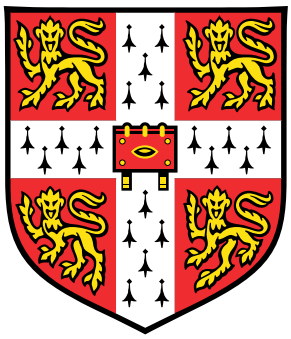
Searches for New Physics at the Large Hadron Collider



Letizia Palmas

Department of Physics

University of Cambridge

This dissertation is submitted for the degree of
MPhil in Data Intensive Science

June 2024
word count: 6946

Contents

List of Figures

List of Tables

Abstract

The analysis of particle physics data demands exceptional precision to accurately describe phenomena occurring at high energies. In this work, we reproduce the ML4EFT framework, which leverages the Standard Model Effective Field Theory (SMEFT) for a systematic and comprehensive examination of New Physics beyond the Standard Model. We conduct statistical inference to determine the values of the Wilson coefficients that parametrise these interactions. Our approach involves employing Deep Learning techniques to interpret the data distribution and perform global fits. We demonstrate that, compared to traditional measurements, the integration of unbinned observables leads to higher precision in the inference of theoretical parameters.

1 Introduction

Over the course of history, humans have made a conscious effort to understand and describe the structure of the Universe. This objective has driven scientific discovery and innovation, leading to the formulation of theories that explain the fundamental constituents of matter and the forces governing their interactions. Among these theories, the Standard Model of particle physics (SM) is considered to be one of the most successful and thoroughly tested theories in the history of science. It provides a comprehensive framework that describes three of the fundamental forces of nature with remarkable precision. Despite its success, the SM is still regarded as an incomplete theory. It fails to account for many observed phenomena, such as the nature of dark matter, the matter-antimatter asymmetry in the universe, and the origin of neutrino masses. Moreover, its design does not include a quantum theory of gravity. To address these limitations, years of research in particle physics have sought to identify and uncover new complementary theories that extend beyond the Standard Model (BSM).

Among the various approaches to exploring BSM physics, the Standard Model Effective Field Theory (SMEFT) has gained significant attention. SMEFT provides a systematic framework to parameterise and study potential new physics effects that manifest at energy scales higher than those probed by current experiments. It provides a versatile tool to interpret experimental results gathered from the Large Hadron Collider (LHC) measurements in a model-independent manner.

In order to explore this framework in the most detailed fashion, it is desirable to study different physical processes and explain them within a comprehensive global SMEFT analysis [1]. Many studies in this context depend on binned distributions derived from experimental data, which inherently cause some loss of information due to the binning process. Effective strategies towards an improved SMEFT study involve an efficient exploration of the kinematic information contained within given measurements to carry out parameter inference and compare the results with theoretical predictions. Defining optimal unbinned observables can increase the sensitivity to EFT coefficients by preventing the information loss arising from a binned analysis or a restricted analysis of a subset of kinematic variables.

The objective of this project is to devise an optimal strategy to infer the presence of New Physics (NP) in a toy dataset by performing inference on the EFT parameters. We will consider the process of top-quark pair production as a result of proton-proton collisions: $pp \to \tau\tau \to b\ell^+\nu_\ell\bar{b}\ell^-\bar{\nu}_\ell$, where $\ell^\pm$ are the positive and negative leptons, ν_ℓ are the lepton neutrinos, and $b\bar{b}$ is the b-jet pair. We will initially develop a classical strategy based on binned observables, with the goal of providing an initial estimate of the Wilson coefficients. We will then implement supervised Machine Learning (ML) classification techniques to train different Neural Networks (NN) and extract the likelihood of the data. This will be done on Monte Carlo (MC) simulations at the parton level, which represents a simplified scenario for which the analytical form of the likelihood is known. This will allow us to test the effectiveness of the ML models. Once validated, the NNs will be applied to the complete datasets for a maximally optimal analysis. The ML parameterisation will be integrated into the description of the likelihood of the data. From this, we will be able to infer the posterior probability distributions of the EFT coefficients using Nested Sampling. We will finally compare

the NN implementation with the traditional approach to assess the potential of the ML strategy.

2 Theoretical Background

2.1 The Standard Model of Particle Physics

Nature is governed by four fundamental interactions, operating over different ranges and having different strengths. The strong force has a range of 10^{-15} m and is responsible for interactions at the nuclear level; the weak interaction, associated with radioactive decay, has a range of 10^{-17} m; the electromagnetic force governs most of the macroscopic physics, with infinite range and strength defined by the fine structure constant $\alpha \approx 1/137$ [2]; the gravitational force is the weakest, with low-energy coupling of $\approx 10^{-38}$, but has infinite range [3]. Each of these forces can be described by its own theory. The classical theory of gravity is represented by Newton's laws of universal gravitation, while its relativistic explanation resides in Einstein's general theory of relativity. The electromagnetic force is described by the theory of electrodynamics, originally formulated by Maxwell and whose quantistic definition was provided by Feynman's *quantum electrodynamics* (QED). The theoretical explanation of the weak force, also called *flavour dynamics*, was formulated by Fermi and expanded by Glashow, Weinberg and Salam. Finally, the strong force, described by the theory of *chromodynamics*, was extensively studied by Yukawa [4].

The SM is a theoretical framework that describes, via quantum gauge field theories, the strong, weak and electromagnetic interactions and the known elementary matter constituents. Matter fields are described by quarks and leptons, classified in pairs into three generations. The lightest particles, which constitute the building blocks of all stable matter in the universe, belong to the first generation. Heavier particles, belonging to the next generations, are less stable and hence decay quickly. The 3 generations of quarks, up u and down d, charm c and strange s, top t and bottom b, are also characterised by three colour charges, red, green and blue, that determine how they interact among each other [5]. The six leptons are similarly arranged in three generations: electron e and the electron neutrino ν_e, muon μ and muon neutrino ν_μ, and tau τ and tau neutrino ν_τ. The electron, muon and tau all have an electric charge and distinct mass, whereas the neutrinos are electrically neutral and have negligible mass. Electrons are the least massive among the charged leptons and are as such the most stable, while the heavier muons and taus undergo rapid particle decay. For this reason, electrons commonly occur in the universe, while muons and taus can only be produced in high-energy collisions [6].

These forces are mediated by the exchange of force-carrier particles, the bosons, through which matter particles can exchange discrete amounts of energy. The strong force is mediated by the gluon g, the electromagnetic force by the photon γ, and the weak force by the intermediate vector bosons W and Z. Although not yet found, the graviton is thought to be the corresponding force-carrying particle of the gravitational force [7].

2.2 Physics beyond the Standard Model

Despite being the most successful theory of particle physics to date, the Standard Model still presents some defects. Research in theoretical physics often focuses on proposing new hypotheses beyond the Standard Model. These theories must modify the SM while remaining consistent with observed data, address its flaws, and suggest new experiments to validate any deviations from it. Although the SM can account for the majority of observed phenomena, some unresolved questions demand a more thorough and complete theory:

- **Gravity:** The SM describes three of the four fundamental interactions, failing to combine the microscopic scale of quantum theory with the macroscopic theory of general relativity. The effect of gravity at the particle level is so weak that it can be neglected over such short distances, but becomes dominant in large-scale contexts.

- **Dark matter:** The SM does not account for dark matter, which constitutes approximately 26.8% of the Universe [8]. The existence of dark matter can be inferred from gravitational effects on visible matter, the radiation coming from the cosmic microwave background and galactic rotation curves. The SM successfully describes visible matter but fails to include dark matter particles.

- **Dark energy:** In a similar fashion, the SM cannot explain the presence of dark energy, which constitutes about 68.3% of the Universe and is known to be responsible for its accelerated expansion [9]. The presence and impact of dark energy remains an open question in Cosmology and is not addressed by the theory of the SM.

- **Neutrino masses:** The SM assumes neutrinos are massless fermions that move at the velocity of light. However, numerous experiments have provided compelling evidence that there are three discrete neutrino masses. These values not only represent a fundamental probe of the SM, but are also relevant in Astrophysics and Cosmology. Although neutrinos are very light, they may significantly contribute to the mass density of the Universe. Understanding their masses could explain the role of neutrinos in the evolution of the Universe [10].

- **CP violation:** The strong CP problem arises from the non-observation of charge-parity (CP) violation in the strong interaction. Theoretical predictions suggest that the strong force should exhibit CP violation, but this has not been observed yet experimentally. One of the theories proposed to explain this problem involves the presence of the axion, a hypothetical particle which is not included in the Standard Model [11].

- **Matter-antimatter asymmetry:** The Universe is characterised by a baryon asymmetry, meaning that it contains significantly more matter than antimatter. To explain this phenomenon, the theory should take into account C and CP violation, but the hypotheses coming from the SM are insufficient to fully account for this asymmetry.

The current understanding of the SM suggests that this theory could be a low-energy approximation of a more fundamental concept, but the lack of experimental evidence has impeded any further progress towards the description of a more complete framework. The necessity to discover new BSM physics and to further validate the SM led to a combination of direct and indirect search strategies, which leverage experimental and theoretical tools to probe the fundamental laws of nature.

Direct searches for NP aim at the direct detection of new particles or interactions at high-energy particle colliders and other experimental setups. The largest and most powerful particle accelerator at present is the Large Hadron Collider (LHC), which is capable of colliding protons at energies up to 13 TeV [12]. The production and subsequent decay of new particles can produce distinct and identifiable signatures that can provide immediate evidence of NP. Direct detection also provides a straightforward way to test specific models by comparing theoretical assumptions with observed data. However, the reach of direct searches is limited by the energy of the colliders, since it is physically impossible to observe new physical scenarios at energy scales beyond the collider's capability.

On the other hand, indirect searches focus on precise measurements of SM processes and the detection of deviations from SM predictions, in particular by investigating the tail of the distributions. These strategies are particularly effective when NP occurs at energy scales beyond the reach of current colliders.

2.3 Standard Model Effective Field Theory

Theoretical research plays a critical role in guiding experimental searches and interpreting their results. The development of models and simulations helps predict possible BSM phenomena and design experiments to test these predictions. For instance, the Standard Model Effective Field Theory (SMEFT) extends the SM by including higher-dimensional operators that parametrise the effects of NP at unknown higher energy scales.

