Università degli Studi di Padova
Sede Amministrativa: Università degli Studi di Padova
Dipartimento di Fisica e Astronomia Galileo Galilei

Scuola di Dottorato di Ricerca in Fisica
Ciclo XXXI

PhD Thesis

Statistical Learning and Inference at Particle Collider Experiments

PhD School Coordinator: Prof. Franco Simonetto

Supervisor: Dott. Tommaso Dorigo

PhD Candidate: Pablo de Castro Manzano

Acknowledgements

First and foremost, I would like to thank my PhD supervisor Tommaso Dorigo for his support and guidance during the last three years. In addition to the valuable feedback that he provided on the various projects presented in this document and the much needed help with the seemingly endless bureaucratic endeavours, his advising style was consequential for finding a beneficial balance between exploration of curiosity-driven research ideas and their exploitation. Tommaso is also to blame for many improvements on my writing style as well as the blueprint of a PhD plan including a broad range of training opportunities.

Most of the analysis work within the CMS Collaboration could not have been carried out without the help and support of my postdoc officemates Martino Dall'Osso and Andres Tiko, so I would like to thank them both for their work and their friendship. The collaboration and discussions with Mia Tosi, Alexandra Oliveira and Roberto Rossin were also instrumental for the CMS analysis work. I would like also to express gratitude towards the other members of the CMS group at Padova, that were very welcoming since the beginning and regularly demonstrated interest and gave feedback on the status of my research. I would also like to acknowledge all assistance received by the administrative staff of the INFN - Sezione di Padova.

A substantial fraction of the work presented here was carried out within the CMS Collaboration, and thus was shaped by interactions with some of its thousands of members by different means including insightful comments after presentations, long e-mail discussions, and the review of diverse documents. In addition to thankfully acknowledge the contributions of all my CMS collaborators, I would like to highlight the role of the members of the CMS Higgs group and the HH subgroup in the development of the Higgs pair production analysis as well as of Markus Stoye, Mauro Verzetti, Jan Kieseler and Marcel Rieger regarding the project focussed on the integration DeepJet in the CMS software. Some conversations with Sebastien Wertz, Andre David, Gilles Louppe and Joeri Hermans were also particularly relevant to define some aspects of the work presented in this document.

Within the AMVA4NewPhysics network, which was the research community that motivated many of the research projects of my PhD, I would like to thank all the senior members for their effort in organising the various training and collaboration activities, which were important to create an open and productive research environment. Most importantly, I would like to thank the other network students (aka ESRs), which were always available for stimulating chats about research, life or the universe over coffee or beers. I would also like to express my gratitude towards Sabine Hemmer and Pietro Vischia for the energy they devoted to motivate and manage scientific outreach in the form of blogging and Twitter, which turned out to be very positive experiences.

In addition, I would like to thank Giovanna Menardi and Bruno Scarpa from the UNIPD Statisics department for hosting me for one month and providing feedback on the statistical aspects of my research. Maurizio Sanarico at SDG Consulting was also very accommodating during the months spent in Milan. Also to Daniel Whiteson, Peter Sadowski, and other attendants of the ML for HEP meeting in UCI, which gave initial support for the new machine learning technique described in this thesis. I am also thankful to Kostas Vellidis and IASA for hosting me in Athens. The environment provided by CERN both remotely and during the secondment were essential for carrying out the majority of my research so I would like to thank everyone that works to make such institution function as it does.

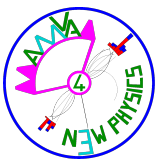
One achievements are nothing but the product of incremental improvements over the work, ideas and shared experiences of others, thus the present section is guaranteed to be incomplete. Hence if a conversation we had, a paper you published, a blog post you produced, a presentation you gave, a question you answered, a book you wrote or some open-source software you developed was helpful to the work presented in this thesis, I am sincerely thankful. Last from an academic perspective, I would like to acknowledge Francisco Matorras and Alexander Read for their willingness for their work as external reviewers of this thesis.

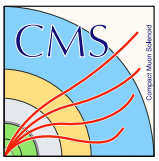
At a personal level, moving to Padova was a bit challenging so I am grateful to have been embraced by the Scambio di Lingue group and their Touch Rugby team, through which I met many new friends and provided a great social context in this marvellous city. I would like to thank my family, particularly my parents and my grandparents, for the unconditional love and support they have provided since I happen to remember. Finally, I would like to thank Ksenija, whom I met halfway through this adventure and has been a wonderful companion since, for her persistent encouragement and affection.

Preface

This document is a summary of the main projects that I have carried out within my PhD in Physics at the University of Padua (UNIPD), between December 2015 and the December 2018. The main research focus, connecting the projects presented here, has been the development and application of new statistical learning techniques in particle collider experiments. Given the interdisciplinary nature of the topics discussed in this thesis, an effort has been made to discuss the research issues and solutions in a domain generic manner, so the links with the fields of statistics and machine learning are more evident. The price of such attempt has likely been a less cohesive narrative, yet I believe that is has been worth the cost for a different take on the data analysis problems at particle colliders.

Most of the work included in this report has been carried out while employed by the INFN - Sezione di Padova as an Early Stage Researcher of the AMVA4NewPhysics MSCA-ITN. AMVA4NewPhysics is a European research network (EU Horizon 2020 Grant Agreement 675440) that provided the funding and context for the ventures described in this document. Part of the results presented here were joint work with other collaborators at CMS experiment at LHC, which is based at the European Organisation for Nuclear Research (CERN).





Abstract

Advances in data analysis techniques may play a decisive role in the discovery reach of particle collider experiments. However, the importing of expertise and methods from other data-centric disciplines such as machine learning and statistics faces significant hurdles, mainly due to the established use of different language and constructs. A large part of this document, also conceived as an introduction to the description of an analysis searching for non-resonant Higgs pair production in data collected by the CMS detector at the Large Hadron Collider (LHC), is therefore devoted to a broad redefinition of the relevant concepts for problems in experimental particle physics. The aim is to better connect these issues with those in other fields of research, so the solutions found can be repurposed.

The formal exploration of the properties of the statistical models at particle colliders is useful to highlight the main challenges posed by statistical inference in this context: the multi-dimensional nature of the models, which can be studied only in a generative manner via forward simulation of observations, and the effect of nuisance parameters. The first issue can be tackled with likelihood-free inference methods coupled with the use of low-dimensional summary statistics, which may be constructed either with machine learning techniques or through physically motivated variables (e.g. event reconstruction). The second, i.e. the misspecification of the generative model which is addressed by the inclusion of nuisance parameters, reduces the effectiveness of summary statistics constructed with machine-learning techniques.

A subset of the data analysis techniques formally discussed in the introductory part of the document are also exploited to study the non-resonant production process $\mathrm{pp} \to \mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ at the LHC in the context of the Standard Model (SM) and its extensions in effective fields theories (EFT), based on anomalous couplings of the Higgs field. Data collected in 2016 by the CMS detector and corresponding to a total of 35.9 fb^{-1} of proton-proton collisions are used to set an 95% confidence upper limit at 847 fb on the production cross section $\sigma\big(\mathrm{pp} \to \mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}\big)$ in the SM. Upper limits are also obtained for the cross sections corresponding to a representative set

of points of the parameter space of EFT. The combination of those results with the ones obtained from the study of other decay channels of HH pairs is also discussed.

In addition, the exercise of reformulating the goals of high energy physics analysis as a statistical inference problem is combined with modern machine learning technologies to develop a new technique, referred to as inference-aware neural optimisation. The technique produces summary statistics which directly minimise the expected uncertainty on the parameters of interest, optimally accounting for the effect of nuisance parameters. The application of this technique to a synthetic problem demonstrates that the obtained summary statistics are considerable more effective than those obtained with standard supervised learning methods, when the effect of the nuisance parameters is significant. Assuming its scalability to LHC data scenarios, this technique has ground-breaking potential for analyses dominated by systematic uncertainties.

Contents

Introduction

> Every new beginning
> comes from some other beginning's end.
>
> Seneca the Younger

Humans strive for understanding the world by seeking explanations to the varied natural phenomena happening around them, and accumulating the resulting knowledge in models that can be used to predict and shape the future reality. The scientific method provides a formal framework for carrying out these investigations and checking the validity of the current description of our environment. Recorded experiences of assumed known origin, also known as data, have a central role in updating these explicative theories, because they can provide quantitative or qualitative support to some candidate explanations over others.

Direct sensory perception and personal information processing have a limited investigative reach and are easily affected by subjective conditions. Well understood and calibrated measurement instruments can be used instead for data acquisition, in controlled settings referred to as scientific experiments, so that quantifiability and precision are enhanced. The same applies to theoretical modelling and experimental data analysis, where robust mathematical and computational procedures empower researchers to construct more accurate descriptions of the world we live in. These establish a strong coupling between technology and science, by which technical and conceptual innovations allow the development of better tools, which in turn lead to more scientific knowledge.

The universe is filled with an abundance of interesting phenomena occurring at very different time and space scales, so curious observers might face a difficult choice when deciding what to focus their scientific attention on. Nevertheless, there seems to be a complexity hierarchy whereby larger physical systems are composed by simpler parts, and the properties of the former can be explained by means of those of the latter. Hence, a worthy path of exploration can start with the study of the most fundamental components of nature and their dynamics. At our current level of

understanding, we can reason this would be a quest motivated solely by curiosity, pushed by our desire of making sense of the structure of reality, and not a pragmatic proxy for the development of technological applications. That will be our motivation to delve into experimental particle physics, a discipline dealing with the practical study of the most elementary constituents of matter and their interactions.

The elementary quality of the chosen subject of study does not imply that the journey towards valuable scientific knowledge in this area will be a simple one. On the contrary, as the following chapters will make evident, this undertaking poses grand technical and non-technical challenges which in many cases require novel solutions. Furthermore, the problems at hand are often closely related with those present in other research or technological fields, so their findings and innovations can be repurposed. Oftentimes this can even be a bidirectional relation, where the obstacles are challenging or original enough that solutions have to go beyond the state of the art in the relevant applied domain. In general, the pursuance of fundamental explanations does require solutions to a multitude of practical problems.

Advances and expertise from other disciplines can accelerate significantly the rate of progress in a fundamental research domain such as experimental particle physics. This is specially relevant in areas such as data analysis, where the infrastructure changes required in evolving environments are low. Yet, some barriers exist against the proliferation of interdisciplinarity, such as field specific language (also known as jargon) and seemingly unclear problem descriptions for collaborators with different backgrounds. This document, in addition to presenting the main research results of the projects I have been involved in the recent past, will attempt to reduce this communication gap by trying to clearly state the main data analysis challenges we face in experimental particle physics in a way they can be linked to other data-centric disciplines such as statistics and machine learning.

The general methodology considered in this work consists on breaking the main research goals in a series of applied problems, express them in a domain-generic way, and understand what is their role in view of the final aim. When possible, the presented concepts and methods will be illustrated with simple use cases when these can help understanding their working principles. The mentioned perspective shift combined with the use of practical but minimal examples has been really useful to identify possible shortcomings on the way data analysis is carried out at the LHC, as well as to develop new techniques capable of addressing them. Nevertheless, we believe that the projects mentioned and presented here are nothing but the first step of what is possible; and the evolution of data analysis techniques and tools could be

a promising route for the advancement of our understanding of the basic building blocks of the universe.

This thesis is organised as follows. Chapter 1 provides an overview of our current comprehension of the properties and interactions of the fundamental constituents of nature, followed by a summary of the limitations of our understanding together with the main proposed testable alternative explanations. The links between the mathematical description of our universe and the computation of experimental observables will be highlighted when describing the theoretical foundations.

The focus shifts in Chapter 2 towards how these theories can be experimentally validated through scientific experiments. In particular, the discussion revolves around how the design and characteristics of general purpose experiments at high-energy colliders are relevant for the attainment of valuable data that yields new insights on the fundamental properties of the cosmos. The Compact Muon Experiment (CMS) detector at the Large Hadron Collider (LHC) serves as the default example of such an instrument, because it is the scientific experiment that provided the academic context during my graduate (and late undergraduate) years and the main driver of some of the projects included in this report. Experimental modelling and simulation will be emphasised in this chapter, due to their importance when extracting knowledge from the acquired data.

Indeed, the problem of obtaining useful information from data is so involved in modern scientific experiments that a standalone chapter will be centered on statistical inference concepts and techniques. Inference is the ultimate goal of particle physics experiments, providing a key connection between theory and experiment. In Chapter 3 we review the problem at hand in particle colliders form a formal statistical perspective as well list the main approaches for making quantitative statements based on data and their shortcomings. Two domain-specific aspects of data analysis in high energy physics will be remarked: the generative-only characteristic of accurate experimental models and the challenges of dealing with known unknowns we are not interested in, commonly referred as nuisance parameters.

Advancements in computational power coupled with extensive research effort at the intersection between computer science and statistics during the past few decades have contributed to the development of techniques that deal with the automatic improvement of certain objective tasks given some data. An introduction to this family of methods, generally referred to as machine learning techniques, and a review of their usefulness for tackling some common data analysis problem in experimental particle physics, are included in Chapter 4. Some non-trivial connections between

the use of those techniques and the details of the underlying statistical issues will be stressed.

The first four chapters, as outlined above, offer a multi-disciplinary survey of the theoretical and experimental foundations of our understanding of nature and the relevant techniques that allow the extract valuable information from the data. In contrast, Chapter 5 presents a complete example of an analysis at the LHC that applies those techniques to a real-world scenario. Specifically, the use case will be the search for evidence of anomalous non-resonant Higgs boson pair production using CMS data at the LHC, which can be a smoking gun pointing to alternative explanations to the current theoretical comprehension of the fundamental interactions and constituents of the universe.

The aforementioned example will be useful to epitomise the main statistical and methodological challenges on the way LHC analyses are carried out. In Chaper 6, we try to shed some light on these issues, and demonstrate how a novel machine learning technique we have developed can deal with one of the most relevant concerns: learning summary statistics using inference-aware losses that account for the effect of nuisance parameters. The limitations of the proposed method as well as alternative solutions to increase the discovery potential of the LHC will be explored.

This document will conclude with Chapter 7, where the main contributions and outcomes of this work will be summarised together with some ideas for future extensions and improvements.

1 Theory of Fundamental Interactions

> Nothing in life is to be feared.
> It is only to be understood.
>
> Marie Skłodowska Curie

Scientific theories are frameworks describing natural phenomena that are capable of making experimentally testable predictions. Oftentimes, they are specified using mathematical language and built on previous observational knowledge and basic properties of the system under study. At the most fundamental scales known to date, the Standard Model (SM) of particle physics is a scientific theory that provides a very accurate description of most of the observed properties and dynamics of the universe around us. It is constructed upon an innovative theoretical framework, generally referred as quantum field theory (QFT), and principles regarding fundamental symmetries of the laws of nature. In this chapter, a non-exhaustive introduction to this theory and its descriptive reach will be provided together with a summary of the known limitations and possible extensions or alternatives. Given the experimental character of the research discussed in the following chapters, the aim of this chapter is not solely the discussion of the basic structure and properties of the theory, but also the methodology followed to compute predictions for observables that can be contrasted with empirical data.

1.1 The Standard Model

The Standard Model (SM) of particle physics is a mathematically self-consistent gauge field theory that classifies all known types of elementary particles and describes their electromagnetic, weak and strong interactions. Within this fundamental theory, all known matter and energy phenomena can be explained in terms of the kinematics and interactions of elementary particles, which can in turn be understood as local excitations of different fields that permeate our universe.

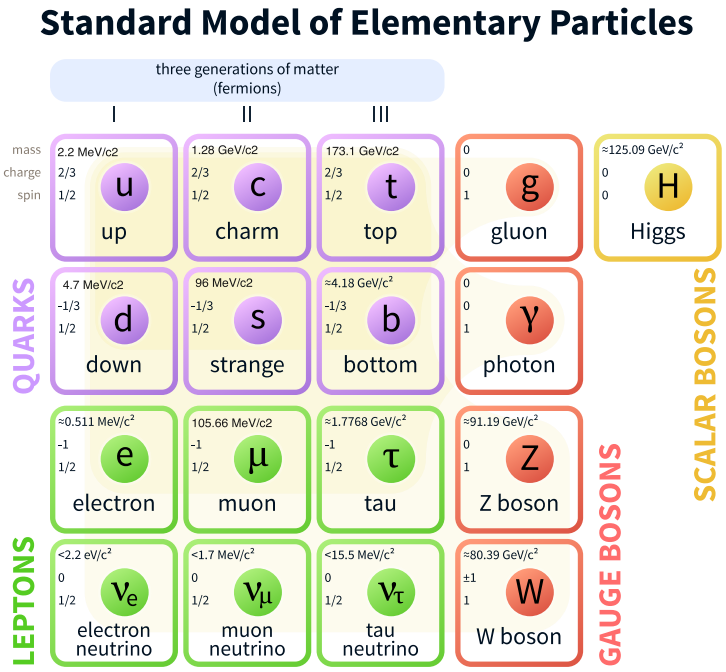


Figure 1.1: Schematic overview of the particle content within the SM. Fundamental particles include fermions, further subdivided in quarks and leptons, and fundamental bosons, including the force mediators and the Higgs boson. Diagram adapted from MissMJ (CC BY 3.0 license).

From a historical perspective, this theory is the product of a succession of important theoretical developments and experimental discoveries over the last century [1], culminating with the discovery of the Higgs boson in 2012 [2, 3]. If a more principled viewpoint is taken, the SM can be thought of as the most general but mathematically consistent theory that respects a set of symmetries, namely a global Poincaré group symmetry (translational, rotational and relativistic boost invariance) and a local

$$G_{\mathrm{SM}} = SU(3)_C \otimes SU(2)_L \otimes U(1)_Y \tag{1.1}$$

gauge group symmetry. The G_{SM} symmetry group is essential to describe three of the four fundamental interactions observed in nature: strong interaction, weak interaction and electromagnetic interaction. In fact, the $SU(3)_C$ is associated the strong force and the conservation of its charge, called colour, while the $SU(2)_L \otimes U(1)_Y$ symmetry instead is related with electroweak interactions (i.e. unification of weak and electromagnetic) and the conservation of isospin and weak hypercharge. The SM is typically specified using the Lagrangian formalism and depends on a total of 19 parameters (not accounting for neutrino masses and mixing angles), which are not predicted by the theory from first principles, and thus can only be determined through experimental measurements.

In the context of the SM, excitations of the fundamental fields give rise to two types of elementary particles: fermions (characterised by having half-integer spin) and bosons (characterised by having integer spin). Fermions are the fundamental constituents of matter, and they are further subdivided into leptons and quarks depending on their interactions. A schematic overview of the fundamental particles of the SM and their properties is provided in Figure 1.1. Three particle generations are known for both quarks and leptons, each containing a pair of particles with different masses. For quarks, the heavier is referred to as up-type and the lighter as down-type. Instead, for leptons we distinguish the heavier charged particles (electron, muon and tau) from their corresponding light and uncharged neutrinos.

Regular matter is largely made of the first generation of quarks and electrons, given that higher generations rapidly decay quickly to lower generations characterised by smaller masses. All fermions interact via the weak force but only quarks carry colour charge and are subjected to the strong force. For each fermion in the SM, there is a another particle with identical properties but opposite quantum numbers, globally referred to as antimatter, and denoted for each particle with the anti prefix and a bar over the symbol (e.g. up antiquark $\bar{u}$) or by explicitly denoting the charge sign

(e.g. positron e^+). Neutrinos are the only fermions that do not carry electrical charge and might be their own antiparticle.

The mediators of the strong, weak and electromagnetic fundamental interactions are referred to as gauge bosons, and are characterised by having spin 1. To model the strong interaction colour charge exchanges, a total of eight independent strong massless force mediators, or *gluons*, are needed. Gluons carry colour charge themselves and thus participate in colour interactions with other gluons, which leads to a phenomenon known as *colour confinement*, which will be discussed in Section 1.1.2 in more detail. The massless and neutral *photon* is the mediator of the electromagnetic force, while instead the massive Z, W^+ and W^- bosons mediate weak interactions. The last piece in the SM is the *Higgs boson*, the only fundamental known particle with spin 0. The Higgs boson is the quantum excitation of the *Higgs field*, which also couples with other fundamental particles such as the gauge bosons of the weak force, effectively generating their mass through their interaction. The Higgs boson and Higgs field play an essential role in the electroweak symmetry breaking (EWSB) mechanism, which will be discussed in more detail in Section 1.1.4.

The rest of this section will be devoted a more mathematically exhaustive review of the different components of the Standard Model, starting by reviewing the basic formalism of quantum field theories and incrementally building on it do describe the characteristics of both the strong and electroweak interactions that give rise to the diverse interactions dynamics of relevance in particle physics experiments. The mentioned review is heavily inspired by standard bibliographical references on the topic [4, 5], and which are recommended directly for a more detailed survey on the subject.

1.1.1 Essentials of Quantum Field Theory

As hinted in the previous section, in quantum field theory (QFT), observed particles are understood as excitations of fields that extend through the whole universe. Quantum field theory unifies the physical foundations of quantum mechanics and special relativity, and can be used to accurately describe phenomena in systems where relativistic and quantum effects are relevant, such as interactions between highly relativistic particles. In QFT, all the known physical processes in the universe are explained in terms of the state and dynamics of a set of fundamental tensor fields. A tensor field can be defined as a continuous and differentiable set of values, such a scalar or a vector, that exist for any given location and time. For simplicity, the

fields in QFT are usually defined in a relativistic coordinate system $x = (t, \boldsymbol{x})$ in order treat space $\boldsymbol{x}$ and time t jointly.

To exemplify the fundamentals of the QFT framework, let us consider the simplest case, e.g. a single field that does not interact with any other field, which will be denoted as $\phi(x)$. The dynamics of a field (or several fields) in QFT are specified by using the *Lagrangian formalism*, similarly to what can be done for systems in classical mechanics. However, instead of considering the Lagrangian L which depends the generalised coordinate vector $\boldsymbol{q}(t)$ and its time derivatives $\dot{\boldsymbol{q}}(t)$, in QFT the Lagrangian density $\mathcal{L}$ is commonly used, which depends only on the field $\phi(x)$ and its first derivative $\partial_\mu \phi(x)$. In an analogous manner to what is done in classical mechanics to define the action functional $S_{\text{classical}}$, we can define the action of the quantum field S_{QFT} as a function of the Lagrangian density $\mathcal{L}$ as follows:

$$S_{\text{classical}} = \int L(\boldsymbol{q}(t), \dot{\boldsymbol{q}}(t)) dt \quad \Rightarrow \quad S_{\text{QFT}} = \int \mathcal{L}(\phi, \partial_\mu \phi)\ d^4x \tag{1.2}$$

noting that the previous definition would also be valid when the Lagrangian depends on multiple fields and their derivatives instead of a single free field. Identically to what is done in classical systems, we can attempt to solve for the field that minimises the action, i.e. $\delta S = 0$. With the help of some functional calculus [6], it is possible to obtain the relativistic field theory version of the Euler-Langrange equation:

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0 \tag{1.3}$$

where $\partial_\mu = \partial / \partial x_\mu$ and the repetition of the coordinate index $\mu \in \{0, 1, 2, 3\}$ means summation over the product. The previous relation would still apply to each field in the case a Lagrangian including several fields was considered; therefore, given a Lagrangian, we can use Equation 1.3 to obtain their equations of motion. As an example, let us consider the following Lagrangian $\mathcal{L}_{\text{Dirac}}$, which is a function of a bispinor field ψ, a 4-dimensional complex vector field that can represent a field whose excitations behave like fermions of mass m:

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi \tag{1.4}$$

where γ^μ are the gamma matrices and $\bar{\psi} = \psi^\dagger \gamma^0$ is the spinor adjoint. As the chosen naming for the previous Lagrangian $\mathcal{L}_{\text{Dirac}}$ gave away, the Euler-Lagrange relation

obtained by minimising the action $\delta S = 0$ can be used to obtain field equations of motion that correspond to the Dirac equation [4] for the spinor field and its adjoint:

$$i\gamma^\mu \partial_\mu \psi - m\psi = 0 \quad \text{and} \quad i\gamma^\mu \bar{\psi} \partial_\mu + m\bar{\psi} = 0 \tag{1.5}$$

as well as the well-known Klein-Gordon equation component-wise $(\partial^\mu \partial_\mu + m^2)\psi = 0$, where $\partial^\mu = \partial/\partial x^\mu$. Both Dirac and Klein-Gordon equations were proposed in the context of a relativistic formulation of quantum mechanics.

To shed some light on how a field like ψ can represent actual fermions in the universe, such as electrons or positrons, the field can be quantised by considering a plane wave expansion and defining annihilation operators $a^s_{\boldsymbol{p}}$ and $b^s_{\boldsymbol{p}}$, as well as creation $a^{s\dagger}_{\boldsymbol{p}}$ and $b^{s\dagger}_{\boldsymbol{p}}$ operators. The field and its adjoint, which can be thought of directly as operators instead of fields in this context, may then be expressed as:

$$\psi(x) = \int \frac{d^3p}{(2\pi)^3} \frac{1}{\sqrt{2E_{\boldsymbol{p}}}} \sum_s \left(a^s_{\boldsymbol{p}} u^s(p) e^{-ipx} + b^{s\dagger}_{\boldsymbol{p}} u^s(p) e^{ipx} \right) \tag{1.6}$$

$$\bar{\psi}(x) = \int \frac{d^3p}{(2\pi)^3} \frac{1}{\sqrt{2E_{\boldsymbol{p}}}} \sum_s \left(b^s_{\boldsymbol{p}} \bar{v}^s(p) e^{-ipx} + a^{s\dagger}_{\boldsymbol{p}} \bar{u}^s(p) e^{ipx} \right) \tag{1.7}$$

where $u^s(p)$ and $v^s(p)$ and its adjoints are the free particle solutions of the Dirac equation, s is their spin and $E_{\boldsymbol{p}}$ their energy. The operators in the previous quantisations can be used to define arbitrary many-particle states. The vacuum state $|0\rangle$ can be defined as the state for which $a^s_{\boldsymbol{p}}|0\rangle = b^s_{\boldsymbol{p}}|0\rangle = 0$. A single free fermion state of momenta $\boldsymbol{p}$ and spin s can be obtained by applying the creation operators on the vacuum state $|\boldsymbol{p}, s\rangle = \sqrt{2E_{\boldsymbol{p}}} a^{s\dagger}_{\boldsymbol{p}}|0\rangle$ - or alternatively an anti-fermion if the $b^{s\dagger}_{\boldsymbol{p}}$ is used instead. Multi-particle free states in momenta representation can analogously be defined by the successive application of creation operators over momenta space.

In particle colliders, we are instead interested in interacting theories rather than free theories, given the we aim to compute total and differential cross sections. Interacting theories can also be characterised by their Hamiltonian density $\mathcal{H} = \mathcal{H}_{\text{free}} + \mathcal{H}_{\text{int}}$, which can be expressed as a function the Lagrangian density $\mathcal{H} = \pi^a \dot{\psi}_a - \mathcal{L}$, where $\dot{\psi}_a$ is the time derivative of the field and π^a is the conjugate momentum. The Hamiltonian density can divided in $\mathcal{H}_{\text{free}}$, that is the part corresponding to the free theory, and $\mathcal{H}_{\text{int}}$ that are the additional terms due to interactions. In interacting theories, time-dependence becomes more important and depends only on the $\mathcal{H}_{\text{int}}$ component. Additionally, the ground state $|\Omega\rangle$ can be different in interacting theories from the free theory vacuum state $|0\rangle$.

Let us denote by $|i\rangle = |\psi(t \to -\infty)\rangle$ and $|f\rangle = |\psi(t \to +\infty)\rangle$ some arbitrary initial and final multi-particle states, temporarily far before and after the actual interaction being studied happened (i.e. around $t = 0$), respectively. The observables of interest, which are discussed in Section 1.3, are a function of the transition amplitude $\langle i|\mathcal{S}|f\rangle$ over all possible initial and final states, where $\mathcal{S}$ is an operator describing the transition. The transition probability, which is expressed as the modulus square of the amplitude $|\langle i|\mathcal{S}|f\rangle|^2$, is therefore also a function of $\mathcal{S}$, fully describing the time-evolution from the initial the final state. The $\mathcal{S}$ operator may be expressed as a perturbative series using the Dyson expansion:

$$\begin{aligned}\mathcal{S} &= T\left[\exp\left(-i\int_{-\infty}^{\infty} d^4x \mathcal{H}_{\text{int}}(x)\right)\right] \\ &= \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int_{-\infty}^{\infty} d^4x_1 ... \int_{-\infty}^{\infty} d^4x_n T[\mathcal{H}_{\text{int}}(x_1)...\mathcal{H}_{\text{int}}(x_n)]\end{aligned} \tag{1.8}$$

where T is an operator ensuring that the Hamiltonian density factors $\mathcal{H}_{\text{int}}(x_i)$ are ordered in time. Each time-ordered term in the series can be written as a sum of normal (i.e. not time ordered) products of permutations using Wicks theorem [7], which can become rather tedious for high orders. The formalism of Feynman diagrams can be used to simplify the computation of observables at a given order in the perturbative expansion.

Based on the previous perturbative series expansion, the transition amplitude $\langle i|\mathcal{S}|f\rangle$ can be easily linked with scattering observables when denoted as:

$$\langle i|\mathcal{S}|f\rangle = \langle i|\mathbf{1}|f\rangle + i\mathcal{M}(2\pi)^4\delta^4\left(\sum p_i - \sum p_f\right) \tag{1.9}$$

where the first term corresponds to no interaction occurring, and the second includes the matrix element $\mathcal{M}$ including all orders in the perturbative orders, and multiplied by a factor making explicit the conservation of momentum between the initial and final state particles. The matrix element $\mathcal{M}$, which can be computed perturbatively as a function of the momenta of the particles given final state considered, can be used to define the differential cross section:

$$\frac{d\sigma}{d\Phi} \sim |\mathcal{M}|^2 \text{ where } d\Phi = (2\pi)^4\delta^4\left(\sum p_i - \sum p_f\right)\prod_f \frac{1}{2E_f}\frac{d^3\boldsymbol{p}_f}{(2\pi)^3} \tag{1.10}$$

where the proportionality factor is a function of the initial state particles momenta and $d\Phi$ is the full phase space differential element for which can be generally ex-

pressed as a product of the final state particle momenta differential elements. Total scattering rates can be obtained by summing over possible initial and final states and integrating over final states. Both differential and total cross sections can be truncated at a given perturbative order. The lowest expansion order is referred as leading order (LO), yet considering additional expansion can greatly increase the prediction accuracy so one (NLO) or two (NNLO) orders are often considered, higher orders often being too computationally challenging. A truncation at an additional order n, relative to the lowest interaction order, will provide corrections proportional to $\alpha = g^2/(4\pi)$, where g is the coupling constant characteristic of the interaction.

1.1.2 Quantum Chromodynamics

In a hadron collider such as the LHC, strong interactions between quark and gluons are dominant, and they can be modelled using quantum chromodynamics (QCD). The theory of QCD can be linked to a $SU(3)$ symmetry group and is described by the following gauge invariant Lagrangian density:

$$\mathcal{L}_{\text{QCD}} = \bar{\psi}(\gamma^\mu D_\mu - m_f)\psi - \frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a, \quad \psi = \begin{bmatrix} \psi_r \\ \psi_g \\ \psi_b \end{bmatrix} \tag{1.11}$$

where ψ is a spinor quark field for a given flavour $f \in \{\text{u}, \text{d}, \text{s}, \text{c}, \text{b}, \text{t}\}$ and quark mass m_f, and each vector component represents a colour degree of freedom. Assuming that the Gell-Mann matrices λ^a are used to define a basis for the gluon field $A_\mu = 1/2\lambda^a \sum A^a_\mu$, the covariant derivative can be defined as $D_\mu = \partial_\mu - i g_s A_\mu$, where g_s is the strong interaction coupling. In turn, the gluon field strength tensor $G^a_{\mu\nu}$ is also related with the gluon field components:

$$G^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g_s f^{abc} A^b_\mu A^c_\nu \tag{1.12}$$

where f^{abc} are the structure constants of the $SU(3)$ gauge group. The last term accounts for the self-interaction of the gluon, which are the massless and electrically neutral mediators of the strong force. There are two properties of QCD that play an important role from a phenomenological standpoint: *confinement* and *asymptotic freedom*.

The property of confinement has been postulated to explain why isolated quarks and gluons are not found in nature. Quarks have only been found as part of hadrons, that are colour-neutral composite particles. Even though confinement has not been

understood from first principles, because the observables of bound states in QCD at low-energies cannot be computed in a perturbative manner, there exist extensive evidence both from lattice QCD calculations and experiments. In a bound state between quarks, the effective potential includes a term that increases proportional to their distance, so when the quarks are separated by an external energetic interaction, the additional potential energy generates an additional quark-antiquark pair, leading to the formation of bound states. Similar phenomena occur for isolated gluons, which generally are referred as hadronization, and can be understood as a consequence of colour confinement. In particle colliders, successive hadronization and radiation processes led to parton showers (see Section 1.3.4).

Quark are then only found in bound states, referred to as hadrons, which can either be mesons or baryons. Mesons are formed by quark-antiquark pairs $q\bar{q}$, while baryons are composed of three quarks qqq. Charged and neutral pions π^+ ($u\bar{d}$) and π^0 ($(u\bar{u} - d\bar{d})/\sqrt{2}$), kaons K^+ ($u\bar{s}$) and K^0 ($d\bar{s}$) and the J/Ψ ($c\bar{c}$) are among the most common mesons produced at particle colliders. Baryons instead include the well-known proton (uud) and neutron (udd) that together with electrons are the constituents of most of the known matter in the universe. Many more short-lived baryons exist [8], in addition to the recently discovered exotic bound states referred as tetraquarks [9] and pentaquarks [10]. A detailed description of the compositeness of proton is an essential element for computing LHC observables, as reviewed in Section 1.3.2.

Asymptotic freedom is instead linked with the strength reduction of the strong coupling constant when higher energy scales are considered. Let us consider a renormalisation energy scale μ_R^2, which has to be often defined in order to compute physical observables which otherwise would be divergent due higher order perturbative corrections which cannot be easily calculated. This effect can be also understood as a coupling that varies with the energy scale, which is referred to as a "running" coupling constant. The strong force coupling $\alpha_s = g_s^2/(4\pi)$ can thus be approximated as a function of the renormalisation energy scale μ_R^2 as follows:

$$\alpha_s(\mu_R^2) = \frac{\alpha_s(\mu_0^2)}{1 + \alpha_s(\mu_0^2)\frac{33-2n_f}{12\pi}\ln\left(\frac{\mu_R^2}{\mu_0^2}\right)} \tag{1.13}$$

where $\alpha_s(\mu_0^2)$ is the measured coupling at a given energy and n_f is total number of quark flavours which are assumed to be massless in this approximation. The strong interaction thus becomes weaker at higher energies (or short distances) allowing

the perturbative computation of observables related with high-energy interactions, as discussed in Section 1.3. The approximation from Equation 1.13 also provides a lower bound for the energy scale at which QCD can be treated perturbatively, i.e. the denominator becomes zero for an energy scale around 200 MeV, leading to a diverging coupling constant.

1.1.3 Electroweak Interactions

The remaining two fundamental interactions between elementary particles are the electromagnetic and the weak force. The description of the electromagnetic interaction in terms of quantum fields and gauge symmetries, leading to the development of quantum electrodynamics (QED) in the late 1940s, prompted a quest for an analogous theory for the weak force. The weak force, known to be responsible for the beta decay at the time, could effectively be modelled using Fermi theory using four-fermion interactions [11] but was not renormalisable and lacked the predictive capabilities and elegance of QED. A large theoretical effort lead to an alternative description based on a $SU(2) \otimes U(1)$ symmetry, which unified electromagnetic and weak interactions [12, 13], and where the weak interaction was mediated by means of charged $W^{\pm}$ and neutral Z massive vector bosons. Nevertheless, the theory did not provide an explanation for the mass of the weak mediators, until the so-called Brout-Englert-Higgs [14, 15, 16] mechanism for spontaneous symmetry breaking (SSB) was conceived. Higgs also noted explicitly that the mechanism would effectively create an additional scalar field, associated with a new scalar boson, whose existence could experimentally testable. The SSB mechanism was then combined with $SU(2) \otimes U(1)$ unified theory [17] to give rise to what is now known as *electroweak theory*, which was then proved to be renormalisable [18].

The different testable properties of electroweak phenomena were verified by experiments including the existence of weakly-interacting neutral and charged currents [19] and the discovery of the massive $W^{\pm}$ [20, 21] and Z [22, 23] bosons. Experimental evidence also showed that weak interactions were parity violating [24], thus in the electroweak theory the fermion fields are separated in their left-handed ψ_{L} and right-handed ψ_{R} chiral components as follows:

$$\psi_{\mathrm{L}} = \mathrm{P}_{\mathrm{L}}\psi = \frac{1}{2}(1 - \gamma_5)\psi \quad \psi_{\mathrm{R}} = \mathrm{P}_{\mathrm{R}}\psi = \frac{1}{2}(1 + \gamma_5)\psi \tag{1.14}$$

where P_{L} and P_{R} are the chiral projection operators and $\gamma_5 = i\gamma_0\gamma_1\gamma_2\gamma_3$ is the product of the gamma or Dirac matrices. For massless particles, chirality is equal to

the helicity $H = (\boldsymbol{p} \cdot \boldsymbol{s})/|\boldsymbol{p}|$ which is the sign of the scalar product of momenta and spin. For massive particles, chirality is still defined but is not identical to helicity which cannot be invariantly defined.

Within the electroweak theory, fermion fields are broken in into their left-handed components, which can be expressed as doublets that would transform under $SU(2)$, and can be denoted as:

$$L_q = \left\{ \begin{pmatrix} u \\ d \end{pmatrix}_L, \begin{pmatrix} c \\ s \end{pmatrix}_L, \begin{pmatrix} t \\ b \end{pmatrix}_L \right\} \quad L_l = \left\{ \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, \begin{pmatrix} \mu_\tau \\ \tau \end{pmatrix}_L \right\} \tag{1.15}$$

and their right handed components, that instead can be expressed as singlets only transforming under $U(1)$:

$$R_u = \{u_R, c_R, t_R\} \quad R_d = \{d_R, s_R, b_R\} \quad R_l = \{e_R, \mu_R, \tau_R\} \tag{1.16}$$

where the right-handed neutrino components are omitted in the electroweak theory (and the SM), given they are electrically neutral and would not interact weakly when right-handed.

The electroweak interactions then can be made explicit by introducing additional boson fields $W = \{W^1, W^2, W^3\}$ and B which will interact with the fermions. Similarly in structure to QED (and also QCD as described in Section 1.1.2), the electroweak Lagrangian before spontaneous symmetry breaking is composed by interaction terms for the previous doublet and singlet fields, characterised by a covariant derivative, and kinematic terms for both boson fields:

$$\begin{aligned} \mathcal{L}_{\text{EW}} = & \sum^{\psi \in \{L_q, L_l\}} \bar{\psi}(i\gamma_\mu D_L^\mu)\psi + \sum^{\psi \in \{L_q, L_l\}} \bar{\psi}(i\gamma_\mu D_R^\mu)\psi \\ & - \frac{1}{4} W_{\mu\nu} W^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \end{aligned} \tag{1.17}$$

where the covariant derivatives for left-handed D_L^μ and right-handed D_R^μ fermion fields are respectively defined as:

$$\begin{aligned} D_L^\mu &= \partial^\mu - \frac{1}{2} g_B Y B_\mu - \frac{1}{2} g_W \sigma W_\mu \\ D_R^\mu &= \partial^\mu - \frac{1}{2} g_B Y B_\mu \end{aligned} \tag{1.18}$$

where $\sigma = \{\sigma_1, \sigma_2, \sigma_3\}$ are the Pauli matrices and g_B and g_W are the coupling constants. The $W_{\mu\nu}$ and $B_{\mu\nu}$ field strength tensors from kinematic terms can in turn be obtained as:

$$\begin{aligned} W^i_{\mu\nu} &= \partial_\mu W^i_\nu - \partial_\mu W^i_\mu - g_W \epsilon^{ijk} W^i_\mu W^k_\nu \\ B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\mu B_\mu \end{aligned} \tag{1.19}$$

where ϵ^{ijk} is the Levi-Civita symbol for each permutation, which is the structure constant for $SU(2)$.

1.1.4 Symmetry Breaking and the Higgs Boson

The problem with the electroweak theory as described by the Lagrangian from Equation 1.17, which is based on Yang-Mills gauge theory formulation, is that it is not possible to directly add mass term for the fermions nor the weak bosons to the Lagrangian density without breaking the $SU(2)$ invariance. At the time the mentioned theory was developed, there was extensive evidence not only for lepton masses but also for the weak bosons being massive; the mass required to explain why the weak interaction was short-ranged. The issue of lacking a theoretical mechanism that could explain the mass of fermions and weak boson was solved by the spontaneous symmetry breaking mechanism [14, 15, 16], which is based on postulating the existence of an additional complex scalar field ϕ, which is a $SU(2)$ doublet with the following structure:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \begin{pmatrix} \phi_3 + i\phi_4 \\ \phi_1 + i\phi_2 \end{pmatrix} \tag{1.20}$$

where we made the component notation explicit because it will be relevant later. This scalar field is expected to interact with the electroweak fields W and B by means of the following Lagrangian:

$$\mathcal{L}_{\text{scalar}} = (D^H_\mu \phi)^\dagger (D^\mu \phi) - V(\phi) \tag{1.21}$$

where the covariant derivate in this case is defined as:

$$D^\mu_H = \partial^\mu - \frac{1}{2} i g_B Y B_\mu - \frac{1}{2} i g_W \sigma W_\mu. \tag{1.22}$$

The minimal form for a scalar field potential $V(\phi)$, constructed ad-hoc to provide a degenerate vacuum states and a local maximum - a required condition for spontaneous symmetry breaking, may be expressed as:

$$V(\phi) = -\mu^2\phi^\dagger\phi + \frac{1}{2}\lambda(\phi^\dagger\phi)^2 \tag{1.23}$$

where both the quadratic μ^2 and the quartic λ self-interaction parameters are defined positive with this sign convention. The resulting shape for the potential is often referred as *mexican hat*, and is depicted in Figure 1.2. The presence of a potential minimum different from the origin gives rises to a non-zero vacuum expectation value for the scalar field:

$$\langle\phi\rangle_0 = \frac{\mu^2}{\lambda} = v^2 \tag{1.24}$$

whose values depend on the $V(\phi)$ potential parameters μ^2 and λ, and it is denoted as v^2 for convenience. The non-zero vacuum expectation value is thus said to spontaneously break the the $SU(2)\otimes U(1)$ symmetry, the consequences made more clear when the field is expanded around the minimum:

$$\phi = \frac{1}{\sqrt{2}}\exp(i\frac{\sigma\cdot G}{v})\begin{pmatrix}0\\v+H\end{pmatrix} \tag{1.25}$$

as a product of a scalar field H and a complex exponential of the scalar product of a three-component field $G = \{G_1, G_2, G_3\}$ with the Pauli matrices $\sigma = \{\sigma_1, \sigma_2, \sigma_3\}$. The complex exponential phase can be then removed by a $SU(2)$ group rotation, a transformation that is often referred as *unitary gauge*. The resulting scalar field can simply be expressed as:

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+H\end{pmatrix} \tag{1.26}$$

where three of the four degrees of freedom in Equation 1.20, which correspond the field G which would otherwise give rise to the so-called Goldstone bosons, have been removed after the gauge transformation.

Substituting the rotated scalar field from Equation 1.26 in the Lagrangian described by Equation 1.21 leads to mass-like terms for linear combinations of the W

and B fields. In order to obtain the physical bosons observed in nature, the mass terms have to be made independent by the following transformations:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2) \quad \begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix} \tag{1.27}$$

where the fields W^+ and W^- are associated with the charged weak bosons, the field Z with the neutral weak boson, the electromagnetic field A with the photon, and g_W is the Weinberg angle which is related with the electroweak couplings according the relation $\tan\theta_W = g_B/g_W$. Omitting for now the terms related with the H field, the Lagrangian in Equation 1.21 leads to the following mass terms for the electroweak force mediators after the unitary gauge and the transformation described in Equation 1.27 have been applied:

$$\begin{aligned} \mathcal{L}_{\text{EW bosons}} = & \frac{1}{2}\underbrace{\left(\frac{g_W^2 v^2}{4}\right)}_{m_{W^+}^2} W_\mu^+ W^{+\mu} + \frac{1}{2}\underbrace{\left(\frac{g_W^2 v^2}{4}\right)}_{m_{W^-}^2} W_\mu^- W^{-\mu} + \\ & \frac{1}{2}\underbrace{\left(\frac{g_W^2 v^2}{4\cos\theta_W}\right)}_{m_Z^2} Z_\mu Z^\mu + \frac{1}{2}\underbrace{(\ 0\ )}_{m_\gamma^2} A_\mu A^\mu \end{aligned} \tag{1.28}$$

resulting in mass terms for the massive weak bosons which depend to the weak coupling, the Weinberg angle and the vacuum expectation value of the Higgs field. The last term for the electromagnetic field has only been included to make explicit that no mass term is associated with the electromagnetic force carrier γ. The terms related with the scalar H field (and Higgs boson) are discussed later independently.

In addition to providing a mechanism that leads to mass terms for the weak force bosons, additional interactions of the various fermion fields with the scalar field ϕ can explain their masses. These gauge invariant terms are generally referred to as Yukawa interactions, and correspond to the following Lagrangian terms:

$$\begin{aligned} \mathcal{L}_{\text{Yukawa}} = & -\lambda_l(\bar{L}_l \phi R_l + \bar{R}_l \phi^\dagger L_l) \\ & -\lambda_d(\bar{L}_q \phi R_d + \bar{R}_d \phi^\dagger L_q) \\ & -\lambda_u(\bar{L}_q i\sigma_2 \phi^\dagger R_u + \bar{R}_u i\sigma_2 \phi L_q) \end{aligned} \tag{1.29}$$

where λ_l , λ_d and λ_u are the Yukawa coupling parameters. A charge-conjugate transformation $\phi \to i\sigma_2\phi^\dagger$ is used to give mass to up-type quarks. For the quark sector, the λ_u and λ_d couplings can be expressed by a single non diagonal matrix

in the flavour basis, referred to as Cabibbo-Kobayashi-Maskawa (CKM matrix) [25, 26], which can in turn be parametrised by three angles and a complex phase. The fact that the matrix is not diagonal leads to flavour mixing, due to the mass eigenstates being different from flavour eigenstates. Another relevant property of fermion masses is that after spontaneous symmetry breaking, the fermion mass is effectively proportional to its coupling with the Higgs scalar field, which is useful to intuitively understand the dominant interactions and decays of the Higgs boson.

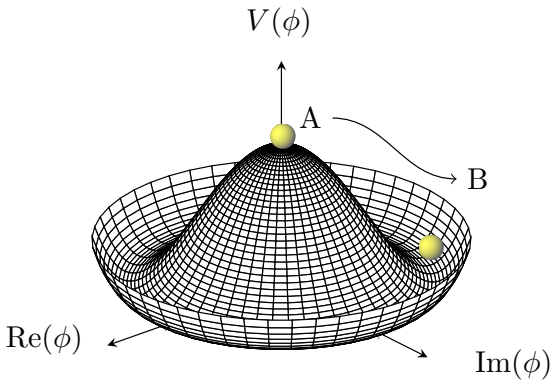


Figure 1.2: Graphical depiction[1] of the mexican hat potential for the scalar field ϕ. A local maximum is present at the origin, but lower energy degenerate minima exist around it.

In addition of giving masses to both weak bosons and fermions, the remaining degree of freedom after electroweak symmetry breaking gives rise to a scalar field H. The terms of the Lagrangian concerning only H may be obtained substituting Equation 1.26 in Equation 1.21, leading to the following expression:

$$\mathcal{L}_H = \frac{1}{2}\partial_\mu H \partial^\mu H - \mu^2 H^2 - \lambda v H^3 - \frac{\lambda}{4} H^4 \tag{1.30}$$

where the second (quadratic term) can be interpreted as a scalar boson with a mass $\sqrt{2\mu^2}$, which is commonly referred as the Higgs boson. A particle with a mass of 125.09(24) GeV [27] and consistent with the expected properties for the Higgs boson was discovered in 2012 by the CMS and ATLAS collaborations [3, 2]. The cubic λv and quartic λ terms will give rise to self-interaction interaction vertices. The so-called cubic or trilinear Higgs coupling is discussed in a Higgs pair search using data from the CMS experiment in Chapter 5. The direct determination of the Higgs

self-coupling is an relevant missing piece, and an important proof of consistency of the spontaneous symmetry breaking mechanism.

1.2 Beyond the Standard Model

The experimental success of the Standard Model and its main subcomponents QED, QCD, and EW unification and symmetry breaking is clearly incontestable, ranging from the confirmation of theoretical prognostication of the existence and some the properties of new particles (e.g. Z, $W^{\pm}$ and Higgs bosons or top quark) to the agreement of precise predictions with meticulous experimental observations. The fine structure constant α at zero energy scale is an example of the latter, with its experimentally determined value consistent among independent physical measurements when the Standard Model based theoretical correction are accounted, down to 12 significant digits [28, 29]. In addition to describing natural phenomena with unprecedented accuracy, the SM is a self-consistent theory that provides non-divergent predictions at the highest energy scales probed to date.

1.2.1 Known Limitations

In spite of the successes mentioned above, several shortcomings of the Standard Model are known and hence the theory is not considered as a complete theory of natural phenomena at the most fundamental scales. Those concerns include unexplained empirically observed phenomena such as gravitational interactions, neutrino masses or dark matter particle candidates, theoretical considerations regarding the stability of vacuum or aesthetic principles such as naturalness. Hence, it is presumed that the Standard Model is an effective theory, able to successfully describe fundamental processes within a range of energies as an approximation of a more complete unified theory. For completeness, the main empirical and theoretical concerns are summarised:

- **Omission of gravitational interactions**: the current formulation of the SM completely disregards the effect of gravity in fundamental interactions, because no consistent quantum descriptions for gravity matching the experimental predictions of the well-established theory of general relativity [30] have been developed to date. While several theoretical efforts are ongoing, such as loop quantum gravity [31] or string theory [32], the coupling for gravitational interactions at the current experimental high-energy reach is expected to be

more than 30 times weaker than for weak interaction, and hence can be safely ignored when computing theoretical predictions.

- **Lack of a viable Dark Matter candidate**: through a variety of astrophysical observations, including the observed galaxy rotation curves [33], gravitational lensing [34] and the Cosmic Microwave Background (CMB) [35], there is clear evidence indicating the presence of more gravitational interacting matter in the universe than what is expected by contrasting with the electromagnetic spectra. It has been thus estimated that about 85% of massive existing matter in the universe does not notably interact with ordinary matter and radiation, and therefore is referred as *Dark Matter*. While its particular nature is still unknown, scientific consensus seems to favour long-lived cold non-baryonic matter as an explanation, predominantly weakly-interacting massive particles (WIMPs). The three neutrino types are the only WIMP within the Standard Model, but considering the known upper limits on their masses, they can only account for a very small fraction of the total mass of dark matter in the universe.

- **Unexplained matter-antimatter asymmetry**: as discussed in Section 1.1, each matter particle in the Standard Model has an identical anti-matter possessing opposite quantum numbers. Because pair creation and annihilation processes are symmetric, but our universe is manifestly dominated by what we refer as matter, some asymmetric interaction processes ought to exist. Within the SM, some electroweak processes are known to violate CP-symmetry and potentially explain a small part of the observed matter-antimatter asymmetry. New unknown CP-symmetry processes, potentially through interactions not included in the SM, are needed to resolve the mentioned disparity.

- **Origin of neutrino masses**: the Standard Model was developed assuming that neutrinos were massless, yet is currently well established that neutrinos oscillate between different flavour eigenstates [36, 37], implying that flavour states mix and hence that neutrino masses are very small but different from zero. The SM Lagrangian can be extended to account for the masses of neutrinos in a similar fashion to what is done for leptons and quarks, but their Yukawa coupling has to be much smaller than of any of the other particles, and it requires the existence of very weakly interacting right-handed neutrinos. An alternative mechanism for including neutrino masses exists, and it is based on assuming that these particles are Majorana fermions and hence they are their

own anti-particle. This hypothesis is currently being experimentally tested. It also worth noting that in order to explain the smallness of neutrino masses in a principled way, the Seesaw mechanism [38] has been proposed, which implicitly assumes that the SM is only a low-energy scale effective theory of a more complete unified theory.

- **Mismatch between vacuum energy and Dark Energy**: in addition of providing evidence for dark matter, astrophysical observations such as studies of the properties of the Cosmic Microwave Background [35] or the redshift of type Ia supernovae [39], consistently point to the hypothesis of an accelerating expansion of the current universe. The simplest way to account for this in cosmological models is to include a cosmological constant, which should be understood as an intrinsic energy density of the vacuum, exerting a negative pressure and therefore driving the observed expansion of the universe. In fact, in order to reconcile the theoretical models with experimental observations, about 68% of the total energy in the present universe would correspond to this type of unknown energy density, generally referred to as *Dark Energy*. In most quantum field theories, such as the Standard Model, some non-zero zero-point energy originating from quantum fluctuations is expected. However, modern attempts to predict energy densities from QFT are at variance with the observed energy vacuum energy density, some of them differing by 120 orders of magnitude [40].

- **Naturalness, hierarchy and fine-tuning concerns**: as discussed at the beginning of Section 1.1, the SM can be thought of the most general theory based on a set symmetries, and its 19 parameters (or 26 accounting for neutrino masses and mixing angles) are not obtained from first principles but measured experimentally. Having such a large number of free parameters and observing large differences among their relative magnitude has been viewed as a theoretical concern from an aesthetic perspective. A related issue is why the electroweak energy scale (epitomised by the Higgs mass) is much smaller than the assumed cut-off scale of the SM, where gravitational interactions become relevant at $M_{\text{Planck}} \approx 10^{19}\text{GeV}$, which is generally referred as the *hierarchy problem*. In the absence of New Physics or additional interaction mechanisms, the only way to obtain the observed Higgs mass from the bare Higgs mass (at zero energies) is through a very precise cancellation of divergences, which is regarded as an *unnatural* or *fine-tuned* property of the SM theory.

Other possible issues, in some cases related with those discussed, have also been raised. One of them is the apparent vacuum meta-stability [41] and other the so-called strong CP problem [42]. Many of these questions can be clarified once the higher precision measurements of the SM become available, which are mainly obtained in particle collider experiments.

1.2.2 Possible Extensions

The known limitations stated in the previous section have motivated the development of alternative theories for describing fundamental interactions. Given the quantitative success of the Standard Model, most of the known proposed theoretical models are either extensions of the SM or its associated predictions can be effectively reduced to those of the SM at the energy range current being explored in particle physics experiments. The set of alternatives that have been proposed is too substantial to be exhaustively listed here, especially given that many of the alternatives include additional free parameters that greatly modify the expected theoretical observables.

Precision Measurements of the SM

Due the existing large space of alternatives to the SM from a theoretical standpoint, the exploration of all possibilities through dedicated searches becomes unattainable. An alternative way to possibly obtain quantitative information pointing to extension of the SM is to measure its most relevant observables with high precision. If significant discrepancies are found between the experimental measurement and the theoretical prediction of those observables, it could be evidence pointing to New Physics outside the SM.

Effective Field Theories

In addition to carrying out precision measurements and model-specific searches, there exists a practical way to consider possible extensions due to New Physics phenomena occurring at a higher energy scale Λ than the one being probed, which will be denoted by E. The model-independent approach often referred to as *effective field theory* (EFT) [43, 44] allows to compute observables by extending the SM Lagrangian terms from Section 1.1 with additional operators:

$$\mathcal{L}_{\text{EFT}} = \mathcal{L}_{\text{SM}} + \sum_i \frac{c_i}{\Lambda^{d_i-4}} \mathcal{O}_i \tag{1.31}$$

where $\mathcal{O}_i$ are referred to as *effective operators*, describing the characteristics of the new interactions that are considered in the extended theory and c_i are the the *EFT or Wilson coefficients* that parametrise the strength of those new interactions. The integer d_i defines the dimension of the operator $\dim(\mathcal{O}_i) = [E]^{d_i}$, and while in principle an infinite set of operators with any dimension $d_i > 4$ can be considered, their effects is expected to be suppressed by $(E/\Lambda)^{d_i-4}$ thus high-dimensional operators may be neglected when studying the dominant effects of an EFT extension of the SM.

If all the EFT coefficients c_i are zero or the new energy scale Λ is infinite, the EFT theory reduces to the SM Lagrangian. Instead, if $\Lambda \approx E$, the effective approximation in Equation 1.31 does not hold, and the interactions have to be realistically modelled using a complete theoretical description of the New Physics scenario under study. While in general effective field theories are not renormalisable, observables and higher-order corrections can be computed, because of the well-defined cutoff energy scale Λ. The best-known example of an EFT that has been used in practice is Fermi theory, which is a useful simplification to compute EW observables at low-energies $E \approx 10$ MeV rather than an extension of the SM, given that the detailed structure of electroweak interactions due to $\mathrm{W}^{\pm}$ boson mediating β decays was unknown at the time.

At the LHC and other collider experiments, the main use case of EFT is to describe generic extensions of the SM that could arise due to New Physics at energy scales that are not directly accessible. From an experimental standpoint, the goal is thus to constraint the values of the EFT operator coefficients using experimental data. Because the for $d_i = 5$ the only possible operator is relevant for neutrino phenomenology [45], the set of Lagrangian operators of interest at collider experiments often corresponds to $d_i = 6$ dimension operators. The large set of possible dimension six operators can be greatly reduced by requiring that the main experimentally verified properties of the SM are respected, such as the gauge and Poincaré symmetries, or baryon number conservation. In Chapter 5, a subset of dimension six EFT operators are used to study non-resonant extensions of Higgs pair production in a model-independent manner.

1.3 Phenomenology of Proton Collisions

Once the properties and limitations of the theoretical model that best describes the current understanding of the fundamental structure and dynamics of nature have

been described, we can delve into how to model proton-proton collisions from a quantitative perspective, so theoretical predictions can be contrasted with experimental results at the LHC. The focus of this section then is to make sense of the various outcomes of high-energy proton-proton collisions and how we can predict their relative rates of occurring given some initial state conditions of the interaction.

1.3.1 Main Observables

A related consideration that is useful as an introduction to the aforementioned topic is the question of what outcomes can originate as a result of proton-proton collisions. An answer somehow circular but compatible with our current interpretation of the universe is that everything that could be produced would be produced, meaning that any outcome that can happen in a way that is consistent with the underlying properties of nature is possible. Even though probably the true description of the properties of nature is not known, as discussed in Section 1.1, the Standard Model provides an effective model and restricts considerably the space of possible outcomes, in a way that can be compared with experimental observations. It is worth noting that alternative descriptions of nature, such as those motivated by the known limitations of the SM and reviewed in Section 1.2, may provide alternative mechanisms for the production of outcomes that are not allowed by the SM, and hence often drive the experimental searches for evidence of New Physics.

For those physical processes that could happen as a product of a proton-proton collision, under the assumption of validity of a particular theoretical model, their total expected rate of occurrence is one the most relevant quantities to predict and compare with observations. To ease its experimental interpretation, the rate of occurrence of any given subnuclear process is commonly expressed as a cross section σ, which has dimensions of area and is typically expressed in submultiples of barn ($1\ \text{barn} = 10^{-28}\text{m}^2$). The advantage of cross sections over rates is that their value is independent from the density of the incident particle fluxes. The rate, or probability per unit of time, of a process occurring can be computed simply by multiplying its cross section by the instantaneous luminosity $\mathcal{L}_{\text{inst}}$, which corresponds to the number of particles per unit of area per unit of time crossing in opposite directions in the collision volume.

Another related concept, which is especially important for simulating interactions, is the differential cross section $d\sigma$. While the initial state conditions are fixed, the rate of occurrence of a physical process can be expressed as a function of some final-state variables, such as the angle and energy of outgoing particles. While these

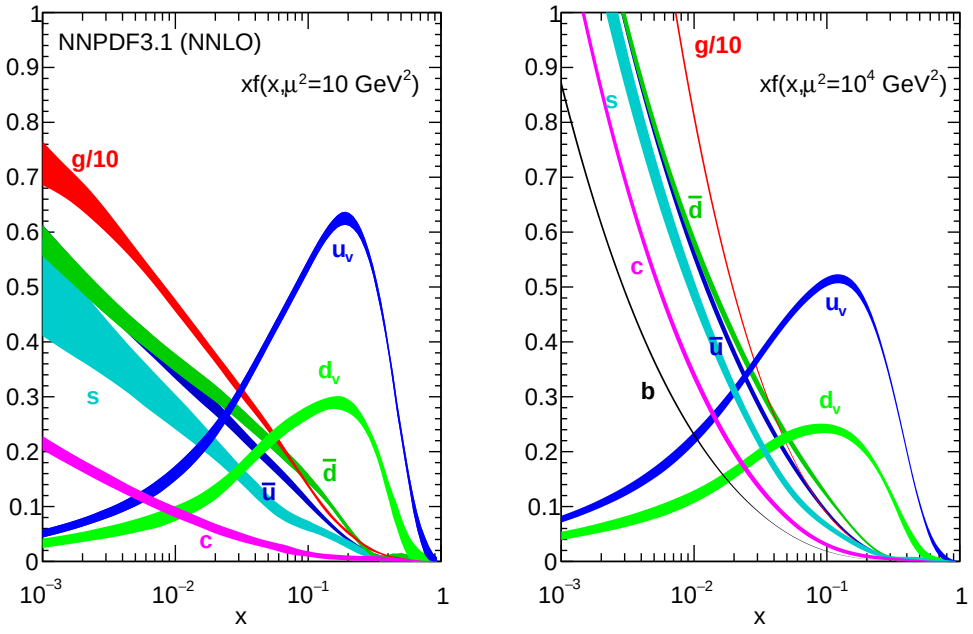
variables can be integrated over to compute total cross sections σ, the integrand is proportional to the probability density of each outcome happening as a function of final-state variables, hence its evaluation is crucial for a correct modelling of their multi-dimensional distributions via random sampling. In fact, we will be dealing with differential cross sections instead of total process cross section in this section for generality.

