancing the discrimination power in the two-component likelihood fit.

A similar outward effect is also induced by the point-source utility U_{PS}, although its

strength is more strongly dependent on the gamma-ray energy being considered. The U_{IR} term, which captures integrated energy resolution, likewise favors array expansion (see Figure 2.9, bottom), but its gradient profile varies more smoothly with radial distance.

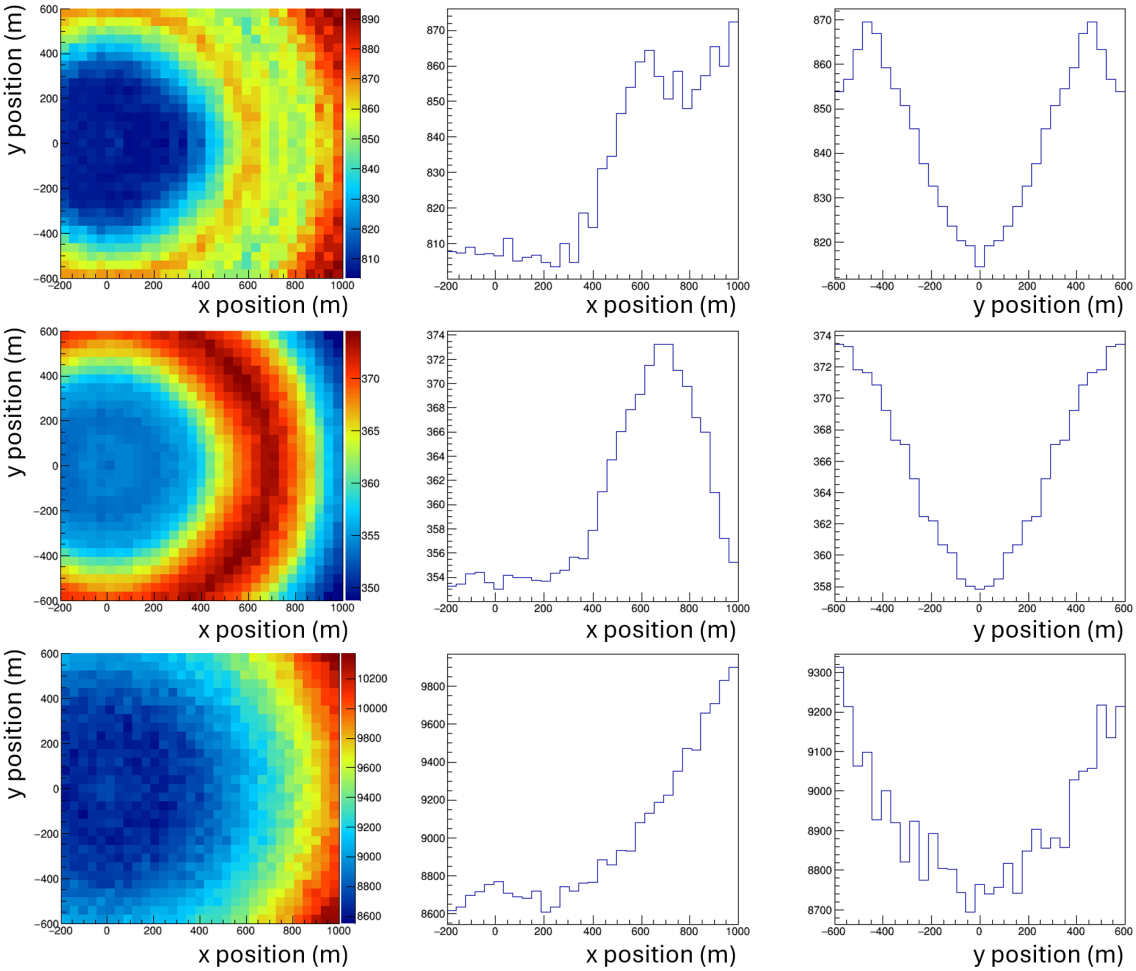


Figure 2.9: Top: temperature map (left) of the value of the U_{GF} utility term as a function of the position of a 19-unit aggregate detector complementing a circular array of 37 other macro-units arranged in a hexagonal pattern spanning a circular area of 400m radius. The center (right) histograms show the utility as a function of the x (resp., y) coordinate of the unit, for $y = 0$ (resp., $x = 0$). Center, left: temperature map of the value of U_{IR} utility term for the same array; the center and right histograms show the value of U_{IR} as a function of x and y as above. Showers used in these simulations have an energy $E = 1$ PeV. Bottom: the same graphs are shown for the U_{PS} utility, given a point source of 2 PeV energy at a polar angle of 45 degrees.

Comparing optimizations performed with only U_{PR} or U_{IR} reveals contrasting preferences (Figure 2.10). While both terms promote spread-out configurations, U_{IR} favors homogeneous distributions, whereas U_{PR} tends to generate structured, anisotropic lay-

outs that better capture the timing features of the particle front, especially under the constraint of a minimum number of triggering units.

These observations underscore the potential tension between competing optimization criteria —particularly between flux sensitivity and resolution terms— and reinforce the need for thoughtful utility composition. Assessing optimal layouts cannot rely on speculative reasoning alone but must reflect explicit scientific priorities encoded directly in the utility function.

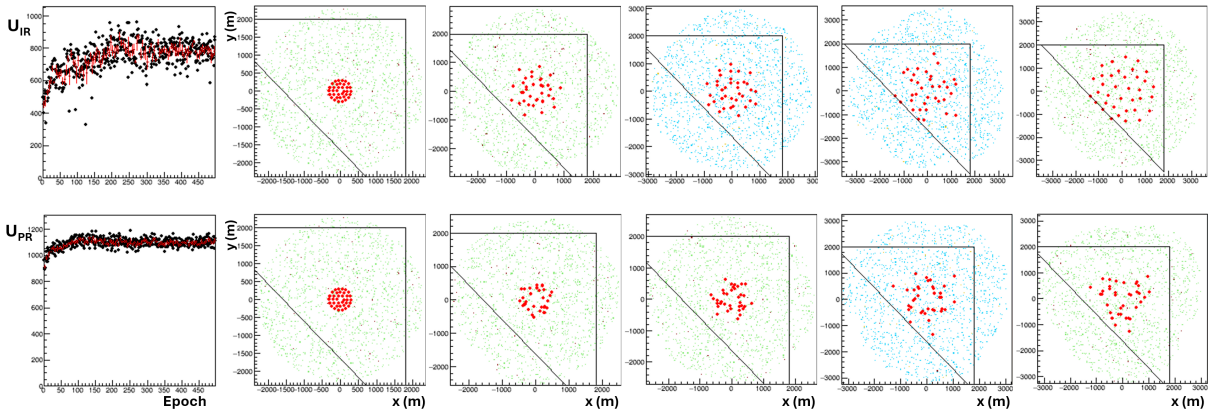


Figure 2.10: Comparison of the optimization of U_{IR} and U_{PR} terms, for an array of 36 macro-tanks comprising 19 units each. The initial array is set in a packed circle, centered within the triangular constraints of the Pampa la Bola site. The top left (bottom left) panel show the U_{IR} (U_{PR}) utility as a function of epoch. The five graphs to the right in each line show the progressive optimization of the U_{IR} (U_{PR}) terms at epoch 1, 20, 50, 100, and 500, respectively. The background green and cyan points show the center of generated showers that pass the triggering criterion.

Radial density of tanks A second recurring outcome of the optimization process—observed regardless of the initial array shape—is the systematic depopulation of the central region in radially symmetric configurations. In stark contrast with the dense cores typical of benchmark layouts (see Section 2.2), the optimizer consistently identifies no benefit in maintaining high central density. Instead, the gradients induce a shear effect that pushes central units outward more strongly than those at the periphery. This behavior, while distinct from the overall expansion trend discussed earlier, arises from the same underlying principle: the transverse scale of photon-induced showers imposes a natural dimensional structure on the problem.

The rationale is straightforward. The array center is a point of measure zero, and the probability of a shower core falling exactly there is negligible. Even for low-energy gamma rays—which generate more compact footprints—dense central tiling adds little

value. Given that shower core density grows quadratically with radial distance, distributing detectors outward leads to more statistically valuable sampling. The result is a preference for "donut-like" layouts over solid discs—a robust feature that consistently emerges from the utility gradients. Only in a regime of infinite integration time and exclusive focus on high-precision reconstruction of a few near-center events would a compact layout be justified. Under realistic conditions, an array that "looks outward" captures more useful information and ultimately improves sensitivity.

Symmetries In its default configuration, the optimizer enforces triangular symmetry by grouping detector units into equilateral triplets through a 120-degree rotational constraint (`CommonMode=3`). While this symmetry restricts the configuration space, it still allows for non-trivial layouts within each sector. The emergence of coherent, regular patterns after several optimization epochs signals that deterministic gradients are prevailing over stochastic fluctuations.

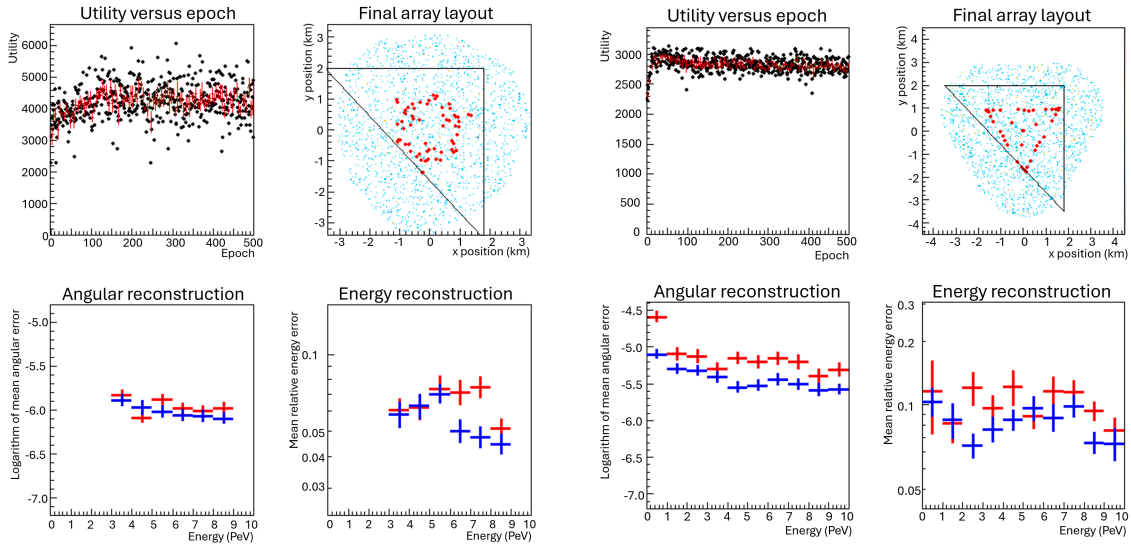
As shown in Figure 2.11b, a 500-epoch optimization run of the standard utility—starting from a densely packed circular layout of 63 macro-units and constrained within a strongly penalized area—converges to a highly symmetrical triangular configuration. In this case, flux sensitivity dominates the utility, driving units toward the outer boundary. Nonetheless, energy and pointing resolution also improve compared to the initial layout. A similar run maximizing U_{PS} at 6 PeV (Figure 2.11a) yields a richer but still symmetrical structure, highlighting the role of the chosen utility in shaping the final geometry.

Although convergence to symmetry is a sign of optimization stability, it does not guarantee global optimality. The utility landscape remains rich with local maxima, and escaping them may require more aggressive learning schedules—an investigation deferred to future, more refined studies. Among tested symmetry constraints, triangular patterns obtained with `CommonMode=3` remain particularly illustrative, offering insight into how structured solutions emerge from simple initial conditions.

Utility choice comparisons To assess the influence of different optimization targets, we performed three runs starting from the same initial configuration: a tightly packed circular layout of 330 macro-tanks (equivalent to 6270 individual units), using `CommonMode=3` to enforce triangular symmetry. The three utility definitions tested were:

1. the standard composite utility U_1 (Equation 2.12), with a flat shower energy spectrum from 100 TeV to 10 PeV;

2. the point-source utility U_{PS} evaluated at 2 PeV;



(a) Top left: utility versus epoch, with the center of triggering showers at the last epoch shown as cyan dots; top right: final layout (macro-units are the red points); bottom left: logarithm of the mean pointing error as a function of gamma energy, for the initial (red) and the optimized configuration (blue); bottom right: mean relative energy error as a function of gamma energy, for the initial (red) and the optimized configuration (blue).

(b) Top left: utility versus epoch; top right: final layout (macro-units are the red points), with in light cyan the center of triggering showers at the last epoch; bottom left: logarithm of the mean pointing error as a function of gamma energy, for the initial (red) and the optimized configuration (blue); bottom right: mean relative energy error as a function of gamma energy, for the initial (red) and the optimized configuration (blue).

Figure 2.11: Optimization of the U_{PS}(a) -with a point-source energy of $E = 6$ TeV- and U_1 (b) utility function for 61 19-unit macro-tanks in the Pampa la Bola site.

3. the same U_{PS}, but at 6 PeV.

As shown in Figure 2.12, several consistent features emerge. First, the U_1 optimization produces a layout where most units are pushed toward the array boundary, forming an equilateral triangle. This is driven by the restriction that showers must originate within 2 km of a detector: a triangle covers more sky under this constraint than a circular shape. The area penalty was intentionally strong in this run, but the flux-driven expansion still dominates, resulting in a final layout roughly 1.8 times larger than the initial one.

Second, the final layout under U_1 displays near-uniform filling within the allowed region, consistent with the behavior seen when maximizing the energy resolution utility U_{IR}. Third, all three runs show a clear depletion of the central region—an effect discussed previously—as the optimization favors configurations that better sample off-center showers. Lastly, both versions of the U_{PS} utility yield similar final layouts, with a larger central void for the 6 PeV case, reflecting the broader spatial footprint of higher-energy showers.

Table 2.4 reports utility gains ranging from 62% to 75%. While some of this increase

comes from larger effective area, the emergence of structured, non-uniform configurations plays an important role. Interestingly, energy resolution improves only modestly, while position resolution improves by 34–41%.

	U_{in}	U_{fi}	U_{fi}/U_{in}	$U_{GF,fi}/U_{GF,in}$	$U_{IR,fi}/U_{IR,in}$	$U_{PR,fi}/U_{PR,in}$
U_1	2178 ± 20	3572 ± 10	1.64 ± 0.02	2.20 ± 0.02	1.06 ± 0.02	1.34 ± 0.01
$U_{PS}\ (2PeV)$	1737 ± 52	2822 ± 25	1.62 ± 0.04	1.64 ± 0.04	1.07 ± 0.01	1.41 ± 0.01
$U_{PS}\ (6PeV)$	1732 ± 34	3071 ± 17	1.77 ± 0.04	1.50 ± 0.04	1.05 ± 0.01	1.40 ± 0.01

Table 2.4: Initial and final value of the utility for the three runs described in the text. The improvement in the utility from the initial packed circle configuration to the final ones shown in Figure 2.12 is shown for the separate factors making up the U_1 utility also for the runs maximizing the U_{PS} utility, for comparison.

For a rough comparison of the utility values with the benchmarks descrived in Section 2.2, we recreated the A7 benchmark using macro-tank equivalents matched in fill factor and radius (Figure 2.13) and evaluated it using the same utility definitions. As shown in Table 2.5, the optimized layouts outperform the A7-matched layout by 25.5 ± 0.5% in U_1, 4.6 ± 1.4% in U_{PS}(2 PeV), and 12.1 ± 1.0% in U_{PS}(6 PeV). These results highlight the potential of the optimization pipeline, even in its current form with several simplifications. A more detailed comparison of realistic layouts is left to future studies once model fidelity and reconstruction accuracy are improved.

	A7-like
U_1	2845 ± 9
$U_{PS}\ (E = 2PeV)$	2698 ± 27
$U_{PS}\ (E = 6PeV)$	2739 ± 19

Table 2.5: Utilities computed with the same running parameters of the optimization runs discussed in the text, for an array of 330 19-unit macro-tanks arranged to mimic the layout of the A7 configuration. Results for the final utilities of the optimized results are reported for comparison.

2.6 Conclusions and outlook

In the absence of assumptions about the angular distribution of gamma-ray sources, SWGO must be designed to observe showers arriving from all directions. This leads naturally to a preference for axially symmetric layouts, as implied by the $p(\theta) \propto \sin\theta$ distribution of arrival angles. However, with a finite number of detector units, perfect axial symmetry is unattainable, and discrete n-fold symmetries are the practical alter-

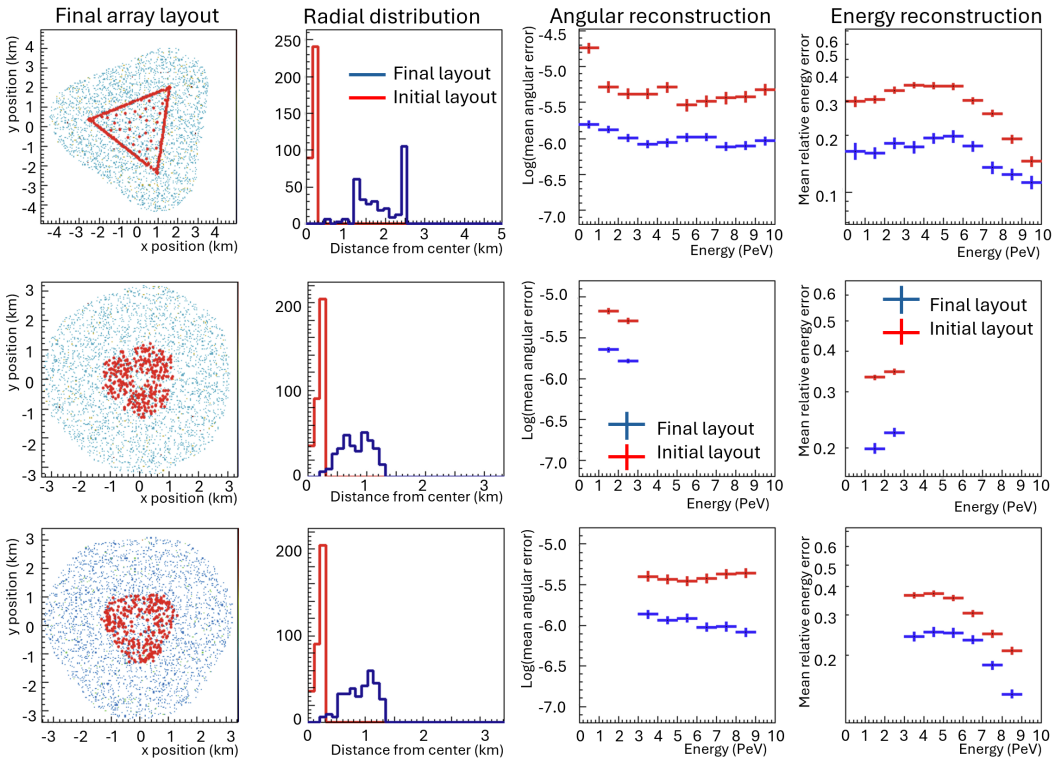


Figure 2.12: Final layouts, radial distribution, and relative pointing and energy errors achieved by optimization runs maximizing the U_1 utility of Eq. 2.6 (top row), the U_{PS} utility with a point source energy of 2 PeV (middle row), and the U_{PS} utility with a point source of 6 PeV (bottom row). The arrays, initially set in a tightly-packed configuration of 330 macro-units in a circle of 300 m radius, converge to different layouts after 2000 iterations.

native. While this has little impact at high detector counts, it imposes real constraints during optimization and provides a useful consistency check across solutions.

Our results show that different symmetric configurations often converge toward similar final layouts, though effects like triggering thresholds or site boundaries can introduce meaningful deviations. The specific formulation of the utility and reconstruction pipeline also plays a key role in guiding solutions. While this dependency might seem at odds with the notion of an ”end-to-end” approach, our tests so far use simplified but valid models intended to study algorithm behavior. A full optimization campaign will require finalized utility definitions, detailed detector modeling, and inclusion of real-world constraints—a task we reserve for future work.

While the presented optimization framework already yields meaningful and physically motivated layouts, several improvements are needed before it can serve as a robust tool for evaluating the scientific merit of alternative configurations. Below we outline key areas for future development:

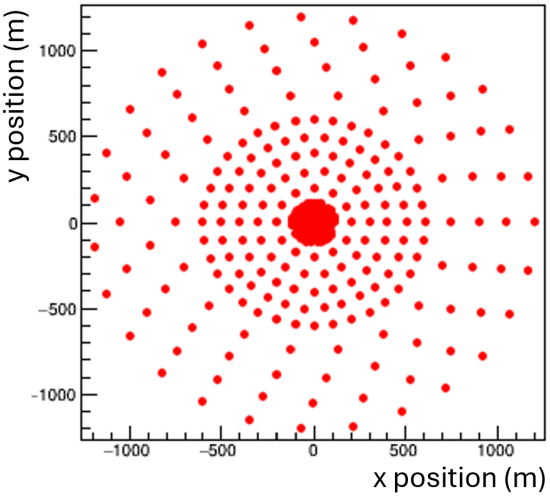


Figure 2.13: ”A7-like” layout of the 330 19-unit macro-tanks mimicking the fill factors of the A7 array described in the text.

- The current parametric model can be enhanced with more accurate treatments of primary angular distributions, spatial and temporal structure of the shower front, and the distinct signatures of secondary components. These refinements will directly impact the precision of reconstructed shower parameters.

- As the specifications of the detection units are finalized, the model should incorporate realistic angular and energy-dependent efficiency profiles, and account for measurement uncertainties at high fluence. The current approach, based on species-wise particle counting, is a placeholder pending these upgrades.

- The likelihood-based reconstruction used here is closely tied to the assumed model and can be improved in both speed and accuracy. In the longer term, we envision replacing it with a machine learning–based method, such as a deep neural network trained to perform both reconstruction and gamma/hadron discrimination. Development of such a model is underway.

- While stochastic gradient descent (SGD) is effective and well understood, it is not necessarily optimal for this class of problems. Challenges include inconsistent gradient magnitudes, reliance on fully parametric models, and limited robustness to model approximations. Alternatives worth exploring include genetic algorithms and sparsity-enforcing methods, such as Lasso regularization applied to an overpopulated starting array.

- Our current approach relies on analytically derived gradients, which offers trans-

parency and flexibility. However, as the model grows in complexity, automatic differentiation will become essential for scalability and maintainability, especially when optimizing over increasingly detailed detector response models.

Ultimately, the success of SWGO will depend on an interplay of choices—layout, detection units, site, and resource allocation—all of which should be evaluated in the context of maximizing scientific return under realistic constraints. Among these, the layout is especially suitable for early-stage isolated optimization, due to its relatively loose coupling with other design parameters. However, the full potential of layout optimization can only be realized once a baseline, experiment-wide utility function is agreed upon—one that synthesizes the scientific goals motivating SWGO as a unique addition to the global astroparticle physics program. The more clearly this utility is defined, the more impactful and actionable future optimization studies will become.

This chapter has presented a first working example of a differentiable optimization pipeline for astroparticle detector design, applied to the SWGO array layout problem. Despite the use of simplified models and surrogate reconstruction, the method already yields non-trivial and interpretable solutions, demonstrating its ability to balance competing physical objectives through a unified, gradient-based framework. Importantly, this approach remains flexible: additional constraints and improved modeling can be integrated progressively without altering the core optimization mechanism.

The remaining chapters of this thesis build directly on these concepts and extend them to a more complex setting: the calorimeter design for a future Muon Collider. There, we transition from a symbolic gradient engine to full TensorFlow-based implementations, leveraging modern tools from the machine learning ecosystem to couple detailed simulation with learnable geometry and reconstruction pipelines. This shift enables scalability to higher-dimensional problems and closer integration with data-driven surrogate models.

Chapter 3

Muon Collider in the bestiary of Future Colliders

The Standard Model has proven remarkably successful in describing fundamental interactions, but its limitations are now well established. The discovery of the Higgs boson at the LHC confirmed the last piece of the SM puzzle but raised new and deeper questions: Why is the Higgs so light, given the large corrections its mass receives from quantum loops? What is the nature of dark matter? Are there new forces or symmetries beyond those observed? These questions motivate a new generation of particle colliders.

The LHC, and its upcoming high-luminosity phase (HL-LHC), is expected to deliver 3 ab^{-1} at 14 TeV by the mid-2030s. This dataset will provide critical constraints on Higgs couplings and enable rare process searches. Yet, the intrinsic limitations of hadron collisions — from the steeply falling parton luminosities to the large QCD background and high pileup — restrict the reach of the HL-LHC to new physics at a few TeV, and limit the ultimate achievable precision.

This motivates the design of a new collider that can extend both the energy and precision frontiers. The current international roadmap includes a variety of options — linear and circular lepton colliders, higher energy hadron machines, and the Muon Collider. Each proposal represents a different balance between feasibility, cost, and physics potential. Figure 3.1 summarizes the timeline for the R&D and design work currently under way to establish the Muon Collider as a viable candidate.