SMEFT modifies the Standard Model to account for New Physics while preserving its exact symmetries and field content [13]. In this framework, it is assumed that the effect of new massive particles with mass scale $M \simeq \Lambda$ can be parametrised at lower energies, $E \ll \Lambda$, in a model-independent manner and in terms of higher-dimensional operators [14]. In this context, Λ is the energy scale of NP, typically assumed to be much larger than the electroweak scale. The effective SM Lagrangian can therefore be expanded as

$$\mathcal{L}_{\text{smeft}} = \mathcal{L}_{\text{sm}} + \sum_i \frac{c_i}{\Lambda^2}\mathcal{O}_i^{(6)} + \sum_i \frac{c_i}{\Lambda^4}\mathcal{O}_i^{(8)} + \cdots \tag{1}$$

where $\mathcal{O}_i^{(d)}$ are the d-dimensional higher operators, and c_i are the Wilson coefficients that parametrise the strength of each operator. The leading-order effects of NP are usually captured by dimension-6 operators, which modify the interactions and properties of the SM. These can be constructed from the SM fields and satisfy the $\mathrm{SU_C}(3) \times \mathrm{SU_L}(2) \times \mathrm{U_Y}(1)$ gauge symmetry [14]. This group dictates how the fundamental particles interact through the strong, weak, and electromagnetic forces. $\mathrm{SU_C}(3)$ represents the strong interaction and its mediators, the gluons. $\mathrm{SU_L}(2)$ describes the symmetry group of the weak interaction, with 3 gauge bosons: W^+, W^-, Z.

Finally, $U_Y(1)$ represents the symmetry of the electromagnetic force, associated with the hypercharge [15].

3 Traditional analysis

For a dataset $\mathcal{D}$, the theory $\mathcal{T}$ describing the distribution of the data depends on n parameters of the model, say $\mathbf{c} = \{c_1, c_2, ..., c_n\}$. Bayesian statistics describes the degree of belief in a given theory or hypothesis. To this extent, we define the likelihood function as the probability of observing the measured data given some value of the parameters as $\mathcal{L}(\mathbf{c}) = P(\mathcal{D}|\mathcal{T}(\mathbf{c}))$ [16]. In this section, we will conduct a Bayesian analysis of the data to perform parametric inference following the traditional binned approach on a subset of features.

3.1 Dataset overview

The observed dataset contains 10,560 measured events characterised by 14 features, which represent the kinematic properties of each state. These properties are:

- **Transverse momentum:** The momentum component perpendicular to the beam line [4]. This refers to the positive and negative leptons $p_T^{\ell\pm}$, the lepton pair $p_T^{\ell\ell}$, and the b-jet pair p_T^{bb};
- **Pseudorapidity:** Rapidity, y, typically indicates the angle with respect to the axis of the colliding beams. This quantity is well-defined for particles with velocity close to the speed of light, and it is equal to 0 for particle trajectories that are perpendicular to the beam. Pseudorapidity is an approximation of y that can be more easily calculated from the Cartesian angle between the particle direction above or below the beam line [17]. The dataset contains pseudorapidity measures of the positive and negative leptons $\eta_{\ell\pm}$, as well as the separation $\Delta\eta_\ell$;
- **Invariant mass:** This refers to the mass of a particle before its decay and can be calculated from the energies and momenta of the decay products. This inferred value is independent of the reference frame in which the energies and momenta are measured, which is why it is defined as *invariant*. The data presents calculations of the invariant masses of the lepton pair $m_{\ell\ell}$ and of the b-jet pair m_{bb};
- **Azimuthal angle separation:** This describes the angular separation of the two leptons $\Delta\phi_{\ell\ell}$ along the axis of the beam;
- **Missing transverse energy (MET):** This refers to the energy that is not identified by the detector but is expected to exist due to conservation laws. MET is commonly used to infer the presence of non-detectable particles.

3.2 Binned Likelihoods

In addition to the observed data, we consider 5 datasets of weighted MC samples, one for the SM predictions, and 4 related to the EFT Wilson coefficients: two rep-

resent the linear form of $c_{\mathrm{dt}}^{(8)}$ and $c_{\mathrm{qt}}^{(8)}$, two describe their quadratic versions. The MC samples present the same features of the observed data and an additional list of weights corresponding to the contribution to the total cross-section of each event. The events i are combined into N_b bins. It is important to subdivide each feature into the correct number of bins to avoid as much loss of information as possible. For each observable, we devise an appropriate binning and investigate the distribution of the data compared to the estimated values. We choose to study the distribution of transverse momentum of the positive leptons, $p_T^{\ell+}$. The effective number of events for the MC samples is calculated from the value of the cross-section and the luminosity L of the detector: $N_{\mathrm{ev}} = \sigma L$. We apply the same binning to both the observed and MC-generated datasets and calculate the measurement errors σ_i by assuming that the events follow a Poisson distribution: $\sigma_i = \sqrt{X_i}$, where X_i corresponds to the observed values. On the other hand, we define the theoretical values as T_i. The binned distribution of both the MC and observed data for $p_T^{\ell+}$ is illustrated in Figure 1.

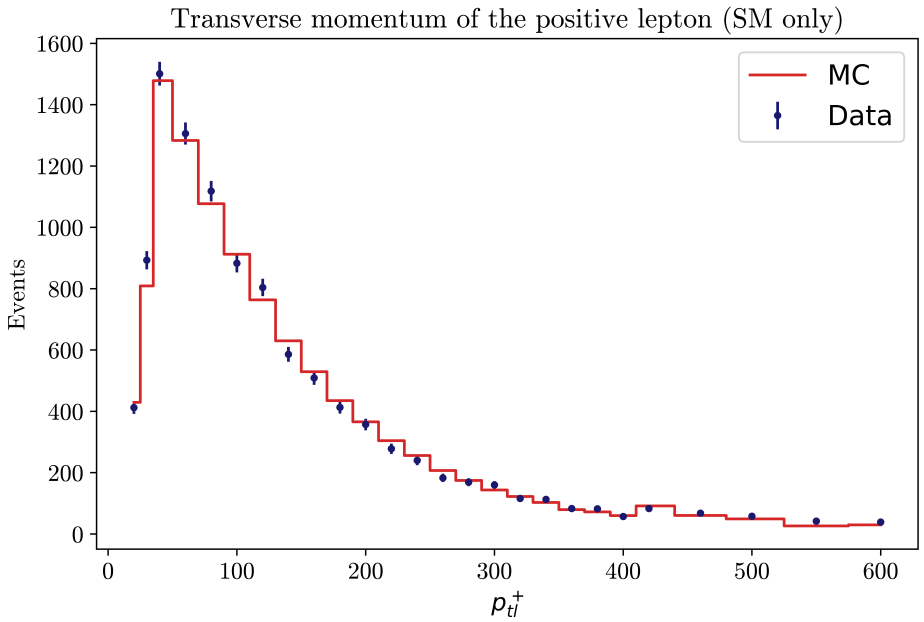


Figure 1: Distribution of the transverse momentum of the positive lepton, $p_T^{\ell+}$ for the binned Monte Carlo samples and the observed data.

By introducing the MC samples of the EFT parametrisation, we can combine and fit the datasets to infer the values of the Wilson coefficients $c_{\mathrm{dt}}^{(8)}, c_{\mathrm{qt}}^{(8)}$ and their squared terms. Our model can be defined by the following equation:

$$E_i = T_{\mathrm{sm}} + c_{\mathrm{dt}}^{8} T_{\mathrm{c8dt}} + c_{\mathrm{qt}}^{8} T_{\mathrm{c8qt}} + (c_{\mathrm{dt}}^{8})^2 T_{\mathrm{c8dt2}} + (c_{\mathrm{qt}}^{8})^2 T_{\mathrm{c8qt2}} \tag{2}$$

where E_i reflects our expectations and is estimated by considering both the SM (T_{sm}) and the EFT theories. In this case, the variables correspond to the transverse momentum, $p_T^{\ell+}$, but the formula applies equally to all observables, provided a pertinent binning. To proceed with the Wilson coefficients inference, we will implement the Nested Sampling (NS) algorithm to identify the points that maximise the likelihood of the data.

3.3 Nested Sampling for Parametric Inference

When we compare our model to the data, our interest lies in constraining its parameters and eventually performing hypothesis testing by comparing our theory to other assumptions. Different techniques have been devised to explore model parameter spaces. The `UltraNest` package implements the Nested Sampling algorithm, which allows us to perform Bayesian inference on an arbitrarily defined likelihood and construct the posterior probability distribution to constrain the parameters describing the data [18]. The evidence can be expressed as an integral in parameter space Θ of the form

$$Z = \int_{\Theta} \mathcal{L}(x|\theta)\pi(\theta)d\theta \tag{3}$$

where $\mathcal{L}(x|\theta)$ is the likelihood of observing data x given parameters θ, and $\pi(\theta)$ represents the prior distribution. Analytically computing the evidence is generally only possible when the posterior is a simple and well-known probability distribution. This scenario only arises in low-dimensional problems, when one can find a conjugate prior to the likelihood [19]. Many methods to calculate the evidence have been devised over time (such as thermodynamic integration [20]), but the computational costs involved in obtaining accurate evidence estimates, in addition to the need for extensive fine-tuning of the algorithm, have prevented them from becoming widely used.

Nested Sampling was introduced as a numerical approximation method for computing the evidence integral [21]. One defines the maximum likelihood in parameter space as $\mathcal{L}_{\max}$, where

$$\mathcal{L}_{\max} = \max_{\theta \in \Theta} \mathcal{L}(x|\theta) \tag{4}$$

Let $\xi(L)$ be the prior probability mass associated with likelihoods greater than a given value L:

$$\xi(L) = \int_{\{\theta:\mathcal{L}(x|\theta)>L\}} \pi(\theta)d\theta \tag{5}$$

$$\equiv \int_{\mathcal{L}>L} \pi(\theta)d\theta \tag{6}$$

By noting that $\pi(\theta)$ is a probability density, it follows that $\xi(L)$ is a decreasing function such that

$$\xi(0) = 1 \tag{7}$$

$$\xi(\mathcal{L}_{\max}) = 0 \tag{8}$$

If θ is continuous, the function $\xi(L)$ is strictly decreasing and can be inverted to give $L(\xi)$ [22]. If we denote the quantity $d\xi$ as the prior mass associated with likelihoods in the range $(L, L + dL)$, we realise that $d\xi$ contributes an amount $Ld\xi$ to the total evidence Z. Hence, the sum of all contributions gives

$$Z = \int_0^1 d\xi L(\xi) \tag{9}$$

For a decreasing sequence of points in the prior volume, $0 < \xi_M < \xi_{M-1} < ... < \xi_1 < 1$, and their associated likelihoods $L_\mu = L(\xi_\mu)$, the 1D solution to Equation 9 can be

found using the trapezium rule:

$$Z \approx \frac{1}{2}\sum_{\mu=1}^{M}(L_{\mu-1}+L_{\mu})\Delta\xi_{\mu} \tag{10}$$

where $\Delta\xi_{\mu} = \xi_{\mu} - \xi_{\mu-1}$, and where we have defined $\xi_0 = 1$ with associated likelihood $L_0 = 0$. If we define the new term $w_{\mu} = (\xi_{\mu-1} - \xi_{\mu+1})/2$, we can rewrite Equation 10 as

$$Z \approx \sum_{\mu=1}^{M} w_{\mu}L_{\mu} \tag{11}$$

with $\xi_{M+1} = \xi_M$.