1.3.2 Parton Distribution Functions

A complication that has not been addressed yet is that protons are composite particles, which within a static interpretation can be thought of as the combination of two up-type quarks and one down-type quark bound together via the strong force. The dynamics of proton-proton scattering are then dictated by quantum chromodynamics (see sec. 1.1.2), which cannot be addressed using perturbation theory for low energies, limiting the first principles computation of relevant observables for the most common interactions. That said, predictions regarding the interaction outcomes from the hard scattering of proton constituents (referred to as partons) can be perturbatively approximated under the assumption of asymptotic freedom at high energies. This allows the modelling of very high energy collisions at particle colliders, which are the focus of most LHC analyses, even if the details about the parton structure cannot be calculated.

When modelling hard (i.e. high energy) scattering processes, a non-perturbative input is required, mainly the probability of finding a particular proton constituent with a certain momentum fraction inside each of the colliding protons, referred to as the parton distribution function (PDF). The model of the proton as three quarks coupled by strong force is too simplistic for modelling proton-proton scattering realistically, especially at high energies. The continuous exchange of gluons between the three constituent quarks effectively generates a sea of virtual quark-antiquark pairs from which other partons can scatter off. Consequently, in the interaction of two protons, not only the constituent quarks, referred as to valence quarks, can take part in the hard scattering process but also gluons and sea quarks.

At the time of writing, PDFs are not computable from first principles so they have to be parametrised and extrapolated from various experimental sources including fixed-target proton deep inelastic scattering (DIS) and previous collider studies. It is worth noting that the distribution functions depend strongly on the energy scale of the process, yet the evolution for parton densities can be modelled theoretically [47, 48, 49]. Given their relevance for computing observables in high-energy colliders,



(a) Low Energy Scale $\mu^2 = 10\text{GeV}^2$ (b) High Energy Scale $\mu^2 = 10^4\text{GeV}^2$

Figure 1.3: Distribution functions for the different partons at low and high energies. The contribution from gluons shown is 1/10 of the actual contribution. Image adapted from the NNPDF collaboration [46].

several research collaborations such as NNPDF [46] provide accurate estimations that can be readily used for simulation and prediction. In Figure 1.3 are shown the parton distribution functions at two different energy scales estimated by one of those collaborations, at lower energy scales the valence quarks (up and down) dominate while when we extrapolate at higher energies, gluon scattering become the most likely outcome for the interaction.

1.3.3 Factorisation and Generation of Hard Processes

Let us consider the computation of the differential cross section for a hard scattering process $pp \to X$, which will be denoted as $d\sigma(pp \to X)$, for two protons colliding head on at centre of mass energy s. Here X denotes a possible outcome for the interaction, not necessarily a single particle and the proton remnants (e.g. a Higgs boson $X = H + \text{other}$), but a set of particles (e.g. a bottom quark-antiquark pair $X = \text{b}\bar{\text{b}} + \text{other}$). According to the QCD factorisation theorem [50], the differential cross section for $d\sigma(pp \to X)$ can be expressed as a sum of functions of the partonic cross section $d\hat{\sigma}_{ij\to X}$:

$$d\sigma(pp \to X) = \sum_{i,j} \int f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) d\hat{\sigma}_{ij\to X}(s x_1 x_2, \mu_R^2, \mu_F^2) dx_1 dx_2 \qquad (1.32)$$

where i and j indicate the partons involved (e.g. a certain type of quark or a gluon), $f_i(x_1, \mu_F^2)$ and $f_j(x_2, \mu_F^2)$ are their parton distribution functions for given momentum fractions x_1 and x_2 respectively, μ_F is the factorisation scale and μ_R is the renormalisation scale. The differential partonic cross section $d\hat{\sigma}_{ij\to X}$ for a centre of mass energy of the interacting partons $\hat{s} = s x_1 x_2$, can be calculated perturbatively at different expansion orders from the Lagrangian density as hinted in Section 1.1. The total cross section $\sigma(pp \to X)$ can then be attained by integrating out all final state quantities, commonly referred as phase space variables, in the differential total cross section element $d\sigma(pp \to X)$. It is worth pointing out that for simple cases (small number of particles in the final state) is often possible to integrate out the final state phase space variables directly in the partonic differential cross section $d\sigma(ij \to X)$, and thus directly compute the total cross section by a similar parton distribution function integration as the one used in Equation 1.32.

As more more complex final states or higher perturbative orders are considered, the final state phase space integration over many particles can rapidly become in-

tractable. This motivates the use of *Monte Carlo integration* techniques, especially those based on importance sampling such as VEGAS [51], which provide convergence rates that scale well with the integral dimensionality by randomly sampling the multi-dimensional space. In fact, the initial state integration over parton types and momenta fractions can also be carried out jointly with these methods, greatly simplifying the computation procedure. The resulting weighted random samples can be used to estimate not only the total cross section, but also any other observable or distribution that is a function of the differential cross section $d\sigma(pp \to X)$. A common observable that is often used in experimental high energy physics is the efficiency ϵ, or fraction of observations from a specific process $pp \to X$ that are expected to satisfy a given condition that is a function of the final state details.

In collider experiments typically we cannot measure directly the properties of final states produced in the hard scattering, either because of the characteristics of the detector, the decay/hadronisation of particles producing other secondary particles, or due to additional physical effects occurring in a bunch crossing not accounted in Equation 1.32, such as additional collision products due to multiple interactions or processes comprising the proton remnants. Thus it is very useful in the construction of the complete model to consider the problem of generation of realistic collision products.

Taking into consideration that some of the computational techniques for including subsequent physical processes and the detailed simulation of the detectors are considerably resource intensive, as will be detailed in Section 1.3.4 and Section 2.3.2 respectively, the use of weighted samples is not a very efficient approach. Hence, for the generation of simulated products of high-energy collisions, also referred to as *event generation*, an acceptance-rejection sampling step is carried out to obtain an unweighted sample, where the relative frequency of each simulated outcome is expected to match its theoretical prediction. After such procedure, the calculation of all observables is also simplified, because the weight of all samples can be taken as a constant, e.g. a unitary weight $w = 1$, so the computation of quantities of interest such as efficiencies becomes trivial.

1.3.4 Hadronization and Parton Showers

In order to link the hard scattering process outcome with the actual observable quantities that can be detected in an experiment, it is necessary to account for the radiation of soft gluons or quarks form the initial or final state partons in the collision, as well as the formation of hadrons from any free parton due to colour

confinement (see Section 1.1.2). Additional processes that affect the collision outcome include secondary interactions between the protons, as well as the decays of all generated unstable particles. An example of the typical complexity of the physical processes occurring as a result of a single high-energy proton-proton scattering is provided in Figure 1.4. These and additional minor effects (e.g. colour reconnection) are accounted by *parton showering* (PS) programs, that take as the input the generated particle outcome of the hard scattering and return a set of the resulting stable particles that would propagate through the detector.

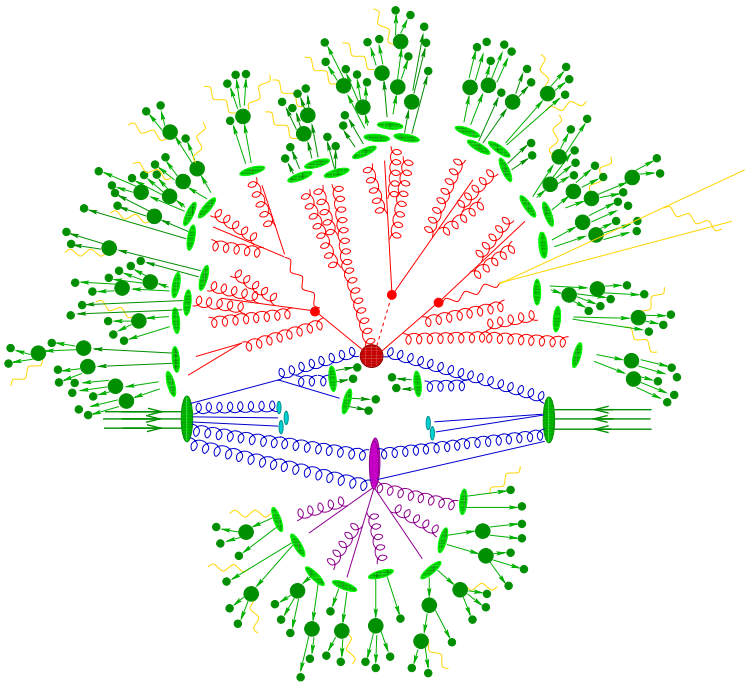


Figure 1.4: Diagram of a proton-proton collision and the underlying physical processes occurring therein, adapted from [52]. The dark green ellipses following the three parallel arrows represent the incoming hadrons. The main interaction between partons is shown in red colour, producing a tree-like structure of decays, in turn producing partons that rapidly transition to hadrons (light green ellipses) and decay (dark green circles) as well as soft photon radiation (yellow lines). The blue lines represent the interaction between partons and the path of the the initial hadron remnants followed by light blue ellipses. For completeness, an additional hard interaction within the same hadron-hadron process is shown in purple, which often has to be accounted to obtain realistic simulations.

2 Experiments at Particle Colliders

> Measure what is measurable
> and make measurable what is not so.
>
> Galileo Galilei (attributed)

In Chapter 1, we reviewed the most successful testable theory to date describing the properties and dynamics of our universe at the most fundamental scales. Clear limitations of the Standard Model as it is currently formulated are known, such as the complete omission of gravity forces or the absence of viable dark matter candidates, motivating the quest for alternative unified descriptions of the physical world. A direct path to verify the predictions of the Standard Model up to high accuracy and test alternative theoretical models is to collide high energy particles in a controlled setting and quantitatively study the properties of the particles produced as an outcome of the scattering. That is the aim of the Large Hadron Collider (LHC) and the experiments set up around its collision points. In this chapter, the main design characteristics of a general purpose high-energy physics experiment, namely the Compact Muon Solenoid (CMS) detector at the LHC, will be explored. Given the data-centric nature of the next chapters, particular significance will be given to the acquisition, processing and simulation of individual experimental observations, commonly referred to as events.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the largest and most powerful particle accelerator on operation at the time of writing. Its main purpose is to accelerate bunches of protons and other heavier nuclei in opposite directions to ultra-relativistic velocities, so they can be collimated and made interact at high energies in several specified collision points inside specially designed detectors. The LHC machine complex is located at the European Organisation for Nuclear Research (CERN) laboratories at the

Switzerland-France border near Geneva, its most distinctive element being a circular ring of superconductive magnets and accelerating structures installed inside a 26.7 km long underground tunnel inherited from the Large Electron Positron (LEP) collider, as depicted in Figure 2.1. The setup was designed to achieve center-of-mass energies up to 14 TeV for nominal instantaneous luminosities reaching $1 \times 10^{34}\text{cm}^{-1}\text{s}^{-1}$ for proton-proton collisions, and hence explore the high-energy frontier of particle physics, extending by a factor of seven the reach at the highest collision energy, formerly achieved by the Tevatron collider at Fermilab.

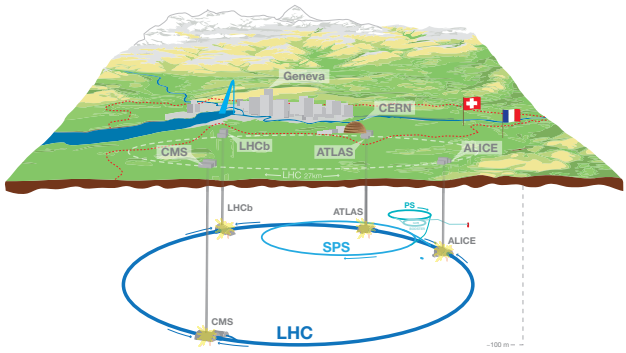


Figure 2.1: Depiction of the placement of LHC tunnel and the main experiments places at its collision points (ATLAS, ALICE, CMS and LHCb) relative to Geneva and the French-Swiss border. The image has been adapted from [53].

The main reason for building a high-energy proton-proton collider such as the LHC instead of an electron-positron more powerful than LEP, given the difficulties when computing observables due to protons being composite particles as described in Section 1.3, is that protons are considerably more massive and thus their synchrotron radiation loss is greatly reduced, so they can be accelerated to higher energies more efficiently. Another practical advantage of proton colliders is that very high collisions rates (i.e. instantaneous luminosities) are technically achievable, which makes them suitable for the discovery of rare but interesting physical processes. While the LHC and most of its detectors can also be used to study collisions of nuclei from heavier atoms, such as Pb, Au or Xe ions, which have important scientific use cases such as recreating the conditions present in the early universe, in this work we will be focussing on proton-proton collisions.

2.1.1 Injection and Acceleration Chain

In order to achieve beam energies of the TeV order, protons have to follow several stages of synchronised accelerations through a variety subcomponents of the CERN accelerator complex, whose main subcomponents as of 2018 are summarised in Figure 2.2. The purpose of this section is to outline the sequence of steps followed to obtain the high energy proton bunches that are used for high-energy collisions at the LHC.

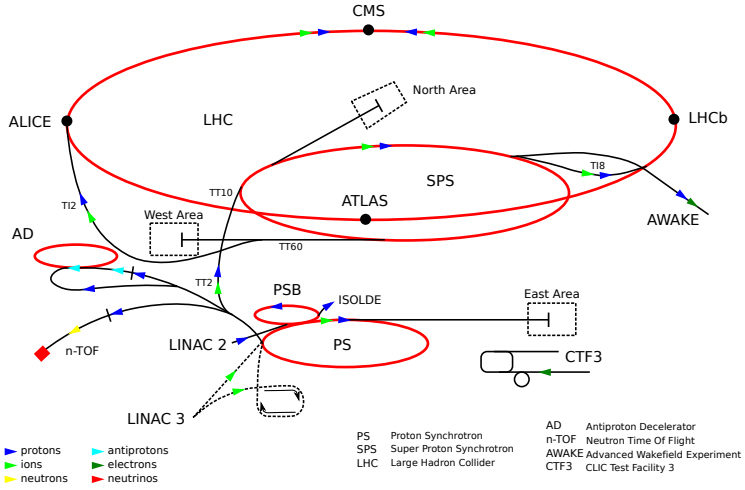


Figure 2.2: Schematic representation of the CERN Accerator Complex, including the the relative placement of the experiments as well as the main elements of the LHC accelerating chain: LINAC2, PSB, PS, SPS and the LHC ring. Figure credit to Forthommel(CC BY-SA 3.0 license).

The process begins with the extraction of a low-energy beam of protons by filling a duoplasmatron device [54] with gas from a hydrogen H_2 bottle. Those protons are then injected into to a linear accelerator, named LINAC2, which boosts them to an energy of 50 MeV. The next step of acceleration occurs at the Proton Synchrotron Booster (PSB), which receives beams split from the LINAC2 beam line and increases their energy to 1.4 GeV using four superimposed synchrotron rings. Promptly after, the Proton Synchrotron (PS) further splits and boosts the energy of proton bunches to 25 GeV. The penultimate step of the chain is the Super Proton Synchrotron (SPS) which accelerates the proton bunches to 450 GeV and injects them in opposite directions in the LHC ring.

The main LHC machine is composed by two adjacent proton beam lines (also referred as beam pipes) kept at an ultra-high vacuum ($10^{-10} - 10^{-11}$ mbar), in order to reduce the likelihood of spurious collisions of the highly-boosted hadrons with gas molecules. The proton trajectories are bent around the ring using a total of 1232

super-conducting dipole electromagnets, each 15 m long and kept at a temperature of 1.9 K using superfluid helium, capable of providing very strong magnetic fields (up to 8.3 T for a 11.8 kA current). For collimation of the proton bunches, 392 additional quadrupole magnets are placed around the ring. Higher-order multipoles are also interleaved to provide finer corrections of the beam direction and field geometry. Additional energy is provided to the protons in each revolution using 8 radio frequency (RF) cavities per beam line, until the protons reach the desired energy (6.5 TeV during the Run II of the LHC, which took place between 2015-2018). Given that each cavity can provide about 60 keV per revolution, it takes about 20 minutes of *ramp* time to reach collision energies.

During the whole acceleration process, specialised dipole magnets are used to keep the beams separated at the four interactions points (IPs) and hence avoid collisions during that time. With the purpose of maximising the interaction rates, the beams are made more compact (commonly referred as *squeezed*) at the interaction region right before switching to collision mode. Once the characteristics of the proton beams are suitable, the quadrupoles focus the beam trajectories and collisions begin. A stable configuration is then adopted by the LHC machine, providing about 7 keV of energy per turn to the beam to account for synchrotron radiation losses using the RF cavities. In the absence of problems, the proton beams are kept circling the LHC ring and colliding at the IPs for several hours until the bunch properties are degraded beyond correction, a period that typically is referred as a LHC *fill*. The *fill* is terminated when some problem occurs or when all the proton bunches inside the ring are *dumped* (made collide) against graphite absorbers tangent to the beam pipes.

2.1.2 Operation Parameters

One of the most relevant parameters for a particle collider is the instantaneous luminosity $\mathcal{L}_{\text{inst}}(t)$, which already appeared in Section 1.3 and corresponds to the number of particles per unit of area per unit of time crossing each other in the interaction volume. Given a certain physical process characterised by a cross section σ, the number of collisions n_c expected to occur by unit of time, also known as the rate of such collisions, can be expressed as:

$$\frac{dn_c}{dt} = \mathcal{L}(t) \cdot \sigma \tag{2.1}$$

thus the luminosity $\mathcal{L}$ is proportional to the number of expected interactions of any given process. For studing rare scattering processes, corresponding to very small cross sections σ, the luminosity is a crucial factor, because it determines the expected total amount such collisions produced per time unit. The instantaneous luminosity at the interaction region at a given time can be estimated from the characteristics of the proton beams as:

$$\mathcal{L}_{\text{inst}} = \frac{n_p^2 n_b f_r \gamma_r}{4\pi\epsilon_n \beta^*}\mathcal{F} \tag{2.2}$$

where n_p is the number of particles per bunch, n_b is the number of bunches per beam, f_r is the beam revolution frequency, γ_r is a relativistic suppression factor, ϵ_n is the normalised beam emittance, β^* is the transverse size of the beam, and $\mathcal{F}$ is an additional luminosity reduction factor. The main contribution to the reduction factor $\mathcal{F}$ comes from a small tilt of the beams at the crossing point, characterised by the crossing angle ϕ_c, which avoids parasitic interactions between bunches but reduces the luminosity by approximately:

$$\mathcal{F} = \left(1 + \left(\frac{\phi_c \sigma_z}{2\sigma^*}\right)^2\right)^{-1/2} \tag{2.3}$$

where σ_z is the root mean square (RMS) bunch length and σ^* is the RMS of the beam in the transverse direction at the interaction volume. The peak instantaneous luminosities per day for the different years of proton-proton data acquisition periods (also known as *runs*) at the LHC are summarised in Figure 2.3, those numbers can be compared with the peak design luminosity of the LHC of $\mathcal{L}_{\text{design}} = 10^{34}\ \text{cm}^{-2}\text{s}^{-1} = 10\ \text{Hz/nb}$.

From Equation 2.2 it can be inferred that that value of instantaneous luminosity varies between LHC fills depending on the beam parameters. In fact, it also varies within a single fill with time, mainly because the number of average protons per bunch n_p decreases due to the collisions at all the interaction points. For convenience, a quantity referred as integrated luminosity $\mathcal{L}_{\text{int}}$ that is computed by integrating over the instantaneous luminosity for a given time period $\Delta T = t_1 - t_0$ within a fill, is used:

$$\mathcal{L}_{\text{int}} = \int_{t_0}^{t_1} \mathcal{L}(t)dt \tag{2.4}$$

which is proportional to the number of collisions for a given process during that period and thus can be used to quantify the amount of data acquired. When studying data from different time periods jointly, integrated luminosity is additive, even if the

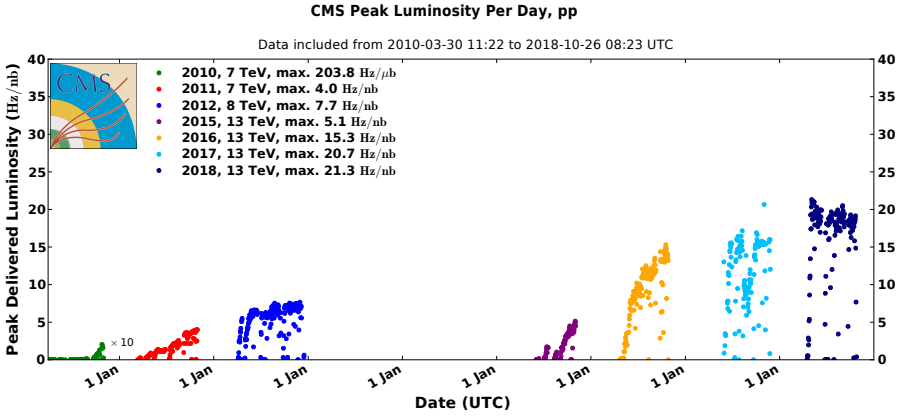


Figure 2.3: Peak luminosity per day as measured using the CMS detector for the all the proton-proton data-taking periods of the LHC to date. Figure from CMS Public Luminosity Results.

beam conditions (e.g. proton density) are different as long as the beam energies are matching. Such notion will be particularly useful when talking about the amount of data collected by a detector during a year or a longer data acquisition period.

2.1.3 Multiple Hadron Interactions

Given the high density of protons in each bunch at the collision points, every bunch crossing generates a few dozen proton-proton interactions, a phenomenon that is commonly referred to as *pileup*. The products of all these interactions go through the surrounding detectors at almost the same time, which complicates the interpretation of the detector readouts as the product of a single interaction. The number of proton-proton interactions for each crossing is effectively a random variable, however its expected value is proportional to the instantaneous luminosity and the total cross section of processes that produce detectable remnants in the detectors, mainly originating from low-energy inelastic proton scattering processes.

In fact, at the collision point of one of the general purpose detectors at the LHC, the most likely outcome of any given bunch crossing at the nominal design luminosity of $1 \times 10^{34} \mathrm{cm}^{-1}\mathrm{s}^{-1}$ is about 25 *soft* scattering interactions (i.e. ones characterised by a low momentum transfer), producing hundreds of low energy particles all around the collision region, as depicted in Figure 2.4. Quite rarely, given the small relative cross section of *hard* scattering processes in comparison with the total scattering cross section as discussed in Section 1.3, one of the produced interactions might involve a large momentum transfer between partons, which is characteristic of the funda-

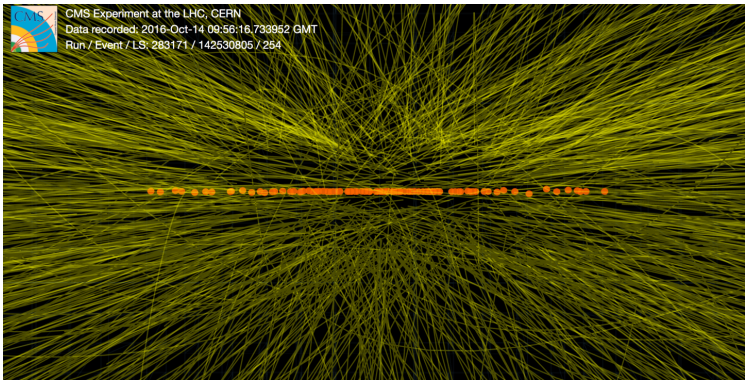


Figure 2.4: Multiple interactions in a single bunch crossing as recorded by the CMS detector during a special high-pile up luminosity at the end of 2016 [55]. The reconstructed primary interaction vertices are shown using orange circles while the yellow lines represent the trajectories of charged particles.

mental physical processes of special interest at the LHC, such as the production of a Higgs boson. The probability of two or more hard interactions happening in the same bunch crossing is really low, and can be safely neglected for any practical purposes. Nevertheless, the outcome of each hard interaction of interest will be overlapping in the detector volume with the product of all other soft interaction that occurred on the same bunch crossing, greatly complicating the task of *event reconstruction* as will be discussed in Section 2.3. This also motivates the use of *pileup mitigation* techniques, heavily based on accurate detectors that can extrapolate and differentiate the primary interaction vertices of the collisions from the charged particle trajectories.

In addition to multiple hadron interactions per bunch crossing, the goal of recording the outcome of a very high number of proton interactions leads to a different experimental complication. As illustrated in Equation 2.1, a simple way to increase the luminosity is to increase the number of total proton bunches per beam n_b. This fact is exploited in the nominal proton fill scheme of the LHC by having a total of 2808 proton bunches in each beam, corresponding to a separation between most of the bunches of only approximately 7.5 m. Hence the time separation between consecutive bunch crossing is about 25 ns, which is of the same order as the response time of many of the detector elements used at the LHC. The readout from a a particular bunch crossing can therefore be affected by the detector occupation caused by the previous or subsequent crossings, in what is referred to as *out-of-time pileup*, that becomes an important consideration for detector design in high-luminosity environments.

2.1.4 Experiments

Around the collision volume at each of the interaction points, large detectors are positioned in order to reveal and quantitatively study the outcomes of the highly-energetic particle scattering, which can in turn be used to obtain information about the properties of fundamental interactions. Four large particle experiments are installed at the LHC interaction points:

- **ATLAS** (A Toroidal LHC ApparatuS) [56]: the largest experiment at the LHC, designed as a general-purpose detector to study the various products of high-energy interactions, especially those of high-luminosity proton-proton collisions. While one of the most important scientific goals of the ATLAS experiment was to discover Higgs boson and provide a detailed study of its properties, it was also built with the aim of extensive testing of Beyond the Standard Model (BSM) theories.

- **CMS** (Compact Muon Solenoid) [57]: the other general-purpose experiment at the LHC, sharing most of the research goals with ATLAS, but opting for an alternative design and a different choice of detector technologies making it considerably more compact. It is the detector that collected the data use in the analysis in Chapter 5 and hence is described extensively in Section 2.2.

- **LHCb** (Large Hadron Collider beauty) [58]: operating at a lower range of luminosity than ATLAS or CMS by deliberately separating the beams, this experiment focusses on very accurate precision measurements of the properties and rate decays of b-quark and c-quark hadrons as well as the search for indirect evidence of new physics leading to CP violation in heavy flavour physics phenomena.

- **ALICE** (A Large Ion Collider Experiment) [59]: a heavy-ion collisions detector, designed to study the dynamics quark-gluon plasma, a high energy density state of strongly interacting matter, as it expands and cools down. Such studies can lead to a better understanding of colour confinement and other relevant QCD problems, as well as shedding some light on the processes that occurred a few microseconds after the Big Bang.

Additionally, three smaller experiments are built around the mentioned detectors with specific research purposes: TOTEM [60], LHCf [61] and MoEDAL [62]. Both TOTEM and LHCf have been designed to investigate features of forward physics

interactions, where scattering products remain the original proton trajectories, and hence they are set up tangent to the LHC beam line at the sides of CMS and ATLAS interactions points respectively. MoeDAL is instead built at the same experimental space than LHCb and its main aim is to search for evidence of production of magnetic monopoles and other highly ionising stable massive particles.

2.2 The Compact Muon Solenoid

The Compact Muon Solenoid (CMS) is a general purpose detector placed about 100 meters underground around one of the collision points of the Large Hadron Collider (LHC) ring. It has been designed to carry out experimental research on a wide range of high-energy physics phenomena, including searching for the Higgs boson and studying its properties, testing alternative explanations of nature such as extra dimensions or supersymmetry, and looking for evidence of direct production of particle dark matter candidates.

In spite of having such ambitious research goals, the principle of operation of CMS is rather simple, as it can be reduced to the detection of the outgoing particles produced as a result of high-energy interactions between protons and the identification and measurement of their most relevant properties, such as momenta and energies. These is done by putting together the information acquired by a large number of simple detecting elements, placed in layers around the collision region. The properties and kinematics of several of those final state detected particles can often be combined to compute observables of more complex objects, such as the invariant mass of an intermediate particle. After collecting data from a large number of collisions, a subset of relevance of the data can be compared with the expected theoretical predictions, and statistical inference in the form of interval estimates on parameters of interest or hypothesis testing of alternative explanations can be performed.

The CMS detector is built inside and around a large cylindrical coil of superconductive wire, forming a 6 m diameter solenoid magnet that can provide an homogenous magnetic field of 3.8 T. Particle detection and identification are achieved using several layers of sub-detectors with specialised functions, almost covering the full solid angle around the interaction region, as depicted in Figure 2.5. Inside the solenoid volume, a particle tracker made of silicon pixel and strip detectors, a lead tungstate crystal electromagnetic calorimeter (ECAL) and a brass-scintillator hadronic calorimeter (HCAL) are placed, each of them composed of a barrel and two endcap sections. A large muon detection system, composed of cathode strip chambers (CSC), resistive

plate chambers (RPC) and drift tubes (DT), is embedded in the steel flux-return yoke outside the solenoid. Furthermore, extensive forward calorimetry complements the coverage provided by the barrel and endcap sections. A more detailed review of the detection principles and capabilities for each detector component are included in the following sections, yet the detector performance technical design report [63] and references therein are recommended for a more comprehensive account.

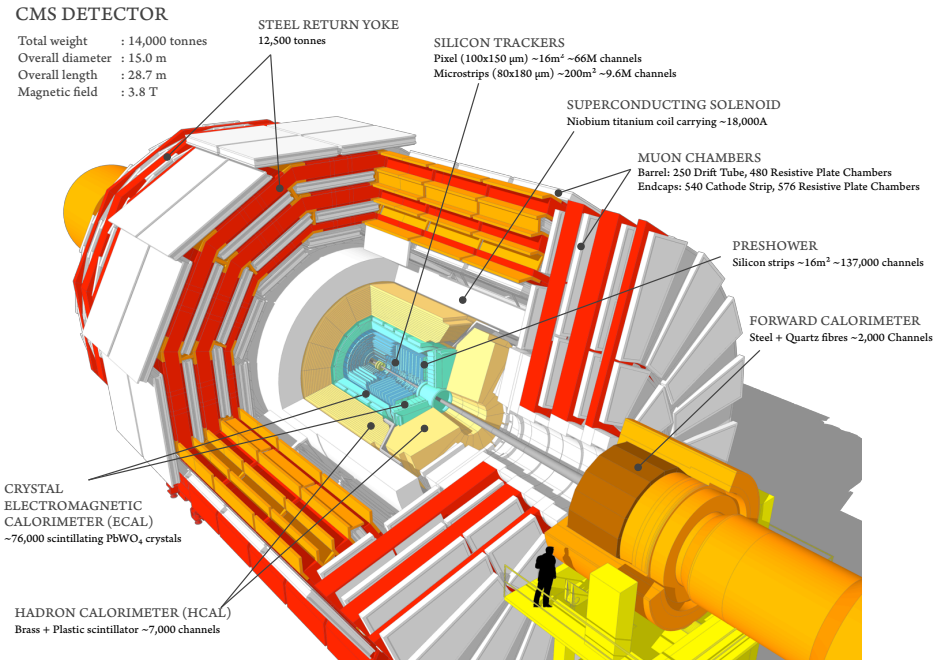


Figure 2.5: Cutaway view of the CMS detector, based on a three-dimensional representation, an highlighting the main detecting systems and characteristics. The image has been adapted from [64].

2.2.1 Experimental Geometry

Given the geometry of the detector, the coordinate system used is centred at the nominal interaction point inside the detector. The x axis point inwards towards the LHC ring origin, while the y axis points vertically upward toward the terrestrial surface. The z axis is thus tangent to the beam line, increasing in the counter-clockwise direction when looking at the LHC ring from above. Considering the expected symmetries for particle production, spherical coordinates are a convenient representation, where ϕ is the angle from the x axis in transverse plane (i.e. x-y

plane), and θ is the polar angle with respect to the LHC plane using a sign convention consistent with the previous definition of the z and y axes.

As mentioned before, particle momentum is the main observable of the detected particles. The energy is simply a function of the momentum and the mass of the particle, as shown by the relation $E^2 = p^2 + m^2$, expressed in natural units ($c = 1$). Because the x and y momentum components are insensitive to the initial state boost in the z direction due to the stochastic differences in parton momenta in the initial state, and are measured more accurately as a result of the design of the detector, it is common to refer separately to the total transverse momentum quantity $p_T = \sqrt{p_x^2 + p_y^2} = |p| \sin\theta$ and its transverse plane angle ϕ. While the z component of the momentum could be specified directly either by using p_z or by the angle θ, the differences of any of those observables between two particles detected on an event depend on the initial parton state boost β on the z direction, which varies between different collisions and it is hard to estimate precisely in the laboratory frame of reference.

Since the dependence on the initial state z boost would complicate the statistical analysis and the definition of derived observables, an alternative observable is used. The rapidity y is defined as:

$$y = \frac{1}{2} \ln\left(\frac{E + p_z}{E - p_z}\right) \tag{2.5}$$

and its value under a z-axis boost β is easily obtained by adding an additive factor $y' = y - \tanh^{-1}\beta$. Hence differences in rapidity between two particles in a collision $\Delta y = |y_b - y_a|$ are invariant to Lorentz boost in the z direction. Because the rapidity depends on the total energy/momentum of the particle, which might not be possible to measure with high precision in hadron collider detectors, it is more suitable to approximate it. The approximation is referred to as the *pseudo-rapidity* η, and can be defined as:

$$\eta = \frac{1}{2} \ln\left(\frac{p + p_z}{E - p_z}\right) = \ln\left(\tan\frac{\theta}{2}\right) \tag{2.6}$$

that only depends on the polar angle θ with respect to the LHC plane. The *pseudo-rapidity* η is equal to the rapidity y for massless particles, and is a very effective approximation in the highly-relativistic limit, when $E \gg m$. It is useful observing that for particles produced in the transverse plane (i.e. $\theta = \pi/2$), their *pseudo-rapidity* is $\eta = 0$. Instead, in the limit of fully forward particles, when $\theta \to 0$ or $\theta \to \pi$, their *pseudo-rapidity* becomes $\eta \to +\infty$ and $\eta \to -\infty$, respectively.

Oftentimes, angular distances between two particles are very useful observables in an collision to cluster observed particles or isolate interesting collisions. The distances between two particles, indentified with a and b subindexes, in the transverse $\Delta\phi$ and forward direction $\Delta\eta$ can be computed as:

$$\Delta\phi = \min(|\phi_b - \phi_a|, 2\pi - |\phi_b - \phi_a|) \quad \text{and} \quad \Delta\eta = |\eta_b - \eta_a| \tag{2.7}$$

while the total angular distance ΔR between the two particles is instead defined as:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{2.8}$$

which is invariant to boosts in the z direction in the highly-relativistic limit, and is particularly practical to cluster the products of the hadronization of quarks and gluons as detailed in Section 2.3.

2.2.2 Magnet

The purpose of the CMS magnet is to curve the trajectories of charged particles coming out the interaction region, so their transverse momenta p_T can be accurately estimated, and the sign of their charge determined. In order to understand how such momentum measurement can be carried out, let us assume a solenoidal magnetic field that is fully homogenous and pointing in the z direction $\vec{B} = B\hat{z}$. Due to Lorentz force, a particle with a transverse momentum p_T and a forward momentum p_z would describe an helicoidal trajectory, where the curvature radius in the transverse plane r_T and the transverse momentum are related:

$$r_T = \frac{p_T}{qB} \quad \Longrightarrow \quad p_T[\text{GeV/c}] = 0.3 \cdot q[\text{e}] \cdot B[\text{T}] \cdot r_T[\text{m}] \tag{2.9}$$

where q is the particle charge, and the second equation corresponds to a simplification using units denoted inside the brackets adjacent to each quantity (e are electron charge units). This simple proportionality relation indicates that the higher the momentum of a particle, the larger its radius of curvature. Furthermore, the direction of the curvature is determined by the sign of the particle charge. For more realistic scenarios, like the magnetic field not being completely homogenous or the particle momentum decreasing along its trajectory due to interaction with the detecting elements, Equation 2.9 is only an approximation and the trajectory path can be obtained by solving the relevant differential equation.

In the case of CMS, the magnetic field is generated by a large superconducting solenoid, contained inside a hollow cylinder about 13 m long and with an outer radius of 3 m. Very high currents, up to 19 kA, circulate along NbTi wires kept at 4.5 K using a liquid helium cooling system, providing an almost homogenous field at the centre of the solenoid up to 3.8 T in the z direction. In addition to the solenoid, the magnetic flux lines are closed by a 10000 ton return yoke, composed by a series of magnetised iron blocks interleaved with the muon detectors in the outer part of CMS, providing a magnetic field about 2T in the opposite direction. The remaining elements of the CMS magnetic spectrometer, referring to the detector systems used to estimate the curved particle trajectories are reviewed in Sections 2.2.3 and 2.2.6.

2.2.3 Tracking System

The inner tracking system is the detector that is the closest to the interaction point, and its functions include the estimation of the charged particles trajectories, used to provide a measurement of their momenta as described in Section 2.2.2, as well as allowing the positional determination of interaction or decay vertices by extrapolating the trajectories inside the interaction region. The detection of charged particle trajectories, or *tracks* for short, is carried out by several silicon detector layers placed non-uniformly around the collision volume, as shown in Figure 2.6. The placement of layers is symmetric in ϕ, the outermost layers contained within a supporting cylindrical structure of 2.5 m of diameter and 5.8 m of length.

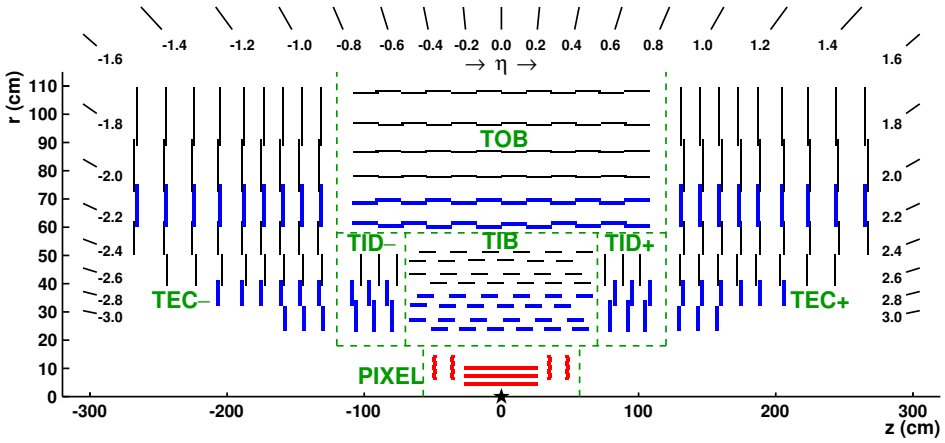


Figure 2.6: Cross sectional view of the CMS detector inner tracker detector in the $r-z$ plane, detailing the position of detecting layers as well as the main detector subcomponents. The tracker is approximately symmetric around $r = 0$, so only the top half is shown. Figure has been adapted from [65].

The detector is composed of two main parts: a silicon pixel detector system situated very close to the interaction point and a much larger strip detector arrangement placed outside the former. The disposition on the detecting layers allows to detect tracks within a pseudo-rapidity range defined by $|\eta| < 2.5$. Both systems have to deal with the efficient tracking of hundred of charged particles, at a rate of 40 MHz, produced from each bunch crossing. A successful apparatus in such a environment requires a short response time, as well as to be composed of many small detecting elements. The latter property is commonly referred as *high granularity*, and allows to keep the number of detected track points (i.e. *hits*) per detector unit at acceptable levels.

Being so close to the collision region, the set-up has also to sustain very high particle fluxes during long periods of time, up to $1\text{MHz}/\text{mm}^2$ at the first pixel layer. Therefore, resistance to radiation damage of the detecting elements and the accompanying electronics, dubbed as *radiation-hardness*, is an essential specification. Additionally, the amount of material present in the particle trajectories has to be kept to a minimum, to avoid stochastic secondary interactions that would degrade the precision and efficiency of track determination. The use of silicon semiconductor detector technologies [66] in the CMS tracking system is thus motivated by a combination of all previously mentioned reasons. In total, the CMS tracking system is composed of 1440 pixel detector modules and 15148 strip detector modules, accounting for an active area over 200m^2.

The pixel detector, the innermost detecting system of the CMS experiment, is comprised by a total of 66 million silicon cells placed in 1440 modules around the collision region. Each pixel cell has an area of $100\times150\mu\text{m}^2$ and a thickness of 285μm, providing two-dimensional local track hit coordinates with a resolution around in the cell surface plane about $20\,\mu$m, that can in turn be used to compute the global three-dimensional hit location with high accuracy after accounting for the precise location of the detecting module. As depicted in Figure 2.6, the pixel detector is composed by three *barrel* layers (i.e. placed around the collision region in an cylindrical arrangement), located at radii of 4.4 cm, 7.3 cm and 10.2 cm respectively, and two forward disks at each side at distance of 34.5 cm and 46.6 cm from the nominal interaction point.

The rest of the tracking system, placed outside the pixel detector, is constituted of several silicon strip detector modules organised in four different sub-detectors, referred as TIB, TID, TOB and TEC in Figure 2.6. The inner part of the strip tracker, adjacent to the pixel detector, is composed of four barrel layers of strip

modules constituting the tracker inner barrel (TIB) section, and three module layers arranged in disks at at each side forming the tracker inner disk (TID). Further away from the interaction region, the outer strip tracker, comprising of six barrel layers in the tracker outer barrel (TOB) and nine disks at each side forming the tracker endcaps (TEC). The strip specifications varies depending on the sub-detector, with thicknesses ranging from 320μm to 500μm, and pitches (i.e. distances between strips) from 80 μm to 184 μm.

The strips are placed longitudinally parallel to the beam line in the barrel modules and radially in the perpendicular plane in the endcap disks, with silicon strip lengths ranging from 10 cm to 20 cm, and in an overlapping tiled setting (see Figure 2.6) Each strip layer provides a single local coordinate for a particle track hit, aligned with ϕ both the barrel and the endcap disk. A second coordinate can be easily obtained taking into account the placement on the module, thus obtaining the r coordinate in the barrel and z in the endcap disks. In order to provide information on the unknown coordinate in each case, some layers of the tracker (in blue colour in Figure 2.6) are composed of two modules instead on one, with a small tilt of 0.1 rad that allows to obtain a precise 3D coordinate for a track hit by combining the two local coordinates and their module positions.

2.2.4 Electromagnetic Calorimeter

The function of the CMS Electronic Calorimeter (ECAL) is to measure the total energy of the electrons, positrons and photons that reach that part of the detector, by means of their *electromagnetic showers*. In order attain such task, scintillating lead tungstate $PbWO_4$ transparent crystals are placed inside the solenoid magnet, right outside the tracking system, covering the solid angle around the interaction point as depicted in Figure 2.7. When a high energy electron or a positron enters the dense crystal material, it rapidly decelerates and emits photons through bremsstrahlung radiation. High energy photons from electron/positron deceleration or directly coming from the collision region produce positron-electron pairs through matter interaction, that in turn radiate more photon through bremsstrahlung processes. These chain of processes, referred as *electromagnetic shower* keeps occurring until the energy of the photons goes below the pair production threshold or the energy loss of the electrons/positrons happens through alternative mechanisms. The resulting low energy photons from the electromagnetic shower produce visible range light in the scintillating but transparent crystal, which is detected, amplified and collected by photodetectors placed at the end of each lead tungstate crystal.

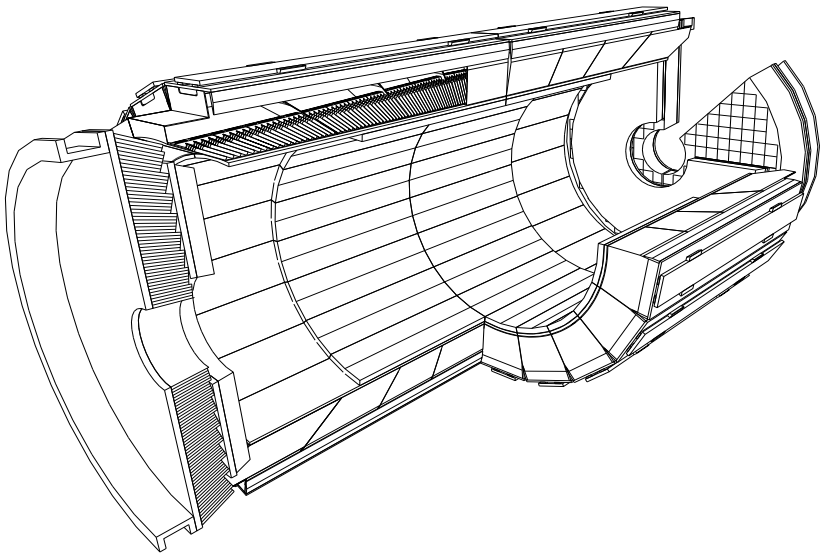


Figure 2.7: Cutaway view of the CMS electromagnetic calorimeter, based on a tree-dimensional model of the detector geometry. The placement of the lead tungstate crystal is shown for part of the barrel and endcaps. The figure has been adapted from [67].

The ECAL is composed of two main parts, the barrel calorimeter (EB) section covering pseudo-rapidities up to $|\eta| < 1.479$, and two symmetrically positioned endcap calorimeters (EE) further extending the coverage to $|\eta| < 3.0$. The trapezoid-shaped crystals are placed radially around the collision region, a total of 61200 blocks in the EB and another 7324 blocks for each EE part. The sides facing the IP in the barrel section have dimensions of 22×22 mm^2 and a length of 23 cm, while the front-facing sides of those in the endcaps are slightly larger at 28.6×28.6 mm^2 with a length of 22 cm. The advantages of using lead tungstate crystal include its very short radiation length $\mathcal{X}_0 = 0.89$cm - which characterises the longitudinal energy loss profile $E(E) = E_0 e^{x/\mathcal{X}_0}$ - as well as its small Moliere radius of 2.19 cm - defining the radius containing average transversal radius containing 90% of the shower energy - leading to narrow showers which contribute to improved position and energy resolution. The lengths of the crystal blocks in the EB and EE amount to $25.8\mathcal{X}_0$ and $24.7\mathcal{X}_0$, which ensures that all the energy is effectively deposited inside the detectors.

Another advantage of lead tungstate crystals is that $PbWO_4$ is also a scintillating material, thus the resulting shower energy is absorbed and partially emitted back as visible light, with a yield spectrum maximum in the blue-violet range around 430 nm. The reemission process is also very fast, since about 80% of the scintillating light is emitted within 25 ns of absorption, which is the time until the next LHC bunch crossing occurs. The scintillator light propagates effectively through the crystal due to its high transparency, and reaches the photodetectors attached to the end of the crystal trapezoids. Avalanche photodiodes (APD) are used for light detection and amplification at the barrel crystals while vacuum phototriodes (VPT) are used for the endcaps, given their different radiation hardness and sensitivity to magnetic fields.

In addition to the EE and EB, a sampling detector referred as pre-shower electromagnetic calorimeter, based on two layers of lead absorber followed by two layers of silicon strip detectors, is placed right before the lead tungstate crystals in the endcap to provide higher granularity in the forward region. The main purpose of the pre-shower extension is to distinguish high-energy photons coming directly from the collision region and high energy neutral pions that have decayed into two closely-spaced photons.

2.2.5 Hadronic Calorimeter

The purpose of the hadron calorimeter (HCAL) is to measure the energy and position of all long-lived neutral or charged mesons and baryons produced as a result of the collision, typically including pions, kaons, protons and neutrons. The main detecting

elements of this sub-detector are an assortment of sampling calorimeters, interleaving brass plates as absorber material and plastic scintillator tiles as active medium; the former causing the deposition of energy in the form of secondary particles by means of interactions with the material nuclei and the latter converting a part of that energy to visible light. The light from each tile is captured by a thin optical fibre and carried to a photodetector, producing an electric signal that can be used to measure the total amount of deposited energy once it has been carefully calibrated.

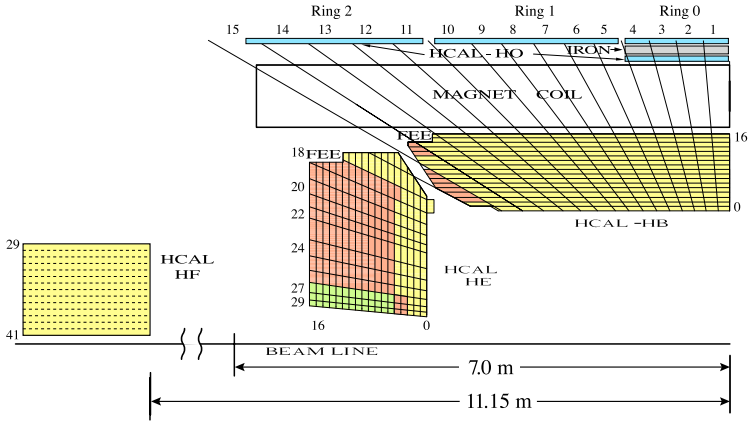


Figure 2.8: Cross sectional view of the CMS hadronic calorimeter (HCAL) detector in the $r-z$ plane, depicting the positioning of the various detector segments relative to the beam line and the solenoid magnet. The HCAL is symmetric around $r = 0$, so only the top half is shown. The figure adapted from [68].

The different segments of the CMS HCAL are shown in Figure 2.8. The barrel section of the hadronic calorimeter (HB) as well as two endcap sections (HE) at each side are placed after the ECAL but still inside the solenoid volume, providing pseudo-rapidity coverages of $|\eta| < 1.3$ and $1.3 < |\eta| < 3.0$, respectively. Both the HB and HE sections are composed of a stack of brass plates with plastic scintillator tiles in between, providing a total of $5.6\lambda_I$ at $\eta = 0$ and $11.8\lambda_I$ at $\eta = 3$, where λ_I is the hadronic interaction length. Given the limited space inside the solenoid and the fact that about $11\lambda_I$ are required to absorb about 99% of the total energy of the hadrons at the expected energy ranges, the hadronic calorimeter system is complemented by an outer detector (HO) outside of the solenoid. The HO is composed of five rings of scintillator tiles, effectively using the solenoid material as absorbing material. Because the absorbing material path length is shorter around $\eta = 0$, the central ring is shielded by large iron plates and an additional layer of scintillating material, yielding a total absorber length over $11.8\lambda_I$ and therefore improving its measuring capabilities.

Over 70000 thin plastic scintillator tiles are placed between and after absorber plates. The size of those plates depends on their geometrical placement and are aligned according to their angular coordinates between layers, so each longitudinal projection corresponds to an approximate area $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ within the HB coverage region and $\Delta\eta \times \Delta\phi = 0.17 \times 0.17$ outside it. When secondary particles go through the scintillating tiles, part of the energy is absorbed and promptly released as violet-blue visible light, over 65% of the total amount of emitted light within 25 ns. The light is collected and guided through thin optical wavelength-shifting fibres that change the light to the green spectrum region, then through standard optical fibres until reaching readout boxes that contain hybrid photodiodes (HPD). The optical signal for each alignment of tiles are added optically to a single readout for most of the radial projections, with the exception of those in the intersections between the barrel and endcaps, that are kept in two or three separate channels in order to ease calibration procedures.

The last element in the HCAL system is the forward hadronic calorimeter (HF), situated 11.15 m at each side of the interaction point, adjacent to the beam pipe, and providing detection capabilities for particles with pseudo-rapidities in the range $3.0 < |\eta| < 5.2$. The HF greatly increases the pseudo-rapidity energy measurement for charged and neutral particles, allowing a near hermetic (full solid angle) coverage, and hence allowing the estimation of missing energy in the event such that corresponding to neutrinos leaving CMS undetected, as will be discussed in Section 2.3. Because the radiation fluxes are extremely high in the forward region and there are no depth constraints, a different detector design is used, based on 165 cm of steel absorber plates and quartz fibres aligned of the z-axis, each with an effective detecting area of $\Delta\eta \times \Delta\phi = 0.17 \times 0.17$.

The fibres running along the HF detect and guide the Cherenkov light of the charged secondary particles produced in the showers to photomultipliers tubes (PMT) placed behind a 40 cm thick steel and polyethylene shield. In this pseudo-rapidity range, the HF serves also as an electromagnetic calorimeter. The HF detector has been designed in a specific way to disentangle the energy contributions from electromagnetic and hadronic showers, which is useful for many physics data analyses use cases. Only half of the fibres start close to the face of the absorber plates closest to the IP, the rest starting at a depth of 22 cm. By comparing the readouts from the long and short fibres the type of shower can be inferred, given that electromagnetic showers are much shorter than hadronic showers.

2.2.6 Muon System

The scientific objective of the CMS muon sub-system, or outer tracker, is to identify, determine the charge and measure the momenta of high energy muons, which are the only type of charged particles capable of passing through all the other detector systems without a significant energy loss. While their trajectories can be detected in the inner tracker, the amount of energy loss due to bremsstrahlung is much smaller than those of electrons or positrons due to its much heavier mass (given that the emission probability scales with $1/m^2$) and hence the do not deposit a significant fraction of their energy in the ECAL or the HCAL. The simplest way then to augment the amount of information about muons obtained from the tracker is to place additional tracking detectors outside the solenoid, while sustaining a high magnetic field that can curve the muon trajectories by using large blocks of ferromagnetic material as *flux-return yokes*.

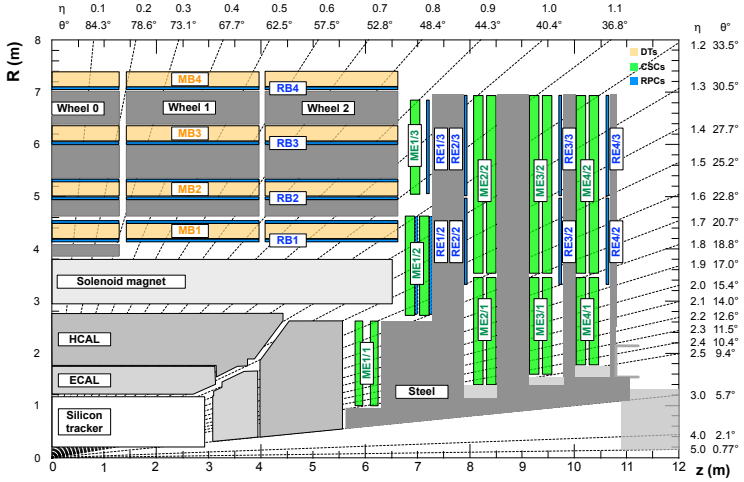


Figure 2.9: Cross sectional view of the layout of CMS detector in the $r-z$ plane, focussing on the components of muon system. The detector is symmetric around $r = 0$ axis and the $z = 0$ plane, so only the top quarter is shown. Figure adapted from [69].

The muon system is the most external sub-detector of CMS and it is based on gaseous tracking detector technologies, given the enormous volumes covered. The principle of action of gaseous detectors is rather simple: charged particles passing through the gas ionise gas molecules in their path, which start moving due to a high electric field between conducing wires, producing an electrical signal that can be read out. The time dependence of the signal on the different readout wires is used to infer the particle trajectory with high precision, and in some cases built-in

signal amplification can be achieved due to secondary ionisation by the choice of a gas mixture combined with high electric field gradients.

An overview of the various detectors of the muons system and their geometrical placement around the solenoid magnet cylinder is depicted in Figure 2.9. Due to a combination of criteria regarding uniformity and strength of the magnetic field, expected radiation fluxes and signal readout times, three different types of gaseous detectors are used: drift tubes (DT), cathode strip chambers (CSC) and resistive place chambers (RPC). In the barrel section where the particle flux is not expected to be very high, four layers of drift tubes (DT) are arranged cylindrically around the solenoid magnet, covering a pseudo-rapidity range $|\eta| < 1.2$. On the endcap section instead, due to higher radiation fluxes and magnetic field non-uniformity, multi-wire cathode strip chambers (CSC) are used, with a detecting pseudo-rapidity coverage of $0.9 < |\eta| < 2.4$. Both DT and CSC detectors can achieve very high position resolution, but their signal readout time and time resolution is not as good, thus a series of fast resistive plate chambers (RPC) are positioned both in the barrel and the endcap sections, up to pseudo-rapidities $|\eta| < 1.6$.

2.2.7 Trigger and Data Acquisition

As discussed in Section 1.3, the occurrence of relevant processes that may provide information about the physical properties of fundamental interactions in proton-proton collisions is purely stochastic given some initial conditions, plus their relative frequency is very rare compared with known phenomena. In order increase the expected chances of recording interesting phenomena, the LHC collides 40 million high-density proton bunches every second inside the CMS detector. Furthermore, as discussed in Section 2.1.3, tens of proton-proton interactions typically happen within each bunch crossing. The CMS sub-systems are hence detecting a good fraction of 100s of particles produced as a result of the interactions at each bunch crossing, in addition of being subjected to instrumental noise or external radiation sources such as cosmic rays.

The combined readout of all sub-detectors each 25 ns amounts to a large data size, due to the total number of sub-system channels, even if efficient techniques for representation and compression of information are used. Given that technical limitations on the amount of data that can be recorded exist, a practical choice for data acquisition is to keep only the detailed detector information of collisions that could be maximally useful to study the properties of fundamental interactions in subsequent data analyses. The decision system that makes the choice of whether to

record or filter out the detailed detector readouts for a given collision, is commonly referred as *trigger*, and is based on a fast and possibly asynchronous analysis of those readouts. In particular, such decision criteria is typically focussed on the most relevant properties of one or a subset of detected particles, such as their type, charge or the magnitude and direction of their momenta.

A flexible and sparse representation of all CMS detector readouts for a given collision that keeps sufficient information for detailed analyses is of the order of a few megabytes (i.e. $\mathcal{O}(1)$ MB). Because of the technical capabilities of the storage system, the total data acquisition rate is limited to less that 10 Gb/s, hence the trigger system has to reduce the rate of collision readouts from 40 MHz to about 1 kHz. As a compromise between processing speed and requirement adaptability, the trigger system of CMS is divided in two stages: the level 1 trigger (L1), which is a custom-hardware based solution that reduces the detector readout rate to 100 kHz, and the high-level trigger (HLT), a second step reducing it to the required 1 kHz and that is instead carried out by a commodity computer farm.

2.3 Event Simulation and Reconstruction

The raw account of the readout of all detectors after a single bunch crossing, as well as any derived representation of it, is commonly referred to as *an event*, and is the most fundamental type of observation in high-energy data analyses. All approaches to extract useful conclusions from CMS data are based on this information unit or simplifications thereof. This is because for practical purposes, statistical independence between events can be assumed, barring possible caveats (e.g. out-of-time pile-up or detector malfunctioning). Therefore, data analyses are reduced to the task of comparison between the observations and the predicted frequencies of events with different characteristics.

The dimensionality of an event evidently depends on its data representation, simpler representations being lower-dimensional and easing the comparison with theoretical predictions, at the cost of possibly losing some useful information. A principled way to obtain lower dimensional representations of an event given its raw detector readouts is to attempt to reconstruct all the primary particles that were produced in the main proton-proton interaction of the collision and estimate their main properties, through a process generally referred as *event reconstruction*. Nevertheless, for carrying out successfully the aforesaid task it is convenient to be able to have a detailed model of the detector readout output expected for a given set of parti-

cles produced in a collision. Realistic modelling of high-energy physics collisions in high-dimensional representations can be achieved through simulation.

In this section, a generative view of the main physical mechanisms that are happening both in the proton-proton collisions and when particles propagate through the CMS detector is first discussed. Such overview doubles as an introduction of the next section, where a description of how realistic simulations of the detector readouts (i.e. events) can be obtained using computational tools is provided. Afterwards, the inverse process is tackled, which is considerably harder and often ill-defined, namely how can we estimate the set of primary particles that were produced in the collision given the detector readout, through event reconstruction techniques.

2.3.1 A Generative View

When two high-density proton bunches travelling in opposite directions pass through each other inside the collision region of CMS, several proton-proton interactions can occur as discussed in Section 2.1.3. While most of the interactions will correspond to a small energy transfer between the interacting partons, given that the total interaction cross section is heavily dominated by soft scattering processes, a small fraction of collisions would include physically interesting process such as the production of heavy particles (e.g. a Higgs boson). The absolute and differential rates for such *hard* processes can be predicted as outlined in Section 1.3.3. Therefore, for a specific process in a proton-proton interaction, realistic high-dimensional modelling of the intermediate particles can be obtained by repeated sampling of the parton distribution functions and phase space differential cross sections. Subsequent decay, hadronization and radiation processes as well as more subtle effects and higher order corrections, can be then accounted for using the methods mentioned in Section 1.3.4, generally referred to as *Monte Carlo event generation* techniques. The end result of the mentioned procedures is a large dataset of simulated particle outcomes for a specific process, each example including a set of stable or sufficiently long-lived particles and their kinematics properties that would propagate through the detector.