As emphasized in the 2020 update of the European Strategy for Particle Physics [95], the muon collider was identified as a promising long-term avenue, provided feasibility could be demonstrated. In response, the International Muon Collider Collaboration (IMCC) was formed and has since produced detailed conceptual and

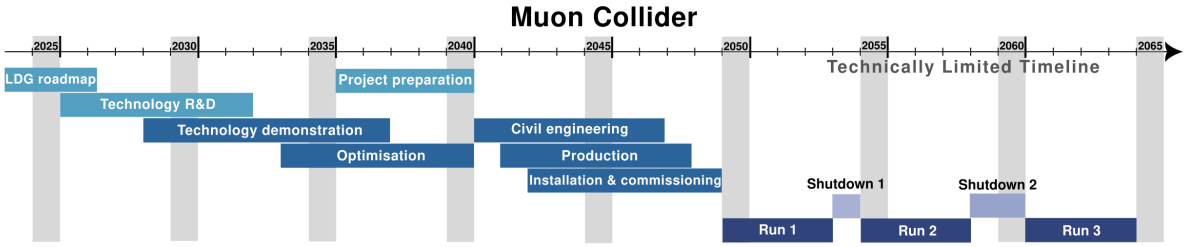


Figure 3.1: Proposed R&D and construction milestones needed to enable a first 3 TeV stage by 2050, assuming a successful demonstration of cooling, magnets, and detectors

performance studies, culminating in the ESPPU Muon Collider Report submitted for the 2026 strategy update[96].

Unlike electrons, muons radiate very little synchrotron energy, which allows their acceleration and collision in a compact circular collider. This makes it possible to achieve multi-TeV center-of-mass energies while retaining the clean environment typical of lepton collisions. The resulting machine offers access to both precision Higgs physics and direct production of new heavy states in a single infrastructure — a qualitative leap in design philosophy.

At the same time, the technical challenges of a Muon Collider are formidable. Chief among them is the production, capture, cooling, and rapid acceleration of muons, whose 2.2 µs lifetime demands an ultra-fast machine design. Furthermore, the decay of muons in the ring produces a dense flux of secondary particles — the beam-induced background (BIB) — which must be mitigated at the detector level through advanced shielding, timing, and reconstruction strategies.

A detailed staging strategy has been proposed, which foresees:

- An initial 3 TeV collider based on 11-T Nb_3Sn magnets;
- A later 10 TeV upgrade reusing most of the infrastructure but replacing the collider ring with a higher-field design;
- Optional luminosity staging strategies that frontload energy but at reduced event rates.

These options are discussed in the ESPPU report, where two complementary approaches — energy staging and luminosity staging — are evaluated with respect to feasibility, risk, and timescale.

The Muon Collider is not just an accelerator proposal; it is a long-term vision for a versatile facility capable of supporting a comprehensive program of electroweak precision studies, rare decay searches, and new particle discovery. Its emergence in the

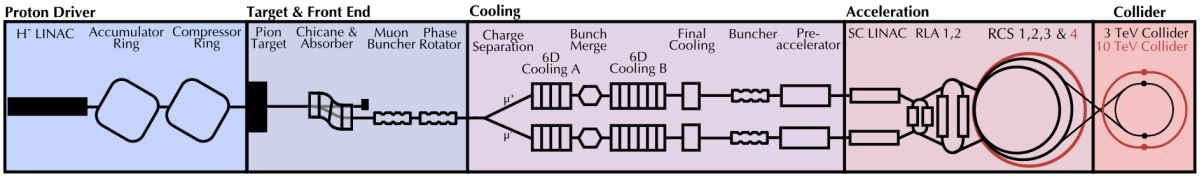


Figure 3.2: Schematic layout of the Muon Collider system. From [96]

strategic debate reflects a growing consensus that new design philosophies — and not just bigger versions of existing colliders — will be needed to push beyond the current frontiers.

3.1 The Muon Collider Experimental Landscape

Muon beam production and beamline concept

The Muon Collider design concept stems from the development studies of the Muon Accelerator Programme (MAP) [97]. After 2017 the International Muon Collider Collaboration is progressing the design, of which Figure 3.2 presents a schematic layout.

The beamline itself consists of

1. A *Proton driver*, proving an initial high-intensity and short-pulsed beam of protons;
2. A *target* for the proton beam, to produce pions that are later driven into the decay channel. Here a muon beam is formed thanks to a buncher and phase rotator;
3. A *cooling* stage, consisting of several steps of absorbers and RF cavities in high magnetic field. This is crucial to reduce longitudinal and transverse emittance of the beam;
4. An *accelerating* stage, with a linac, two recirculating linacs and a sequence of high-energy accelerating rings. This system brings the beam energy up to 5 TeV;
5. a final collider ring, where the muons actually circulate and collisions are produced.

Table 3.1 list a set of target parameters for 3 TeV and 10 TeV center-of-mass collision scenarios. Note that as such their purpose is to set an objective to drive technology research and explore design limits. However, if we assume those parameters are fully met, the targeted integrated luminosity can be reached within five years of continuous

Parameter	Symbol	Unit	Scenario 1		Scenario 2	
			Stage 1	Stage 2	Stage 1	Stage 2
Center-of-mass energy	E_{cm}	TeV	3	10	10	10
Target integrated luminosity	$\int \mathcal{L}_{target}$	ab^{-1}	1	10	10	10
Estimated luminosity	$\mathcal{L}_{estimated}$	10^{34}cm^{-2}s^{-1}	2.1	21	5(tbc)	14
Collider circumference	C_{coll}	km	4.5	10	15	15
Collider arc peak field	B_{arc}	T	11	16	11	11
Luminosity lifetime	N_{turn}	turns	1039	1558	1040	1040
Muons/bunch	N	10^{12}	2.2	1.8	1.8	1.8
Repetition rate	f_r	Hz	5	5	5	5
Beam power	P_{coll}	MW	5.3	14.4	14.4	14.4
RMS longitudinal emittance	$\epsilon_{\parallel}$	eV	0.025	0.025	0.025	0.025
RMS transverse emittance	$\epsilon_{\perp}$	μm	25	25	25	25
IP bunch length	σ_z	mm	5	1.5	tbc	1.5
IP beta function	β	mm	5	1.5	tbc	1.5
IP beam stile	σ	μm	3	0.9	tbc	0.9
Protons on target/bunch	N_p	10^{14}	5	5	5	5
Proton energy in target	E_p	GeV	5	5	5	5

Table 3.1: Tentative target parameters for a Muon Collider at different energies. Adapted from [96]

run at full luminosity. This approach allows for flexibility in design and a realistic luminosity ramp-up, accommodating ongoing technology developments. An additional advantage is the possibility of earlier implementation through initial stages operating at reduced luminosity, while retaining the full physics potential through subsequent upgrades.

Site and timescale

The Muon Collider Collaboration (MCC) is developing a site-independent design concept, allowing deployment at multiple potential host laboratories. At present, two specific implementation studies are underway: one at CERN and one at Fermilab. These proposed layouts are shown in Figure 3.3. While both options aim to leverage existing infrastructure—such as the SPS and LHC tunnels at CERN—the collider design remains modular and does not fundamentally depend on legacy components.

At CERN, civil engineering studies have confirmed the feasibility of constructing the surface installations entirely within CERN-owned land. The SPS and LHC tunnels could be reused to host the lower-energy accelerator rings, minimizing excavation requirements. In this configuration, the proton driver would be located on the Meyrin site.

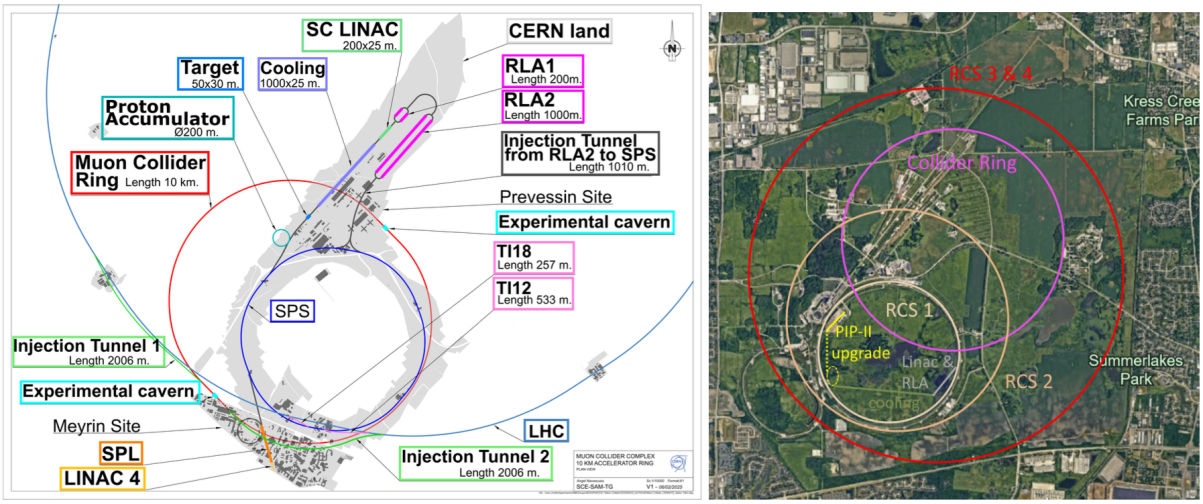


Figure 3.3: Plan of potential beamline implementations at CERN (Left) and Fermilab (Right). From [96].

The proton beam would travel through the SPS tunnel to the Prevessin campus, where cooling and initial linacs would be installed in new cut-and-cover tunnels. The acceleration sequence would involve injection into the SPS, then into the LHC, and finally into a new 10 km collider ring.

This baseline configuration could enable collisions at a center-of-mass energy of 7.6 TeV using existing magnet technology. Further upgrades are under consideration, such as installing high-field HTS fast-ramping magnets in the final Rapid Cycling Synchrotron (RCS) stage, or implementing hybrid RCS modules in both the SPS and LHC tunnels. These solutions would extend the energy reach of the collider while maintaining compatibility with the CERN accelerator complex. The possibility of operating the SPS concurrently with the Muon Collider is also under evaluation.

A similar siting study is being conducted for Fermilab, examining how the existing and planned infrastructure can be incorporated into the collider complex. A preliminary layout, designed to fit within the boundaries of the Fermilab site, is shown in Figure 3.3 (right). The exact parameters of a Fermilab-based collider remain under refinement, pending the outcomes of ongoing integration and feasibility studies.

Muon decays produce high-energy neutrinos that propagate through the earth and may emerge far from the accelerator complex. This feature enables a novel experimental program, well beyond the reach of current initiatives such as FASERv. In such a setup, neutrino detectors would be installed at the surface where the beam exits the ground. A suitable implementation for this concept has already been identified at CERN.

At the same time, it is essential to ensure that neutrinos produced throughout the col-

lider infrastructure do not generate unacceptable levels of off-site radiation. To mitigate this, a system of magnet movers is foreseen in the collider arcs, allowing for vertical displacement of the magnets to dilute the neutrino flux. Preliminary studies suggest that this strategy can reduce the environmental impact to negligible levels—comparable to that of the LHC. Further optimization, along with a complete assessment of the full facility, is ongoing.

The ability of a Muon Collider to achieve 10 TeV parton-parton collisions with high luminosity makes it an exceptional opportunity for both near- and long-term exploration. Its development must be guided by a commitment to environmental responsibility and by strategic staging scenarios. Key enabling technologies—ranging from high-temperature superconducting magnets to machine learning and AI-based control systems—should be advanced in synergy with broader technological trends. Together with strong international collaboration, these elements will be essential to deliver a technically ambitious yet practical facility for the next era of particle physics.

Not all implementation scenarios can be pursued in parallel, and the International Muon Collider Collaboration (IMCC) has outlined the timeline in Figure 3.1 aimed at enabling first operations by 2050. Achieving this objective requires that all critical technologies reach maturity within the next 15 years. The proposed schedule is technically driven, assuming both successful R&D and adequate funding throughout the development phases.

While optimistic, this success-oriented timeline provides a useful reference for the community and for decision makers. It defines a clear target for coordinating technological development, guiding strategic planning, and framing the resource commitments required to realise a Muon Collider within a realistic timeframe.

Current timeline assessments have identified three key technological developments that define the minimum time needed to realise a Muon Collider:

- Muon production and cooling, including a demonstrator test facility. With sustained support, this technology could reach maturity within 15 years, enabling an informed decision on implementation.

- Magnet systems, in particular those based on high-temperature superconductors (HTS). The required HTS solenoids for the production and cooling stages, as well as fast-ramping magnets, are expected to be available within the same timeframe. For the collider ring, 11 T Nb_3Sn magnets with 16 cm aperture are projected to be feasible, though more advanced HTS or hybrid designs may require a longer development cycle.

- Detector technologies, particularly those impacting background rejection and measurement precision. These are also expected to reach readiness within 15 years.

Ensuring adequate funding across all R&D lines will be essential to prevent auxiliary systems or less mature technologies from becoming bottlenecks in the overall timeline.

A staged implementation strategy based on Nb_3Sn collider ring magnets could enable first operations of a Muon Collider by 2050. Two main approaches are currently under study (see Table 3.1), each offering different trade-offs in cost, timeline, and physics scope.

- *Energy staging* involves constructing an initial collider stage at lower energy, such as 3 TeV, where a strong physics program already exists. This option substantially reduces the cost of the first phase and may accelerate decision-making. The 3 TeV design is compatible with 11 T Nb_3Sn dipoles. In the second stage, the full complex is reused—except for the collider ring, which is replaced along with the addition of a Rapid Cycling Synchrotron (RCS) to achieve higher beam energies;

- *Luminosity staging* retains the full 10 TeV center-of-mass energy from the outset but uses less performant magnets in the collider ring, reducing luminosity. Replacing 16 T HTS dipoles with 11 T Nb_3Sn magnets increases the ring circumference and decreases luminosity by approximately a factor of three, compounded by limitations in the interaction region optics. While future upgrades to the interaction region magnets can partially restore performance, the rest of the collider ring is unlikely to be replaced. This scenario requires most of the full project cost upfront, which may influence the timeline and funding trajectory

The choice between energy and luminosity staging will depend on physics priorities, technological readiness, and funding availability. Accelerated progress in HTS magnet development, combined with strong financial support, could make a direct path to the 10 TeV collider more attractive and feasible within the desired timeframe.

3.1.1 Machine-detector Interface

Beam-induced background (BIB) represents one of the principal challenges for detector performance at a multi-TeV Muon Collider. The dominant source of background is the decay of circulating muons, which produce a continuous flux of secondary particles. Additional contributions arise from incoherent electron–positron pair production—triggered

by real or virtual photon interactions between the counter-rotating muon bunches—as well as from possible beam halo losses on machine apertures.

The resulting radiation field is complex and mixed. It includes electrons, positrons, and photons, as well as hadrons from photo-nuclear interactions and muons produced via Bethe–Heitler pair production. These secondary particles interact with the detector and surrounding materials, significantly increasing hit occupancy and radiation levels.

In the absence of dedicated mitigation strategies, BIB would compromise the reconstruction of physics events and lead to unacceptable radiation damage in sensitive detector components. Addressing this issue is a core element of both machine–detector interface design and detector technology development.

An optimised machine–detector interface (MDI) is essential to minimise the impact of beam-induced background on detector performance and to control radiation damage in sensitive components. The MDI must incorporate a carefully engineered absorber system, including masks within the final focus region and a conical shielding structure that extends deep into the detector volume—reaching within a few centimetres of the interaction point.

The design and optimisation of these elements must be fully integrated with the detector layout and interaction region optics. A conceptual MDI configuration for a 1.5 TeV Muon Collider was originally developed by the MAP collaboration using the MARS Monte Carlo code [98]–[100]. This served as a starting point for higher-energy designs developed within the IMCC. Studies for 3 TeV and 10 TeV configurations were carried out using the FLUKA particle transport framework [101]–[103], with the 10 TeV design based on a newly developed interaction region lattice [104]. The 3 TeV studies remain anchored to the MAP optics [105].

As demonstrated in earlier studies by the MAP collaboration at 1.5 TeV [106], the flux of secondary particles entering the detector can be reduced by several orders of magnitude through the use of massive absorbers placed close to the interaction point. The central element of this shielding system is a nozzle-shaped absorber, which defines the detector's inner envelope and sets its minimum angular acceptance—10° in the MAP design. The nozzle's geometry and material composition determine both the rate and spectrum of particles reaching the detector. It must be constructed from a high-Z material to suppress electromagnetic showers from muon decays and beam halo losses. To reduce the neutron flux arising from photo-nuclear interactions, the design includes a layer of borated polyethylene or an equivalent neutron-moderating material.

Beyond shielding efficiency, the nozzle presents substantial engineering challenges. These include the design of mechanical support and alignment systems, ther-

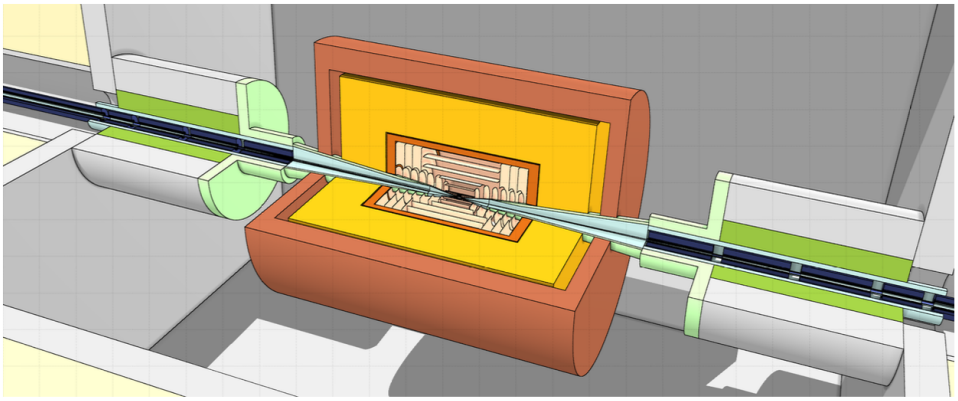


Figure 3.4: FLUKA rendering of the detector systems and interaction region. From [96]

mal management to handle energy deposition from beam-related backgrounds, and integration with beam instrumentation and vacuum components. The final design must also be optimised for different center-of-mass energies and tailored to the specific constraints of the detector environment.

While the particle spectra of the beam-induced background are largely determined by the shielding nozzle, the total background flux is also influenced by the layout of the interaction region (IR)—in particular by the lattice design and the strategic placement of mask-like absorbers within the IR magnets. Previous MAP studies at 1.5 TeV [105] showed that introducing a dipolar component in the final focus can help reduce the flux of secondary particles entering the detector.

Optimising the IR lattice is one of the central challenges in collider design, especially at 10 TeV. Earlier layouts at 3 TeV and 6 TeV were developed within MAP [107], although background simulations for those configurations have not been published. Beyond background mitigation, the IR must meet stringent beam optics requirements. At 10 TeV, the design calls for a very small $\beta^*$of 1.5 mm at the interaction point, corresponding to a highly focused beam and resulting in significant chromatic aberrations. This in turn leads to large β-functions in the final focus region, requiring quadrupoles with large apertures and high field gradients. To protect these magnets from the intense secondary radiation field, a few centimetres of internal shielding must be included to mitigate heat loads and limit cumulative radiation damage from muon decays.

Even with an optimised machine–detector interface and interaction region design - a rendering of which is presented in Figure 3.4-, effective background mitigation also relies on detector technology choices and reconstruction strategies. Background particles often display characteristics that distinguish them from signal events—for example, delayed arrival times relative to the bunch crossing or non-collision-like incident angles.

Ongoing efforts are being carried to further optimize the nozzle structure, also with the aid of automatic differentiation techniques. The most recent achievements are presented in detail in Section 3.2 of [96].

3.1.2 Detector systems

The development of dedicated detectors for a Muon Collider has rapidly progressed in recent years, fostering an active design environment and leading to a first generation of technically grounded concepts. Although detector design remains in its early stages, robust conclusions can already be drawn regarding performance targets, technological requirements, and future opportunities for optimisation. A central challenge in all detector concepts is the mitigation of beam-induced backgrounds (BIB) arising from muon decays in the collider ring. These backgrounds are typically addressed through the use of shielding elements placed close to the beamline, which in turn reduce angular acceptance. Balancing background suppression with acceptance and resolution performance requires a careful, system-level optimization.

Two detector concepts have been proposed to instrument the collider's two interaction points: MUSIC (MUon System for Interesting Collisions) and MAIA (Muon Accelerator Instrumented Apparatus)[108]. Both follow a similar structural blueprint, consisting of a cylindrical detector 11.4 m long and 12.8 m in diameter. The primary subsystems include a central tracking detector, electromagnetic and hadronic calorimeters (ECAL and HCAL), and an outer muon identification system. A superconducting solenoid provides the magnetic field required for momentum measurement in the tracker.

In this work, we focus on the MUSIC concept. MUSIC builds on the CLICdet layout [109], adapted to the specific needs of a Muon Collider operating at $\sqrt{s} = 3$ TeV. A full simulation study [110] has demonstrated that, under realistic BIB conditions, the MUSIC design supports the core physics programme of the collider and achieves performances compatible with precision Higgs and electroweak measurements. This design serves as the current baseline for full simulation studies at both 3 TeV and 10 TeV. The modularity of the MUSIC design enables flexible adaptation to future staging scenarios while maintaining compatibility with high-precision measurements and robust BIB mitigation strategies.

While further optimisation of the detector design and reconstruction algorithms is still ongoing—particularly to achieve the precision targets required at high energy—recent efforts have extended the original 3 TeV detector studies to the 10 TeV regime. These developments build on the same foundational layout, adapted to meet the increased demands of higher beam energy and background rates.

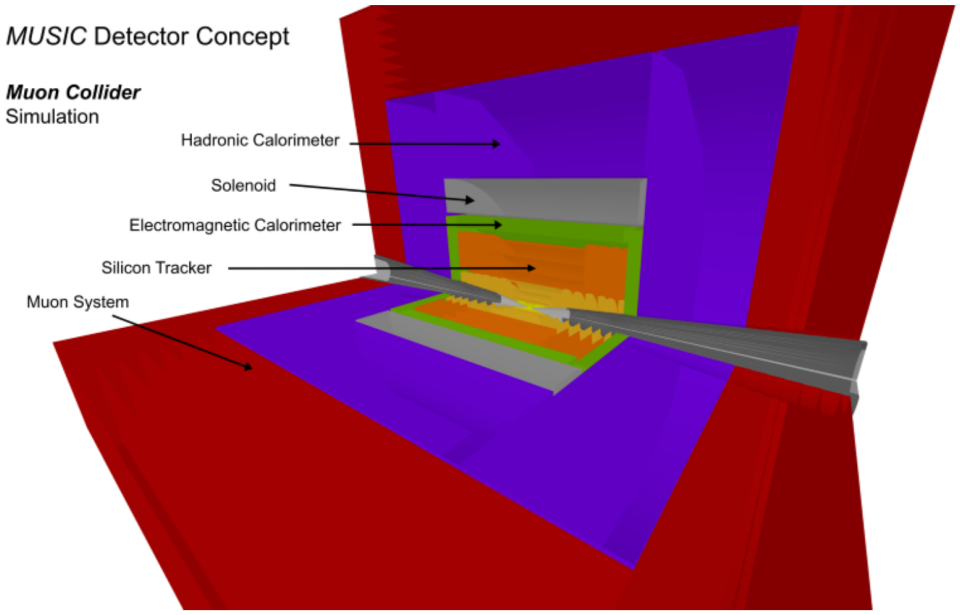


Figure 3.5: MUSIC detector concept

MUSIC The layout of the MUSIC detector is shown in Figure 3.5 In the following, we summarize the main features of its sub-detectors, beginning with the central tracking system.

The tracker is composed of three main subsystems covering a polar angle range from 10° to 170°: the Vertex Detector (VXD), the Inner Tracker (IT), and the Outer Tracker (OT). All three systems follow a cylindrical geometry in the barrel region and are complemented by end-cap disks placed perpendicular to the beam axis in the forward and backward regions.

The VXD is based on silicon pixel planar modules, arranged in five barrel layers of 26 cm length, with radii spanning from 2.9 cm to 10.1 cm. Each end-cap includes four disks placed between $|z|$ =18 and 36.6cm from the interaction point. The entire system features $25 \times 25 \mu$m^2 silicon pixels, delivering a spatial resolution of 5µm×5µm and a time resolution of 30 ps per hit.

The Inner Tracker consists of three barrel layers, located at radii between 16.4 cm and 55.4 cm, and seven disks per side at $|z|$ positions ranging from 60.4 cm to 219.0 cm. The first barrel layer extends over 96.3 cm in z, while the second and third reach 138.5 cm.

The Outer Tracker comprises three barrel layers, each 252.8 cm long, placed at radii between 81.9 cm and 148.6 cm. Each end-cap includes four disks from $|z| = 141.0$cm to 219.0 cm. The IT and OT subsystems employ $50 \times 1000 \mu$m^2 macropixels, offering spatial resolutions of 7 µm $\times$ 90 µm and a timing resolution of 60 ps.

The MUSIC calorimeter system is composed of an electromagnetic calorimeter (ECAL) positioned inside the solenoid, and a hadronic calorimeter (HCAL) located outside the magnet bore. Together, the calorimeters provide hermetic coverage over a polar angle range from approximately 7° to 173°, ensuring full containment for electromagnetic and hadronic showers.

The ECAL consists of a 4.4 m long central barrel with an inner radius of 1.69 m, closed by two end-caps positioned at $|z|$ =2.3 m. Both the barrel and end-caps are 27 cm thick. The system adopts a semi-homogeneous crystal design (CRILIN), using $10 \times 10 \times 45$mm^2 lead-fluorite crystals segmented longitudinally in six layers, corresponding to a total of 26.5 radiation lengths. This design enables high-resolution reconstruction of electrons and photons.