We initialise the parameter space by drawing a number N_{live} of independent *live points* from the prior: $\theta_j \sim \pi$, where $j = 0, 1, ..., N_{\text{live}} - 1$. The algorithm then proceeds iteratively. At each iteration, the live point with the lowest value of the likelihood is replaced by a new live point with a higher likelihood, again drawn from the prior but subject to the constraint that $\mathcal{L}(x|\theta)$ is larger than the likelihood of the point that was previously discarded. With this strategy, the collection of live points should cluster around the maximum likelihood value, $\mathcal{L}_{\text{max}}$, over time. This process results in an increasing sequence of likelihood values $L_i < L_{i+1}$ from the deceased points. The algorithm continues until a stopping condition is satisfied, for example when there are minimal changes in the approximation of the evidence $\hat{Z}$, or until the algorithm reaches the maximum value of $L(\theta)$, when it is known [23].

We implement the `ReactiveNestedSampler` algorithm from `UltraNest` by specifying appropriate prior and likelihood functions. The likelihood is built around our expected results E_i, defined by Equation 2 and dependent on the Wilson coefficients, and the observed data X_i. For a sufficiently large number of samples, we can model the distribution as a Gaussian and define the distribution of data $\mathcal{D}$ given our theoretical prediction $\mathcal{T}(\mathbf{c})$ as

$$\mathcal{L}(\mathcal{D}|\mathcal{T}(\mathbf{c})) = \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\frac{[X_i - E_i(\mathbf{c})]^2}{\sigma_i^2}\right] \tag{12}$$

where $\sigma_i = \sqrt{X_i}$ are the measurement-associated errors. We can therefore notice that the log-likelihood function takes the form of a χ^2 :

$$-2\log\mathcal{L} = \sum_{i=1}^{N}\frac{[X_i - E_i(\mathbf{c})]^2}{\sigma_i^2} \equiv \chi^2 \tag{13}$$

We choose the prior to be uniform to reflect our ignorance of the true distribution of the Wilson coefficients, but restrict the range of possible values between -2 and 2. Once these two functions have been defined, we explore the space of the $c_{\text{dt}}^{(8)}$ and $c_{\text{qt}}^{(8)}$ parameters to find the values that maximise the likelihood in the measured data. We initialise the algorithm with 1000 live points and display the results of the analysis in a corner plot, shown in Figure 3. The diagonal elements of the corner plot display the marginal posterior distributions of the individual Wilson coefficients through one-dimensional histograms. The 2D contour plot, off the diagonal, presents

the joint distribution between both parameters. The contour lines indicate regions of increasing probability density.

Based on the results obtained from NS, we proceed with our investigation and assess the effectiveness of our analysis. We evaluate the log-likelihood of the SM, for which $c_{\rm dt}^{(8)} = c_{\rm qt}^{(8)} = 0$, and of our model, whose coefficients are obtained from `UltraNest`, and compute the χ^2. This comparison allows us to measure the tension between the two models and determine the extent to which the SMEFT parameters provide a better explanation of the data than the SM. We calculate the $\Delta\chi^2 = \chi^2_{\rm sm} - \chi^2_{\rm best} = 6.183$. Given that a lower χ^2 is indicative of a better fit to the data, this positive difference indicates a preference for the best-fit (EFT) model. The critical χ^2 value at the 95% confidence level, for 2 degrees of freedom, is 5.991 [24]. The $\Delta\chi^2$ associated with our analysis surpasses this value, indicating statistical significance at this confidence level. Based on this, we derive the p-value and conclude that our observations define a tension from the SM at 2σ standard deviations.

The maximum likelihood points determined by `UltraNest` yield values of $c_{\rm dt}^{(8)} = -0.669$ and $c_{\rm qt}^{(8)} = -0.856$, with 95% confidence intervals of $[-1.256, 0.324]$ for $c_{\rm dt}^{(8)}$, and $[-0.952, 0.021]$ for $c_{\rm qt}^{(8)}$. These intervals, derived from the posterior distribution, represent the range of values within which the true parameters lie with 95% probability, given the observed data and the prior information. Figure 2 further illustrates the 2D distributions of the parameters at the 68%, 95% and 99% confidence levels. Each region in the figure indicates where the true parameter values are likely to lie within the corresponding probability. These regions are not centred around (0,0), as would be expected if the SM could fully account for the observed events. Instead, the observed deviations from the SM, as reflected in the statistical calculations, suggest that the SMEFT model provides a more accurate description of the observed dataset.

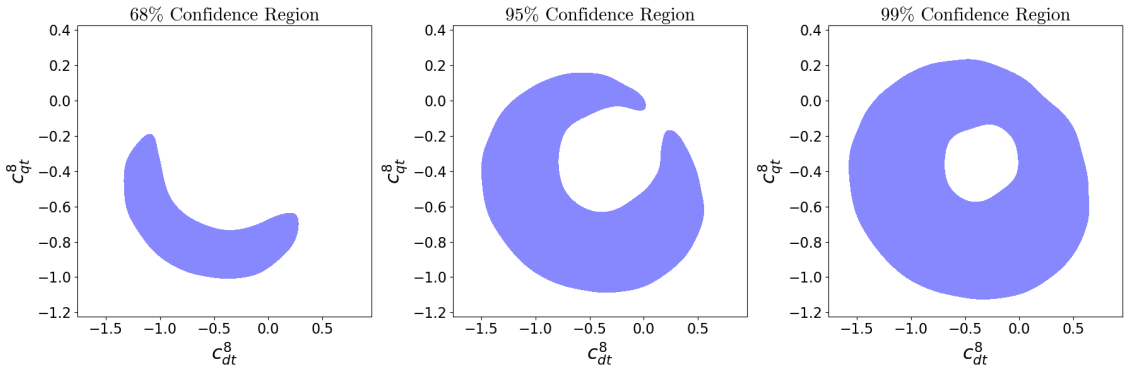


Figure 2: Confidence regions resulting from the binned analysis.

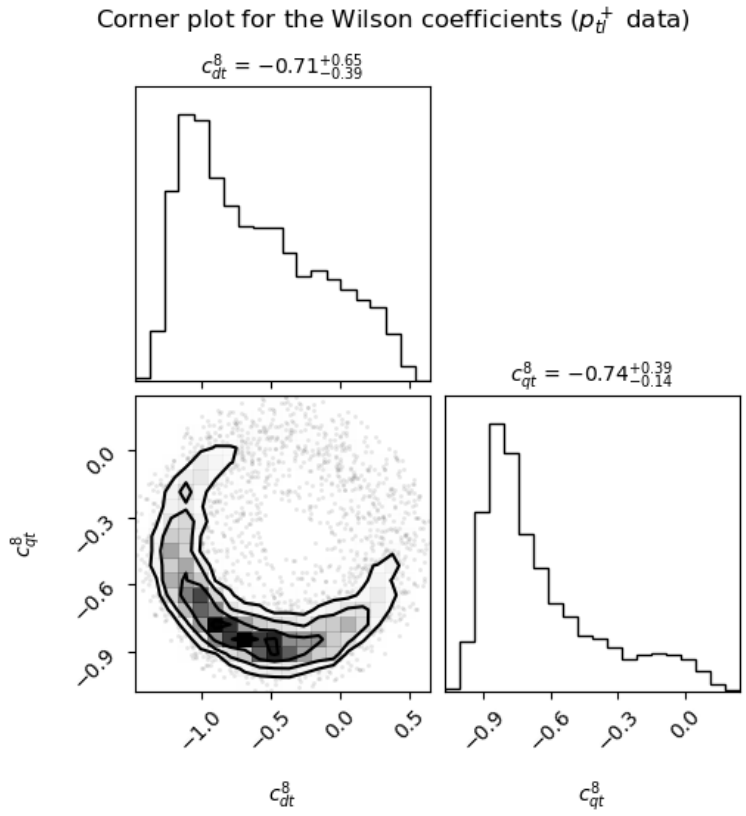


Figure 3: Corner plot obtained by applying NS to the binned measurements of the transverse momentum of the positive lepton.

4 Machine Learning Framework

4.1 Advantages of an unbinned analysis

One significant limitation of a binned likelihood analysis is the inevitable loss of information due to the averaging of data points falling within the same bins. To mitigate this issue, one can construct unbinned likelihoods, where each event in the dataset is examined individually. In this case, $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, ..., x_{i,n}\}$ denotes the array of values of the n final-state variables [1]. While we still define the theoretical framework as $\mathcal{T}(\mathbf{c})$, for $\mathbf{c}$ model parameters, we construct the likelihood from the probability density function $f_\sigma(\mathbf{x}, \mathbf{c})$, assuming the random variables are independent and identically distributed:

$$\mathcal{L}(\mathbf{c}) = \prod_{i=1}^{N_{\text{ev}}} f_\sigma(\mathbf{x}_i, \mathbf{c}) \tag{14}$$

where N_{ev} is the total number of events [1]. The distribution $f_\sigma(\mathbf{x}, \mathbf{c})$ is given by the normalised differential cross-section

$$f_\sigma(\mathbf{x}, \mathbf{c}) = \frac{1}{\sigma_{\text{tot}}} \frac{d\sigma(\mathbf{x}, \mathbf{c})}{d\mathbf{x}} \tag{15}$$

Where σ_{tot} is the total cross-section of the kinematic variables, dependent on the Wilson coefficients and derived from the contributions of the SM and the EFT parameters. More specifically, $\sigma_{\text{tot}} = \sigma_{\text{sm}} + \sum_i c_i \sigma_{\text{eft}}^{(i)}$, where the first term corresponds to

the cross-section of the SM observables, and the second term is the total cross-section of the EFT parameters, with corresponding coefficients c_i.