In addition to the set of particles in the hard proton-proton interaction, the effect of pileup interactions can be accounted for by adding the particle outcome of a random number of randomly sampled soft interactions matching their approximately expected distribution in the collisions given the instantaneous luminosity conditions. This final set of long-lived particles produced in the interaction region represents a possible particle outcome for a collision assuming a given *hard* process occurred. While they cannot be directly observed, but only indirectly inferred through the

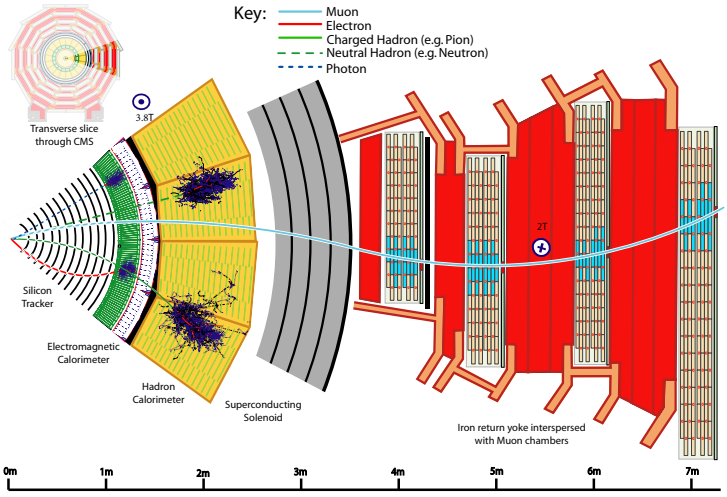


Figure 2.10: Transverse view of a section of the CMS detector and the interactions of the various particle types with the detecting sub-components. The figure has been adapted from [70].

detector readouts, it is assumed that an analogous set of particles is produced as result of each collision in the actual experiment. Based on the expected readout that they produce in the different CMS detector subcomponents, five main types of detectable particles are distinguished: muons, electrons, charged hadrons, neutral hadrons and photons.

The traces that each of the mentioned particle types leave in each detector subsystem are depicted in Figure 2.10. Even though muons are unstable particles, their long mean lifetime $\tau_\mu = 2.2\mu s$ allows them to travel very large distances when highly boosted, as is the case for all the high-energy muons coming out from the interaction region. Hence, for the purposes of studying LHC collisions they can be considered stable, given the unlikeliness of their decay in the detector volume at the range of energies studied. Because muons are charged particles, they leave hits in detector layers of the inner tracker following their curved trajectories However, due to their high mass, energy loss due to bremsstrahlung is not high enough to produce significant EM showering in the ECAL. After passing through the HCAL without interacting notably, muons reach the outer tracking system providing additional trajectory points.

The trajectories of high-energy electrons are also recorded by the CMS inner tracker, but as mentioned in Section 2.2.4, their interactions differ from those caused by muons because electrons lose energy rapidly due to bremsstrahlung when they

reach the ECAL, producing subsequent electromagnetic showers. It is worth noting that within CMS reconstruction and analysis, it is common to simply use the term *electron* to refer both to electrons and positrons, their charge inferred from the curvature sign of their trajectories. Charged hadrons, the term here largely referring to charged pions, kaons and protons, behave similarly to electrons in the tracking detector[1] but instead generate much larger hadronic showers in the hadronic calorimeter.

Long-lived neutral hadrons, including neutrons and the neutral kaon K_L^0, follow instead straight lines in the inner detector volume because they are not affected by the magnetic field neither leave any traces when passing through the tracking detectors. It is not until neutral hadrons reach the calorimeter detectors, chiefly the HCAL, that nuclear interactions produce large hadronic showers producing measurable signals that can be correlated with the energy deposited. Photons are massless and neutral particles, and at the ranges of energies of interest characteristic of the outcome of particle collisions at the LHC are not expected to deposit enough energy in the thin inner tracking layers to produce significant signal, thus they also follow a straight line trajectory to the calorimetry sub-systems. In contrast, when photons reach the electromagnetic calorimeter, electron-positron pair-production processes are bound to occur, producing in turn electromagnetic showers which can be readout as a ECAL detector signal.

The previous classification of particles based on their detectable energy remnants in the different detectors, patently disregards a common outcome of high energy collisions: neutrinos. Neutrinos only interact via weak and gravitational forces, hence the probability of interaction with the detecting elements of CMS is negligible. They thus escape the experimental area undiscovered. The production of high-energy neutrinos, or other weakly-interacting unknown hypothetical particles (e.g. dark matter candidates), can nevertheless be inferred by the total transverse energy imbalance. While the initial longitudinal momentum in the laboratory frame is unknown due to the proton compositeness, the initial total transverse momentum is very close to zero given that the collisions occur head-on. Because detecting structures of CMS have a near complete angular coverage around the interaction points, with the exception of very low transverse momentum particles that are lost near the beam pipe, the total transverse collision momentum of all detectable particles can be obtained simply by summing the estimation of their transverse momenta estimation. Ergo, the quantity $E_T^{\text{miss}} = || - \sum \vec{p}_T ||$ is referred to as total missing transverse energy or by

[1] Tracks from electrons and positrons are different due to bremsstrahlung, the radiated photons often recovered in the ECAL.

the acronym MET, and can be used to infer the production of non-detected particles particles such as neutrinos.

In summary, the physical characteristics of each category of particle previously stated cause different signatures in the various detector sub-systems, that often can be used to distinguish between each type. It is also worth pointing out the main attributes each individual detector element readout, which are principally the angular position in η and ϕ, the distance to the interaction point which is given by the detecting element placement or the z coordinate, and the amount of deposited energy. The latter is especially relevant for calorimeter detecting units. The precision of the angular location coordinates greatly varies between different detector types depending in their granularity, tracking detectors providing more accurate position measurements given that they extract information directly from the particle trajectories.

2.3.2 Detector Simulation

While the simplified map between the particle outcome of a given collision and the corresponding detector readouts presented in the previous section is extremely useful for obtaining a general understanding the operation of the CMS detector, it is not detailed enough to realistically model the detector readouts given a set of particles generated in a collision. Most of the relevant dynamics for modelling, such as interactions between protons, the produced particles and the detector material or the detector response, are of stochastic nature, hence they have to be specified either by sampling approximated probability distributions or by a complex probabilistic program that goes through a mechanistic simulation of the underlying physical processes actually occurring.

A detailed simulation is found to be the most accurate approach, given the many subtleties affecting the detector readout for a given set of generated particles, including various possible particle decays and material interactions that can occur when the particle is travelling through the detector, the non-uniformity of the magnetic field and its effect on the particle trajectories, and the intricacy of the detector geometry and the electrical response of its components. All these effects can be accounted for, to a high degree of validity, in a simulator program considering the non-deterministic propagation of the particles produce through the detector volume. The propagation of each particle through magnetic and electric fields can often be treated independently though a stochastic chain of time steps, that can an any point branch out to produce new particles through decays and other secondary particle generating

physical processes, so local energy deposits in the different detector structures can be recorded. After propagating all particles, the combination of all energy deposits in the detecting volumes can be used to produce realistic detector responses.

Such type of detector simulation is referred to as *full simulation*, or *fullsim* for short, and it is carried out for CMS generated events using a custom implementation of the geometry, properties and response of the different detectors as well as the magnetic field details, heavily reusing components from the GEANT4 toolkit [71] for the simulation of the passage of particles through matter. Additional modules are used to incorporate relevant modelling details such as the distribution of the interaction vertices in the interaction region, referred to as *vertex smearing*, and the addition of particles coming from additional soft interactions in the same collision or from adjacent bunch crossings, denoted as *pileup mixing*, which can affect the readouts and subsequent interpretation due to the overlapping of detector deposits and detector sensitivity dead-times.

As can be conjectured by its level of detail, such simulation processes are very time consuming, taking several minutes of CPU time given currently available computing technologies, for producing a realistic detector readout for each initial set of particles produced at a primary *hard* interaction. Given that oftentimes billions of generated events (i.e. simulated observations) of common processes are needed in order to obtain a realistic modelling of known types of interactions, alternative simulation techniques are sometimes used. By trading off some accuracy with simulation speed, the modelling of the physical processes and detector responses can be simplified, reducing running times considerably, up to two orders of magnitude [72]. Alternatively, as initially stated at the beginning of this section, detailed simulated observations can be used to directly parametrise low-dimensional summaries of the detector readout, such as the reconstructed main quantities that will be presented in the next section, by using approximate conditional probability density functions. While this approach, implemented in software packages such as DELPHES [73], is limited by the flexibility and accuracy of the modelling of the conditional probabilities, it is very useful as a very fast substitute of the full simulation chain for simplified studies that aim to obtain an approximate estimate the expected sensitivity reach or measurement accuracy of a given analysis. Peripherally related with the focus of this work, the use of unsupervised machine learning techniques structurally similar to those describe in Section 4.2.2 is being investigated to provide a fast simulation alternative without relying in a simplistic parameterisations [74, 75].

2.3.3 Event Reconstruction

In the previous sections, the generative mechanisms by which particles produce signals in the different detectors, as well as the techniques used to procedurally simulate them with high fidelity, were summarised. In contrast with simulated events, the set of underlying particles that were produced in the interaction region, and subsequently detected, are not known a priori in real collisions. A very helpful task to understand the nature of the fundamental interaction that likely happened in a collision is to infer the type and properties of the particles that were probably produced on a given collision given the detector output. Such procedure is generally referred to as *event reconstruction*. The underlying problem for achieving such goal is the assignment of detector readouts to the produced particle. This is not a simple problem, because the total number or the relative multiplicities of the different particle categories in a given event is unknown and variable, however expected to be large given the high-energy and luminosity conditions of proton-proton collisions.

Reconstruction at CMS: Particle-Flow Algorithm

A hierarchical strategy is followed to perform event reconstruction at the CMS experiment. First, the combined properties of small groups of low-level readouts for each sub-detector in each collision are used to construct higher-level summaries that distill the information regarding the origin, direction or energy of the particles. In a second step, such high level constructs are linked by an algorithm based on the expected properties of each particle type, to obtain a list of *physics objects* and their relevant attributes, which would probably correspond to those that actually were generated in the collision. Such approach, that is referred to as *particle-flow* (PF) event reconstruction [70] within CMS, has proven very effective to obtain a lower dimensional transformation of the detector readout that greatly simplifies the interpretation and categorisation of events based on their particle content.

As mentioned before, the first reconstruction stage encompasses the combination of detector traces in each sub-detector system to create higher level constructs. In the tracking detector, this amounts to the association of location estimates for the signals detected in all layers of the pixel and strip detector, referred to as *hits*, to trajectories of charged particles, simply called *tracks*. This inverse measurement problem is approached in CMS by using a combinatorial extension of the Kalman Filter algorithm [76, 77, 65]. In broad terms, the algorithm starts by selection sets of two-hit and three-hit associations from the inner layers, referred to as *seeds*, which

are then extrapolated outwards and used to gather hits in the the other layers by consecutive prediction and update steps, keeping all combinations that are deemed compatible. An additional step is then carried out, that filters out all candidate tracks under some pre-defined quality threshold and removes possible duplicates. Once the set of hits that define each track are found, their parameters are fitted again using a more detailed prediction step in the Kalman filter, thus obtaining more accurate estimates for their origin, momentum and direction.

The reconstructed charged particle trajectories can be used to identify the spatial locations where proton-proton interactions occurred in each bunch crossing, dubbed *primary vertices*, by extrapolating them back to the collision region and looking for overlapping subsets. In practice, a custom algorithm for vertex adaptive fitting [78] is used in combination with deterministic annealing, to identify and compute the vertices location and their uncertainty more accurately. Most primary vertices correspond to soft scattering processes (pileup), and can be used to characterise the position and size of the interaction region. In collisions where a hard interaction occurs, the main primary vertex may effectively be identified with the one whose linked tracks transverse momenta squared sum $\sum p_T^2$ is the largest. The distinction of a main primary vertex is useful to mitigate the effect of pile-up interactions in reconstruction by removing the contributions from particles linked to pileup vertices.

Regarding the calorimeter detector readouts, the initial step comprises the clustering of low-level deposits in each sub-detector, so as to identify the energy remnants left by each individual particle. The clustering procedure starts by finding the calorimeter cells where the amount of deposited energy are local maximal, referred to as *seed* deposits. The deposits from contiguous energy cells are grouped together until their energy is smaller than twice the expected noise level, forming larger groups referred to as *topological clusters*. Because such clusters might be the result of the overlapping of the energy deposited by two or more particles, the final clusters are identified by fitting a Gaussian-mixture model via the expectation-maximisation algorithm, using the number of initial seeds present in the cluster as the number of Gaussian components in the mixture. The fitted cluster amplitudes are thus expected to be heavily correlated with the energy deposited by an individual particle, however extensive calibration based on a detailed simulation of the detector and the assumed particle type is needed for accurate energy estimates. The resulting calibrated clusters in each sub-detector (ECAL, HCAL and HF) is instrumental for improve the energy measurement of charged hadrons, identifying and measuring the energy of

neutral hadrons and photons, and to facilitate the identification and reconstruction of electrons.

Once the basic elements for event reconstruction have been constructed, charged particle tracks and calorimeter clusters are linked together to form *blocks*. This step is an attempt to group the different traces that particle can leave in the various sub-detectors, by linking pairs of elements based on their distance in the (η, ϕ) plane and other properties depending of the specific sub-systems considered. When considering links between the inner tracker and calorimeter clusters, the curvature of the tracks and other details regarding the detector geometry are taken into account. Calorimeter cluster-to-cluster links between the HCAL and ECAL, and between the ECAL and the pre-shower clusters are also sought. Additionally, ECAL clusters possibly created by bremsstrahlung photons can also be linked to electron-like tracks if they are consistent with an extrapolation of the track tangent. Finally, links between two tracks due subsequent photon conversion via pair production are also considered if the sum of track momenta matches the mentioned electron-like track tangent.

The outcome of the aforementioned procedure is a set of blocks of elements for a given collision readout, formed by associating elements that have been directly linked or share a common link with other elements. The following reconstruction step is referred to as *object identification*, and it is based in the association of blocks to a list of particle candidates, also known as *physics objects*. This is done sequentially, starting out by the objects that more easily identified (e.g. muons) and progressively masking out the blocks that are considered for each object until all particles candidates have been reconstructed. The reconstruction process is rather conservative, given that most CMS data analysis share the same reconstructed *physics objects*, therefore it is common to specify additional selection criteria on the resulting set of objects based on their properties within each analysis to reduce the rate of fake or wrong reconstruction. The rest of this section is devoted to discuss in more detail the identification, calibration and common selection requirements on the main reconstructed objects that are used within physical analyses.

Muon Reconstruction

Muons can be thought of as the easiest object to identify given the observed detector readouts, because they are the only particle expected to reach the outer tracking systems (i.e. muon detecting system). Furthermore, the detecting volume far away from the interaction region is much larger and hence the density of particle trajectories is considerably lower. The sparse particle hits in each of the muon detector systems

are linked to form tracks that can be combined using a Kalman filter, similarly to what is done for the inner tracker as described earlier this section. To increase the measurement accuracy and reduce the fake rate, for analyses directly studying final states including muons, oftentimes a matching between the track segments in the muon detectors and a those in the inner tracker is required. The details and performance of the reconstruction procedure depend on the momenta of the muon, and are described in more detail the following reference [65].

The main challenges of muon reconstruction include the dismissal of muons produced by cosmic rays hitting the atmosphere and going through the CMS detector, simply dubbed as *cosmic muons*, as well as the rejection of signals from very energetic hadrons produced in the collision that are able to transverse the dense calorimeter and magnet section and still produce a response in the muon detectors, that are referred to as *punch-through* hadrons. In addition, muons are a common product of the decay of hadrons and it is thus important to differentiate between muons produced in the primary interaction, or *prompt muons*, and those produced in a secondary decay of another particle. The amount of energy deposited around the muon trajectory, called *muon isolation*, as well as the distance to the primary vertex are important variables for such distinction.

Electron and Photon Reconstruction

Electron reconstruction is more challenging because it uses the readouts from the inner tracker and the ECAL, both detectors being sensitive to additional charged particles coming out from the interaction volume, and the latter also to high-energy photons. Furthermore, electrons lose energy in their curved trajectories through the tracker, thereby complicating an accurate track reconstruction. The latter can be accounted for during the track reconstruction by using a Gaussian-Sum filter extension fo the Kalman filter [79] algorithm, which can be used to model the previously mentioned non-linearities. The procedural details of the identification and property measurement for electrons depend on their transverse momenta. Lower energy electrons are more accurately indentified using the inner tracker hits, while the electromagnetic calorimeter is more useful at higher energy ranges. These and other details regarding electron reconstruction are discussed in the following reference [80].

The electron momentum direction is measured using the track information, while the energy is estimated by combining both information from the tracking and calorimeter detectors. In order to obtain precise energy and momentum estimates, under 5% in the full pseudo-rapidity range, a calibration step is required to correct for non-

clustered energy deposits and pile-up contributions. Similarly to what is done for muons, additional quality criteria can be applied to distinguish between the electrons produced in the primary interaction and those coming from hadronic decays or converted photons, including conditions on several track-based and calorimeter-based observables as well as isolation requirements, the latter ensuring that no significant energy from hadrons was deposited around the electron trajectory.

High-energy photons are identified and reconstructed using only the calorimeter [81], when the energy distribution in the ECAL calorimeter cells is consistent with that expected from a photon shower. Energy isolation requirements are also essential to distinguish photons coming from hadrons or secondary radiative decays, which will be discussed together with hadrons, from those originated as a direct product of the primary interaction. Additional quality and fine-tuned calibration is often used, for example in the $H \to \gamma\gamma$ analysis, to reduce the fake rate and obtain higher momentum resolution.

Jet Reconstruction and B-Tagging

Once muons, electrons and isolated photons in the event have been identified, the remaining particle-flow blocks (i.e. linked tracks and/or calorimeters deposits) are interpreted either as neutral or charged *PF candidates* [70]. These physics object candidates account for charged and neutral hadrons coming from the hadronisation of partons produced in the collision or their subsequent decays, as well as non-isolated photons radiated during those processes. When the aim is studying high-energy fundamental interactions that produce partons or other parton-decaying intermediate particles (e.g. $H \to b\bar{b}$), such reconstructed objects are not directly practical because their individual momenta cannot be linked with original parton momentum. This is because the processes of fragmentation, hadronisation, decays and associated radiation are stochastic, producing tree-like structures with multiple leafs as discussed in Section 1.3.4, difficulting most attempts to uniquely identify each parton with its decay chain. In addition, contributions from additional soft pileup interactions may further complicate the mentioned assignment, while this factor is lessen by charged hadron subtraction techniques (CHS) [82] based on removing candidates not associated with a primary vertex.

A possible way to construct simpler observables that can be linked with the original partons is to create composite objects based the remaining candidates through clustering. These objects, referred to as *jets*, are an attempt to represent the chain of hadrons and radiated energy produced, so the original parton energy and momentum

can be recovered from the summed of the components. They can be geometrically viewed as cones coming from the interaction region, covering an angular area ΔR of a given size in an outwards direction, that contains a collimated set of hadrons and radiated photons flying away a direction similar to the original parton. Several jet clustering algorithms exist, each characterised by a given a size or resolution parameter R and a recombination scheme, defining how candidates are combined to create the composite clustered object.

Due to the properties of hadronisation and QCD radiation processes, a common requirement for such clustering algorithms is that they do not change significantly when a particle is split in two collinear ones (i.e. they are *collinear safe*) or additional soft radiation is produced by one of the clustered particles (i.e. they are *infrared safe*), which greatly simplifies direct comparison with generation level observables. In particular, in the analysis described in Chapter 5, the default jet CMS reconstruction is extensively used, which is based on the anti-k$_T$ algorithm [83]. This is a sequential algorithm, also referred to as hierarchical agglomerative clustering in statistical language. The algorithm starts by assigning each candidate to each own cluster and successively merging them according to the following distances between two jets indexes as i and j respectively:

$$d_{ij} = \min(p_{Ti}^{2a}, p_{Tj}^{2a}) \frac{\Delta R_{ij}^2}{R^2} \quad \text{and} \quad d_{iB} = p_{Ti}^{2a} \tag{2.10}$$

where ΔR_{ij}^2 is the $\eta - \phi$ plane distance as defined in Section 2.2.1, p_{Ti}^{2a} and p_{Tj}^{2a} are the transverse momenta of each jet, R is the size parameter, and $a = -1$ for the anti-k$_T$ algorithm. The algorithm starts by computing all distances d_{ij} and d_{iB} for all initial candidates, which are placed in a list. If the minimum corresponds to a given distance between two candidates d_{ij} then both candidates are removed from the candidate list and group together by summing their four momenta forming a composite object, which is in turn added to the list. Alternatively, if the minimum distance is d_{iB}, the i candidate is assigned as a jet and removed from the list. Such procedure is recursively applied until the list is empty, because all single and composite candidates have been grouped with other candidates or defined as a jets of a given size R. The choice of the parameter R has to provide a balance between covering all the radiation from the initial parton and being increasingly affected by noise produced by soft particles. During the data taking period considered in Chapter 5, a cone size $R = 0.4$ was used for the default jet collection, used in the analysis. Larger jet (e.g. $R = 0.8$) cones are used in analyses that include final states with highly boosted intermediate

particles, that produce a collimated set of hadrons and radiation when they decay, commonly with internal structure that can be exploited to improve the sensitivity. Various sequential clustering algorithms can be defined by considering a different value of a in Equation 2.10. If a negative choice for the exponent a, as used in the anti-k$_T$ algorithm, higher transverse momenta particles are clustered first and thus the final jet outcome is less sensitive to soft pileup contributions and radiation.

The energy and momenta of the resulting jets is not expected to match accurately that of the original partons, due to the compound effect of detector readout and or non-linearities, as well as effect from pileup contribution. This motivates the application of a set of corrections, referred to as *jet energy corrections* (JECs) [84], that greatly reduce this discrepancies by sequentially shifting and rescaling the jet four-momenta based on extensive calibrations obtained from simulation.

So far, jets have been defined as an experimental simplification of hadronisation, decay and fragmentation chains in order to estimate the energy and the momenta of initial partons produced in the collision, and we have ignored other properties of the original parton. In particular, information regarding the flavour of the initial parton can be instrumental to distinguish event containing jets coming from high-energy processes with physical interesting intermediate particles like a Higgs boson H or top quarks/antiquarks, which predominantly decay to b quarks. Heavy flavour b quarks, and to a lesser extent also for c quarks, hadronise producing B (and D) hadrons that have lifetimes long enough to fly away from the primary vertex before decaying.

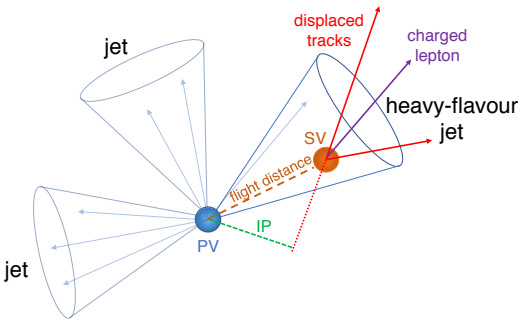


Figure 2.11: Schematic representation of the features of a heavy-flavour jet that can be used for jet tagging including the presence charged tracks, with a large impact parameter (IP), that is not compatible with the primary vertex (PV), and a reconstructed secondary vertex (SV), both due to the decay of B or C hadrons. The figure has been adapted from [85].

Some properties of the decay of B and D hadrons can be used to distinguish heavy flavour jets from those produced by light quarks and gluon hadronisation processes. The lifetimes of heavy flavour hadrons are often long, e.g. 1.638 ± 0.004 ps and 1.519 ± 0.005 ps for B^+ and B^0 [8], respectively. When long-lived hadrons are highly boosted, they can move several millimetres away from the primary vertex where they were produced before decaying. Thus, heavy flavour jets are associated with the presence of displaced charged tracks and secondary vertices (SV) within the jet, as depicted by Figure 2.11. In addition, both B or D hadron decays are characterised by a large decay multiplicity (average 5 charged daughters) and a high probability (36%) of producing a lepton in their decays chain. Flavour tagging techniques, often referred to as b-tagging or c-tagging when the purpose is to identify a jets originating from a particular type of parton, combine quantitative information related with the various properties previously mentioned to distinguish the flavour of the parton that generated a given jet.

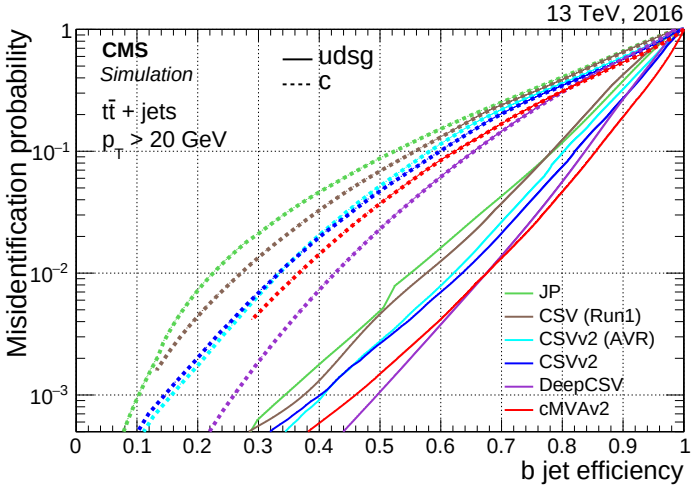


Figure 2.12: Misidentification probability (in log scale) for jets originating from c (dashed line) and light quarks or gluons (solid line) versus b-tagging efficiency, for different b-tagging algorithms available in CMS during 2016. The misidentification probability and efficiencies are obtained from the subset of reconstructed jets with a $p_T > 20$ GeV from a large $t\bar{t}$ simulated sample. The figure has been adapted from [85].

Heavy flavour tagging, particularly b-tagging can very useful for analyses considering jets in final states, such as the search for Higgs pair production with CMS data described in Chapter 5. The misidentification versus efficiency curve of the main b-tagging algorithms that were available in 2016 for high-energy jets is shown

in Figure 2.12. They differ in the subset of information associated to the jet that is considered and the specifics of the multivariate techniques used to construct the final discriminator. The simplest b-tagging algorithm, referred to as jet probability (JP) is only based a calibrated estimation of the displaced track probabilities. The b-tagging discriminators pertaining to the combined secondary vertex (CSV) family combine displaced track information with reconstructed secondary vertex. The improvement between different CSV-based b-tagging algorithms is due to the use of more advanced statistical learning techniques and additional discriminating variables [85]. The CMVAv2 algorithm, which is used in the analysis included in Chapter 5, combines the output from JP and CSVv2 algorithms with two taggers that summarise the information from non-isolated electrons and muons inside the jet.

In Section 4.3.2, the role of recent advances in machine learning techniques for particle identification and regression are discussed in more detail, focussing on the development and integration on a new deep learning based multi-category jet tagger referred as DeepJet. The DeepJet tagger outperforms both CMVAv2 and DeepCSV (which also leverages deep learning technologies), while providing additional discrimination capabilities (e.g. gluon-quark separation). It is worth mentioning that jet tagging techniques can also be applied for identifying substructure in larger radius jets, which are very relevant for analyses where highly boosted intermediate objects are expected, but are not discussed in this work.

Missing Transverse Energy

As hinted in Section 2.3.1, neutrinos can be produced at high-energy proton-proton collisions, and they leave the detector undetected. Nevertheless, the presence of neutrinos (or other hypothetically weakly-interacting particles) can be inferred by the total momentum imbalance in the transverse plane of the event. Within the Particle-Flow reconstruction framework, this accounts to computing the vectorial sum of the transverse momenta of all PF reconstructed objects:

$$\vec{p}_T^{\,\text{miss}} = \sum \vec{p}_{Ti}^{\,\text{miss}} \tag{2.11}$$

where $\vec{p}_T^{\,\text{miss}}$ is the total missing transverse momentum, whose Euclidean norm modulo is the missing transverse energy E_T^{miss}, and $\vec{p}_{Ti}^{\,\text{miss}}$ is the transverse momentum each PF candidate.

It is worth remarking that some hadron decay processes can produce neutrinos, therefore a non-zero transverse missing energy E_T^{miss} does not necessarily mean that

weakly-interacting particles were produced in the hard interaction or by its direct products. Furthermore, any mis-detections or mis-measurements of the momenta of some of the produced particles can lead to transverse energy imbalances.

3 Statistical Modelling and Inference at the LHC

> Life is complicated,
> but not uninteresting.
>
> Jerzy Neyman

In this chapter we will consider the problem of extracting quantitative information about the validity or properties of the different theoretical models (see Chapter 1), which can be made given the experimental data acquired in a controlled setting (see Chapter 2). We will begin by formally defining the properties and structure of the statistical models used to link the parameters of interest with the experimental data, followed by a description of the inference problems in experimental high-energy physics and how they can be tackled with statistical techniques. Some relevant particularities of the inference problems typically of interest of the LHC experiments will be discussed, mainly the generative-only nature of the simulation models and the high dimensionality of the data. As we will see, these issues are intimately related, the former requiring the use of likelihood-free inference techniques such as constructing non-parametric sample likelihoods, which in turn demand for lower dimensional summary statistics.

3.1 Statistical Modelling

An essential element for carrying out statistical inference is the availability of an adequate statistical model. In this section, the main characteristics of the statistical models used in particle collider analyses will be formally developed from first principles. This methodology allows a mathematical approach to their structure and factorisation. This will prove useful to establish a formal link between the techniques discussed in the next chapters and the simulation-based generative models that are often used to describe the data. Additionally, the role and importance of event selection, event reconstruction and dimensionality reduction - i.e. the compression of the

relevant information from high-dimensional data into a lower-dimensional representation, such as the output of a multivariate classifier - will be described in the larger statistical framework of an LHC analysis. Lastly, the main approaches commonly followed to construct synthetic[1] likelihoods that efficiently connect summaries of the detector observation with the parameters of interest will be illustrated.

3.1.1 Overview

Let us suppose that we record a collection of raw detector readouts $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$ for a total of n bunch crossings at a particle collider experiment, such as CMS at the LHC (see Section 2.2). Note that vector notation is used for each individual readout, also referred to as event, because for mathematical simplification we will be assuming that each detector observation can be embedded - in the mathematical sense - as a member of a fixed size d-dimensional space, i.e. $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, even though variable-size sets or tree-like structures might be a more compact and useful representation in practice, as will be discussed later. As a starting point, let us assume for simplicity that the detector readout for every bunch crossing is recorded, i.e. no trigger filtering system as the one described in Section 2.2.7 is in place, hence after each bunch crossing i a given raw detector readout $\boldsymbol{x}_i$ will be obtained. From here onwards, each event/observation/readout will be assumed to be independent and identically distributed (i.i.d.), a reasonable approximation if the experimental conditions are stable during the acquisition period as discussed at the beginning of Section 2.3; consequently the event ordering or index i are not relevant.

Experiment Outcome

Within the above framework, we could begin by posing the question of how we expect the readout output, which can be effectively treated as a random variable $\boldsymbol{x}$, to be distributed and how such distribution is related with the (theoretical) parameters we are interested in measuring or the model extensions we are interested in testing using the experiment. We would like then to model the probability density distribution function generating a given observation $\boldsymbol{x}_i$ conditional on the parameters of interest, that is:

$$\boldsymbol{x}_i \sim p(\boldsymbol{x}|\boldsymbol{\theta}) \tag{3.1}$$

[1] In this work, synthetic likelihood will be used to refer to likelihoods that are not based on the probability distribution function of the generative model, but on non-parametric approximations using low-dimensional summaries of the data.

where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ denotes all the parameters we are interested in and affects the detector outcome of collisions. As will be extensively discussed in this chapter, an analytical or even tractable approximation of $p(\boldsymbol{x}|\boldsymbol{\theta})$ is not attainable, given that we are considering $\boldsymbol{x}$ to be a representation of the raw readout of all sub-detectors, thus its dimensionality d can be of the order $\mathcal{O}(10^8)$. It is worth mentioning that even d is very high, each observation is usually extremely sparse given that most of the detectors would not sense any signal. The total number of observations n is also very large at modern colliders, e.g. a collision occurs each 25 ns at the LHC. Furthermore, the known interactions that produce the set of particles of the event as well as the subsequent physical processes that generate the readouts in the detectors are overly complex, and realistic modelling can only be obtained through simulation, as jointly reviewed in Section 1.3 and Section 2.3.

Mixture Structure

While a detailed closed-form description of $p(\boldsymbol{x}|\boldsymbol{\theta})$ cannot be obtained, we can safely make a very useful remark about its basic structure, which is fundamental for simplyfing the statistical treatment of particle collider observations and simulations, and was already hinted at in Section 1.3.1 when discussing the possible outcomes of fundamental proton-proton interactions. The underlying process generating $\boldsymbol{x}$ can be treated as a *mixture model*, which can be expressed as the probabilistic composition of samples from multiple probabilistic distributions corresponding to different types of interaction processes occurring in the collision. If we knew the probabilistic distribution function of each mixture component $p_j(\boldsymbol{x}|\boldsymbol{\theta})$ then $p(\boldsymbol{x}|\boldsymbol{\theta})$ could be expressed as:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=0}^{K-1} \phi_j \; p_j(\boldsymbol{x}|\boldsymbol{\theta}) \tag{3.2}$$

where K is the number of mixture components and ϕ_j is the mixture weight/fraction, i.e. probability for a sample to be originated from each mixture component j. The specifics of the mixture expansion as well as the total number of mixture components are not uniquely defined, but are based on the independence of groups of physical processes, as will be discussed later. Practically, each $p_j(\boldsymbol{x}|\boldsymbol{\theta})$ will be intractable due to the exact same reason that $p(\boldsymbol{x}|\boldsymbol{\theta})$ is intractable, thus a more sensible description of the mixture model is its generative definition, described by the following two-step sampling procedure:

$$z_i \sim \text{Categorical}(\boldsymbol{\phi}) \quad \longrightarrow \quad \boldsymbol{x}_i \sim p_{z_i}(\boldsymbol{x}|\boldsymbol{\theta}) \tag{3.3}$$

describing the sampling of random integer $z_i \in \{0, \ldots, K-1\}$ from a random categorical[2] distribution and the subsequent sampling of the corresponding mixture component indexed by z_i, where $\boldsymbol{\phi} = \{\phi_0, \ldots, \phi_{K-1}\}$ is the vector of probabilities for each of the mixture components. For here onwards, mixture models might in some cases be portrayed by using the analytical depiction as in Equation 3.2, always noting that the generative approach might be more convenient for the actual estimation of expectation values when the mixture component distributions $p_j(\boldsymbol{x}|\boldsymbol{\theta})$ are not tractable.

Mixture Components

The mixture model structure can be directly linked to the physical processes happening in fundamental proton-proton collisions and within the detectors used to study them, as described in previous chapters. As an additional simplification for now, let us neglect the effect of multiple particle interactions, described in Section 2.1.3. For each proton bunch crossing, hard interactions (i.e. ones associated with a large characteristic energy scale Q^2, whose cut-off does not have be specified for this particular argument) between partons might or might not occur, given the stochastic nature of the scattering processes. We could nevertheless associate a probability for a hard interaction happening ϕ_{hard}, as well to it not happening $\phi_{\text{not-hard}} = 1 - \phi_{\text{hard}}$. Given the proton colliding conditions at the LHC, the latter case is much more likely, i.e. $\phi_{\text{not-hard}} \gg \phi_{\text{hard}}$, yet the relative probabilities depend on the energy scale cut-off considered.

We can further break each previously mentioned category in sub-components corresponding to different types of processes. The hard interaction category can itself be expressed as a mixture of groups of physical interactions that can produce a hard scattering[3], so the probability ϕ_{hard} can be expresses as the following sum:

$$\phi_{\text{hard}} = \phi_0 + \cdots + \phi_{K-2} = \sum_{k \in H} \phi_k \tag{3.4}$$

[2] Here categorical distribution refers to the special case of the multinomial distribution were the number of trials is one.

[3] The term *group/type of interactions* here generally refers to a set of processes that could be generatively modelled independently, not to quantum mechanical amplitudes or intensities of a process. For example, each group can correspond to a group of processes with a given final state $pp \to X$ which could be modelled by sampling its differential cross section from Equation 1.32 followed by parton showering and detector simulation. The group category is a latent/hidden variable for each event, i.e. it is not observed.

where H represents a given set of independent contributions k, each characterised by a distribution $p_j(\boldsymbol{x}|\boldsymbol{\theta})$, which depends on the group j of processes that produce hard scatterings. Such a set is not uniquely defined nor its the number of elements, given that any two components a and b in H can be substituted by c, where $\phi_c = \phi_a + \phi_b$ and

$$p_c(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\phi_a}{\phi_a + \phi_b}\, p_a(\boldsymbol{x}|\boldsymbol{\theta}) + \frac{\phi_b}{\phi_a + \phi_b}\, p_b(\boldsymbol{x}|\boldsymbol{\theta}) \tag{3.5}$$

which can be applied recursively to alter the number of components in the set. Independently on the basis chosen for the mixture expansion, in general it is not possible to infer the latent category z_i (see Equation 3.2 given an observation $\boldsymbol{x}_i$, because $\boldsymbol{x}_i$ may be in the support of several mixture components $p_j(\boldsymbol{x}|\boldsymbol{\theta})$. Only probabilistic statements about the generative group j can be made based on the observations.

A convenient definition for the set H is one that is aligned with the way theoretical calculations are carried out, given that the relative probability for a given process $\phi_{pp\to X}$ will be proportional to its total cross section $\sigma(pp \to X)$, while its readout distribution will depend on its differential cross section $d\sigma(pp \to X)$ and its support (i.e. subset of the function domain not mapped to zero). In fact, given that the total and differential cross sections are proportional to the matrix element squared (see Section 1.1.1) of a given process $d\sigma(pp \to X) \propto |\mathcal{M}|^2$, it is often possible to further divide each process into the cross product of Feynman diagram expansions (including interference terms). This can be a very useful notion for some analysis use cases, and is related with the approach that will be used in Chapter 5.

Signal and Background

Oftentimes, we are interested in studying a subset $S \subset H$ of all the hard interaction processes, which will be referred to as signal set in what follows. This can be a single type of physical process $\sigma(pp \to X)$, e.g. the inclusive production of a pair of Higgs bosons $\sigma(pp \to \mathrm{HH} + \text{other})$, or several, which it can be effectively viewed as one mixture component using Equation 3.5. We can accordingly define the background subset $B = H - S$, as the result of all other generating processes in H that we are not interested in, a definition which could also be extended to include collisions where non-hard processes occurred if needed. Such distinction between generating processes of interest S and background B is at the root of every analysis at the LHC and it is motivated by the fact that small changes of the parameters of the SM or its

theoretical extensions/alternatives affect only a subset of the produced processes at leading order, those that are governed by the interactions linked to the parameter.

As a matter of a fact, customarily statistical inference at the LHC is not carried out directly on the parameters of the SM or the extension being studied, but on the relative frequency of the set of processes of interest ϕ_S or the properties of its distribution $p_S(\boldsymbol{x}|\boldsymbol{\theta})$. As previously mentioned, the former is proportional to the cross section of the signal processes σ_S (see Section 1.3) while the latter can include properties such as the mass of an intermediate particle resonance[4] (e.g. the Higgs mass m_{H}) or the general behaviour of the differential distribution (i.e. using unfolding methods to remove the experimental effects, which are not discussed in this work). Those parametric proxies can then be used by comparing them with the theoretical predictions of the SM or the alternative considered, in order to exclude it or constrain its fundamental parameters (i.e. those that appear in the Lagrangian).

Event Selection

Given the mixture model structure expected for $p(\boldsymbol{x}|\boldsymbol{\theta})$ and the fact we are only interested in a small amount of the readout generating processes for each collision, because in general $\phi_S \ll \phi_B \ll \phi_{\text{not-hard}}$, the effect of trigger or any other *event selection* should be considered. The role of event selection is to reduce the fraction of events that do not contain useful information for the inference task of interest. Trigger selection can be thought of as a technical requirement, reducing the total rate of detector readouts recorded to match the available hardware for data acquisition, as discussed in Section 2.2.7. The purpose of analysis selection, as will be discussed in Chapter 5, is instead to reduce the expected contribution of background processes that are not well-modelled by simulation, as well as to the increase the expected fraction of signal events in synthetic counting likelihoods, such as those which will be detailed in Section 3.1.3.

In general mathematical terms, any deterministic event selection can be thought of as an indicator function $\mathbb{1}_{\mathcal{C}} : \mathcal{X} \longrightarrow \{0, 1\}$, of a given subset of the set of possible detector readouts $\mathcal{C} \subseteq \mathcal{X}$. The indicator function $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ can be defined as:

$$\mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} 1 \text{ if } \mathbf{x} \in C \\ 0 \text{ if } \mathbf{x} \notin C \end{cases} \tag{3.6}$$

[4] In particle physics, a resonance is a peak around a certain energy in the differential cross section associated with the production of subatomic particles.

where the specific definition of such function depends on the definition of the subset $\mathcal{C}$, e.g. a simple cut on a one-dimensional function $f : \mathcal{X} \longrightarrow T \subseteq \mathcal{R}$ of the readout $f(\boldsymbol{x}) > t_{\text{cut}}$. Any indicator function can also be viewed as a boolean predicate function, so the event selection can also be expressed as a combination of selection functions, i.e. if the set $\mathcal{C} = \mathcal{A} \cap \mathcal{B}$ is the intersection between two subsets, the indicator function of C can be simply expressed as the product $\mathbb{1}_{\mathcal{C}} = \mathbb{1}_{\mathcal{A}} \cdot \mathbb{1}_{\mathcal{B}}$. This framework is flexible enough to represent all deterministic event selections, and it could also be extended by an independent non-deterministic term without affecting the rest of the considerations presented in this chapter. A non-deterministic factor could be useful to model for example trigger prescales, which are trigger decisions based on randomly selecting a fraction of all the selected events to be recorded, ensuring that the total rate is manageable.

In practice, a given selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$, likely based on a composition of simple criteria, would have been imposed on the recorded detector readouts before any statistical analysis is carried out. The structure of the statistical model $g(\boldsymbol{x}|\boldsymbol{\theta})$ resulting after applying an arbitrary selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ on a mixture model as the one described in Equation 3.1 can be obtained by multiplying the probability density by $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$. After including the relevant normalisation term, the resulting probability distribution can be expressed as:

$$g(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) \sum_{j=0}^{K-1} \phi_j \; p_j(\boldsymbol{x}|\boldsymbol{\theta})}{\int \left(\mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) \sum_{j=0}^{K-1} \phi_j \; p_j(\boldsymbol{x}|\boldsymbol{\theta}) \right) d\boldsymbol{x}} = \sum_{j=0}^{K-1} \left(\frac{\phi_j \epsilon_j}{\sum_{j=0}^{K-1} \phi_j \epsilon_j} \right) g_j(\boldsymbol{x}|\boldsymbol{\theta}) \tag{3.7}$$

where $g_j(\boldsymbol{x}|\boldsymbol{\theta}) = \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) p_j(\boldsymbol{x}|\boldsymbol{\theta}) / \epsilon_j$ is the probability density function of each mixture component after the selection, $\epsilon_j = \int \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) p_j(\boldsymbol{x}|\boldsymbol{\theta})$ is the *efficiency* on the selection on each mixture, and the integral sign in the denominator in the last expression has been simplified by noting that $\int g_j(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} = 1$. From Equation 3.7 it becomes clear that the statistical model after any event selection is also a mixture model, whose mixture components are $g_j(\boldsymbol{x}|\boldsymbol{\theta})$ and mixture fractions are $\chi_j = \phi_j \epsilon_j / \sum_{j=0}^{K-1} \phi_j \epsilon_j$. This fact will be very relevant to build statistical models of the observed data after an event event selection is in place.

So far, no explicit assumptions on the probability distribution functions of each mixture component j or the details of the event selection function $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ have been considered, in order to keep the previously developed modelling framework as general as possible. In the next sections, it will become increasingly clear how $p_j(\boldsymbol{x}|\boldsymbol{\theta})$, and in

turn $g_j(\boldsymbol{x}|\boldsymbol{\theta})$ and the efficiency ϵ_j, can be modelled by generating simulated detector readouts produced by a given process j.

3.1.2 Simulation as Generative Modelling

The physical principles underlying the simulation of detector readouts, or events, for a given hard proton-proton interaction process were reviewed in Section 1.3 and Section 2.3. Instead of focussing on the procedural details of event generation, the focus of this section is the study of the simulation chain as a generative statistical model, together with its basic structure and properties, that will be useful later to understand many analysis techniques that are commonly used in experimental particle physics.

For simplicity, we will be considering the statistical model describing the distribution of observations of detector readouts before any event selection, what was referred to as $p(\boldsymbol{x}|\boldsymbol{\theta})$ in the previous section. Always taking into account that the distribution after any arbitrary deterministic event selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ is also a mixture model (see Equation 3.7) and samples under the corresponding probability distribution functions $g_j(\boldsymbol{x}|\boldsymbol{\theta})$ and mixture fractions χ_j can easily obtained from the non-selected simulated events, as it is actually done in practice.

Observable and Latent Variables

The first step to build a generative statistical model is to define what are the observed variables and what are the hidden quantities, referred to as *latent variables*, that explain the structure of the data. For particle collider experiments, we may consider the full detector readout $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ as the only observable variable, given that any other observable can be expressed as a function of the raw readout, as will be discussed in Section 3.1.3. The probability density function of the data $p(\boldsymbol{x}|\boldsymbol{\theta})$ from a generative standpoint can be written as an integration of the joint distribution $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$ over all latent variables $\boldsymbol{z}$ of an event:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \int p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{z} \tag{3.8}$$

where $\boldsymbol{\theta}$ is a vector with all model parameters, which normally are global (same for all the observations) and include the theory parameters of interest as well as any other parameter that affect the detector readouts. While the true generative model of the data $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$ is unknown, knowledge about the underlying physical

processes described in in Section 1.3 and Section 2.3 can be used to build a generative approximation of $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$ which can describe the observed data realistically and be used to carry out inference on the parameters of interest.

In fact, one of the most relevant latent variables at particle colliders has been already introduced with the generative definition of a mixture model in Equation 3.3, the mixture assignment integer $z_i \in \{0, \ldots, K-1\}$. This latent variable represents which type of fundamental interaction occurred in the event, and is useful to exemplify the main property of latent variables: that they are not observed but can only (at most) be inferred. Let us consider the problem of finding out the type of interaction j that caused a single detector readout observation $\boldsymbol{x}_i$. As long as $\boldsymbol{x}_i$ is in the support space of more than one of the mixture components $p_j(\boldsymbol{x}|\boldsymbol{\theta})$, which is almost always the case, only probabilistic statements about the type of interaction originating $\boldsymbol{x}_i$ can be made, even if the $p_j(\boldsymbol{x}|\boldsymbol{\theta})$ are known. In practice, $p_j(\boldsymbol{x}|\boldsymbol{\theta})$ are not known analytically so probabilistic classification techniques can be used to estimate the conditional probabilities based on simulated samples, as discussed in Chapter 4.

Structure of Generative Model

Other than the basic mixture model structure, our understanding of the underlying physical process occurring in proton-proton collisions can be used to recognise additional structure in the generative model by means of factorising the joint distribution $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$ in conditional factors matching the various simulation steps and their dependencies:

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p}) \sum_{j=0}^{K-1} p(z_i = j|\boldsymbol{\theta})p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}, z_i = j) \tag{3.9}$$

where each factor can be defined as follows:

- $p(z_i = j|\boldsymbol{\theta}) = \phi_j(\boldsymbol{\theta})$ is the probability of a given type of process j occurring, which is usually a function of theory parameters $\boldsymbol{\theta}$.
- $p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}, z_i = j)$ is the conditional probability density of a given set of parton-level four-momenta particles (characterised by the latent representation $\boldsymbol{z}_\mathrm{p} \in \mathcal{Z}_\mathrm{p}$) of being the outcome of a group of fundamental proton interaction processes $pp \longrightarrow X$ indexed by the latent variable $z_i \in \mathcal{Z}_i$, which might also be a function of theory parameters $\boldsymbol{\theta}$.

- $p(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p})$ is the conditional density of a given parton-shower outcome. $\boldsymbol{z}_\mathrm{s} \in \mathcal{Z}_\mathrm{d}$ as a function of the parton-level outcome.
- $p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s})$ is the conditional density of a set of detector interactions and readout noise $\boldsymbol{z}_\mathrm{d} \in \mathcal{Z}_\mathrm{d}$ as a function of the parton-shower output.
- $p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d})$ is the conditional density of a given detector readout $\boldsymbol{x} \in \mathcal{X}$ as a function of the detector material interactions and detector readout noise.

The dimensionality of the latent space greatly increases with each simulation step, from a single integer for $\mathcal{Z}_i$, to $\mathcal{O}(10)$ parton four-momenta variables within $\mathcal{Z}_p$, to $\mathcal{O}(100)$ after the parton-shower $\mathcal{Z}_s$, and finally to $\mathcal{O}(10^8)$ in the detector interaction latent space $\mathcal{Z}_d$ and also the observable readout space $\mathcal{X}$. In the factorisation presented in Equation 3.9, the dependence on the parameters has only has been made explicit for $p(z_i|\boldsymbol{\theta})$ and $p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}, z_i)$, that is because the theoretical parameters of interest $\boldsymbol{\theta}$ often only affect the rate of the different fundamental processes and their differential distributions, which correspond to the mentioned conditional probability distributions. In the actual simulation chain, all conditional factors typically depend on additional parameters which might be uncertain, and whose effect and modelling will be discussed in Section 3.1.4.

As previously mentioned, computer programs can be used to realistically simulate detector observations. For simulated observations, not only the final readout is observed, but all latent variables can be obtained from the intermediate steps of the generative chain. These variables, in particular $\boldsymbol{z}_\mathrm{p}$ and $\boldsymbol{z}_\mathrm{s}$, are commonly referred as *generator level observables*, and are extremely useful to construct techniques that approximate the latent variables from the detector readouts. In fact, the whole simulation chain can be viewed as a probabilistic program [86, 87], thus each of the factors in Equation 3.9 can be further broken down as a sequence of random samples, which can be used to speed up latent variable inference based on the execution traces, i.e. recorded sequences of random numbers generated for each observation.

Some joint factorisations are particularly useful for data analysis and simulation, such as the one making explicit the dependence between the differential partonic cross sections and the parton configuration in the collision, because it allows to factor out the density of the latent variables $\boldsymbol{z}_\mathrm{PDF}$ associated with the parton components (i.e. flavour and momenta of each interacting parton and factorisation scale μ_F^2, as depicted in Section 1.3.3). Each mixture component j in Equation 3.9, which represents a group of fundamental interactions between protons $pp \longrightarrow X$, can be expressed as the product of the probability of a given parton configuration $p(\boldsymbol{z}_\mathrm{PDF}|\boldsymbol{\theta}_\mathrm{PDF})$ and a

mixture over all parton configurations that can that produce $pp \longrightarrow X$, referred as L in the following expression:

$$p(z_i|\boldsymbol{\theta})\ p(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta}, z_i) = p(\boldsymbol{z}_{\mathrm{PDF}}|\boldsymbol{\theta}_{\mathrm{PDF}}) \sum^{g \in L} p(z_f = g|\boldsymbol{\theta}, z_{\mathrm{PDF}}) p(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta}, z_f = g) \tag{3.10}$$

where $p(z_f = g|\boldsymbol{\theta}, z_{\mathrm{PDF}})$ is the relative probability of given partonic process g given a parton configuration $\boldsymbol{z}_{\mathrm{PDF}}$ and $p(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta}, z_f = g)$ is the probability distribution function of the parton-level particles produced as a result of the interaction for a given partonic process g, which is proportional to the partonic differential cross section $d\sigma(ij \to X)$. This factorisation is basically a probabilistic model version of Equation 1.32, dealing with the QCD factorisation of the parton distribution functions and the hard process differential cross section.

Another relevant phenomenon that can be made explicit in the joint distribution $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$ is the effect of multiple hadron interactions in the collision, or pileup, as discussed in Section 2.1.3. Given that each proton-proton interaction is independent from the others, the effect of pileup interactions can be considered by augmenting the factor representing the conditional probability density of the detector interaction and noise as a function of the hard interaction parton shower output $p(\boldsymbol{z}_{\mathrm{d}}|\boldsymbol{z}_{\mathrm{s}})$ as follows:

$$p(\boldsymbol{z}_{\mathrm{d}}|\boldsymbol{z}_{\mathrm{s}}) = p(\boldsymbol{z}_{\mathrm{d}}|\boldsymbol{z}_{\mathrm{s}}, \boldsymbol{z}_{\mathrm{pileup}}) p(\boldsymbol{z}_{\mathrm{pileup}}|\boldsymbol{\theta}_{\mathrm{pileup}}) \tag{3.11}$$

where $\boldsymbol{z}_{\mathrm{pileup}}$ is a latent variable representing the details about the pileup interactions that happened in a given collision (i.e. number of interactions and their corresponding particle outcome), and $\boldsymbol{\theta}_{\mathrm{pileup}}$ are the bunch crossing and luminosity parameters that affect the pileup distribution.

Further structure in the generative model can be often found, depending on the process being generated, the modelling assumptions, and the latent space representation chosen. As an example, it is often useful to factorise out the latent subspace that depends directly on the subset of parameters of interest from those that do not. The conditional observations in that latent subspace can sometimes be analytically expressed, or their dimensionality may be low enough to use non-parametric density estimation techniques effectively, which can greatly simplify the modelling of changes in the parameters of interest.

Simulated Observations

The mentioned mixing structure of the probability distribution function $p(\boldsymbol{x}|\boldsymbol{\theta})$ greatly simplifies the simulation of realistic observations, because large datasets $S_j = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_m\}$ of simulated observations for each type of interaction j can be simulated before any event selection. The expected value of any measurable function of the detector read-out $f(\boldsymbol{x})$ for events coming from a given process j can be expressed as:

$$\mathop{\mathbb{E}}_{x\sim p_j(\boldsymbol{x}|\boldsymbol{\theta})}[f(\boldsymbol{x})] = \int f(\boldsymbol{x}) p_j(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \approx \frac{1}{m} \sum^{\boldsymbol{x}_s \in S_j} f(\boldsymbol{x}_s) \tag{3.12}$$

where the last terms approximates the integral as the sum over all stochastic simulations for a given process. The previous Monte Carlo approximation can be used to estimate the selection efficiency ϵ_j, as defined in Equation 3.7, after any deterministic event selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$:

$$\epsilon_j = \mathop{\mathbb{E}}_{x\sim p_j(\boldsymbol{x}|\boldsymbol{\theta})}[\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})] = \int \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) p_j(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \approx \frac{1}{m} \sum^{\boldsymbol{x}_s \in S_j} \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) \tag{3.13}$$

which simply corresponds to the number of simulated observations that pass the selection divided by the total number of simulated observations m. Lastly, the expected value of any measurable function $f(\boldsymbol{x})$ after a given event selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ for events generated by a given process j can be approximated by:

$$\mathop{\mathbb{E}}_{x\sim g_j(\boldsymbol{x}|\boldsymbol{\theta})}[f(\boldsymbol{x})] = \frac{1}{\epsilon_j} \int f(\boldsymbol{x}) \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) p_j(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \approx \frac{1}{\epsilon_j m} \sum^{\boldsymbol{x}_s \in S_j} f(\boldsymbol{x}_s) \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) \tag{3.14}$$

which corresponds to the mean of $f(\boldsymbol{x})$ for all the events that passed the selection, noting that if all the events passed the selection (i.e. $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) = 1$), then Equation 3.12 would be recovered.

While we have been dealing independently with the estimation of arbitrary expected values for a given mixture component j, the computation of expected values of any measurable function $f(\boldsymbol{x})$ under the total mixture distribution can be easily be expressed as function of expectations of mixture components:

$$\mathop{\mathbb{E}}_{x\sim g(\boldsymbol{x}|\boldsymbol{\theta})}[f(\boldsymbol{x})] = \int f(\boldsymbol{x}) \sum_{j=0}^{K-1} \chi_j g_j(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \approx \sum_{j=0}^{K-1} \chi_j \mathop{\mathbb{E}}_{x\sim g_j(\boldsymbol{x}|\boldsymbol{\theta})}[f(\boldsymbol{x})] \tag{3.15}$$

where $\chi_j = \phi_j \epsilon_j / \sum_{j=0}^{K-1} \phi_j \epsilon_j$ is the mixture fraction after selection (see Equation 3.7). While the problem of estimation of expected values might seem unrelated to the inference problem at hand, in Chapter 3.1.3 it will become evident that the construction of non-parametric likelihoods of summary statistics can be reduced to the estimation of expected values.

Oftentimes, the simulated observations are generated using a somewhat different probability distribution than that of experimental data, maybe because some of the generating parameters are not known precisely beforehand (e.g. the properties of pileup interactions). Alternatively, we might want to use a single set of simulated observations to realistically model observables corresponding to a different value of the parameters $\boldsymbol{\theta}$ or even to compute observables under a different process j. Let us suppose that the samples were generated under $p_Q(\boldsymbol{x}|\boldsymbol{\theta}_Q)$ while we want to model samples under $p_R(\boldsymbol{x}|\boldsymbol{\theta}_R)$. In that case, if both distributions have the same support, we can express the expectation value under the desired distribution as:

$$\mathop{\mathbb{E}}_{x\sim p_R(\boldsymbol{x}|\boldsymbol{\theta}_Q)}[f(\boldsymbol{x})] = \frac{\int f(\boldsymbol{x}) \frac{p_R(\boldsymbol{x}|\boldsymbol{\theta}_R)}{p_Q(\boldsymbol{x}|\boldsymbol{\theta}_Q)} p_Q(\boldsymbol{x}|\boldsymbol{\theta}_Q) d\boldsymbol{x}}{\int \frac{p_R(\boldsymbol{x}|\boldsymbol{\theta}_R)}{p_Q(\boldsymbol{x}|\boldsymbol{\theta}_Q)} p_Q(\boldsymbol{x}|\boldsymbol{\theta}_Q) d\boldsymbol{x}} \approx \frac{\sum^{\boldsymbol{x}_s \in S_j} w(\boldsymbol{x}_s) f(\boldsymbol{x}_s)}{\sum^{\boldsymbol{x}_s \in S_j} w(\boldsymbol{x}_s)} \tag{3.16}$$

which is analogous to what was done in Equation 3.12, but accounting for a weight $w(\boldsymbol{x}_s) = p_R(\boldsymbol{x}_s|\boldsymbol{\theta}_R)/p_Q(\boldsymbol{x}_s|\boldsymbol{\theta}_Q)$ for each simulated observation. This technique can be also used together with an arbitrary event selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ simply by considering as event weight the product $w_{\mathcal{C}}(\boldsymbol{x}_s) = \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) w(\boldsymbol{x}_s)$, which amounts to summing over the selected events. In particle physics experiments, the probability distribution functions $p_Q(\boldsymbol{x}|\boldsymbol{\theta}_R)$ and $p_Q(\boldsymbol{x}|\boldsymbol{\theta}_R)$ are most likely intractable, thus estimation of $w_{\mathcal{C}}(\boldsymbol{x}_s)$ has either to be carried out by non-parametric density estimation in a lower dimensional-space of the detector readouts (discussed in Section 3.1.3) or by directly estimating the density ratio via probabilistic classification as will be discussed in Chapter 4.

As previously mentioned, an advantage of using simulated observations is that the latent variables $\mathcal{H}_j = \{\boldsymbol{z}_0, ..., \boldsymbol{z}_m\}$ for a given simulated set of observations $S_j = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1}\}$ are known. This allows to rewrite the weight $w(\boldsymbol{x}_s, \boldsymbol{z}_s)$ for a given event as the ratio of joint distributions:

$$w(\boldsymbol{x}_s, \boldsymbol{z}_s) = \frac{p_R(\boldsymbol{x}_s, \boldsymbol{z}_s|\boldsymbol{\theta}_R)}{p_Q(\boldsymbol{x}_s, \boldsymbol{z}_s|\boldsymbol{\theta}_Q)} = \frac{p_R(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d}) p_R(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s}) p_R(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p}) p_R(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}_R)}{p_Q(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d}) p_Q(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s}) p_Q(\boldsymbol{z}_\mathrm{s}|\boldsymbol{z}_\mathrm{p}) p_Q(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta}_Q)} \tag{3.17}$$

where the last term is an expansion of each joint distribution as a product of the conditional distributions discussed in Equation 3.9. If the difference between $p_R(\boldsymbol{x}|\boldsymbol{\theta}_R)$ and $p_Q(\boldsymbol{x}|\boldsymbol{\theta}_Q)$ is contained in one of the factors of the joint distribution, which is often the case, most of the factors in Equation 3.17 cancel out and we are left with a much simpler problem of density ratio estimation in the latent space. This is often what is done to model the effect of a different pileup distribution or alternative parton distribution functions, further factoring the joint distribution to include explicit dependencies with respect to $\boldsymbol{z}_{\text{pileup}}$ or $\boldsymbol{z}_{\text{PDF}}$, as done in Equation 3.11 and Equation 3.10 respectively. The case when the difference between distributions is contained in a subset of the parton-level latent variables is one of special relevance, because the event weight for a given event $w(\boldsymbol{z}_s)$ can be expressed as the ratio:

$$w(\boldsymbol{z}_s) = \frac{p_R(\boldsymbol{z}_{\text{p}}|\boldsymbol{\theta}_R)}{p_Q(\boldsymbol{z}_{\text{p}}|\boldsymbol{\theta}_Q)} \tag{3.18}$$

which is referred to as *generator-level re-weighting*, a procedure that in some cases can even be done analytically. The concept of *re-weighting* will be useful to model different parameter points in Chapter 5 with a single set of simulated observations as well as to understand how the effect of varying parameters can be modelled via differentiable transformations in Chapter 6.

3.1.3 Dimensionality Reduction

In the previous overview of the basic statistical modelling principles of experimental high-energy physics, the structure and properties of the probability distribution of the full detector readout $\boldsymbol{x} \in \mathcal{X}$ have been considered. The consideration of the detector readout as single observable variable $\boldsymbol{x}$ in the generative model greatly simplifies the modelling narrative, plus also allows to include the effect of any arbitrary event selection as a deterministic function $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$. Nevertheless, the high-dimensionality of the readout space $\boldsymbol{x} \in \mathcal{X}$ (i.e. $\mathcal{O}(10^8)$) complicates its direct use when comparing simulated and recorded observations, which is crucial when carrying out any type of statistical inference.