The HCAL surrounds the ECAL with a 5m long barrel at an inner radius of 2.9m, complemented by two end-caps at $|z|$ =2.58 m, each 1.86m thick. It is an iron–scintillator sampling calorimeter composed of 70 alternating layers of 20mm iron and 30×30mm^2 scintillator tiles, providing roughly seven nuclear interaction lengths. The iron structure also serves as the magnetic return yoke.

Beyond the HCAL, the muon system includes seven barrel layers and six end-cap layers, covering the same polar angle range. Although a specific technology has not yet been selected, the role of these detectors is to identify muons, as no magnetic field is present outside the calorimeters.

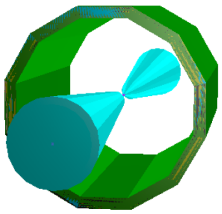
Crilin The challenges posed by the Beam-Induced Background (BIB) at a Muon Collider demand a calorimeter capable of high-resolution energy reconstruction in a radiation-dense, time-structured environment. To meet these requirements, the Crilin geometry [111], [112] was developed as a voxel-based electromagnetic calorimeter (ECAL) design optimized for timing, granularity, and ML-readiness.

The Crilin setup consists of modular arrays of PbF_2 crystals, each shaped into $1 \times 1 \times 4.5$ cm^3 voxels. These dimensions were selected to balance compactness, light yield, and longitudinal sampling. PbF_2 was chosen due to its fast Cherenkov response and radiation hardness, critical for operation in a BIB-heavy environment.

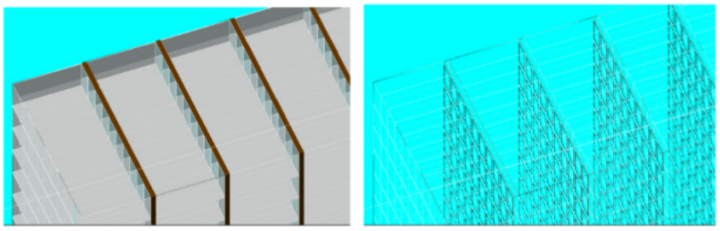
The ECAL barrel is arranged in a dodecahedral layout, with each edge composed of five layers of 2D voxel arrays. This structure provides:

- Longitudinal segmentation for improved shower profiling;
- Regular voxel tiling to support CNN- and GNN-based reconstruction;
- Mechanical modularity, simplifying construction and cooling integration.

A rendering of the Crilin barrel and edge structure is provided in Figure 3.6.



(a) ECAL barrel with nozzle.



(b) Voxel structure at calorimeter edge.

Figure 3.6: Visualization of the Crilin ECAL geometry: full barrel layout (a), and detailed voxel segmentation (b) [111].

Compared to traditional sampling calorimeters, Crilin offers:

- Lower Molière radius (owing to dense PbF_2);
- Sharper angular resolution due to voxel regularity;
- Compatibility with fine-grained ML reconstruction methods such as Object Condensation and Graph Neural Networks.

The BIB rejection efficiency is particularly notable: even under full overlay at 1.5 TeV $\sqrt{s}$, photon signals as low as 25 GeV retain reconstructibility with $\Delta E/E \leq 25\%$, with minimal contamination from background [111].

While Crilin offers a high-performance baseline, its uniform voxel tiling presents challenges in regions with inhomogeneous BIB deposition. Energy deposits are more concentrated along certain layers and geometrical boundaries — hinting that a spatially adaptive voxel configuration could further improve signal-to-noise ratios. This observation motivates the study of end-to-end differentiable optimization techniques, which are introduced in the following chapters to refine the Crilin layout beyond uniform segmentation.

3.2 Physics benchmarks

We conclude the discussion about the Muon Collider development by showcasing a selection for benchmark results in terms of potential for new physics discoveries and exclusions. Again, most of the material presented has been prepared for the submission for the 2026 European Strategy Update [96], which discusses in larger details these and many other physics cases.

The Standard Model not only accounts for the masses of the electroweak gauge bosons but also for those of quarks and leptons, embedding the Higgs mechanism in a framework of remarkable predictive power. As a consequence, electroweak interactions can be described up to energy scales vastly beyond those currently probed, with no internal inconsistencies.

Yet the very nature of the Higgs boson raises deep theoretical questions. Unlike fermions and vector bosons, a fundamental scalar field is not protected from large quantum corrections. The fact that the Higgs mass appears stable and light—in spite of these corrections—poses a serious challenge to the naturalness principle as formulated within the Wilsonian effective field theory framework. This tension, crystallized in the so-called Naturalness Paradox, continues to shape the agenda for both experimental and theoretical exploration beyond the Standard Model.

The document submitted by the International Muon Collider Collaboration for the 20206 update of the European Strategy of Particle Physics [96], on which this chapter is largely based, highlights four tracks to explore in terms of physics benchmark results to assess the performance of the machinery:

- *Energy track*, thanks to the high center-of-mass energy of the collider, and the absence of partons absorbing fractions of the total collision energy;
- *Precision track*, motivated by the high cross section for electroweak processes;
- *Precision from energy track*, triggered by the linear scaling in luminosity with energy, allowing for high precision measurements also at multi-TeV scales;
- *Muon and neutrinos track*, a novel direction opened by the availability of multi-TeV muon beams for the first time in history, as well as the possibility to use intense neutrino beams in parasite experiments.

In this section we tackle the main physics benchmarks of the Muon Collider, summarizing the theoretical challenges and future requirements, and illustrating the four highlighted paths. A more detailed discussion is presented in [96, part V].

New Standard Model physics at the Muon Collider

While current technologies allow for the construction of low-energy electron-positron colliders capable of measuring Higgs and electroweak observables with remarkable precision, they fall short of probing the dynamics that underpin the very mechanism of electroweak symmetry breaking. Precision alone is insufficient to access the regime

where the spontaneous breaking of a fundamental gauge symmetry becomes manifest in the structure of scattering amplitudes. Such a regime lies at high energy, where the interactions among electroweak bosons reveal their non-Abelian origin and the role of the Higgs as the unitarizing agent of the theory.

To enter this domain, one requires not only high energies, but also the clarity of leptonic initial states—free from the partonic uncertainties and QCD backgrounds that characterize hadron colliders. Among the known leptons, muons offer a uniquely promising path. Their higher mass suppresses synchrotron radiation, making it possible to construct a circular lepton collider with multi-TeV reach. Accessing the high-energy regime of electroweak-Higgs unification—where the symmetry structure of the Standard Model can be tested in its most direct form—constitutes the central and most distinctive physics motivation for the Muon Collider programme.

As the center-of-mass energy rises well above the electroweak scale, the Standard Model predicts a series of increasingly distinctive phenomena. Among the most critical is the unitarization of electroweak boson scattering, where the virtual exchange of the Higgs boson cancels the energy-growing contributions from longitudinal vector boson exchange. This mechanism is not merely a feature of the theory—it is its defining prediction at high energies, ensuring the consistency of the Electroweak-Higgs framework.

At the LHC, accessing this regime is challenging: large QCD backgrounds and limited phase space at high invariant masses obscure the signal. A muon collider, by contrast, offers a clean environment and enhanced collinear vector boson emission from energetic muons, enabling precise measurements of vector boson scattering processes in the multi-TeV range. The confirmation of unitarization thus becomes experimentally accessible, anchoring one of the most fundamental aspects of electroweak symmetry breaking.

The presence of massive vector bosons inside the massless muon at high energies is itself a striking manifestation of broken gauge symmetry—an effect the Muon Collider is uniquely positioned to establish experimentally. These processes reveal the “quantum compositeness” of elementary particles: not as classically structured objects, but as quantum states composed of electroweak partons. Unlike QCD, electroweak theory is perturbative and features a mass gap, allowing particles to be cleanly defined as asymptotic states and their dynamics to be computed reliably.

At 10 TeV, the Muon Collider will probe the internal electroweak structure of leptons and establish high-precision tests of symmetry restoration, including the Goldstone Boson Equivalence Theorem, which links the Higgs field to the longitudinal modes of the W and Z bosons.

The electroweak structure of the muon also includes its neutrino content—a direct consequence of SU(2) unification—which enables unique access to neutrino-neutrino and neutrino-muon collisions. These can be observed both via electroweak parton interactions and through forward detectors placed to intercept neutrinos emitted in muon decays, enabling the study of collider neutrinos at high energies for the first time. Finally, electroweak radiation at 10 TeV will expose the partonic structure of neutrinos themselves, including their effective W and electron content, with the striking possibility of producing observable “neutrino jets”—a qualitatively new collider signature

Physics beyond the Standard Model at the Muon Collider

The observation of novel Standard Model phenomena is a guaranteed outcome of a Muon Collider, but so too is its sensitivity to physics beyond the Standard Model (BSM). In many cases, the same measurements that reveal new electroweak dynamics also probe BSM scenarios, offering both discovery potential and model-independent constraints.

The relatively high effective vector boson content of the muon enables searches for new interactions in high-energy boson scattering, and for BSM states coupled via Higgs or gauge portals. Such scenarios address central questions about dark matter, naturalness, and the electroweak phase transition. These same vector bosons are responsible for the large Higgs production rate at a Muon Collider—both single and pair production—leading to per-mille precision on several Higgs couplings, and few-percent sensitivity to the trilinear self-coupling. Overall, the collider achieves a level of precision on Higgs properties that is comparable to low-energy electron-positron Higgs factories, while providing complementary observables and direct access to the self-coupling at high energy.

Beyond indirect effects on Higgs couplings, the high energy reach of the Muon Collider enables direct probes of the new physics responsible for those deviations. A striking case is that of Higgs portal models inducing a first-order electroweak phase transition. In large regions of parameter space, such scenarios can be uncovered at the Muon Collider both through shifts in single and double Higgs couplings and by direct resonant production of the new scalar states via vector boson fusion.

More generally, many BSM models that alter Higgs couplings include electroweak-charged particles. The Muon Collider can discover such states up to masses of 5 TeV—beyond the reach of the LHC and even a 100 TeV proton-proton collider, assuming the new particles are not QCD-charged. This 5 TeV threshold applies to particles decaying promptly into visible final states, but sensitivity extends also to more elusive

signatures.

Dedicated studies, including full beam-induced background simulation, have demonstrated the Muon Collider's ability to detect minimal WIMP dark matter candidates such as Higgsinos and Winos. In these models, relic abundance is fixed by mass and electroweak charge, setting sharp mass targets. The collider can meet—and in some cases exceed—these targets using techniques such as disappearing or soft track searches, benefiting from the cleanliness of the leptonic environment. Together, these strategies offer broad coverage of minimal electroweak dark matter models.

The Muon Collider's unique combination of precision and energy unlocks discovery avenues that go beyond treating direct searches and precision measurements as separate programs. In field theory, these two modes of probing new physics are inherently linked. At a Muon Collider, this coherence becomes experimentally accessible: weakly coupled BSM scenarios often produce deviations in high-energy observables at the same scale where new particles may also be directly produced.

This overlap is more than a convenience—it is a powerful diagnostic. The simultaneous observation of resonant new states and deviations in scattering cross sections provides complementary information on the quantum numbers, couplings, and structure of the underlying theory. In UV-complete frameworks, the collider's full observable set acts collectively, probing both the infrared manifestations and the ultraviolet origin of new physics.

More than a superposition of capabilities, the Muon Collider uniquely benefits from the intersection of energy and precision. It enables accurate measurements of scattering cross sections at 10 TeV, which are sensitive to BSM physics at much higher mass scales. For theories where new effects scale as $(E/M)^2$, measuring cross sections at percent level corresponds to probing new physics at scales approaching 100 TeV.

A historical precedent illustrates the point: the finite size of the proton—its compositeness—was revealed not by extreme precision at low energy, but by moderate accuracy at higher momentum transfer. Likewise, the Muon Collider can test Higgs compositeness at 40 TeV, and in the absence of deviations, exclude compositeness scales up to 60 TeV. This reach surpasses that of any other planned facility, making the Muon Collider a uniquely powerful instrument for uncovering the substructure of the Higgs sector.

The Muon Collider's discovery potential extends beyond specific models and can be framed systematically through Effective Field Theory (EFT). A broad class of flavour-universal, energy-growing operators—affecting dilepton, dijet, vector boson, and Higgs final states—can be probed up to scales nearly 50 times larger than those accessi-

ble today. This reach exceeds that of electron-positron Z factories, and relies not on sub-permille precision, but on percent-level deviations in high-energy cross sections. Such a strategy confers robustness to the Muon Collider's projections and is further strengthened by the variety of accessible observables, including angular and differential distributions.

Through 10 TeV cross-section measurements, vector boson scattering, and Higgs observables, the Muon Collider can significantly extend the program of electroweak and Higgs precision tests (EWPT). These tests are enabled by a methodology that is less sensitive to theoretical and experimental systematics than traditional low-energy fits, and more naturally suited to characterising new physics once observed.

The same measurements also benefit flavour physics. Neutral current flavour transitions—rare in the SM and suppressed at low energy—become accessible at 10 TeV through processes such as flavour-violating dilepton or dijet production. As before, the energy enhancement amplifies otherwise negligible BSM effects, raising them above background. For many operators in both the quark and lepton sectors, the sensitivity matches or surpasses the projected reach of future low-energy experiments, opening a complementary path to probing the flavour structure of new physics.

Physics highlights

Energy track The high center-of-mass energy of a Muon Collider enables direct searches for new heavy particles with mass reach far beyond that of the LHC. This advantage stems from the fundamental nature of the colliding particles: muons are elementary, and their full beam energy is available for particle production. In contrast, protons are composite, and the steep fall-off of parton distribution functions limits the effective energy available in any given pp collision. As a result, a 10 TeV Muon Collider can access heavier states than the 14 TeV LHC.

This is illustrated in Figure 3.7, which compares the mass reach of the 10 TeV Muon Collider to that of the HL-LHC and the proposed 100 TeV proton collider FCC-hh. The plot shows projected exclusion limits for a range of hypothetical electroweakly produced BSM particles. The Muon Collider reach is shown for a center-of-mass energy of 10 TeV and 10 ab^{-1} integrated luminosity. In models permitting single production, the true reach can be significantly higher (see e.g. [113]). For the Wino and Higgsino, the "ΩDM" label marks the thermal relic mass required to reproduce the observed dark matter abundance. At the Muon Collider, such particles can be pair-produced via electroweak interactions, with cross sections that depend only on their spin and gauge quantum numbers. For masses up to the 5 TeV kinematic threshold, cross sections range from

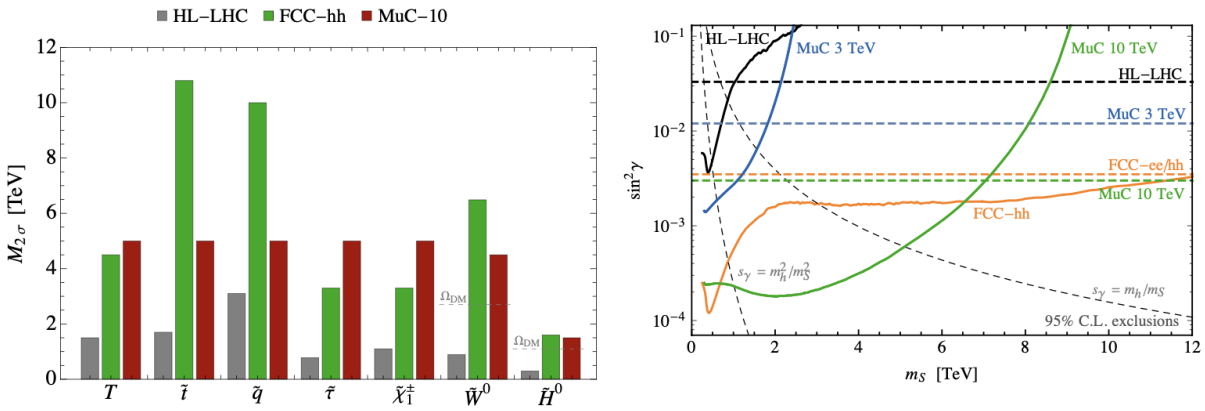


Figure 3.7: Left: Projected 95% CL exclusion limits on the mass of several BSM particles at future colliders, based on electroweak pair production alone. Right: Exclusion contour [110] for a scalar singlet of mass m_S mixing with the Higgs boson via an angle $\sin\gamma$, illustrating the sensitivity to Higgs-portal interactions. From [96]

0.1 to 10 fb, yielding more than 1000 events at an integrated luminosity of 10 ab^{-1} sufficient for discovery in most scenarios with prompt and visible decays

The high-energy reach of the Muon Collider enables direct searches for new particles across a broad range of BSM scenarios, from minimal extensions of the Higgs sector to solutions of long-standing anomalies. Previous studies have explored its sensitivity to WIMP dark matter, heavy neutral leptons, composite resonances, extended scalar sectors, and models addressing the muon $g-2$ anomaly.

These results are typically based on detailed simulations, accounting for SM backgrounds and realistic detector performance using the IMCC DELPHES card [114].

A well-studied benchmark is the scalar singlet model [8–10], in which an additional electroweak-neutral scalar mixes with the Higgs via a portal interaction. This scenario is motivated by its ability to induce a strong first-order electroweak phase transition, and appears in many extensions of the SM. The 10 TeV Muon Collider provides superior sensitivity to this scenario compared to FCC-hh, particularly in the region of parameter space relevant for baryogenesis. Even when indirect constraints from Higgs couplings are included, the MuC remains dominant.

This is illustrated in the right panel of Figure 3.7, which shows the exclusion reach in the plane of singlet mass and its mixing angle with the Higgs boson. The advantage comes from the enhanced cross section for Higgs portal interactions in vector boson fusion at a Muon Collider. Similar conclusions extend to other portal-coupled scenarios, reinforcing the MuC's role as the optimal probe for this class of new physics at multi-TeV energies.

A series of studies [115]–[128] have examined the prospects for discovering WIMP

dark matter candidates at the Muon Collider (see also [110] for a summary). Search strategies include mono-photon and more general mono-X signatures, indirect effects from virtual particles in loops, and direct detection of disappearing tracks associated with the charged partners of the dark matter state.

Among these, disappearing track searches are especially promising, though particularly sensitive to beam-induced background (BIB) from muon decays. Since this background is difficult to model reliably in fast simulations, a dedicated study was carried out in Ref. [117] using full Monte Carlo simulation of the BIB. It demonstrated that disappearing tracks from long-lived Winos and Higgsinos can be observed up to the thermal relic mass.

Complementary strategies based on soft track reconstruction were explored in Ref. [119], which showed that the thermal Higgsino could also be discovered through soft final states. Importantly, this second method enables discovery even at a first-stage Muon Collider operating at 3 TeV.

The mass reach reported in Figure 3.7 for the Higgsino and Wino at the 10 TeV Muon Collider corresponds to a mono-photon search with one disappearing track [117], [123]. These results confirm that minimal WIMP scenarios can be robustly probed across their full cosmologically motivated parameter space.

Precision track Precision measurements in the electroweak and Higgs sectors are a central component of the Muon Collider physics program. Thanks to the large rate of electroweak-scale scattering processes initiated by effective vector bosons, and the low QCD background intrinsic to leptonic collisions, the MuC offers an ideal environment for high-accuracy studies.

With 10 million Higgs bosons produced via vector boson fusion at 10 TeV, detailed projections have been developed for signal strength measurements [129]. These are based on fast simulation and validated against available full simulation results [130]. The resulting sensitivities were used in [110] to perform a global Higgs coupling fit, using the same setup employed in [131] for the comparative assessment of collider scenarios submitted to the 2020 European Strategy process.

The outcome, shown in the left panel of Figure 3.8, confirms that single Higgs couplings can be determined at the per-mille level, matching the performance of proposed low-energy e^+e^- Higgs factories. A closer look reveals strong complementarity between the MuC and low-energy machines in terms of which couplings are best constrained.

These results correspond to the so-called k_0 fit [206], where BSM Higgs decays are assumed absent. To go beyond this and close the more general k_3 fit, at least one

	HL-LHC	HL-LHC +10 TeV	HL-LHC +10 TeV + ee
κ_W	1.7	0.1	0.1
κ_Z	1.5	0.2	0.1
κ_g	2.3	0.5	0.5
κ_γ	1.9	0.7	0.7
$\kappa_{Z\gamma}$	10	5.2	3.9
κ_c	-	1.9	0.9
κ_b	3.6	0.4	0.4
κ_μ	4.6	2.4	2.2
κ_τ	1.9	0.5	0.3
κ_t^*	3.3	3.0	3.0

* No input used for the MuC

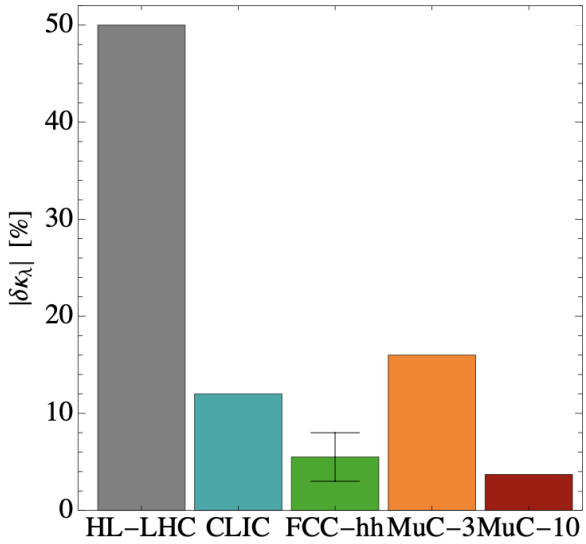


Figure 3.8: Left: Projected 1σ sensitivities (in percent) from a 10-parameter Higgs coupling fit in the k-framework for a 10 TeV Muon Collider with 10 ab^{-1}, compared to the HL-LHC. Also shown is the impact of measurements from a 240 GeV e^+e^- Higgs factory. Right: Sensitivity to the Higgs trilinear coupling modifier δk_λat various future colliders. The performance of the 3 TeV and 10 TeV Muon Collider stages (MuC-3, MuC-10) is compared with that of HL-LHC, CLIC, and FCC-hh. Plots adapted from Ref. [1]

absolute coupling measurement is required—one that does not depend on assumptions about the total Higgs width. The Muon Collider can provide such a constraint via a 0.5% precision measurement of the total inclusive Higgs cross section in Z-boson fusion [105, 112], enabling a 1% determination of the Higgs width.

However, this measurement depends on detecting muons in the far-forward region, requiring a dedicated detector whose feasibility remains under study. If this turns out to be unachievable, the closure of the k_3 fit will rely on inclusive measurements from a low-energy e^+e^- collider —an additional example of complementarity. Alternatively, the flat direction in the fit could be lifted at high energy using mild theoretical assumptions [139].

Unlike low-energy Higgs factories, the Muon Collider can access double-Higgs production and measure the trilinear Higgs coupling directly. As shown in the right panel of Figure 3.8, a 10 TeV MuC with 10 ab^{-1} can achieve percent-level precision on δk_λ [98, 102]. This result is based on a parametric detector model assuming CLIC-like performance, and has been validated against the full simulation projections developed for CLIC [207].

For comparison, Figure 2.2.3 also includes the expected sensitivity of FCC-hh,

which ranges between 3% and 8% depending on detector assumptions [208]. The Muon Collider result shows less dependence on detector performance, thanks to the lower background and cleaner environment provided by lepton collisions.

Precision from energy track A defining feature of the Muon Collider is its ability to combine energy and precision in the same observables. This enables percent-level measurements of scattering cross sections at multi-TeV energies—specifically, in final states involving quarks, leptons, vector bosons, and Higgs pairs with invariant masses around 10 TeV.

High-energy observables naturally amplify the effects of short-distance physics. For BSM interactions that scale with $(E/M)^2$, measurements at 10 TeV probe new physics up to 100 TeV. By contrast, the same effects at the electroweak scale produce deviations suppressed at the part-per-million level, well below the sensitivity of even the most precise low-energy experiments. This strategy mirrors historical precedents, such as the discovery of nucleon structure via deviations in electron scattering. It provides the Muon Collider with a unique advantage over both low-energy lepton colliders and even a 100 TeV hadron collider, which cannot directly access such high mass scales in model-independent terms.

Although the $\mu^+\mu^-$ initial state is electrically neutral, the Muon Collider retains sensitivity to charged final states. At 10 TeV, the probability of emitting soft-collinear W bosons becomes significant due to Sudakov enhancement, enabling high-rate production of charged final states such as WZ, hZ and $t\bar{b}$. These processes effectively probe charged scattering amplitudes initiated by muon-neutrino collisions and open up sensitivity to a wide class of interactions that would otherwise be inaccessible.

By systematically exploiting electroweak radiation, one can define a broad set of observable cross sections that disentangle different BSM effects and allow structured EFT analyses [191]. This expands the range of accessible processes well beyond what might be expected from a neutral initial state, further enhancing the Muon Collider's versatility as a precision and discovery machine.