In the SMEFT design, the differential cross-section exhibits a quadratic dependence on the Wilson coefficients, meaning that the probability density function $f_\sigma(\mathbf{x}, \mathbf{c})$ can be expanded as

$$f_{\text{sm}}(\mathbf{x}, \mathbf{0}) + \sum_{j=1}^{n} f_\sigma^{(j)}(\mathbf{x}) c_j + \sum_{j=1}^{n} \sum_{k \geq j}^{n} f_\sigma^{(j,k)}(\mathbf{x}) c_j c_k \tag{16}$$

Where $f_{\text{SM}}(\mathbf{x}, \mathbf{0})$ denotes the theoretical framework of the SM only, for which $\mathbf{c} = 0$, $f_\sigma^{(j)}$ represents the linear corrections and $f_\sigma^{(j,k)}$ the quadratic corrections associated with the EFT [1]. In our case, we can express the cross-section parametrisation as

$$\frac{d\sigma}{d\mathbf{x}}(\mathbf{x}, \mathbf{c}) = \frac{d\sigma_{\text{sm}}}{d\mathbf{x}}(\mathbf{x}) + c_1 \frac{d\sigma_1}{d\mathbf{x}}(\mathbf{x}, \mathbf{c}) + c_2 \frac{d\sigma_2}{d\mathbf{x}}(\mathbf{x}, \mathbf{c}) + c_1^2 \frac{d\sigma_{11}}{d\mathbf{x}}(\mathbf{x}, \mathbf{c}) + c_2^2 \frac{d\sigma_{22}}{d\mathbf{x}}(\mathbf{x}, \mathbf{c}) \tag{17}$$

In our case, $c_1 = c_{\text{dt}}^{(8)}, c_2 = c_{\text{qt}}^{(8)}$ and each cross-section can be independently derived from the different datasets of MC-generated samples for both the SM and the EFT coefficients. However, evaluating the unbinned likelihood is computationally intensive due to its reliance on numerical simulations [1]. To address this, deep learning methods can be employed to parametrise this probability density, therefore enabling a coherent SMEFT analysis. In this regard, we focus on the design of Neural Networks that could reproduce the cross-section ratios in a supervised ML setting. For this purpose, we introduce the term $r_\sigma(\mathbf{x}, \mathbf{c})$:

$$r_\sigma(\mathbf{x}, \mathbf{c}) = \frac{f_\sigma(\mathbf{x}, \mathbf{c})}{f_\sigma(\mathbf{x}, \mathbf{0})} = 1 + \sum_{j=1}^{n} r_\sigma^{(j)}(\mathbf{x}) c_j + \sum_{j=1}^{n} \sum_{k \geq j}^{n} r_\sigma^{(j,k)}(\mathbf{x}) c_j c_k \tag{18}$$

Where the right-hand side denotes a reformulated version of Equation 16, and

$$r_\sigma^{(j)}(\mathbf{x}) = \frac{f_\sigma^{(j)}(\mathbf{x})}{f_\sigma(\mathbf{x}, \mathbf{0})} \tag{19}$$

$$r_\sigma^{(j,k)}(\mathbf{x}) = \frac{f_\sigma^{(j,k)}(\mathbf{x})}{f_\sigma(\mathbf{x}, \mathbf{0})} \tag{20}$$

correspond to the ratios between the EFT and SM cross-sections.

4.2 Unbinned analysis at the parton level

Feed-forward deep NN can be implemented to perform statistical inference and to investigate the deviation of the EFT hypothesis from the SM. We focus our analysis on 5 datasets of MC-generated samples at the parton level, depicting a simplified scenario that we can describe analytically. We consider pairs of datasets with the same number of events, in which one always describes the SM theory $\mathcal{T}(\mathbf{0})$, and the other represents one of the coefficients under the EFT assumption $\mathcal{T}(\mathbf{c})$, where $\mathbf{c} = \{c_{\text{dt}}^{(8)}, c_{\text{qt}}^{(8)}, (c_{\text{dt}}^{(8)})^2, (c_{\text{qt}}^{(8)})^2\}$. We denote these datasets $\mathcal{D}_{\text{sm}}$ and $\mathcal{D}_{\text{eft}}(\mathbf{c})$, respectively. They contain 200,000 samples with 3 features: the invariant mass of the top-quark pair, $m_{t\bar{t}}$, the rapidity of the top-quark pair, $y_{t\bar{t}}$, and the fixed weights corresponding

to one or the other theoretical framework. We label the data by assigning value 1 to the SM data, and 0 to the EFT. For each pair of datasets, we train a binary classifier whose aim is to delineate a decision boundary $g(\mathbf{x}, \mathbf{c})$ that should separate an event $\mathbf{x}$ as either coming from $\mathcal{D}_{\text{sm}}$ or $\mathcal{D}_{\text{eft}}$. The NNs are optimised through the Binary Cross-Entropy (BCE) loss function [25]:

$$L[g(\mathbf{x}, \mathbf{c})] = -\int d\mathbf{x} \frac{d\sigma(\mathbf{x}, \mathbf{c})}{d\mathbf{x}} \log\left(1 - g(\mathbf{x}, \mathbf{c})\right) - \int d\mathbf{x} \frac{d\sigma(\mathbf{x}, \mathbf{0})}{d\mathbf{x}} \log g(\mathbf{x}, \mathbf{c}) \tag{21}$$

For a large enough number of samples, as $N \to \infty$, we can differentiate Equation 21 with respect to the decision boundary $g(\mathbf{x}, \mathbf{c})$ and minimise it to derive an equation for g, namely:

$$\frac{\partial L}{\partial g} = 0 \to g(\mathbf{x}, \mathbf{c}) = \left(1 + \frac{1}{\sigma_{\text{eft}}} \frac{d\sigma_{\text{eft}}}{d\mathbf{x}} \Big/ \frac{1}{\sigma_{\text{sm}}} \frac{d\sigma_{\text{sm}}}{d\mathbf{x}}\right)^{-1} = \frac{1}{1 + r_\sigma(\mathbf{x}, \mathbf{c})} \tag{22}$$

Equation 22 hence demonstrates that it is possible to derive the cross-section ratios $r_\sigma(\mathbf{x}, \mathbf{c})$ from the NN outputs [1].

We design a Multi-Layer Perceptron (MLP) and train it for each combination of datasets. The architecture remains unchanged, but we modify the hyperparameters to ensure maximum accuracy between the analytical solution and the NN output. Our model takes 2 features as inputs, $m_{t\bar{t}}$ and $y_{t\bar{t}}$, and produces a one-dimensional output corresponding to the decision boundary $g(\mathbf{x}, \mathbf{c})$. The weights are excluded from training as their values carry too much information and would therefore prevent the NN from learning the true underlying distribution of the parameters, leading to poor generalisation on unseen data. The MLP consists of 6 linear layers with a variable number of nodes. The forward pass is constructed as follows:

$$\begin{aligned}
\mathbf{h}_1 &= \mathbf{a}[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}] \\
\mathbf{h}_2 &= \mathbf{a}[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1] \\
\mathbf{h}_3 &= \mathbf{a}[\boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2] \\
\mathbf{h}_4 &= \mathbf{a}[\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3] \\
\mathbf{h}_5 &= \mathbf{a}[\boldsymbol{\beta}_4 + \boldsymbol{\Omega}_4 \mathbf{h}_4] \\
y &= \mathbf{Sigmoid}[\mathbf{h}_5]
\end{aligned}$$

Where $\mathbf{h}_n$ corresponds to each layer of the NN, $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ represent the learnable parameters, and y is the output of the NN. Each layer of the MLP has 50, 70, 150, 100, and 50 nodes respectively. The output of the first 5 layers is passed through a Rectified Linear Unit (ReLU) activation function [26], which is defined as

$$\mathbf{a}[x] = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{23}$$

The implementation of a non-linear activation function allows for complex relationships within the data to be learned by the network, while still avoiding the problem of vanishing gradients that often arises in deep neural networks. To ensure that the output of the NN aligns with our desired boundary and falls within the range $[0, 1]$, we apply the logistic sigmoid function [25] to the final layer:

$$\mathbf{Sigmoid}[x] = \frac{1}{1 + e^{-x}} \tag{24}$$

We split our combined dataset into training and test sets, allocating 20% of the data for testing. By exploring the distribution of the features, we notice a significant difference in the range of values taken by the momentum and rapidity. For this reason, we scale the data by implementing a `RobustScaler` from `sklearn`. This subtracts the median from the data and scales it according to the Interquartile Range (IQR), which is defined as the difference between the first quartile (25^{th} quantile) and the third quartile (75^{th} quantile):

$$x_{\text{new}} = \frac{x - x_{\text{median}}}{\text{IQR}} \tag{25}$$

Common standardisation techniques subtract the mean and scale the data to unit variance [27]. However, this method is less robust to outliers, whereas using the median and IQR often yields better results. Once our dataset has been pre-processed and split, we instantiate our NN and define our loss function according to Equation 21. We use AdamW as our optimisation algorithm, which is a variant of Adam that modifies the typical implementation of weight decay by decoupling it from the optimisation steps taken with respect to the loss function. L_2 regularisation in Adam is usually implemented as follows

$$\mathbf{g}_t = \nabla f(\boldsymbol{\theta}_{t-1}) + w_t \boldsymbol{\theta}_{t-1} \tag{26}$$

where w_t corresponds to the weight decay at time t. It has been shown that L_2 regularisation is not effective for adaptive gradient methods, such as Adam, as it is for Stochastic Gradient Descent [28]. When Adam updates its parameters according to a loss function f plus L_2 regularization, weights that tend to have large gradients in f do not get regularised as much as they would with decoupled weight decay, since the gradient of the regulariser gets scaled along with the gradient of f. This therefore implies that L_2 regularisation and decoupled weight decay are not the same in adaptive gradient algorithms. AdamW adjusts the weight decay term to appear in the gradient update, as shown in line 12 of Algorithm 1:

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta_t \left(\frac{1}{\sqrt{\hat{\boldsymbol{\nu}}_t} + \epsilon} \cdot \alpha \hat{\mathbf{m}}_t + \lambda \boldsymbol{\theta}_{t-1} \right) \tag{27}$$

where α is the learning rate, λ is the regulariser penalty, β_1 and β_2 are factors of the momentum vectors $\mathbf{m}$ and $\boldsymbol{\nu}$, and η_t is a scheduled learning rate multiplier. The learning rate multiplier is introduced to improve performance. In adaptive gradient algorithms such as Adam, the learning rate gets updated for each parameter. However, to account for the scheduling of both α and λ, the global scheduler η_t is introduced through a user-defined procedure `SetScheduleMultiplier(t)` [28].