The high-dimensionality of the raw detector readout space $\boldsymbol{x} \in \mathcal{X}$ also makes it very difficult to specify an effective event selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ that is able to reduce the contributions from non-interesting or not well-modelled background processes. This motivates the use of a dimensionality reduction function $\boldsymbol{f}(\boldsymbol{x}) : \mathcal{X} \longrightarrow \mathcal{Y}$, from the raw detector readout space $\mathcal{X} \subseteq \mathbb{R}^d$ to a lower dimensional space $\mathcal{Y} \subseteq \mathbb{R}^b$. Here $\boldsymbol{f}(\boldsymbol{x})$

represents any deterministic function of the detector readout, but in practice it can be implemented by a series of consecutive transformations.

Let us denote as $\boldsymbol{y} \in \mathcal{Y}$ the resulting variable after the transformation $\boldsymbol{f}(\boldsymbol{x})$ is applied to the observed detector readout. If the function $\boldsymbol{f}$ is differentiable and bijective (i.e. there is a one-to-one correspondence between $\boldsymbol{x}$ and $\boldsymbol{y}$), the probability density distribution function of $\boldsymbol{y}$ could be obtained as:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta}) \left| \det \frac{d\boldsymbol{x}}{d\boldsymbol{y}} \right| \tag{3.19}$$

where the last term is the Jacobian determinant of the inverse of $\boldsymbol{f}$. The transformations commonly used in particle colliders are non-bijective and sometimes non-differentiable, plus Equation 3.19 is in any case of little use when $p(\boldsymbol{x}|\boldsymbol{\theta})$ is intractable. However, the expectation value of $\boldsymbol{y}$ as well any other deterministic transformation of the detector readout $\boldsymbol{x}$ after any arbitrary event selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ can be obtained using simulated samples for a given interaction process as shown in Equation 3.14, independently of whether the transformation is invertible or differentiable. In the rest of this section, the main procedures followed to reduce the dimensionality of the observable space and its objectives from a statistical perspective will be discussed.

Event Reconstruction

The methods of event reconstruction, as described in Section 2.3.3, provide a very efficient way to transform the high-dimensional detector readout to a lower-dimensional space that can more easily be interpreted from a physical standpoint. In fact, reconstruction can be viewed as a complex procedural technique of inference on a subset of the latent variables given the detector readout $\boldsymbol{x}$ of an event. These methods attempt to walk back the generative chain described in Equation 3.9 to recover the subset of the parton-level $\boldsymbol{z}_{\mathrm{p}}$ (and $\boldsymbol{z}_{\mathrm{s}}$ or $\boldsymbol{z}_{\mathrm{d}}$ in some cases) that strongly depends on the detector readouts, providing a compressed summary of the information in the event about the parameters of interest $\boldsymbol{\theta}$. The dimensionality of the output of the reconstruction procedure $\boldsymbol{y}_{\mathrm{reco}}$ depends on the subset of variables considered for each physical object, which typically amounts to a total of $\mathcal{O}(100)$ dimensions, which is a significant reduction from $\dim(\mathcal{X}) \approx \mathcal{O}(10^8)$.

Due to the detector noise and characteristics, the reconstruction function $\boldsymbol{f}_{\mathrm{reco}}(\boldsymbol{x}) : \mathcal{X} \longrightarrow \mathcal{Y}_{\mathrm{reco}}$ cannot fully recover $\boldsymbol{z}_{\mathrm{p}} \in \mathcal{Z}_{\mathrm{p}}$. This is the case for neutrinos that leave the detector undetected, when the measured four-momentum of a given particle differs from the real value, or when the reconstructed particle does not even exist in $\boldsymbol{z}_{\mathrm{p}}$.

Simulated events can then be used to make calibrated probabilistic statements of the resulting reconstructed physical objects and their relation with the actual unobserved particles going through the detector. Particle identification (e.g. jet b-tagging) and fine-tuned momentum regressions on the reconstructed objects can also be thought of as inference of latent variables, which amounts to using the additional detector information around an object to measure more precisely its properties. These properties include the type of particle that produced the detector readouts clustered for particle identification, or a more precise determination of the momentum for particle regression.

One aspect of the generative model that complicates both reconstruction and statistical inference which has not been discussed yet is that efficient representations of the latent space of simulated events are not easily represented as a fixed-size real vector $\boldsymbol{z} \in \mathcal{Z} \subseteq \mathbb{R}^b$. Let us consider as an example the parton-level latent information $\boldsymbol{z}_\mathrm{p}$, which amounts to a short list of produced particles. The total number of particles and the number of particles of each type are variable, thus $\boldsymbol{z}_\mathrm{p}$ is better represented by a set (or several sets, one for each particle type):

$$\boldsymbol{z}_\mathrm{p}^\mathrm{set} = \{\boldsymbol{z}_\mathrm{p}^i \mid i \in \{1, ..., n_\mathrm{p}\}\} \tag{3.20}$$

where n_p is the total number of particles produced at parton-level and $\boldsymbol{z}_\mathrm{p}^\mathrm{i}$ are the latent variables associated to each particle (i.e. type, four-momenta, charge, colour and spin). A similar set structure can be attributed to latent variables describing long-lived particles after the parton-shower $\boldsymbol{z}_\mathrm{s}$, while additional variables might be associated to each particle (e.g. production vertex) and total number and type diversity would be considerably larger. Because the number of particles and their type greatly varies between different interaction processes, the mapping this structure to observable variable space is very useful. In fact, the result of general event reconstruction process at CMS can be expressed also as a set of physical objects:

$$\boldsymbol{y}_\mathrm{reco}^\mathrm{set} = \{\boldsymbol{y}_\mathrm{reco}^\mathrm{i} \mid i \in \{1, ..., n_\mathrm{reco}\}\} \tag{3.21}$$

where n_reco is the total number of particles, $\boldsymbol{y}_\mathrm{reco}^\mathrm{i}$ are the reconstructed variables for each physical object (i.e. reconstructed type, reconstructed four-momenta, reconstructed charge and any other reconstructed attributes). The calibration between the reconstructed physical objects $\boldsymbol{y}_\mathrm{reco}^\mathrm{set}$ and the actual particles produced in the collision $\boldsymbol{z}_\mathrm{p/s}^\mathrm{set}$ hence amounts to matching set elements (typically based on a ΔR

distance criterion, see Section 2.2.1) and the comparison of their reconstructed and generated attributes.

The fact that both reconstructed and latent spaces have a variable-size set structure greatly complicates the application of inference and learning techniques directly based on $\boldsymbol{y}_{\text{reco}}^{\text{set}}$, because they often can only deal with a fixed-size vector of real numbers $\mathbb{R}^b$. Similarly to what is done for event selection, often the elements in the set of reconstructed objects in an event are reduced by imposing a given condition based on their attributes (e.g. type, isolation or momenta). There exist naive ways to embed a set such as $\boldsymbol{y}_{\text{reco}}^{\text{set}}$ as a fixed-size vector $\mathbb{R}^b$, such as taking the relevant attributes of the first n_{sel} objects according to a specific ordering convention after a given *object selection* and possibly padding with zeros or alternative numerical values the elements that do not exist for a given event. Some of the newer machine learning techniques that will be presented in Chapter 4 can deal with variable-size input, such as sequences, sets or graphs inputs, by *embedding* them in vector representations internally, providing new ways to deal with the mentioned representational issue.

Summary Statistics

The attributes of the subset of reconstructed objects selected in an event for a given analysis, often as a fixed-size vector representation $\boldsymbol{y}_{\text{sel}} \in \mathcal{Z}_{\text{sel}} \subseteq \mathbb{R}^b$, are often still too high-dimensional to be considered directly for statistical inference. The effectiveness of the likelihood-free techniques that will be presented later in this chapter strongly depend on the dimensionality of the observable space considered. Hence, it is desirable to further combine the reconstructed outputs in a lower dimensional *summary statistic*, which can be either a function of each single observation or a set of multiple observations, so simpler statistical models that relate the parameters of interest with the observations can be constructed.

Until now, we have been dealing with the problem of how a single event is distributed $p(\boldsymbol{x}|\boldsymbol{\theta})$, however in practice a collection $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$ of events is considered for inference. Let us first consider again the set D, before any trigger or event selection, similarly to what was done at the beginning of Section 3.1.1. Because of the independence between events, the probability density of a given set D can be expressed as the product of individual probability densities for each event $\boldsymbol{x}_i$:

$$p(D|\boldsymbol{\theta}) = \prod_{\boldsymbol{x}_i \in D} p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{3.22}$$

where $p(\boldsymbol{x}_i|\boldsymbol{\theta})$ can only be modelled realistically by forward simulation, and has the mixture model structure and latent factorisation discussed before. After an arbitrary event selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$, only a subset of events $D_{\mathcal{C}} = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_{n_{\mathcal{C}}}\} \subseteq D$ remain. These events are also independent, so their probability density can be expressed as:

$$g(D_{\mathcal{C}}|\boldsymbol{\theta}) = \prod_{\boldsymbol{x}_i \in D_{\mathcal{C}}} g(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{3.23}$$

where the dependence between the distribution function after the event selection $g(\boldsymbol{x}_i|\boldsymbol{\theta})$ and that before $p(\boldsymbol{x}_i|\boldsymbol{\theta})$ was already described in Equation 3.7. If we are only focussed on the probability distribution of the events in $D_{\mathcal{C}}$, we would be neglecting an important quantity that can also provide information about the parameters of interest: the total number of events that pass the event selection $n_{\mathcal{C}}$. Because this quantity depends on the set of recorded readouts $\mathcal{D}$, where each individual readout $\boldsymbol{x}_i$ is assumed to be an independent and identically distributed variable, the total number of selected events n_C after a deterministic selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ can be modelled using a binomial distribution:

$$p(n_{\mathcal{C}}|n, \boldsymbol{\theta}) = \text{Binomial}(n, \epsilon) \approx \text{Poisson}(n\epsilon) \tag{3.24}$$

where n is the total number of events, and the dependence on the parameters is contained in the total efficiency ϵ, i.e. probability $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})$ of passing the selection criteria, that can be defined as $\epsilon = \int \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{\theta})$. The Poisson approximation is justified because the number of trials n is sufficiently large (i.e. 40 million bunch crossings per second) and the total selection efficiencies $\epsilon \leq 0.000025$ already at trigger level, as discussed in Section 2.2.7. This type of stochastic process is also referred to in the literature as multi-dimensional homogenous Poisson point process [88]. The expected value of n_C coincides with the Poisson mean $n\epsilon$. It could be more intuitively linked with the parameters of interest θ by making explicit the contributions from the different mixture processes:

$$\mathop{\mathbb{E}}_{D \sim p(D|\boldsymbol{\theta})}[n_{\mathcal{C}}] = n \sum_{j=0}^{K-1} \phi_j \mathop{\mathbb{E}}_{x \sim p_j(\boldsymbol{x}|\boldsymbol{\theta})}[\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})] = n \sum_{j=0}^{K-1} \phi_j \epsilon_j \tag{3.25}$$

where the efficiency for each process $\epsilon_j = \int \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) p_j(\boldsymbol{x}|\boldsymbol{\theta})$ can be estimated using simulated observations as shown in Equation 3.13. In principle, all possible processes j that could occur have to be considered, i.e. cases when no hard collision occurred as well as the inclusive contribution of each possible hard process, as described in

Equation 3.4. However, if the product of the expected probability of a given process occurring ϕ_j and the event selection efficiency ϵ_j is low enough relative to the total efficiency $\epsilon = \sum_{j=0}^{K-1} \phi_j \epsilon_j$, the effect of those mixture components can be safely neglected.

The situation discussed above is often the case for events where no hard collision occurred after some basic event selection, that is $\epsilon_{\text{not-hard}} \approx 0$ so it can thus be neglected. For the subset of bunch crossings where hard interactions occur, the probability of a given type of interaction before any event selection might be expressed as the product of its cross section σ_j by the total integrated luminosity during the data taking period $\mathcal{L}_{\text{int}}$ divided by the total number of bunch crossings, thus the expected value for number of observations $n_{\mathcal{C}}$ after an event selection that reduces to a negligible fraction the contribution of non-hard processes $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$ can also be expressed as:

$$\mathop{\mathbb{E}}_{D\sim p(D|\boldsymbol{\theta})}[n_{\mathcal{C}}] = n \sum_{j=0}^{K-1} \frac{\mathcal{L}_{\text{int}}\sigma_j}{n}\epsilon_j = \mathcal{L}_{\text{int}} \sum_{j=0}^{K-1} \sigma_j \, \epsilon_j \tag{3.26}$$

where $n_j = \mathcal{L}_{\text{int}} \, \sigma_j \, \epsilon_j$ is the expected number of events coming from a given process j, that can be estimated with theoretical input regarding σ_j, simulated observations to estimate ϵ_j and an experimental measurement of the luminosity $\mathcal{L}$.

The number of observations $n_{\mathcal{C}}$ that pass a given event selection $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$, which normally includes trigger and some additional analysis dependent selection, is the quantity that serves as the basis of the simplest statistical model used in particle physics to link theoretical parameters and observations. This type of summary statistic is very effective when the parameter of interest is the cross section of a single process σ_S and the rest of background processes are well modelled by theoretical predictions and simulated observations. In that case, if all parameters but σ_S are known, a *cut-and-count* sample-based likelihood can be built based on Equation 3.24, corresponding to the following probability density function:

$$p(n_{\mathcal{C}}|\sigma_S) = \text{Poisson}\left(\sigma_s\epsilon_s + \sum^{j\in B} \sigma_j\epsilon_j\right) \tag{3.27}$$

which can be used to carry out statistical inference about σ_S given an observed number of events that pass the event selection $n_{\mathcal{C}}^{\text{obs}}$, using classical techniques.

The previous concept can be applied to several disjoint subsets of $\mathcal{X}$ simultaneously $T = \{\mathcal{C}_0, ..., \mathcal{C}_b\}$, each characterised by a different indicator function $\mathbb{1}_{\mathcal{C}_t}(\boldsymbol{x})$ defining an arbitrary event selection, as long as their intersection is null. The probability

function for the variable $\boldsymbol{n}_T = \{n_{\mathcal{C}_0}, ..., n_{\mathcal{C}_b}\}$, given that each $n_{\mathcal{C}_i}$ is independent, can be obtained as:

$$p(\boldsymbol{n}_T|\boldsymbol{\theta}) = \prod^{\mathcal{C}_i \in T} \text{Poisson}\left(\sum^{j \in H} n_j^{\mathcal{C}_i}(\boldsymbol{\theta})\right) \tag{3.28}$$

where $n_j^{\mathcal{C}_i}(\boldsymbol{\theta})$ is the expected number of observed events coming from process j after the selection $\mathcal{C}_i$. As long as a parametrisation of $n_j^{\mathcal{C}_i}(\boldsymbol{\theta})$ exists, which can be often estimated as $n_j^{\mathcal{C}_i}(\boldsymbol{\theta}) = \mathcal{L}\, \sigma_j\, \epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta})$, Equation 3.28 can be used to construct a likelihood to carry out inference on the parameters $\boldsymbol{\theta}$ based on the observed value of the sample summary statistic $\boldsymbol{n}_T^{\text{obs}}$.

Sufficient Statistics

The selection count vector $\boldsymbol{n}_T^{\text{obs}}(D)$, which has not been specified yet, could be also written as a sum over a function $\boldsymbol{n}_T(\boldsymbol{x}) : \mathcal{X} \subseteq \mathbb{R}^d \longrightarrow \mathcal{Y} \subseteq \{0,1\}^b \subset \mathbb{R}^b$ applied for each event in $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$ as follows:

$$\boldsymbol{n}_T^{\text{obs}}(D) = \sum^{\boldsymbol{x}_i \in D} \boldsymbol{n}_T(\boldsymbol{x}) \tag{3.29}$$

where $\boldsymbol{n}_T^{\text{obs}}(D)$ could be described as a summary statistic of the whole collection of observations while $\boldsymbol{n}_T(\boldsymbol{x}_i)$ summarises a single event $\boldsymbol{x}_i$.

There are infinite ways to choose a lower-dimensional summary statistic of the detector readout $\boldsymbol{s}(\boldsymbol{x}) : \mathcal{X} \subseteq \mathbb{R}^d \longrightarrow \mathcal{Y} \subseteq \mathbb{R}^b$. Functions of the type $\boldsymbol{n}_T(\boldsymbol{x})$ are only a reduced subset, yet still infinite, of the possible space of functions. Regardless of the likelihood-free inference methods considered (see Section 3.2), the need of a low-dimensional summary statistic can be thought as an effective consequence of the *curse of dimensionality*, because the number of simulated observations required to realistically model the probability density function or compute useful distance measures rapidly increases with the number of dimensions.

In general, the selection of a summary statistic $\boldsymbol{s}(\boldsymbol{x})$ is far from trivial, and naive choices can lead to large losses of useful information about the parameters of interest $\boldsymbol{\theta}$. From classical statistics, we can define a *sufficient summary statistic* as the function of the set of observations that can be used for carrying out inference about the model parameters $\boldsymbol{\theta}$ of a given statistical model in place of the original dataset without losing information [89]. Such a sufficient statistic contains all the information in the observed sample useful to compute any estimate on the model parameters, and no complementary statistic can add any additional information about $\boldsymbol{\theta}$ contained in

the set of observations. Sufficient statistics can be formally characterised using the Fisher-Neyman factorisation criterion, which states that a summary statistic $\boldsymbol{s}(\boldsymbol{x})$ is sufficient for the parameter vector $\boldsymbol{\theta}$ if and only if the probability distribution function of $\boldsymbol{x}$ can be factorised as follows:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = q(\boldsymbol{x})r(\boldsymbol{s}(\boldsymbol{x})|\boldsymbol{\theta}) \tag{3.30}$$

where $q(\boldsymbol{x})$ is a non-negative function that does not depend on the parameters and $r(\boldsymbol{x})$ is also a non-negative function for which the dependence on the parameters $\boldsymbol{\theta}$ is a function of the summary statistic $\boldsymbol{s}(\boldsymbol{x})$. The identity function $\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{x}$ is a sufficient summary statistic according to the theorem in Equation 3.30, however we are only interested in summaries that reduce the original data dimensionality without losing of useful information about the parameters $\boldsymbol{\theta}$.

The definition of sufficiency can also be applied to a set of observations $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$. In fact if we assume they are independent and identically distributed, and $\boldsymbol{s}(\boldsymbol{x})$ is sufficient for each observation $\boldsymbol{x}_i$, we may rewrite Equation 3.22 as:

$$p(D|\boldsymbol{\theta}) = \prod^{\boldsymbol{x}_i \in D} q(\boldsymbol{x}) \prod^{\boldsymbol{x}_i \in D} r(\boldsymbol{s}(\boldsymbol{x}_i)|\boldsymbol{\theta}) = q(D)r(\boldsymbol{s}(D)|\boldsymbol{\theta})$$

where the set of sufficient summary statistics for each observation is a sufficient summary statistic for the whole dataset $\boldsymbol{s}(D) = \{\ \boldsymbol{s}(\boldsymbol{x}_i) \mid \forall \boldsymbol{x}_i \in D\}$ and the dependence on the summary statistic is contained as the product of independent factors for each observation.

Because $p(\boldsymbol{x}|\boldsymbol{\theta})$ is not available in closed form in particle collider experiments, the general task of finding a sufficient summary statistic that reduces the dimensionality cannot be tackled directly by analytic means. However, for finite mixture models where the only model parameters are a function of the mixture coefficients ϕ_j, probabilistic classification can be used to obtain (approximate) sufficient summary statistics. We will return to this topic in Chapter 4. When the parameters of interest or additional unknown parameters affect the mixture components $p_j(\boldsymbol{x}|\boldsymbol{\theta})$, the construction of sufficient summary statistics cannot be addressed directly, thus a fraction information about the parameters $\boldsymbol{\theta}$ is very likely to be lost in the dimensionality reduction step. An automated way to obtain powerful summary statistics in those cases using machine learning techniques will be presented in Chapter 6.

Synthetic Likelihood

The advantage of using lower-dimensional summary statistics $\boldsymbol{s}(D) : \mathcal{X}_D \subseteq \mathbb{R}^{d \times n} \longrightarrow \mathcal{Y}_D \subseteq \mathbb{R}^{b \times n}$ of the detector readout collected by the experiment is that often the generative model of $p(\boldsymbol{x}|\boldsymbol{\theta})$ can be used to build non-parametric likelihoods of $s(D)$ that link the observations with the model parameters, so classical inference algorithms can be used. This likelihoods are referred here as synthetic because they are not based on the actual generative model of $\boldsymbol{x}$ but on approximations constructed using simulated observations.

For summary statistics of the type $\boldsymbol{n}_T^{\text{obs}}(D) : \mathcal{X}_D \subseteq \mathbb{R}^{d \times n} \longrightarrow \mathcal{Y}_D \subseteq \{0,1\}^b$ the likelihood can be expressed as a product of independent Poisson count likelihoods as shown in Equation 3.28. Such likelihood can be evaluated for the observed data D and specific parameters $\boldsymbol{\theta}_R$, even in the case that $\boldsymbol{\theta}$ modifies the distribution of the mixture components $p_j(\boldsymbol{x}|\boldsymbol{\theta})$, by forward approximating $n_j^{\mathcal{C}_i}(\boldsymbol{\theta}_R)$ (or alternatively $\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_R)$) using simulated observations for each process j generated for $\boldsymbol{\theta}_R$. This process would rapidly become computationally very demanding if it had to be repeated for each likelihood evaluation during the whole inference process. Re-weighting procedures such as those described in Equation 3.18 can often be applied to re-use already simulated events using $\boldsymbol{\theta}_R$ to model events corresponding to different values of the parameters $\boldsymbol{\theta}_Q$.

A more economical approach, commonly used in LHC analyses that use binned Poisson likelihoods based on the formalism introduced in Equation 3.28, is to parametrise the effect of varying parameters by interpolating between the values of the $\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_k)$ (or directly $n_j^{\mathcal{C}_i}(\boldsymbol{\theta}_k)$) for different values of k. Such parametrisation allows the analytical approximation of the likelihood originated by Equation 3.28, and simplifies the computation of gradients with respect to the parameters. This is particularly relevant to model the effect of *nuisance parameters*, which are uncertain but not of direct interest, and have to be accounted for in the inference procedure; this issue will be discussed in Section 3.1.4. Different interpolation conventions exist [90], but they are normally based on the marginal one-dimensional interpolation between the effect of a single parameter $\theta_i \in \boldsymbol{\theta}$ at three equally spaced values (the nominal parameter values and the up/down variations). In that case the total effect on $\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_k)$ is accounted by adding absolute shifts or multiplying marginal effects.

Even assuming that the marginal description when a single parameter of interest varies is accurate, which is not ensured by the interpolation, and the effect of each parameter is factorised in $p_j(\boldsymbol{x}|\boldsymbol{\theta})$, the integral definition of $\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_k)$ from Equation 3.13 does not ensure that the correlated effect of the variation of multiple $\theta_i \in \boldsymbol{\theta}$ is

accurately modelled. This issue can be easily exemplified, considering the product of relative variations in the two parameter case $\boldsymbol{\theta}_R = (\theta_0^R, \theta_1^R)$. Let us consider the expected value for the efficiency after a given selection $\mathbb{1}_{\mathcal{C}_i}(\boldsymbol{x})$:

$$\begin{aligned} \epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_R) &= \int \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) p_j(\boldsymbol{x}|\boldsymbol{\theta}_R) d\boldsymbol{x} \\ &= \int \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) p_j(\boldsymbol{x}|\boldsymbol{\theta}_Q) \frac{p_j(\boldsymbol{x}|(\theta_0^R, \theta_1^Q))}{p_j(\boldsymbol{x}|\boldsymbol{\theta}_Q)} \frac{p_j(\boldsymbol{x}|(\theta_0^Q, \theta_1^R))}{p_j(\boldsymbol{x}|\boldsymbol{\theta}_Q)} d\boldsymbol{x} \end{aligned} \tag{3.31}$$

where $\boldsymbol{\theta}_R$ is the parameter point we want to simulate by interpolating around a nominal point $\boldsymbol{\theta}_Q$. The last expression in Equation 3.31 is only correct when the effect of each parameter is independent, i.e. the underlying probability density function can be factorised as the product of independent factors. However, it becomes evident that the previous expression does not simplify:

$$\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_R) \neq \epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_Q) \frac{\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_R)}{\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_Q)} \frac{\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_R)}{\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta}_Q)} \tag{3.32}$$

because the integral of the product of functions is not product of integrals, unless the volume of the selected region C is infinitesimally small - an irrelevant case as it would correspond to null efficiencies. This effect also applies if additive variations are considered and can be more notable when more parameters are considered.

The previously mentioned modelling issue, even though to the best of our knowledge has not been made explicit in the literature before, affects a multitude of analyses at the LHC, i.e. those that use *template interpolation*, as implemented in the standard statistical libraries used in particle physics experiments [91, 92]. A possible solution would include doing a multi-dimensional interpolation, but it would naively require evaluating at least all 3-point combinatorial variations of the parameters, amounting to a minimum of 3^p evaluations of $\epsilon_j^{\mathcal{C}_i}(\boldsymbol{\theta})$, where p is the number of parameters. If the effect of the parameters can be factored out in the joint distribution and the same simulated event set can be modified to describe each marginal variation, as reviewed around Equation 3.17, the non-marginal terms can be estimated from the product of per-event marginal terms by considering the finite sum approximation of the last expression in Equation 3.31, which would only require $(2p + 1)$ parameter variation evaluations. Alternatively, the basis of the approach presented in Chapter 6, where the variation of the parameters and its derivatives are computed in place over the simulated observations by specifying the full computational graph,

could also be used in analyses where the discussed assumption fails to realistically describe the data.

3.1.4 Known Unknowns

So far we have assumed that the simulated observations can model the data and the only parameters $\boldsymbol{\theta}$ that affect the generative model are those we are interested in carrying out inference on. However, simulated observations effectively depend on the modelling of the physical processes occurring in the proton-proton collisions and the detector, of which we only have an approximate description. Those mis-modelling effects have to be accounted in the inference procedure to obtain unbiased estimates, and are accounted by additional *nuisance parameters* in the statistical model when their effect is known and can be approximated. For cases where simulation does not provide the desired level of accuracy, the contribution from some of the mixture components can often be estimated from data directly, using what are referred to as *data-driven estimation* techniques.

Nuisance Parameters

The general definition of nuisance parameters in a statistical model refers to all the uncertain parameters of the statistical model that are not of intermediate interest but have to be accounted for in the inference procedure. These parameters can include uncertain theoretical parameters (e.g. top quark mass or expected background rate), account for limitation on the experimentally measured parameterisations of certain phenomena (e.g. parton density functions uncertainties) or represent the accuracy limits of calibration between data and simulation. Nuisance parameters can also represent additional degrees of freedom of the model that cover for possible wrong assumptions or quantify imprecisions due to the limited number of simulated observations.

Because the actual generative process for the experimental data is not known perfectly, the simulation-based model is extended with additional parameters that portray the possible variability on the distribution of the detector readouts. The formalism developed in the previous part of Section 3.1 still applies, noting that the parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\iota, \boldsymbol{\theta}_\nu\}$ now includes both parameters of interest $\boldsymbol{\theta}_\iota$ and nuisance parameters $\boldsymbol{\theta}_\nu$. While the effect of (usually theoretical) parameters of interest typically only affects the parton-level latent factor $p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})$, some nuisance parame-

ters account for possible mis-modelling in subsequent steps of the simulation thus can affect the other factors in Equation 3.9.

The effect of variation of nuisance parameters for any observable or summary statistic considered in a given analysis can be estimated by simulating again the affected observation with the chosen parameters - often prohibitively expensive - or by re-weighting already simulated observations as described in Equation 3.17 - which is much faster and reduces the statistical fluctuations between variations associated with the random sampling of the full latent space. Unprincipled modelling shortcuts, such as considering the additive or multiplicative effect of marginal efficiencies to account for combined effects, are also frequently used for count vector observables $n_j^{\mathcal{C}_i}(\boldsymbol{\theta})$, as discussed in Section 3.1.3 together with possible solutions to some of the associated issues.

The re-weighting approach from Equation 3.16 is extremely effective to model the effect of parameters in the conditional factor that deal with low-dimensional latent variables, such as $p(\boldsymbol{z}_\mathrm{p}|\boldsymbol{\theta})$, because the rest of the factors in the joint distribution simplify and we are left with a low-dimensionality density estimation problem (even analytically tractable in some cases). For conditional factors that deal with higher dimensional latent or observable spaces, such as $p(\boldsymbol{z}_\mathrm{d}|\boldsymbol{z}_\mathrm{s}, \boldsymbol{\theta})$ or $p(\boldsymbol{x}|\boldsymbol{z}_\mathrm{d}, \boldsymbol{\theta})$, the ratio can be very hard to estimate unless additional simplifications are possible. For those nuisance parameters, it is easier to consider the effect on the lower-dimensional summary statistic instead of the detector readout x, because the ratio:

$$w(\boldsymbol{s}(\boldsymbol{x})) = \frac{p_R(\boldsymbol{s}(\boldsymbol{x})|\boldsymbol{\theta}_R)}{p_Q(\boldsymbol{s}(\boldsymbol{x})|\boldsymbol{\theta}_Q)} \tag{3.33}$$

can be simpler to estimate through density estimation or approximately factorise if the summary statistic is chosen carefully. This fact motivates an alternative way to model the effect of some of the nuisance parameters, especially those related with the differences in the reconstructed objects observables between simulation and data after calibration. Let us consider the case where summary statistics $\boldsymbol{s}(\boldsymbol{x}) : \mathcal{X} \subseteq \mathbb{R}^d \longrightarrow \mathcal{Y}_\mathrm{sum} \subseteq \mathbb{R}^b$ are effectively a function of the reconstructed objects and its properties $\boldsymbol{y}_\mathrm{reco} \in \mathcal{Y}_\mathrm{reco}$, which can be schematically represented by the following function composition chain:

$$\mathcal{X} \xrightarrow{g} \mathcal{Y}_\mathrm{reco} \xrightarrow{h} \mathcal{Y}_\mathrm{sum} \tag{3.34}$$

where $\boldsymbol{y}_{\text{reco}} = g(\boldsymbol{x})$ and $\boldsymbol{y}_{\text{sum}} = h(\boldsymbol{y}_{\text{reco}})$. This compositional approach can be extended to include also event selection at trigger or analysis level, or other intermediate summaries of $\boldsymbol{x}$ complementary to reconstruction, as part of the definition of the summary statistic $\boldsymbol{s}(\boldsymbol{x})$. In all cases where $s(\boldsymbol{x})$ is a deterministic function, all differences between simulated observations and data in any expected observables originate from the differences between the simulation-based generative definition of $p(\boldsymbol{x}|\boldsymbol{\theta})$ and the true unknown generative process $p_{\text{true}}(\boldsymbol{x})$. While the task of evaluating and parametrising these differences directly by studying the raw detector output is quite convoluted, the differences can be corrected and their uncertainty assessed for the lower-dimensional intermediate states of the composition chain depicted in Equation 3.34.

For example, if the momenta of a certain subset of the reconstructed objects $\boldsymbol{y}_{\text{reco}}$ statistically differ between experimental data and the simulated observations, based on a subset of the data that is assumed to be well-modelled, the momenta of simulated observations can be corrected to better model the data. The statistical accuracy of such procedure due to the different factors leads to a set of nuisance parameters that describe the limit of the mentioned calibration as a function of the value of $\boldsymbol{y}_{\text{reco}}$. The effect of these type of nuisance parameters often be modelled in the simulation by using a function of the simulated intermediate outputs, e.g. in the case of reconstructed objects:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p(\boldsymbol{x}|\boldsymbol{\theta})}[\boldsymbol{s}(\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{y}_{\text{reco}}\sim p(\boldsymbol{y}_{\text{reco}}|\boldsymbol{\theta}_o)}[h(r(\boldsymbol{y}_{\text{reco}}, \boldsymbol{\theta}_\rho))] \tag{3.35}$$

so $p(\boldsymbol{x}|\boldsymbol{\theta})$ can be approximated by computing observables after applying the re-parametrisation $r(\boldsymbol{y}_{\text{reco}}, \boldsymbol{\theta}_\rho)$ to the simulated observations, where $\boldsymbol{\theta}_\rho$ is the vector of parameters representing the different uncertainty factors.

In general, the effects of all relevant nuisance parameters can be modelled by a combination of simulated observations re-weighting by $w(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}_w)$ and transformations of intermediate simulated observations $\boldsymbol{y}_{\text{new}} = r(\boldsymbol{y}_{\text{sim}}, \boldsymbol{z}_i|\boldsymbol{\theta}_\rho)$. The former is based on importance sampling [93] to estimate the properties of a different distribution than the one sampled originally from, while the latter assumes that the mis-modelling can be accounted by a parametrisation of the simulated intermediate observables. If the functions $w(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}_w)$ and $r(\boldsymbol{y}_{\text{sim}}, \boldsymbol{z}_i|\boldsymbol{\theta}_\rho)$ are differentiable or can be approximated by differentiable functions, the gradient (and higher order derivatives) with respect to the parameters $\boldsymbol{\theta}$ of any expectation value can be very efficiently approximated. This can be very useful for statistical inference (e.g. likelihood maximisation), while

it has not been used so far in LHC analysis to our knowledge. This is one of the core concepts of the technique to construct summary statistics presented in Chapter 6.

The inference results of a given analysis depend strongly on the assumptions implicit in the statistical model. The determination, assessment and practical definition of the effect of nuisance parameters that are relevant for a given analysis is one the most challenging yet important aspects in experimental particle physics at the LHC. When nuisance parameters are quantitatively taken into account in the statistical model, they lead to an increase of the uncertainty on the parameters of interest and larger interval estimates (or exclusion limits) on the parameters of interest. The choice of summary statistics may also affect significantly subsequent inference, and while nuisance parameters are usually qualitatively considered when building simple summary statistics by physics-inspired combinations of reconstructed variables, they are not regarded at all when the automatic multi-variate techniques described in Chapter 4 are applied to construct complex non-linear observables. This issue is addressed by the method proposed in Chapter 6.

Data-Driven Estimation

For some fundamental processes, the generative modelling provided by simulated observations might not be accurate enough for the purposes of a given LHC analysis. In a subset of those cases, the simulated observations can be calibrated to better describe the observations in well-modelled data regions, as mentioned in the previous section. However, if the description of the summary statistics considered in the analysis provided by the simulated observations from process j is substandard, e.g. the number of simulated observations that could be realistically simulated is not sufficient, then the contribution from the mentioned mixture component might have to be estimated from experimental observations directly.

The actual procedure used for modelling the contribution for a given mixture component j from data depend on the specifics of the process as well the details of the analysis, but often includes some re-weighting factor obtained from simulated observations or additional experimental observations with an orthogonal selection criterion. Such data-driven estimation techniques are often used for the background processes, but are hard to combine with the non-linear summary statistics reconstructed by machine learning techniques such as those described in Chapter 4. In the CMS analysis presented in Chapter 5, we describe and utilise a fully data-driven background estimation technique fine-tuned for the modelling of the QCD-based

multiple jet background for the search of Higgs pair production decaying to four b-quarks.

3.2 Statistical Inference

In the previous section, the main characteristics of the generative statistical model $p(D|\boldsymbol{\theta})$ relating the parameters $\boldsymbol{\theta}$ with the set of observations $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$ have been reviewed. In addition, we discussed the role of lower dimensional summary statistics as functional transformations of each detector readout $\boldsymbol{s}(\boldsymbol{x}_i)$ or even the whole dataset $\boldsymbol{s}(D)$, as well as how the effect of additional uncertain parameters can be included in the simulation-based generative model of the data. In this section, we deal with the actual problem of inference of the subset of parameters of interest $\boldsymbol{\theta}_\iota$ once a summary statistic has already been chosen and the final statistical model $p(\boldsymbol{s}(D)|\boldsymbol{\theta})$ has been fully specified.

3.2.1 Likelihood-Free Inference

One of the main properties of the statistical models at particle colliders we focussed on in the last section was their generative-only nature, whereby their probability density $p(\boldsymbol{x}|\boldsymbol{\theta})$ cannot be expressed analytically, but only by means of forward simulated observation. This fact greatly complicates the application of standard inference techniques which require the explicit definition of a likelihood

$$L(\boldsymbol{\theta}|D) = \prod_{\boldsymbol{x}_i \in D} p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{3.36}$$

in order to make quantitative statements about the parameters of interest, because it expresses the extent to which a set of values for the model parameters are consistent with the observed data. Problems where the likelihood cannot be expressed directly are common in many scientific disciplines, because a link between observations and the underlying parameters can often only be provided by a probabilistic computer program. This is frequently the case when the system under study is complex, e.g. can only be described by a hierarchy or a sequence of stochastic processes.

The evaluation of the likelihood for complex generative models rapidly becomes impractical, especially when the dimensionality of the observations or the parameter space is very high. Various statistical techniques for dealing with these cases exist, generally referred to as *likelihood-free* or *simulation-based* inference techniques. A well established group of techniques for inference when the likelihood function is

unknown is referred to as Approximate Bayesian Computation (ABC) [94, 95]. The fundamental concept behind ABC is the generation of a simulated sample $S_0 = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1}\}$ using a given vector of parameters $\boldsymbol{\theta}_0$, which is then compared using a distance criterion to the actual observed dataset D. If the data and the simulation are close enough, then $\boldsymbol{\theta}_0$ is retained as sample from the posterior. The process is repeated until the posterior is estimated with the desired accuracy. The quality of the posterior approximation produced by ABC techniques, as well as the number of sampling steps required to reach a given accuracy, strongly depend on the distance definition. When the dimensionality of the output is high, a summary statistic vector $\boldsymbol{s}(D)$ has to be used in practice to increase the computational efficiency of the previous procedure, which would be otherwise intractable.

The approach commonly used when carrying out inference at particle physics experiments at the LHC is somehow related with the mentioned family of techniques. The observations are also reduced to a lower-dimensional summary statistic space, but then a non-parametric likelihood is constructed so that standard inference techniques can be applied. The likelihood is often based on the product of Poisson count terms, as depicted in Equation 3.27 and Equation 3.28, where the dependence on the expectations on the parameters is based on the simulation and the mixture structure. Alternative approaches include the use of a simple one-dimensional parametrisation for a continuous background and a bump-like signal, which is common when the reconstructed mass of an intermediate object is used as summary statistic and its distribution is well-controlled, e.g. a Higgs bosons decaying to two photons. An additional alternative approach, which has not been used in LHC analyses to date, could be to use non-parametric density estimation techniques to obtain an unbinned likelihood directly from simulated data. This approach has been recently referred as Approximate Frequentist Computation (AFC) [96], and can be also combined with the technique presented in Chapter 6.

3.2.2 Hypothesis Testing

Statistical inference within experimental particle physics is often framed as a hypothesis testing problem. The goal of statistical testing is to make a quantitative statement about how well observed data agrees with an underlying model or prediction, which is often referred to as a *hypothesis*. The statistical model under consideration is often referred to as *null hypothesis* H_0. Classical statistical testing techniques often require the definition of an *alternative hypothesis* H_1, whose agreement with the data is compared with that of the null. A hypothesis is said to be *simple*, when all the

distribution (or generative model) parameters are fully specified, i.e. $p(\boldsymbol{x}|H_s) = f(\boldsymbol{x})$ does not depend on any non-fixed parameter. A *composite* hypothesis instead depends on one or more parameters $\boldsymbol{\theta}$, i.e. the distribution under the hypothesis can be expressed as $p(\boldsymbol{x}|H_c) = f(\boldsymbol{x}, \boldsymbol{\theta})$.

In order to carry out hypothesis testing based on a set of observations $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$, a *test statistic* $t(D)$ that is a function of the observations is constructed. The choice of test statistic is especially challenging when $\boldsymbol{x}$ is high-dimensional and $p(\boldsymbol{x}_i|\boldsymbol{\theta})$ is not known. The concepts of test statistic and summary statistic, the latter discussed in Section 3.1.3, are very related. A test statistic is in fact a sample summary statistic[5] $s(D)$, that is used within an statistical test to accept or reject hypothesis, so all the concerns regarding sufficiency from Section 3.1.3 also apply. Regarding the dimensionality of $t(D) : \mathcal{X}_D \subseteq \mathbb{R}^{d \times n} \longrightarrow \mathcal{T}$, while it can be a multi-dimensional vector (e.g. could even use $t(D) = (\boldsymbol{x}_0, ..., \boldsymbol{x}_n)$), a one dimensional variable is usually considered in order to simplify the process of making calibrated statistical statements.

Let us refer to the test statistic for the set of observations as t_{obs} from here onwards. The result of the statistical test is whether the hypothesis H_0 can be rejected in favour of H_1 if the null is unlikely enough. In practice, in order to make a principled decision, a critical region $\mathcal{T}_C \subseteq \mathcal{T}$ in the space of the test statistic has to be defined before looking at the set of observations. Once the critical region has been chosen, a test can be then characterised by its *significance* level α and *power* $1 - \beta$. The significance, which is also referred to as the *Type I error rate*, is directly related with the probability of rejecting H_0 when it is actually true. For a given test based on the summary statistic $t(D)$ and its critical region $\mathcal{T}_C$, the significance level can be defined as:

$$\alpha = P(t \in \mathcal{T}_C|H_0) = \int_{\mathcal{T}_C} g(t|H_0)dt \overset{\text{1D}}{\Rightarrow} \int_{t_{\mathbf{cut}}}^{\infty} g(t|H_0)dt \tag{3.37}$$

where $g(t|H_0)$ is the distribution of the test statistic under the null hypothesis H_0, and the latter simplification applies for one-dimensional summary statistics where the critical region is defined based on a given threshold t_{cut}. The power of a test $1 - \beta$ is instead defined by the probability of not rejecting the null hypothesis when the alternative is actually true, which often referred as *type II error rate* β. The type

[5] Here a statistic is a function of observations, and *sample summary statistic* refers to statistics the summarise a set of observations $\boldsymbol{s}(D) : \mathcal{X}_D \subseteq \mathbb{R}^{d \times n} \longrightarrow \mathcal{Y}_D \subseteq \mathbb{R}^{b \times n}$.

II error rate β can be defined as the probability of not being in the critical region under the alternative hypothesis:

$$\beta = P(t \notin \mathcal{T}_C | H_1) = 1 - \int_{\mathcal{T}_C} g(t|H_1)dt \overset{\text{1D}}{\Rightarrow} 1 - \int_{-\infty}^{t_{\mathbf{cut}}} g(t|H_1)dt \tag{3.38}$$

where $g(t|H_0)$ is the distribution of the test statistic under the alternative hypothesis H_1, and the last terms corresponds to the one dimensional case based on a threshold. Both the significance level and the power of a hypothesis test depend on the definition of its test statistic and the critical region. The significance level of a test α is often fixed at a given value in order to reject the null in favour of an alternative. It is then beneficial to design the test so its power is as high as possible, which is equivalent to having a Type II error rate as low as possible.

From the definition of Type I and Type II error rates in Equation 3.37 and Equation 3.38, it is evident that either the probability distribution function of the test statistic under both the null and alternative hypothesis or a way to estimate the integrals from simulated observations are required. The main advantage of one-dimensional test statistics, similarly to the low-dimensional summary statistics discussed in Section 3.1.3, is that they allow for an efficient estimation of the probability distribution function using non-parametric techniques. When both the null H_0 and alternative hypothesis H_1 are simple, the Neyman-Pearson lemma [97] states that the *likelihood ratio*, which is a one-dimensional test statistic defined as:

$$\Lambda(\mathcal{D}; H_0, H_1) = \frac{p(D|H_0)}{p(D|H_1)} = \prod_{\boldsymbol{x} \in \mathcal{D}} \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{x}|H_1)} \tag{3.39}$$

is the most powerful test statistic at any threshold t_{cut}, which is associated with a significance $\alpha = P(\Lambda(\mathcal{D}; H_0, H_1) \leq t_{\text{cut}})$. The last expansion requires independence between the different observations. While the likelihood ratio can be proven to be the most powerful test statistic, it cannot be evaluated exactly if the likelihood is not known, which often the case for LHC inference problems as discussed in Section 3.2.1. The alternative hypothesis is usually composite in particle colliders because the signal mixture fraction μ (or its cross section equivalently) is one of the parameters of interest. The likelihood ratio test can nevertheless be expressed in this case as a function the parameter μ, which will be the most powerful test for a given μ if it is the only unknown parameter.

It is worth noting that while the likelihood ratio defined in Equation 3.39 defines the most powerful test, the likelihood ratio based on a summary statistic $\boldsymbol{s}(D)$ can

also be defined, but it is not the most powerful test for inference based on D unless $\boldsymbol{s}(D)$ is a sufficient summary statistic with respect to the parameters $\boldsymbol{\theta}$ which fully define the null $p(\boldsymbol{x}|H_0) = p(\boldsymbol{x}|\boldsymbol{\theta}_0)$ and alternate $p(\boldsymbol{x}|H_1) = p(\boldsymbol{x}|\boldsymbol{\theta}_1)$ hypotheses. This fact motivates the use of machine learning techniques to approximate the likelihood ratio directly based on simulated observations as discussed in Section 4.3.1. The likelihood-ratio can then be calibrated by means of non-parametric probability density estimation techniques or count-based likelihoods.

Another relevant issue when defining test statistics is that hypotheses are rarely simple (or with a composite alternate in the way previously described). Let us suppose the μ is the parameter of interest, e.g. the mixture coefficient for the signal. The statistical model often depends on additional nuisance parameters $\boldsymbol{\theta}$, as discussed in Section 3.1.4. The likelihood ratio from Equation 3.39 is not guaranteed to be the most powerful test statistic when the hypotheses are composite. In this case, often summary statistics based on the *profile likelihood ratio* are used, that can be defined for LHC searches as:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} \tag{3.40}$$

where $\hat{\hat{\boldsymbol{\theta}}}$ at the numerator refers to the value of the nuisance parameter that maximises the likelihood for a given μ, and $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ at the denominator are the standard maximum likelihood estimators. The property that motivates the use of the profile likelihood ratio, other than its convergence to the likelihood ratio when the hypotheses are simple, is that the distribution for large numbers of observations can be effectively approximated, as demonstrated by Wilks and Wald [98, 99].

For a discussion of the different test statistics based on the profiled likelihood ratio as well as their asymptotic approximations, the following reference is recommended [100]. In particular, the use of the *Asimov dataset*, where the observed sample summary statistic of the type outlined Equation 3.29 is assumed to be equal to the expectation, is instrumental for the technique described in Chapter 6. The statistical framework of hypothesis testing is used to decide whether to reject or not reject the null hypothesis in favour of the alternate. Alternatively it can also be useful to estimate the probability of obtaining the observed data (or test statistic) under the null hypothesis, which is simply referred to as the p-value or alternatively as Z-value when standard deviation units are used. When the null hypothesis is not rejected H_0, the statistical test can be recast to obtain *exclusion upper limits* at a given confidence level (usually 95% is used), as is done in the non-resonant Higgs production search included in Chapter 5.

For obtaining exclusion upper limits, it is useful to define a modified test statistic $\widetilde{q}(\mu)$:

$$\widetilde{q}(\mu) = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\boldsymbol{\theta}}}(\mu))} & \text{if } \hat{\mu} < 0 \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\hat{\boldsymbol{\theta}}}(\mu))} & \text{if } 0 \leq \hat{\mu} \leq \mu \\ 0 & \text{if } \hat{\mu} > \mu \end{cases}$$

which does not regard negative background fluctuations or cases where $\hat{\mu} > \mu$ as evidence against μ. When using $\widetilde{q}(\mu)$ or similar profile-likelihood-based one-dimensional test statistics, the observed exclusion upper upper limit can be defined as the largest value of μ for which the probability of obtaining a test statistic value is equal or larger than a given confidence level (e.g. $\alpha = 0.05$ for 95% confidence intervals), which can be expressed as the following integral:

$$P(\widetilde{q}(\mu) \geq \alpha|\mu) = \int_{\widetilde{q}_{\text{obs}}(\mu)}^{\infty} g(\widetilde{q}(\mu)|\mu)dq \tag{3.41}$$

where $\widetilde{q}_{\text{obs}}(\mu)$ is the observed test statistic and $g(\widetilde{q}(\mu)|\mu)$ is the distribution under the alternate when the signal fraction is μ. This integral can be approximated using Monte Carlo simulations or by the asymptotic approximations described in [100]. A different upper limit definition is often used to avoid excluding an alternative hypothesis with a fixed probably α even when the analysis has no sensitivity, referred to as CLs procedure [101, 102], in which the exclusion limit is defined as the value of μ for which $P(\widetilde{q}(\mu) \geq \alpha|\mu)/P(\widetilde{q}(\mu) \geq \alpha|0) \geq \alpha)$, which solves the mentioned issue at the cost of over-coverage.

Most data analyses at the LHC, and particularly searches such as the one discussed in Chapter 5, are carried out in blinded manner to reduce the experimenter's bias, i.e. the subset of observations or results relevant for statistical inference are not considered (or concealed) until all the analysis procedures have defined. In order to optimise the various analysis components (e.g. selection or summary statistic), it is useful to compute a figure of merit that is representative of the prospective sensitivity of the analysis. The *expected significance*, is the expectation value for the probability value from Equation 3.37 under the alternative hypothesis. Instead, the median instead of the expectation is often considered to preserve monotonicity with Z-values, and several approximations exist for simple cut-and-count likelihoods. Both the expected and median significance depend on the signal fraction μ assumed, so they are particularly useful to optimise analyses where the order of magnitude expected for μ is known, e.g. cross section measurements of SM processes.

Alternatively, the expected median upper limit can be defined as the exclusion upper limit using the median test statistic $\widetilde{q}_{\text{med}}(\mu)$ under the null hypothesis instead of the observed statistic. In addition to the median expected limit, it is common practice in LHC searches to also compute the so-called 1-sigma and 2-sigma bands, that correspond to the 50.0 ± 34.1 and 50.0 ± 47.7 percentiles instead of the median. The upper limit bands provide a quantitive estimation of the possible limit variation if no signal is present in the data. Both the expected significance and the expected upper limit can be estimated asymptotically for summary statistics like the one described in Equation 3.29. The effect of nuisance parameters can be also included in both in the asymptotic approximations or the Monte Carlo based estimation. The asymptotic approximation are found to be good empirically, within 10% to 30% (for situations where the number of events is small) of the Monte Carlo based estimation, and thus are frequently used for obtaining limits and significances in New Physics searches.

3.2.3 Parameter Estimation

Another inference problem that can be defined based on the observed data, is parameter estimation, whose goal can generally be defined as the determination of the possible or optimal values that the parameters of a statistical model in relation to a set of observations. Two types of parameter estimation problems are often considered: point estimation and interval estimation. If the aim is to obtain the best estimate (i.e. a single value) of a vector of parameter based on a set of observations, it is referred to as a *point estimation* problem. When we are instead interested on using a set of observations to make statistical statements about a range or region for the values that the statistical model parameters, we are dealing with an *interval estimation* problem.

Parameter estimation can be addressed either from a classical (i.e. also known as frequentist) standpoint where the true values of the parameter are assumed to be fixed but unknown, and intervals represent the region of parameters for which the set of observed data could be obtained upon repeated sampling; or from a Bayesian perspective, where probabilistic statements representing the degree of belief on the values for the parameters are updated based on the set of observations. A classical inference approach is predominantly adopted in this document, where the definition of probability is based on the relative frequency of the outcome when repeated trials are carried out. Classical interval estimation, often referred to as *confidence interval* estimation is strongly related with hypothesis testing, as reviewed in Section 3.2.2.

The $100(1-\alpha)\%$ confidence interval (CI) for a one-dimensional parameter θ can be defined as the interval $[\hat{\theta}^-, \hat{\theta}^+]$:, such that:

$$P(\hat{\theta}^- \leq \theta \leq \hat{\theta}^+) = 1 - \alpha \tag{3.42}$$

where $\hat{\theta}^-$ and $\hat{\theta}^+$ are referred as the lower and upper limits. The definition of confidence interval in the context of classical parameter estimation is the range of values for a given parameter which, upon repeated trials, would contain the true value $100(1-\alpha)\%$ of the times. The concept of confidence interval can also be extended to confidence region when a multi-dimensional parameter vector or several disjoint intervals are considered. While the definition of confidence interval based on its coverage properties is rather simple, its construction based on a set of observations $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$ can be quite challenging. It is worth noting that both upper and lower limit are estimators, quantities calculated by applying a given produce to the set of observations, and thus $\hat{\theta}^-(D)$ and $\hat{\theta}^+(D)$ explicitly depend on the set of data.

The Neyman construction [103] provides a principled procedure to define $100(1-\alpha)\%$ confidence intervals which guarantee the property defined in Equation 3.42, by inverting an ensemble of hypothesis tests (as defined in Section 3.2.2), by using simulated datasets for the different values that parameter θ can take. Confidence intervals can be one-sided, e.g. such as the exclusion upper limits defined in Equation 3.41, or two-sided as the definition provided in Equation 3.42. In particle collider analyses, there is often a dichotomy between one-sided intervals for null results and two-sided intervals for non-null results, which can be solved by extending the Neyman construction with a likelihood-ratio ordering criterion [104].

Confidence interval procedures based on the Neyman construction work very well for simple statistical models with one or two parameters, however rapidly become computationally intractable for larger number of parameters. Even though the number of parameters of interest at LHC analyses is usually small, nuisance parameters play an important role in inference as reviewed in Section 3.1.4, and cannot be accounted for in a straightforward manner in the previous procedure. Thus when the total number of parameters is high, confidence intervals are usually computed based on alternative approximations, often based of some of the properties of the profiled likelihood ratio discussed in Section 3.2.2.

Before discussing the fundamentals of the confidence interval approximations, it is useful to formally define the *maximum likelihood estimator* of a parameter $\boldsymbol{\theta}_{\text{ML}}$ based on a set of observations $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$ as:

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} L(D; \boldsymbol{\theta}) \tag{3.43}$$

where $L(D; \boldsymbol{\theta})$ is the likelihood function given the set of observations D which is a function of the model parameters $\boldsymbol{\theta}$. The maximum likelihood estimator of model parameters was already used to define the profile likelihood ratio test statistic in Equation 3.40, an it is a very common point estimator because it is asymptotically consistent and efficient. In addition, the maximum likelihood estimator coincides with the *maximum a posteriori* (MAP) point estimator in Bayesian inference when the parameter priors are uniform, because the evidence is proportional to the likelihood.

The shape of the likelihood function around the maximum likelihood estimator $\boldsymbol{\theta}_{\mathrm{ML}}$ can be used to approximate confidence intervals. Using asymptotic theory developed by Wilks [98], the $100(1-\alpha)\%$ confidence region for the parameter vector $\boldsymbol{\theta}$ can be determined using the following relation:

$$-\ln L(D; \boldsymbol{\theta}) \leq -\ln L(D; \boldsymbol{\theta}_{\mathrm{ML}}) + \Delta \ln L \tag{3.44}$$

where $\ln L(D; \boldsymbol{\theta}_{\mathrm{ML}})$ is the natural logarithm of the likelihood for the maximum likelihood estimator and $\Delta \ln L$ depends on the number of parameter dimensions and the desired coverage $1 - \alpha$. For example, the values of $\boldsymbol{\theta}$ inside the 68.27% (i.e. 1-sigma) confidence region and for one dimensional parameter are those for which the previous relation is verified using $\Delta \ln L = 0.5$. If $\boldsymbol{\theta}$ is one-dimensional and the function $L(D; \boldsymbol{\theta})$ is convex, the confidence interval limits $\hat{\theta}^{-}(D)$ and $\hat{\theta}^{+}(D)$ can be obtained by finding the most extreme values of θ that verify Equation 3.44 at each side of the maximum likelihood estimator $\boldsymbol{\theta}_{\mathrm{ML}}$.

As discussed in Section 3.1.4, we are often interested on confidence intervals for a subset of interest of the statistical model $\boldsymbol{\theta}_{\iota}$, while regarding the others as nuisance parameters $\boldsymbol{\theta}_{\nu}$. The previous procedure can be extended for computing approximate confidence interval for the parameters of interest, by considering the profiled likelihood [105] $\hat{L}(D; \boldsymbol{\theta}_{\iota})$ instead of the full likelihood in Equation 3.44, which is defined as:

$$\hat{L}(D; \boldsymbol{\theta}_{\iota}) = \arg\max_{\theta_{\nu} \in \Theta_{\nu}} L(D; \boldsymbol{\theta}_{\iota}, \boldsymbol{\theta}_{\nu}) \tag{3.45}$$

so the nuisance parameters $\boldsymbol{\theta}_{\nu}$ are profiled by considering their values that would maximise the likelihood conditional for each value of the parameters of interest $\boldsymbol{\theta}_{\iota}$.

Noting that a constant denominator in the likelihood would cancel out at each side of Equation 3.44, and similarly when using the profiled likelihood from Equation 3.45. Both procedures can be theoretically linked with the profile-likelihood ratio test statistic defined in Equation 3.40. Algorithms for likelihood maximisation and computation of intervals based on the profiled likelihood are implemented in the MINOS routine as part of the MINUIT software library [106], which can also account for bounded parameters. Confidence intervals based on the profiled likelihood will be used for benchmarking different ways for constructing summary statistics in Chapter 6.

Another important subtlety when dealing with nuisance parameters (which also applies to a lesser degree to the combination of measurements), is that oftentimes they are constrained by theory or external measurement. This can be included in the previous likelihood-based techniques by considering the likelihood as a product of the likelihood derived from the statistical model for the set of observations $L_D(D;\boldsymbol{\theta})$ with the available constraints $L_C^i(\boldsymbol{\theta})$, as follows:

$$L(D;\boldsymbol{\theta}) = L_D(D;\boldsymbol{\theta}) \prod_{i=0}^{c} L_C^i(\boldsymbol{\theta}) \tag{3.46}$$

where simplified likelihoods (e.g. a normal approximation) are often used in the constrain terms $L_C^i(\boldsymbol{\theta})$ but they could in principle also depend on an independent set of observations. The constrain terms could be also understood as prior probability distributions in a Bayesian setting, obtained from previous evidence.

In order to obtain approximate confidence intervals from the shape of the likelihood or profile likelihood function around the maximum likelihood, several likelihood evaluations (together with a constrained optimisation problem if $\hat{L}(D;\boldsymbol{\theta}_\iota)$ is used) are often required to estimate accurately a confidence interval. A cruder but often useful approximation can be obtained from the curvature of the negative log-likelihood function at $\boldsymbol{\theta}_{\mathrm{ML}}$. In more than one dimension, the local curvature can be expressed by the Hessian matrix $\boldsymbol{H}$. The expectation of hessian of the $-\ln L(D;\boldsymbol{\theta})$ is also referred as the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$ [107] and it is defined as:

$$\boldsymbol{I}(\boldsymbol{\theta})_{ij} = \boldsymbol{H}(\boldsymbol{\theta})_{ij} = \mathop{\mathbb{E}}_{D\sim p(D|\boldsymbol{\theta})}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j}(-\ln L(D;\boldsymbol{\theta}))\right] \tag{3.47}$$

which can be evaluated at any given $\boldsymbol{\theta}$, e.g. by using numerical differentiation. The Cramér-Rao lower bound [108, 109] provides a link between the inverse of the Fisher information matrix and the covariance of an unbiased estimator $\hat{\boldsymbol{\theta}}$:

$$\mathrm{cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \geq I(\boldsymbol{\theta})^{-1} \tag{3.48}$$

which becomes an equality in the large-sample limit for an efficient parameter estimator such as the maximum likelihood estimator $\boldsymbol{\theta}_{\mathrm{ML}}$. The diagonal elements of the inverse of the information matrix $\sigma_i^2 = \left(I(\boldsymbol{\theta})^{-1}\right)_{ii}$ may be used to construct a 68.3% confidence interval for θ_i parameter where the effect of the rest of parameters has been profiled as $[\boldsymbol{\theta}_{\mathrm{ML}} - \sigma_i, \boldsymbol{\theta}_{\mathrm{ML}} + \sigma_i]$. This approximation is equivalent to profiling assuming that the $-\ln L(D;\boldsymbol{\theta})$ can be described by a multi-dimensional parabola centered at $\boldsymbol{\theta}_{\mathrm{ML}}$, and thus leads to symmetric intervals. In Bayesian literature, an analogous approach is used to extend MAP estimation in order obtain a multi-dimensional normal approximation for the posterior, which is often referred to as Laplace approximation [110]. An important advantage of this approximation, that will be used in Chapter 6 to construct an inference-aware machine learning loss function, is that can be interpreted both in the context of classical and Bayesian inference.

4 Machine Learning in High-Energy Physics

> Computers are useless.
> They can only give you answers.
>
> Pablo Picasso

Machine learning is an interdisciplinary field that deals with the general problem of how computers can automatically improve at certain tasks given data. The usefulness and range of applicability of such techniques has surged in the last decades due to the increase on accessible computational power and the amount of useful data available. In this section, a general overview of machine learning methods as well as the main tasks that can be addressed with them will be provided. Subsequently, the technical basis of two specific types of machine learning methods used in the next chapters will be explored: boosted decision trees and neural networks. Last but not least, we will go through a brief review of the common past use cases of these techniques at high energy physics experiments, especially focussing on those cases where they can be used to address some of the statistical inference and modelling issues from Chapter 3.

4.1 Problem Description

Machine learning is the field that deals with algorithms, as described by computer programs, that are able to *learn* from data. A more formal definition of learning, yet general and useful in the context of this work, can be found in the literature [111]: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E". The previous sentence clearly denotes the three key elements for learning in the context of computer algorithms: the task (or class of task) that to be accomplished T, a quantitative and robust way to measure the

performance on those tasks P and a set of data that the algorithm can experience in order to improve E.