The sensitivity of high-energy measurements at the Muon Collider to concrete BSM scenarios is illustrated in Figure 3.9. The left panel shows the discovery and exclusion reach for a heavy neutral gauge boson Z', coupled to the Standard Model hypercharge current [210]. At 10 TeV, the Muon Collider can discover such a state with masses up to 100 TeV for gauge couplings comparable to those of the SM. For larger couplings —up to the perturbative limit $g_{Z'} \simeq 1.5$- the exclusion reach extends to nearly 500 TeV. The figure also shows the comparative sensitivity of CLIC, FCC-ee, and FCC-hh [201],

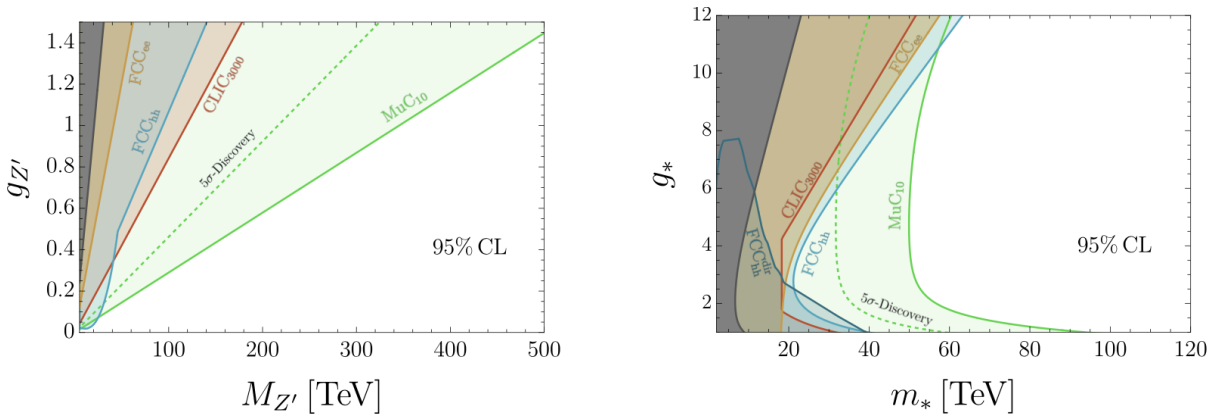


Figure 3.9: Left: Projected 95% CL exclusion limits on a minimal Z' gauge boson from future colliders, with the 5σ discovery reach for muon colliders shown as dashed lines. Right: Combined sensitivity of the Muon Collider to Higgs compositeness in the (m_*, g_*) parameter space, based on deviations in Higgs couplings, di-fermion cross sections, and vector–Higgs final states. Plots adapted from Ref. [191].

which fall well short of this scale. In this scenario, a multi-TeV Muon Collider emerges as the only realistic experimental path to probe such ultra-heavy neutral currents.

The Muon Collider's sensitivity to Higgs compositeness arises from three complementary classes of measurements, whose combined reach is shown in Figure 3.9 in the (m_*, g_*) plane. At large values of the coupling g_*, deviations in Higgs couplings dominate the sensitivity, enabling a mass reach that grows with the strength of the new sector. For small $g_* \sim 1$, constraints from 10 TeV di-fermion cross sections—sensitive to electroweak gauge interaction modifications—become more relevant. A third, dominant contribution comes from di-boson and Higgs-plus-boson final states. These processes, accessible at high energy with percent-level precision, probe interactions tied directly to the compositeness of the Higgs doublet. Their strength depends only on the compositeness scale m_*, making them effective across the entire (m_*, g_*) plane.

Altogether, the Muon Collider can discover Higgs compositeness up to 35 TeV, or exclude it up to 50 TeV even in the most conservative coupling scenarios. As shown in Figure 3.9, this reach exceeds that of all other proposed colliders [201, 206], establishing the MuC as the leading facility for probing the substructure of the Higgs sector.

The R&D trajectory of the Muon Collider also opens the possibility of extending the center-of-mass energy beyond 10 TeV. If the luminosity scales quadratically with energy—as assumed in the IMCC target scenario—the sensitivity to new physics scales Λ would grow linearly. This provides a clear path to further enhancing the collider's reach.

Such scaling is difficult to replicate in e^+e^- colliders, where the precision of key ob-

servables is ultimately limited by systematic and theoretical uncertainties rather than statistics. The Muon Collider thus retains a long-term advantage in pushing the sensitivity to high-scale phenomena through both energy and luminosity upgrades.

Muons and neutrinos track Energetic muon beams in the multi-TeV range will be made available for the first time at a Muon Collider. Colliding muons opens up discovery opportunities for new physics preferentially coupled to the second lepton generation—scenarios that would be inaccessible at electron-based facilities. While motivated by general exploration, this sensitivity is also theoretically grounded.

The muon's Yukawa coupling to the Higgs is significantly larger than that of the electron, implying enhanced couplings to any BSM dynamics associated with electroweak symmetry breaking. This effect has been studied, for example, in extended Higgs sectors [163]. More broadly, the larger Yukawa couplings of second-generation fermions suggest stronger interactions with hypothetical sectors responsible for the origin of flavour. In many UV models featuring flavour non-universal gauge interactions, new physics effects preferentially manifest in muons over electrons [214–222].

High-energy muon collisions are also complementary to low-energy tests of the Standard Model, such as the anomalous magnetic moment of the muon [153–162] or lepton-flavour violating decays like $\tau \to 3\mu$. Should tensions with the SM persist in these measurements, a Muon Collider will be instrumental in confirming their origin and characterizing the underlying dynamics.

In addition to collisions, the decay of high-energy muon beams naturally produces an intense and collimated beam of neutrinos. This offers the possibility of a dedicated neutrino experiment with unique capabilities, extending the scope of the collider beyond electroweak and BSM physics.

Chapter 4

Reconstruction with Object Condensation

4.1 Photon reconstruction in high - granularity calorimeters

High-granularity calorimeters represent a significant evolution in detector technology, designed to meet the increasing demands of precision measurements in high-energy physics experiments. The CMS experiment at the LHC is undergoing a major upgrade in this direction, replacing its endcap calorimeters with a new High-Granularity Calorimeter (HGCAL) for the High-Luminosity LHC (HL-LHC) era [132]. This upgrade features unprecedented spatial and longitudinal resolution, with silicon sensors finely segmented in both transverse and depth dimensions to provide detailed 3D imaging of particle showers.

Similar advances are being explored for future collider concepts, including detector designs for electron-positron machines and proposed muon colliders, where fine-grained calorimetry is crucial to disentangle signal from intense backgrounds. These designs share a common goal: to preserve excellent shower resolution and particle identification capabilities even in environments with extremely high occupancies.

Photon reconstruction in such calorimeters poses distinct challenges. Photons initiate purely electromagnetic showers, but in dense, high-background conditions —such as those anticipated in the HL-LHC or a Muon Collider— accurately isolating the shower core and attributing energy deposits becomes non-trivial. Traditional clustering-based reconstruction approaches, which rely on geometric proximity and hand-crafted features, often degrade under pile-up or background fluctuations. These limitations mo-

tivate the development of machine learning–based methods that can leverage the full spatial structure of the energy deposits and learn to assign them to individual shower sources in a more robust and data-driven manner.

Object Condensation (OC) has emerged as a promising paradigm in this context. It allows for simultaneous clustering and regression by learning a low-dimensional representation in which hits can be dynamically assigned to particle-level objects. The approach is particularly well suited for high-granularity detectors, as it makes full use of the detailed spatial information and naturally scales to variable numbers of hits and objects per event. This chapter focuses on the application of Object Condensation to the problem of photon reconstruction in high-granularity calorimeters, using the DeepJetCore framework [29] and simulation data prepared specifically for this task.

4.1.1 From clustering to condensation: evolution of reconstruction paradigms

The problem of reconstructing particle-level objects from calorimeter data has traditionally been approached through clustering algorithms. In these methods, energy deposits—referred to as “hits” or “voxels” in high-granularity calorimeters—are grouped based on spatial proximity and local energy patterns. A typical workflow involves nearest-neighbor searches, density estimators, or seeded growing techniques, often supplemented by hand-engineered features that exploit the expected longitudinal or transverse shape of electromagnetic showers.

While such techniques have proven effective in relatively sparse environments, their limitations become evident as detector occupancy increases. In future collider environments, particularly in the high pile-up regime of the HL-LHC or the intense beam-induced background of a muon collider, spatial overlap between showers becomes common. In these cases, geometric clustering alone fails to provide unambiguous object associations, leading to degraded energy resolution, object misidentification, and missed photons.

Machine learning techniques offer a new strategy for tackling this challenge. Rather than relying solely on geometric heuristics, learning-based approaches can model more complex correlations between hits and can incorporate both spatial and semantic information. In this setting, Object Condensation (OC) [29] provides a conceptually elegant and operationally powerful solution. OC recasts the reconstruction task as a combination of clustering and regression performed in a learned feature space: each hit is embedded in a low-dimensional latent space, and a set of learned variables determine

its likelihood of belonging to a specific object.

A key innovation of OC is the condensation mechanism, whereby hits are simultaneously attracted to learned object centers and repelled from competing clusters through a loss function that enforces both association and separation. This approach is inherently end-to-end: a single network learns to output both the object assignments and the relevant physical quantities, such as energy or position. Moreover, OC supports variable-length inputs and outputs, making it ideally suited to the sparse and irregular nature of calorimeter data.

In this work, we apply Object Condensation to photon reconstruction in a the Crilin ECal barrel, based on the high-granularity design discussed in Section 3.1.2. Using the HGCalML implementation within DeepJetCore -hosted at `https://github.com/cms-pepr/HGCalML`-, we adapt the model to identify photon-induced showers and estimate their properties on a voxelized input representation. The integration of this approach with a realistic simulation and event generation pipeline is described in the following sections.

4.2 Object condensation with DeepJetCore

The algorithm is based on the idea that all information about an object can be condensed in a number of vertices that is at most the number of voxels in our dataset (and at least 1). It generalizes the use of the GravNet Graph Neural Network [133] in image segmentation to the higher-dimensionality space of a physics detector data, relying on filtering layers to reduce the dimensionality of the dataset and introduce a new parameter space (clustering space).

The network is trained to condensate high-level information on the events into a subset of point in this space, which is to say reconstruct a physical object by clustering together the points belonging to the same objects around a condensation vertex. The clustering itself is performed through the individuation of a condensation point for each object in the dataset and the minimization of a loss function. This loss function usually scales as the distance of each vertex from the condensation point.

In this specific case, the loss is interpreted as a physical potential V. A scalar $\beta_i \in (0,1)$ is predicted by the network, and represent the likelihood for a particular vertex i to be a condensation point. From this quantity a charge is defined through a monotonic function, ensuring a well defined minimum:

$$q_i := \text{arctan}^2\, \beta_i + q_{min}.$$

The force pushing each vertex j towards the object k is then

$$q_i \nabla V_k(x_j) = q_j \nabla \sum_{i=1}^{N} \delta_k^i V_{ik}(x_i, x_j)$$

where x_i are the cluster-space coordinates of vertex i, the sum runs over all the N vertices and $delta = 1$ if vertex i belongs to object k (zero otherwise). The sum over all contributions is approximated by the potential from the vertex of object k with highest charge $q_{\alpha k}$, and consists of an attractive contribution $||x - x_\alpha|| q_{\alpha k}$ and a repulsive contribution max$(0, 1 - ||x - x_\alpha||) q_{\alpha k}$. Summing on all objects k and all vertices j, the total condensation loss is given by Equation 6 in [29]:

$$L_V = \frac{1}{N} \sum_{j=1}^{N} q_j \sum_{k=1}^{K} \Big(\delta_{jk} ||x_j - x_\alpha|| q_{\alpha k} + (1 - \delta_{jk}) \, \text{max}(0, 1 - ||x_j - x_\alpha||) q_{\alpha k} \Big).$$

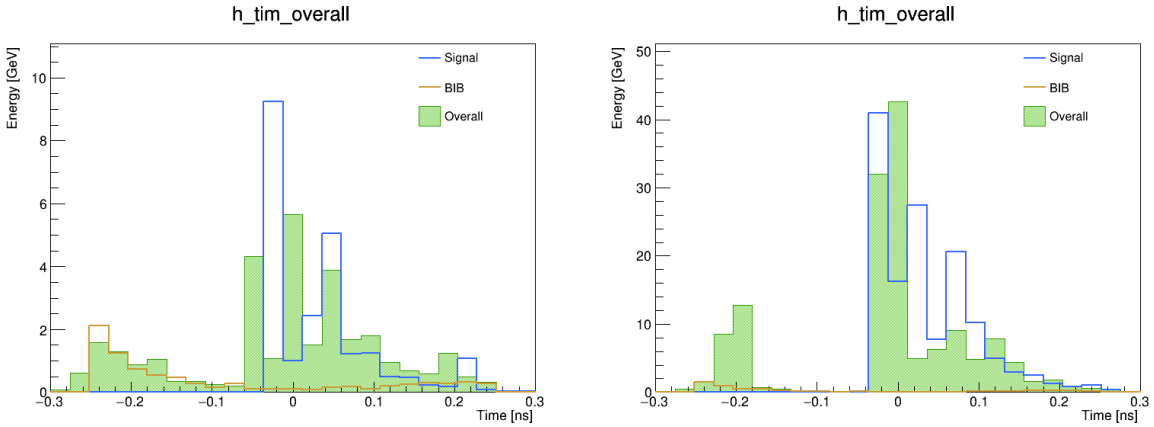
The remaining subset of features we want the network to predict can then be introduced in linear combination as extra terms of the total loss.
An accurate description of the network architecture is provided in [134] together with more details on the additional losses

4.3 Dataset generation

The data chosen to train the algorithm consists of a set of 10k single-photon events, generated with Geant4[135] on the MUSIC detector geometry developed by the Muon Collider Collaboration discussed in Chapter 3.2. The energy of the photons is uniformly sampled in the interval [10,175]GeV, while they hit the barrel orthogonally ($\eta = 0$) with a transverse direction ϕ sampled isotropically.

The BIB deposits from a 1.5TeV center of mass simulation with Geant4 [135]–[137] are subsequently added with an overlay script -available at `https://github.com/FedericoNardi/CrilinTuples`- that implements some basic event digitization: a timing cut in a 500ps window from the bunch crossing has been implemented to reduce the BIB contribution, and the time value assigned to the cell deposit corresponds to the first deposit (BIB or signal) happening within the window in that cell. The timing deposit for the single datasets and for the overlay result is shown in figure 4.1.

Note that the overall distribution used later in the training features is not the simple sum of the signal and BIB deposits. Rather, a simple accumulation strategy is



(a) Timing Histogram for 25 GeV

(b) Timing Histogram for 150 GeV

Figure 4.1: Timing histograms for different energy levels.

implemented to mock triggering and digitization: the time coordinate for each voxel is assigned as the time of the first registered hit, and the corresponding energy is the resulting of integration of all deposits in the voxel within the 500ps time window. Gaussian smearing of 20ps has been added to the final time value in order to reproduce the response of electronics.

Figure 4.2 offers some visualizations of different energy showers with and without superimposed BIB. At lower energies the signal information is almost completely lost in the noise of the initial layers.

4.4 Training and inference

The neural network has been trained in two configurations, including and omitting timing information in the training features. This allows us to see whether the effect of including timing distribution allows better energy reconstruction. The algorithm is trained to assign cells either to a signal or background cluster, as well as to infer a signal deposit for each cell. The training losses are shown in figure 4.3

In order to verify the reconstruction performance, we test the network on a dataset of photon showers in fixed energy points 10,25,50,75,100,125,150,175GeV. For each point, the energies of those cells labeled as signal are summed together and plotted in a distribution which is fitted to a CrystalBall function to obtain a mean and sigma value. The result of these fits are shown in figure 4.4.

The reconstructed energies for each point are plotted in figure 4.5, both for the case with and without timing. They are also compared with the reconstruction performed

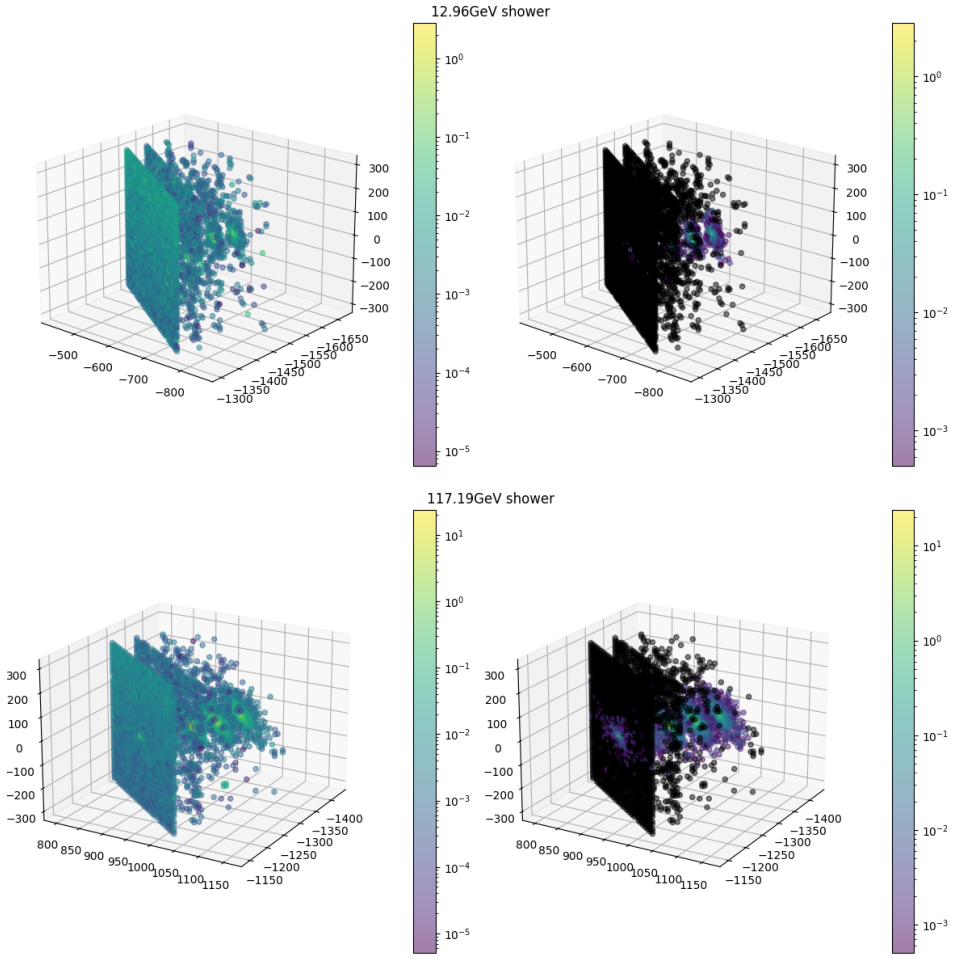
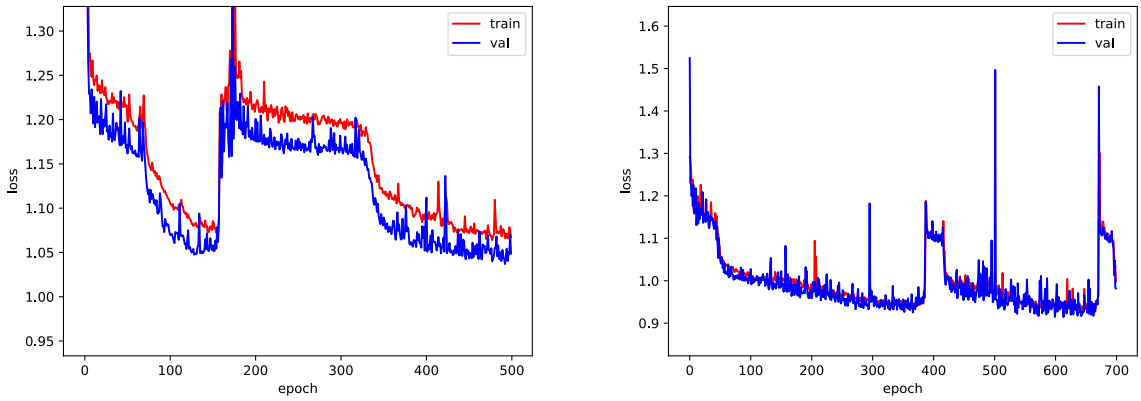


Figure 4.2: Visualizations of simulated showers with overlayed BIB (right) and signal-only deposits (right).

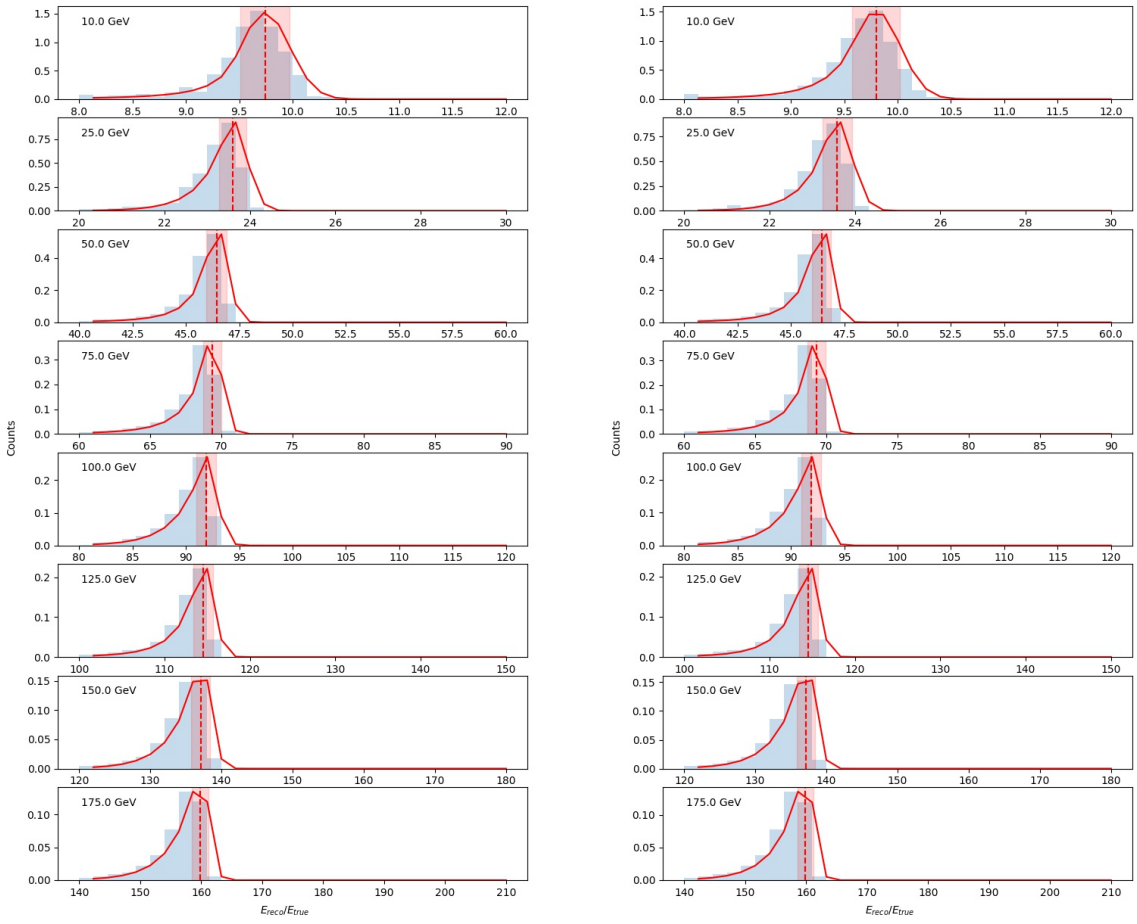
using the Pandora and ParticleFlow algorithms integrated in the Muon Collider collaboration framework [138].

We note immediately the significant improvement in reconstruction performance, especially at lower energies where most of the signal information is lost in the background. We can also see that the addition of time information to the features does not seem to be particularly relevant to improve reconstruction in our case, highlighting the fact that the time window cut might already make good use of it and its addition in the training features might just introduce computational complexity.



(a) (b)

Figure 4.3: Training losses for object condensation with timing information (b) and without (a). Sudden fluctuations are related to the activation of different loss components. The final model weights are taken as the ones corresponding to the minimum loss value for each configuration.



(a) (b)

Figure 4.4: Reconstructed energies and fitted CrystalBall functions with (b) and without timing information (a).

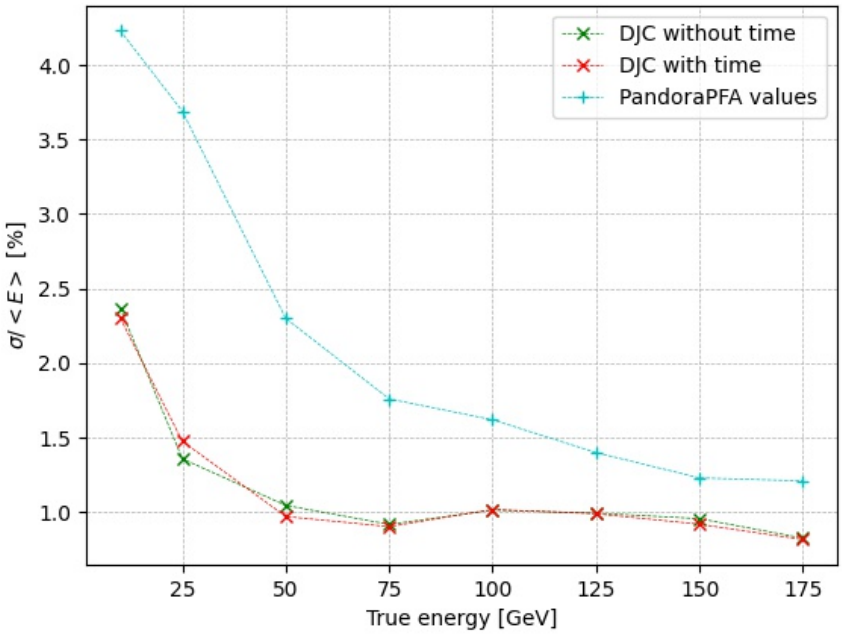


Figure 4.5: Energy reconstruction performance for object condensation, compared with collaboration results

Chapter 5

Pipeline and Surrogate Models

Optimizing the layout of an electromagnetic calorimeter (ECAL) requires balancing competing constraints between physics performance and instrumentation cost. Within the MODE collaboration [62], we approach this challenge from a differentiable programming perspective. The aim is to build an end-to-end pipeline capable of guiding ECAL design choices by jointly modeling detector response, reconstruction performance, and geometric complexity

The structure of this pipeline is outlined in Figure 5.1. It is composed of three main components:

- A differentiable simulation model that approximates the detector response for given physics benchmark events, producing a structured representation of how the energy is deposited within the calorimeter volume.
- A reconstruction algorithm that extracts high-level physical objects from the simulated response. For this purpose, we employ an Object Condensation network developed for the CMS High-Granularity Calorimeter upgrade [29], which clusters the deposited energy into signal candidates.
- A cost model that evaluates the instrumentation overhead associated with a given detector layout, accounting for segmentation, readout density, and spatial coverage.