We implement AdamW and keep the following parameters fixed:

- $\beta_1 = 0.9$; $\beta_2 = 0.999$
- $\epsilon = 10^{-8}$
- $\lambda = 0.01$

For each NN, we tune the learning rate of the optimiser to achieve optimal outcomes. We also modify the batch size of the train and test datasets accordingly.

Algorithm 1 Adam with decoupled weight decay - AdamW

Given: $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \lambda \in \mathbb{R}$
Initialise: time step $t \leftarrow 0$, parameter vector $\boldsymbol{\theta}_{t=0} \in \mathbb{R}$, first moment vector $\mathbf{m}_{t=0} \leftarrow 0$, second moment vector $\boldsymbol{\nu}_{t=0} \leftarrow 0$, schedule multiplier $\eta_{t=0} \in \mathbb{R}$
repeat
    $t \leftarrow t + 1$
    $\nabla f_t(\boldsymbol{\theta}_{t-1}) \leftarrow \text{SelectBatch}(\boldsymbol{\theta}_{t-1})$ ▷ Select batch and return gradient
    $\mathbf{g}_t \leftarrow \nabla f_t(\boldsymbol{\theta}_{t-1}) + \lambda \boldsymbol{\theta}_{t-1}$ ▷ $\lambda\boldsymbol{\theta}_{t-1}$ term only implemented in Adam
    $\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t$ ▷ Element-wise operations
    $\boldsymbol{\nu}_t \leftarrow \beta_1 \boldsymbol{\nu}_{t-1} + (1 - \beta_2)\mathbf{g}_t^2$
    $\hat{\mathbf{m}}_t \leftarrow \mathbf{m}_t / (1 - \beta_1^t)$
    $\hat{\boldsymbol{\nu}}_t \leftarrow \boldsymbol{\nu}_t / (1 - \beta_2^t)$
    $\eta_t \leftarrow \text{SetScheduleMultiplier}(t)$ ▷ ScheduleMultiplier can be fixed or decay
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta_t \left(\alpha \hat{\mathbf{m}}_t / (\sqrt{\hat{\boldsymbol{\nu}}_t} + \epsilon) + \lambda \boldsymbol{\theta}_{t-1} \right)$
until stopping criterion is met
Return optimised parameters $\boldsymbol{\theta}_t$

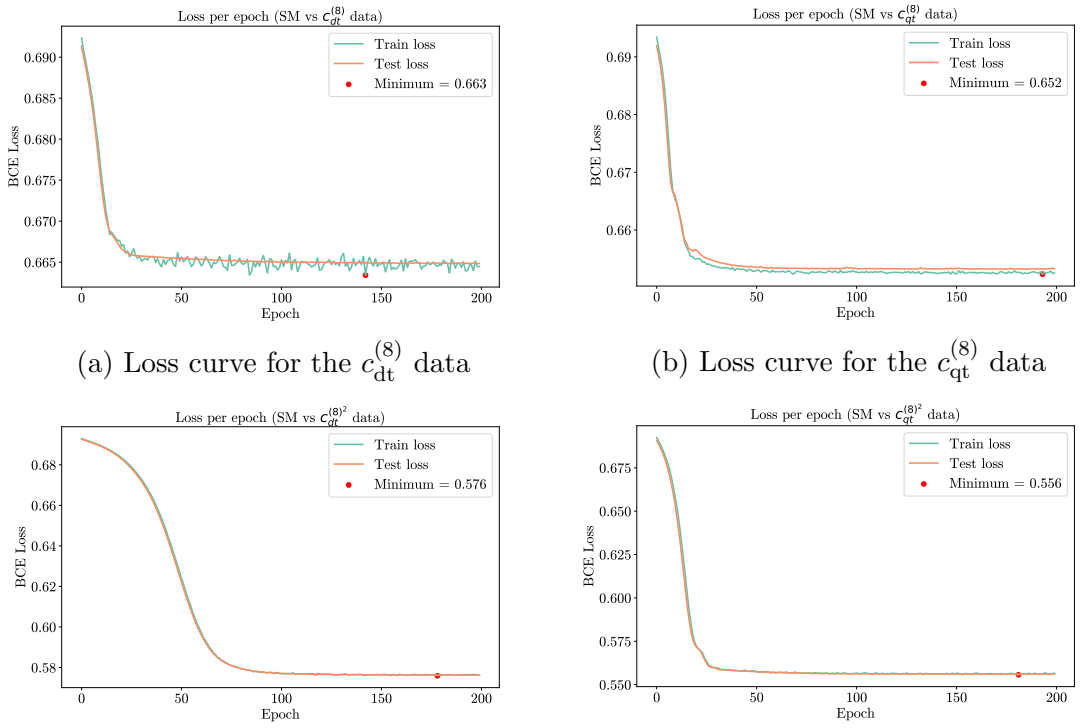
Table 1 illustrates the batch size and learning rates chosen for each dataset. We perform mini-batch gradient descent to allow for a more robust convergence; this method also contributes to avoiding local minima in support of a more complete exploration of the parameter space [29]. We favoured larger batch sizes since the implementation of smaller batches led to poor model performance. For instance, for the NN trained on batches of size 256 to explore the $c_{qt}^{(8)}$ parameter space, the training loss decreased by only 0.0002 units over 200 epochs, while the test loss exhibited significant fluctuations over time, indicating a lack of convergence.

Coefficient	**Batch size**	**Learning Rate**
$c_{dt}^{(8)}$	150,000	0.0003
$c_{qt}^{(8)}$	200,000	0.001
$c_{dt}^{(8)^2}$	200,000	0.0001
$c_{qt}^{(8)^2}$	200,000	0.0005

Table 1: Hyperparameters adopted for each Neural Network trained at the parton level.

We train each model independently for 200 epochs and subsequently plot the train and test loss curves to analyse their trends. As depicted in Figure 4, the losses reach a plateau after approximately 100 epochs. Each Figure also displays the minimum loss achieved during training, showing that the models trained on the quadratic coefficients generally perform better.

By applying each NN to the datasets, we aim to assess whether each model has learned to distinguish and classify the two sets of events. We do not anticipate that the loss will reach zero due to the similarity in distributions of the input parameters, $m_{t\bar{t}}$ and $y_{t\bar{t}}$, across both the SM and EFT events, which makes separating the two datasets challenging. Our analysis of the model outputs is focused on revealing a consistent trend among the neural networks: classifying SM theory instances towards



(a) Loss curve for the $c_{\text{dt}}^{(8)}$ data

(b) Loss curve for the $c_{\text{qt}}^{(8)}$ data

(c) Loss curve for $(c_{\text{dt}}^{(8)})^2$

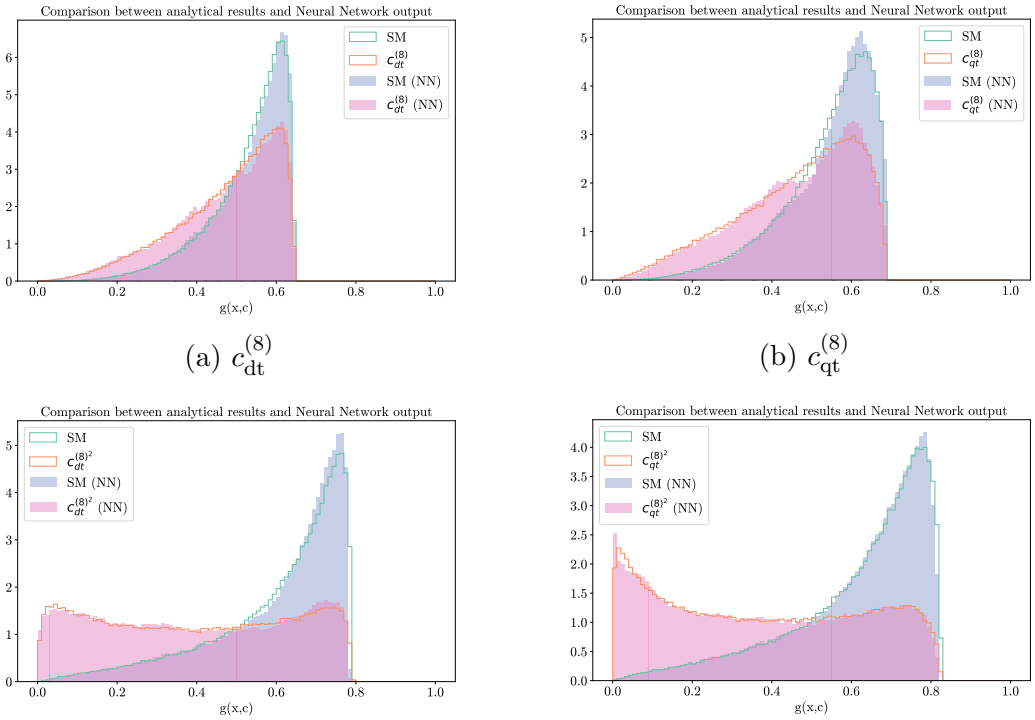
(d) Loss curve for $(c_{\text{qt}}^{(8)})^2$

Figure 4: Loss curves for the Neural Networks trained on the combined SM and EFT datasets at the parton level.

a value of 1, and EFT data towards 0. We consider this result to be sufficient to demonstrate the Neural Networks' ability to learn the decision boundary.

For processes at the parton level we can verify the correct implementation of the NNs by comparing their outputs to the analytical cross-sections [30]. We calculate the total cross-sections σ_{eft} and σ_{sm} as the sum of the weights in the corresponding datasets; we then compute $d\sigma_{\text{eft}}/d\mathbf{x}$ and $d\sigma_{\text{sm}}/d\mathbf{x}$ on the MC generated data. Since each NN returns the decision boundary $g(\mathbf{x})$, we use Equation 22 to rearrange the analytical formulae such that their outcomes match those of the neural networks. We apply this procedure 4 times, one for each coefficient, then plot the distribution of NN outputs together with the theoretical evaluations. The comparisons are shown in Figure 5.