The first step in order to tackle a problem with machine learning techniques is the formal definition of the task T, together with a quantifiable metric that scores the accuracy on such task P. In this section, the most common machine learning tasks that are of relevance for their possible use in particle collider experiments and similar scientific contexts are introduced. Simultaneously with the description of the tasks, performance measures and data, the main general machine learning concepts are reviewed.

4.1.1 Probabilistic Classification and Regression

One of the conceptually simple, yet versatile, tasks that can be addressed with machine learning algorithms is *classification*. A classifier or a classification rule is a function $f(\boldsymbol{x}) : \mathcal{X} \longrightarrow \mathcal{Y}$ that predicts a label $y \in \{0, ..., k-1\}$, denoting correspondence to one category in a set of of k categories, for each input $\boldsymbol{x} \in \mathcal{X}$. The task of classification, in the context of machine learning algorithms, is to produce classification functions $f(\boldsymbol{x})$ that perform well on an unobserved set of data.

Classification is often framed as belonging to a larger category of tasks referred to as *supervised learning*, where the goal is predicting the value of an output variable $\boldsymbol{y}$ (here a multi-dimensional vector for generality) based on the observed values of the input variables $\boldsymbol{x}$, based on a *learning set* of n input vectors with known output values $S = \{(\boldsymbol{x}_0, \boldsymbol{y}_0), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$. The output values $\boldsymbol{y}$ are known in the learning set, because they were previously determined by an external method, typically a teacher or supervisor looking at past observations, thus explaining the name of these family of techniques.

From a statistical standpoint, the input observations and target values from the learning set can be viewed as random variables sampled from a joint probability distribution $p(\boldsymbol{x}, \boldsymbol{y})$, which is typically unknown. The family of supervised learning tasks also includes *regression*, which amounts to construct a $f(\boldsymbol{x})$ that can to predict a numerical target output $\boldsymbol{y}$, and *structured output* tasks where the output vector $\boldsymbol{y}$ is a vector or a complex data structure where its elements are tightly interrelated. As will be reviewed in Section 4.3, most analysis problems amenable by machine learning in high-energy physics experiments are framed as classification and regression tasks, while the use of structured output tasks is instead not quite extended. The reconstruction of the set and properties of physical objects in an event directly from the detector readout could be framed as a structured output task, if it was to

be approached directly using machine learning algorithms instead of the procedures described in Section 2.3.3.

The goal of supervised learning is not to perform well on the learning set S used for improving at the specified task, but rather to perform well on additional unseen observations sampled from the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$. Supervised learning algorithms exploit the conditional relations between the input and the output variables, in order to classify new observations better than a random classification rule that does not depend on the value of $\boldsymbol{x}$. When using machine learning techniques in data analysis at the LHC, as will be reviewed in Section 4.3, simulated observations are used instead of expert-labelled past observations. Simulated observations correspond to random samples of the joint distribution over the latent variables for the generative model $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$, as described in Section 3.1.

In fact, the problem of inferring a subset of latent variables $\boldsymbol{z}$ of the statistical model for the raw detector readouts of a collider experiment $\boldsymbol{x}$, or from any deterministic function of it $\boldsymbol{s}(\boldsymbol{x})$, can be cast as a supervised learning problem. The learning set S would consist of simulated observations $\boldsymbol{x}_i$ (or a summary of it $\boldsymbol{s}(\boldsymbol{x}_i)$), and a matching subset of interest of the latent variables $\boldsymbol{y}_i \in \mathcal{Y} \subseteq \mathcal{Z}$. The supervised learning task can then be viewed as the estimation of the conditional expectation value $\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x}=\boldsymbol{x}_i)}[\boldsymbol{y}]$ for each given input observation $\boldsymbol{x}_i$, thus characterising the probability distribution $p(\boldsymbol{y}|\boldsymbol{x})$.

While several performance measures P are possible for a given task T, for supervised learning is common to use performance measures that estimate the expected prediction error, or risk R, of a given predictor function $f(\boldsymbol{x})$, which can normally be expressed as:

$$R(f) = \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{x},\boldsymbol{y})} [L(\boldsymbol{y}, f(\boldsymbol{x}))] \tag{4.1}$$

where L is a *loss function* that quantifies the discrepancy between the true output and the prediction. The quantity defined in Equation 4.1 is often also referred to as *risk*, *test error*, or also as *generalisation error*.

The optimal model for a given task T thus depends on the definition of its loss function L, if the objective is minimising the expected prediction error. In practice, the expected prediction error cannot be estimated analytically because $p(\boldsymbol{x}, \boldsymbol{y})$ is not known, or not tractable in the case of a generative simulation model. The

generalisation error has thus to be estimated from a subset of labelled samples $S' = \{(\boldsymbol{x}_0, \boldsymbol{y}_0), ..., (\boldsymbol{x}_{n'}, \boldsymbol{y}_{n'})\}$ as follows:

$$R(f) \approx R_{S'} = \frac{1}{n'} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S'} L(\boldsymbol{y}_i, f(\boldsymbol{x}_i)) \tag{4.2}$$

which is also commonly referred to as *empirical risk* approximation $R_{S'}(f)$ based on the set S'. The supervised learning problem can then be stated as one of finding the function $\hat{f}$ from a class of functions $\mathcal{F}$, which depends on the particularities of the algorithm, that minimises the empirical risk over the learning set S:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} R_S(f) \tag{4.3}$$

which is referred to as empirical risk minimisation (ERM) [112], and it is at core of most of the existing learning techniques, such as those described in Section 4.2. However, the ultimate goal of a learning algorithm is to find a function f^* that minimises the risk or expected prediction error $R(f)$:

$$f^* = \arg\min_{f \in \mathcal{F}} R(f) \tag{4.4}$$

where $R(f)$ is the quantity defined in Equation 4.1, corresponding to the generalisation error, or average expected performance on unseen observations sampled from $p(\boldsymbol{x}, \boldsymbol{y})$. The previous equation can be used to define the optimal prediction function $f_B(\boldsymbol{x})$, also referred as *Bayes model*, which represents the minimal error that any supervised learning algorithm can achieve due to the intrinsic statistical fluctuations and properties in the data. The Bayes model can be expressed as:

$$f_B(\boldsymbol{x}) = \arg\min_{\boldsymbol{y} \in \mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x})} [L(\boldsymbol{y}, f(\boldsymbol{x}))] \tag{4.5}$$

where the last term indicates the optimal choice of target $\boldsymbol{y}$ for each value of $\boldsymbol{x}$. The previous expression can be obtained by explicitly considering the conditional expectation in the risk term described in Equation 4.4, that is $R(h) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{y})} \left[\mathbb{E}_{\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x})} [L(\boldsymbol{y}, f(\boldsymbol{x}))] \right]$, that can be obtained using Bayes theorem. The Bayes model $f_B(\boldsymbol{x})$, and its corresponding risk $R(f_B)$, also referred as *residual error*, can only be estimated if $p(\boldsymbol{x}, \boldsymbol{y})$ is known and the expectation can be computed analytically. Even though the Bayes optimal model cannot be obtained for real world

problems, it can be useful nevertheless when benchmarking techniques in synthetic datasets or for theoretical studies.

Because most learning algorithms optimise f, or its parameters, using the learning set S, the empirical risk $R_{\mathrm{S}}(f)$ is not a good estimator of the expected generalisation error $R(f)$. In general, $R_{\mathrm{S}}(f)$ underestimates $R_{\mathrm{S}}(f)$ because the statistical fluctuations of the finite number of observations in S can be learnt to increase the performance on S, while they are not useful for prediction on a new set of observations. If the family of functions $\mathcal{F}$ considered in the learning algorithm is flexible enough, which is often the case, it is possible to achieve $R_{\mathrm{S}}(f) = 0$ for the learning set S while the generalisation error $R(f)$ is well away from zero. This effect can actually lead to a degradation of the generalisation error while the empirical risk in the learning set is decreasing during the learning procedure, which is often referred to as *over-fitting*.

To compare different prediction functions or to realistically evaluate the generalised performance of a given prediction model f, it is useful to be able to compute unbiased estimates of $R(f)$. The simplest way to obtain such estimate is to divide the learning set S into two disjoint random subsets S_{train} and S_{test}. The train subset S_{train} will be used by the learning algorithm to optimise the prediction function f by means of empirical risk minimisation, as described in Equation 4.3. The hold-out or test subset S_{test} can then be used to obtain an unbiased estimation of the performance of f on unseen observations.

For many learning algorithms, the learning process, or *training*, is iterative: the function f is optimised incrementally based on the training data. In those cases, an estimation of the generalisation error as the training evolves may be useful to stop the training procedure and avoid the degradation of generalisation due over-fitting, in what is referred as *early stopping*. In those cases, as well as to compare and ensemble the results of various predictor functions and model configurations, is useful to hold out a fraction of S_{train} which is commonly referred as validation set S_{valid}. Alternative approaches to estimate the generalisation error exist, including *cross-validation* and its variations [113], which are usually preferred when the amount of training data is reduced.

Another important concept for most machine learning techniques is that of *hyper-parameters*. The majority of machine learning algorithms depend on a set of parameters that regulate the flexibility of the family of functions $\mathcal{F}$ to consider for empirical risk minimisation as well as the details of the optimisation procedure followed to solve the task presented in Equation 4.3. The expected performance of a

given model depends on these parameters, however their optimal value depends on the particularities of the data (e.g. number of input dimensions or number of size of the data size). This motivates the notion of *hyper-parameter optimisation*, where the performance of the various choices of hyper-parameters is evaluated on the validation set or by means of cross-validation techniques, in order to select the best configuration.

The loss function L of a supervised learning algorithm, which quantifies the discrepancies between the prediction and the true output target, depends on the task T and formally defines it. A principled loss function for classification is the *zero-one loss*, which is defined as zero when the prediction $f(\boldsymbol{x})$ matches the target y and one otherwise. The zero-one risk can then be expressed as:

$$R_{0-1}(f) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)} [\mathbb{1}(y \neq f(\boldsymbol{x}))] \tag{4.6}$$

where $\mathbb{1}(y \neq f(\boldsymbol{x}))$ is an indicator function, which was defined in Equation 3.6. The zero-one loss is non-differentiable when $y = f(\boldsymbol{x})$ and its gradients are zero elsewhere; in addition, it is not convex, a property which makes the minimisation task in Equation 4.3 hard to tackle by optimisation algorithms. In fact, it can be proven that finding the function f in F that minimises directly the R_{0-1} empirical risk with a training sample is a NP-hard problem [114]. The Bayes optimal classifier for the 0-1 loss can nevertheless be easily obtained from Equation 4.7 as a function of the conditional expectation:

$$f_B(\boldsymbol{x}) = \arg\min_{y\in\mathcal{Y}} \mathop{\mathbb{E}}_{y\sim p(\boldsymbol{y}|\boldsymbol{x})} [\mathbb{1}(y \neq f(\boldsymbol{x}))] = \arg\max_{y\in\mathcal{Y}} p(y|\boldsymbol{x}) \tag{4.7}$$

thus the optimal classifier amounts to the prediction of the most likely output category y for a given input $\boldsymbol{x}$. The previous problem is normally referred to as *hard classification*, where the objective is to assign a category for each input observation. Because most problem in high-energy physics that can be cast as supervised learning are ultimate inference problems as will be reviewed in Section 4.3, it is generally more useful to consider the problem of *soft classification*, which instead amounts to estimate the class probability for each input $\boldsymbol{x}$.

Soft classification is especially useful when the classes are not separable, which is often the case for applications in collider experiments. Luckily, soft classification is also a consequence of most convex relaxations of the zero-one loss of Equation

4.6. For a two-class classification problem, e.g signal versus background, a useful approximation of the zero-one loss is the binary cross entropy, defined as:

$$L_{\mathrm{BCE}}(y, f(\boldsymbol{x})) = -y \log(f(\boldsymbol{x})) - (1-y)\log(1-f(\boldsymbol{x})) \tag{4.8}$$

where now the one-dimensional output prediction $f(\boldsymbol{x})$, when bounded between 0 and 1 (e.g. using a sigmoid/logistic function), will effectively approximate the conditional probability $p(\boldsymbol{y}=1|\boldsymbol{x})$. In fact, the Bayes optimal model for a binary cross-entropy classifier is:

$$\begin{aligned} f_B(\boldsymbol{x}) &= \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)}[L_{\mathrm{BCE}}(y, f(\boldsymbol{x}))] = p(y=1|\boldsymbol{x}) \\ &= \frac{p(\boldsymbol{x}|y=1)p(y=1)}{\sum_{\forall y_i \in \{0,1\}} p(\boldsymbol{x}|y=y_i)p(y=y_i)} = \left(1 + \frac{p(\boldsymbol{x}|y=0)p(y=0)}{p(\boldsymbol{x}|y=1)p(y=1)}\right)^{-1} \end{aligned} \tag{4.9}$$

where the second line in the equation is a direct consequence of Bayes theorem and from the last term it can be clearly seen that the prediction output is monotonic with the density ratio between the probability density functions for each category. Similar results can be obtained for the Bayes optimal classifier when using other soft relaxations of the zero-one function. Machine learning binary classifiers will effectively approximate this quantity directly from empirical samples, where the prior probabilities of each class represent the relative presence of observations from each category.

Binary cross entropy is a subclass of the more general *cross entropy* loss function, that can be used for k-categories classification, commonly referred to as multi-class classification. In these cases, a k-dimensional vector target $\boldsymbol{y}$ is often constructed, where each component y_i is one if the observation belongs to the class i or zero otherwise, and the output of the prediction function $\hat{\boldsymbol{y}} = f(\boldsymbol{x})$ is also a vector of k components. Within this framework, the cross entropy loss can then be defined as:

$$L_{\mathrm{CE}}(\boldsymbol{y}, f(\boldsymbol{x})) = -\sum_i y_i \log \hat{y}_i \tag{4.10}$$

which can be used to recover Equation 4.8 when $k = 2$, considering the one-dimensional target and prediction as the i=1 elements and that $y_0 = 1 - y$ and $\hat{y}_0 = 1 - f(x)$. If the prediction output is to generally represent exclusive class probabilities, as is the goal of soft classification, the prediction sum is expected to be one. A simple way to ensure the aforementioned property is to apply a function that ensures that the prediction outputs are in the range $[0, 1]$ and normalised so $\sum_i \hat{y}_i = 1$.

The *softmax function* is a common choice to achieving the mentioned transformation within the field of machine learning. It is a generalisation of the logistic function to k dimensions, and is defined as:

$$\hat{y}_i = \frac{e^{f_i(\boldsymbol{x})/\tau}}{\sum_{j=0}^{k} e^{f_j(\boldsymbol{x})/\tau}} \tag{4.11}$$

where f_i and f_j refer to the i and j elements of the vector function $f(\boldsymbol{x})$ and τ is the temperature, a parameter that regulates the softness of the operator which is often omitted (i.e. $\tau = 1$). In the limit of $\tau \to 0^+$, the probability of the largest component will tend to 1 while others to 0. The softmax output can be used to represent the probability distribution of a categorical distribution in a differentiable way, where the outcome represent the probabilities of each of the k possible outcomes. We will make use of this function in Chapter 6. When the softmax function and the cross entropy loss are used together for multiclass classification, the optimal Bayes model is:

$$\begin{aligned} f_{B,i}(\boldsymbol{x}) &= \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)} [L_{\mathrm{CE}}(y, f(\boldsymbol{x}))] = p(y = y_i|\boldsymbol{x}) \\ &= \frac{p(\boldsymbol{x}|y = y_i)p(y = y_i)}{\sum_{\forall y_i \in \{0,\dots,k-1\}} p(\boldsymbol{x}|y = y_i)p(y = y_i)} \end{aligned} \tag{4.12}$$

which can also be expressed as a function of a sum of density ratios of the categories.

4.2 Machine Learning Techniques

While the focus of the previous section was defining the main problems and properties that can be addressed with machine learning techniques, details about the actual computational and statistical procedures used for learning were not provided. In this chapter, the basis of the two classes of algorithms that are used elsewhere in this work will be described in detail: boosted decision trees and artificial neural networks. These families of learning methods are also those that are most commonly used in machine learning within experimental particle physics, mostly to solve supervised learning problems, as will be described in Section 4.3. The overview included here is by no means comprehensive about the mentioned approaches or alternative popular statistical learning techniques such as random forests or support vector machines, for which the following references provided a more extensive review [113, 115, 116].

4.2.1 Boosted Decision Trees

The term *boosted decision trees* (BDT) refers to a large family of algorithms that are based on additively constructing ensembles of decision trees for supervised learning tasks [117, 118, 119] as those described in Section 4.1.1. A subset of these techniques, which is often referred as *gradient boosting*, are particularly useful for classification and regression problems. The basis for these methods is that a strong model can be obtained by combining the outcome of a set of weak models, e.g. shallow binary decision trees, if they are built to minimise the residual error at each stage. Gradient boosting algorithms can be applied to any supervised task as long as it can be specified by a differentiable loss function, and they can be understood as *gradient descent* (which will be discussed in Section 4.2.2) in function space [120].

While it can be applied to other weak learners, gradient boosting is often used to learn ensembles of decision trees. A decision tree is hierarchical branched structure that associates an outcome for each input $\boldsymbol{x} \in \mathcal{X}$ by means of partitioning the input space in different disjoint subsets $R = (\mathcal{X}_0, ..., \mathcal{X}_L)$, each associated with a constant prediction w_r for each leaf. A generic type of decision trees, which is referred to as classification and regression trees (CART) [121] can be expressed as a function of the input $t(\boldsymbol{x})$ as a sum over the indicator function $\mathbb{1}^r_{\mathcal{X}}(\boldsymbol{x})$ of each subspace (see Equation 3.6) as follows:

$$t(\boldsymbol{x}) = \sum^{\mathcal{X}_r \in R} w_r \mathbb{1}_{\mathcal{X}_r}(\boldsymbol{x}) \tag{4.13}$$

where w_r is the outcome for each subspace, noting the summands will be zero for all subsets $\mathcal{X}^r$ except for one because their are disjoint. The indicator function $\mathbb{1}_{\mathcal{X}_r}(\boldsymbol{x})$ of a given subspace is specified by a series of binary decisions on a single feature. If the leaf predictions w_r are categorical, the resulting model $t(\boldsymbol{x})$ is referred as a classification tree. If w_r are numerical, $t(\boldsymbol{x})$ is a regression tree. In the context of gradient boosting, regression trees are often more useful, even for classification tasks, i.e. regression trees can be used in conjunction with soft classification loss functions (e.g. cross entropy). For the rest of this section, we will then focus on gradient boosting with regression trees. A schematic representation of a regression tree is provided in Figure 4.1, which corresponds to the first tree in the ensemble used for signal versus background classification in the analysis described in Chapter 5.

Given its structural limitations, a single CART tree of small maximum depth d performs rather poorly a given supervised learning task for complex non-linear problems. If d is very large, the problem of learning an optimal tree based on data is computationally very demanding, and the resulting model would not generalise well

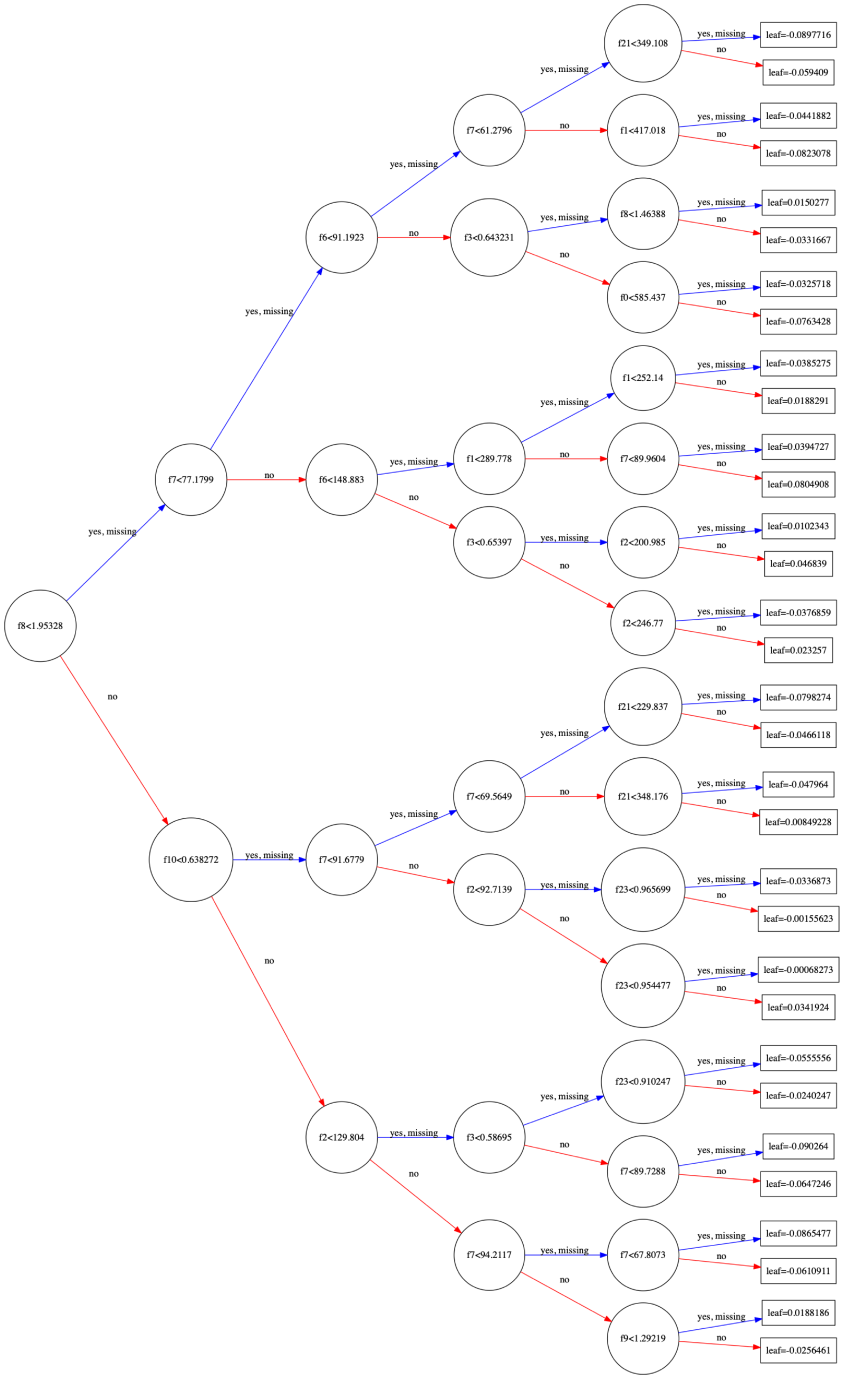


Figure 4.1: Graphical representation of a regression tree. At each node that is not a leaf node, the tree is split in two depending on wether based on whether a boolean condition is met, which based on a threshold for the input variable indexed by the number indicated. This tree corresponds to the first on the ensemble of trees used for classification in Chapter 5, which was trained using binary cross entropy as loss function.

to unseen data. This motivates the use of tree ensembles, where the final prediction is composed by the combined predictions of several small trees. For an ensemble of K CART trees, the final model prediction $T(\boldsymbol{x})$ can be expressed as:

$$T(\boldsymbol{x}) = \sum_{j=1}^{K} t_j(\boldsymbol{x}) \tag{4.14}$$

where each $t_j(\boldsymbol{x})$ is a CART model, as described in Equation 4.13. Other regression tree ensembles based on alternative methods such as bagging [122] can also be expressed by a similar combination of predictions. The learning problem can be expressed as empirical risk minimisation in the space of possible tree ensembles over the learning set of labelled observations $S = \{(\boldsymbol{x}_0, \boldsymbol{y}_0), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, as discussed in Equation 4.3. The total empirical risk functional $R(T)$ for an ensemble of K trees can usually be written as:

$$R(T) = \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S} L(\boldsymbol{y}_i, T(\boldsymbol{x}_i)) + \sum_{j=1}^{K} \Omega(t_j) \tag{4.15}$$

where $L(\boldsymbol{y}_i, T(\boldsymbol{x}_i))$ is the preferred loss function for the task (e.g. binary cross entropy as defined in Equation 4.8) and $\Omega(t_j)$ is a regularisation term that depends on the properties of each tree and controls the complexity of the model in order to avoid overfitting.

Because learning the structure and leaf weights w_r of all trees in the ensemble at the same time is intractable, boosting is based on sequentially learning trees. At each step, a tree t_j is built to improve over the previously ensemble of trees $T_{(j-1)}(\boldsymbol{x})$, the prediction for each observation in the learning set a given step j of the training procedure can then be expressed as:

$$T_j(\boldsymbol{x}_i) = T_{(j-1)}(\boldsymbol{x}_i) + t_j(\boldsymbol{x}_i) \tag{4.16}$$

which can be used to redefine the equivalent risk from Equation 4.15 at each training step, where the tree $t_j(\boldsymbol{x})$ is being created as:

$$R(T_j) = \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S} L(\boldsymbol{y}_i, T_{(j-1)}(\boldsymbol{x}_i) + t_j(\boldsymbol{x}_i)) + \sum_{j=1}^{K} \Omega(t_j) \tag{4.17}$$

where the loss $L(\boldsymbol{y}_i, T_{(j-1)}(\boldsymbol{x}_i)$ can be expanded as a Taylor series assuming that at the step j the ensemble $T_{(j-1)}(\boldsymbol{x})$ is constant. Omitting constant terms, which do not play any role in risk minimisation, the risk at a given training step can be expressed as:

$$R(T_j) \sim \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S} \Bigg(\underbrace{\frac{\partial L(\boldsymbol{y}_i, T_{(j-1)}(\boldsymbol{x}_i))}{\partial T_{(j-1)}(\boldsymbol{x}_i)}}_{g_i} t_j(\boldsymbol{x}_i) + \frac{1}{2} \underbrace{\frac{\partial^2 L(\boldsymbol{y}_i, T_{(j-1)}(\boldsymbol{x}_i))}{\partial T^2_{(j-1)}(\boldsymbol{x}_i)}}_{h_i} t_j^2(\boldsymbol{x}_i) \Bigg) + \Omega(t_j) \tag{4.18}$$

where g_i and h_i are so-called gradient statistics, computed using the first and second partial derivatives of the loss function with respect to the ensemble prediction at the previous step $T_{(j-1)}(\boldsymbol{x}_i)$. At each step the learning problem can then be reduced to choosing a tree structure and weights, characterised by the function t_j, that minimises $R(T_j)$. This technique can therefore be applied to any supervised learning tasks as long the associated loss function is differentiable.

A common regularisation term, that is used by the XGBOOST library [123] used for training the classifier in Chapter 5, is a combination of the number of leaves L and the squared sum of the leaf weights w_r for all the leaves:

$$\Omega(t_j) = \gamma L + \frac{1}{2} \lambda \sum^{\mathcal{X}_r \in R} w_r^2 \tag{4.19}$$

where γ and λ are constants that regulate the relative importance of each regularisation component. Using the previous regularisation term, it is possible to redefine the risk of a given tree structure and set of leaf weight at given training step as:

$$R(T_j) \sim \sum^{\mathcal{X}_r \in R} \left(w_r \underbrace{\sum^{\boldsymbol{x}_i \in S} g_i \mathbb{1}_{\mathcal{X}_r}(\boldsymbol{x}_i)}_{G_r} + \frac{1}{2} w_r^2 \underbrace{\sum^{\boldsymbol{x}_i \in S} (h_i + \lambda) \mathbb{1}_{\mathcal{X}_r}(\boldsymbol{x}_i)}_{H_r + \lambda} \right) + \gamma L \tag{4.20}$$

where G_r and H_r represent the sum of g_i and h_i over all the samples in the learning set that correspond to the leaf indexed by r. The previous expression can in turn be used to obtain the optimal leaf weight $w_r^\star$ and simplify the risk at a given step as follows:

$$w_r^\star = -\frac{G_r}{H_r + \lambda} \Rightarrow R(T_j) = -\frac{1}{2} \sum^{\mathcal{X}_r \in R} \frac{G_r^2}{H_j + \lambda} + \gamma T \tag{4.21}$$

where $\mathcal{X}_r$ are the subsets of the input space corresponding to each leaf of the last tree j. The last expression for $R(T_j)$ can be used to compare tree structures to be added to the ensemble in a principled manner.

In practice, the number of possible tree structures is infinite so the problem of finding the optimal tree at each step is still intractable. A greedy heuristic is instead used, which proceeds one level of the tree at time. For each input feature, the optimal splitting at a given level can be found by maximising the splitting gain, which can be done very efficiently by sorting the observations in that feature and finding the threshold that maximises the gain $\mathcal{G}$, that is defined as:

$$\mathcal{G} = \frac{1}{2}\left(\frac{G_L}{H_L+\lambda} + \frac{G_R}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) + \gamma \tag{4.22}$$

where G_L and H_L are the sum of gradient statistics left of the threshold and G_R and H_R are those right of the threshold. If the gain is negative for the whole, no splitting is preferred in the considered features. Once the optimal splitting is determined for all the features, the featurs that provides the minimal risk as defined in Equation 4.21 is chosen. The algorithm then proceeds to the next tree level until the maximum tree depth is reached or any additional splitting degrades the performance.

Boosted tree ensembles are prone to overfitting to the learning set, so additional heuristics are often used to improve generalisation. A common approach after each step that produces a tree t_j by the procedure outlined before, is to define ensemble for the next step by weighting the constribution from the last three $T_j(\boldsymbol{x}_i) = T_{(j-1)}(\boldsymbol{x}_i) + \eta\, t_j(\boldsymbol{x}_i)$, where η is referred as learning rate or shrinkage. The use of $\eta < 1$ produces a less efficient learning procedure, so additional trees are required in the ensemble, however the resulting model is less prone to overfitting. Other policies against overfitting include subsampling the set of observations or the feature vector dimensions. Early stopping, as defined in Section 4.1.1, can also be trivially applied to boosted tree ensembles simply by leaving out the last n trees in the summation so the risk over validation set is maximised.

4.2.2 Artificial Neural Networks

An alternative way to carry out empirical risk minimisation is based on consider function $f(\boldsymbol{x};\boldsymbol{\phi})$, which depends on a vector of parameters $\boldsymbol{\phi}$, and attempt to find the values of $\boldsymbol{\phi}$ that minimise the risk $R_S(f)$ over the learning set $S = \{(\boldsymbol{x}_0, \boldsymbol{y}_0), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$. If $f(\boldsymbol{x};\boldsymbol{\phi})$ is differentiable with respect to the parameter vector $\boldsymbol{\phi}$, the minimisation from Equation 4.4, can be attempted with gradient-based methods. The simplest

gradient-based optimisation technique is referred to as *gradient descent* (GD), and can be applied to the previous problem by initialising the parameter vector at random $\boldsymbol{\phi}^0$ and then iteratively updating the model parameters $\boldsymbol{\phi}$ at each step t according to:

$$\boldsymbol{\phi}^{t+1} = \eta(t)\nabla_{\boldsymbol{\phi}} R_S(\boldsymbol{\phi}^t) = \eta(t)\nabla_{\boldsymbol{\phi}} \frac{1}{n} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i)\in S} \big(L(\boldsymbol{y}_i, f(\boldsymbol{x}_i; \boldsymbol{\phi}^t)) + \Omega(\boldsymbol{\phi}^t)\big) \tag{4.23}$$

where $\nabla_{\boldsymbol{\phi}}$ is the gradient operator with respect the model parameters, $\eta(t)$ is the learning rate or step size and $\Omega(\boldsymbol{\phi})$ is a generic generalisation term added to the loss to constrain model complexity. Many other gradient-based optimisation methods exist [124], e.g. using second-order derivative information. The previous flavour of gradient descent is often referred as batch gradient descent, because the whole learning set S is used to compute the parameter updates at each step. Batch gradient descent can be very computationally demanding when the number of observations in S is large and the computation of the gradient of the loss for each labelled observation is costly. In addition, batch gradient descent is a deterministic optimisation method and likely to get stuck at a local minima if the optimisation surface is non-convex.

A variation of the previous technique, that is referred to as stochastic gradient descent (SGD) [125], overcomes the mentioned issues by using a random subset $B = \{(\boldsymbol{x}_0, \boldsymbol{y}_0), ..., (\boldsymbol{x}_m, \boldsymbol{y}_m)\}$ of m observations from the training set S at each step. If m is small the updates can be computed much faster, the trade-off being more noisy estimates of $\mathbb{E}_{(\boldsymbol{x}_i, \boldsymbol{y}_i)\in S} \nabla_{\boldsymbol{\phi}} \big[L(\boldsymbol{y}_i, f(\boldsymbol{x}_i; \boldsymbol{\phi}^t)\big]$. The parameter update rule from Equation 4.23 in SGD can be instead be expressed as:

$$\boldsymbol{\phi}^{t+1} = \eta(t)\nabla_{\boldsymbol{\phi}} R_S(\boldsymbol{\phi}^t) = \eta(t)\nabla_{\boldsymbol{\phi}} \frac{1}{m} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i)\in B} \big(L(\boldsymbol{y}_i, f(\boldsymbol{x}_i; \boldsymbol{\phi}^t)) + \Omega(\boldsymbol{\phi}^t)\big) \tag{4.24}$$

where B is a random subset of size m of the learning set S. In the original formulation $m = 1$, yet nowadays a larger value for m is often used in what is referred to as mini-batch SGD to obtain balance the estimate noise and take advantage of vectorised computations. Several variations of SGD exist, which in some cases can provide convergence advantages over the previous update rule by using adaptive learning rates or momentum in the update dynamics [126]. Stochastic gradient descent methods are a key element for training complex differentiate machine models $f(\boldsymbol{x}; \boldsymbol{\phi})$ as artificial neural networks, which will be discussed in the rest of this section. SGD in combination with a non-decomposable loss function is also used in Chapter 6 to learn inference-aware summary statistics.

A particularly promising family of parametric functions $f(\boldsymbol{x};\boldsymbol{\phi})$ is referred to as *artificial neural networks*. Artificial neural networks are differentiable functions based on the composition of simple (and possibly non-linear) operations. The simplest type of artificial neural network is depicted in Figure 4.2, which is referred as *feed-forward neural network*, that maps a input $\boldsymbol{x}$ to an output $\boldsymbol{y}$ by means of a series of forward transformations, referred as neural network layers. In the simplest configuration, the values at a given layer k other than the input layer can be computed as non-linear transformation of the result of a linear combination of the output of the previous layer after the addition of a bias term. The previous transformation can be expressed very compactly in matrix form as:

$$\boldsymbol{a}^k = g((\boldsymbol{W}^k)^T\boldsymbol{a}^{k-1} + \boldsymbol{b}^k) \tag{4.25}$$

where $\boldsymbol{a}^k$ is the outcome in vector notation after the layer transformation, $\boldsymbol{a}^{k-1}$ is the vector of values from the previous transformation (or $\boldsymbol{a}^0 = \boldsymbol{x}$ if it is the first layer after the input), $\boldsymbol{W}^k$ a matrix with all the linear combination coefficients and $\boldsymbol{b}^k$ is the bias vector that is added after linear combination. The activation function $g(\boldsymbol{z})$ is applied element-wise, and it is often based on a simple non-linear function. The sigmoid function $\sigma(z) = 1/(1+e^z)$ used to be a common choice for the activation function, but nowadays the rectified linear unit (ReLU) function $g(\boldsymbol{z}) = \max(0, z)$ and its variants are most frequently used instead.

The full feed-forward model $f(\boldsymbol{x};\boldsymbol{\phi})$ is based on the composition of transformation of the type described in Equation 4.25. When a single transformation is applied, i.e. $\boldsymbol{y} = g((\boldsymbol{W})^T\boldsymbol{x} + \boldsymbol{b})$, the model can be referred to as perceptron. If the model is instead based on the composition of several transformations, it can also be called multi-layer perceptron (MLP), and each of the intermediate transformations (which can be composed by an arbitrary number of computational units) is referred as hidden layers. The model in Figure 4.2 is a MLP. The advantage of using models based on feed-forward neural networks with hidden layers is that they can be used to model any arbitrary function due to the universal approximation theorem [127]. In fact, while it is still the focus of theoretical research, the use of a large number of hidden layers is found to increase the expressivity and facilitate the training of powerful neural network models. The experimental success of these family techniques has led to the concept of *deep learning*, where multiple transformations layers are used for learning data representations in many learning tasks.

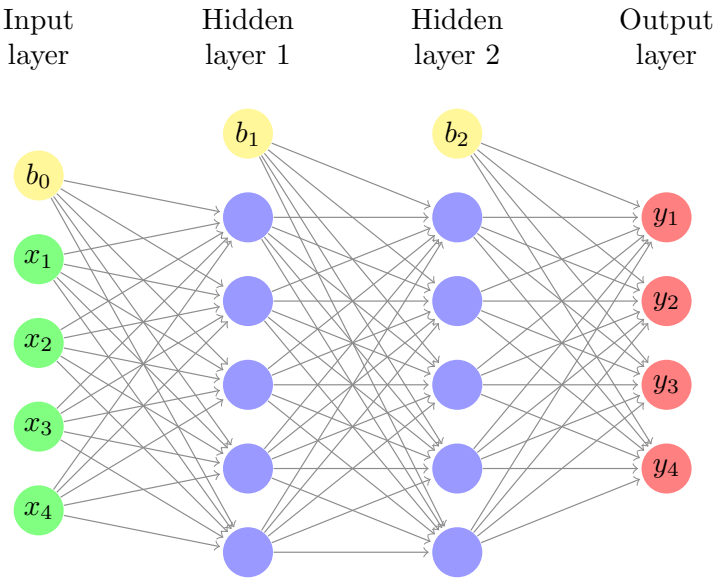


Figure 4.2: Graphical representation of a feed-forward neural network with two hidden layers, which is a function mapping and input $\boldsymbol{x}$ to an output $\boldsymbol{y}$ by means simple non-linear transformations. The output value of a node each layer (other than the input layer) is the result of applying an activation function g to a linear combination of the previous layer outputs plus possibly a bias term.

A good choice for depth and overall structure for a neural network model depends on the problem at hand as well as the characteristics and size of the learning set available, thus it frequently has to be defined by trial-and-error, based on the performance on a validation set as discussed in Equation 4.1.1. The output size and choice of activation function in the last transformation often depends on the task at hand. For binary classification classification tasks, it is practical to use the sigmoid function $\sigma(z) = 1/(1 + e^z)$ as the activation function of the last layer, in combination with a loss function for soft classification (e.g. binary cross entropy from Equation 4.8). For multi-class classification problems, such as the one discussed in Section 4.3.2, the size of the output vector usually matches the number of the categories given that the softmax function (see Equation 4.11) is often used in the last layer to approximate conditional class probabilities in combination with a cross entropy loss (see Equation 4.10). For learning tasks different from classification, different output structures and constraints might be used, e.g. the output vector size in the use case in Chapter 6 corresponds to the number of dimensions of the resulting summary statistic, that is based on a transformation of the input using a multi-layer neural network.

The SDG update rule from Equation 4.24 requires the computation of the gradients of the loss function with respect to the model parameters. For complex models, e.g. those put together by stacking layers as those described in Equation 4.25, the computation of derivatives by numerical finite differences or symbolic differentiation may become rather challenging. The former requires the evaluation of the loss function after variations for at least twice the number of parameters and are affected by round-off and truncation errors, and a naive use of the later could instead lead to very large expressions for the exact derivative that cannot be easily simplified. Given that a numerical function as implemented in a computer program is a sequence of simple operations (e.g. addition, subtraction, exponentiation, etc.), it is possible to efficiently obtain gradients and other derivatives by applying the chain rule repeatedly based on the structure of the program, the derivatives of the simple operations and a record of the intermediate values.

The previous family of techniques, which will not be discussed in depth in this work, are referred as *automatic differentiation* (AD) [128]. The most efficient way of computing the gradients of a one-dimensional function that depends on many parameters, as the gradient of the empirical risk for a batch of observations from Equation 4.24 is by means of reverse-mode automatic differentiation, which is also referred to as the *backpropagation* in the context of neural network training. The computational cost of computing the full gradient of the loss to numerical precision using backpropagation is of the same order than a single forward evaluation of the loss, which provides a great advantage relative to finite differences. In addition, when implemented in a computation framework, it can be generally applied to any numerical function as long as can be expressed as a computational graph, e.g. an arbitrary program containing control flow statements, without requiring complex expression simplification as would be the case for symbolic differentiation. In fact, modern computational that include automatic differenciation such as TensorFlow [129] or PyTorch [130] may also be used to compute higher-order gradients (e.g. Hessian matrix elements), which are useful in Chapter 6 to build a differentiable approximation the covariance matrix based on a summary statistic.

As mentioned before, reverse mode automatic differentiation can be used to computed the gradients of an arbitrary function as long as it can be represented as a computational graph containing differentiable simple operations. Thus the neural network model $f(\boldsymbol{x};\boldsymbol{\phi})$ is not restricted to the composition of layers of the type described in Equation 4.25, which are often referred as fully connected or dense layers. Alternative function components are useful for dealing with data cannot be repre-

sented by a fixed-length vector [115], e.g. convolutional layers are often useful for working with 2D images while recurrent layers extend the application of neural networks to sequences that vary in length between observations. Both convolutional and recurrent layers are used in the neural network model for jet flavour-tagging described in Section 4.3.2. Other differentiable neural network components have also been developed to deal with permutation invariant sets [131] or graphs [132] as input data structures, which could have promising applications in particle collider experiments analyses.

4.3 Applications in High Energy Physics

Machine learning techniques, in particular supervised learning, are increasingly being used in experimental particle physics analysis at the LHC [133]. In this section, the main use cases are described, linking the learning task with the statistical problems and properties which were described in Chapter 3. In broad terms, most supervised learning at collider experiments can be viewed as a way to approximate the latent variables of the generative model based on simulated observations. Those latent variable approximations are often very informative about the parameters of interest and then can be used to construct summary statistics of the observations, which allow to carry out likelihood-free inference efficiently.

4.3.1 Signal vs Background Classification

The mixture structure of the statistical model for the outcome of collisions, discussed in Chapter 3, facilitates its framing as a classification problem. Intuitively, the classification objective could be stated as the separation of detector outcomes coming from processes that contain information about the parameters of interest from those that do not, which will be referred as signal and background respectively, following the same nomenclature from Section 3.1.1. The two classes are often non-separable - i.e. a given detector outcome $\boldsymbol{x}$ (or any function of it) could have been produced either by signal or background processes, and only probabilistic statements of class assignment can be made.

In order to use supervised machine learning techniques to classify detector outcomes, labelled samples are required, yet only the detector readout $\boldsymbol{x}$ is known for collected data. Realistic simulated observations, generated specifically to model events from a given set processes (e.g. signal and background) can instead be used as training data, where the categorical latent variable z_i that represents a given set of processes

can effectively used as classification label. If the simulator model is misspecified, e.g. due to the effect of known unknowns as discussed in Section 3.1.4, the resulting classifiers would be trained to optimise the classification objective for different distributions.

To understand the role of classification in the larger goal of statistical inference of a subset of parameters of interest in a mixture model, let us consider the general problem of inference for a two-component mixture problem. One of the components will be denoted as signal $p_s(\boldsymbol{x}|\boldsymbol{\theta})$ and the other as background $p_b(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are of all parameters the distributions might depend on. As discussed in Section 3.1.1, it is often the case that $f_s(\boldsymbol{x}|\boldsymbol{\theta})$ and $f_b(\boldsymbol{x}|\boldsymbol{\theta})$ are not known, observations can only be simulated, which will not affect the validity the following discussion. The probability distribution function of the mixture can be expressed as:

$$p(\boldsymbol{x}|\mu, \boldsymbol{\theta}) = (1 - \mu)p_b(\boldsymbol{x}|\boldsymbol{\theta}) + \mu p_s(\boldsymbol{x}|\boldsymbol{\theta}) \tag{4.26}$$

where μ is a parameter corresponding to the signal mixture fraction, which will be the only parameter of interest for the time being. As discussed in Section 3.1.1, most of the parameters of interest in analyses at the LHC, such as cross sections, are proportional to the mixture coefficient of the signal in the statistical model. The results presented here would also be also be valid if alternative mixture coefficient parametrisations such as the one considered in Section 6.5.1 are used, e.g. $\mu = s/(s + b)$ where s and b is the expected number of events for signal and background respectively, as long as b is known and fixed and s is the only parameter of interest.

Likelihood Ratio Approximation

Probabilistic classification techniques will effectively approximate the conditional probability of each class, as discussed in Equation 4.9 for the binary classification. A way to approximate the density ratio $r(\boldsymbol{x})$ between two arbitrary distribution functions $\rho(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is then to train a classifier - e.g. a neural network optimising cross-entropy. If samples from $\rho(\boldsymbol{x})$ are labelled as $y = 1$, while $y = 0$ is used for observations from $q(\boldsymbol{x})$, the density ratio can be approximated from the soft BCE classifier output $s(\boldsymbol{x})$ as:

$$\frac{s(\boldsymbol{x})}{1 - s(\boldsymbol{x})} \approx \frac{p(y = 1|\boldsymbol{x})}{p(y = 0|\boldsymbol{x})} = \frac{p(\boldsymbol{x}|y = 1)p(y = 1)}{p(\boldsymbol{x}|y = 0)p(y = 0)} = r(\boldsymbol{x})\frac{p(y = 1)}{p(y = 0)} \tag{4.27}$$

thus the density ratio $r(\boldsymbol{x})$ can be approximated by a simple function of the trained classifier output directly from samples of observations. The factor $p(y=1)/p(y=0)$ is independent on $\boldsymbol{x}$, and can be simply estimated as the ratio between the total number of observations from each category in the training dataset - i.e. equal to 1 if the latter is balanced.

Density ratios are very useful for inference, particularly for hypothesis testing, given that the likelihood ratio Λ from Equation 3.39 is the most powerful test statistic to distinguish between two simple hypothesis and can be expressed as a function of density ratios. Returning to the two component mixture from Equation 4.26, for discovery the null hypothesis H_0 corresponds to background-only $p(\boldsymbol{x}|\mu=0,\boldsymbol{\theta})$ while the alternate is often a given mixture of signal and background $p(\boldsymbol{x}|\mu=\mu_0,\boldsymbol{\theta})$, where μ_0 is fixed. For the time being, the other distribution parameters $\boldsymbol{\theta}$ will be assumed to be known and fixed to the same values for both hypothesis. The likelihood ratio in this case can be expressed as:

$$\Lambda(\mathcal{D};H_0,H_1)=\prod_{\boldsymbol{x}\in\mathcal{D}}\frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{x}|H_1)}=\prod_{\boldsymbol{x}\in\mathcal{D}}\frac{p(\boldsymbol{x}|\mu=0,\boldsymbol{\theta})}{p(\boldsymbol{x}|\mu=\mu_0,\boldsymbol{\theta})} \tag{4.28}$$

where the $p(\boldsymbol{x}|\mu=0,\boldsymbol{\theta})/p(\boldsymbol{x}|\mu_0,\boldsymbol{\theta})$ factor could be approximated from the output of a probabilistic classifier trained to distinguish observations from $p(\boldsymbol{x}|\mu=0,\boldsymbol{\theta})$ and those from $p(\boldsymbol{x}|\mu=\mu_0,\boldsymbol{\theta})$. A certain μ_0 would have to be specified to generate $p(\boldsymbol{x}|\mu=\mu_0,\boldsymbol{\theta})$ observations in order to train the classifier. The same classifier output could be repurposed to model the likelihood ratio when H_1 is $p(\boldsymbol{x}|\mu=\mu_1,\boldsymbol{\theta})$ with a simple transformation, yet the mixture structure of the problem allows for a more direct density ratio estimation alternative, which is the one regularly used in particle physics analyses.

Let us consider instead the inverse of the likelihood ratio Λ from Equation 4.28, each factor term is thus proportional to the following ratio:

$$\Lambda^{-1}\sim\frac{p(\boldsymbol{x}|H_1)}{p(\boldsymbol{x}|H_0)}=\frac{(1-\mu_0)p_\mathrm{b}(\boldsymbol{x}|\boldsymbol{\theta})+\mu_0 p_\mathrm{s}(\boldsymbol{x}|\boldsymbol{\theta})}{p_\mathrm{b}(\boldsymbol{x}|\boldsymbol{\theta})} \tag{4.29}$$

which can in turn be be expressed as:

$$\Lambda^{-1}\sim(1-\mu)\left(\frac{p_\mathrm{s}(\boldsymbol{x}|\boldsymbol{\theta})}{p_\mathrm{b}(\boldsymbol{x}|\boldsymbol{\theta})}-1\right) \tag{4.30}$$

thus each factor in the likelihood ratio is a bijective function of the ratio $p_\mathrm{s}(\boldsymbol{x}|\boldsymbol{\theta})/p_\mathrm{b}(\boldsymbol{x}|\boldsymbol{\theta})$. The previous density ratio can be approximated by training a classifier to distinguish

signal and background observations, which is computationally more efficient and easier to interpret intuitively than the direct $p(\boldsymbol{x}|H_0)/p(\boldsymbol{x}|H_1)$ approximation mentioned before.

From a statistical inference point of view, supervised machine learning framed as the classification of signal versus background can be viewed as a way to approximate the likelihood ratio directly from simulated samples, bypassing the need of a tractable density function (see Section 3.2.1). It is worth noting that because it is only an approximation, in order to be useful for inference it requires careful calibration. Such calibration is usually carried out using a histogram and an holdout dataset of simulated observations, effectively building a synthetic likelihood of the whole classifier output range or the number of observed events after cut in the classifier is imposed (see Section 3.1.3). Alternative density estimation techniques could also be used for the calibration step, which could reduce the loss of information due to the histogram binning.

The effect of nuisance parameters, due to known unknowns, have also to be accounted for during the calibration step. The true density ratio between signal and background depends on any parameter $\boldsymbol{\theta}$ that modifies the signal $p_s(\boldsymbol{x}|\boldsymbol{\theta})$ or background $p_b(\boldsymbol{x}|\boldsymbol{\theta})$ probability densities, thus its approximation using machine learning classification can become complicated. In practice, the classifier can be trained for the most probable likely value of the nuisance parameters and their effect can be adequately accounted during calibration, yet the resulting inference will be degraded. While this issue can be somehow ameliorated using parametrised classifiers [134], the main motivation for using the likelihood ratio - i.e. the Neyman-Pearson lemma - does not apply because the hypothesis considered are not simple when nuisance parameters are present.

Sufficient Statistics Interpretation

Another interpretation of the use of signal versus background classifiers, which more generally applies to any type of statistical inference, is based on applying the concept of statistical sufficiency (see Section 3.1.3). Starting from the mixture distribution function in Equation 4.26, and both dividing and multiplying by $p_b(\boldsymbol{x}|\boldsymbol{\theta})$ we obtain:

$$p(\boldsymbol{x}|\mu, \boldsymbol{\theta}) = p_b(\boldsymbol{x}|\boldsymbol{\theta}) \left(1 - \mu + \mu \frac{p_s(\boldsymbol{x}|\boldsymbol{\theta})}{p_b(\boldsymbol{x}|\boldsymbol{\theta})}\right) \tag{4.31}$$

from which we can already prove that the density ratio $s_{s/b}(\boldsymbol{x}) = p_s(\boldsymbol{x}|\boldsymbol{\theta})/p_b(\boldsymbol{x}|\boldsymbol{\theta})$ (or alternatively its inverse) is a sufficient summary statistic for the mixture coef-

ficient parameter μ, according the Fisher-Neyman factorisation criterion defined in Equation 3.30. The density ratio can be approximated directly from signal versus background classification as indicated in Equation 4.27.

In the analysis presented in Chapter 5 and in the synthetic problem considered in Section 6.5.1, as well as for most LHC analysis using classifiers to construct summary statistics, the summary statistic

$$s_{s/(s+b)} = \frac{p_s(\boldsymbol{x}|\boldsymbol{\theta})}{p_s(\boldsymbol{x}|\boldsymbol{\theta}) + p_b(\boldsymbol{x}|\boldsymbol{\theta})}$$

is used instead of $s_{s/b}(\boldsymbol{x})$. The advantage of $s_{s/(s+b)}(\boldsymbol{x})$ is that it represents the conditional probability of one observation $\boldsymbol{x}$ coming from the signal assuming a balanced mixture, so it can be approximated by simply taking the classifier output. In addition, being a probability it is bounded between zero and one which greatly simplifies its visualisation and non-parametric likelihood estimation. Taking Equation 4.31 and manipulating the subexpression depending on μ by adding and subtracting μ we have:

$$p(\boldsymbol{x}|\mu, \boldsymbol{\theta}) = p_b(\boldsymbol{x}|\boldsymbol{\theta})\left(1 - 2\mu + \mu\frac{p_s(\boldsymbol{x}|\boldsymbol{\theta}) + p_b(\boldsymbol{x}|\boldsymbol{\theta})}{p_b(\boldsymbol{x}|\boldsymbol{\theta})}\right) \tag{4.32}$$

which can in turn can be expressed as:

$$p(\boldsymbol{x}|\mu, \boldsymbol{\theta}) = p_b(\boldsymbol{x}|\boldsymbol{\theta})\left(1 - 2\mu + \mu\left(1 - \frac{p_s(\boldsymbol{x}|\boldsymbol{\theta})}{p_s(\boldsymbol{x}|\boldsymbol{\theta}) + p_b(\boldsymbol{x}|\boldsymbol{\theta})}\right)^{-1}\right) \tag{4.33}$$

hence proving that $s_{s/(s+b)}(\boldsymbol{x})$ is also a sufficient statistic and theoretically justifying its use for inference about μ. The advantage of both $s_{s/(s+b)}(\boldsymbol{x})$ and $s_{s/b}(\boldsymbol{x})$ is that they are one-dimensional and do not depend on the dimensionality of $\boldsymbol{x}$ hence allowing much more efficient non-parametric density estimation from simulated samples. Note that we have been only discussing sufficiency with respect to the mixture coefficients and not the additional distribution parameters $\boldsymbol{\theta}$. In fact, if a subset of $\boldsymbol{\theta}$ parameters are also relevant for inference (e.g. they are nuisance parameters) then $s_{s/(s+b)}(\boldsymbol{x})$ and $s_{s/b}(\boldsymbol{x})$ are not sufficient statistics unless the $p_s(\boldsymbol{x}|\boldsymbol{\theta})$ and $p_b(\boldsymbol{x}|\boldsymbol{\theta})$ have very specific functional form that allows a similar factorisation.

In summary, probabilistic signal versus background classification is an effective proxy to construct summary statistic that asymptotically approximate sufficient statistics directly from simulated samples, when the distributions of signal and background are fully defined and μ (or s in the alternative parametrisation mentioned before) is the only unknown parameter. If the statistical model depends on addi-

tional nuisance parameters, probabilistic classification does not provide any sufficiency guarantees, so useful information about that can used to constrain the parameters of interest might be lost if a low-dimensional classification-based summary statistic is used in place of $\boldsymbol{x}$. This theoretical observation will be observed in practice in Chapter 6, where a new technique is proposed to construct summary statistics, that is not based on classification, but accounts for the effect of nuisance parameters is presented.

4.3.2 Particle Identification and Regression

While the categorical latent variable z_i, denoting the interaction process that occurred in a given collision, is very useful to define an event selection or directly as a summary statistic, information about other latent variables can also be recovered using supervised machine learning. As discussed in Section 2.3.3, event reconstruction techniques are used to cluster the raw detector output so the various readouts are associated with a list of particles produced in the collision. It is possible that in the near future the algorithmic reconstruction procedure might be substituted by supervised learning techniques, training directly on simulated data to predict the set of latent variables at parton level, especially given the recent progress with sequences and other non-tabular data structures. For the time being, machine learning techniques are instead often used to augment the event reconstruction output, mainly for particle identification and fine-tuned regression.

The set of physics objects obtained from event reconstruction, when adequately calibrated using simulation, can estimate effectively a subset of the latent variables $\boldsymbol{z}$ associated with the resulting parton level particles, such as their transverse momenta and direction. Due to the limitations of the hand-crafted algorithms used, some latent information is lost in the standard reconstruction process, particularly for composite objects such as jets. Supervised machine learning techniques can be used to regress some of these latent variables, using simulated data and considering both low-level and high-level features associated with the relevant reconstructed objects. This information could be used to complement the reconstruction output for each object and design better summary statistics, e.g. adding it as an input to the classifiers discussed in Section 4.3.1.

The details of the application of machine learning techniques in particle identification and regression depend on the particle type and the relevant physics case. In the remainder of this section, the application of new deep learning techniques to jet tagging within CMS is discussed in more detail. The integration of deep learning jet

taggers with the CMS experiment software infrastructure was one of the secondary research goals of the project embodied in this document. Leveraging better machine learning techniques for jet tagging and regression could substantially increase the discovery reach of analyses at the LHC that are based on final states containing jets, such as the search for Higgs boson pair production described in Section 5.

Deep Learning for Jet Tagging

The concept of jet tagging, introduced in Section 2.3.3, is based on augmenting the information of reconstructed jets based on their properties to provide additional details about latent variables associated to the physics object which were not provided by the standard reconstruction procedure. Heavy flavour tagging, and in particular b-tagging, is extremely useful to distinguish and select events containing final states from relevant physical interactions. The efficiency of b-tagging algorithms in CMS has been gradually improving for each successive data taking period since the first collisions in 2010. The advance in b-tagging performance, which was already exemplified by Figure 2.12, is mainly due the combined effect of using additional or more accurate jet associated information (e.g. secondary vertex reconstruction or lepton information) and better statistical techniques.

Jet tagging can generally be posed as a supervised machine learning classification problem. Let us take for example the case of b-tagging, i.e. distinguishing jets originating from b-quarks from those originating from lighter quarks or gluon, which can be framed as binary classification problem: predicting wether a jet is coming from a b-quark or not given a set of inputs associated to each jet. The truth label is available for simulated samples, which are used to train the classifier. The CSVv2 b-tagging algorithm (and older variants) mentioned in Section 2.3.3 is based on the output of supervised classifiers trained from simulation, i.e. the combination of three shallow neural network combination depending on vertex information for CSVv2. The CMVAv2 tagger, which is used in the CMS analysis included in Section 5, is instead based on a boosted decision tree binary classifier that uses other simpler b-tagging algorithm outputs as input. Similar algorithms based on binary classification have been also developed for charm quark tagging and double b-quark tagging for large radius jets.

The first attempt to use some of the recent advances in neural networks (see Section 4.2.2) for jet tagging within CMS was commissioned using 2016 data, and it is referred to as DeepCSV tagger. The purpose for the development of this tagger was to quantify the performance gain due to the use of deep neural networks for jet

tagging in CMS, which was demonstrated effective using a simplified detector simulation framework [135, 136]. Thus, a classifier based on a 5-layer neural network, each layer with 100 nodes using ReLU activation functions, was trained based on the information considered for the CSVv2 tagger. A vector of variables from up to six charged tracks, one secondary vertex and 12 global variables was considered as an input, amounting to 66 variables in total. Another change with respect to previous taggers is that flavour tagging is posed as a multi-class classification problem, which is a principled and simple for tacking the various flavour tagging problems simultaneously.

Five exclusive categories were defined based different on the generator level hadron information[1]: the jet contains exactly one B hadron, at least two B hadrons, exactly one C hadrons and no B hadrons, at least two C hadrons and no B hadrons, or none of the previously defined categories. The softmax operator (see Equation 4.11) was used to normalise the category output as probabilities and construct a loss function based on cross entropy (see Equation 4.10). As was shown in Figure 2.12 for b-tagging performance, the DeepCSV tagger is considerably better than CSVv2 for the b-jet efficiency/misidentification range - e.g. about 25% more efficient at light jet and gluon mistag rate of 10^{-3}. In fact, DeepCSV outperforms the CMVAv2 super-combined tagger, which uses additional leptonic information. While not shown in this document, the performance for c-tagging was found also comparable with dedicated c-taggers [85].

The very favourable results obtained for DeepCSV motivated the use of newer machine learning technologies, such as convolutional and recurrent layers, which were readily available in open-source software libraries [137, 129], as well as advances in hardware (i.e. more powerful GPUs for training). The large amount of jets available in simulated data, e.g. in 2016 about 10^9 $t\bar{t}$ events were simulated for CMS (each with two b-quarks and probably several light quarks), conceptually justifies the use of more complex machine learning models because over-fitting is unlikely. Thus, a new multi-class jet tagger referred to as DeepJet (formerly know as DeepFlavour) was developed, whose architecture is depicted in Figure 4.3, that can be characterised by a more involved input structure and both convolutional and recurrent layers.

Instead of a fixed input vector, optionally padded with zeroes for the elements that did not exist (e.g. not reconstructed secondary vertex has been reconstructed), a

[1] Here by B and C hadrons we refer to hadrons containing b-quarks c-quarks as valence quarks respectively, which often have a lifetime large enough to fly away from the primary vertex as discussed in Section 2.3.3.

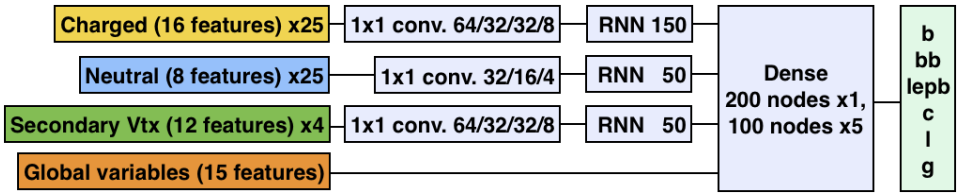


Figure 4.3: Scheme of DeepJet tagger architecture. Four different sets of inputs are considered: a sequence of charged candidates, a sequence of neutral candidates, a sequence of secondary vertices and a 15 global variables. Sequences go first through a series of 1x1 convolution filter that learn a more compact feature representation and then through a recurrent layer that summarises the information of the sequence to in a fixed size vector. All the inputs are then feed to a 7-layer dense network. A total of six exclusive output categories are considered depending on the generator-level components: b, bb, leptonic b, c, light or gluon. Figure adapted from [138].

complex input object is considered for DeepJet. Variable-size sequences are directly taken as input for charged candidates, neutral candidates and secondary vertices; each element in the sequence characterised by 16, 8 and 12 features respectively. Each of the three input sequences go through a 3-layers of 1x1 convolutions in order to obtain a more compact element representation, 8-dimensional for charged candidates and secondary vertices and 4-dimensional for neutral candidates. The output of the convolutional layers is connected with a recurrent layer, which transforms a variable-size input to fixed-size embedding. The fixed-size outputs after the recurrent layer, as well as a set of 15 global jet variables, are feed into a 6-layer dense network with 100 (200 for the first layer) cells with ReLU activation functions per layer.