These models are integrated into a unified, differentiable workflow where a global loss function—capturing both signal sensitivity and cost—can be minimized using gradient-based optimization tools such as PyTorch [139] and TensorFlow [140]. The outcome is not a definitive detector design, but rather a parameter-efficient framework capable of exploring the design space and suggesting geometries with improved trade-offs.

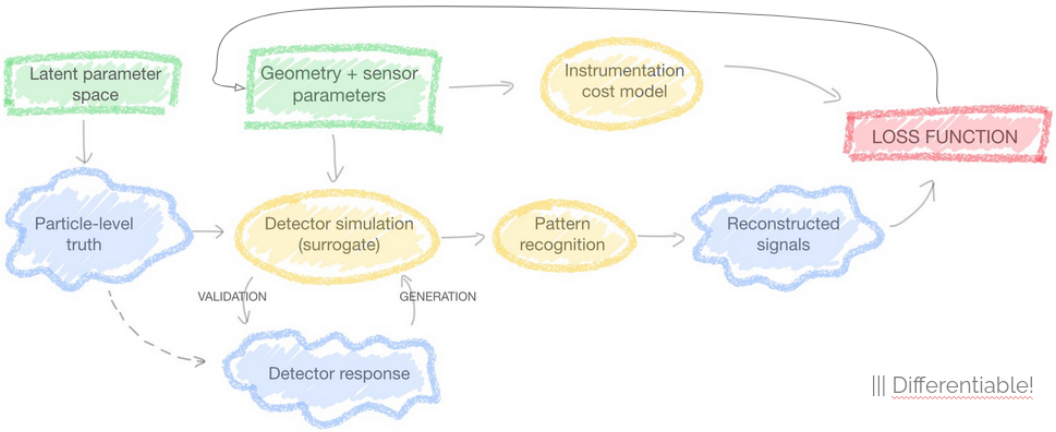


Figure 5.1: Sketch of the workflow towards the development of a differentiable pipeline for the geometric optimization of the ECAL

The remainder of this chapter details each component of the pipeline. We begin by introducing the surrogate models used to emulate the detector response—one for beam-induced backgrounds (BIB) and one for electromagnetic showers from signal photons—along with the differentiable event generator that combines them. These elements are implemented in the repository folder at [141], which contains the full pipeline logic including geometry tessellation and spatial mapping. We then present the reconstruction strategy based on Object Condensation, and discuss how reconstruction performance feeds back into the optimization of the detector geometry.

5.1 Signal generation surrogate

In order generate and sample single-photon events on a continuous range of energies while still preserving code differentiability, we need to develop a surrogate that mimics the energy deposition of the showers in a block of PbF_2 material. We choose to set this up in two steps, to maximize code modularity and allow for future patches and upgrades, as well as to simplify the submodel structure:

1. A basic GNN to generate 2D shower images;

2. A regression algorithm that allows us to recover the 3D structure of the shower.

In the following parts we describe in detail these two algorithms, underlying their advantages and limitations within our framework.

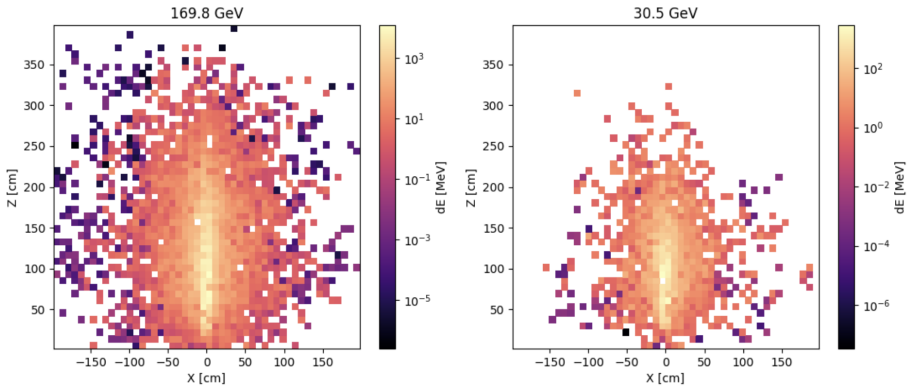


Figure 5.2: Visualization of two Geant4-generated events from the signal dataset.

5.1.1 GNN architecture and training pipeline

To predict the local energy deposition pattern dE of electromagnetic showers in a calorimeter slice, we implemented a graph-inspired neural network based on the GravNet architecture[133] using TensorFlow[140]. The model operates on a variable-sized, sparsely sampled set of hit points in the calorimeter, each represented by its x, z, and primary energy E_0 coordinates. The aim is to regress the corresponding energy deposits per hit, dE, in a supervised manner.

Input and preprocessing The input dataset consists of HDF5 files, each containing 2D calorimeter events in the form of arrays with columns (x, y, E_0, dE). Each event consists of 2D histograms of energy deposits in a block of PbF_2 generated with Geant4, with energy uniformly sampled in the range $[10, 175]$GeV. A visualization of two of those events is proposed in Figure 5.2.

The data is processed as follows:

- A subsampling routine randomly selects a fixed number of hits (default: 5000) per event. This keeps memory usage under control when generating (100x100) pixel images;
- For each hit the (x, z, E_0) values are retained as features;
- the target for regression is the deposited energy dE.

The dataset is loaded using a Python generator wrapped into a `tf.data.Dataset` pipeline, enabling efficient batched and prefetched input.

GravNet-based model The neural network is based on a GravNet-like message-passing architecture tailored for spatial point clouds:

1. Two consecutive GravNet layers serve as the core of the architecture. Each layer performs the following:

    - Extracts a coordinate subspace -in our case (x,z)- to compute Euclidean distances between hits;
    - Identifies k-nearest neighbors (with $k = 8$) for each point;
    - Aggregates neighbor features via average pooling;
    - Applies a small feedforward MLP to the aggregated features;
    - Concatenates the original spatial positions with the learned features;

2. Feature Mixing and Regression Head:

    - The output from the two GravNet layers is concatenated and passed through a stack of dense layers with decreasing dimensionality: $120 \to 64 \to 32$ units;
    - Batch normalization is applied before the final layer;
    - The last layer is a fully connected dense layer with a ReLU activation, producing one predicted dE value per hit.

This structure allows the model to learn localized patterns of energy deposition while respecting the non-grid nature of the data.

Loss function and optimization The model is trained using the Mean Squared Error (MSE) loss between the predicted and true deposited energy values. Although a placeholder for a custom loss combining MSE and normalization terms exists, the default training uses standard MSE:

$$\mathcal{L} = \frac{1}{N} \sum_i (dE_i - \hat{dE_i})^2.$$

The optimizer used is Nadam with a learning rate of 10^{-3}. Training is conducted for 1000 epochs using GPU acceleration.

5.1.2 Validation and Inference

At regular intervals, the following diagnostics are computed:

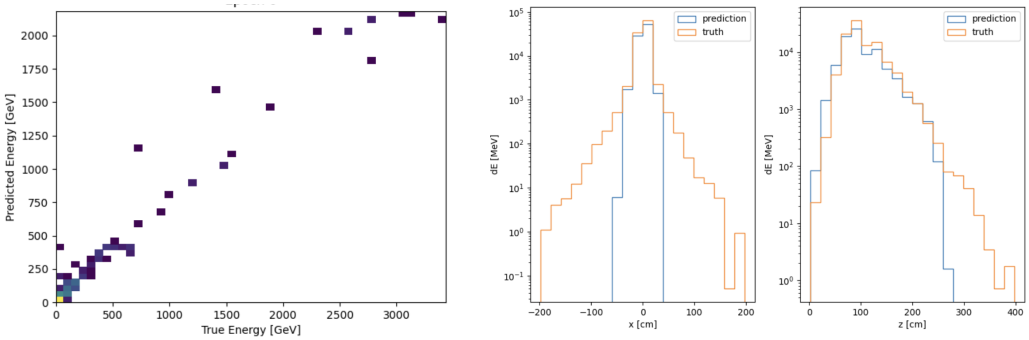
- 2D correlation plots between true and predicted dE values across all hits in a validation event

- Longitudinal and transverse profiles, showing the energy deposited along x and z directions, for both truth and prediction.

Figure 5.3 shows these plots for a random validation event after the training epochs. The correlation plot (a) has approached the diagonal, providing accurate predictions for most voxels -despite some sparse outliers still remaining. Looking at the shape development of the shower (b), significant deviations start to appear at the order of 10 MeV deposits, however the bulk of the shower both transversely and longitudinally is already decently captured.

It is clear that some more fine-tuning for the model is needed, however this already provides a satisfying description for our purpose -which, at this phase, is not to develop an accurate fast-simulation for photon showers, but rather obtain a differentiable surrogate to set up an optimization pipeline.

Currently a study within MODE [142] is tackling the development of Conditional Diffusion Models to provide accurate differentiable surrogates that can patch the current signal generation model, significantly improving physical accuracy.



(a) ERRATA: labels are MeV (b)

Figure 5.3: Validation plots for event generation using out basic GNN model. Right: correlation plot of predictions vs targets, Left: transversal and longitudinal shape of predicted and target showers.

5.2 Inference of a 3D distribution from a marginal

The nature of our shower data surrogate introduces the need to infer a 3-dimensional distribution given what *de facto* is a marginal distribution provided by our model. Specifically, we observe a function $I(x,z)$, which corresponds to the integral of the true three-dimensional energy density $E(x,y,z)$ E(x,y,z) over the unobserved transverse coordi-

nate y:

$$I(x,z) := \int_{\text{dom}(y)} dy\, E(x,y,z). \tag{5.1}$$

Reconstructing the full volumetric distribution $E(x,y,z)$ from this marginal constitutes an ill-posed inverse problem. To make the problem tractable, we exploit the physical symmetry of the underlying phenomenon: electromagnetic showers in homogeneous calorimeters are approximately symmetric around the incoming particle direction, which we align with the z-axis. This motivates the assumption of cylindrical symmetry, allowing us to reduce the problem to the recovery of a radial-longitudinal density function $\rho(r,z)$, where $r = \sqrt{x^2 + y^2}$.

Given the finely binned yet discrete nature of our output, we choose to frame this as a minimization problem. The aim is to learn the elements of a density matrix $R(r,z)$ with elements ρ_{ij}, where we exploit cylindrical symmetry with respect to the shower axis (z-axis in our case) to reduce the dimensionality of our problem.
Passing from the coordinate couplet (x,y) to $r = \sqrt{x^2 + y^2}$ requires the definition of a new matrix $F(r,x,y,z)$ that provides the map between coordinates. This matrix can be interpreted as a Jacobian-like mapping between the two coordinate systems, computed via stochastic sampling of each Cartesian bin to estimate its contribution to concentric cylindrical bins. Each element f_{ijkl} is calculated as the fraction of volume of klth cylindrical bin that falls within the ijlth cartesian bin.
The cartesian energy deposits E_{ijk} can thus be obtained as

$$E_{ijk} = \sum_m \sum_l f_{ijkl}\, \rho_{lm} \tag{5.2}$$

and its marginal by simply collapsing one of the first two indices equivalently

$$\hat{I}_{ik} := \sum_j E_{ijk}. \tag{5.3}$$

We use the difference between this latter quantity 5.3 and the original (binned) marginal 5.1 to define a χ^2 quantity to be minimized

$$\chi^2 := \sum_{ij} \left(\hat{I}_{ij} - I_{ij}\right)^2. \tag{5.4}$$

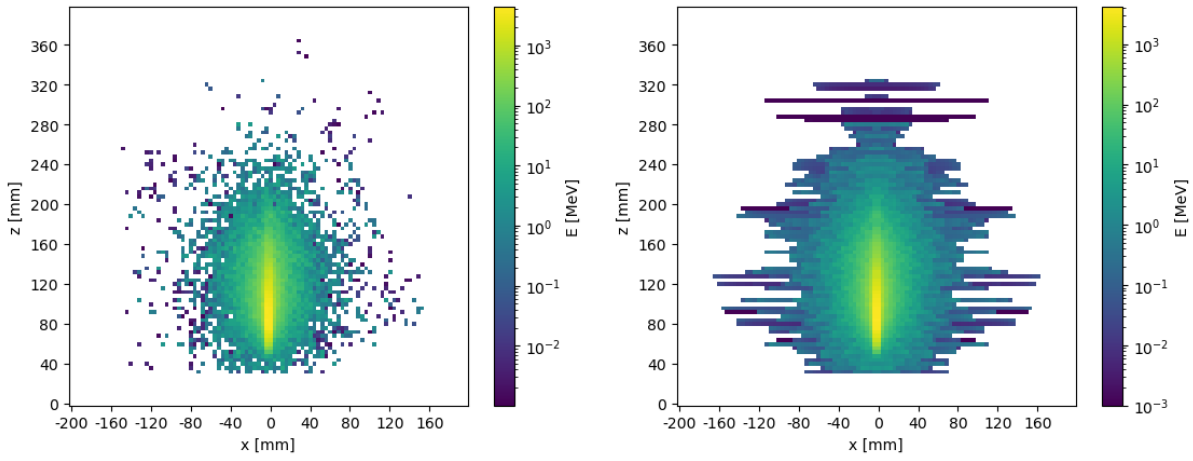
This formulation enables us to recast the reconstruction task as a differentiable optimization problem in TensorFlow[140]. The parameter tensor ρ is updated using gradient descent with the Adam optimizer. To maintain physical consistency, negative val-

ues of ρ are clipped to zero at each step. The full reconstruction routine is summarized in Algorithm 2:

Algorithm 2 Reconstruction of 3D Energy Distribution from a Marginal using Cylindrical Symmetry

Input: Target marginal image $f_Y(x,z) \in \mathbb{R}^{n\times m}$
Output: Predicted 3D distribution $E(x,y,z) \in \mathbb{R}^{n\times m\times p}$

procedure InflateShower(f_Y)
  Precompute cylindrical ring fractions f via uniform sampling
  Initialize density tensor ρ as trainable variable
  for $t = 1$ to T (training steps) **do**
    Compute predicted marginal: $\hat{f}_{Y\,i,j} = \sum_k \sum_m \sum_l f_{ijkl}\,\rho_{lm}$
    Compute loss: $L = \sum_{ij} \left(\hat{f}_{Y\,ij} - I_{ij}\right)^2$.
    Update ρ using Adam optimizer and gradients from L
    Enforce physicality: $\rho \leftarrow \max\{\rho,0\}$
  end for
  Reconstruct full 3D energy volume: $E_{ijk} = \sum_m \sum_l f_{ijkl}\,\rho_{lm}$
  return E
end procedure



(a) (b)

Figure 5.4: Geant4-generated marginal distribution (a) vs predicted marginal from χ^2 minimization (b).

Figure 5.4 shows the marginal distribution of a test event of 142GeV generated with Geant4 (a) compared with the predicted marginal from our code (b). Despite the presence of low-energy rings originated from sparse hits outside of the original bulk distribution both images are consistent in terms of shape and total deposition, as Figure 5.5 highlights.

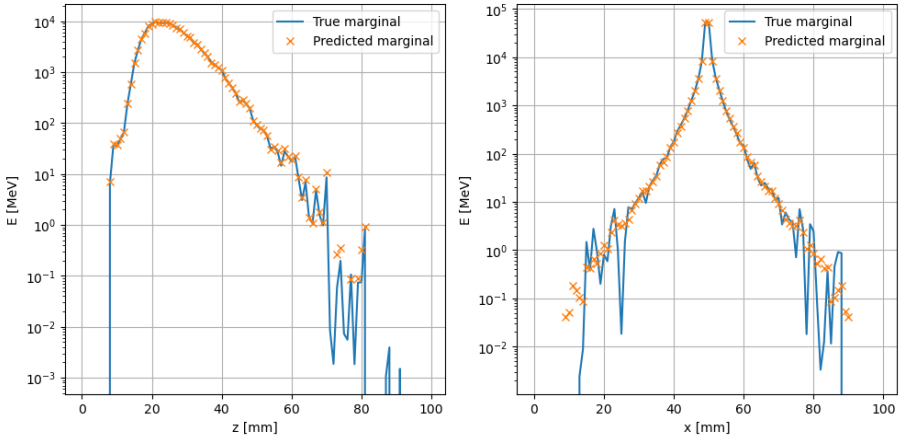


Figure 5.5: Energy deposit distribution along transverse coordinate (x) and z for both true and predicted shower.

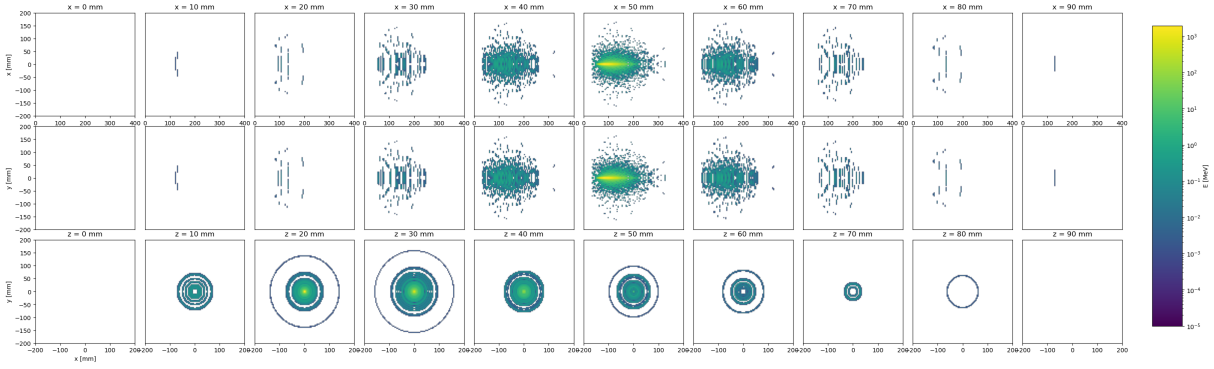


Figure 5.6: Volumetric sections of the inferred distribution $E(x,y,z)$ with planes spaced 10 cm. Top row: xz-views, central row: yz-views, bottom row: xy-views.

To qualitatively assess the reconstructed 3D structure and verify that cylindrical symmetry is preserved, we visualize orthogonal 2D slices of the inferred $E(x,y,z)$ volume in Figure 5.6. The top row shows xz-planes, the middle row yz-planes, and the bottom row xy-planes at regular 10 cm intervals. The visual consistency between orthogonal projections supports the validity of the cylindrical assumption in the reconstruction, and in general the consistency of our model.

5.3 BIB generator surrogate

In parallel to the reconstruction of signal showers, we developed a probabilistic model to describe the average background energy deposition due to beam-induced background (BIB). The aim is to infer a smooth two-dimensional map of energy density as a function of calorimeter depth z and transverse layer index y, based on simulation-derived

histograms. This approach provides a reference BIB template that can be used for subtraction, comparison, or as input to later simulation stages.

5.3.1 Data processing and Gaussian Process Regression

Dataset description The model ingests ROOT histograms corresponding to energy depositions per calorimeter wedge and layer. Those histograms are obtained from the 750GeV Geant4 BIB simulation used also in [130]. It consists of a set of detector hits describing the interaction of the muon decays originating by a single incoming beam (from the left side of the detector) after being shielded by the nozzle. Note that to describe a full event, this flux needs to be symmetrized to simulate the effects from the other side of the detector. From this dataset we isolate the ECal hits and store them in independent files to avoid multiple computationally expensive BIB preprocessing cycles. Each histogram, labeled as `h_layer_<wedge>_<layer>`, contains energy deposition as a function of z. For each wedge (12 in total) and each of the 5 radial layers, we average over the transverse x-axis to extract a one-dimensional energy profile. The central axis positions are computed from bin edges and correspond to the original Crilin design of Section 3.1.2. The final input data is built by:

- Averaging the energy deposits across all wedges, producing a $(435,5)$ matrix;
- Creating coordinate grids (z,y) where $y \in \{0, 62.5, 125, 187.5, 250\}mm$ and $z \in$ `range`$[-2175,2175]$mm with 1cm spacing;
- Normalizing the resulting energy values into a volumetric density in GeV/mm^3.

This results in a dataset of shape (435,2) for the input coordinates and (435,1) for the target densities for each of the 5 layer quotas y.

Model description To infer a continuous $f(X)$, smooth distribution from the discretized input (especially along the y coordinate), we employ a Gaussian Process Regression (GPR) model. GPR is a non-parametric Bayesian approach that models the target function as a distribution over functions [143, Section 45]. Given a set of noisy observations, it provides both the posterior mean and variance of the predicted function.

Let $X = (z,y)$ denote the input space. The GP prior assumes:

$$f(X) \sim \text{Norm}(0, K(X,X'))$$

where K is the covariance kernel. We use a Radial Basis Function (RBF) kernel defined as:

$$K_{RBF} := \sigma^2 \exp\big(-\frac{1}{2}||X_i - X_j||_L^2\big),$$

where σ^2 is the signal variance, and $L = (l_z, l_y)$ the length scales along the two input axes. This kernel encodes the smoothness and spatial correlation of the energy deposition.

To incorporate observational noise, we define the full covariance matrix as

$$K_{\text{train}} = K(X,X) + \sigma^2_{\text{noise}} I,$$

where σ^2_{noise} is the noise variance. Inference is performed using the standard GP closed-form expressions. The predictive mean and variance for a test input X^* are:

$$\mu(X^*) = K(X,X^*)^T K^{-1}_{\text{train}} Y, \tag{5.5}$$

$$Var(X^*) = K(X^*,X) - K(X,X^*)^T K^{-1}_{\text{train}} K(X,X^*). \tag{5.6}$$

The matrix inversion is performed efficiently via Cholesky decomposition. Once computed, predictions at arbitrary points are available at negligible cost.

5.3.2 Implementation and inference

The model is implemented in TensorFlow and compiled using XLA with `@tf.function` decorator. Key parameters of the kernel — the signal variance, two length scales, and noise level — are chosen fixed as $l_y = 5.0$mm, $l_y = 20.0$mm and $\sigma_{\text{noise}} = 10^{-3}$.

The prediction function is precompiled and supports batched evaluation. This enables fast interpolation of background energy density maps onto arbitrary detector layouts or binning schemes.

Figure 5.7 shows the 95% confidence region for the predicted volumetric density function $f(X)$ evaluated at the voxel centroids of the original Crilin design. The validation check to match the total deposition from the original simulations is passed.

To test the full strength of the model however we need to verify its interpolation power and therefore whether it is able to yield sensible results when inferring density distributions at intermediate y quotas. The Geant4+FLUKA simulation does not provide us access to those intermediate values, that become however crucial when studying alternative geometries.

Figure 5.8 provides the shapes of sampled distributions obtained evaluating our GP at intermediate layer quotas. Since we lack a validation dataset to assess the

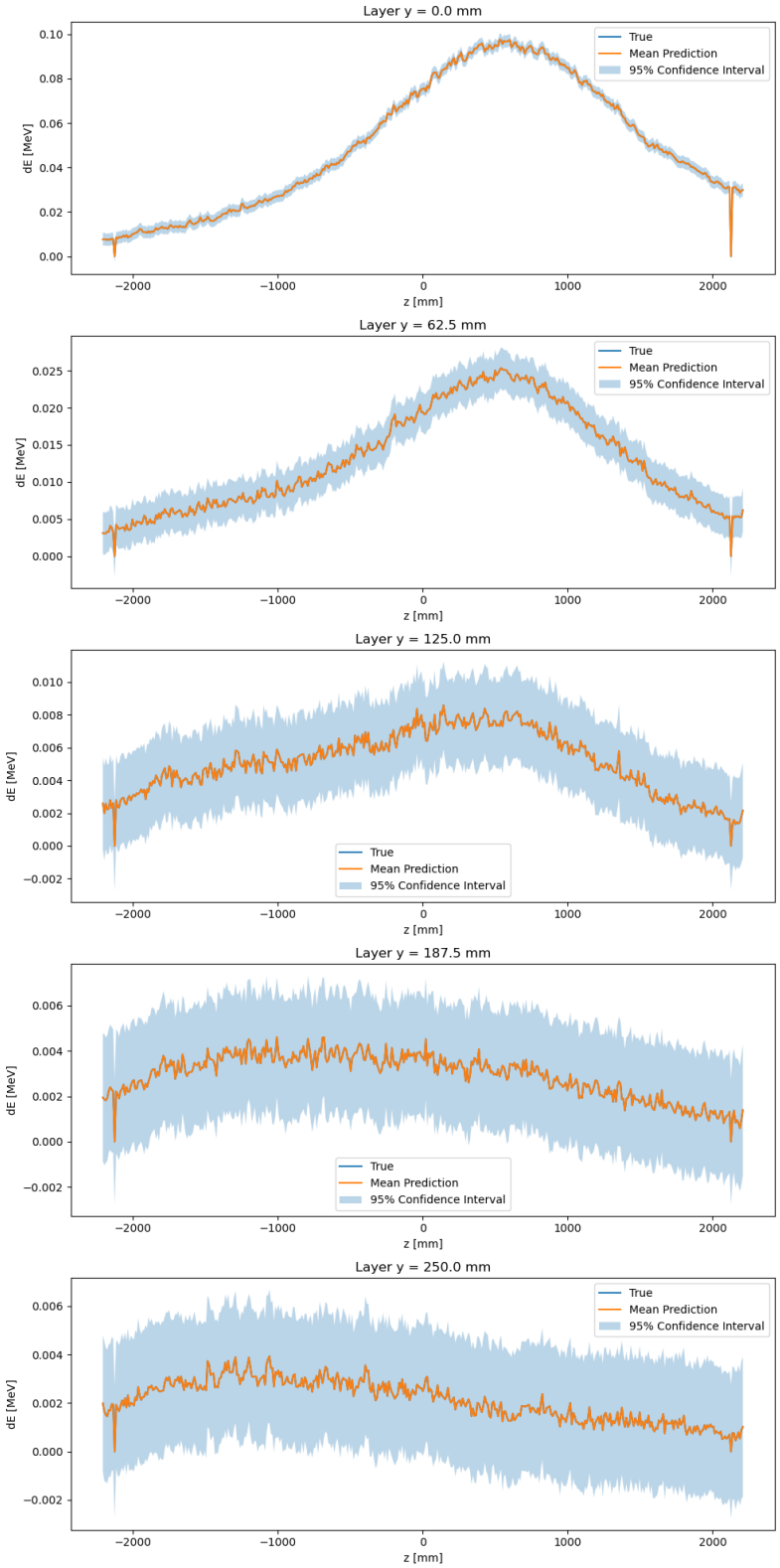


Figure 5.7: 95% Confidence region for $f(Y)$ inferred through GP model at the original Crilin layer quotas. The original distribution from the simulation dataset is superimposed too.

predictions we need to rely on consistency checks, such as verifying that energy spans continuously decrease when going further from the interaction region (increasing the y quota), and that the total deposition in the same volume matches the original. Both checks are consistent.