From the plots, we observe that the ML parametrisation is in good agreement with the analytical evaluations, especially for the coefficients in their quadratic form, as we were expecting from the behaviour of the loss functions. This result suggests that the NN trained at the parton level have the potential to deliver a good performance on a higher-dimensional set of features.



(a) $c_{\text{dt}}^{(8)}$ (b) $c_{\text{qt}}^{(8)}$

(c) $c_{\text{dt}}^{(8)}$ - quadratic (d) $c_{\text{qt}}^{(8)}$ - quadratic

Figure 5: Comparison of the Neural Networks outputs and analytical solutions provided at the parton level.

4.3 Complete unbinned analysis

4.3.1 Construction of the unbinned likelihood

We now extend our analysis to include the complete set of kinematic features. In this section, we aim at combining the observables used in the traditional analysis with the ML strategies implemented at the parton level. Our goal is to perform statistical inference to determine the values of the coefficients $c_{dt}^{(8)}$ and $c_{qt}^{(8)}$ in their linear and quadratic form by training a NN for each element of equation 16 and construct informative confidence regions to constrain our results. Similarly to the parton-level analysis, we will examine simulated data from both the SM and EFT. However, we face the additional challenge of not having the analytical calculations implemented in Section 4.2 available for cross-checking.

As mentioned previously (Equation 14), in the unbinned case the likelihood can be expressed as the product of the individual contributions from each event. We are also aware that each NN learns a decision boundary $g(\mathbf{x}, \mathbf{c})$, from which one can derive the cross-section ratios $r_\sigma(\mathbf{x}, \mathbf{c})$ relative to each Wilson coefficient.

In reality, the measured number of events in the dataset, N_{ev}, is not fixed but is instead distributed according to a Poisson distribution with mean $\nu_{\text{tot}}(\mathbf{c})$. This leads us to define an extended unbinned likelihood:

$$\mathcal{L}(\mathbf{c}) = \frac{\nu_{\text{tot}}(\mathbf{c})^{N_{\text{ev}}}}{N_{\text{ev}}!} e^{-\nu_{\text{tot}}(\mathbf{c})} \prod_{i=1}^{N_{\text{ev}}} f_\sigma(\mathbf{x}_i, \mathbf{c}) \tag{28}$$

with corresponding log-likelihood

$$\log \mathcal{L}(\mathbf{c}) = -\nu_{\text{tot}}(\mathbf{c}) + N_{\text{ev}} \log \nu_{\text{tot}}(\mathbf{c}) + \sum_{i=1}^{N_{\text{ev}}} \log f_\sigma(\mathbf{x}_i, \mathbf{c}) \tag{29}$$

where $f_\sigma(\mathbf{x}_i, \mathbf{c})$ is defined in Equation 15. We can manipulate and rearrange the log-likelihood to adapt it to our dataset of separate MC samples for the SM and EFT scenarios. We focus on the third term on the right-hand side of Equation 29:

$$\begin{aligned}
\sum_{i=1}^{N_{\text{ev}}} \log f_\sigma(\mathbf{x}_i, \mathbf{c}) &= \sum_{i=1}^{N_{\text{ev}}} \log \left(\frac{f_\sigma(\mathbf{x}_i, \mathbf{c})}{\frac{1}{\sigma_{\text{sm}}} \frac{d\sigma_{\text{sm}}}{d\mathbf{x}_i}} \frac{1}{\sigma_{\text{sm}}} \frac{d\sigma_{\text{sm}}}{d\mathbf{x}_i} \right) \\
&= \sum_{i=1}^{N_{\text{ev}}} \left(\log \frac{1}{\sigma_{\text{tot}}(\mathbf{c})} \frac{d\sigma_{\text{tot}}(\mathbf{c})}{d\mathbf{x}_i} \Big/ \frac{1}{\sigma_{\text{sm}}} \frac{d\sigma_{\text{sm}}}{d\mathbf{x}_i} + \log \frac{1}{\sigma_{\text{sm}}} \frac{d\sigma_{\text{sm}}}{d\mathbf{x}_i} \right) \\
&= \sum_{i=1}^{N_{\text{ev}}} \log \left(\frac{d\sigma_{\text{tot}}(\mathbf{c})}{d\mathbf{x}_i} \Big/ \frac{d\sigma_{\text{sm}}}{d\mathbf{x}_i} \right) - \sum_{i=1}^{N_{\text{ev}}} \log \frac{\sigma_{\text{tot}}(\mathbf{c})}{\sigma_{\text{sm}}} \\
&= \sum_{i=1}^{N_{\text{ev}}} \log \left(1 + \mathbf{c}_i \frac{d\sigma^i_{\text{eft}}}{d\mathbf{x}} \Big/ \frac{d\sigma_{\text{sm}}}{d\mathbf{x}} \right) - N_{\text{ev}} \log \sigma_{\text{tot}}(\mathbf{c}) + N_{\text{ev}} \log \sigma_{\text{sm}}
\end{aligned}$$

where from the first to the second line we have substituted $f_\sigma(\mathbf{x}_i, \mathbf{c})$ with its definition (Equation 15); from the second to the third line we have eliminated the last logarithmic term relative to the SM cross-section since it did not depend on the Wilson coefficients, and equally for the final term in line 4; in the third line we separated the terms that did not depend on $\mathbf{x}$.
We define a new term, r'_i, which is related to the cross-section ratio r_σ by

$$r'_i = \frac{\sigma^{(i)}_{\text{eft}} / \sigma^{(i)}_{\text{eft}}}{\sigma_{\text{sm}} / \sigma_{\text{sm}}} \frac{\frac{d\sigma^{(i)}_{\text{eft}}}{d\mathbf{x}}}{\frac{d\sigma_{\text{sm}}}{d\mathbf{x}}} = \frac{\sigma^{(i)}_{\text{eft}}}{\sigma_{\text{sm}}} r^{(i)}_\sigma \tag{30}$$

Recall that $r^{(i)}_\sigma$ is related to the decision boundary $g_i(\mathbf{x}, \mathbf{c})$ through

$$g_i(\mathbf{x}, \mathbf{c}) = \frac{1}{1 + r^{(i)}_\sigma(\mathbf{x}, \mathbf{c})} \tag{31}$$

$$\Rightarrow r^{(i)}_\sigma(\mathbf{x}, \mathbf{c}) = \frac{1}{g_i(\mathbf{x}, \mathbf{c})} - 1 \tag{32}$$

We further notice that

$$\begin{aligned}
N_{\text{ev}} \log \sigma_{\text{tot}}(\mathbf{c}) &= N_{\text{ev}} \log \sigma_{\text{tot}}(\mathbf{c}) \frac{L}{L} \\
&= N_{\text{ev}} \log \sigma_{\text{tot}}(\mathbf{c}) L - N_{\text{ev}} \log L \\
&= N_{\text{ev}} \log \nu(\mathbf{c})
\end{aligned}$$

where L is the luminosity of the detector and we have ignored the final term which did not depend on the coefficients $\mathbf{c}$. We finally combine our results to obtain an equation for our log-likelihood function:

$$\log \mathcal{L}(\mathbf{c}) = -\nu_{\text{tot}}(\mathbf{c}) + \sum_{i=1}^{N_{\text{ev}}} \log \left[1 + \mathbf{c}_i \frac{\sigma^{(i)}_{\text{eft}}}{\sigma_{\text{sm}}} \left(\frac{1}{g_i(\mathbf{x}, \mathbf{c})} - 1 \right) \right] \tag{33}$$

where $\mathbf{c}_i = c_{\text{dt}}^{(8)}, c_{\text{qt}}^{(8)}, (c_{\text{dt}}^{(8)})^2, (c_{\text{qt}}^{(8)})^2$ are the Wilson coefficients, $\sigma_{\text{eft}}^{(i)}/\sigma_{\text{sm}}$ are the individual EFT cross-sections normalised by the SM calculated from the sum of the weights; $g_i(\mathbf{x}, \mathbf{c})$ represent the output of the Neural Networks for the linear and quadratic representations of the Wilson coefficients.

If we focus our attention on the argument of the logarithm in Equation 33, we recognise that a negative term would lead to an undefined logarithm, eventually invalidating the whole calculation. This suggests that the second term between square brackets should always be less than 1 to avoid a departure from the region of validity of our prediction. In order to suppress these numerical complications, we restrict the range of possible values of the second term such that the maximum value it can take is ≈ -1.

4.3.2 Neural Network training

We train the MLP models with the complete set of features from the MC-generated data. The observables are the same as described in section 3.1, excluding the weights parameters for the same reasons outlined in section 4.2. These features will represent the training data X, with labels 1 for the SM and 0 for the EFT.

In this scenario, we adopt a different scaling method to ensure that the feature values are mapped to the same range when they are passed through the NNs. Even if each NN is trained and applied independently for each SM-EFT pair, these results are combined and applied to the observed data when passed through the NS algorithm for parameter inference. For this reason, we need to ensure that values of similar ranges are passed onto both the training models and the NS likelihood calculations. We consequently standardise each dataset by subtracting the mean of the SM data and dividing it by its standard deviation. This standardisation strategy ensures that the features are approximately in the same range across all datasets.

The MLP architecture remains the same as the one we implemented in section 4.2, but takes as input a 14-dimensional array corresponding to the complete set of features. The output always represents the decision boundary $g(\mathbf{x}, \mathbf{c})$, therefore the model is trained on the same BCE loss function and optimised using AdamW using similar hyperparameters. More specifically, we adopt the following learning rates and batch sizes:

Coefficient	**Batch size**	**Learning Rate**
$c_{\text{dt}}^{(8)}$	200,000	0.0003
$c_{\text{qt}}^{(8)}$	200,000	0.0001
$c_{\text{dt}}^{(8)^2}$	200,000	0.0001
$c_{\text{qt}}^{(8)^2}$	200,000	0.0005

Table 2: Hyperparameters adopted for each Neural Network trained on the full dataset.