A total of six mutually exclusive output categories are considered based on the generator-level particle content associated to the jet:

- b - exactly one B hadron that does not decay to a lepton.
- bb - at least two B hadrons.
- $lepb$ - one hadron B decaying to a soft lepton
- c - at least one C hadron and no B hadrons
- l - no heavy hadrons but originated from a light quark
- g - no heavy hadrons but was originated from a gluon.

The DeepJet tagger aims to provide gluon-quark discrimination in addition to b-tagging, c-tagging and double b-tagging. The output probabilities are normalised by using the softmax operator (see Equation 4.11). The training loss function was constructed based on cross entropy (see Equation 4.10). Additional details regarding the architecture and training procedure are available at [139].

The b-tagging performance of DeepJet, by means of the misidentification versus efficiency curve compared with the DeepCSV tagger, is shown in Figure 4.4. The additional model complexity and input variables lead to a clear performance improvement, about a 20% additional efficiency at a mistag rate of 10^{-3} for light quark and gluon originated jets. Larger relative enhancements with respect to DeepCSV are seen for b-jet versus c-jet identification. The performance for c-tagging and quark-gluon discrimination is slightly improved in comparison with dedicated approaches, with the advantage of using a single model for all the flavour tagging variations. The expected relative performance boost, especially when compared non deep learning based taggers (CSVv2 or CMVA) can increase significantly the discovery potential for analyses targeting final states containing several b-tagged jets, such as the one presented in Chapter 5. In addition similar model architectures have since been successfully applied to large radius jet tagging [140] and could be also extended to other jet related tasks, as providing a better jet momenta estimation by means of a regression output.

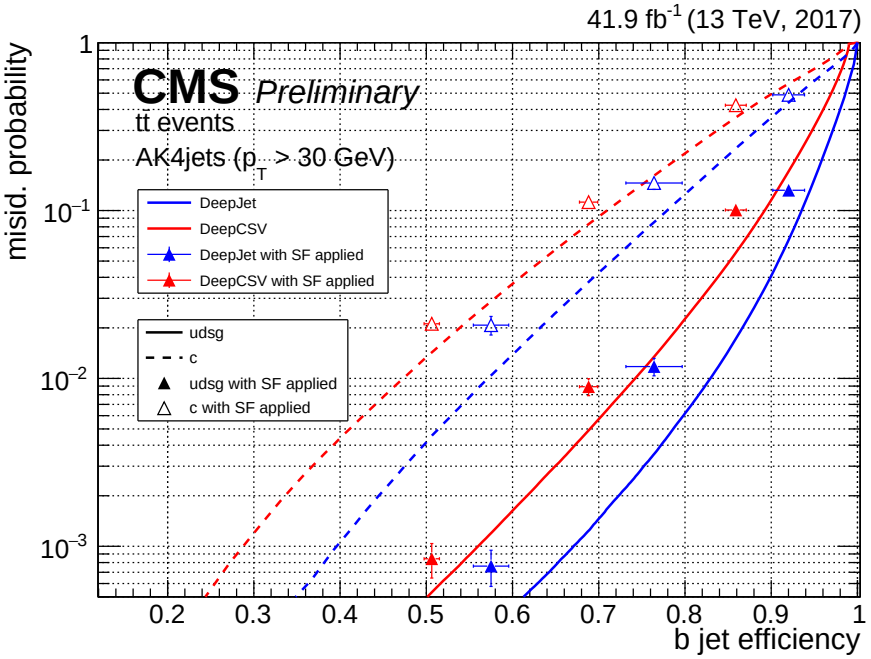


Figure 4.4: Misidentification probability (in log scale) for jets originating from c quarks (dashed lines) or light quarks and gluons (solid lines) as a function of the b-tagging efficiency for both DeepCSV and DeepJet taggers. The corrected mistag/efficiency and its uncertainty for the loose, medium and tight working points are also included. Figure adapted from [138].

While both advances in model architecture and the addition of input features allow notable jet tagging performance gains, they can complicate the integration of these tools within the CMS experiment software framework [141], which is often referred as CMSSW. Training and performance evaluation of both DeepCSV and DeepJet was carried out using the Keras [137] and TensorFlow [129] open-source libraries. In order to integrate jet tagging models in the standard CMS reconstruction sequence, which has rather stringent CPU and memory requirements per event because it is run for both acquired and simulated data in commodity hardware in a distributed manner around the world in the LHC computing grid [142]. In addition, the lwtnn open-source library [143], a low-overhead C++ based interface used for the integration of DeepCSV did not support multi-input models with recurrent layers at the time.

An alternative path to integrate DeepJet into production was thus required. Given than TensorFlow backend is based on the C++ programming language and a basic interface to evaluating training was also provided, the direct evaluation of machine learning model using its native TensorFlow backend was chosen as the best alternative. In addition, this way the integration effort and basic interface developed could be re-used in future deep learning use cases in the CMS experiment (e.g. large radius jet tagging), leading to the development of the CMSSSW-DNN module [144]. The integration process was made more challenging due to the difficulty recovering the same features at reconstruction level, the strict memory requirements and multi-threading conflicts. After resolving all the mentioned issues [145], the output of the DeepJet model at production was verified to match that of the training framework [146] to numerical precision. The successful integration, that is currently in use, facilitated the measurement of DeepJet b-tagging performance on data for the main discriminator working points, as shown in Figure 4.4.

5 Search for Anomalous Higgs Pair Production with CMS

All Life is Problem Solving.

Karl Popper

In this chapter, the concepts and techniques from the previous sections are applied in the search for non-resonant production of Higgs boson pairs, using data from proton-proton collisions at a centre-of-mass energy of 13 TeV collected in 2016 by the CMS detector at the LHC, corresponding to a total integrated luminosity of $35.9\,\mathrm{fb}^{-1}$. The most probable decay channel for the Higgs boson pairs, where each Higgs boson leads to a $\mathrm{b\bar{b}}$, is considered. While the aforementioned final state is the most frequent by a considerable margin, a large background of similar events is expected from multi-jet QCD processes, which motivates the use of machine learning techniques to construct a summary statistic that can exploit the fine differences between signal and background for statistical inference. In fact, the expected background is so copious that is not possible to generate a sufficiently large number of simulated observations to obtain the required level of modelling accuracy, thus we have to resort to the development of a new data-driven background estimation technique referred to as hemisphere mixing [147]. In addition to setting upper limits on the Standard Model (SM) production of Higgs boson pairs, the data analysis framework is also used to set upper limits in the context of effective field theories (EFT) of anomalous couplings, that parametrise possible deviations from the SM. The main results presented in this section have been carried out within the CMS Collaboration, and have been made public and published [148].

5.1 Introduction

After the discovery of the Higgs boson (H) in 2012 with the LHC experiments [2, 3, 149], the detailed study of its properties has become one of the most important

topics in fundamental physics. The experimental determinations of its couplings and production production rates by the CMS and ATLAS collaborations [27, 150], including the recent observations of the associated production of the Higgs boson with a $t\bar{t}$ quark pair [151, 152], are found to be compatible with the Standard Model (SM) theoretical predictions. That said, several predicted properties remain unmeasured because of the difficulty of their experimental determination. Among them, the Higgs boson self-coupling being one of the most relevant parameters since it can be modified by physics beyond the standard model (BSM) [153, 154, 155, 156, 157].

A principled way to determine the Higgs self-coupling, and thus reconstruct the scalar potential of the Higgs field that is responsible for spontaneous symmetry breaking described in Section 1.1.4, is to measure the production of Higgs boson pairs (HH) [158]. The SM prediction for the inclusive HH production cross section for 13 TeV proton-proton collisions, assuming $m_{\mathrm{H}} = 125.09$ GeV [27, 159], can be theoretically calculated [160, 161, 162, 163, 164] obtaining:

$$\sigma(pp \to HH + jets) = 33.49^{+4.3\%}_{-6.0\%}(\text{scale}) \pm 2.3\%(\alpha_S) \pm 2.1\%(\text{PDF}) \text{ fb} \qquad (5.1)$$

where the listed sources of uncertainties correspond to factorisation μ_{R} and renormalisation μ_{F} scales, uncertainties in the strong coupling constant α_S, and the uncertainty associated with the parton distribution functions (PDF), respectively. The predicted cross section of the HH production process in the SM is very small, several orders or magnitude smaller than that of single Higgs production, and thus has not been directly observed the LHC data yet and will require targeted studies at the HL-LHC or other future colliders. New physics effects beyond the SM can enhance the HH production cross sections, e.g. as can be modelled by effective theories of anomalous couplings [165], in a way so HH production could be observed with the data already collected at the LHC.

The search of possible beyond the SM enhancements of HH production motivated early searches using $\sqrt{s} = 8$ TeV LHC data [166, 167], as well as several analyses using data collected during 2015 and 2016 at the LHC experiments, including the one presented in this work. Several analyses looking for an enhancement of resonant HH production, leading to a peak in the reconstructed invariant mass of the Higgs pair due to decay of the hypothetical mediating particle, have also been performed, Such mechanism for the production of Higgs boson pairs is not considered in this analysis. Regarding non-resonant production of HH pairs at at $\sqrt{s} = 13$ TeV, both ATLAS and CMS collaborations have carried out searches for different decay channels

including $b\bar{b}b\bar{b}$ [168], $b\bar{b}l\nu l\nu$ [169], $b\bar{b}\tau\tau$ [170] and $b\bar{b}\gamma\gamma$ [171]. In all the mentioned analyses, one of the Higgs bosons decays to a $b\bar{b}$ quark pair, which its the most likely decay model (with a branching fraction of 57.7% for $m_H = 125$ GeV), in order to consider a large fraction of expected HH decays. The CMS Collaboration has also carried out an analysis complementary to the one presented here, where one of the $b\bar{b}$ is highly boosted and thus reconstructed as a single large-area jet [172]. The most stringent expected upper limit on the SM HH production cross section to date, which corresponds to a 95% C.L. exclusion for rates about 19 times the SM prediction, was obtained by the CMS $b\bar{b}\gamma\gamma$ channel search [170], which yielded an observed upper limit of 22 times the SM. The ATLAS $b\bar{b}b\bar{b}$ channel search has a similar experimental reach [168], studying the same final state considered in this analysis, however with a different methodology regarding their summary statistic and background estimation.

A detailed description of the main characteristics and results of an analysis searching for HH production using CMS experiment data, with both Higgs bosons decaying into $b\bar{b}$ quark pairs, is included in this chapter. The data considered was acquired by the CMS detector during the year 2016, corresponding to an integrated luminosity of 35.9 fb^{-1}. In the final state considered here, each of the four b quark results in a distinct reconstructed jet. While it is the most likely decay mode for the Higgs pair, a much larger quantity of similar events with four or more jets are expected from hard quantum chromodynamics (QCD) interactions. The differences between signal and background are used to increase the sensitivity by using as a summary statistic the prediction of a multivariate probabilistic classifier. Because the expected contribution from the QCD multi-jet processes is so abundant, it could not be modelled with the required precision with the available simulations. To address this issues, a method for carrying out a fully data-driven background estimation was developed, that is described in Section 5.6.

5.2 Higgs Pair Production and Anomalous Couplings

At proton-proton colliders, the main production mechanism for a Higgs pair is *gluon fusion*. The gluon fusion interaction at leading order includes a fermion loop as depicted in the top diagrams of Figure 5.1, which is largely dominated by the contribution from top and bottom quarks, and thus explaining the low expected production rate listed in Equation 5.1. The most common production mode, labelled as (b) in Figure 5.1, features a triangular fermion loop followed by the production of an off-self Higgs boson, that in turn decays on two on-shell Higgs bosons via a triple Higgs bo-

son interaction vertex. In addition, within the SM is also possible to produce a pair of Higgs bosons at leading order through a fermion box loop, as shown in diagram (a) of Figure 5.1, which evidently does not depend on the Higgs self-coupling. Both box and triangle loop contributions interfere destructively in the SM amplitude to give rise to the total HH production.

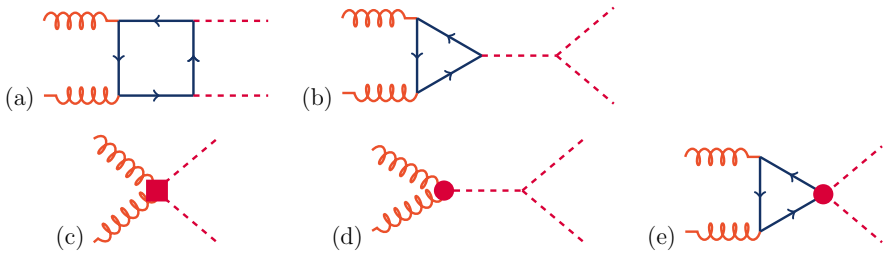


Figure 5.1: Set of HH production Feynman diagrams, representing all gluon-induced processes at leading order. The interactions depicted by (a) and (b) represent processes that are expected within the SM, while the contact interactions between the Higgs bosons and gluons (c) and (d), as well the contact interaction of two Higgs bosons with top quarks (e), are effective diagrams of BSM interactions. Figure adapted from [148].

New physics at higher energy scales can affect processes and observables at the electroweak scale, such as Higgs pair production. As reviewed in Section 1.2.2, the effective field theory (EFT) approach is a way to calculate observables of possible extensions of the SM without being tied to a certain class of BSM model, by adding non-renormalisable local interactions. In the context of Higgs pair production, the effect of new operators can be parametrised by the following effective Lagrangian:

$$\begin{aligned}\mathcal{L}_{\mathrm{H}} = &\frac{1}{2}\partial_\mu \mathrm{H}\,\partial^\mu \mathrm{H} - \frac{1}{2}m_{\mathrm{H}}^2\mathrm{H}^2 - \kappa_\lambda\,\lambda_{\mathrm{SM}} v\,\mathrm{H}^3 \\ &-\frac{m_{\mathrm{t}}}{v}(v + \kappa_{\mathrm{t}}\,\mathrm{H} + \frac{c_2}{v}\mathrm{HH})\,(\bar{\mathrm{t}}_{\mathrm{L}}\mathrm{t}_{\mathrm{R}} + \mathrm{h.c.}) \\ &+\frac{1}{4}\frac{\alpha_{\mathrm{S}}}{3\pi v}(c_{\mathrm{g}}\,\mathrm{H} - \frac{c_{2\mathrm{g}}}{2v}\,\mathrm{HH})\,G^{\mu\nu}G_{\mu\nu}\end{aligned} \tag{5.2}$$

where $v = 246$ GeV is the vacuum expectation value of the Higgs field. After neglecting the enhanced coupling of the Higgs boson with bottom quarks due its experimental constraints and the presence of new light particles, a total of five EFT parameters remain, which are highlighted by using red colour in Equation 5.2. The factors $\kappa_\lambda = \lambda_{\mathrm{HHH}}/\lambda_{\mathrm{SM}}$ and $\kappa_{\mathrm{t}} = y_{\mathrm{t}}/y_{\mathrm{SM}}$ account for possible deviations from the

SM of the Higgs boson trilinear coupling and the top quark Yukawa coupling, thus effectively modifying the relative weight of the SM Feynman diagrams described at the beginning of the section. The absolute parameters c_g, c_{2g} and c_2 instead lead to new contact interactions not expected within the SM, represented in the (c), (d) and (e) Feynman diagrams of Figure 5.1, and which could arise by mediation of heavy particles beyond the electroweak scale. The previous parametrisation is commonly referred to as dimension-six non-linear or anomalous couplings EFT, however alternative approaches exist, such as the so-called linear EFT [173] which is more appropriate to model smaller BSM effects.

A theoretical prediction for the differential and total cross section for each point in the mentioned five-dimensional EFT parameter space $(\kappa_\lambda, \kappa_\mathrm{t}, c_2, c_\mathrm{g}, c_\mathrm{2g})$ can be computed as outlined in Section 1.3. The distribution of the final state kinematical variables, i.e. the relative angles and momenta of the Higgs pair, can depend substantially on the value of some of these couplings. A naive grid or random scan of the full five-dimensional space would require simulated samples of observations at too many EFT points and hence it is not feasible. While this signal modelling issue could be tackled by means of event re-weighting, as described in Section 3.1.2, it is useful to consider a different methodology to represent the main properties of the anomalous couplings parameter space where only a reduced number of EFT points are considered.

For the analysis presented in this work, a total of twelve EFT points referred to as *benchmarks* are considered, which have been chosen via a agglomerative clustering procedure so they represent the main kinematical topologies in the parameter space. The details of the clustering methodology are detailed in [174], but they amount to the construction of a distance between the main kinematic distributions at generator level of each pair of EFT points. The parameters corresponding to each of the benchmarks, as well as those corresponding to the SM model and the case where Higgs boson self coupling is zero, are included in Table 5.1.

5.3 Analysis Strategy

The goal of this analysis is to carry out statistical inference on the occurrence of $\mathrm{pp} \to \mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$, as predicted by the SM or in BSM effective field theory extensions, based on experimental data acquired by the CMS detector on 2016. The type of statistical inference applicable to this search is hypothesis testing, as introduced in Section 3.2.2. In principle, we would like to test whether the null hypothesis

Table 5.1: Effective field theory parameters for the anomalous couplings benchmarks considered in this analysis, as defined in [174], as well as the modified couplings corresponding to the Standard Model.

Benchmark point	κ_λ	κ_t	c_2	c_g	c_{2g}
1	7.5	1.0	-1.0	0.0	0.0
2	1.0	1.0	0.5	-0.8	0.6
3	1.0	1.0	-1.5	0.0	-0.8
4	-3.5	1.5	-3.0	0.0	0.0
5	1.0	1.0	0.0	0.8	-1.0
6	2.4	1.0	0.0	0.2	-0.2
7	5.0	1.0	0.0	0.2	-0.2
8	15.0	1.0	0.0	-1.0	1.0
9	1.0	1.0	1.0	-0.6	0.6
10	10.0	1.5	-1.0	0.0	0.0
11	2.4	1.0	0.0	1.0	-1.0
12	15.0	1.0	1.0	0.0	0.0
Box	0.0	1.0	0.0	0.0	0.0
SM	1.0	1.0	0.0	0.0	0.0

H_0 corresponding to the SM without HH production hypothesis can be rejected. Several alternate hypothesis H_1 are considered, which are based on the SM including HH production processes, either coming from SM production models or from EFT extensions. However we do not expect to reject the H_0 hypothesis, so the objective is the one of setting exclusion upper limits on the signal cross section for a given model including Higgs pair production. This, we would like to adopt an analysis strategy that maximises the sensitivity to the presence of HH production, which amounts to minimising the Type II error rate for a given fixed Type I error rate in statistical terms. The Type II error rate would in turn depend on the alternate hypothesis H_1 considered, which for the optimisation of the analysis strategy would be the SM including HH production through SM processes at an enhanced rate.

The event selection in this analysis will include some custom online requirements, which were set at trigger level to reduce the total rate of data collection while keeping a large fraction of events relevant for this analysis, as well as an offline selection to reduce the contribution of background processes that are not well modelled, in order to simplify the construction of powerful summary statistics. The online trigger requirements as well as the characteristics of the datasets considered in this analysis are described in Section 5.4, while the adopted event selection is described in detail in Section 5.5.

After a basic event selection, mainly comprising the filtering of events with four or more b-tagged jets[1], a subset including four of the reconstructed jets within each event is paired to construct two *di-jet candidates*, as an attempt to recover the kinematic properties of the Higgs bosons, including their reconstructed masses. The information from the two di-jet candidates can in turn be combined to compute variables that can approximate the features of the Higgs pair system, which are also quite useful for inference. A set of variables from the selected jets, the H candidates and the HH system, are combined in a single discriminating variable obtained by training a probabilistic classification model, specifically machine learning model based on boosted decision trees (see Section 4.2.1), to separate signal from background, in a analogous manner to what was described in Section 4.3.1.

The statistical inference in this analysis is based on constructing a binned likelihood of the expected distribution of the classifier output for events originated from signal and background processes. This likelihood, which also accounts for the effect of nuisance parameters as discussed in Section 3.1.3, is used to extract information the about the parameter of interest (i.e. HH production cross section times the branching ratio) based on the observed data. While both the SM and the various BSM signal models can be modelled using simulated observations, the main background of the analysis, multi-jet QCD production, is hard to model by simulation. Thus a data-driven background estimation method, described in detail in Section 5.6, is used both for training the probabilistic classifier and for modelling the background contribution in the binned likelihood.

After including the effect of the relevant sources of systematic uncertainty, which are listed in Section 5.7, upper limits are obtained for the $pp \to HH \to b\bar{b}b\bar{b}$ cross section for each of the benchmarks listed in Table 5.1, as well as for the SM HH production process. The results, which are contained in Section 5.8, include the upper limit on the mentioned cross section a function of the Higgs self-coupling factor parameter κ_λ when $\kappa_t = 1$ and the other EFT parameters are null. While the analysis could be redone for any arbitrary EFT point by recomputing the limits for the particular model, given that the benchmarks have been constructed to represented the main differential cross section differences in a large part of the EFT parameter space, approximate limits can be obtained by considering the limit obtained for the closest benchmark using the distance measure from [174].

[1] Events with a different b-tagged jet definition will be also used to define a data control region, as will be discussed in Section 5.6.2.

5.4 Trigger and Datasets

The experimental data considered in this analysis was collected by the CMS detector in 2016 from proton-proton collisions at centre-of-mass energy $\sqrt{s} = 13$ TeV. The total integrated luminosity at the CMS interaction point corresponding to the certified set of datasets used in this analysis is 35.9 fb^{-1}, which is the subset of data corresponding to periods when the relevant detecting systems were running regularly and no problematic anomalies were discovered during data quality monitoring (DQM).

Because the rates for the main background processes of this analysis - events originating from QCD multi-jet events - are expected to be much higher that those of the signal, an efficient online trigger selection is essential for maximising the sensitivity of the analysis. While the set of standard CMS trigger path includes ones that select events with several high-energy jets, a more practical strategy is to include some b-tagging requirements within the high-level trigger sequence. Hence, this analysis re-uses the multi-jet trigger paths that were developed for the search of the resonant process $\text{pp} \to \text{X} \to \text{HH} \to \text{b}\bar{\text{b}}\text{b}\bar{\text{b}}$ [175], where X is a heavy mediating particle. These two paths both require that at least three jets have are b-tagged by the online version of the Combined Secondary Vertex (CSV) algorithm [85].

The full specification trigger selection used is rather complex, however it may be represented by a logical OR of the following two HLT trigger paths that were in place during the CMS 2016 data taking period:

- HLT_DoubleJet90_Double30_TripleBTagCSV_p087
- HLT_QuadJet45_TripleBTagCSV_p087

which represent a particular online selection sequence at the HLT. The sequence is preceded by a given set of L1 trigger seeds, as conceptually reviewed in Section 2.2.7. The L1 trigger paths are different for each of the HLT paths, but are based on the logical OR between several conditions requiring a certain number of L1 jets over a given energy or the total deposited energy on the calorimeter H_T to be over a certain threshold. At the HLT, both paths require some quality criteria on the reconstructed primary vertex and at least 4 reconstructed jets within a pseudo-rapidity range defined by $|\eta| < 2.6$. The first path in addition requires that the momenta of two of the reconstructed jets satisfy the requirement $p_T > 90$ GeV, while two other jets are required to have $p_T > 30$ GeV. The second path instead requires that the event contains at least four reconstructed jets with $p_T > 45$ GeV. As mentioned, both paths include a b-tagging requirement, chiefly that the value of

the online CSV discriminator is larger than the value of 0.87, which is defined as the "medium working point" of the algorithm, for three of the eight most energetic reconstructed jets in the event.

Samples of simulated observations from Higgs pair production are generated using MadGraph5_aMC@NLO [176] at leading-order, following the relevant prescriptions, including the loop factor on an event-by-event basis detailed in [177]. A total of 300,000 events have been simulated for the SM model production component, as well as an older version of the clustering benchmarks discussed in Section 5.2 and the $\kappa_\lambda = 0$ box model. Regading the parton distribution function used for generation, the NNPDF30_LO_AS_0130_NF_4 n set [178] was used for all samples.

The datasets for the benchmark points listed in Table 5.1, or any other EFT point for that matter, can be generated from the previous samples by means of generator re-weighting. As described in Section 3.1.2, the latent variables of the simulator can be used to model a different point of the parameter space of the the underlying theory by computing observables after assigning to each event a weight proportional to the ratio between probability density functions. In this case, the effect of varying EFT parameters in Equation 5.2 can be fully characterised by two parton variables at leading order: the Higgs pair invariant mass m_{HH} and the $|\cos\theta^*|$, where θ^* is the polar angle of any one of the Higgs bosons with the respect to the beam axis. Once these two variables are specified, the rest of the simulation does not depend on the EFT parameters. A set of HH production simulated events generated for a given vector of EFT parameters $\boldsymbol{\theta}_{\text{EFT}} = (\kappa_\lambda, \kappa_{\text{t}}, c_2, c_{\text{g}}, c_{2\text{g}})$ re-weighted by:

$$w(m_{\text{HH}}, |\cos\theta^*|) = \frac{p\big(m_{\text{HH}}, |\cos\theta^*| \;\big|\; \boldsymbol{\theta}'_{\text{EFT}}\big)}{p(m_{\text{HH}}, |\cos\theta^*| \;\big|\; \boldsymbol{\theta}_{\text{EFT}})} \tag{5.3}$$

could be used to model events generated at the EFT point $\boldsymbol{\theta}'_{\text{EFT}}$, as long as the both the numerator and denominator are not zero. The previous concept can be extended to any arbitrary probability distribution of $p(m_{\text{HH}}, |\cos\theta^*|)$, e.g. a large sample uniformly distributed in the mentioned 2D-space could be re-weighted to model any EFT parameter point. While the density ratio in Equation 5.3 can also be estimated exactly as the ratio between the matrix elements [179], a non-parametric density estimation approach was adopted in this analysis.

A large sample of HH production events was formed by concatenating all non-resonant Higgs pair events simulated from each of the 14 samples, creating what will be referred to as the *pangea* sample. For all the EFT points of interest, 50,000 events (300,000 for the SM production) were generated at parton level, which is

rather inexpensive. The per-event weight in Equation 5.3 is estimated by the ratio of 2D-histograms, which effectively approximate the mentioned density ratio. The *weighted pangea* sample can represent any EFT parameter point at leading order by this procedure, so it is used to model the signal characteristics of all the models considered in this work.

5.5 Event Selection

Given that the final state studied in this analysis is characterised by the presence of four highly energetic b quarks, the physics objects of relevance are reconstructed jets. The details of the reconstruction procedure at CMS were already discussed in Section 2.3.3. Advanced jet flavour tagging, in particular b-tagging, is also essential to distinguish jets that originate from b quarks from those originating from lighter quarks and gluons, and thus very useful to reduce the contribution from a large number of QCD multi-jet processes.

The subset of collected events that pass the trigger requirements, as well as all the simulated events, as listed at the beginning of Section 5.4 undergo a process of event reconstruction, producing a representation of the detector readout that attempts to recover the latent particle features at parton level, as discussed in Section 3.1.3. The first step of the offline event selection is to consider for each event the set of reconstructed particle-flow jets with $p_T > 30$ GeV and $|\eta| < 2.4$. An event is only selected if four or more jet passing those requirement are found.

After filtering out jets with lower energy or falling out of the tracker acceptance, at least four of the remaining jets are required to be b-tagged to consider the event in the final selection. The medium working point of the CMVA discriminator [85], defined as the value of the discriminator for which the expected mis-identification of light quarks and gluons is 1%, is used as b-tagging criteria. The object selection efficiency for jets originating from the b quarks produced in the decay of the Higgs boson pairs has been estimated from simulated samples to be around 65%. For the SM HH production process, the absolute and relative selection efficiencies of the trigger and offline selection, and the total number of expected events per fb^{-1}, are included in Table 5.2, as estimated from the simulated events.

The goal of the previous selection is to reduce the contribution from QCD multi-jet processes and to isolate the set of signal events where all the jets from the Higgs pair decays can be fully reconstructed. After such selection, the most often occurring value for the number of jets in the selected subset of events is five. The four jets

Table 5.2: Event selection efficiency and number of events expected per each integrated fb^{-1} of integrated luminosity for the Standard Model $pp \to HH \to b\bar{b}b\bar{b}$ production process, as estimated using simulated events.

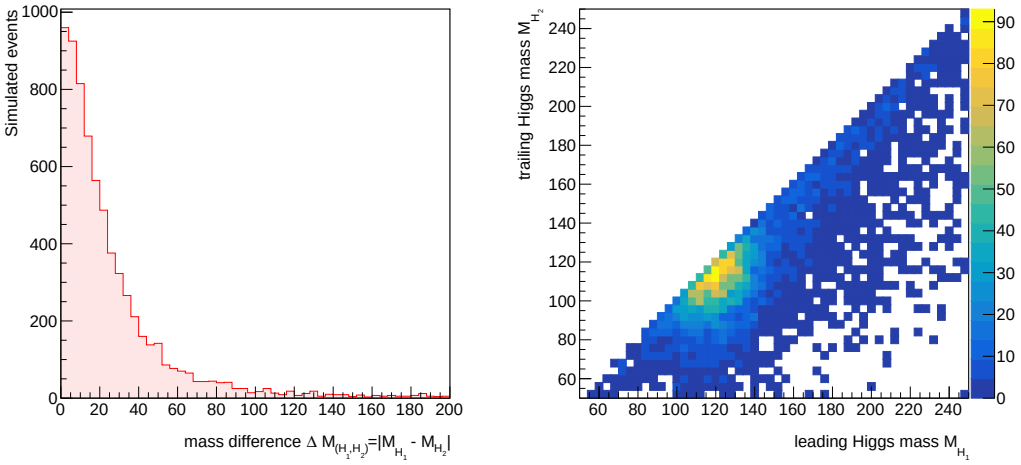
	Produced	Trigger	≥ 4 btags
N events / fb	11.4	3.9	0.22
Relative eff.		34%	5.6%
Efficiency		34%	1.9%

with highest CMVA discriminant are chosen as candidate decay products of the Higgs bosons. In order to reconstruct features of the Higgs boson candidates, a pairing between the selected jets has to be defined. The pairing used in this analysis is rather simple, the invariant masses for the two Higgs candidates M_{H_1} and M_{H_2} are computed for the three possible combinations of the four decay candidate jets, and the invariant mass difference $\Delta M_{(H_1,H_2)}$ is computed for each combination:

$$\Delta M_{(H_1,H_2)} = |M_{H_1} - M_{H_2}| \tag{5.4}$$

so the combination with the smallest mass difference is taken. Alternative decay candidates selection and pairing techniques were considered and tested. The fact that the chosen procedure does not explicitly use the mass of the Higgs boson makes it very effective to avoid conditioning also the distributions of the background processes. The aforementioned procedure correctly pairs the jets to form Higgs candidates in approximately 54% of the events. The distribution of $\Delta M_{(H_1,H_2)}$ and M_{H_1} versus M_{H_2} is shown in Figure 5.2. To distinguish between the two Higgs candidates during the rest of this chapter, the term leading Higgs H_1 will be used for the reconstructed Higgs candidates with the largest invariant mass while trailing Higgs H_1 for the other candidate.

In this analysis, the final summary statistic considered for inference is based on the output of classifier that discriminates signal and background observations, which will approximate the likelihood ratio or a sufficient summary statistic if the signal and background components are fully specified, as discussed in Section 4.3.1. The machine learning classification technique used is based on gradient boosted decision trees (BDT), a technique that was summarised in Section 4.2.1. The implementation from the XGBOOST software library [123] was used to train a probabilistic classifier using a set of simulated events corresponding to SM Higgs pair production (i.e. 60% of the weighted pangea observations) and background artificial events resulting from the data-driven procedure which will be described in Section 5.6.



(a) mass difference $\Delta M_{(\mathrm{H}_1,\mathrm{H}_2)}$ (b) M_{H_1} versus M_{H_2}

Figure 5.2: Mass difference $\Delta M_{(\mathrm{H}_1,\mathrm{H}_2)}$ (left) and 2D histogram of M_{H_1} versus M_{H_2} (right) for simulated signal observations. Only the lower right part of the right plots includes observations because the Higgs candidates are ordered by mass.

The set of features, or input variables, which are fed to the probabilistic classifier are listed in Table 5.3. The set of variables are divided in three subgroups, the first corresponding to variables related with the properties of the reconstructed Higgs pair HH system, which are compared for signal and background in Figure 5.3, including its invariant mass M_{HH}, its total transverse momentum $p_T^{\mathrm{H_1H_2}}$ and $\cos\theta^{\star}_{\mathrm{H_1H_2-H_1}}$, where $\theta^{\star}_{\mathrm{H_1H_2-H_1}}$ is the angle between the HH system and the leading Higgs boson candidate. Another feature that is found to increase the discrimination power of the classifier is the M_{X} variable, defined as:

$$M_{\mathrm{X}} = M_{\mathrm{HH}} - (M_{\mathrm{H}_1} - M_{\mathrm{H}}) - (M_{\mathrm{H}_2} - M_{\mathrm{H}}) \tag{5.5}$$

where $M_{\mathrm{H}} = 125$ GeV is the Higgs boson mass. The second group of features includes variables associated individually with each Higgs boson candidate (see Figure 5.4 for a comparison of marginal distributions), such as the reconstructed mass of each paired di-jet system M_{H_1} and M_{H_2}. The reconstructed Higgs candidate masses have the largest discrimination power, because their marginal distributions are expected to peak around $M_{\mathrm{H}} = 125$ GeV for the subset of well-paired signal events while more spread for background events. Other features in this sub-group include the transverse momenta of the reconstructed Higgs candidates $p_T^{\mathrm{H}_1}$ and $p_T^{\mathrm{H}_2}$, the angular distances

between their component jets $\Delta R_{jj}^{\mathrm{H}_1}$, $\Delta R_{jj}^{\mathrm{H}_2}$, $\Delta\phi_{jj}^{\mathrm{H}_1}$, $\Delta\phi_{jj}^{\mathrm{H}_2}$, and $\cos\theta^{\star}_{\mathrm{H}_1\mathrm{H}_2-\mathrm{H}_1}$, where $\theta^{\star}_{\mathrm{H}_1\mathrm{H}_2-\mathrm{H}_1}$ is the angle between the leading Higgs boson candidate and the leading jet. The last group includes variables directly associated to the reconstructed jets, including the transverse momenta $p_{T_j}^{(i=1-4)}$ and pseudo-rapidity $\eta^{(i=1-4)}$ of the first four jets, ordered by their value of the CMVA b-tagging discriminant as well as the scalar sum of their transverse momenta H_T. Finally, the scalar p_T sum of all the jets that were not used for the reconstruction of the Higgs pair system H_T^{rest} and the b-tagging CMVA discriminant value for the third and fourth jet CMVA_3, CMVA_4 are also used. The marginal comparison of the distributions of signal and background for jet-based based variables is shown in Figure 5.5.

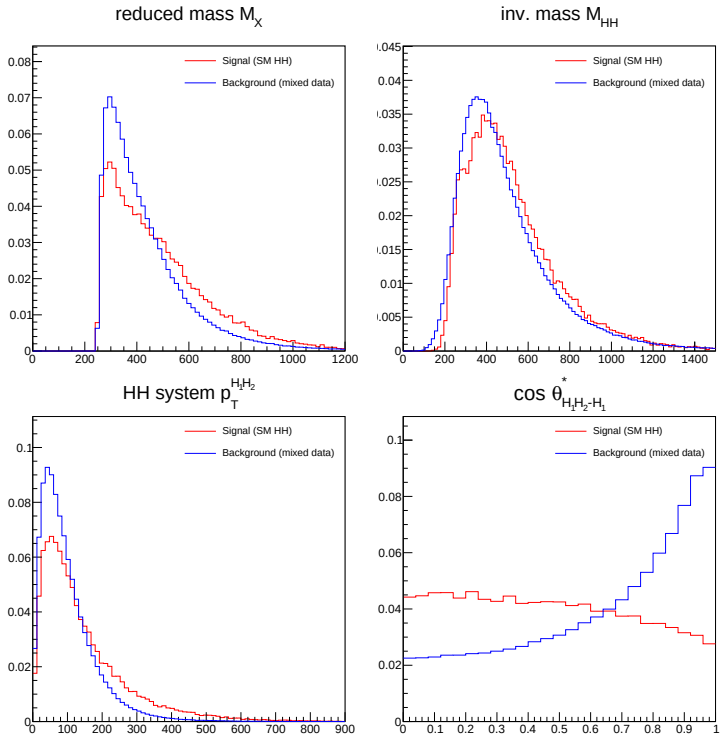


Figure 5.3: Comparison of the signal (SM HH production) and background (mixed data) distributions for the HH system features considered in the probabilistic classifier. See Table 5.3 and associated text for more details.

The trained classifier combines the 25 variables from Table 5.3 in a single scalar value, that approximates the conditional probability of belonging to the signal conditional on the input $p(y = 1|\boldsymbol{x})$, which depends on the relative frequencies of signal and background events in the training dataset, as discussed in Section 4.1.1. For training, signal and background observations were weighted so as to represent the

Table 5.3: List of reconstruction-based features used as input of the probabilistic classifier.

HH system	H candidates	Jet variables
M_{X}, M_{HH},	M_{H_1}, M_{H_2}	$p_{T_j}^{(i=1-4)}$, $\eta^{(i=1-4)}$,
$p_T^{\mathrm{H}_1\mathrm{H}_2}$	$p_T^{\mathrm{H}_1}$, $p_T^{\mathrm{H}_2}$	H_T^{rest}, H_T
$\cos\theta^{\star}_{\mathrm{H}_1\mathrm{H}_2-\mathrm{H}_1}$	$\cos\theta^{\star}_{\mathrm{H}_1-j_1}$	CMVA_3, CMVA_4,
	$\Delta R_{jj}^{\mathrm{H}_1}$, $\Delta R_{jj}^{\mathrm{H}_2}$, $\Delta\phi_{jj}^{\mathrm{H}_1}$, $\Delta\phi_{jj}^{\mathrm{H}_2}$	

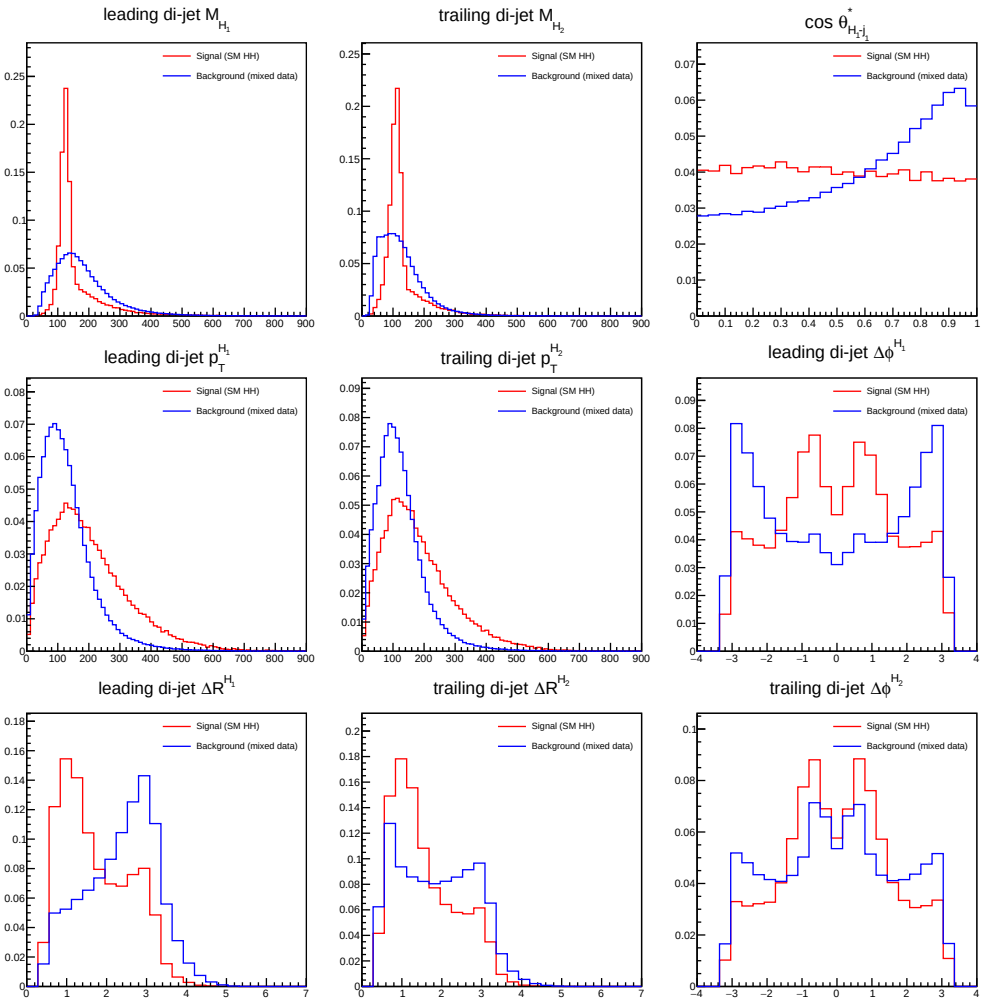


Figure 5.4: Comparison of the signal (SM HH production) and background (mixed data) distributions for the di-jet features considered in the probabilistic classifier. Di-jet candidates are ordered by their mass value. See Table 5.3 and associated text for more details.

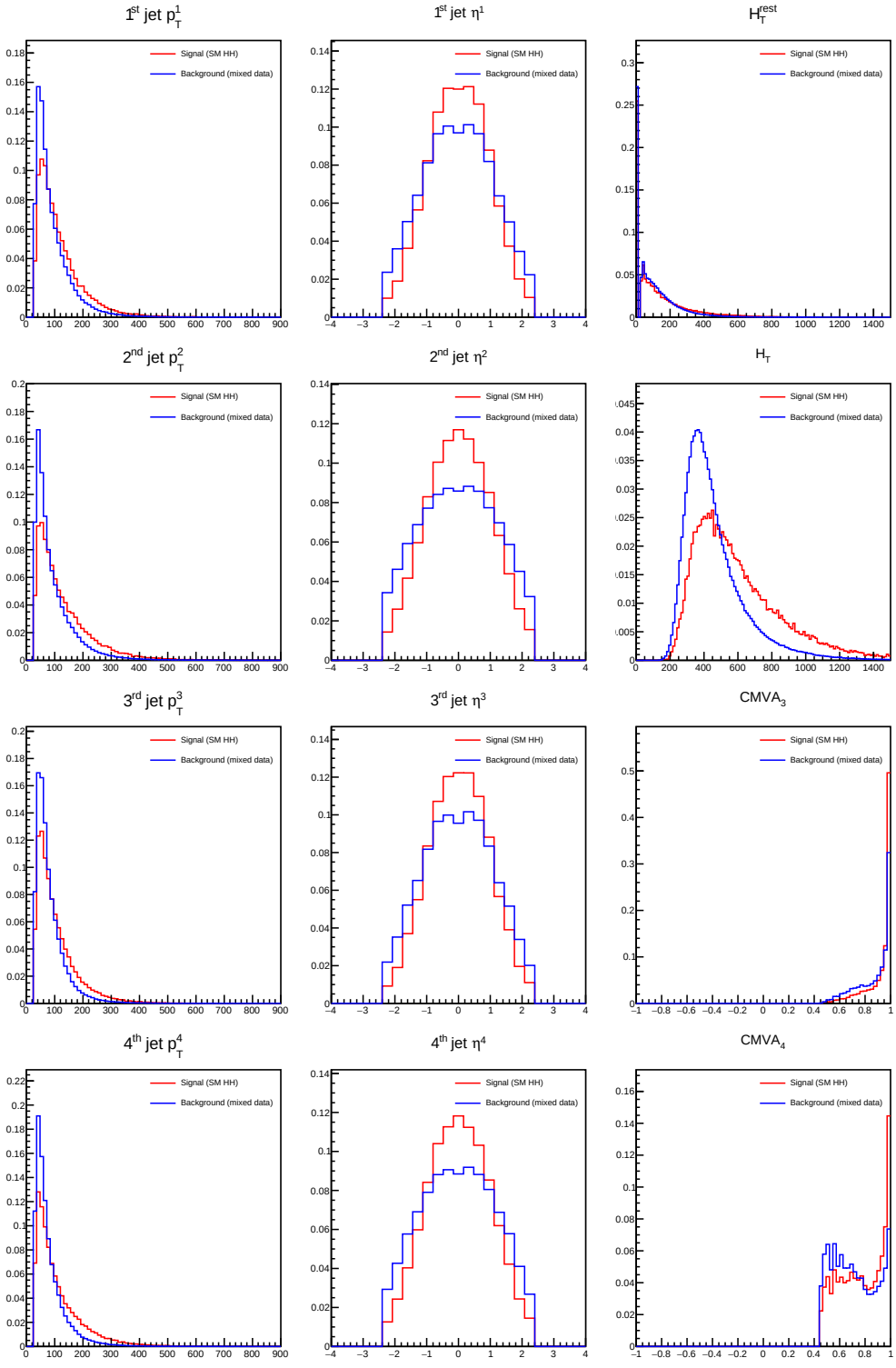


Figure 5.5: Comparison of the signal (SM HH production) and background (mixed data) distributions for the jet-based features considered in the probabilistic classifier. Jet are ordered by CMVA value. See Table 5.3 and associated text for more details.

same prior probability and balance the classification problem. The hyper-parameters have been chosen based on a simple grid search, with the help of the scikit-learn software library [180], based on the area under the curve (AUC) of the resulting classifiers on a validation hold-out dataset.

5.6 Data-Driven Background Estimation

The principal background of this analysis is composed of events with several jets coming from multiple quarks and gluon production from QCD processes. While simulated observations of multi-jet QCD processes can be generated, and were in fact readily available at the time this analysis was carried out, they are in practice not useful to realistically model the background contribution for the purposes of this work. Large datasets modelling inclusive QCD multi-jet production were produced in the CMS simulation campaign, divided in various consecutive range of total generator level scalar transverse momenta sum H_T^{gen}. Leaving aside issues regarding the accuracy of the modelling of high jet multiplicity event provided by current leading order plus parton shower generators, the main obstacle for using the simulated samples is that their equivalent luminosity in the H_T^{gen} relevant for this analysis is several orders of magnitude smaller than the actual luminosity.

As a rule of thumb, to accurately model a mixture component using simulated samples, the number of simulated events has to be at least 10 times more than the number of expected events, or the modelling uncertainty due to the limited simulation statistics will greatly degrade subsequent inference. This problem is made worse when a significant fraction of the simulated dataset has to be used for training a probabilistic classifier and thus cannot be used for computing any expected value, because they might lead to biased estimations. A naive solution could be to simulate more events, but given the large cross section of low energy QCD processes, the total number of QCD inclusive simulated events required would be well over 1 billion which is too a large number given the total simulation budget available for the CMS experiment.

Another option, which was initially explored for modelling the QCD background in this analysis, was to only simulate events that pass a selection at parton level, e.g. with two or more high energy b-quarks. This could provide a radical reduction on the total computing time needed for simulation, especially if combined with the approximate simulation techniques described in Section 2.3.2, because the associated cross section can be greatly reduced. However, such generator level filtering is difficult

to implement in a way that relevant events are not omitted after the event selection. Because of that, the desired level of modelling accuracy could not be achieved with this method.

The previously mentioned reasons motivate the direct used of real data to estimate the background contribution, as discussed in Section 3.1.4. Data-driven background estimation can be notoriously difficult and often several assumptions about the properties of the background have to be made. For example, the corresponding search by the ATLAS collaboration [168], models the background contribution with an independent data sample characterised by the same trigger and selection but for the looser requirement of only two b-tagged reconstructed jets. These events are then re-weighted using a factor that accounts for the probability that QCD processes produce two additional b-tagged jets, where the mentioned weight is also obtained from a data side band where not significant signal is expected. While that approach is proven effective when using the reconstructed M_{H} distribution for inference, it cannot be easily extended to a situation where all the multi-dimensional features of the data require to be precisely modelled, as is the case when the output of a probabilistic classifier is used as the summary statistic.

In the analysis presented here a different path was followed, based on developing a new data-driven background estimation method based on the concept of hemisphere mixing and some assumptions of the phase space characteristics of QCD multi-jet processes [147]. The technique, which is described in Section 5.6.1, directly attempts to create an artificial dataset using the the whole original dataset as input, hence can be used both for training the probabilistic classifier and to model the distribution of the final summary statistic used for inference. Because some aspects of the method are ad-hoc and cannot be formally demonstrated, it has been calibrated and then validated using a signal-depleted control region, a procedure that is discussed in Section 5.7.

5.6.1 Hemisphere Mixing

The basis of the data-driven background estimation method here proposed is to divide each event in two parts, referred to as hemispheres, so each can be substituted by an hemisphere from a different event in order to produce an artificial dataset. A graphical illustration of the hemisphere mixing technique used in this work is provided in Figure 5.6. The transverse thrust axis, defined as the axis in the $x - y$ plane for which the absolute value sum of the projections of the transverse momenta of the selected subset of reconstructed jets is maximal, is used as a reference to

divide each original event in two halves perpendicularly to the mentioned axis. This procedure is carried out for all the collected events that pass the selection described in Section 5.5, creating a dataset (or library) of hemispheres with as many rows as twice as many rows as the number of original events. Each half, or hemisphere, can be basically reduced to a set of reconstructed jets with their directions relative to the thrust axis. Once the hemisphere library has been created, each hemisphere in the original event can be substituted by a similar one by from a different event, once an appropriate distance metric has been defined. The procedure results in an artificial dataset that can be used to model the background component.

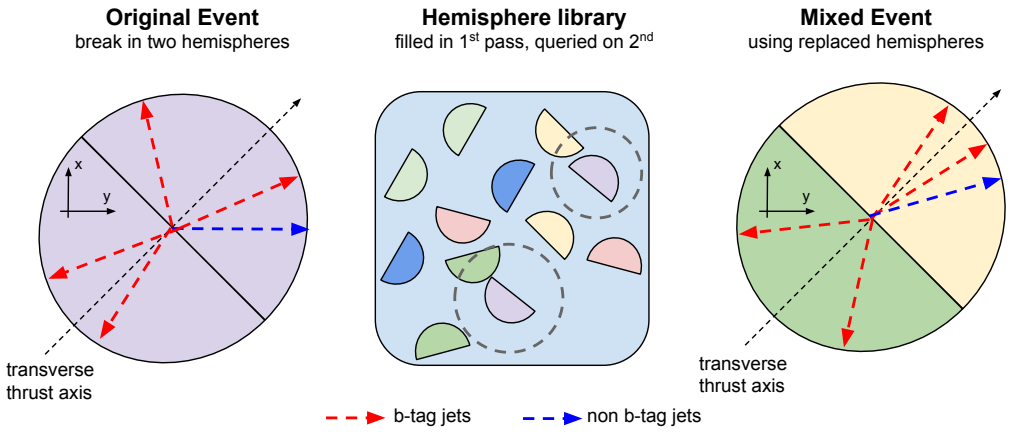


Figure 5.6: Schematic depiction of the hemisphere mixing background estimation procedure. The red arrows represent b-tagged jets and the blue arrows represent jets that were not b-tagged in an event. The first step includes finding the thrust axis in the $x-y$ plane. The event is then divided in two hemispheres, each composed of a set of jets, by the plane perpendicular to the thrust axis. All these hemispheres are used to create a dataset (or library) of hemispheres. For each original event, a artificial event can be created by substituting each original hemisphere with its closest neighbours, once a distance metric for hemispheres has been defined. Figure adapted from [148].

The matching between the original and the replacement hemisphere is done by finding the pair minimising a inter-hemisphere distance. The mentioned distance is a function of the set of reconstructed jets contained within each hemisphere, and it is a combination of discrete and continuous variables. The discrete requirement for matching original hemispheres with those in the library is that they have the same number of jets N_j^h and b-tagged jets N_b^h, which ensures a similar jet multiplicity distributions for the artificial data. The previous condition also avoids creating artificial events that do not pass the event selection, e.g. by combining an hemisphere

with 2 b-tagged jets with another one including only one b-tagged jet, which would result in the artificial events having less that four b-tagged jets. For infrequent jet and b-jet multiplicity categories, the discrete condition is relaxed by considering a unique category. This is for example the case when four jets or b-jets are present in the hemisphere. In addition to the mentioned categorisation, the following continuous distance metric between the original hemisphere $\boldsymbol{h}_o$ and each hemisphere from the library $\boldsymbol{h}_q$ is defined as a measure of similarity:

$$\begin{aligned} d(\boldsymbol{h}_o, \boldsymbol{h}_q)^2 = & \frac{\left(M_{\mathrm{t}}(\boldsymbol{h}_o) - M_{\mathrm{t}}(\boldsymbol{h}_q)\right)^2}{\mathrm{Var}(M_{\mathrm{t}})} + \frac{\left(T(\boldsymbol{h}_o) - T(\boldsymbol{h}_q)\right)^2}{\mathrm{Var}(T)} \\ & + \frac{\left(T_a(\boldsymbol{h}_o) - T_a(\boldsymbol{h}_q)\right)^2}{\mathrm{Var}(T_a)} + \frac{\left(P_z(\boldsymbol{h}_o) - P_z(\boldsymbol{h}_q)\right)^2}{\mathrm{Var}(P_z)} \end{aligned} \tag{5.6}$$

where $M_{\mathrm{t}}(\boldsymbol{h})$ is the invariant mass of the system composed of all the jets contained in the hemisphere, $T(\boldsymbol{h})$ is the scalar sum of all the transverse momenta projection of all jets of an hemisphere to the thrust axis, $T_a(\boldsymbol{h})$ is the scalar sum of the transverse momenta projections over a axis orthogonal to the thrust axis, and $P_z(\boldsymbol{h})$ is the absolute value of the projection of the vectorial sum of the jet momenta along the beam axis. The denominators in Equation 5.6 are the variances of each of the variables and discrete category, as estimated directly from the library of hemispheres. This normalisation factor is included in order to reduce the effect of the scale of the magnitude of each component to the distance metric.

The substitute for each original hemisphere is found by finding the k^{th} nearest-neighbour hemisphere in the library. The closest hemisphere ($k = 0$), corresponding to zero distance, would be the very same original hemisphere which is present in the library. Rather, the hemisphere is substituted with its k^{th} nearest neighbour, only considering $k \geq 1$. Assuming forward-backward symmetry in the z direction and ϕ rotational symmetry, and given that the distance metric $d(\boldsymbol{h}_o, \boldsymbol{h}_q)^2$ does not depend on the sign and absolute magnitude of those quantities, all the jets in the hemisphere can be rotated in ϕ or their p_z sign to match the original hemisphere properties. It is possible to considering different k neighbours for each hemisphere, obtaining a different artificial dataset in each case. Each of this artificial datasets can be labelled by a tuple (k_1, k_2), where k_1 indicate the ordinal of the neighbour used as the substitute for the original hemisphere corresponding to a $\Delta\phi > 0$ with respect to the thrust vector rotated $\pi/2$ clock-wise, and k_2 corresponds to the ordinal of the neighbour substituting the other original hemisphere. Consequently, if up to k_{max} neighbours are considered for each hemispheres, a total of k_{max}^2 artificial datasets,

each of the same size of the original dataset, could be composed by considering all the permutations.

The rationale of the above technique rests on the fact that QCD multi-jet production at leading-order corresponds to a $2 \to 2$ parton scattering process, which is then affected by higher order corrections such such as QCD radiation, pileup or multiple interactions. By breaking the event in two hemispheres using the transverse thrust, the aim is to separate the outcome of the processes associated with each of the two final state partons in the mentioned $2 \to 2$ approximation. The hemisphere distance metric attempts to preserve the main properties of the event, while avoiding strong correlations between jets in the two hemispheres. The goal of the hemisphere mixing procedure is then to obtain an artificial dataset where the effect of the signal present in the original dataset are effectively removed. This has been tested by injecting up to 100 times the expected SM contribution of simulated HH production events to a dataset of simulated QCD multi-jet events [147]. The distributions of the various variables after hemisphere mixing are not affected by the presence of signal, and are compatible with the QCD multi-jet component, which is the majority component. The level of agreement for the variables used as input of the probabilistic classifier in a control region will be discussed in more detail in Section 5.6.2.

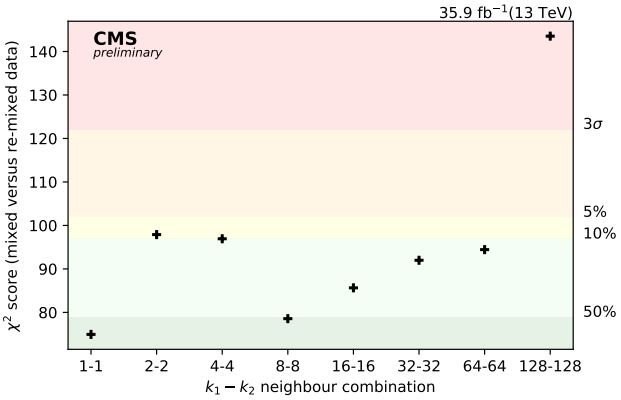


Figure 5.7: Comparison (χ^2 score) of the mixed and re-mixed data (see Section 5.6.2) as a function of the neighbour combination (k_1, k_2). The test score has been calculated based on the binned distribution of the probabilistic classifier. The one-sided confidence bands for the test score are also included for guidance. Figure adapted from [148].

The hemisphere mixing technique is applied to the data events passing the selection described in Section 5.5. Artificial datasets up to $k_{\text{max}} = 10$ have been considered, given that good modelling was observed until very large values of k_{max}.

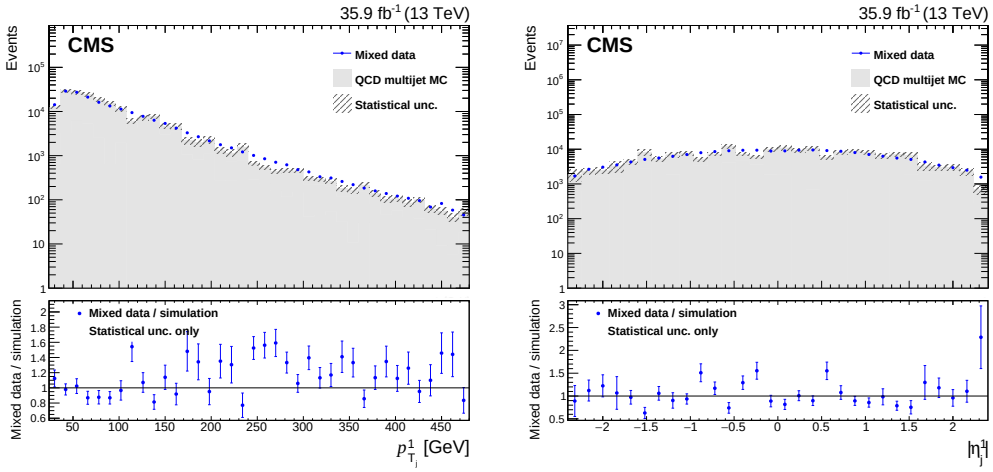
The test score of the compatibility between the mixed artificial data as a function of the combination label is included in Figure 5.7, modelling breaks only at high values, e.g. $k = 128$. All the neighbour combinations up to $k_{\max} = 10$ are sub-divided in three sets used for training the probabilistic classifier (training), validating and optimised the classifier (validation) and to estimate the background distribution of the final summary statistic (application). The last dataset is referred to as application instead of test set because its purpose is not to obtain unbiased estimates of the classifier performance, but rather to extract unbiased estimates of the classifier output distribution of background events. All the artificial datasets are not independent, e.g. the $(1, 1)$ and $(1, 2)$ dataset use the same first hemisphere, thus some careful choices are required when splitting the mixed datasets. The dataset splitting considered in this analysis, using the (k_1, k_2) notation described before, correspond to:

- *training set*: concatenation of $(1, 1)$, $(1, 2)$, $(2, 1)$ and $(2, 2)$ mixed datasets
- *validation set*: concatenation of $(3, 4)$, $(5, 6)$, $(7, 8)$ and $(9, 10)$ mixed datasets
- *application set*: concatenation of $(4, 3)$, $(6, 5)$, $(8, 7)$ and $(10, 9)$ mixed datasets

noting that the observation in the training set are not fully independent, but it is expected that reusing hemispheres in the training sample at most might degrade slightly the classifier performance, but does not bias in any way the inference results if an independent set is used. The next section is devoted to the validation of the background model in data control regions and the development of a methodology to correct for possible biases in the final summary statistic expectations. For completeness, a comparison of the distribution of relevant variables, that are used as input to the probabilistic classifier, between the QCD multi-jet simulations available and those estimated using hemisphere mixing, are shown in Figure 5.8. The overall agreement is good, as expected from the discussion at beginning of this section, the statistical uncertainties coming from the low H_T range simulated QCD dataset are large.