As a final step, in order to reproduce the physical BIB flux inside the calorimeter, this distribution needs to be symmetrized and summed over the original one. This allows us to simulate contributions coming from both beams, and thus decays along the beamline incoming from the left and right side of the detector.

5.4 Overlaying signal and background: from surrogates to events

Once both the BIB and signal models have been introduced, the next step in the simulation pipeline is to combine them into complete calorimetric events. This is achieved by coherently embedding the two contributions—background and shower—within a common spatial discretization of the detector volume. The resulting event representation is not only differentiable with respect to the underlying model parameters (e.g. centroid positions, source direction), but is also structured in a way that makes it compatible with downstream reconstruction and optimization steps.

At the core of this overlaying process lies a tessellation of the sensitive detector volume into irregular, non-overlapping regions, each associated with a ”centroid”. These centroids define the spatial units used to accumulate energy, and serve as the shared basis for both the background model and the shower generation procedure. In the current implementation, centroids are sampled once at the beginning of each optimization cycle and are held fixed during the forward simulation of BIB and signal components.

The combined event must reflect not only the spatial heterogeneity of the calorimeter layout, but also the intrinsic variability of particle incidence. To that end, each generated event is randomized in both the primary particle direction and the transverse entry point. The direction is specified in spherical coordinates (θ, ϕ), while the transverse entry point is randomized within a fixed acceptance window (aligned with the calorimeter face). This ensures that each event samples a different incidence geometry, consistent with a realistic distribution of collision products at the Muon Collider.

The end result is a structured event composed of a set of 3D centroids, a vector of energy deposits (in units of MeV), and associated volumetric metadata. The complete generation process—including spatial discretization, energy assignment, and overlay—is implemented as a differentiable pipeline and is fully compatible with GPU

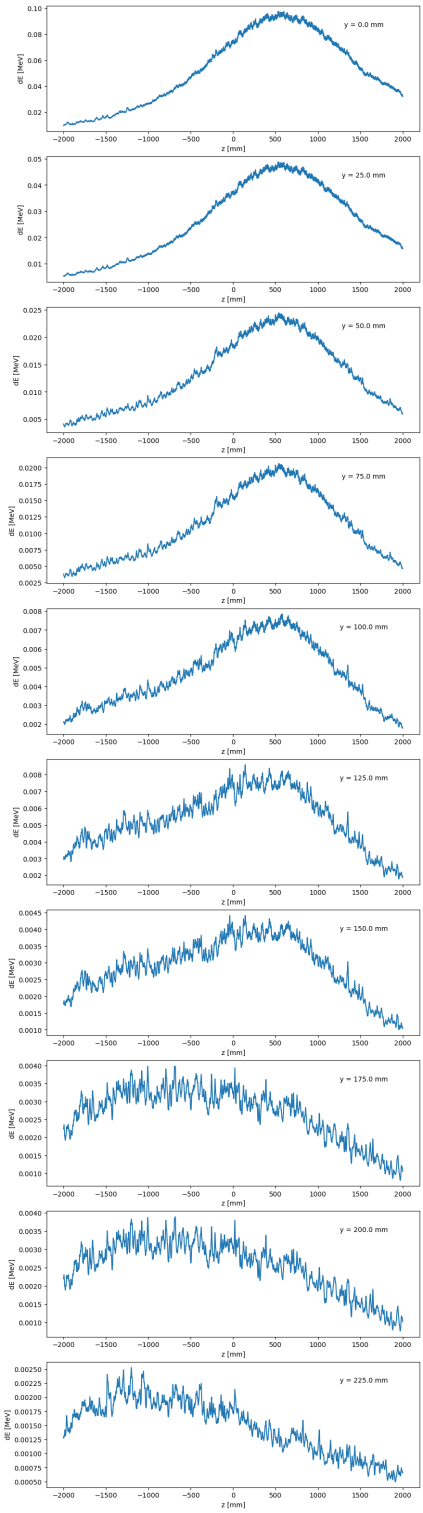


Figure 5.8: Interpolated BIB flux densities at intermediate layer quotas.

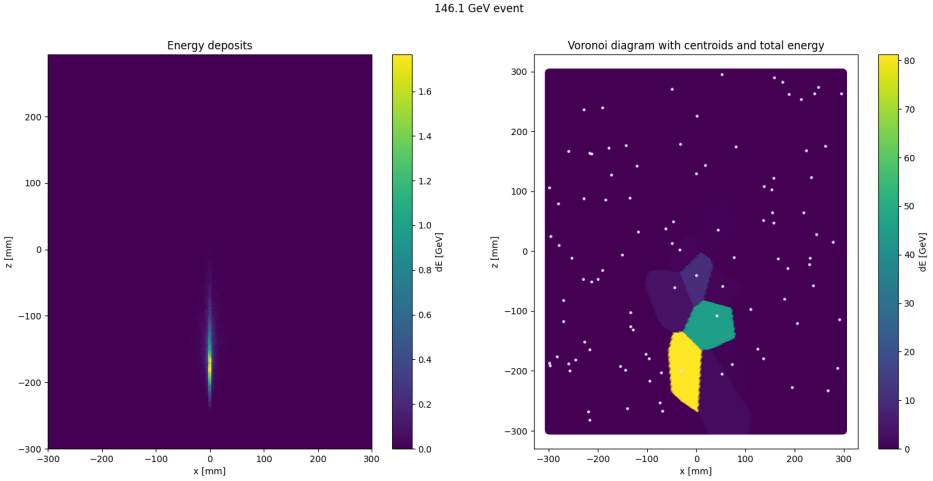


Figure 5.9: 2D projection of a photon shower (left) and respective Voronoi assignments (right), given a random distribution of centroids.

execution.

5.4.1 Voronoi regions and energy mapping

To construct a meaningful representation of calorimetric activity, both the background and signal contributions must be embedded within a common spatial structure. In our framework, this structure is defined by a centroid-based Voronoi tessellation of the calorimeter volume. The use of irregular volumetric cells allows for flexible and differentiable spatial binning, capable of adapting to learned configurations during optimization.

The Voronoi tessellation is computed over a predefined 3D bounding box, using the centroids as generating sites. Each voxel in the calorimeter is assigned to the nearest centroid according to Euclidean distance, resulting in a partition of space into convex polyhedral regions. The implementation, provided in the `Voronoi3D` class, is fully vectorized and supports GPU execution via TensorFlow-compatible operations. Although expensive at high resolution, this step is performed once per event and provides both the cell volumes and an assignment map for later energy aggregation.

A planar slice of such a tessellation is shown in Figure 5.9, where each polygon represents the intersection of a Voronoi region with a constant plane. The centroid locations are distributed non-uniformly, and the resulting cell shapes reflect the underlying spatial flexibility of the representation. This discretization is held fixed throughout a single forward pass, and its resolution and stability are critical for the accuracy of downstream energy modeling.

The two surrogate models introduced earlier—one for the beam-induced back-

ground (BIB) and one for the photon-induced electromagnetic shower—differ not only in their physical interpretation but also in the structure of their outputs. The BIB surrogate returns a continuous 3D energy density flux, defined on a regular voxel grid spanning the entire calorimeter volume. This density is interpreted as a volumetric energy distribution, which is integrated over each Voronoi region. Each voxel's contribution is weighted by its cell volume and accumulated into the corresponding centroid's energy sum. The result is a per-region energy deposit vector reflecting diffuse, low-energy background radiation.

In contrast, the shower surrogate —via the 3D regressor mechanism— produces a fixed, coarse grid of energy deposits that directly correspond to pre-tabulated voxels. These deposits are not interpolated densities but rather localized energy lumps positioned in the shower's intrinsic coordinate system. Before being mapped to the calorimeter, the shower is subjected to a rigid rototranslation: its axis is rotated to align with a randomly sampled incident direction (θ, ϕ), and its origin is shifted such that the axis intersects the calorimeter's front face at a randomly chosen entry point on the $z = 0$ plane.

Each of these transformed energy deposits is then mapped to the corresponding Voronoi region using the previously computed index map. As with the BIB, energy is accumulated per region, normalizing the volumetric density by the region volume (estimated via a simple Monte Carlo sampling). The final energy vector for the event is the sum of BIB and shower deposits, sharing the same centroid-based spatial structure.

This mapping process remains differentiable throughout, including with respect to the centroid coordinates and the shower's angular and positional parameters. It therefore enables gradient propagation across spatial geometry, energy allocation, and event composition—crucial for end-to-end optimization of calorimeter design and reconstruction.

5.4.2 Event assembly with `EventGenerator` class

The final assembly of a simulated calorimetric event—incorporating both background and signal components—is managed by the `EventGenerator` class in [141]. This class encapsulates the logic for transforming surrogate model outputs into a complete event structure, ready for use in training or inference. The generation process is modular, reproducible, and compatible with differentiable pipelines.

At its core, the class takes as input a fixed set of centroids and a randomized set of kinematic parameters describing the incoming photon: the direction of incidence, defined by spherical angles (θ, ϕ) uniformly sampled between 0 and 30deg, and the

transverse entry point, sampled from a predefined region on the $z = 0$ plane. These parameters define the spatial transformation applied to the signal shower, ensuring variation across generated events while preserving the geometric consistency of the simulation.

The event generation proceeds in the following steps:

1. Background Projection: The BIB surrogate is queried to produce a smooth 3D energy density over the calorimeter volume. This density is then integrated across the Voronoi regions using the assignment map generated by the `Voronoi3D` method. The result is a vector of per-region background energy deposits $E_{BIB} \in \mathbb{R}^N$, where N is the number of centroids.

2. Shower Transformation and Mapping: The fixed, coarse grid of energy deposits produced by the shower surrogate is transformed into calorimeter coordinates. This includes:

   - a rotation aligning the shower axis with the sampled (θ, ϕ) direction;
   - a translation placing the shower origin such that it intersects the calorimeter at the randomized entry point.

   The transformed voxel coordinates are then assigned to Voronoi regions, and the corresponding energy values are summed to obtain the signal contribution E_{shower}.

3. Overlay and Output Packaging: The total event energy is computed as the sum $E = E_{BIB} + E_{shower}$, yielding a vector of deposits aligned with the input centroids. The quantity E_{shower} is also retained to calculate the signal fraction per hit, serving as the target for the reconstruction method.

This procedure is applied independently to each event in the batch, and supports parallel execution across events for efficient throughput. Crucially, the entire process is implemented using TensorFlow primitives, enabling automatic differentiation through all components—from centroid placement to energy accumulation. This makes the generator fully compatible with reconstruction loss backpropagation, allowing joint optimization of both geometry and reconstruction model.

Debugging and inspection utilities are embedded within the class, including routines for visualizing energy projections along the xz and yz planes. These have proven useful for verifying alignment, shower placement, and BIB-background consistency across events.

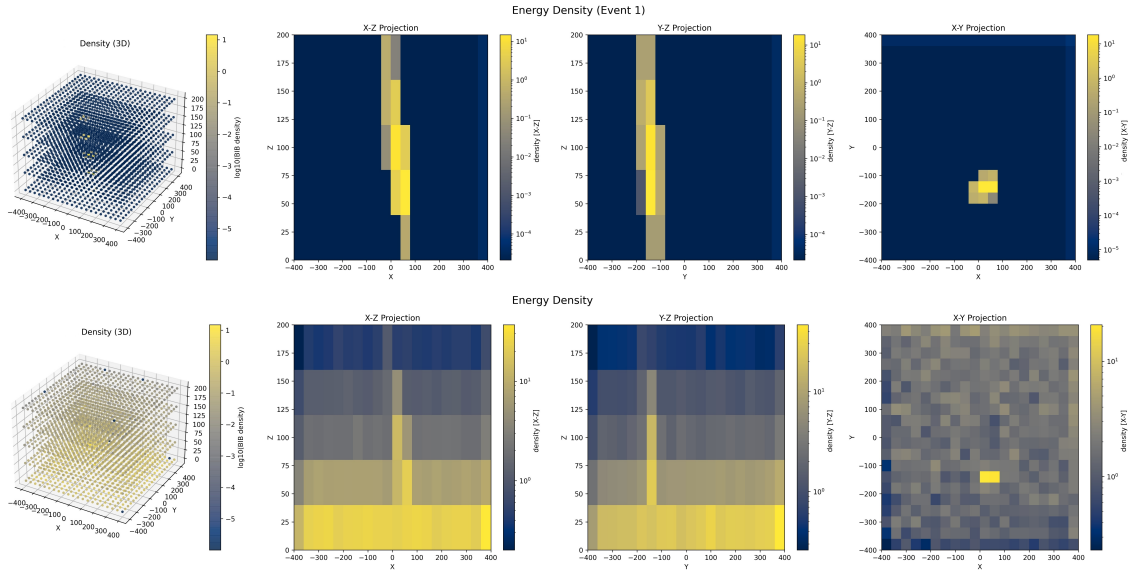


Figure 5.10: Visualization of photon shower generated with the developed framework. Top row: signal only, bottom row: signal with overlaid BIB

Figure 5.10 illustrates a generated event within this framework, where for visualization purposes we grouped the Voronoi regions in histogram bins. The centroid distribution is arranged in 5 regular layers, reproducing the original Crilin configuration of Section 3.1.2. Once the BIB is overlaid to the signal, most of the information on the lowest voxels is lost, but the shower shape becomes clearly distinguishable in the center of the detection region.

5.5 Summary and interface reconstruction

Each event generated by the pipeline consists of a set of centroids, the corresponding energy deposits, and their Voronoi cell volumes. This structured output forms the natural interface to the reconstruction stage, where centroids act as spatial carriers of information, and the energy vector E becomes the primary input to the downstream learning model. The volumetric metadata V may optionally be used to normalize or reweight contributions based on heterogeneous cell sizes.

By construction, the entire generation process is differentiable with respect to both the geometric parameters (e.g. centroid positions) and the physical parameters of the incoming particle (e.g. direction and entry point). This allows gradients computed during reconstruction loss backpropagation to flow upstream, enabling joint optimization of the detector geometry and the reconstruction algorithm in a unified training loop.

The generated events capture both stochastic and structured aspects of calorimetric

activity. The signal component is randomized in its angular incidence and transverse entry point, ensuring diversity in spatial configurations. The background component is sampled from a smooth, learned distribution that models the diffuse flux of beam-induced particles. The combined output is therefore statistically rich enough to support training, while remaining computationally lightweight compared to full Geant4 or Fluka simulations.

In the next section, we describe how this surrogate-generated dataset is used to train reconstruction models and to drive end-to-end optimization of the calorimeter layout. We will also explore how the performance of the learned reconstruction informs design choices and contributes to the broader goal of detector co-optimization.

Chapter 6

Optimization runs and remarks

6.1 Pipeline implementation

The optimization workflow described in chapter 5 is implemented in a unified executable script openly available at https://github.com/FedericoNardi/HGCalML/blob/master/scripts/test_pipeline.py. This serves as the main entry point for launching a full differentiable optimization of the Muon Collider ECAL layout. The script connects all components of the pipeline -from event generation to loss evaluation- into a single, reproducible process where graph structure is preserved and gradients can be calculated for differentiation purposes.

Built on top of the differentiable tools provided in the `genModules` package [141], the script orchestrates the generation of synthetic calorimetric events using surrogate models for both beam-induced backgrounds (BIB) and photon showers. These events are then processed by a reconstruction network, specifically an adaptation of the Object Condensation (OC) model described in Chapter 4. In this context, the OC architecture is employed with a simplified loss head that predicts the signal fraction for each centroid (i.e., the likelihood that a given region contains part of the primary photon-induced shower). This scalar output per hit serves as the basis for the optimization objective.

The focus of this implementation is on tractable optimization under realistic constraints. Rather than simulating the full detector volume at native resolution, a reduced subregion of the calorimeter is considered. To further reduce memory consumption, the inter-centroid spacing along the z and x direction is doubled with respect to the target resolution, trading off granularity for computational feasibility. These simplifications allow meaningful optimization experiments to be carried out on a single GPU, while maintaining the essential structure and differentiability of the simulation.

In this section we describe the configuration logic of the input parameters, the con-

struction of the centroid grid, and the main forward-pass operations that define a single optimization step.

6.1.1 Input configuration and geometry setup

The execution of the pipeline begins with the definition of the simulated detector geometry and the configuration of the physical parameters associated with each event. These include the spatial layout of centroids used to discretize the calorimeter, as well as the kinematic properties of the incoming signal particles.

The detector volume is represented by a regular grid of centroids spanning a fixed cuboidal region within the ECAL. The grid is arranged in 5 layers to mock the original Crilin structure. Each centroid defines the generator of a Voronoi region, which acts as the fundamental unit of spatial resolution for the event simulation. Due to memory limitations, only a partial slice of the full detector geometry is simulated. In particular, the spacing between centroids is doubled along the beam axis z and x (horizontal) directions, reducing the total number of regions while preserving the overall structure of the pipeline. This coarse discretization allows the simulation to fit comfortably within the memory budget of a single GPU, while still capturing the spatial features needed to guide the optimization process.

In addition to geometry, the script specifies the generation of randomized signal events. Each photon is assigned a direction of incidence, sampled from a distribution over (θ, ϕ) spherical angles, and an entry point drawn from a uniform distribution over the front face of the simulated volume. These parameters are used to rototranslate the shower surrogate into calorimeter coordinates, ensuring that each generated event represents a physically plausible configuration. The background energy, on the other hand, is drawn from a fixed surrogate model representing the beam-induced background (BIB), which returns a smooth, event-independent 3D energy density.

The centroids are initialized as a trainable TensorFlow variable, enabling direct optimization of their spatial coordinates through gradient descent. Once the geometry is established, the surrogate models are queried: the BIB model provides a smooth density that is integrated over Voronoi volumes, while the signal model returns a coarse but structured grid of energy deposits, which is transformed and mapped onto the same Voronoi layout.

The resulting event —represented as a set of centroid coordinates, energy deposits, and cell volumes— is ready to be passed through the reconstruction model.

6.1.2 Pipeline workflow overview

Once a full event has been generated—comprising centroid coordinates, energy deposits, and Voronoi volumes—it is passed to the reconstruction model for signal identification and energy recovery. In the current implementation, the reconstruction stage is handled by the Object Condensation (OC) network introduced in Chapter 4, with a modification to its output structure tailored to the optimization task at hand.

Rather than predicting clustered object properties or condensation scores, the OC model is configured to produce a single scalar value per input centroid: the predicted signal fraction. This value lies in the range $[0,1]$ and reflects the model's estimate of how much of the energy deposited in that region originates from the primary signal shower, as opposed to the background. The simplified architecture allows for more stable training in the early stages of optimization, while still capturing the essential information needed to guide geometry refinement. This simplified model has been trained on a dataset of 10000 events generated with the surrogate described in Chapter 5. Gaussian smearing was applied from the regular starting grid, shown also in Figure 5.10, make the model more solid when handling irregular geometries.

The reconstruction of the signal energy is then performed by applying the predicted signal fractions to the original energy deposits. For each centroid i, the reconstructed signal energy is computed as

$$\hat{E}^{(i)}_{signal} = \sigma^{(i)} E^{(i)}$$

where $E^{(i)}$ is the total energy assigned to the i-th centroid and $\sigma^{(i)}$ is the predicted signal fraction. The total reconstructed signal energy is obtained by summing over all the N centroids:

$$\hat{E}_{signal} = \sum_{i=1}^{N} \sigma^{(i)} E^{(i)}. \tag{6.1}$$

This reconstruction is compared to the known true signal energy, which is available from the simulation since the photon and background component are generated independently.

The loss $\mathcal{L}$ is computed as the relative absolute deviation between the predicted and true signal energies:

$$\mathcal{L} = \frac{|\hat{E}_{signal} - E_{signal}|}{E_{signal}}.$$

This scalar loss is fully differentiable with respect to the signal fraction predictions and, crucially, with respect to the centroid positions themselves. As such, gradients can be propagated not only through the OC network but also through the event generation pipeline, allowing the calorimeter layout to be refined in a direction that improves signal

recovery.

This setup defines the core of the forward pass for a single event. The next stage handles the batching of events, application of gradient updates, and accumulation of training metrics across multiple optimization steps.

6.1.3 Optimization loop and diagnostics

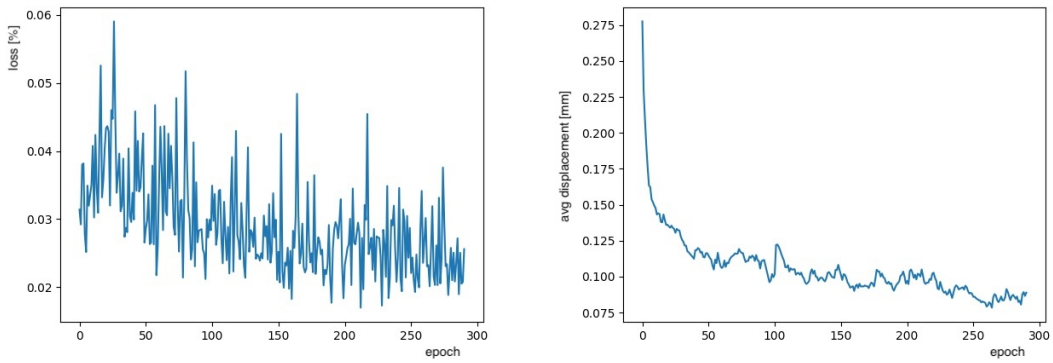
With the forward pass and loss function defined, the pipeline proceeds to execute the optimization loop that drives the refinement of the calorimeter geometry. This loop updates the positions of the centroids by minimizing the reconstruction loss using gradient-based optimization.

To evaluate reconstruction performance, a batch of 128 synthetic events is simulated for each optimization epoch. Each event corresponds to a unique configuration of primary photon direction and entry point, while the BIB background remains fixed. These parameters are used by the event generator to produce per-centroid energy deposits, which are then passed through the Object Condensation network to predict the signal fraction per region. The reconstructed signal energy is computed and compared to the known true value, yielding a scalar loss per event. The centroids themselves are trainable TensorFlow variables, updated using an Adam optimizer.

Due to memory constraints, the full batch cannot be processed simultaneously. Instead, events are handled one at a time, and the corresponding gradients are accumulated across multiple forward-backward passes. Only once all 128 events have been processed are the accumulated gradients applied in a single optimizer step. This strategy preserves the statistical benefits of larger batch sizes while maintaining the memory footprint of single-event inference. The accumulation loop is implemented using native TensorFlow constructs to ensure compatibility with automatic differentiation and GPU execution.

Throughout the training process, key diagnostics such as reconstruction loss, predicted signal energy, and centroid displacements are logged to disk. Event visualizations can be saved periodically to monitor qualitative progress, and model checkpoints are stored at regular intervals to enable recovery or retrospective analysis.

This loop forms the core mechanism by which the detector geometry is optimized in response to reconstruction performance. It enables the exploration of learned layouts that improve signal separation and energy recovery, opening the door to principled, gradient-driven detector design studies.



(a) (b)

Figure 6.1: Validation plots. Training loss (a) and Average centroid displacement (b) during optimization cycle.