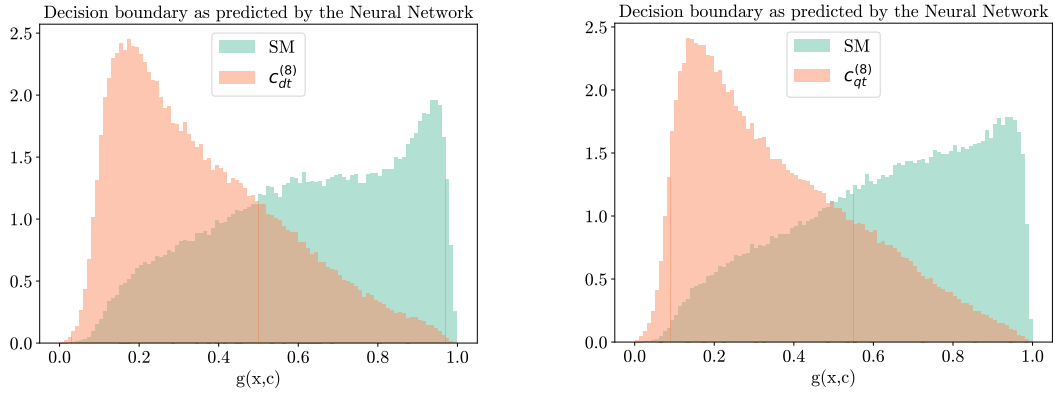
We train our models for 400 epochs to ensure convergence. Table 3 displays the minimum losses achieved by the model in both the parton-level and the full-feature data. Compared to the parton case, the models required more time to explore the full parameter space, but ultimately achieved lower loss minima. This behaviour can

be attributed to the larger set of input features provided to the Neural Networks. Although this increased the model's complexity and computational time, it also enabled the networks to explore a more comprehensive feature space, allowing for a more detailed learning of the underlying data distribution.

Training dataset	**Parton Level**	**Complete Case**
$c_{dt}^{(8)}$	0.663	0.535
$c_{qt}^{(8)}$	0.652	0.538
$c_{dt}^{(8)^2}$	0.576	0.503
$c_{qt}^{(8)^2}$	0.556	0.483

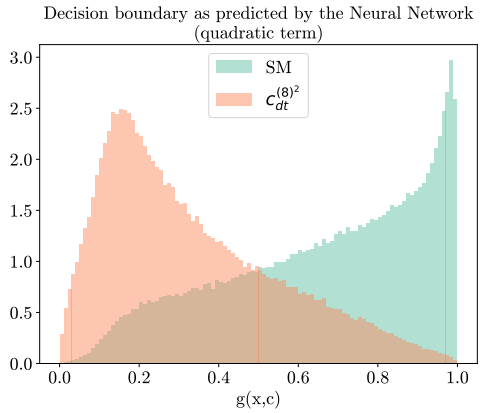
Table 3: Comparison of the minimum losses achieved by each Neural Network at the parton level and in the complete scenario.

We further evaluate the performance of the NNs by applying them to the SM and EFT data and displaying their distributions. Figure 6 highlights that the separation between the SM and EFT datasets is much sharper than in the parton case (Figure 5). This suggests that the Neural Networks have learned to classify the data with significantly greater precision. At the parton level, the models had more difficulty learning the parameter space of the linear coefficients, with a marginal improvement in the quadratic scenario. We still observe this inclination in the complete framework, but recognise an upgraded performance in both the linear and quadratic settings, with a much clearer decision boundary. Furthermore, the $g(\mathbf{x}, \mathbf{c})$ values in this case span the entire range from 0 to 1, further supporting the revised classification strategy of the integrated datasets. The increased accuracy achieved by the full models suggests that the likelihood computations will yield more rigorous results, potentially implying a better performance of the Nested Sampling algorithm.

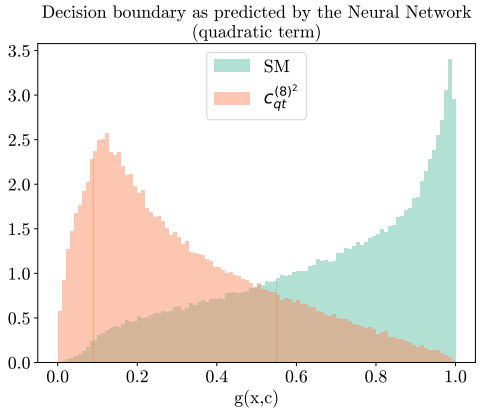


(a) $g(\mathbf{x}, \mathbf{c})$ for $c_{\mathrm{dt}}^{(8)}$

(b) $g(\mathbf{x}, \mathbf{c})$ for $c_{\mathrm{qt}}^{(8)}$



(c) $g(\mathbf{x}, \mathbf{c})$ for $(c_{\mathrm{dt}}^{(8)})^2$



(d) $g(\mathbf{x}, \mathbf{c})$ for $(c_{\mathrm{qt}}^{(8)})^2$

Figure 6: Learned classifications between the full-feature SM and EFT data, parametrised by the decision boundary function $g(\mathbf{x}, \mathbf{c})$.

4.3.3 Nested Sampling implementation

The EFT parameter inference based on the output of the NNs is obtained through Nested Sampling as with the binned case. The likelihood function, defined in Equation 33, includes the total fiducial cross-section ratios calculated from the weights of each sample and factors in the NN outputs when applied to the measured data. The algorithm will explore the complete parameter space on the dataset of measured events and will estimate the values of $c_{\mathrm{dt}}^{(8)}$ and $c_{\mathrm{qt}}^{(8)}$. From these results, we will potentially establish the presence of New Physics in our data by calculating the observed deviation from the SM.

We implement `UltraNest` with 1000 live points and use the observed data as input for the log-likelihood function. Instead of enforcing a specific prior, we maintain a uniform prior with equal range as in the traditional analysis to ensure an unbiased comparison. Table 4 shows a comparison of the Maximum Likelihood Estimates and the 95% confidence intervals obtained by `UltraNest` in the traditional analysis and with the ML approach.

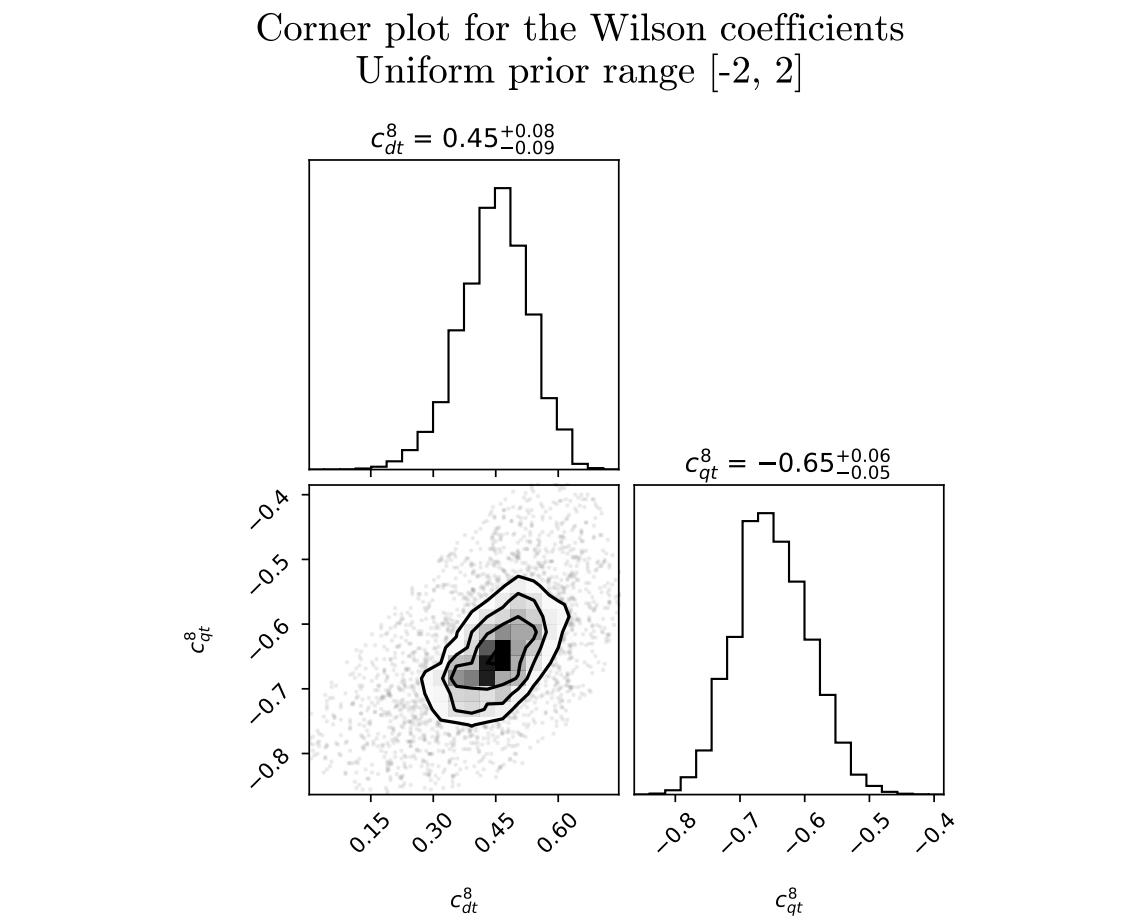
By analysing the output of the NS algorithm, we observe significantly narrower confidence intervals compared to the traditional analysis. This indicates higher precision in the parameter estimation. Confidence regions are a crucial metric for assess-

	Coefficient	Traditional Analysis	ML Parametrisation
MLE	$\boldsymbol{c_{dt}^{(8)}}$	-0.669	0.453
	$\boldsymbol{c_{qt}^{(8)}}$	-0.856	-0.653
95% CI	$\boldsymbol{c_{dt}^{(8)}}$	$[-1.256, 0.324]$	$[0.271, 0.602]$
	$\boldsymbol{c_{qt}^{(8)}}$	$[-0.952, 0.021]$	$[-0.753, -0.536]$

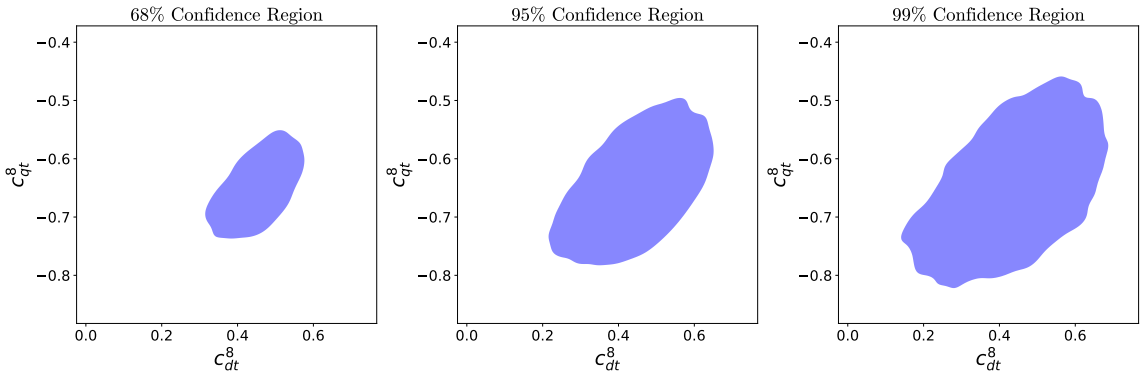
Table 4: Comparison of the Nested Sampling results following the posterior evaluation of the parameter space using binned likelihoods and the Machine Learning approach.

ing the performance of the NN and for comparing this strategy with the traditional method described in section 3.3. Since parametric inference at this stage demands maximal precision, we are primarily interested in the level of *confidence* that we can achieve through both approaches.