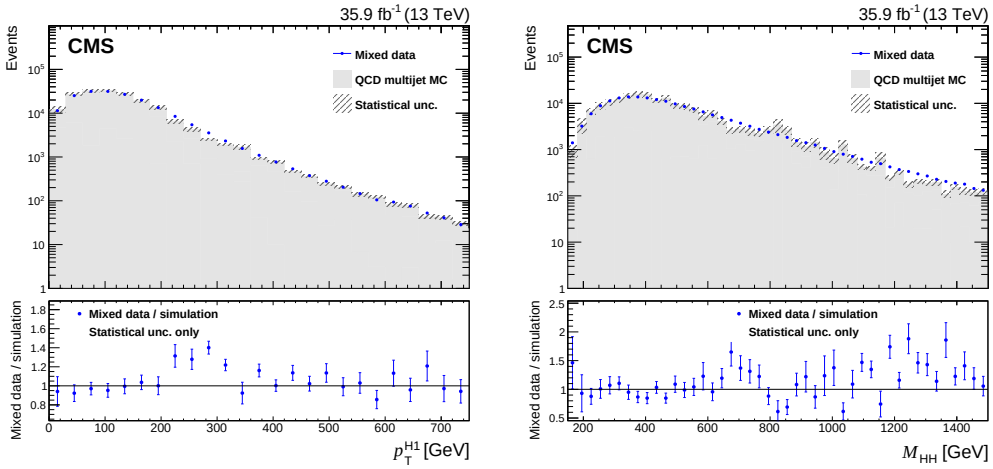
5.6.2 Background Validation

One of the drawbacks of using data-driven methods, is that they are often based on a series of implicit assumptions regarding the underlying statistical model of the data, which are difficult to demonstrate directly. Therefore, a more practical approach to verify the validity of a given background model is usually taken, studying its validity in a set of data control region where the component under study dominates and the



(a) first jet transverse momenta $p^1_{T_j}$

(b) first jet pseudo-rapidity η^1_j



(c) Higgs candidate transverse momenta $p_T^{H_1}$

(d) Higgs candidate pair system mass M_{HH}

Figure 5.8: Comparison between the background model obtained with the hemisphere mixing technique and the simulated observations from QCD processes for a set of relevant reconstructed variables. A correction factor obtained from the binned classifier distribution, as described in Section 5.6.2, has been applied as a weight to the mixed dataset. Only statistical uncertainties are shown. Figures adapted from [148].

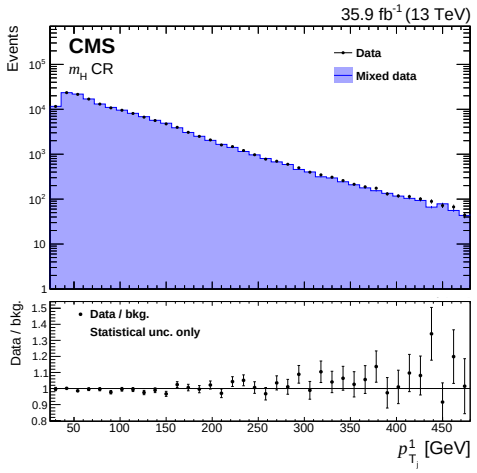
contribution from the signal is negligible. For the purpose of studying the hemisphere mixing method in this analysis, two data control regions (CRs) are defined:

- *mass control region* (M_{H} CR): this dataset is obtained using the same selection described in Section 5.5, but removing all events around the Higgs candidate masses $90 < M_{\mathrm{H_1}} < 150$ GeV and $80 < M_{\mathrm{H_2}} < 140$ GeV. This cut in the reconstructed Higgs masses plane considerably reduces the signal contribution, which is expected to peak around $M_{\mathrm{H}} = 125$ GeV.

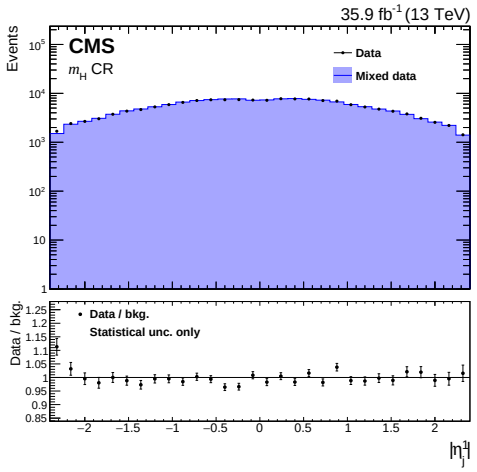
- *b-tag control region* (b-tag CR): this dataset is obtained using the same selection described in Section 5.5 but b-tagged jets are defined using the loose working point of CMVA, which has a misidentification rate of 10% and a b-tagging efficiency around 85% for jets originating from the Higgs pair decay, while filtering out events with any jet above the medium working point of the CMVA discriminator.

The relative signal contribution in each of these control regions is greatly reduced, e.g the expected $n_{\mathrm{S}}/n_{\mathrm{B}}$ ratio in the mass (b-tag) control region is only a 16%(17%) of that of those events inside the $90 < M_{\mathrm{H_1}} < 150$ GeV and $80 < M_{\mathrm{H_1}} < 140$ GeV region. The multi-jet QCD component is still the dominant background in both control regions. While for carrying out the mass control region comparison is enough to apply an additional cut over the selection, the b-tag control region study requires redoing the hemisphere mixing procedure on the new set of event with different b-tag jet selection. For both control regions, all the relevant one-dimensional marginal distributions are found to be in good agreement, as shown for a reduced number of important variables that used as input for the classifier in Figure 5.9 and Figure 5.10.

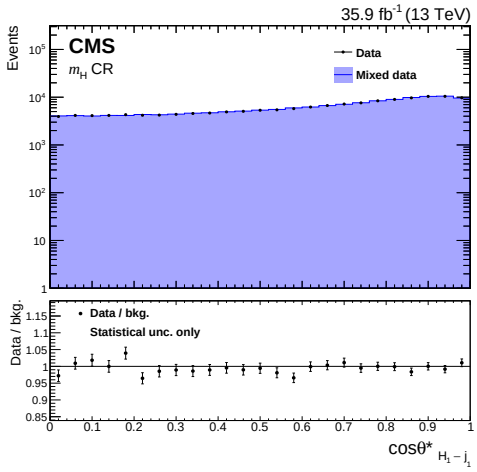
While the marginal distributions of each variable are well-modelled, the goal of the technique is rather to obtain an adequate modelling accuracy in the higher dimensional space considered as input of the probabilistic classifier. A way to check the quality of such modelling is to compare the classifier output distribution for the control region data with the background model. This comparison is shown for the M_{H} control region in Figure 5.11. The same comparison is not straightforward to carry out for the b-tag control region, because the classifier was trained using the lowest value of the CMVA classifiers, which was lower bounded by the medium working point for the standard selection which instead is upper bounded by same working point in the b-tag CR. While Figure 5.11 shows a reasonable agreement overall, a slight background model excess seems to exist in the lower classifier output range.



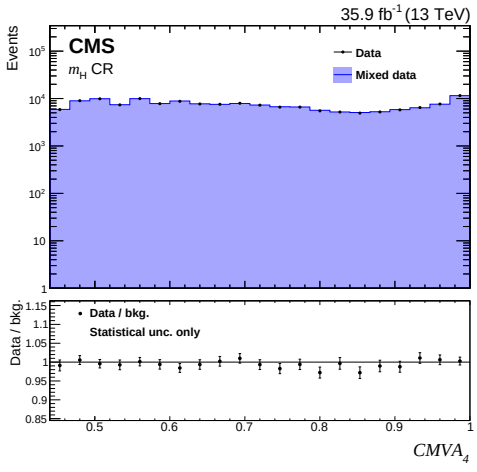
(a) first jet transverse momenta $p^1_{T_j}$



(b) first jet pseudo-rapidity η^1_j

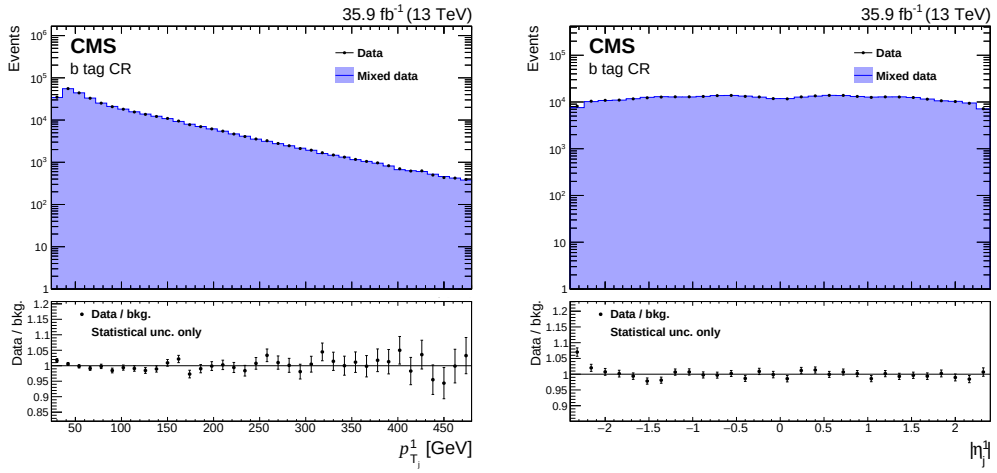


(c) first Higgs candidate - first jet $\cos\theta^*_{\mathrm{H}_1-j_1}$



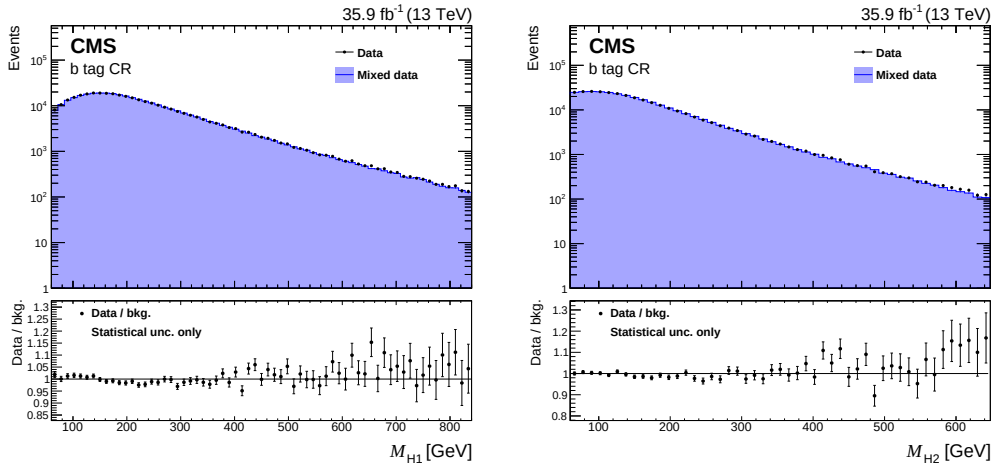
(d) lowest CMVA discriminator CMVA_4

Figure 5.9: Comparison between the background model obtained with the hemisphere mixing technique and the data for the M_{H} control region for a set of reconstructed variables used as input of the classifier. A correction factor obtained from the binned classifier distribution, as described in Section 5.6.2, has been applied as a weight to the mixed dataset. Only statistical uncertainties are shown. Figures adapted from [148].



(a) first jet transverse momenta $p^1_{T_j}$

(b) first jet pseudo-rapidity η^1_j



(c) first Higgs candidate mass M_{H_1}

(d) second Higgs candidate mass M_{H_2}

Figure 5.10: Comparison between the background model obtained with the hemisphere mixing technique and the data for the b-tag control region for a set of reconstructed variables used as input of the classifier. A correction factor obtained from the binned classifier distribution, as described in Section 5.6.2, has been applied as a weight to the mixed dataset. Only statistical uncertainties are shown. Figures adapted from [148].

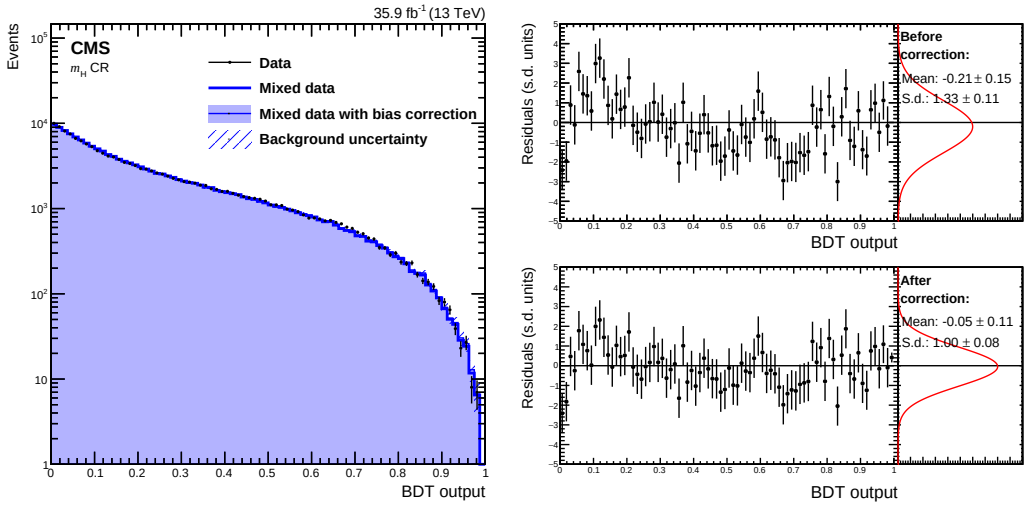


Figure 5.11: Left: Comparison of the BDT classifier output for data in the M_{H} control region, with the same output computed using an artificial dataset by hemisphere mixing. Right: bin-by-bin differences between the control region data and the hemisphere mixing estimation, divided by their uncertainty, both before (top right) and after the bias correction procedure. The pull distributions and their parameters when fitted by a Gaussian are also shown. The uncertainty after the bias correction has been increased conservatively in order to obtain a unit standard deviation for the residual pull distribution. Figures adapted from [148].

The previous mentioned issue has motivated a quantitative study to assess and potentially correct the hemisphere mixing based background model for the classifier output. The bias assessment procedure, schematically depicted in Figure 5.12, starts by constructing a very large artificial sample M by concatenating all the permutations of the (k_1, k_2) datasets up to a $k_{\max} = 10$, except those used for training the classifier. A total of 200 smaller datasets, referred as replicas M_i, with the same number of events of the original data are obtained by subsampling without replacement N times from the large mixed dataset M. Each replica dataset is treated in an analogous manner to the original dataset, thus the hemisphere mixing procedure is applied again to create a set of new artificial datasets R_i. The classifier output distribution is obtained for all the new artificial datasets R_i and compared with the reference distribution of the large sample M, considering a histogram with 80 bins of equal width in the full range of the classifier output $[0.0, 1.0]$.

The median difference between the distribution of the classifier output between the large dataset M and each of the mixed replicas R_i is shown in Figure 5.13 for the final event selection. A small bias is found in the recovered distribution, which is directly used as a correction to hemisphere mixing technique prediction. Similar results are obtained in the previously mentioned control region. The effect of the correction in the classifier output distribution and pulls in the M_{H} control region is also shown in Figure 5.11. The mean of the predicted values minus the observed values are compatible with zero in both control regions, while the root-mean-squared of the pull distribution is not compatible with one in the M_{H}. In order to conservatively account for the mentioned discrepancy, the variation due to the nuisance parameters added per bin to account for the limited statistics of the artificial background sample is multiplied by a factor $\alpha = 1.9$ so the previous pull distribution root-mean-square becomes one.

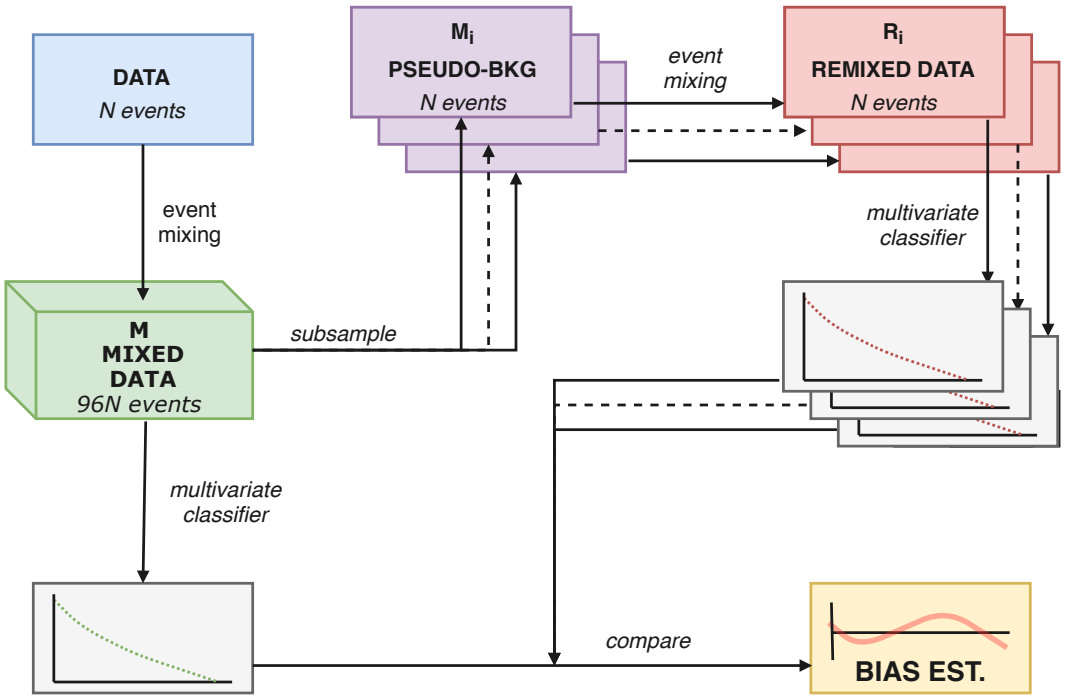


Figure 5.12: Diagram describing the procedure used to estimate the background bias correction. All possible combinations of mixed hemispheres except those used for training are added together to create a large sample N of $96N$ events from which we repeatedly subsample without replacement 200 replicas M_i of N events. The hemisphere mixing procedure is then carried out again for each of this replicas to produce a set of re-mixed data replicas R_i. The trained multivariate classifier is then evaluated over all the events of M and each R_i and the histograms of the classifier output are compared to obtain the differences for each of the replicas. The median difference is taken as bias correction. Figure adapted from [148].

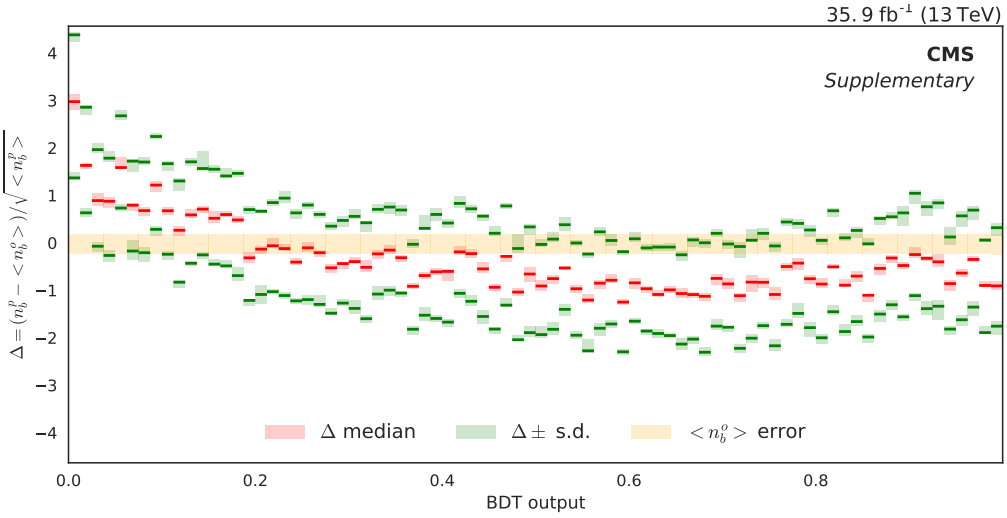


Figure 5.13: Bias estimation obtained by the resampling technique described in the text, in relative units of the statistical uncertainty of the predicted background, used to correct the background estimation. The median (red line) and the upper and lower one s.d. quantiles (green lines) have been computed from 200 subsamples of the re-mixed data comparing the predicted background n_b^p with the observed n_b^o. The variability due to the limited number of subsamples is estimated by bootstrap and it is shown for each estimation using a coloured shadow around the quantile estimation. The light yellow shadow represents the uncertainty due to the limited statistics of the reference observed sample. The separation between the one s.d. quantiles is compatible with the expected variance if the estimation was Poisson or Gaussian distributed. Figure adapted from [148].

5.7 Systematic Uncertainties

Both the signal model based on simulated observations and data-driven background model in this analysis are not perfectly known, hence a set of nuisance parameter have to be considered in the statistical model to account for such lack of certainty, as generally discussed in Section 3.1.4. Each nuisance parameter, which can affect the signal, the background component or both, effectively leads to an increase of the uncertainty on the parameters of interest. For analysis where upper limits are set such as this, the presence of these unknown parameters increases the total interval width. The effect of these parameters in the final statistical estimates is also often referred as systematic uncertainty. A list of the sources of systematic uncertainty considered in this analysis, and their estimated relative effect in the expected upper limit for the SM Higgs pair production, is provided on the Table 5.4.

Table 5.4: List of systematic uncertainties considered in this analysis, and their relative impact on the expected limit for the SM HH production. The relative impact is obtained by fixing the nuisance parameters corresponding to each source and recalculating the expected limit.

Source	Affects	Exp. limit variation
Bkg. shape	bkg.	30%
Bkg. norm.	bkg.	8.6%
b-tagging eff.	sig	2.8%
Pileup	sig	<0.01%
Jet energy res.	sig	<0.01%
Jet energy scale	sig	<0.01%
Int. luminosity	sig	<0.01%
Trigger eff.	sig	<0.01%
μ_F and μ_F scales	sig	<0.01%
PDF	sig	<0.01%

The main sources of uncertainty in this analysis are those associated with the data-driven background model. For each classifier output bin, an independent nuisance parameter is included that accounts for the possible variation of the background prediction due to the limited data statistics of the artificial events used for building the background model and the accuracy limitations found during the bias correction procedure described in Section 5.6.2. Because the data-driven technique described in the previous section does not provide a way to estimate the normalisation of the background, the background normalisation is added a nuisance parameter that is left fully unconstrained.

Regarding systematic uncertainties due to nuisance parameters of the simulation-based signal distribution, the most relevant factors are the uncertainties in the measure differences between data and simulation in b-tagging efficiencies. These are estimated by recomputing the signal distribution weighted by a factor that accounts for a one standard deviation for each of the relevant nuisance parameters and interpolating in-between as described in Section 3.1.3. The uncertainty due to the modelling of the pile-up contribution is included by considering the different effect of pile-up reweighting when a $\pm 4.6\%$ variation on the total inelastic cross section value at 13 TeV is allowed [181]. The effect due to the modelling uncertainties in jet energy resolution and scale are estimated by smearing or shifting the reconstructed jet energy respectively, according to their corresponding uncertainties as a function of the jet p_T and $|\eta|$, and evaluating the effect on the final summary statistic. For all the mentioned sources of uncertainty, both the effect on the classifier output distribution and its normalisation have been considered.

After a correction by the observed discrepancies between the data and simulation, the uncertainty on the trigger efficiency after to a 2% effect on the signal normalisation. The total signal component normalisation is also affected by the uncertainty in the measurement of the integrated luminosity $\mathcal{L}_{\text{int}}$, which has been estimated during the 2016 data-taking period to be 2.5%[182]. The effect of theoretical uncertainties that affect the simulation samples are modelled using per-event weights provided by the simulation software. In particular, the effect of a variation of the renormalisation μ_{R} and factorisation μ_{F} scales on the signal efficiency are estimated by taking the maximum and the minimum difference with respect to the nominal efficiency when varying μ_{R} and μ_{F} each individually as well as both together up and down by a factor of two. For estimating the total signal efficiency variation due to parton distribution function (PDF) uncertainties, the PDF4LHC recommendations [183] are followed, computing the variation as the standard deviation of a set of 100 MC replicas of the NNPDF 3.0 set [178].

5.8 Analysis Results

This section includes the experimental results of the search of non-resonant Higgs pair production with CMS data collected during 2016 at the LHC. The final summary statistic is the distribution of a probabilistic classifier output, which was trained on simulated events of SM HH production and events resulting from the data-driven background estimation technique described in Section 5.6. Specifically, a

non-parametric sample likelihood composed by a product of Poisson count likelihoods is used, where each Poisson factor represents a bin of the distribution of the classifier output, in an analogous manner to Equation 3.28. The classifier distribution was initially divided in 80 equal sized bins, and the expected number of counts from each mixture component and their variations due to nuisance parameters were estimated using simulated observations under the SM hypothesis and each of the BSM EFT points considered for the signal, and from the bias corrected distribution for the data-driven background dataset.

Given the slight mis-modelling observed in the lower range of the classifier output on the control regions discussed in Section 5.6.2, a study studying the variation of the expected limit when a non-zero minimum value is considered in the likelihood binning was carried out. It was found that restricting the fit to classifier output values larger than 0.2 resulted on a negligible loss on sensitivity (i.e. smaller than 2%) while greatly improving the overall data-background compatibility. For this reason, onty the rightmost 64 of the initial 80 bins of the classifier distribution are used to build the Poisson likelihood used for statistical inference. The best-fit distributions for signal, background and data for the classifier output are shown in Figure 5.14, while the those corresponding to the reconstructed Higgs boson masses are shown in Figure 5.15.

Only two mixture components are considered in the final statistical model, signal representing $pp \rightarrow HH \rightarrow b\bar{b}b\bar{b}$, and background estimated from data and dominated by QCD multi-jet processes and secondarily by top quark production with additional jets. The contribution from other hard processes that can produce four b-quarks, such as $t\bar{t}H$, ZH, $b\bar{b}H$, and single Higgs boson production was estimated from simulated samples and found to be negligible in comparison with the considered background uncertainties at the current level of experiment sensitivity.

The same statistical model is used to obtain he observed and expected 95% confidence level (CL) upper limits for non-resonant $pp \rightarrow HH \rightarrow b\bar{b}b\bar{b}$ production, using the asymptotic approximation [100] of the CL_s criterion [101, 102, 184], and the so-called LHC test statistic, that is based on the profile likelihood ratio. All the nuisance parameters are treated by profiling the likelihood. The median expected and observed upper limits for the SM Higgs pair production, as well as the expected limit 1 and 2 standard deviation intervals around the median are included in Table 5.5. The median expected limit obtained for SM HH production is 419 fb, which corresponds to approximately 37 times the SM expectation, which can be obtained by taking the cross section from Equation 5.1 and multiplying it by the $b\bar{b}b\bar{b}$ decay branching frac-

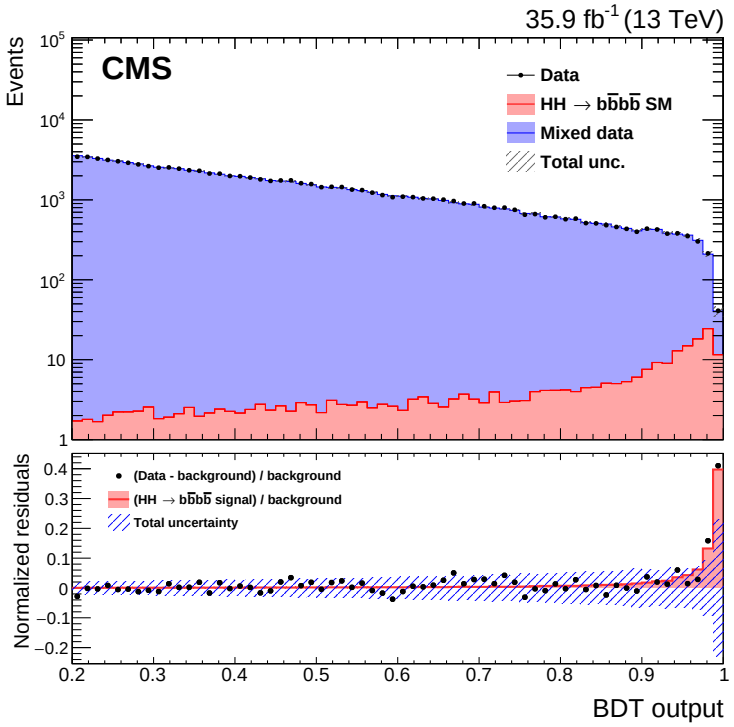
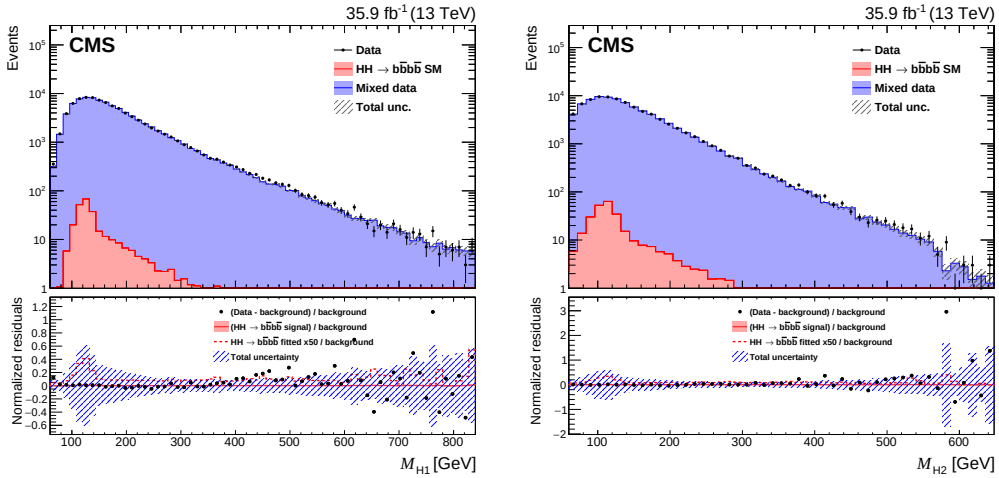


Figure 5.14: Results of the for best fit of the statistical model of BDT classifier output distribution for the SM HH production signal for the observed data. In the lower panel a comparison is shown between the best fit signal and best fit background subtracted from measured data. The dashed band in the lower panel, centred at zero, shows the total uncertainty. Figure adapted from [148].



(a) leading Higgs mass M_{H_1} (b) trailing Higgs mass M_{H_2}

Figure 5.15: Distributions of the reconstructed Higgs masses for the best fit. A correction factor obtained from the binned classifier distribution, as described in Section 5.6.2, has been applied as a weight to the mixed dataset. Figures adapted from [148].

tion. The obtained observed limit is 847 fb, which is about two standard deviations above the expected limit. To facilitate the comparison with the analyses carried out in other channels, the observed limit corresponds to $\sigma(\mathrm{pp} \to \mathrm{HH}) = 2496$ SM as an upper limit the inclusive HH production cross section of SM-like processes.

Table 5.5: Observed and expected upper limits on $\sigma(\mathrm{pp} \to \mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}})$ in the SM at 95% CL in units of fb.

Category HH → bb̄bb̄	Observed	Expected	-2 s.d.	-1 s.d.	+1 s.d.	+2 s.d.
SM	847	419	221	297	601	834

The same procedure was carried out for each of the EFT benchmarks previously listed in Table 5.1, by re-weighting the simulated HH production observation as discusses and evaluating the signal distribution under each BSM model considered. The observed and expected limits obtained for each of the benchmark points are provided in Table 5.6. The observed and expected limits are also graphically compared between the various EFT points and the SM in Figure 5.16. The observed limits are also found about two standard deviations over the median expected limits, which can be explained by taking into account that the same classifier and thus the same background model (and its associated fluctuations) is considered in the statistical

model for all the inference procedures. In particular, the last bins of the classifier distribution for the data-driven background prediction has a small deficit compared to the observed data, as can be seen in Figure 5.14.

Table 5.6: Observed and expected upper limits on the $\sigma(pp \to HH \to b\bar{b}b\bar{b})$ cross section for the 13 BSM benchmark models listed in Table 5.1 at 95% CL in units of fb.

Benchmark point	Observed	Expected	-2 s.d.	-1 s.d.	+1 s.d.	+2 s.d.
1	602	295	155	209	424	592
2	554	269	141	190	389	548
3	705	346	182	245	497	691
4	939	461	244	327	662	920
5	508	248	131	176	357	501
6	937	457	240	323	657	916
7	3510	1710	905	1210	2440	3390
8	686	336	177	238	483	674
9	529	259	136	183	373	520
10	2090	1000	527	709	1440	2010
11	1080	525	277	372	755	1050
12	1744	859	455	611	1230	1710
Box	1090	542	286	384	775	1080

In addition to the BSM benchmarks, limits are also obtained for the cross section times branching ratio of Higgs pair production processes in the EFT framework, varying κ_λ in the range $[-20, 20]$, while assuming that $\kappa_t = 1$ and the rest of the couplings are zero. The results are shown in Figure 5.17, noting that the upper limit changes considerably in this range because the distribution of the final state properties change considerably, and consequently the associated efficiency for the process also varies. The EFT cross section prediction as a function of κ_λ and keeping $\kappa_t = 1$ is also shown in the previous figure, noting that no values of κ_λ can be excluded at the current level of experimental sensitivity.

5.9 Combination with Other Decay Channels

The results of the search presented here have been combined with other Higgs pair searches carried out by the CMS collaboration for other decay channels for the same data collection period at $\sqrt{13}$ TeV. For the combination, another three decay modes are considered in addition to the $b\bar{b}b\bar{b}$, where one the Higgs decays to a $b\bar{b}$ pair where the other decays into $\gamma\gamma$, $\tau\bar{\tau}$ or a pair or vector bosons, respectively. Combined upper limits where obtained by considering the product of the likelihoods, which depend on

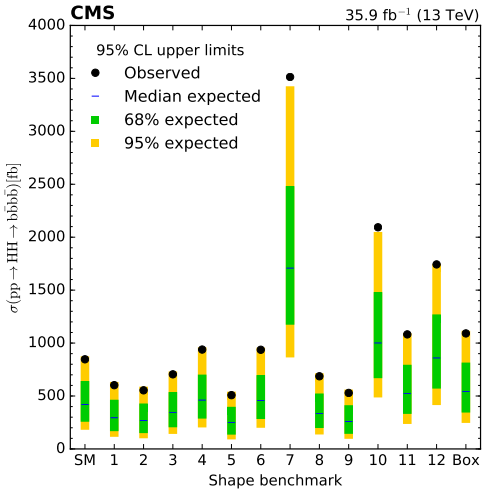


Figure 5.16: Graphical comparison between the observed and expected upper limits at 95% CL on the $\sigma(pp \rightarrow HH \rightarrow b\bar{b}b\bar{b})$ cross section for the SM and the 13 BSM models investigated. The inner green bands and the outer yellow bands and correspond to the range of percentiles around the median that contain the 68% and 95% times the upper limit under the background-only hypothesis. See Table 5.1 for their respective EFT parameter values. Figure adapted from [148].

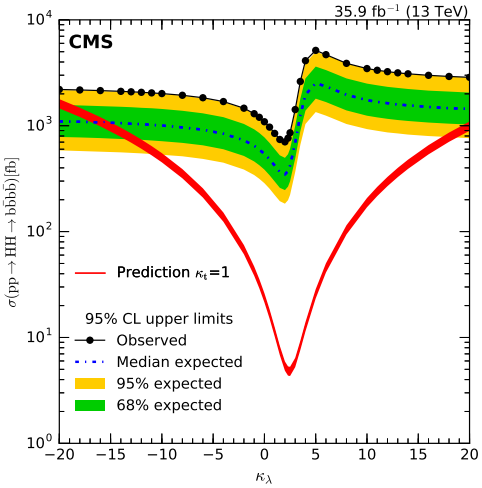


Figure 5.17: Observed and expected $\sigma(pp \rightarrow HH \rightarrow b\bar{b}b\bar{b})$ cross section limits at 95% CL for values of κ_λ in the [-20,20] range, assuming $\kappa_t = 1$. The inner green band and the outer yellow bands correspond to the range of percentiles around the median that contain the 68% and 95% times the upper limit under the background-only hypothesis. the theoretical prediction with $\kappa_t = 1$ is also shown in red colour. Figure adapted from [148].

the HH cross section and several nuisance parameters. Some sources of uncertainty that are correlated between different channels, such as the luminosity or b-tagging uncertainty, where modelled using the same nuisance parameters in each individual likelihood. More details on the combination procedure are included in the following CMS Public Analysis Note [185].

The 95% C.L. upper limits for the Higgs pair non-resonant production cross section $\sigma(\mathrm{pp} \to \mathrm{HH})$ from the $\mathrm{pp} \to \mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ can be compared with those obtained by the other searches in Figure 5.18. In the same figure, the upper limits for the combination of the four decay modes are also shown. The combination results are statistically compatible with the SM background contribution. An median expected limit of 12.8 times the SM expectation is obtained from the combination. The combined observed upper limit is 22.2 times the SM expectation, which is well-within the expected variation under the background only hypothesis. Analogously to what was done in Section 5.8, upper limits are also obtained for the cross section times branching ratio of Higgs pair production processes in the EFT framework, varying κ_λ in the range $[-20, 20]$, while assuming that $\kappa_\mathrm{t} = 1$ and the rest of the couplings are zero. This results are shown graphically in Figure 5.19; values for the anomalous self-coupling κ_λ in the range $-11.8 < \kappa_\lambda < 18.8$ are not excluded by the data ($-7.1 < \kappa_\lambda < 13.6$ was the expected interval). The aforementioned results make this combination analysis the most sensitive search to date at the LHC for non-resonant HH production. Substantial improvements can be expected due to the extensions of each analysis to the full Run II dataset.

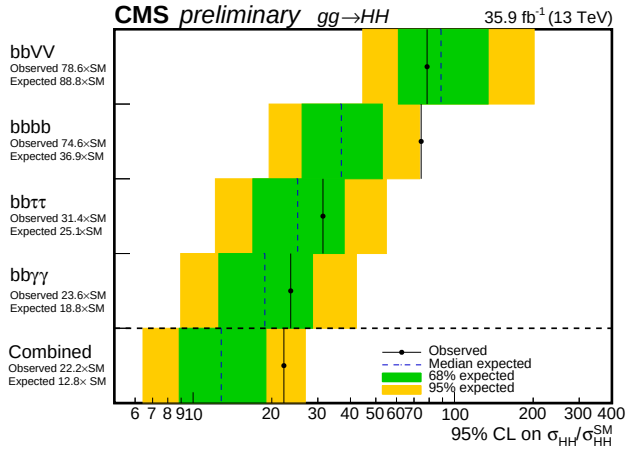


Figure 5.18: Observed and expected $\sigma(\mathrm{pp} \to \mathrm{HH})$ cross section limits relative to the SM for the combination of searches for Higgs boson pair production at 95% CL for values of κ_λ in the [-20,20] range, assuming $\kappa_\mathrm{t} = 1$. The inner green band and the outer yellow bands correspond to the range of percentiles around the median that contain the 68% and 95% times the upper limit under the background-only hypothesis. Figure adapted from [185].

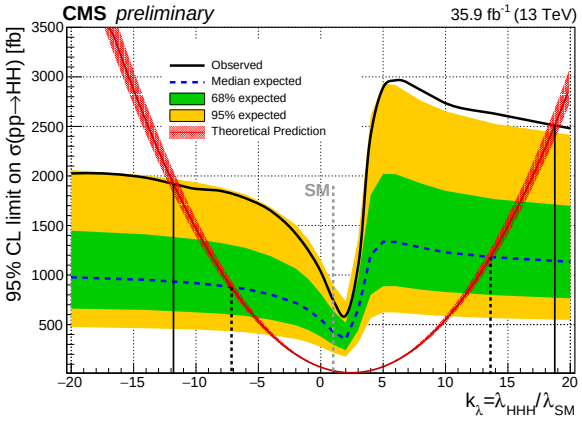


Figure 5.19: Observed and expected $\sigma(\mathrm{pp} \to \mathrm{HH})$ cross section limits for the combination of searches for Higgs boson pair production at 95% CL for values of κ_λ in the [-20,20] range, assuming $\kappa_\mathrm{t} = 1$. The inner green band and the outer yellow bands correspond to the range of percentiles around the median that contain the 68% and 95% times the upper limit under the background-only hypothesis. The anomalous couplings theoretical prediction with $\kappa_\mathrm{t} = 1$ is also shown in red colour. Figure adapted from [185].

6 Inference-Aware Neural Optimisation

> An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.
>
> John Tukey

By this point, it should be evident that powerful statistical inference is the ultimate objective of all experimental high-energy analyses. Supervised learning based on simulated observations or acquired data from control regions, and in particular probabilistic classification, provides a way to extract and approximate estimate of the latent variables of the generative model. Those latent variable estimates are in turn very useful to construct powerful summary for statistical inference. While this approach is very often encountered in experimental high energy physics, complex computer simulations are also required for many other scientific disciplines, making inference very challenging due to the intractability of the likelihood evaluation for the observed data. Summary statistics based on a supervised learning algorithms can be asymptotically optimal if the generative model is fully defined, as is the case for the output of soft classification for mixture models where we are interested in the mixture coefficients, as demonstrated in Section 4.3.1. Unfortunately, their usefulness can rapidly decrease when additional uncertain parameters affect the generative model.

As a practical example, in the analysis presented in Chapter 5, the limiting factor for experimental sensitivity was not in the choice of summary statistics but rather on the lack of detailed knowledge about the expected contribution from background processes, which had to be addressed by the inclusion of nuisance parameters. The technique presented in this chapter, referred to as INFERNO and published in [186], is an attempt to tackle directly the problem of constructing non-linear summary statistics from a statistical perspective that directly addresses the goal of the final inference question. The key contribution required for achieving such goal is to lever-

age the technology that has been developed for recent machine learning techniques, to build inference-aware loss functions that approximate the expected uncertainty on the parameters of interest, accounting for the effect of nuisance parameters.

6.1 Introduction

Simulator-based inference is currently at the core of many scientific fields, such as population genetics, epidemiology, and experimental particle physics. In many cases the implicit generative procedure defined in the simulation is stochastic and/or lacks a tractable probability density $p(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ is the vector of model parameters. Given some experimental observations $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$, a problem of special relevance for these disciplines is statistical inference on a subset of model parameters $\boldsymbol{\omega} \in \Omega \subseteq \Theta$. This can be approached via likelihood-free inference algorithms such as Approximate Bayesian Computation (ABC) [95], simplified synthetic likelihoods [187] or density estimation-by-comparison approaches [188].

Because the relation between the parameters of the model and the data is only available via forward simulation, most likelihood-free inference algorithms tend to be computationally expensive due to the need of repeated simulations to cover the parameter space. When data are high-dimensional, likelihood-free inference can rapidly become inefficient, so low-dimensional summary statistics $\boldsymbol{s}(D)$ are used instead of the raw data for tractability. The choice of summary statistics for such cases becomes critical, given that naive choices might cause loss of relevant information and a corresponding degradation of the power of resulting statistical inference.

For the particular problem of high energy physics data analyses at the LHC, the properties of the underlying generative model discussed in Chapter 3 make the likelihood intractable, but its structure facilitates the construction of simulation-based likelihoods of low-dimensional summary statistics that approximate latent variables. The ultimate aim is nevertheless to extract information about Nature from the large amounts of high-dimensional data on the subatomic particles produced by energetic collision of protons, and acquired by highly complex detectors built around the collision point. Accurate data modelling is only available via stochastic simulation of a complicated chain of physical processes, from the underlying fundamental interaction to the subsequent particle interactions with the detector elements and their readout. As a result, the density $p(\boldsymbol{x}|\boldsymbol{\theta})$ cannot be analytically computed.

Due to the high dimensionality of the observed data, a low-dimensional summary statistic has to be constructed in order to perform inference. A well-known result of

classical statistics, which was also discussed in Section 3.2.2 as the Neyman-Pearson lemma[97], establishes that the likelihood-ratio $\Lambda(\boldsymbol{x}) = p(\boldsymbol{x}|H_0)/p(\boldsymbol{x}|H_1)$ is the most powerful test when two simple hypotheses are considered. As $p(\boldsymbol{x}|H_0)$ and $p(\boldsymbol{x}|H_1)$ are not available, simulated samples are used in practice to obtain an approximation of the likelihood ratio by casting the problem as supervised learning classification.

Within high energy physics analysis, the nature of the generative model (a mixture of different processes) allows the treatment of the problem as signal (S) versus background (B) classification [189], when the task becomes one of effectively estimating an approximation of $p_S(\boldsymbol{x})/p_B(\boldsymbol{x})$ which will vary monotonically with the likelihood ratio. This has been discussed at great lengths in Section 4.3.1. While the use of classifiers to learn a summary statistic can be effective and increase the discovery sensitivity, the simulations used to generate the samples which are needed to train the classifier often depend on additional uncertain parameters (commonly referred to as nuisance parameters). These nuisance parameters are not of immediate interest but have to be accounted for in order to make quantitative statements about the model parameters based on the available data. Classification-based summary statistics cannot easily account for those effects, so their inference power is degraded when nuisance parameters are finally taken into account.

In this chapter, we present a new machine learning method to construct non-linear sample summary statistics that directly optimises the expected amount of information about the subset of parameters of interest using simulated samples, by explicitly and directly taking into account the effect of nuisance parameters. In addition, the learned summary statistics can be used to build synthetic sample-based likelihoods and perform robust and efficient classical or Bayesian inference from the observed data, so they can be readily applied in place of current classification-based or domain-motivated summary statistics in current scientific data analysis workflows.

6.2 Problem Statement

Let us consider a set of n i.i.d. observations $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$ where $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, and a generative model which implicitly defines a probability density $p(\boldsymbol{x}|\boldsymbol{\theta})$ used to model the data. The generative model is a function of the vector of parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, which includes both relevant and nuisance parameters. We want to learn a function $\boldsymbol{s} : \mathcal{D} \subseteq \mathbb{R}^{d \times n} \to \mathcal{S} \subseteq \mathbb{R}^b$ that computes a summary statistic of the dataset and reduces its dimensionality so likelihood-free inference methods can be

applied effectively. From here onwards, b will be used to denote the dimensionality of the summary statistic $\boldsymbol{s}(D)$.

While there might be infinite ways to construct a summary statistic $\boldsymbol{s}(D)$, we are only interested in those that are informative about the subset of interest $\boldsymbol{\omega} \in \Omega \subseteq \Theta$ of the model parameters. The concept of statistical sufficiency is especially useful to evaluate whether summary statistics are informative. In the absence of nuisance parameters, classical sufficiency can be characterised by means of the factorisation criterion (see Section 3.1.3 for more details):

$$p(D|\boldsymbol{\omega}) = h(D)g(\boldsymbol{s}(D)|\boldsymbol{\omega}) \tag{6.1}$$

where h and g are non-negative functions. If $p(D|\boldsymbol{\omega})$ can be factorised as indicated, the summary statistic $\boldsymbol{s}(D)$ will yield the same inference about the parameters $\boldsymbol{\omega}$ as the full set of observations D. When nuisance parameters have to be accounted in the inference procedure, alternate notions of sufficiency are commonly used such as partial or marginal sufficiency [190, 191]. Nonetheless, for the problems of relevance in this work, the probability density is not available in closed form so the general task of finding a sufficient summary statistic cannot be tackled directly. Hence, alternative methods to build summary statistics have to be followed.

For simplicity, let us consider a problem where we are only interested in performing statistical inference on a single one-dimensional model parameter $\boldsymbol{\omega} = \{\omega_0\}$ given some observed data. Be given a summary statistic $\boldsymbol{s}$ and a statistical procedure to obtain an unbiased interval estimate of the parameter of interest which accounts for the effect of nuisance parameters. The resulting interval can be characterised by its width $\Delta\omega_0 = \hat{\omega}_0^+ - \hat{\omega}_0^-$, defined by some criterion so as to contain on average, upon repeated samping, a given fraction of the probability density, e.g. a central 68.3% interval. The expected size of the interval depends on the summary statistic $\boldsymbol{s}$ chosen: in general, summary statistics that are more informative about the parameters of interest will provide narrower confidence or credible intervals on their value. Under this figure of merit, the problem of choosing an optimal summary statistic can be formally expressed as finding a summary statistic $\boldsymbol{s}^*$ that minimises the interval width:

$$\boldsymbol{s}^* = \arg\min_{\boldsymbol{s}} \Delta\omega_0. \tag{6.2}$$

The above construction can be extended to several parameters of interest by considering the interval volume or any other function of the resulting confidence or credible regions.

6.3 Method

In this section, a machine learning technique to learn non-linear sample summary statistics is described in detail. The method seeks to minimise the expected variance of the parameters of interest obtained via a non-parametric simulation-based synthetic likelihood. A graphical description of the technique is depicted on Fig. 6.1. The parameters of a neural network are optimised by stochastic gradient descent within an automatic differentiation framework, where the considered loss function accounts for the details of the statistical model as well as the expected effect of nuisance parameters.

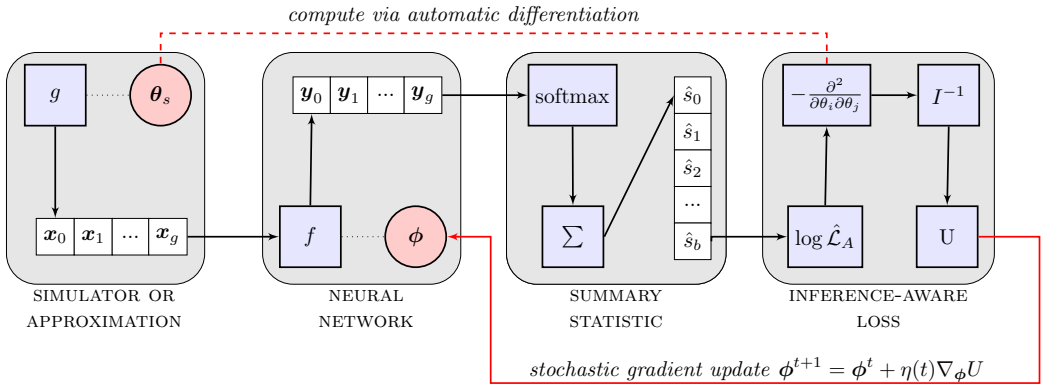


Figure 6.1: Learning inference-aware summary statistics (see text for details).

The family of summary statistics $\boldsymbol{s}(D)$ considered in this work is based on a neural network model applied to each dataset observation $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\phi}) : \mathcal{X} \subseteq \mathbb{R}^d \rightarrow \mathcal{Y} \subseteq \mathbb{R}^b$, whose parameters $\boldsymbol{\phi}$ will be learned during training by means of stochastic gradient descent, as will be discussed later. Therefore, using set-builder notation the considered family of summary statistics considered can be denoted as:

$$\boldsymbol{s}(D, \boldsymbol{\phi}) = \boldsymbol{s}(\ \{\ \boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{\phi})\ |\ \forall\ \boldsymbol{x}_i \in D\ \}\ ) \tag{6.3}$$

where $\boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{\phi})$ will reduce the dimensionality from the input observations space $\mathcal{X}$ to a lower-dimensional space $\mathcal{Y}$. The next step is to map observation outputs to a dataset summary statistic, which will in turn be calibrated and optimised via a non-parametric likelihood $\mathcal{L}(D;\boldsymbol{\theta},\boldsymbol{\phi})$ created using a set of simulated observations $G_s = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_g\}$, generated at a certain instantiation of the simulator parameters $\boldsymbol{\theta}_s$.

In experimental high energy physics experiments, which are the scientific context that initially motivated this work, histograms of observation counts are the most commonly used non-parametric density estimator because the resulting likelihoods can be expressed as the product of Poisson factors, one for each of the considered bins. A naive sample summary statistic can be built from the output of the neural network by simply assigning each observation $\boldsymbol{x}$ to a bin corresponding to the cardinality of the maximum element of $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\phi})$, so each element of the sample summary will correspond to the following sum:

$$s_i(D;\boldsymbol{\phi}) = \sum_{\boldsymbol{x}\in D} \begin{cases} 1 & i = argmax_{j=\{0,\dots,b\}}(f_j(\boldsymbol{x};\boldsymbol{\phi})) \\ 0 & i \neq argmax_{j=\{0,\dots,b\}}(f_j(\boldsymbol{x};\boldsymbol{\phi})) \end{cases} \tag{6.4}$$

which can in turn be used to build the following likelihood, where the expectation for each bin is taken from the simulated sample G_s:

$$\mathcal{L}(D;\boldsymbol{\theta},\boldsymbol{\phi}) = \prod_{i=0}^{b} \text{Pois}\left(s_i(D;\boldsymbol{\phi}) \mid \left(\frac{n}{g}\right) s_i(G_s;\boldsymbol{\phi})\right) \tag{6.5}$$

where the n/g factor accounts for the different number of observations in the simulated samples. In cases where the number of observations is itself a random variable providing information about the parameters of interest, or where the simulated observations are weighted, the choice of normalisation of $\mathcal{L}$ may be slightly more involved and problem specific, but nevertheless amenable. Note the relation between the summary statistics and likelihoods defined in this section and those discussed in Section 3.1.3.

In the above construction, the chosen family of summary statistics is not differentiable due to the *argmax* operator, so gradient-based updates for the parameters cannot be computed. To work around this problem, a differentiable approximation $\hat{\boldsymbol{s}}(D;\boldsymbol{\phi})$ is considered. This function is defined by means of a $softmax$ operator:

$$\hat{s}_i(D;\boldsymbol{\phi}) = \sum_{x\in D} \frac{e^{f_i(\boldsymbol{x};\boldsymbol{\phi})/\tau}}{\sum_{j=0}^{b} e^{f_j(\boldsymbol{x};\boldsymbol{\phi})/\tau}} \tag{6.6}$$

where the temperature hyper-parameter τ will regulate the softness of the operator. In the limit of $\tau \to 0^+$, the probability of the largest component will tend to 1 while others to 0, and therefore $\hat{\boldsymbol{s}}(D;\boldsymbol{\phi}) \to \boldsymbol{s}(D;\boldsymbol{\phi})$. Similarly, let us denote by $\hat{\mathcal{L}}(D;\boldsymbol{\theta},\boldsymbol{\phi})$ the differentiable approximation of the non-parametric likelihood obtained

by substituting $\boldsymbol{s}(D;\boldsymbol{\phi})$ with $\hat{\boldsymbol{s}}(D;\boldsymbol{\phi})$. Instead of using the observed data D, the value of $\hat{\mathcal{L}}$ may be computed when the observation for each bin is equal to its corresponding expectation based on the simulated sample G_s, which is commonly denoted as the Asimov likelihood [100] $\hat{\mathcal{L}}_A$:

$$\hat{\mathcal{L}}_A(\boldsymbol{\theta};\boldsymbol{\phi}) = \prod_{i=0}^{b} \mathrm{Pois}\left(\left(\frac{n}{g}\right)\hat{s}_i(G_s;\boldsymbol{\phi}) \mid \left(\frac{n}{g}\right)\hat{s}_i(G_s;\boldsymbol{\phi})\right) \tag{6.7}$$

for which it can be easily proven that $argmax_{\boldsymbol{\theta}\in\theta}(\hat{\mathcal{L}}_A(\boldsymbol{\theta};\boldsymbol{\phi})) = \boldsymbol{\theta}_s$, so the maximum likelihood estimator (MLE) for the Asimov likelihood is the parameter vector $\boldsymbol{\theta}_s$ used to generate the simulated dataset G_s. In Bayesian terms, if the prior over the parameters is flat in the chosen metric, then $\boldsymbol{\theta}_s$ is also the maximum a posteriori (MAP) estimator. By taking the negative logarithm and expanding in $\boldsymbol{\theta}$ around $\boldsymbol{\theta}_s$, we may obtain the Fisher information matrix [107] for the Asimov likelihood:

$$\boldsymbol{I}(\boldsymbol{\theta})_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(-\log\hat{\mathcal{L}}_A(\boldsymbol{\theta};\boldsymbol{\phi})\right) \tag{6.8}$$

which can be computed via automatic differentiation if the simulation is differentiable and included in the computation graph, or if the effect of varying $\boldsymbol{\theta}$ over the simulated dataset G_s can be effectively approximated. While this requirement does constrain the applicability of the proposed technique to a subset of likelihood-free inference problems, it is quite common in e.g. physical sciences that the effect of the parameters of interest and the main nuisance parameters over a sample can be approximated by the changes of mixture coefficients of mixture models, translations of a subset of features, or conditional density ratio re-weighting.

If $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of the values of $\boldsymbol{\theta}$, the covariance matrix fulfils the Cramér-Rao lower bound [108, 109]:

$$\mathrm{cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \geq I(\boldsymbol{\theta})^{-1} \tag{6.9}$$

and the inverse of the Fisher information can be used as an approximate estimator of the expected variance, given that the bound would become an equality in the asymptotic limit for MLE. If some of the parameters $\boldsymbol{\theta}$ are constrained by independent measurements characterised by their likelihoods $\{\mathcal{L}_C^0(\boldsymbol{\theta}), ..., \mathcal{L}_C^c(\boldsymbol{\theta})\}$, those

constraints can also be easily included in the covariance estimation, simply by considering the augmented likelihood $\hat{\mathcal{L}}'_A$ instead of $\hat{\mathcal{L}}_A$ in Eq. 6.8:

$$\hat{\mathcal{L}}'_A(\boldsymbol{\theta};\boldsymbol{\phi}) = \hat{\mathcal{L}}_A(\boldsymbol{\theta};\boldsymbol{\phi}) \prod_{i=0}^{c} \mathcal{L}_C^i(\boldsymbol{\theta}). \tag{6.10}$$

In Bayesian terminology, this approach is referred to as the Laplace approximation [110] where the logarithm of the joint density (including the priors) is expanded around the MAP to a multi-dimensional normal approximation of the posterior density:

$$p(\boldsymbol{\theta}|D) \approx \text{Normal}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}, I(\hat{\boldsymbol{\theta}})^{-1}) \tag{6.11}$$

which has already been approached by automatic differentiation in probabilistic programming frameworks [192]. While a histogram has been used to construct a Poisson count sample likelihood, non-parametric density estimation techniques can be used in its place to construct a product of observation likelihoods based on the neural network output $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\phi})$ instead. For example, an extension of this technique to use kernel density estimation (KDE) should be straightforward, given its intrinsic differentiability.

The loss function used for stochastic optimisation of the neural network parameters $\boldsymbol{\phi}$ can be any function of the inverse of the Fisher information matrix at $\boldsymbol{\theta}_s$, depending on the ultimate inference aim. The diagonal elements $I_{ii}^{-1}(\boldsymbol{\theta}_s)$ correspond to the expected variance of each of the ϕ_i under the normal approximation mentioned before, so if the aim is efficient inference about one of the parameters $\omega_0 = \theta_k$ a candidate loss function is:

$$U = I_{kk}^{-1}(\boldsymbol{\theta}_s) \tag{6.12}$$

which corresponds to the expected width of the confidence interval for ω_0 accounting also for the effect of the other nuisance parameters in $\boldsymbol{\theta}$. This approach can also be extended when the goal is inference over several parameters of interest $\boldsymbol{\omega} \subseteq \boldsymbol{\theta}$ (e.g. when considering a weighted sum of the relevant variances). A simple version of the approach just described to learn a neural-network based summary statistic employing an inference-aware loss is summarised in Algorithm 1.

Algorithm 1 Inference-Aware Neural Optimisation.

Input 1: differentiable simulator or variational approximation $g(\boldsymbol{\theta})$.
Input 2: initial parameter values $\boldsymbol{\theta}_s$.
Input 3: parameter of interest $\omega_0 = \theta_k$.
Output: learned summary statistic $\boldsymbol{s}(D; \boldsymbol{\phi})$.

1: **for** $i = 1$ to N **do**
2: Sample a representative mini-batch G_s from $g(\boldsymbol{\theta}_s)$.
3: Compute differentiable summary statistic $\hat{\boldsymbol{s}}(G_s; \boldsymbol{\phi})$.
4: Construct Asimov likelihood $\mathcal{L}_A(\boldsymbol{\theta}, \boldsymbol{\phi})$.
5: Get information matrix inverse $I(\boldsymbol{\theta})^{-1} = \boldsymbol{H}_{\boldsymbol{\theta}}^{-1}(\log \mathcal{L}_A(\boldsymbol{\theta}, \boldsymbol{\phi}))$.
6: Obtain loss $U = I_{kk}^{-1}(\boldsymbol{\theta}_s)$.
7: Update network parameters $\boldsymbol{\phi} \to \mathrm{SGD}(\nabla_{\boldsymbol{\phi}} U)$.
8: **end for**

6.4 Related Work

Classification or regression models have been implicitly used to construct summary statistics for inference in several scientific disciplines. For example, in experimental particle physics, the mixture model structure of the problem makes it amenable to supervised classification based on simulated datasets [193, 194]. While a classification objective can be used to learn powerful feature representations and increase the sensitivity of an analysis, it does not take into account the details of the inference procedure or the effect of nuisance parameters like the solution proposed here.

The first known effort to include the effect of nuisance parameters in classification and explain the relation between classification and the likelihood ratio was by Neal [195]. In the mentioned work, Neal proposes training of classifier including a function of nuisance parameter as additional input together with a per-observation regression model of the expectation value for inference. Cranmer et al. [188] improved on this concept by using a parametrised classifier to approximate the likelihood ratio which is then calibrated to perform statistical inference. At variance with the mentioned works, we do not consider a classification objective at all and the neural network is directly optimised based on an inference-aware loss. Additionally, once the summary statistic has been learnt the likelihood can be trivially constructed and used for classical or Bayesian inference without a dedicated calibration step. Furthermore,

the approach presented in this work can also be extended, as done by Baldi et al. [134] by a subset of the inference parameters to obtain a parametrised family of summary statistics with a single model.