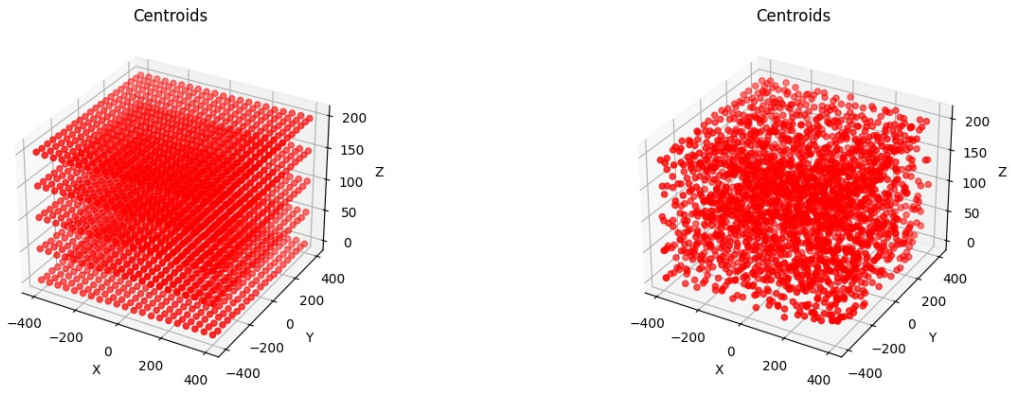
6.2 Optimization run

In order to run the algorithm described above, we first performed some preliminary tuning steps for the learning rate parameter of the optimizer. We find an optimal value lies at $O(10)$, allowing for large enough centroids displacement while still allowing the loss to decrease throughout the epochs.

The run discussed in this section sees the implementation of a linear learning rate scheduler, ranging from 1 to 0.1 across 500 epochs. The energy for the batch events is sampled uniformly in the interval [10,175]GeV. The run in question has been arrested just before epoch 300 due to a maintenance shutdown of the server where the code baseline is hosted and executed. Figure 6.1 shows the reconstruction loss of the optimization run, as well as the average displacement of centroid positions throughout the training run. Despite the still considerable noise fluctuations in the loss, and the relatively small average displacement -suggesting that some more model tuning and longer runs might be required- we present the following result as a proof of concept, serving as a benchmark for a functioning pipeline producing physically sensible results.

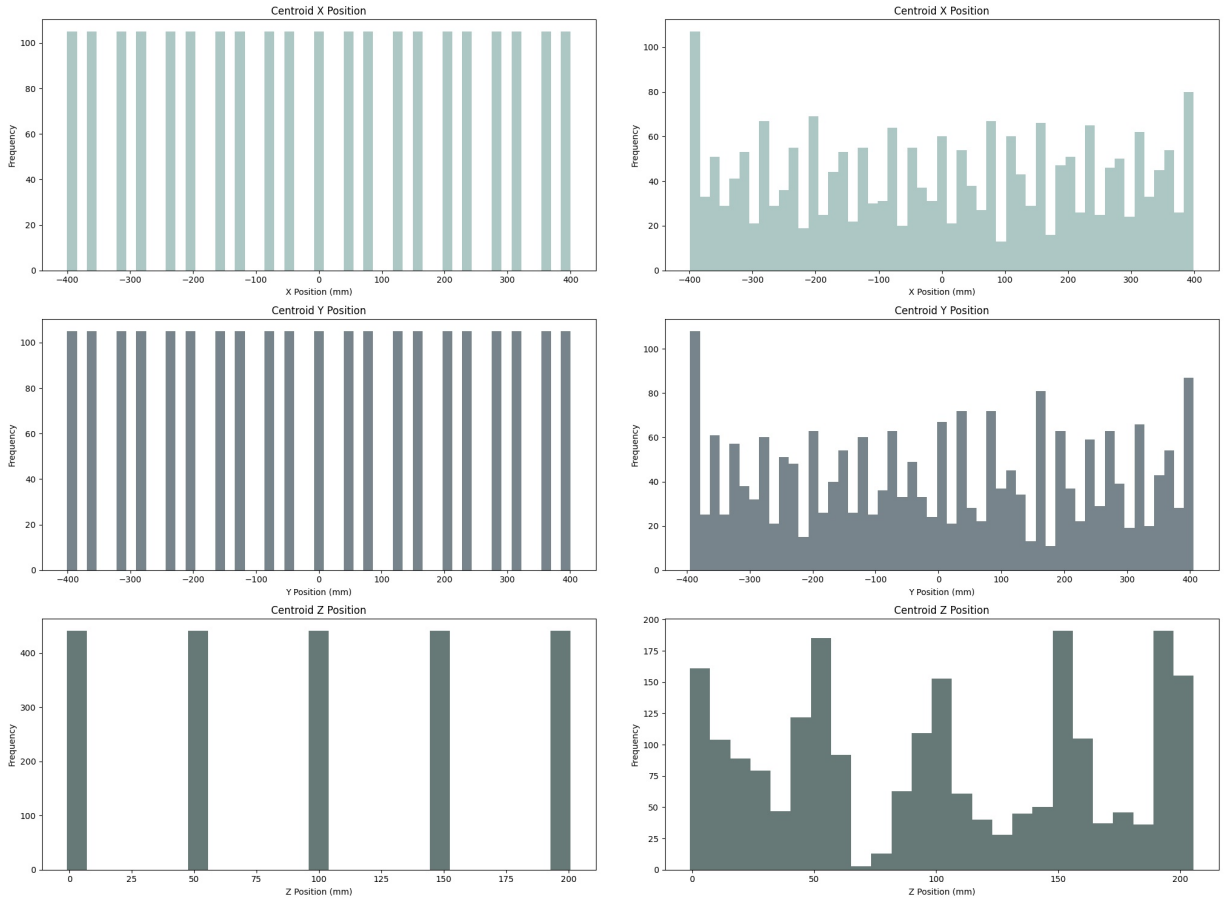
A first visualization of the evolution of the system of centroids is presented in Figure 6.2, from an initial Crilin-like system to a more chaotic -yet physically motivated- arrangement. Such plot however is only informative of the evolution of the system, and the actual ability of the pipeline to move voxel centroids.

More insight is provided in Figure 6.3, showing the voxel distribution along each coordinate before and after the run. The irregularity of the new grid is highlighted especially in the x and y distribution. Looking at the z distribution however, we notice the tendency to reduce the gap between the first two layers.



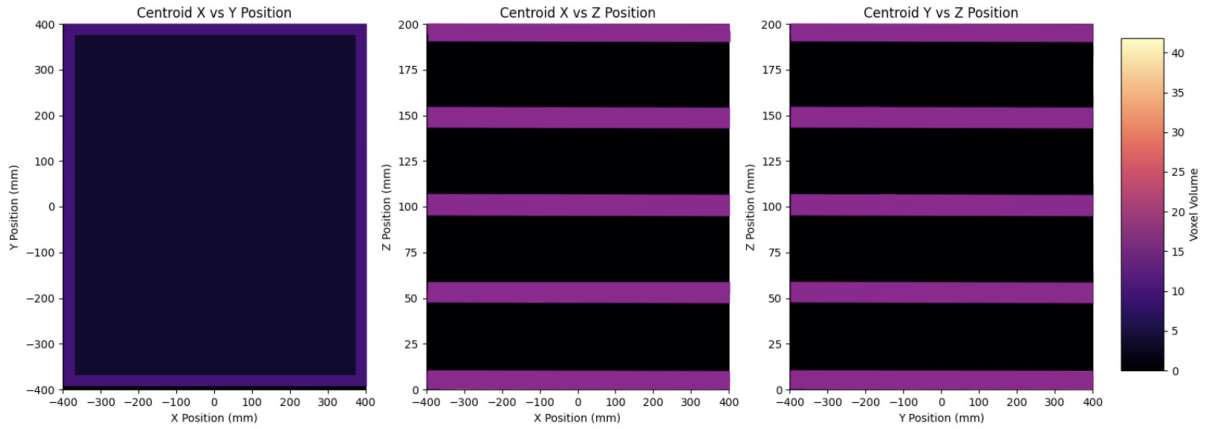
(a) (b)

Figure 6.2: Visualizations of centroid locations at the beginning (a) and end (b) of optimization cycle.

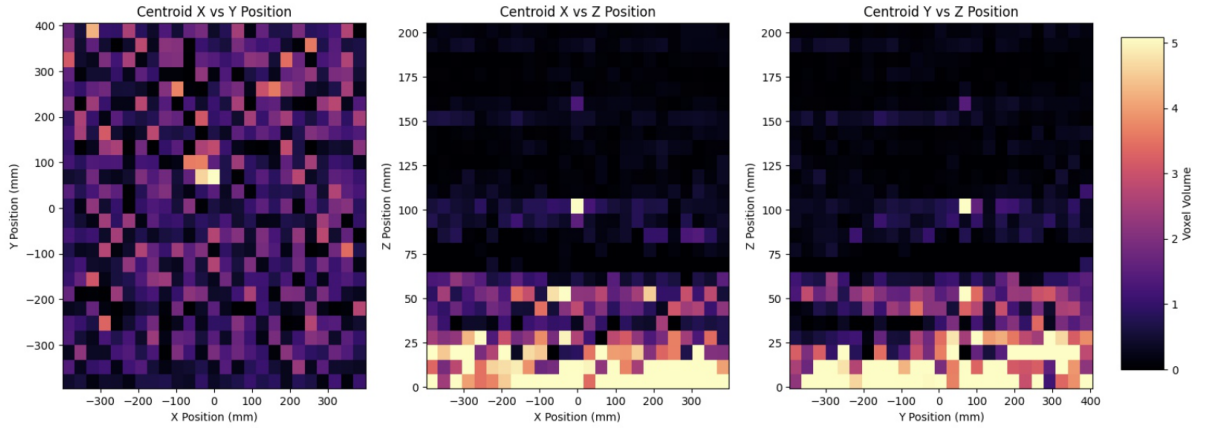


(a) (b)

Figure 6.3: Histograms of centroid locations at the beginning (a) and end (b) of optimization cycle.



(a)



(b)

Figure 6.4: Heat map of voxel volumes at the beginning (a) and end (b) of optimization cycle.

This latter feature is even more evident from Figure 6.4, where a heatmap of the voxel volumes projected along the three cartesian planes are presented. In the lower layers the voxel volumes are significantly larger, meaning that fewer cells are left in the lower layer encouraging the diffusion towards higher z-regions (here the z axis refers to the vertical axis, orthogonal to the beam). Comparing this to the event visualizations from Figure 5.10, we can motivate this displacement as the formation of an initial low-resolution layer, where the voxels act as absorbers and mitigate the presence of BIB in the upper layers, thus not requiring a big spatial resolution. More information on the signal is already available at upper layers (even before the 5cm of the original configuration), where even with some residual BIB the reconstruction seems to yield better results than on a regular grid.

6.3 Final remarks

The results presented in this chapter should be interpreted mainly as a proof of concept, showcasing the effectiveness of Automatic Differentiation approaches to an experimental design optimization study. The identified configuration at this level is not -and neither claims to be- the ideal configuration for a Crilin-like Muon Collider Calorimeter, due to the extreme approximations we took developing the pipeline -among them completely neglecting the presence of passive electronic elements of which each voxel is actually instrumented.

However, this should be considered a benchmark for a different paradigm that could assist the R&D phase and potentially lead to performance gains. Seeing this as the starting point for more advanced studies would be beneficial to analyze the detection system from a more general point of view, and target it to the specific requirements of the collider machine. In particular, more physically motivated results could arise with the implementation of relatively few model improvements:

- *Loss function*: the introduction of new terms in the loss function, such as material cost scaling with voxel volume, or regularization parameters to favor more ordinate configurations. This could push the argument to inspire a setup that is actually physically realizable

- *Multi-headed reconstruction*: related to the introduction of new loss terms is also the definition of other target reconstruction parameters. Adding an output layer that reconstructs the position of the shower as well for instance, could lead to a more fine trade-off between BIB absorption and spatial resolution of the voxel system.

The modularity of the pipeline also allows for further refinements on the surrogate side. Studies to improve the signal generation wit Diffusion Models are already ongoing [142], a differentiable model of BIB at 10TeV (or even a model conditioned to the collision energy) would help assess the potential performance of the target final stage Muon Collider detector.

Chapter 7

Final remarks

Throughout this thesis we give an overview of Automatic Differentiation frameworks, highlighting their advantages and their power when employed to generalize Deep Learning algorithms and set up optimization problems. This is a novel paradigm that would provide a set of tools to approach experimental R&D from a co-design point of view, where the interconnections and shared information between different subsystems are considered, instead of just focusing on the optimization of each single component. It is important to note that such approach does not aim to -and neither should- fully replace 'traditional' development procedures, but rather complement them to suggest starting points, or configurations that might be counterintuitive or otherwise not even considered. The optimization study of the SWGO detector array of Chapter 2 illustrates exactly such a situation, where the optimization process converges to layouts suggesting that a bulk array density close to the center of the detection area is actually unnecessary as it tends to overinstrument a region very close to a null measure point.

After this first example, relying on analytical gradient calculations, we focus on the main part of this thesis' work, which sees the development of the framework to optimize the geometry of the Crilin Muon Collider calorimeter. Such a case is particularly interesting for a series of reasons:

- The debate on Future Colliders is at its peak, as the scientific community needs to agree on a post-LHC program, granting the best possible compromise between high-energy and high-precision reach in the next decades;

- The Muon Collider described in Chapter 3 is a particularly interesting machine for the novel type of collisions offered, allowing for lepton-collider precision at hadron-collider energies. Feasibility studies are needed, especially in terms of BIB mitigation potential, and modern Machine Learning approaches such as the

Object Reconstruction with Graph Neural Networks described in Chapter 4 could provide significant performance gains;

- The Crilin detector, with its modular voxel structure is a great candidate for geometry optimization studies. It allows for straightforward array structure generalizations and it can be resized at prototyping phase for relatively little production costs [112].

We describe in Chapter 5 the efforts to develop surrogates to allow us to circumvent non-differentiability issues in traditional simulators such as Geant4 [135]–[137]. Despite promising ongoing efforts [144], a full differentiable implementation is not yet supported, we must rely on different tools to get a satisfying description of physical processes. An advantage of having independent modular surrogates is that further developments and patches can be implemented without modifying the basic structure of the pipeline, thus allowing for future developments to improve physical realness.

The implementation of the pipeline itself is described in Chapter 6, where we also showcase a first benchmark result from a first optimization look. This serves as a proof of concept to set the basis of more advanced studies, where increasing event complexity and the definition of targeted utility functions would yield configurations that meet the targets discussed within the collaboration. A broader discussion on utilities has proved particularly useful in the SWGO case [145], underlining the need to discuss performance and objectives with the whole group to actually tailor the code to yield the best configuration for the experimental objectives.

Bibliography

[1] MODE Collaboration, *Mode collaboration - machine-learning optimized design of experiments*, https://mode-collaboration.github.io, Accessed: 2025-05-19, 2024.

[2] K. Cranmer, J. Brehmer, and G. Louppe, "The frontier of simulation-based inference," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 055–30 062, 2020. arXiv: 1911.01429 [stat.ML].

[3] Y. Mishnayot, M. Layani, I. Cooperstein, S. Magdassi, and G. Ron, "Three-dimensional printing of scintillating materials," *Rev. Sci. Instrum.*, vol. 85, p. 085 102, 2014. doi: 10.1063/1.4891703. arXiv: 1406.4817 [cond-mat.mtrl-sci].

[4] G. Giacomini, W. Chen, G. D'Amen, and A. Tricoli, "Fabrication and performance of AC-coupled LGADs," *JINST*, vol. 14, no. 09, P09004, 2019. doi: 10.1088/1748-0221/14/09/p09004. arXiv: 1906.11542 [physics.ins-det].

[5] *Particle Track Reconstruction with Deep Learning*, 2017. [Online]. Available: {https://dl4physicalsciences.github.io/files/nips_dlps_2017_28.pdf}.

[6] S. Farrell *et al.*, "Novel deep learning methods for track reconstruction," *4th International Workshop Connecting The Dots 2018*, 2018. arXiv: 1810.06111 [hep-ex].

[7] S. Amrouche *et al.*, "The Tracking Machine Learning Challenge: Accuracy Phase," in *The NeurIPS '18 Competition*, Springer International Publishing, 2020, p. 231, isbn: 978-3-030-29135-8. doi: 10.1007/978-3-030-29135-8_9.

[8] X. Ju *et al.*, "Graph Neural Networks for Particle Reconstruction in High Energy Physics detectors," *33rd Annual Conference on Neural Information Processing Systems*, 2020. arXiv: 2003.11603 [physics.ins-det].

[9] S. Akar, T. J. Boettcher, S. Carl, *et al.*, *An updated hybrid deep learning algorithm for identifying and locating primary vertices*, 2020. arXiv: 2007.01023 [physics.ins-det].

[10] J. Shlomi, S. Ganguly, E. Gross, *et al.*, "Secondary Vertex Finding in Jets with Neural Networks," *Eur. Phys. J. C*, vol. 81, p. 540, 2021. doi: 10.1140/epjc/s10052-021-09342-y.

[11] N. Choma *et al.*, *Track Seeding and Labelling with Embedded-space Graph Neural Networks*, 2020. arXiv: 2007.00149 [physics.ins-det].

[12] F. Siviero, R. Arcidiacono, N. Cartiglia, *et al.*, "First application of machine learning algorithms to the position reconstruction in resistive silicon detectors," *Journal of Instrumentation*, vol. 16, P03019, 2021. doi: 10.1088/1748-0221/16/03/p03019. [Online]. Available: https://doi.org/10.1088/1748-0221/16/03/p03019.

[13] P. J. Fox, S. Huang, J. Isaacson, X. Ju, and B. Nachman, "Beyond 4D Tracking: Using Cluster Shapes for Track Seeding," *JINST*, vol. 16, no. 05, P05001, 2021. doi: 10.1088/1748-0221/16/05/P05001. arXiv: 2012.04533 [physics.ins-det].

[14] S. Amrouche, M. Kiehn, T. Golling, and A. Salzburger, "Hashing and metric learning for charged particle tracking," *33rd Annual Conference on Neural Information Processing Systems*, 2021. arXiv: 2101.06428 [hep-ex].

[15] K. Goto, T. Suehara, T. Yoshioka, *et al.*, *Development of a Vertex Finding Algorithm using Recurrent Neural Network*, 2021. arXiv: {2101.11906} [physics.data-an].

[16] C. Biscarat, S. Caillou, C. Rougier, J. Stark, and J. Zahreddine, "Towards a realistic track reconstruction algorithm based on graph neural networks for the HL-LHC," in *25th International Conference on Computing in High-Energy and Nuclear Physics*, 2021. arXiv: 2103.00916 [physics.ins-det].

[17] Akar, Simon, Atluri, Gowtham, Boettcher, Thomas, *et al.*, "Progress in developing a hybrid deep learning algorithm for identifying and locating primary vertices," *EPJ Web Conf.*, vol. 251, p. 04012, 2021. doi: 10.1051/epjconf/202125104012. [Online]. Available: https://doi.org/10.1051/epjconf/202125104012.

[18] S. Thais and G. DeZoort, *Instance Segmentation GNNs for One-Shot Conformal Tracking at the LHC*, 2021. arXiv: 2103.06509 [cs.CV].

[19] X. Ju *et al.*, "Performance of a geometric deep learning pipeline for HL-LHC particle tracking," *Eur. Phys. J. C*, vol. 81, p. 876, 2021.

[20] G. Dezoort, S. Thais, I. Ojalvo, *et al.*, *Charged particle tracking via edge-classifying interaction networks*, 2021. arXiv: 2103.16701 [hep-ex].

[21] A. Edmonds, D. Brown, L. Vinas, and S. Pagan, "Using machine learning to select high-quality measurements," *Journal of Instrumentation*, vol. 16, T08010, 2021. doi: 10.1088/1748-0221/16/08/t08010. [Online]. Available: https://doi.org/10.1088/1748-0221/16/08/t08010.

[22] E. Lavrik, M. Shiroya, H. Schmidt, A. Toia, and J. Heuser, "Optical inspection of the silicon micro-strip sensors for the cbm experiment employing artificial intelligence," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1021, p. 165 932, 2022, issn: 0168-9002. doi: https://doi.org/10.1016/j.nima.2021.165932. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900221008950.

[23] B. Huth, A. Salzburger, and T. Wettig, "Machine learning for surface prediction in ACTS," in *25th International Conference on Computing in High-Energy and Nuclear Physics*, 2021. arXiv: 2108.03068 [physics.ins-det].

[24] P. Goncharov, E. Schavelev, A. Nikolskaya, and G. Ososkov, "Ariadne: PyTorch Library for Particle Track Reconstruction Using Deep Learning," in *24th International Scientific Conference of Young Scientists and Specialists*, 2021. arXiv: 2109.08982 [physics.data-an].

[25] A. Lazar *et al.*, *Accelerating the Inference of the Exa.TrkX Pipeline*, 2022. arXiv: 2202.06929 [physics.ins-det].

[26] M. Benedikt, A. Blondel, P. Janot, M. Mangano, and F. Zimmermann, "Future Circular Colliders succeeding the LHC," *Nature Phys.*, vol. 16, no. 4, pp. 402–407, 2020. doi: 10.1038/s41567-020-0856-2.

[27] D. Hernandez and T. B. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," *arXiv e-prints*, arXiv:2005.04305, arXiv:2005.04305, May 2020. arXiv: 2005.04305 [cs.LG].

[28] A. G. Baydin, K. Cranmer, P. de Castro Manzano, *et al.*, "Toward machine learning optimization of experimental design," *Nuclear Physics News*, vol. 31, no. 1, pp. 25–28, 2021. doi: 10.1080/10619127.2021.1881364. eprint: https:

//doi.org/10.1080/10619127.2021.1881364. [Online]. Available: https://doi.org/10.1080/10619127.2021.1881364.

[29] J. Kieseler, "Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph and image data," *Eur. Phys. J. C*, vol. 80, no. 9, p. 886, 2020. doi: 10.1140/epjc/s10052-020-08461-2. arXiv: 2002.03605 [physics.data-an].

[30] R. M. Hicks, E. M. Murman, and G. N. Vanderplaats, "An assessment of airfoil design by numerical optimization," NASA Ames Research Center, Tech. Rep., 1974.

[31] R. M. Hicks and P. A. Henne, "Wing design by numerical optimization," *Journal of Aircraft*, vol. 15, no. 7, pp. 407–412, 1978, issn: 0021-8669, 1533-3868. doi: 10.2514/3.58379. [Online]. Available: https://arc.aiaa.org/doi/10.2514/3.58379 (visited on 03/04/2022).

[32] O. Pironneau, "On optimum design in fluid mechanics," *Journal of Fluid Mechanics*, vol. 64, no. 1, pp. 97–110, Jun. 3, 1974, issn: 0022-1120, 1469-7645. doi: 10.1017/S0022112074002023. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0022112074002023/type/journal_article (visited on 03/04/2022).

[33] O. Pironneau, "On optimum profiles in stokes flow," *Journal of Fluid Mechanics*, vol. 59, no. 1, pp. 117–128, Jun. 5, 1973, issn: 0022-1120, 1469-7645. doi: 10.1017/S002211207300145X. [Online]. Available: https://www.cambridge.org/core/product/identifier/S002211207300145X/type/journal_article (visited on 03/04/2022).

[34] A. Jameson, "Aerodynamic design via control theory," *Journal of Scientific Computing*, vol. 3, no. 3, pp. 233–260, Sep. 1988, issn: 0885-7474, 1573-7691. doi: 10.1007/BF01061285. [Online]. Available: http://link.springer.com/10.1007/BF01061285 (visited on 03/04/2022).

[35] M. Towara and U. Naumann, "A discrete adjoint model for OpenFOAM," *Procedia Computer Science*, vol. 18, pp. 429–438, 2013, issn: 18770509. doi: 10.1016/j.procs.2013.05.206. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050913003499 (visited on 03/04/2022).

[36] T. Albring, M. Sagebaum, and N. R. Gauger, "Efficient aerodynamic design using the discrete adjoint method in SU2," *AIAA 2016-3518*, Jul. 2, 2016, published.

[37] M. Luers, M. Sagebaum, S. Mann, J. Backhaus, D. Grossmann, and N. R. Gauger, “Adjoint-based volumetric shape optimization of turbine blades,” in *2018 Multidisciplinary Analysis and Optimization Conference*, Atlanta, Georgia: American Institute of Aeronautics and Astronautics, Jun. 25, 2018, isbn: 9781624105500. doi: 10.2514/6.2018-3638. [Online]. Available: https://arc.aiaa.org/doi/10.2514/6.2018-3638 (visited on 02/25/2022).

[38] A. Nemili, E. Özkaya, N. R. Gauger, F. Kramer, and F. Thiele, “Accurate discrete adjoint approach for optimal active separation control,” *AIAA Journal*, vol. 55, no. 9, pp. 3016–3026, Sep. 2017, issn: 0001-1452, 1533-385X. doi: 10.2514/1.J055009. [Online]. Available: https://arc.aiaa.org/doi/10.2514/1.J055009 (visited on 02/25/2022).

[39] B. Y. Zhou, S. Ryong Koh, N. R. Gauger, M. Meinke, and W. Schöder, “A discrete adjoint framework for trailing-edge noise minimization via porous material,” *Computers & Fluids*, vol. 172, pp. 97–108, Aug. 2018, issn: 00457930. doi: 10.1016/j.compfluid.2018.06.017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0045793018303451 (visited on 02/25/2022).

[40] R. Bombardieri, R. Cavallaro, R. Sanchez, and N. R. Gauger, “Aerostructural wing shape optimization assisted by algorithmic differentiation,” *Structural and Multidisciplinary Optimization*, vol. 64, no. 2, pp. 739–760, Aug. 2021, issn: 1615-147X, 1615-1488. doi: 10.1007/s00158-021-02884-5. [Online]. Available: https://link.springer.com/10.1007/s00158-021-02884-5 (visited on 02/25/2022).