The marginal distributions of possible parameter values for both Wilson coefficients, shown in the corner plot of Figure 7a, are narrower and more concentrated around one value. This behaviour in the distributions is also reflected in the credible regions evaluated at 3 different confidence levels, depicted in Figure 7b. The calculated deviation from the SM is more prominent in this case, providing additional support towards the presence of NP at the level of 8σ standard deviations. This further underlines the increased potential of the unbinned strategy to accurately constrain the values of the Wilson coefficients.



(a) Corner plot of the Wilson coefficients distributions obtained from the unbinned analysis.



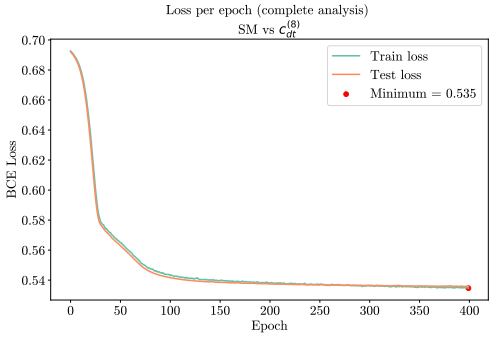
(b) Starting from the left, 68%, 95% and 99% confidence regions of the 2D distribution of the Wilson coefficients.

5 Conclusions

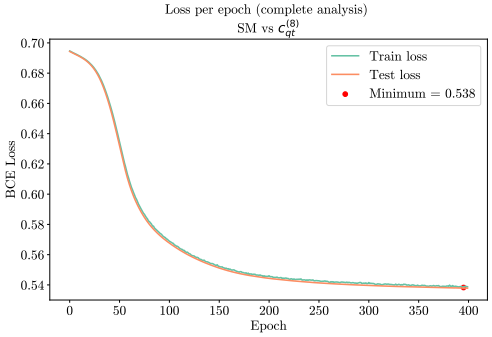
The application of ML to the study of particle physics events holds significant promise for advancing our understanding of the Standard Model Effective Field Theory. Deep learning techniques can efficiently handle the large and complex data measured by particle colliders such as the LHC, and extract more nuanced information than traditional methods. In this project, we have explored how Deep Learning can be employed to optimise the research of physical processes within the SMEFT framework. The analysis of unbinned likelihoods for the construction of Wilson coefficient fits leads to a significant improvement in the constraint of the EFT parameter space. We have shown that using binned measurements results in less accurate fits, making them inadequate for confidently inferring the presence of new physics. To address this concern, we have reproduced and implemented the strategy proposed by the `ML4EFT` framework [1] to parametrise high-dimensional likelihoods and perform statistical inference on the EFT operators. We designed efficient MLPs to derive SMEFT cross-section ratios by accurately modelling the decision boundary between EFT and SM data. An essential aspect of training involved the precise tuning of the Neural Networks, particularly the batch size, to avoid overfitting and obtain optimal results. The analysis at the parton level allowed us to validate our approach, ensuring the reliability of our methods before extending our procedure to a higher-dimensional scenario. From the information learned by the Neural Networks, we described the likelihood of the data and implemented it in `UltraNest` to obtain robust estimates for the two Wilson coefficients, $c_{\mathrm{dt}}^{(8)}$ and $c_{\mathrm{qt}}^{(8)}$. In the comprehensive feature analysis, the Neural Networks have demonstrated to deliver an effective performance, with improved classification between SM and EFT data and lower minimum losses in the parametrisation of higher-dimensional feature spaces. The potential of the Neural Networks implementation lies in their scalability and flexibility: this technique accommodates an arbitrary number of observed features, as MLPs can handle varying input sizes, and an arbitrary number of EFT coefficients, as each model can be trained independently. The global fits analysis achieved better outcomes than traditional approaches for parameter inference. The constrained credible intervals obtained suggest that this revised technique is able to deliver reliable parameter estimates with less uncertainty. This reliability will prove to be crucial in making confident predictions and decisions in the searches for new physics within experimental data.

Appendix A Loss curves - complete analysis

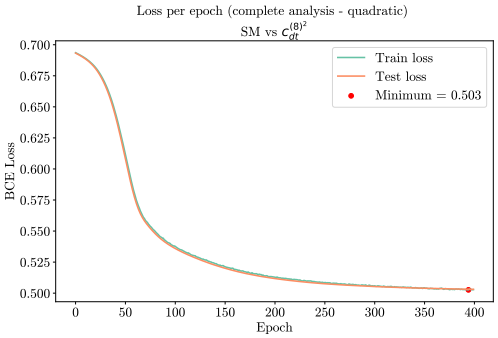
These plots depict the loss curves obtained by training the Neural Networks on the full set of features, using the techniques explained in section 4.3.2.



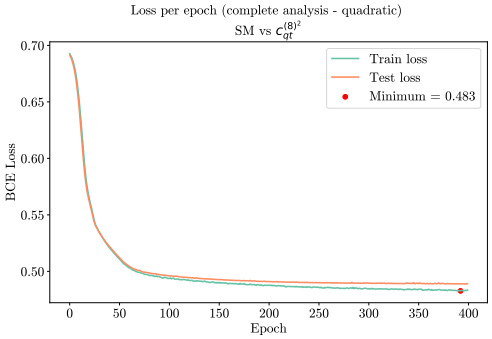
(a) Loss curve for $c_{\text{dt}}^{(8)}$



(b) Loss curve for $c_{\text{qt}}^{(8)}$



(c) Loss curve for $(c_{\text{dt}}^{(8)})^2$



(d) Loss curve for $(c_{\text{qt}}^{(8)})^2$

References

[1] R. G. Ambrosio, J. ter Hoeve, M. Madigan, J. Rojo, and V. Sanz, "Unbinned multivariate observables for global smeft analyses from machine learning," *Journal of High Energy Physics*, vol. 2023, Mar. 2023.

[2] S. L. Adler, "Theories of the fine structure constant α," in *Atomic Physics 3*, (Boston, MA), pp. 73–84, Springer US, 1973.

[3] M. K. Gaillard, P. D. Grannis, and F. J. Sciulli, "The standard model of particle physics," *Reviews of Modern Physics*, vol. 71, no. 2, p. S96, 1999.

[4] D. Griffiths, *Elementary Particle Dynamics*, ch. 2, pp. 55–79. John Wiley and Sons, Ltd, 1987.

[5] G. C. Nayak, "General form of color charge of the quark," *The European Physical Journal C*, vol. 73, pp. 1–25, 2013.

[6] K. Huang, *Quarks, leptons & gauge fields*. World Scientific, 1992.

[7] R. Mann, *An introduction to particle physics and the standard model*. Taylor & Francis, 2010.

[8] D. Laskaroudis, "The energy content of the universe," *Physics Essays*, vol. 29, no. 2, pp. 284–289, 2016.

[9] J. A. Frieman, M. S. Turner, and D. Huterer, "Dark energy and the accelerating universe," *Annu. Rev. Astron. Astrophys.*, vol. 46, no. 1, pp. 385–432, 2008.

[10] C. Weinheimer and K. Zuber, "Neutrino masses," *Annalen der Physik*, vol. 525, no. 8-9, pp. 565–575, 2013.

[11] M. Sozzi, *Discrete symmetries and CP violation: From experiment to theory*. Oxford University Press, 2008.

[12] O. Brüning, H. Burkhardt, and S. Myers, "The large hadron collider," *Progress in Particle and Nuclear Physics*, vol. 67, no. 3, pp. 705–734, 2012.

[13] A. V. Manohar, "Introduction to effective field theories," 2018.

[14] I. Brivio and M. Trott, "The standard model as an effective field theory," *Physics Reports*, vol. 793, p. 1–98, Feb. 2019.

[15] T.-P. Cheng, L.-F. Li, and D. Gross, "Gauge theory of elementary particle physics," 1985.

[16] C. Bailer-Jones, *Statistical models and inference*, p. 55–75. Cambridge University Press, 2017.

[17] C.-Y. Wong, *Introduction to high-energy heavy-ion collisions*. World Scientific, 1994.

[18] J. Buchner, "UltraNest - a robust, general purpose Bayesian inference engine," *The Journal of Open Source Software*, vol. 6, p. 3001, Apr. 2021.

[19] D. Fink, "A compendium of conjugate priors," 1997.

[20] P. M. Goggans and Y. Chi, "Using Thermodynamic Integration to Calculate the Posterior Probability in Bayesian Model Selection Problems," *AIP Conference Proceedings*, vol. 707, p. 59–66, Apr. 2004.

[21] J. Skilling, "Nested sampling for general Bayesian computation.," *Bayesian Analysis*, vol. 1, pp. 833–860, 12 2006.

[22] J. Buchner, "Nested sampling methods," *Statistics Surveys*, vol. 17, Jan. 2023.

[23] N. Chopin and C. P. Robert, "Properties of nested sampling," *Biometrika*, vol. 97, p. 741–755, June 2010.

[24] C. S. Rayat, *Chi-Square Test ($\chi 2$ – Test)*, pp. 69–79. Singapore: Springer Singapore, 2018.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[26] S. J. Prince, *Understanding deep learning*. MIT press, 2023.

[27] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[29] S. Khirirat, H. R. Feyzmahdavian, and M. Johansson, "Mini-batch gradient descent: Faster convergence under data sparsity," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 2880–2887, IEEE, 2017.

[30] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, M. Rüfenacht, M. Schönherr, and G. Watt, "Lhapdf6: parton density access in the lhc precision era," *The European Physical Journal C*, vol. 75, Mar. 2015.