[41] M. Morlighem, D. Goldberg, T. D. dos Santos, J. Lee, and M. Sagebaum, “Mapping the sensitivity of the Amundsen sea embayment to changes in external forcings using automatic differentiation,” *Geophysical Research Letters*, vol. 48, no. 23, Nov. 24, 2021. doi: https://doi.org/10.1029/2021GL095440. (visited on 11/24/2021), published.

[42] J. Andersson, J. Åkesson, and M. Diehl, “Casadi: A symbolic package for automatic differentiation and optimal control,” in *Recent advances in algorithmic differentiation*, Springer, 2012, pp. 297–307.

[43] Y. Achdou and O. Pironneau, *Computational Methods for Option Pricing*. Society for Industrial and Applied Mathematics, Jan. 2005, isbn: 9780898715736 9780898717495. doi: 10.1137/1.9780898717495. [Online]. Available: http://epubs.siam.org/doi/book/10.1137/1.9780898717495 (visited on 03/04/2022).

[44] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[45] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[46] J. Janai, F. Güney, A. Behl, A. Geiger, *et al.*, “Computer vision for autonomous vehicles: Problems, datasets and state of the art,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[47] Y. Goldberg, “Neural network methods for natural language processing,” *Synthesis lectures on human language technologies*, vol. 10, no. 1, pp. 1–309, 2017.

[48] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[49] A. Paszke *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[50] M. Abadi, P. Barham, J. Chen, *et al.*, “{Tensorflow}: A system for {large-scale} machine learning,” in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[51] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, “Automatic differentiation in machine learning: A survey,” *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 5595–5637, 2017, issn: 1532-4435.

[52] C. Olah, *Neural networks, types, and functional programming*, Retrieved on 17-03-2022, 2015. [Online]. Available: http://colah.github.io/posts/2015-09-NN-Types-FP/.

[53] D. Dalrymple, *Differentiable programming as the 2016 most important recent scientific news*, Retrieved on 17-03-2022, 2016. [Online]. Available: https://www.edge.org/response-detail/26794.

[54] Y. LeCun, *Facebook post on differentiable programming*, Retrieved on 17-03-2022, 2018. [Online]. Available: https://www.facebook.com/yann.lecun/posts/10155003011462143.

[55] S. Shirobokov, V. Belavin, M. Kagan, A. Ustyuzhanin, and A. G. Baydin, *Black-Box Optimization with Local Generative Surrogates*, 2020. arXiv: 2002.04632 [cs.LG].

[56] T. Dorigo, “Geometry optimization of a muon-electron scattering detector,” *Physics Open*, vol. 4, p. 100 022, 2020.

[57] F. Ratnikov, “Using machine learning to speed up and improve calorimeter R&D,” *Journal of Instrumentation*, vol. 15, no. 05, p. C05032, 2020.

[58] E. Cisbani *et al.*, “AI-optimized detector design for the future Electron-Ion Collider: the dual-radiator RICH case,” *Journal of Instrumentation*, vol. 15, no. 05, P05009, 2020.

[59] A. Edelen, N. Neveu, M. Frey, Y. Huber, C. Mayes, and A. Adelmann, “Machine learning for orders of magnitude speedup in multiobjective optimization of particle accelerator systems,” *Physical Review Accelerators and Beams*, vol. 23, no. 4, p. 044 601, Apr. 2020, Publisher: American Physical Society. doi: 10.1103/PhysRevAccelBeams.23.044601. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevAccelBeams.23.044601 (visited on 04/09/2020).

[60] D. Koser, L. Waites, D. Winklehner, M. Frey, A. Adelmann, and J. Conrad, “Input beam matching and beam dynamics design optimization of the IsoDAR RFQ using statistical and machine learning techniques,” *arXiv:2112.02579 [physics]*, 2021, (Submitted to Frontiers in Physics). arXiv: 2112.02579. [Online]. Available: http://arxiv.org/abs/2112.02579 (visited on 03/07/2022).

[61] F. Van Der Veken, G. Azzopardi, F. Blanc, *et al.*, “Machine learning in accelerator physics: Applications at the CERN Large Hadron Collider,” in *Proceedings of Artificial Intelligence for Science, Industry and Society PoS(AISIS2019)*, vol. 372, SISSA Medialab, Jul. 2020, p. 044. [Online]. Available: https://pos.sissa.it/372/044/ (visited on 08/31/2020).

[62] T. D. et al., “Toward the end-to-end optimization of particle physics instruments with differentiable programming,” *Reviews in Physics*, vol. 10, p. 100 085, 2023, issn: 2405-4283. doi: https://doi.org/10.1016/j.revip.2023.100085. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405428323000047.

[63] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nature Methods*, vol. 17, pp. 261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[64] D. F. Shanno, “Conditioning of quasi-newton methods for function minimization,” *Mathematics of Computation*, vol. 24, no. 111, pp. 647–656, 1970, issn: 0025-5718, 1088-6842. doi: 10.1090/S0025-5718-1970-0274029-X. [Online].

Available: https://www.ams.org/mcom/1970-24-111/S0025-5718-1970-0274029-X/ (visited on 03/18/2022).

[65] D. Goldfarb, "A family of variable-metric methods derived by variational means," *Mathematics of Computation*, vol. 24, no. 109, pp. 23–26, 1970, issn: 0025-5718, 1088-6842. doi: 10.1090/S0025-5718-1970-0258249-6. [Online]. Available: https://www.ams.org/mcom/1970-24-109/S0025-5718-1970-0258249-6/ (visited on 03/18/2022).

[66] R. Fletcher, "A new approach to variable metric algorithms," *The Computer Journal*, vol. 13, no. 3, pp. 317–322, Mar. 1, 1970, issn: 0010-4620, 1460-2067. doi: 10.1093/comjnl/13.3.317. [Online]. Available: https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/13.3.317 (visited on 03/18/2022).

[67] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, Sep. 1995, issn: 1064-8275, 1095-7197. doi: 10.1137/0916069. [Online]. Available: http://epubs.siam.org/doi/10.1137/0916069 (visited on 03/18/2022).

[68] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software*, vol. 23, no. 4, pp. 550–560, Dec. 1997, issn: 0098-3500, 1557-7295. doi: 10.1145/279232.279236. [Online]. Available: https://dl.acm.org/doi/10.1145/279232.279236 (visited on 03/18/2022).

[69] R. L. Burden, J. D. Faires, and A. M. Burden, *Numerical Analysis*. Cengage Learning, 2015.

[70] Wolfram Research, Inc., *Mathematica 8.0*, version 0.8, 2010. [Online]. Available: https://www.wolfram.com.

[71] A. Griewank and A. Walther, *Evaluating Derivatives* (Other Titles in Applied Mathematics). Society for Industrial and Applied Mathematics, Jan. 1, 2008, 448 pp., isbn: 9780898716597. doi: 10.1137/1.9780898717761. [Online]. Available: https://epubs.siam.org/doi/book/10.1137/1.9780898717761 (visited on 01/06/2022).

[72] R. E. Wengert, "A simple automatic derivative evaluation program," *Communications of the ACM*, vol. 7, no. 8, pp. 463–464, 1964.

[73] S. Linnainmaa, “The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors,” Ph.D. dissertation, Master’s Thesis (in Finnish), Univ. Helsinki, 1970.

[74] B. Speelpenning, “Compiling fast partial derivatives of functions given by algorithms,” Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1980.

[75] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[76] A. Walther and A. Griewank, “Getting started with ADOL-C,” in *Combinatorial Scientific Computing*, U. Naumann and O. Schenk, Eds., Chapman-Hall CRC Computational Science, 2012, ch. 7, pp. 181–202.

[77] R. J. Hogan, “Fast reverse-mode automatic differentiation using expression templates in C++,” *ACM Transactions on Mathematical Software*, vol. 40, no. 4, 26:1–26:24, 2014. [Online]. Available: http://doi.acm.org/10.1145/2560359.

[78] J. Lotz, “Hybrid approaches to adjoint code generation with dco/c++,” Dissertation, Department of Computer Science, RWTH Aachen University, 2016. [Online]. Available: http://publications.rwth-aachen.de/record/667318.

[79] M. Sagebaum, T. Albring, and N. Gauger, “High-performance derivative computations using codipack,” *ACM Transactions on Mathematical Software (TOMS)*, vol. 45, no. 4, Dec. 1, 2019. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3356900, published.

[80] L. Heinrich and M. Kagan, “Differentiable Matrix Elements with MadJax,” in *20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research: AI Decoded - Towards Sustainable, Diverse, Performant and Effective Scientific Computing*, 2022. arXiv: 2203.00057 [hep-ph].

[81] J. Alwall, R. Frederix, S. Frixione, *et al.*, “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations,” *JHEP*, vol. 07, p. 079, 2014. doi: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[82] J. Bradbury, R. Frostig, P. Hawkins, *et al.*, *JAX: Composable transformations of Python+NumPy programs*, http://github.com/google/jax, Version 0.2.5, 2018.

[83] B. Dauvergne and L. Hascoët, "The data-flow equations of checkpointing in reverse automatic differentiation," in *Computational Science – ICCS 2006*, V. N. Alexandrov, G. D. van Albada, P. M. A. Sloot, and J. Dongarra, Eds., red. by D. Hutchison, T. Kanade, J. Kittler, *et al.*, vol. 3994, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 566–573. doi: 10.1007/11758549_78. [Online]. Available: http://link.springer.com/10.1007/11758549_78 (visited on 01/26/2022).

[84] J. Blühdorn, M. Sagebaum, and N. R. Gauger, "Event-based automatic differentiation of OpenMP with OpDiLib," *arXiv:2102.11572 [cs]*, Sep. 27, 2021. arXiv: 2102.11572. [Online]. Available: http://arxiv.org/abs/2102.11572 (visited on 01/28/2022).

[85] M. Aehle *et al.*, *Derivatives in proton ct*, arXiv: 2202.05551, Feb. 14, 2022. [Online]. Available: https://arxiv.org/abs/2202.05551 (visited on 02/14/2022), published.

[86] M. Kasim, D. Watson-Parris, L. Deaconu, *et al.*, "Up to two billion times acceleration of scientific simulations with deep neural architecture search," in *APS Division of Plasma Physics Meeting Abstracts*, vol. 2020, 2020, BO05–001.

[87] S. Collaboration, "Science case and project design," 2023, Internal Benchmark Layouts.

[88] B. L. Dingus and H. Collaboration, "Hawc (high altitude water cherenkov) observatory for surveying the tev sky," *AIP Conference Proceedings*, vol. 921, no. 1, pp. 438–439, Jul. 2007, issn: 0094-243X. doi: 10.1063/1.2757390. eprint: https://pubs.aip.org/aip/acp/article-pdf/921/1/438/11818639/438_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.2757390.

[89] A. e. a. Addazi, "The Large High Altitude Air Shower Observatory (LHAASO) Science Book (2021 Edition)," *Chin. Phys. C*, vol. 46, pp. 035 001–035 007, 2022. arXiv: 1905.02773 [astro-ph.HE].

[90] D. Heck and et al., "Corsika: A monte carlo code to simulate extensive air showers," *FZKA Report*, no. 6019, 1998.

[91] T. Dorigo, M. Aehle, C. Arcaro, *et al.*, "Toward the end-to-end optimization of the swgo array layout," *Nuclear Physics B*, vol. 1017, p. 116 934, Aug. 2025, issn: 0550-3213. doi: 10.1016/j.nuclphysb.2025.116934. [Online]. Available: http://dx.doi.org/10.1016/j.nuclphysb.2025.116934.

[92] J. Neyman and E. Pearson, “Ix. on the problem of the most efficient tests of statistical hypotheses,” *Phil. Trans. R. Soc. Lond. A*, vol. 231, p. 289, 1933. doi: 10.1098/rsta.1933.0009J.

[93] P. Abreu *et al.*, “The southern wide-field gamma-ray observatory (swgo): A next-generation ground-based survey instrument for vhe gamma-ray astronomy,” *arXiv*, 2019. eprint: 1907.07737 (astro-ph.IM).

[94] C. R. Rao, “Information and the accuracy attainable in the estimation of statistical parameters,” *Bull. Calcutta Math. Soc.*, vol. 37, p. 81, 1945.

[95] T. E. S. Group, “Deliberation document on the 2020 Update of the European Strategy for Particle Physics,” CERN, Geneva, Tech. Rep., 2020. doi: 10.17181/ESU2020Deliberation. [Online]. Available: https://cds.cern.ch/record/2720131.

[96] T. I. M. C. Collaboration, “The muon collider - supplementary report to the european strategy for particle physics - 2026 update,” *Submission for the 2026 update of the Europena Strategy of Particle Physics*, 2025.

[97] M. A. Palmer, K. Long, and for the Muon Accelerator Program (MAP), “Muon accelerators for particle physics (muon),” in *Article Collection*, 2016-2021.

[98] Y. I. Alexahin, E. Gianfelice-Wendt, V. V. Kashikhin, N. V. Mokhov, A. V. Zlobin, and V. Y. Alexakhin, “Muon collider interaction region design,” *Phys. Rev. ST Accel. Beams*, vol. 14, p. 061 001, 6 2011. doi: 10.1103/PhysRevSTAB.14.061001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevSTAB.14.061001.

[99] N. V. Mokhov, Y. I. Alexahin, V. V. Kashikhin, S. I. Striganov, and A. V. Zlobin, *Muon collider interaction region and machine-detector interface design*, 2012. arXiv: 1202.3979 [physics.acc-ph]. [Online]. Available: https://arxiv.org/abs/1202.3979.

[100] N. V. Mokhov and S. I. Striganov, *Detector background at muon colliders*, 2012. arXiv: 1204.6721 [physics.ins-det]. [Online]. Available: https://arxiv.org/abs/1204.6721.

[101] D. Calzolari and K. Skoufaris, “Machine-detector interface studies for a multi-TeV muon collider,” in *Proceedings of 41st International Conference on High Energy physics — PoS(ICHEP2022)*, vol. 414, 2022, p. 063. doi: 10.22323/1.414.0063.

[102] e. a. Battistoni G., "Overview of the fluka code," *Annals of Nuclear Energy*, vol. 82, pp. 10–18, 2015. doi: https://doi.org/10.1016/j.anucene.2014.11.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306454914005878.

[103] D. Lucchesi, N. Bartosik, D. Calzolari, L. Castelli, and A. Lechner, "Machine-Detector interface for multi-TeV Muon Collider," *PoS*, vol. EPS-HEP2023, p. 630, 2024. doi: 10.22323/1.449.0630.

[104] K. Skoufaris, C. Carli, and D. Schulte, "10 TeV Center of Mass Energy Muon Collider," in *Proc. IPAC'22*, 2022.

[105] Y. I. Alexahin, E. Gianfelice-Wendt, V. V. Kashikhin, N. V. Mokhov, A. V. Zlobin, and V. Y. Alexakhin, "Muon collider interaction region design," *Phys. Rev. ST Accel. Beams*, vol. 14, p. 061 001, 6 2011. doi: 10.1103/PhysRevSTAB.14.061001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevSTAB.14.061001.

[106] N. Mokhov and S. Striganov, "Detector backgrounds at muon colliders," *Physics Procedia*, vol. 37, pp. 2015–2022, 2012, Proceedings of the 2nd International Conference on Technology and Instrumentation in Particle Physics (TIPP 2011), issn: 1875-3892. doi: https://doi.org/10.1016/j.phpro.2012.03.761. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187538921201927X.

[107] Y. Alexahin, E. Gianfelice-Wendt, and V. Kapin, "Muon collider lattice concepts," *Journal of Instrumentation*, vol. 13, no. 11, P11002, 2018. doi: 10.1088/1748-0221/13/11/P11002. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/13/11/P11002.

[108] C. Bell, D. Calzolari, C. Carli, *et al.*, *Maia: A new detector concept for a 10 tev muon collider*, 2025. arXiv: 2502.00181 [physics.ins-det]. [Online]. Available: https://arxiv.org/abs/2502.00181.

[109] D. Arominski, J.-J. Blaising, E. Brondolin, *et al.*, *A detector for clic: Main parameters and performance*, 2018. arXiv: 1812.07337 [physics.ins-det]. [Online]. Available: https://arxiv.org/abs/1812.07337.

[110] C. Accettura *et al.*, "Towards a muon collider," *Eur. Phys. J. C*, vol. 83, no. 9, p. 864, 2023, [Erratum: Eur.Phys.J.C 84, 36 (2024)]. doi: 10.1140/epjc/s10052-023-11889-x. arXiv: 2303.08533 [physics.acc-ph].

[111] S. C. et al., “Crilin: A crystal calorimeter with longitudinal information for a future muon collider,” *JINST*, vol. 17, no. 09, P09033, 2022. doi: 10.1088/1748-0221/17/09/P09033.

[112] C. Cantone, A. Cemmi, S. Ceravolo, *et al.*, “Developing an alternative calorimeter solution for the future muon collider: The crilin design,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1069, p. 169 973, 2024, issn: 0168-9002. doi: https://doi.org/10.1016/j.nima.2024.169973. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900224008994.

[113] A. Belyaev, R. S. Chivukula, B. Fuks, E. H. Simmons, and X. Wang, “Vectorlike top quark production via an electroweak dipole moment at a muon collider,” *Phys. Rev. D*, vol. 108, p. 035 016, 3 2023. doi: 10.1103/PhysRevD.108.035016. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.108.035016.

[114] M. Selvaggi, “Delphes 3: A modular framework for fast-simulation of generic collider experiments,” *Journal of Physics: Conference Series*, vol. 523, no. 1, p. 012 033, 2014. doi: 10.1088/1742-6596/523/1/012033. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/523/1/012033.

[115] M. Belfkir, T. A. Chowdhury, and S. Nasri, “Doubly-charged scalars of the minimal left-right symmetric model at muon colliders,” *Phys. Lett. B*, vol. 852, p. 138 605, 2024. doi: 10.1016/j.physletb.2024.138605. arXiv: 2307.16111 [hep-ph].

[116] S. Bottaro, M. Cirelli, and F. Sala, “Closing the window on WIMP Dark Matter,” *Eur. Phys. J. C*, vol. 82, no. 1, p. 31, 2022. doi: 10.1140/epjc/s10052-021-09917-9. arXiv: 2107.09688 [hep-ph].

[117] R. Capdevilla, F. Meloni, R. Simoniello, and J. Zurita, “Hunting wino and higgsino dark matter at the muon collider with disappearing tracks,” *JHEP*, vol. 06, p. 133, 2021. doi: 10.1007/JHEP06(2021)133. arXiv: 2102.11292 [hep-ph].

[118] M. Belfkir, A. Jueid, and S. Nasri, “Boosting dark matter searches at muon colliders with machine learning: The mono-Higgs channel as a case study,” *Prog. Theor. Exp. Phys.*, vol. 2023, no. 12, 123B03, 2023. doi: 10.1093/ptep/ptad144. arXiv: 2309.11241 [hep-ph].

[119] R. Capdevilla, F. Meloni, and J. Zurita, "Discovering electroweak interacting dark matter at muon colliders using soft tracks," *Physical Review Letters*, vol. 134, no. 18, p. 181 802, 2025. doi: 10.1103/PhysRevLett.134.181802. arXiv: 2405.08858 [hep-ph].

[120] L. Di Luzio, R. Gröber, and G. Panico, "Probing new electroweak states via precision measurements at the LHC and future colliders," *JHEP*, vol. 01, p. 011, 2019. doi: 10.1007/JHEP01(2019)011. arXiv: 1810.10993 [hep-ph].

[121] R. Franceschini and X. Zhao, "Going all the way in the search for WIMP dark matter at the muon collider through precision measurements," *Eur. Phys. J. C*, vol. 83, no. 6, p. 552, 2023. doi: 10.1140/epjc/s10052-023-11724-3. arXiv: 2212.11900 [hep-ph].

[122] G. Haghighat and M. M. Najafabadi, "Search for lepton-flavor-violating ALPs at a future muon collider and utilization of polarization-induced effects," *Nucl. Phys. B*, vol. 980, p. 115 827, 2022. doi: 10.1016/j.nuclphysb.2022.115827. arXiv: 2106.00505 [hep-ph].

[123] T. Han, Z. Liu, L.-T. Wang, and X. Wang, "WIMPs at High Energy Muon Colliders," *Phys. Rev. D*, vol. 103, no. 7, p. 075 004, 2021. doi: 10.1103/PhysRevD.103.075004. arXiv: 2009.11287 [hep-ph].

[124] S. İnan and A. Kisselev, "Probe of a Randall-Sundrum-like model from muon pair production at high energy muon collider," *arXiv preprint*, 2023. arXiv: 2301.08585 [hep-ph].

[125] K. Korshynska, M. Löschner, M. Marinichenko, K. Mękała, and J. Reuter, "Z' boson mass reach and model discrimination at muon colliders," *Eur. Phys. J. C*, vol. 84, no. 6, p. 568, 2024. doi: 10.1140/epjc/s10052-024-12892-6. arXiv: 2402.18460 [hep-ph].

[126] J. Liu, Z.-L. Han, Y. Jin, and H. Li, "Unraveling the Scotogenic model at muon collider," *JHEP*, vol. 12, p. 057, 2022. doi: 10.1007/JHEP12(2022)057. arXiv: 2207.07382 [hep-ph].

[127] M. Sahin and A. Caliskan, "Excited muon production in muon colliders via contact interaction," *Journal of Physics G: Nuclear and Particle Physics*, vol. 50, no. 2, p. 025 002, 2023. doi: 10.1088/1361-6471/acaffb. arXiv: 2105.01964 [hep-ph].

[128] C. Sen *et al.*, “Displaced Higgs production in Type-III seesaw at the LHC/FCC, MATHUSLA and muon collider,” *Eur. Phys. J. C*, vol. 82, no. 3, p. 230, 2022. doi: 10.1140/epjc/s10052-022-10176-5. arXiv: 2107.12442 [hep-ph].

[129] M. Forslund and P. Meade, “High precision higgs from high energy muon colliders,” *JHEP*, vol. 08, p. 185, 2022. doi: 10.1007/JHEP08(2022)185. arXiv: 2203.09425 [hep-ph].

[130] N. B. et al., “Simulated detector performance at the muon collider,” *Eur. Phys. J. C*, vol. 82, p. 737, 2022. doi: 10.1140/epjc/s10052-022-10665-7.

[131] J. de Blas *et al.*, “Higgs Boson Studies at Future Particle Colliders,” *JHEP*, vol. 01, p. 139, 2020. doi: 10.1007/JHEP01(2020)139. arXiv: 1905.03764 [hep-ph].

[132] A. Martelli, *The cms hgcal detector for hl-lhc upgrade*, 2017. arXiv: 1708.08234 [physics.ins-det]. [Online]. Available: https://arxiv.org/abs/1708.08234.

[133] S. R. Qasim, J. Kieseler, Y. Iiyama, and M. Pierini, “Learning representations of irregular particle-detector geometry with distance-weighted graph networks,” *The European Physical Journal C*, vol. 79, no. 7, p. 608, 2019, issn: 1434-6052. doi: 10.1140/epjc/s10052-019-7113-9. [Online]. Available: https://doi.org/10.1140/epjc/s10052-019-7113-9.

[134] S. R. Q. et al., “End-to-end multi-particle reconstruction in high occupancy imaging calorimeters with graph neural networks,” *The European Physical Journal C*, vol. 82, no. 8, 2022. doi: 10.1140/epjc/s10052-022-10665-7. [Online]. Available: https://doi.org/10.1140%2Fepjc%2Fs10052-022-10665-7.

[135] S. Agostinelli *et al.*, “GEANT4 — a simulation toolkit,” *Nucl. Inst. Meth. A*, vol. 506, p. 250, 2003. doi: 10.1016/S0168-9002(03)01368-8.

[136] J. Allison *et al.*, “Geant4 developments and applications,” *IEEE Transactions on Nuclear Science*, vol. 53, no. 1, pp. 270–278, 2006, issn: 0018-9499. doi: 10.1109/TNS.2006.869826. [Online]. Available: http://ieeexplore.ieee.org/document/1610988/ (visited on 12/08/2021).

[137] J. Allison *et al.*, “Recent developments in Geant4,” en, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 835, pp. 186–225, 2016, issn: 01689002. doi: 10.1016/j.nima.2016.06.125. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168900216306957 (visited on 12/08/2021).

[138] M. Thomson, “Particle flow calorimetry and the pandorapfa algorithm,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 611, no. 1, pp. 25–40, 2009, issn: 0168-9002. doi: 10.1016/j.nima.2009.09.009. [Online]. Available: http://dx.doi.org/10.1016/j.nima.2009.09.009.

[139] A. P. et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[140] M. A. et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[141] F. Nardi, *Genmodules: Differentiable event generation for calorimeter simulation*, https://github.com/FedericoNardi/HGCalML/tree/master/genModules, Accessed: 2025-05-18, 2024.

[142] F. Nardi, J. Donini, L. Chen, N. R. Gauger, T. Dorigo, and X. T. Nguyen, “Differentiable modeling for calorimeter simulation using diffusion models,” in *Book of Abstracts, European AI for Fundamental Physics Conference*, Via dei Giudicati 66, 09131 Cagliari (CA), Italy, 2025, p. 118.

[143] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press, 2003, isbn: 0521642981 9780521642989. [Online]. Available: http://www.inference.phy.cam.ac.uk/mackay/itila/book.html.

[144] M. Aehle, X. T. Nguyen, M. Novák, *et al.*, *Efficient forward-mode algorithmic derivatives of geant4*, 2024. arXiv: 2407.02966 [physics.comp-ph]. [Online]. Available: https://arxiv.org/abs/2407.02966.

[145] T. Dorigo and et al., “On the utility function of experiments in fundamental science,” *Preprint*, 2025, arXiv:Utility2025.