Recently, Brehmer et al. [196, 197, 198] further extended the approach of parametrised classifiers to better exploit the latent-space structure of generative models from complex scientific simulators. Additionally they propose a family of approaches that include a direct regression of the likelihood ratio and/or likelihood score in the training losses. While extremely promising, the most performing solutions are designed for a subset of the inference problems at the LHC and they require considerable changes in the way the inference is carried out. The aim of the algorithm proposed here is different, as we try to learn sample summary statistics that may act as a plug-in replacement of classifier-based dimensionality reduction and can be applied to general likelihood-free problems where the effect of the parameters can be modelled or approximated.

Within the field of Approximate Bayesian Computation (ABC), there have been some attempts to use neural network as a dimensionality reduction step to generate summary statistics. For example, Jiang et al. [199] successfully employ a summary statistic by directly regressing the parameters of interest and therefore approximating the posterior mean given the data, which then can be used directly as a summary statistic.

A different path is taken by Louppe et al. [200], where the authors present a adversarial training procedure to enforce a pivotal property on a predictive model. The main concern we have on the use of that approach is that a classifier which is pivotal with respect to nuisance parameters might not be optimal, neither for classification nor for statistical inference. Instead of aiming for being pivotal, the summary statistics learnt by our algorithm attempt to find a transformation that directly reduces the expected effect of nuisance parameters over the parameters of interest.

6.5 Experiments

In this section, we first study the effectiveness of the inference-aware optimisation in a synthetic mixture problem where the likelihood is known. We then compare our results with those obtained by standard classification-based summary statistics. All the code needed to reproduce the results presented here is available in an online

repository [201], extensively using TensorFlow [129] and TensorFlow Probability [192, 202] software libraries.

6.5.1 3D Synthetic Mixture

In order to exemplify the usage of the proposed approach, evaluate its viability and test its performance by comparing to the use of a classification model proxy, a three-dimensional mixture example with two components is considered. One component will be referred as background $f_b(\boldsymbol{x}|\lambda)$ and the other as signal $f_s(\boldsymbol{x})$; their probability density functions are taken to correspond respectively to:

$$f_b(\boldsymbol{x}|r,\lambda) = \mathcal{N}\left((x_0,x_1)\,\middle|\,(2+r,0),\begin{bmatrix}5 & 0\\ 0 & 9\end{bmatrix}\right) Exp(x_2|\lambda) \tag{6.13}$$

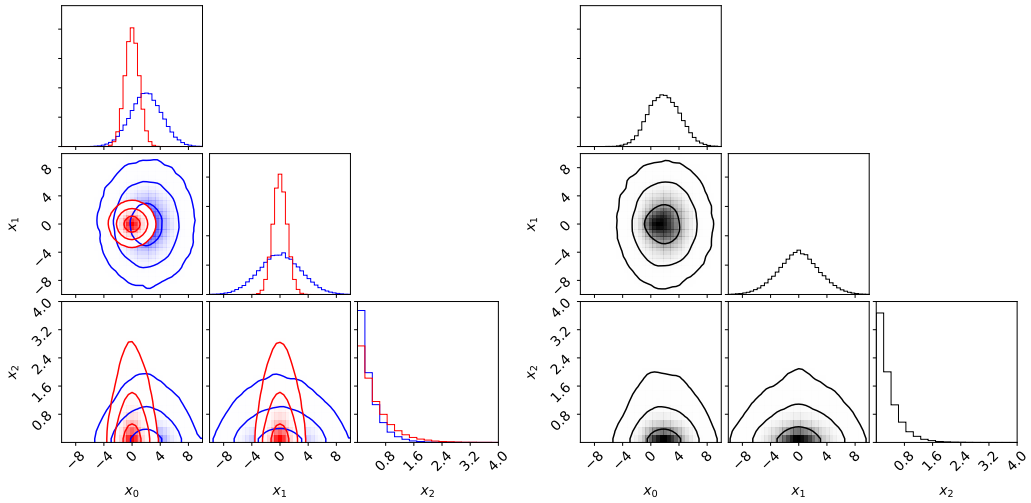
$$f_s(\boldsymbol{x}) = \mathcal{N}\left((x_0,x_1)\,\middle|\,(1,1),\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right) Exp(x_2|2) \tag{6.14}$$

so that (x_0, x_1) are distributed according to a multivariate normal distribution while x_2 follows an independent exponential distribution both for background and signal, as shown in Fig. 6.2a. The signal distribution is fully specified while the background distribution depends on r, a parameter which shifts the mean of the background density, and a parameter λ which specifies the exponential rate in the third dimension. These parameters will be the treated as nuisance parameters when benchmarking different methods. Hence, the probability density function of observations has the following mixture structure:

$$p(\boldsymbol{x}|\mu,r,\lambda) = (1-\mu)f_b(\boldsymbol{x}|r,\lambda) + \mu f_s(\boldsymbol{x}) \tag{6.15}$$

where μ is the parameter corresponding to the mixture weight for the signal and consequently $(1-\mu)$ is the mixture weight for the background. The low-dimensional projections from samples from the mixture distribution for a small $\mu = 50/1050$ is shown in Fig. 6.2b.

Let us assume that we want to carry out inference based on n i.i.d. observations, such that $\mathbb{E}[n_s] = \mu n$ observations of signal and $\mathbb{E}[n_b] = (1-\mu)n$ observations of background are expected, respectively. While the mixture model parametrisation shown in Eq. 6.15 is correct as is, the underlying model could also give information on the expected number of observations as a function of the model parameters. In this toy problem, we consider a case where the underlying model predicts that the



(a) signal (red) and background (blue) (b) mixture distribution (black)

Figure 6.2: Projection in 1D and 2D dimensions of 50000 samples from the synthetic problem considered. The background distribution nuisance parameters used for generating data correspond to $r = 0$ and $\lambda = 3$. For samples the mixture distribution, $s = 50$ and $b = 1000$ were used, hence the mixture coefficient is $\mu = 50/1050$.

total number of observations are Poisson distributed with a mean $s+b$, where s and b are the expected number of signal and background observations. Thus the following parametrisation will be more convenient for building sample-based likelihoods:

$$p(\boldsymbol{x}|s, r, \lambda, b) = \frac{b}{s+b} f_b(\boldsymbol{x}|r, \lambda) + \frac{s}{s+b} f_s(\boldsymbol{x}). \tag{6.16}$$

The parametrisation of Equation 6.16 is common for physics analyses at the LHC, because theoretical calculations provide information about the expected number of observations. If the probability density is known, but the expectation for the number of observed events depends on the model parameters, the likelihood can be extended [203] with a Poisson count term as:

$$\mathcal{L}(s, r, \lambda, b) = \text{Pois}(n|s+b) \prod^{n} p(\boldsymbol{x}|s, r, \lambda, b) \tag{6.17}$$

which will be used to provide an optimal inference baseline when benchmarking the different approaches. Another quantity of relevance is the conditional density ratio,

which would correspond to the optimal classifier (in the Bayes risk sense) separating signal and background events in a balanced dataset (equal priors):

$$s^*(\boldsymbol{x}|r,\lambda) = \frac{f_s(\boldsymbol{x})}{f_s(\boldsymbol{x}) + f_b(\boldsymbol{x}|r,\lambda)} \tag{6.18}$$

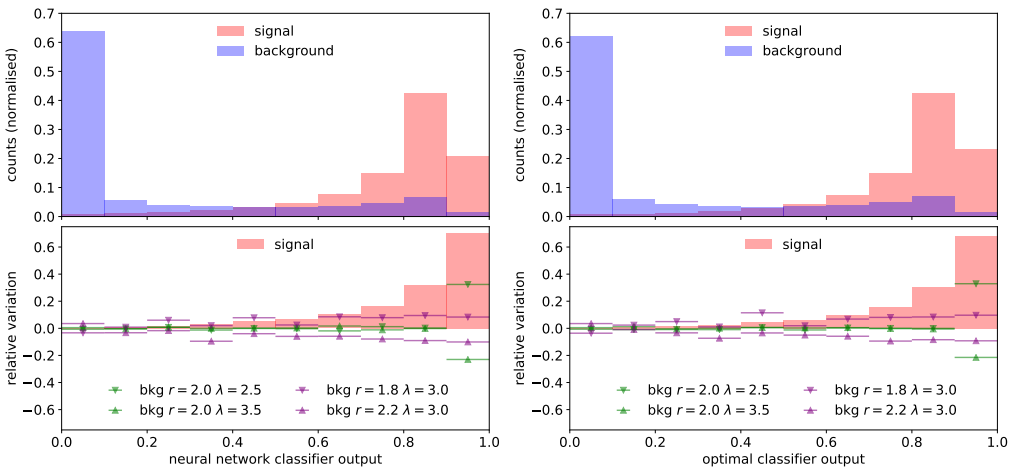
noting that this quantity depends on the parameters that define the background distribution r and λ, but not on s or b that are a function of the mixture coefficients. It can be proven (see Section 4.3.1) that $s^*(\boldsymbol{x})$ is a sufficient summary statistic with respect to an arbitrary two-component mixture model if the only unknown parameter is the signal mixture fraction μ (or alternatively s in the chosen parametrisation). In practice, the probability density functions of signal and background are not known analytically, and only forward samples are available through simulation, so alternative approaches are required.

The synthetic nature of this example allows to rapidly generate training data on demand, yet a training dataset of only 200,000 simulated observations has been considered, in order to study how the proposed method performs when training data is limited. Half of the simulated observations correspond to the signal component and half to the background component. The latter has been generated using $r = 0.0$ and $\lambda = 3.0$. A validation holdout from the training dataset of 200,000 observations is used exclusively for computing relevant metrics during training and to control over-fitting. The final figures of merit that allow to compare different approaches are computed using a larger dataset of 1,000,000 observations. For simplicity, mini-batches for each training step are balanced so the same number of events from each component is taken both when using standard classification or inference-aware losses.

A common treatment of this problem in high-energy physics consist of posing the problem as one of classification based on a simulated dataset, as discussed in Section 4.3.1. A supervised machine learning model such a neural network can be trained to discriminate signal and background observations, considering a fixed parameters r and λ. The output of such a model typically consist in class probabilities c_s and c_b given an observation $\boldsymbol{x}$, which will tend asymptotically to the optimal classifier from Eq. 6.18 given enough data, a flexible enough model and a powerful learning rule. The conditional class probabilities (or alternatively the likelihood ratio $f_s(\boldsymbol{x})/f_b(\boldsymbol{x})$) are powerful learned features that can be used as summary statistic; however their construction ignores the effect of the nuisance parameters r and λ on the background distribution. Furthermore, some kind of non-parametric density estimation (e.g. a histogram) has to be considered in order to build a calibrated statistical model using

the classification-based learned features, which will in turn smooth and reduce the information available for inference.

To exemplify the use of this family of classification-based summary statistics, a histogram of a deep neural network classifier output trained on simulated data and its variation computed for different values of r and λ are shown in Fig. 6.3a. The details of the training procedure will be provided later in this document. The classifier output can be directly compared with $s(\boldsymbol{x}|r = 0.0, \lambda = 3.0)$ evaluated using the analytical distribution function of signal and background according to Eq. 6.18, which is shown in Fig. 6.3b and corresponds to the optimal classifier. The trained classifier approximates very well the optimal classifier. The summary statistic distribution for background depends considerably on the value of the nuisance parameters both for the trained and the optimal classifier, which will in turn cause an important degradation on the subsequent statistical inference.



(a) classifier trained on simulated samples (b) optimal classifier $s(\boldsymbol{x}|r = 0.0, \lambda = 3.0)$

Figure 6.3: Histograms of summary statistics for signal and background (top) and variation for different values of nuisance parameters compared with the expected signal relative to the nominal background magniture (bottom). The classifier was trained using signal and background samples generated for $r = 0.0$ and $\lambda = 3.0$.

The statistical model described above has up to four unknown parameters: the expected number of signal observations s, the background mean shift r, the background exponential rate in the third dimension λ, and the expected number of background observations. The effect of the expected number of signal and background observations s and b can be easily included in the computation graph by weighting the signal and background observations. This is equivalent to scaling the resulting vector

of Poisson counts (or its differentiable approximation) if a non-parametric counting model as the one described in Sec. 6.3 is used. Instead the effect of r and λ, both nuisance parameters that will define the background distribution, is more easily modelled as a transformation of the input data $\boldsymbol{x}$. In particular, r is a nuisance parameter that causes a shift on the background along the first dimension and its effect may be accounted for in the computation graph by simply adding $(r, 0.0, 0.0)$ to each observation in the mini-batch generated from the background distribution. Similarly, the effect of λ can be modelled by multiplying x_2 by the ratio between the λ_0 used for generation and the one being modelled. These transformations are specific for this example, but alternative transformations depending on parameters could also be accounted for as long as they are differentiable or substituted by a differentiable approximation.

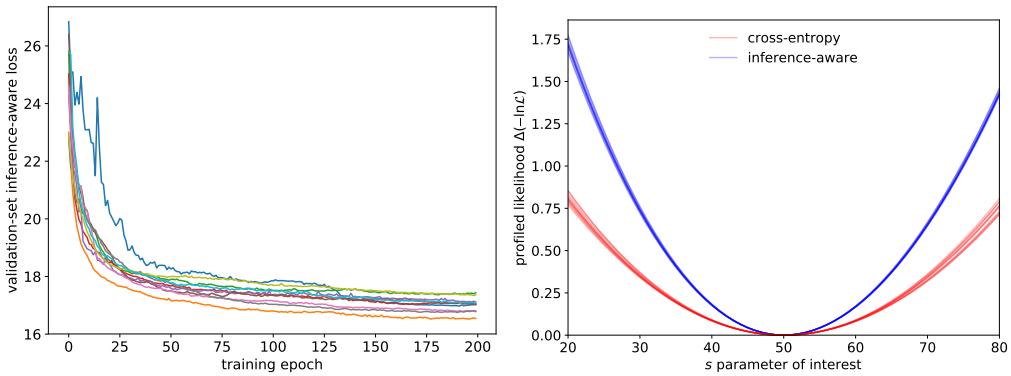
For this problem, we are interested in carrying out statistical inference on the parameter of interest s. In fact, the performance of inference-aware optimisation as described in Sec. 6.3 will be compared with classification-based summary statistics for a series of inference benchmarks based on the synthetic problem described above that vary in the number of nuisance parameters considered and their constraints:

- **Benchmark 0:** no nuisance parameters are considered, both signal and background distributions are taken as fully specified ($r = 0.0$, $\lambda = 3.0$ and $b = 1000.$).
- **Benchmark 1:** r is considered as an unconstrained nuisance parameter, while $\lambda = 3.0$ and $b = 1000$ are fixed.
- **Benchmark 2:** r and λ are considered as unconstrained nuisance parameters, while $b = 1000$ is fixed.
- **Benchmark 3:** r and λ are considered as nuisance parameters but with the following constraints: $\mathcal{N}(r|0.0, 0.4)$ and $\mathcal{N}(\lambda|3.0, 1.0)$, while $b = 1000$ is fixed.
- **Benchmark 4:** all r, λ and b are all considered as nuisance parameters with the following constraints: $\mathcal{N}(r|0.0, 0.4)$, $\mathcal{N}(\lambda|3.0, 1.0)$ and $\mathcal{N}(b|1000., 100.)$.

When using classification-based summary statistics, the construction of a summary statistic does depend on the presence of nuisance parameters, so the same model is trained independently of the benchmark considered. In real-world inference scenarios, nuisance parameters have often to be accounted for and typically are constrained by prior information or auxiliary measurements. For the approach presented here, inference-aware neural optimisation, the effect of the nuisance parameters and their constraints can be taken into account during training. Hence, 5 different train-

ing procedures for INFERNO will be considered, one for each of the benchmarks, denoted by the same number.

The same basic network architecture is used both for cross-entropy and inference-aware training: two hidden layers of 100 nodes followed by ReLU activations. The number of nodes on the output layer is two when classification proxies are used, matching the number of mixture classes in the problem considered. Instead, for inference-aware classification the number of output nodes can be arbitrary and will be denoted with b, corresponding to the dimensionality of the sample summary statistics. The final layer is followed by a softmax activation function and a temperature $\tau = 0.1$ for inference-aware learning in order to ensure that the differentiable approximations are closer to the true expectations. Standard mini-batch stochastic gradient descent (SGD) is used for training and the optimal learning rate is fixed and decided by means of a simple scan; the best choice found is specified together with the results.



(a) inference-aware training loss (b) profile-likelihood comparison

Figure 6.4: Dynamics and results of inference-aware optimisation: (a) square root of inference-loss (i.e. approximated standard deviation of the parameter of interest) as a function of the training step for 10 different random initialisations of the neural network parameters; (b) profiled likelihood around the expectation value for the parameter of interest of 10 trained inference-aware models and 10 trained cross-entropy loss based models. The latter are constructed by building a uniformly binned Poisson count likelihood of the conditional signal probability output. All results correspond to Benchmark 2.

In Fig. 6.4a, the dynamics of inference-aware optimisation are shown by the validation loss, which corresponds to the approximate expected variance of parameter s, as a function of the training step for 10 random-initialised instances of the INFERNO model corresponding to Benchmark 2. All inference-aware models were

trained during 200 epochs with SGD using mini-batches of 2000 observations and a learning rate $\gamma = 10^{-6}$. All the model initialisations converge to summary statistics that provide low variance for the estimator of s when the nuisance parameters are accounted for.

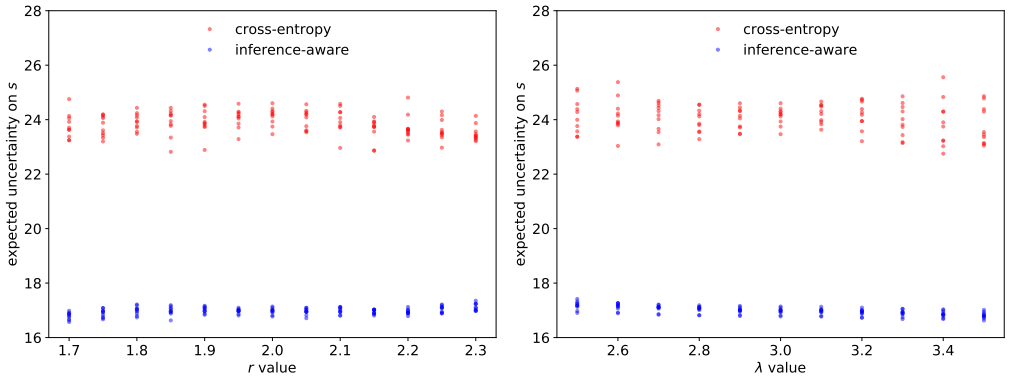
To compare with alternative approaches and verify the validity of the results, the profiled likelihoods obtained for each model are shown in Fig. 6.4b. The expected uncertainty if the trained models are used for subsequent inference on the value of s can be estimated from the profile width when $\Delta\mathcal{L} = 0.5$. Hence, the average width for the profile likelihood using inference-aware training, 16.97 ± 0.11, can be compared with the corresponding one obtained by uniformly binning the output of classification-based models in 10 bins, 24.01 ± 0.36. The models based on cross-entropy loss were trained during 200 epochs using a mini-batch size of 64 and a fixed learning rate of $\gamma = 0.001$.

A more complete study of the improvement provided by the different INFERNO training procedures is provided in Table 6.1, where the median and 1-sigma percentiles on the expected uncertainty on s are provided for 100 random-initialised instances of each model. In addition, results for 100 random-initialised cross-entropy trained models and the optimal classifier and likelihood-based inference are also included for comparison. The confidence intervals obtained using INFERNO-based summary statistics are considerably narrower than those using classification and tend to be much closer to those expected when using the true model likelihood for inference. The only exception being the results obtained for Benchmark 0, where no nuisance parameters are considered, and thus the classification approach is expected to approximate a sufficient summary statistic. Much smaller fluctuations between initialisations are observed for the INFERNO-based cases. The improvement over classification increases when more nuisance parameters are considered. The results also seem to suggest the inclusion of additional information about the inference problem in the INFERNO technique leads to comparable or better results than its omission.

Given that a certain value of the parameters $\boldsymbol{\theta}_s$ has been used to learn the summary statistics as described in Algorithm 1 while their true value is unknown, the expected uncertainty on s has also been computed for cases when the true value of the parameters $\boldsymbol{\theta}_{\text{true}}$ differs. The variation of the expected uncertainty on s when either r or λ is varied for classification and inference-aware summary statistics is shown in Fig. 6.5 for Benchmark 2. The inference-aware summary statistics learnt for $\boldsymbol{\theta}_s$ work well when $\boldsymbol{\theta}_{\text{true}} \neq \boldsymbol{\theta}_s$ in the range of variation explored.

Table 6.1: Expected uncertainty on the parameter of interest s for each of the inference benchmarks considered using a cross-entropy trained neural network model, INFERNO customised for each problem and the optimal classifier and likelihood based results. The results for INFERNO matching each problem are shown with bold characters.

	Benchmark 0	Benchmark 1	Benchmark 2	Benchmark 3	Benchmark 4
NN classifier	$14.99^{+0.02}_{-0.00}$	$18.94^{+0.11}_{-0.05}$	$23.94^{+0.52}_{-0.17}$	$21.54^{+0.27}_{-0.05}$	$26.71^{+0.56}_{-0.11}$
INFERNO 0	$\mathbf{15.51^{+0.09}_{-0.02}}$	$18.34^{+5.17}_{-0.51}$	$23.24^{+6.54}_{-1.22}$	$21.38^{+3.15}_{-0.69}$	$26.38^{+7.63}_{-1.36}$
INFERNO 1	$15.80^{+0.14}_{-0.04}$	$\mathbf{16.79^{+0.17}_{-0.05}}$	$21.41^{+2.00}_{-0.53}$	$20.29^{+1.20}_{-0.39}$	$24.26^{+2.35}_{-0.71}$
INFERNO 2	$15.71^{+0.15}_{-0.04}$	$16.87^{+0.19}_{-0.06}$	$\mathbf{16.95^{+0.18}_{-0.04}}$	$16.88^{+0.17}_{-0.03}$	$18.67^{+0.25}_{-0.05}$
INFERNO 3	$15.70^{+0.21}_{-0.04}$	$16.91^{+0.20}_{-0.05}$	$16.97^{+0.21}_{-0.04}$	$\mathbf{16.89^{+0.18}_{-0.03}}$	$18.69^{+0.27}_{-0.04}$
INFERNO 4	$15.71^{+0.32}_{-0.06}$	$16.89^{+0.30}_{-0.07}$	$16.95^{+0.38}_{-0.05}$	$16.88^{+0.40}_{-0.05}$	$\mathbf{18.68^{+0.58}_{-0.07}}$
Optimal classifier	14.97	19.12	24.93	22.13	27.98
Analytical likelihood	14.71	15.52	15.65	15.62	16.89



(a) different r value (b) different λ value

Figure 6.5: Expected uncertainty when the value of the nuisance parameters is different for 10 learnt summary statistics (different random initialisation) based on cross-entropy classification and inference-aware technique. Results correspond to Benchmark 2.

This synthetic example demonstrates that the direct optimisation of inference-aware losses as those described in the Section 6.3 is effective. The summary statistics learnt accounting for the effect of nuisance parameters compare very favourably to those obtained by using a classification proxy to approximate the likelihood ratio. Of course, more experiments are needed to benchmark the usefulness of this technique for real-world inference problems as those found in High Energy Physics analyses at the LHC.

7 Conclusions and Prospects

So Long,
and Thanks for All the Fish.

Douglas Adams

A large part of this thesis has dealt with the role of statistical learning techniques in the context of particle collider analyses, and their usefulness from a statistical inference perspective. After a broad introduction to the theoretical models of fundamental interactions and a summary of the main characteristics and working principles of the Compact Muon Solenoid (CMS) detector at the Large Hadron Collider (LHC), the fundamentals for statistical modelling at the LHC has been discussed. The relation between the theoretical parameters of interest and the experimental observations can only be modelled accurately by means of a complex simulation chain of the underlying physical processes and expected detector response. The generative-only nature of the simulation-based model combined with its high dimensionality make the definition of the probability density or likelihood function intractable, thus classical inference techniques cannot be applied to carry out statistical inference based on the acquired observations.

The statistical model for particle colliders can be described by a mixture model, each mixture component originating from a group of fundamental physical interactions. The latent variable structure of the generative model can be mapped to the different simulation steps in the simulation: process type, parton-level four-momenta, parton-shower outcome and detector readout. While the dimensionality of the latent space greatly increases for each subsequent step, the joint distribution can be factorised as a product of conditionals, the information about the parameters of interest being compactly expressed by the lowest dimensional latent variables. An efficient way to reduce the dimensionality of the data is thus to approximate the latent variables using the observations. This can be done by a well-calibrated combination of the different detector readouts, as is the case when using event reconstruction is performed, or by directly estimating the latent variables using supervised learning techniques trained on simulated observations.

Recent advances in supervised learning techniques have led to more accurate latent variable estimation that can scale to more data and use advanced non-linear transformations to obtain better performance in complex tasks, both in the context of classification and regression. Signal versus background probabilistic classification, a common conceptual framework for simplifying the event selection problem and constructing low-dimensional summaries in high-energy physics, has been formally proven to produce sufficient summary statistics for the mixture coefficients when the generative model is fully defined. The usefulness of probabilistic classification for such tasks, even in the optimal classifier case, cannot be guaranteed when nuisance parameters affect significantly the distribution of observed samples. In addition, particle identification and regression problems that augment the reconstruction output and can be tackled with machine learning techniques are also discussed. The use of deep learning techniques for advanced jet flavour tagging in CMS are used to exemplify the previous use case, which demonstrates the possible performance improvements due to the combined use of deep neural networks and non-standard input transformations that can deal with sequences. Newer machine learning methodologies that can deal with sets, graphs and other types of non-vector input coupled with powerful parallel hardware could be a promising path to substitute a larger part of the event reconstruction chain by latent variable approximations based on simulated observations, providing higher accuracy and throughput than hand-tuned algorithms.

An analysis using 35.9 fb^{-1} of data collected in 2016 by the CMS detector at the LHC was also included in this work. Proton-proton collisions at a centre-of-mass energy of 13 TeV were used to study the $\text{pp} \to \text{HH} \to \text{b}\bar{\text{b}}\text{b}\bar{\text{b}}$ process in the context of the Standard Model (SM) and anomalous couplings effective field theory (EFT) extensions. The main challenge for this LHC analysis was the large background contribution from multi-jet QCD processes, so numerous that could not be modelled accurately by simulated observations. Hence, a data-driven estimation method, referred to as hemisphere mixing, was developed and validated on control regions to model the background contribution. The final summary statistic used in the analysis is based on the output of a probabilistic classifier, an ensemble of gradient boosted decision trees, trained using simulated signal observations and artificial events produced by the background estimation method. After assessing the different sources of systematic uncertainties and including their effect in the statistical model, a median expected limit obtained for SM HH production of 419 fb was obtained, which corresponds to approximately 37 times the SM expectation. The observed limit obtained

is 847 fb, which is about two standard deviations above the expected limit. Limits were also obtained for a set of EFT benchmarks, which summarise the kinematical properties of a large space of EFT models. The results of the combination of this analysis with other HH decay channels were also included. The estimation of QCD multijet backgrounds will likely remain an important issue for future jet-based analysis at the LHC, given that the biases of the data-driven estimation methods would become increasingly relevant as more data is available.

The ultimate goal of LHC analyses is statistical inference, in the form of hypothesis testing or parameter estimation. Machine learning techniques are useful to approximate latent variables which can then be used to construct powerful summary statistics for inference. In the presence of a generative model that depends on additional uncertain parameters, often referred to as nuisance parameters, the merits of classification or regression based summary statistics are greatly diminished. These concerns have motivated the development of a new family of techniques to construct powerful summary statistics that account directly for the final inference objective. By building and minimising loss functions that approximate the expected uncertainty on the parameters of interest, also accounting for the effect of nuisance parameters, the INFERNO approach can leverage recent machine learning technologies to construct better summary statistics for the inference problem at hand. These techniques were applied to a series of synthetic problems and were found to significantly outperform classification-based summary statistics (e.g. a deep neural network and the optimal classifier) when nuisance parameters are included in the problem. More experiments are needed to evaluate the value of this technique for real-word inference problems, such as those found in particle physics analyses.

As machine learning algorithms become increasingly popular in scientific contexts, it will be more important to formally describe the particularities of the problems we are trying to solve, in order to understand whether the tools at hand are answering the right questions. Otherwise we risk falling for the anti-pattern "if all you have is a hammer, everything looks like a nail", which could significantly slow down the pace of scientific progress. This issue is particularly pressing for particle collider experiments, where the acquired familiarity with a given set of data analysis techniques might hinder the rigour in their application relative to the final objective. Some effort is then required to make sure of the role of a given tool is aligned with the task at hand instead on the subtleties of the tool itself. When using advanced statistical techniques or machine learning, the final analysis goal is of the upmost relevance and cannot be neglected in favour of procedural conventions. If those measures are

coupled with open research practices and a careful use of domain-specific language and constructs in order to promote collaboration with other disciplines, better tools are likely to be developed which could in turn lead to major advancements in this research field.

Bibliography

[1] Steven Weinberg. 'The Making of the standard model'. In: *The European Physical Journal C-Particles and Fields* 34.1 (2004), pp. 5–13.

[2] Serguei Chatrchyan et al. 'Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC'. In: *Physics Letters B* 716.1 (2012), pp. 30–61.

[3] Georges Aad et al. 'Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC'. In: *Physics Letters B* 716.1 (2012), pp. 1–29.

[4] Michael E Peskin and Daniel V Schroeder. *An introduction to quantum field theory*. Includes exercises. Boulder, CO: Westview, 1995. URL: https://cds.cern.ch/record/257493.

[5] Franz Mandl and Graham Shaw. *Quantum Field Theory*. John Wiley & Sons, 2010.

[6] Dave Goldberg. *The standard model in a nutshell*. Princeton, NJ: Princeton University Press, 2017. URL: http://cds.cern.ch/record/2244785.

[7] Gian-Carlo Wick. 'The evaluation of the collision matrix'. In: *Physical review* 80.2 (1950), p. 268.

[8] Particle Data Group. 'Review of Particle Physics'. In: *Phys. Rev. D* 98 (3 Aug. 2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[9] LHCb Collaboration. 'Observation of the Resonant Character of the $Z(4430)^-$ State'. In: *Phys. Rev. Lett.* 112 (22 June 2014), p. 222002. DOI: 10.1103/PhysRevLett.112.222002. URL: https://link.aps.org/doi/10.1103/PhysRevLett.112.222002.

[10] LHCb Collaboration. 'Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to J/\psi K^- p$ Decays'. In: *Phys. Rev. Lett.* 115 (2015), p. 072001. DOI: 10.1103/PhysRevLett.115.072001. arXiv: 1507.03414 [hep-ex].

[11] Enrico Fermi. 'An attempt of a theory of beta radiation.' In: *Z. Phys.* 88.UCRL-TRANS-726 (1934), pp. 161–177.

[12] Sheldon L Glashow. 'Partial-symmetries of weak interactions'. In: *Nuclear Physics* 22.4 (1961), pp. 579–588.

[13] A. Salam and J.C. Ward. 'Electromagnetic and weak interactions'. In: *Physics Letters* 13.2 (1964), pp. 168–171. ISSN: 0031-9163. DOI: https://doi.org/10.1016/0031-9163(64)90711-5. URL: http://www.sciencedirect.com/science/article/pii/0031916364907115.

[14] François Englert and Robert Brout. 'Broken symmetry and the mass of gauge vector mesons'. In: *Physical Review Letters* 13.9 (1964), p. 321.

[15] Peter W Higgs. 'Broken symmetries and the masses of gauge bosons'. In: *Physical Review Letters* 13.16 (1964), p. 508.

[16] Gerald S Guralnik, Carl R Hagen and Thomas WB Kibble. 'Global conservation laws and massless particles'. In: *Physical Review Letters* 13.20 (1964), p. 585.

[17] Steven Weinberg. 'A model of leptons'. In: *Physical review letters* 19.21 (1967), p. 1264.

[18] G. 't Hooft and M. Veltman. 'Regularization and renormalization of gauge fields'. In: *Nuclear Physics B* 44.1 (1972), pp. 189–213. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(72)90279-9. URL: http://www.sciencedirect.com/science/article/pii/0550321372902799.

[19] FJ Hasert et al. 'Observation of neutrino-like interactions without muon or electron in the Gargamelle neutrino experiment'. In: *Nuclear Physics B* 73.1 (1974), pp. 1–22.

[20] UA1 Collaboration. 'Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at s**(1/2) = 540-GeV'. In: *Phys. Lett.* B122 (1983). [,611(1983)], pp. 103–116. DOI: 10.1016/0370-2693(83)91177-2.

[21] UA2 Collaboration. 'Observation of Single Isolated Electrons of High Transverse Momentum in Events with Missing Transverse Energy at the CERN anti-p p Collider'. In: *Phys. Lett.* B122 (1983). [,7.45(1983)], pp. 476–485. DOI: 10.1016/0370-2693(83)91605-2.

[22] UA1 Collaboration. 'Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c**2 at the CERN SPS Collider'. In: *Phys. Lett.* B126 (1983). [,7.55(1983)], pp. 398–410. DOI: 10.1016/0370-2693(83)90188-0.

[23] UA2 Collaboration. 'Evidence for Z0 —> e+ e- at the CERN anti-p p Collider'. In: *Phys. Lett.* B129 (1983). [,7.69(1983)], pp. 130–140. DOI: 10.1016/0370-2693(83)90744-X.

[24] C. S. Wu et al. 'Experimental Test of Parity Conservation in Beta Decay'. In: *Phys. Rev.* 105 (1957), pp. 1413–1414. DOI: 10.1103/PhysRev.105.1413.

[25] Nicola Cabibbo. 'Unitary Symmetry and Leptonic Decays'. In: *Phys. Rev. Lett.* 10 (1963). [,648(1963)], pp. 531–533. DOI: 10.1103/PhysRevLett.10.531.

[26] Makoto Kobayashi and Toshihide Maskawa. 'CP Violation in the Renormalizable Theory of Weak Interaction'. In: *Prog. Theor. Phys.* 49 (1973), pp. 652–657. DOI: 10.1143/PTP.49.652.

[27] ATLAS and CMS Collaborations. 'Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments'. In: *Phys. Rev. Lett.* 114 (2015), p. 191803. DOI: 10.1103/PhysRevLett.114.191803. arXiv: 1503.07589 [hep-ex].

[28] D Hanneke, S Fogwell and G Gabrielse. 'New measurement of the electron magnetic moment and the fine structure constant'. In: *Physical Review Letters* 100.12 (2008), p. 120801.

[29] Richard H Parker et al. 'Measurement of the fine-structure constant as a test of the Standard Model'. In: *Science* 360.6385 (2018), pp. 191–195.

[30] Charles W Misner et al. *Gravitation*. Princeton University Press, 2017.

[31] Carlo Rovelli. 'Loop quantum gravity'. In: *Living reviews in relativity* 11.1 (2008), p. 5.

[32] Joseph Polchinski. *String Theory*. Cambridge monographs on mathematical physics. Cambridge: Cambridge Univ. Press, 1998. URL: http://cds.cern.ch/record/363850.

[33] Edvige Corbelli and Paolo Salucci. 'The extended rotation curve and the dark matter halo of M33'. In: *Monthly Notices of the Royal Astronomical Society* 311.2 (2000), pp. 441–447.

[34] Virginia Trimble. 'Existence and nature of dark matter in the universe'. In: *Annual review of astronomy and astrophysics* 25.1 (1987), pp. 425–472.

[35] Planck. 'Planck 2015 results. XIII. Cosmological parameters'. In: *Astron. Astrophys.* 594 (2016), A13. DOI: 10.1051/0004-6361/201525830. arXiv: 1502.01589 [astro-ph.CO].

[36] Y Fukuda et al. 'Evidence for oscillation of atmospheric neutrinos'. In: *Physical Review Letters* 81.8 (1998), p. 1562.

[37] SNO collaboration et al. 'Measurement of the rate of nu_e+ d–> p+ p+ e^-interactions produced by 8B solar neutrinos at the Sudbury Neutrino Observatory'. In: *arXiv preprint nucl-ex/0106015* (2001).

[38] Evgeny K. Akhmedov, G. C. Branco and M. N. Rebelo. 'Seesaw mechanism and structure of neutrino mass matrix'. In: *Phys. Lett.* B478 (2000), pp. 215–223. DOI: 10.1016/S0370-2693(00)00282-3. arXiv: hep-ph/9911364 [hep-ph].

[39] Adam G Riess et al. 'Type Ia supernova discoveries at z> 1 from the Hubble Space Telescope: Evidence for past deceleration and constraints on dark energy evolution'. In: *The Astrophysical Journal* 607.2 (2004), p. 665.

[40] Ronald J Adler, Brendan Casey and Ovid C Jacob. 'Vacuum catastrophe: An elementary exposition of the cosmological constant problem'. In: *American Journal of Physics* 63.7 (1995), pp. 620–626.

[41] Giuseppe Degrassi et al. 'Higgs mass and vacuum stability in the Standard Model at NNLO'. In: *Journal of High Energy Physics* 2012.8 (2012), p. 98.

[42] Hai-Yang Cheng. 'The strong CP problem revisited'. In: *Physics Reports* 158.1 (1988), pp. 1–89.

[43] Thomas Appelquist and J. Carazzone. 'Infrared singularities and massive fields'. In: *Phys. Rev. D* 11 (10 May 1975), pp. 2856–2861. DOI: 10.1103/PhysRevD.11.2856. URL: https://link.aps.org/doi/10.1103/PhysRevD.11.2856.

[44] W. Buchmuller and D. Wyler. 'Effective Lagrangian Analysis of New Interactions and Flavor Conservation'. In: *Nucl. Phys.* B268 (1986), pp. 621–653. DOI: 10.1016/0550-3213(86)90262-2.

[45] Steven Weinberg. 'Baryon-and lepton-nonconserving processes'. In: *Physical Review Letters* 43.21 (1979), p. 1566.

[46] NNPDF. 'Parton distributions from high-precision collider data'. In: *Eur. Phys. J.* C77.10 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].

[47] Guido Altarelli and G. Parisi. 'Asymptotic Freedom in Parton Language'. In: *Nucl. Phys.* B126 (1977), pp. 298–318. DOI: 10.1016/0550-3213(77)90384-4.

[48] Yuri L. Dokshitzer. 'Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.' In: *Sov. Phys. JETP* 46 (1977). [Zh. Eksp. Teor. Fiz.73,1216(1977)], pp. 641–653.

[49] V. N. Gribov and L. N. Lipatov. 'Deep inelastic e p scattering in perturbation theory'. In: *Sov. J. Nucl. Phys.* 15 (1972). [Yad. Fiz.15,781(1972)], pp. 438–450.

[50] John C. Collins, Davison E. Soper and George F. Sterman. 'Factorization of Hard Processes in QCD'. In: *Adv. Ser. Direct. High Energy Phys.* 5 (1989), pp. 1–91. DOI: 10.1142/9789814503266_0001. arXiv: hep-ph/0409313 [hep-ph].

[51] G. Peter Lepage. 'A New Algorithm for Adaptive Multidimensional Integration'. In: *J. Comput. Phys.* 27 (1978), p. 192. DOI: 10.1016/0021-9991(78)90004-9.

[52] Stefan Höche. 'Introduction to parton-shower event generators'. In: *Proceedings, Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders (TASI 2014): Boulder, Colorado, June 2-27, 2014*. 2015, pp. 235–295. DOI: 10.1142/9789814678766_0005. arXiv: 1411.4085 [hep-ph].

[53] CERN Service graphique. 'Overall view of the LHC. Vue d'ensemble du LHC'. In: (June 2014). General Photo. URL: https://cds.cern.ch/record/1708849.

[54] Bernhard Wolf. *Handbook of ion sources*. CRC press, 2017.

[55] Thomas Mc Cauley. 'Collisions recorded by the CMS detector on 14 Oct 2016 during the high pile-up fill'. CMS Collection. Nov. 2016. URL: https://cds.cern.ch/record/2231915.

[56] ATLAS Collaboration. 'The ATLAS Experiment at the CERN Large Hadron Collider'. In: *JINST* 3 (2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003.

[57] CMS Collaboration. 'The CMS Experiment at the CERN LHC'. In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[58] LHCb Collaboration. 'The LHCb Detector at the LHC'. In: *JINST* 3 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005.

[59] ALICE Collaboration. 'The ALICE experiment at the CERN LHC'. In: *JINST* 3 (2008), S08002. DOI: 10.1088/1748-0221/3/08/S08002.

[60] TOTEM Collaboration. 'The TOTEM experiment at the CERN Large Hadron Collider'. In: *JINST* 3 (2008), S08007. DOI: 10.1088/1748-0221/3/08/S08007.

[61] LHCf Collaboration. 'The LHCf detector at the CERN Large Hadron Collider'. In: *JINST* 3 (2008), S08006. DOI: 10.1088/1748-0221/3/08/S08006.

[62] MoEDAL Collaboration. 'The Physics Programme Of The MoEDAL Experiment At The LHC'. In: *Int. J. Mod. Phys.* A29 (2014), p. 1430050. DOI: 10.1142/S0217751X14300506. arXiv: 1405.7662 [hep-ph].

[63] CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/922757.

[64] Tai Sakuma and Thomas McCauley. 'Detector and event visualization with sketchup at the cms experiment'. In: *Journal of Physics: Conference Series*. Vol. 513. 2. IOP Publishing. 2014, p. 022032.

[65] CMS Collaboration. 'Description and performance of track and primary-vertex reconstruction with the CMS tracker'. In: *JINST* 9.10 (2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det].

[66] Helmuth Spieler. *Semiconductor detector systems*. Vol. 12. Oxford university press, 2005.

[67] *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical Design Report CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/349375.

[68] CMS Collaboration. 'Performance of the CMS Hadron Calorimeter with Cosmic Ray Muons and LHC Beam Data'. In: *JINST* 5 (2010), T03012. DOI: 10.1088/1748-0221/5/03/T03012. arXiv: 0911.4991 [physics.ins-det].

[69] CMS Collaboration. 'Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV'. In: *JINST* 13.06 (2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528 [physics.ins-det].

[70] CMS Collaboration. 'Particle-flow reconstruction and global event description with the CMS detector'. In: *JINST* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].

[71] GEANT4 Collaboration. 'GEANT4: A Simulation toolkit'. In: *Nucl. Instrum. Meth.* A506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[72] CMS Collaboration. 'The fast simulation of the CMS detector at LHC'. In: *J. Phys. Conf. Ser.* 331 (2011), p. 032049. DOI: 10.1088/1742-6596/331/3/032049.

[73] DELPHES 3 Collaboration. 'DELPHES 3, A modular framework for fast simulation of a generic collider experiment'. In: *JHEP* 02 (2014), p. 057. DOI: 10.1007/JHEP02(2014)057. arXiv: 1307.6346 [hep-ex].

[74] Michela Paganini, Luke de Oliveira and Benjamin Nachman. 'Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters'. In: *Phys. Rev. Lett.* 120.4 (2018), p. 042003. DOI: 10.1103/PhysRevLett.120.042003. arXiv: 1705.02355 [hep-ex].

[75] Luke de Oliveira, Michela Paganini and Benjamin Nachman. 'Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis'. In: *Comput. Softw. Big Sci.* 1.1 (2017), p. 4. DOI: 10.1007/s41781-017-0004-6. arXiv: 1701.05927 [stat.ML].

[76] Pierre Billoir and S. Qian. 'Simultaneous pattern recognition and track fitting by the Kalman filtering method'. In: *Nucl. Instrum. Meth.* A294 (1990), pp. 219–228. DOI: 10.1016/0168-9002(90)91835-Y.

[77] R Mankel. *A concurrent track evolution algorithm for pattern recognition in the HERA-B main tracking system*. Tech. rep. DESY-97-054. Hamburg: DESY, Mar. 1997. URL: http://cds.cern.ch/record/334615.

[78] R. Fruhwirth, W. Waltenberger and P. Vanlaer. 'Adaptive vertex fitting'. In: *J. Phys.* G34 (2007), N343. DOI: 10.1088/0954-3899/34/12/N01.

[79] Wolfgang Adam et al. 'Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC'. In: (2005).

[80] CMS Collaboration. 'Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at sqrt(s) = 8 TeV'. In: *JINST* 10.06 (2015), P06005. DOI: 10.1088/1748-0221/10/06/P06005. arXiv: 1502.02701 [physics.ins-det].

[81] CMS Collaboration. 'Performance of Photon Reconstruction and Identification with the CMS Detector in Proton-Proton Collisions at sqrt(s) = 8 TeV'. In: *JINST* 10.08 (2015), P08010. DOI: 10.1088/1748-0221/10/08/P08010. arXiv: 1502.02702 [physics.ins-det].

[82] CMS Collaboration. 'Pileup Removal Algorithms'. In: (2014).

[83] Matteo Cacciari, Gavin P. Salam and Gregory Soyez. 'The anti-k_t jet clustering algorithm'. In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[84] CMS Collaboration. 'Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV'. In: *JINST* 12.02 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex].

[85] CMS Collaboration. 'Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV'. In: *JINST* 13.05 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det].

[86] Mario Lezcano Casado et al. 'Improvements to Inference Compilation for Probabilistic Programming in Large-Scale Scientific Simulators'. In: 2017. arXiv: 1712.07901 [cs.AI].

[87] Atilim Gunes Baydin et al. 'Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model'. In: (2018). arXiv: 1807.07706 [cs.LG].

[88] Crispin W Gardiner. *Handbook of stochastic methods: for physics, chemistry and the natural sciences; 3rd ed.* Springer Series in Synergetics. Berlin: Springer, 2004. URL: https://cds.cern.ch/record/732221.

[89] Robert V Hogg and Allen T Craig. *Introduction to mathematical statistics.(5th edition).* Upper Saddle River, New Jersey: Prentice Hall, 1995.

[90] Kyle Cranmer. 'Practical Statistics for the LHC'. In: *Proceedings, 2011 European School of High-Energy Physics (ESHEP 2011): Cheile Gradistei, Romania, September 7-20, 2011*. [,247(2015)]. 2015, pp. 267–308. DOI: 10.5170/CERN-2015-001.247, 10.5170/CERN-2014-003.267. arXiv: 1503.07622 [physics.data-an].

[91] J. S. Conway. 'Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra'. In: *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011*. 2011, pp. 115–120. DOI: 10.5170/CERN-2011-006.115. arXiv: 1103.0354 [physics.data-an].

[92] Kyle Cranmer et al. 'HistFactory: A tool for creating statistical models for use with RooFit and RooStats'. In: (2012).

[93] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[94] Donald B Rubin. 'Bayesianly justifiable and relevant frequency calculations for the applies statistician'. In: *The Annals of Statistics* (1984), pp. 1151–1172.

[95] Mark A Beaumont, Wenyang Zhang and David J Balding. 'Approximate Bayesian computation in population genetics'. In: *Genetics* 162.4 (2002), pp. 2025–2035.

[96] Johann Brehmer et al. 'A Guide to Constraining Effective Field Theories with Machine Learning'. In: *Phys. Rev.* D98.5 (2018), p. 052004. DOI: 10.1103/PhysRevD.98.052004. arXiv: 1805.00020 [hep-ph].

[97] J. Neyman and E. S. Pearson. 'On the Problem of the Most Efficient Tests of Statistical Hypotheses'. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (1933), pp. 289–337. ISSN: 02643952. URL: http://www.jstor.org/stable/91247.

[98] Samuel S Wilks. 'The large-sample distribution of the likelihood ratio for testing composite hypotheses'. In: *The Annals of Mathematical Statistics* 9.1 (1938), pp. 60–62.

[99] Abraham Wald. 'Tests of statistical hypotheses concerning several parameters when the number of observations is large'. In: *Transactions of the American Mathematical society* 54.3 (1943), pp. 426–482.

[100] Glen Cowan et al. 'Asymptotic formulae for likelihood-based tests of new physics'. In: *Eur. Phys. J.* C71 (2011). [Erratum: Eur. Phys. J.C73,2501(2013)], p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0,10.1140/epjc/s10052-013-2501-z. arXiv: 1007.1727 [physics.data-an].

[101] Alexander L. Read. 'Presentation of search results: The CL(s) technique'. In: *J. Phys.* G28 (2002). [,11(2002)], pp. 2693–2704. DOI: 10.1088/0954-3899/28/10/313.

[102] Thomas Junk. 'Confidence level computation for combining searches with small statistics'. In: *Nucl. Instrum. Meth.* A434 (1999), pp. 435–443. DOI: 10.1016/S0168-9002(99)00498-2. arXiv: hep-ex/9902006 [hep-ex].

[103] J. Neyman. 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability'. In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236.767 (1937), pp. 333–380. ISSN: 00804614. URL: http://www.jstor.org/stable/91337.

[104] Gary J. Feldman and Robert D. Cousins. 'A Unified approach to the classical statistical analysis of small signals'. In: *Phys. Rev.* D57 (1998), pp. 3873–3889. DOI: 10.1103/PhysRevD.57.3873. arXiv: physics/9711021 [physics.data-an].

[105] Wolfgang A. Rolke, Angel M. Lopez and Jan Conrad. 'Limits and confidence intervals in the presence of nuisance parameters'. In: *Nucl. Instrum. Meth.* A551 (2005), pp. 493–503. DOI: 10.1016/j.nima.2005.05.068. arXiv: physics/0403059 [physics].

[106] Fred James and MINUIT Roos. 'MINUIT: a system for function minimization and analysis of the parameter errors and corrections'. In: *Comput. Phys. Commun.* 10.CERN-DD-75-20 (1975), pp. 343–367.

[107] R. A. Fisher. 'Theory of Statistical Estimation'. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 22.5 (1925), pp. 700–725. DOI: 10.1017/S0305004100009580.

[108] Harald Cramér. *Mathematical methods of statistics (PMS-9)*. Vol. 9. Princeton university press, 2016.

[109] C. Radhakrishna Rao. 'Information and the accuracy attainable in the estimation of statistical parameters'. In: *Breakthroughs in statistics*. Springer, 1992, pp. 235–247.

[110] Pierre Simon Laplace. 'Memoir on the probability of the causes of events'. In: *Statistical Science* 1.3 (1986), pp. 364–378.

[111] Thomas M. Mitchell. *Machine Learning*. 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN: 0070428077, 9780070428072.

[112] Vladimir Naumovich Vapnik. 'An overview of statistical learning theory'. In: *IEEE transactions on neural networks* 10.5 (1999), pp. 988–999.

[113] Jerome Friedman, Trevor Hastie and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA: 2001.

[114] Tan Nguyen and Scott Sanner. 'Algorithms for direct 0–1 loss optimization in binary classification'. In: *International Conference on Machine Learning*. 2013, pp. 1085–1093.

[115] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[116] Gilles Louppe. 'Understanding random forests: From theory to practice'. In: *arXiv preprint arXiv:1407.7502* (2014).

[117] Yoav Freund and Robert E Schapire. 'A decision-theoretic generalization of on-line learning and an application to boosting'. In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[118] Jerome Friedman, Trevor Hastie, Robert Tibshirani et al. 'Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)'. In: *The annals of statistics* 28.2 (2000), pp. 337–407.

[119] Jerome H Friedman. 'Greedy function approximation: a gradient boosting machine'. In: *Annals of statistics* (2001), pp. 1189–1232.

[120] Llew Mason et al. 'Boosting algorithms as gradient descent'. In: *Advances in neural information processing systems*. 2000, pp. 512–518.

[121] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[122] Leo Breiman. 'Bagging predictors'. In: *Machine learning* 24.2 (1996), pp. 123–140.

[123] Tianqi Chen and Carlos Guestrin. 'Xgboost: A scalable tree boosting system'. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[124] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. second. New York, NY, USA: Springer, 2006.

[125] Herbert Robbins and Sutton Monro. 'A Stochastic Approximation Method'. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[126] Sebastian Ruder. 'An overview of gradient descent optimization algorithms'. In: *arXiv preprint arXiv:1609.04747* (2016).

[127] George Cybenko. 'Approximation by superpositions of a sigmoidal function'. In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[128] Atilim Gunes Baydin et al. 'Automatic differentiation in machine learning: a survey'. In: *Journal of Marchine Learning Research* 18 (2018), pp. 1–43.

[129] Martin Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[130] Adam Paszke et al. 'Automatic differentiation in PyTorch'. In: *NIPS-W*. 2017.

[131] Manzil Zaheer et al. 'Deep sets'. In: *Advances in Neural Information Processing Systems*. 2017, pp. 3391–3401.

[132] Isaac Henrion et al. 'Neural message passing for jet physics'. In: (2017).

[133] Dan Guest, Kyle Cranmer and Daniel Whiteson. 'Deep Learning and its Application to LHC Physics'. In: *Ann. Rev. Nucl. Part. Sci.* 68 (2018), pp. 161–181. DOI: 10.1146/annurev-nucl-101917-021019. arXiv: 1806.11484 [hep-ex].

[134] Pierre Baldi et al. 'Parameterized neural networks for high-energy physics'. In: *The European Physical Journal C* 76.5 (2016), p. 235.

[135] Daniel Guest et al. 'Jet Flavor Classification in High-Energy Physics with Deep Neural Networks'. In: *Phys. Rev.* D94.11 (2016), p. 112002. DOI: 10.1103/PhysRevD.94.112002. arXiv: 1607.08633 [hep-ex].

[136] Luke de Oliveira et al. 'Jet-images — deep learning edition'. In: *JHEP* 07 (2016), p. 069. DOI: 10.1007/JHEP07(2016)069. arXiv: 1511.05190 [hep-ph].

[137] François Chollet et al. *Keras*. https://keras.io. 2015.

[138] 'Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13TeV with Phase 1 CMS detector'. In: (Nov. 2018). URL: http://cds.cern.ch/record/2646773.

[139] Markus Stoye et al. 'DeepJet: Generic physics object based jet multiclass classification for LHC experiments'. In:

[140] 'Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector'. In: (July 2018). URL: http://cds.cern.ch/record/2630438.

[141] Vincenzo Innocente, L Silvestris, D Stickland et al. 'CMS Software Architecture: Software framework, services and persistency in high level trigger, reconstruction and analysis'. In: *Computer Physics Communications* 140.1-2 (2001), pp. 31–44.

[142] Ian Bird and Roger WL Jones. 'LHC computing grid: technical design report'. In: (2005).

[143] Daniel Hay Guest et al. *lwtnn/lwtnn: Version 2.8*. Nov. 2018. DOI: 10.5281/zenodo.1482645. URL: https://doi.org/10.5281/zenodo.1482645.

[144] Marcel Rieger. *CMSSW-DNN*. https://gitlab.cern.ch/mrieger/CMSSW-DNN. 2017.

[145] Pablo de Castro, Marcel Rieger et al. *DeepJet integration*. https://github.com/cms-sw/cmssw/pull/19893. 2017.

[146] Markus Stoye et al. *DeepJet software framework*. https://github.com/mstoye/DeepJet. 2017.

[147] P. De Castro Manzano et al. 'Hemisphere Mixing: a Fully Data-Driven Model of QCD Multijet Backgrounds for LHC Searches'. In: *PoS* EPS-HEP2017 (2017), p. 370. DOI: 10.22323/1.314.0370. arXiv: 1712.02538 [hep-ex].

[148] CMS Collaboration. 'Search for nonresonant Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state at $\sqrt{s} = 13$ TeV'. In: *Submitted to: JHEP* (2018). arXiv: 1810.11854 [hep-ex].

[149] CMS Collaboration. 'Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV'. In: *JHEP* 06 (2013), p. 081. DOI: 10.1007/JHEP06(2013)081. arXiv: 1303.4571 [hep-ex].

[150] ATLAS, CMS. 'Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV'. In: *JHEP* 08 (2016), p. 045. DOI: 10.1007/JHEP08(2016)045. arXiv: 1606.02266 [hep-ex].

[151] CMS Collaboration. 'Observation of $t\bar{t}H$ production'. In: *Phys. Rev. Lett.* 120.23 (2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801,10.1130/PhysRevLett.120.231801. arXiv: 1804.02610 [hep-ex].

[152] ATLAS Collaboration. 'Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector'. In: *Phys. Lett.* B784 (2018), pp. 173–191. DOI: 10.1016/j.physletb.2018.07.035. arXiv: 1806.00425 [hep-ex].

[153] Claudio O. Dib, Rogerio Rosenfeld and Alfonso Zerwekh. 'Double Higgs production and quadratic divergence cancellation in little Higgs models with T parity'. In: *JHEP* 05 (2006), p. 074. DOI: 10.1088/1126-6708/2006/05/074. arXiv: hep-ph/0509179 [hep-ph].

[154] R. Grober and M. Muhlleitner. 'Composite Higgs Boson Pair Production at the LHC'. In: *JHEP* 06 (2011), p. 020. DOI: 10.1007/JHEP06(2011)020. arXiv: 1012.1562 [hep-ph].

[155] Roberto Contino et al. 'Anomalous Couplings in Double Higgs Production'. In: *JHEP* 08 (2012), p. 154. DOI: 10.1007/JHEP08(2012)154. arXiv: 1205.5444 [hep-ph].

[156] Matthew J. Dolan, Christoph Englert and Michael Spannowsky. 'New Physics in LHC Higgs boson pair production'. In: *Phys. Rev.* D87.5 (2013), p. 055002. DOI: 10.1103/PhysRevD.87.055002. arXiv: 1210.8166 [hep-ph].

[157] S. Dawson, A. Ismail and Ian Low. 'What's in the loop? The anatomy of double Higgs production'. In: *Phys. Rev.* D91.11 (2015), p. 115008. DOI: 10.1103/PhysRevD.91.115008. arXiv: 1504.05596 [hep-ph].

[158] J. Baglio et al. 'The measurement of the Higgs self-coupling at the LHC: theoretical status'. In: *JHEP* 04 (2013), p. 151. DOI: 10.1007/JHEP04(2013)151. arXiv: 1212.5581 [hep-ph].

[159] CMS Collaboration. 'Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV'. In: *JHEP* 11 (2017), p. 047. DOI: 10.1007/JHEP11(2017)047. arXiv: 1706.09936 [hep-ex].

[160] LHC Higgs Cross Section Working Group. 'Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector'. In: (2016). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[161] Daniel de Florian and Javier Mazzitelli. 'Higgs Boson Pair Production at Next-to-Next-to-Leading Order in QCD'. In: *Phys. Rev. Lett.* 111 (2013), p. 201801. DOI: 10.1103/PhysRevLett.111.201801. arXiv: 1309.6594 [hep-ph].

[162] S. Dawson, S. Dittmaier and M. Spira. 'Neutral Higgs boson pair production at hadron colliders: QCD corrections'. In: *Phys. Rev.* D58 (1998), p. 115012. DOI: 10.1103/PhysRevD.58.115012. arXiv: hep-ph/9805244 [hep-ph].

[163] S. Borowka et al. 'Higgs Boson Pair Production in Gluon Fusion at Next-to-Leading Order with Full Top-Quark Mass Dependence'. In: *Phys. Rev. Lett.* 117.1 (2016). [Erratum: Phys. Rev. Lett.117,no.7,079901(2016)], p. 012001. DOI: 10.1103/PhysRevLett.117.079901, 10.1103/PhysRevLett.117.012001. arXiv: 1604.06447 [hep-ph].

[164] Daniel de Florian and Javier Mazzitelli. 'Higgs pair production at next-to-next-to-leading logarithmic accuracy at the LHC'. In: *JHEP* 09 (2015), p. 053. DOI: 10.1007/JHEP09(2015)053. arXiv: 1505.07122 [hep-ph].

[165] Alexandra Carvalho et al. 'Analytical parametrization and shape classification of anomalous HH production in the EFT approach'. In: (2016). arXiv: 1608.06578 [hep-ph].

[166] ATLAS Collaboration. 'Search for Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state from pp collisions at $\sqrt{s} = 8$ TeVwith the ATLAS detector'. In: *Eur. Phys. J.* C75.9 (2015), p. 412. DOI: 10.1140/epjc/s10052-015-3628-x. arXiv: 1506.00285 [hep-ex].

[167] CMS Collaboration. 'Search for Higgs boson pair production in the $bb\tau\tau$ final state in proton-proton collisions at $\sqrt{(s)} = 8$ TeV'. In: *Phys. Rev.* D96.7 (2017), p. 072004. DOI: 10.1103/PhysRevD.96.072004. arXiv: 1707.00350 [hep-ex].

[168] ATLAS Collaboration. 'Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector'. In: (2018). arXiv: 1804.06174 [hep-ex].

[169] CMS Collaboration. 'Search for resonant and nonresonant Higgs boson pair production in the $b\bar{b}\ell\nu\ell\nu$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV'. In: *JHEP* 01 (2018), p. 054. DOI: 10.1007/JHEP01(2018)054. arXiv: 1708.04188 [hep-ex].

[170] CMS Collaboration. 'Search for Higgs boson pair production in events with two bottom quarks and two tau leptons in proton-proton collisions at $\sqrt{s}$ =13TeV'. In: *Phys. Lett.* B778 (2018), pp. 101–127. DOI: 10.1016/j.physletb.2018.01.001. arXiv: 1707.02909 [hep-ex].

[171] CMS Collaboration. 'Search for Higgs boson pair production in the $\gamma\gamma b\bar{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV'. In: (2018). arXiv: 1806.00408 [hep-ex].

[172] CMS Collaboration. 'Search for production of Higgs boson pairs in the four b quark final state using large-area jets in proton-proton collisions at $\sqrt{s} =$ 13 TeV'. In: (2018). arXiv: 1808.01473 [hep-ex].

[173] Adam Falkowski. 'Higgs Basis: Proposal for an EFT basis choice for LHC HXSWG'. In: (Mar. 2015). URL: https://cds.cern.ch/record/2001958.

[174] Alexandra Carvalho et al. 'Higgs Pair Production: Choosing Benchmarks With Cluster Analysis'. In: *JHEP* 04 (2016), p. 126. DOI: 10.1007/JHEP04(2016)126. arXiv: 1507.02245 [hep-ph].

[175] CMS Collaboration. 'Search for resonant pair production of Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at 13 TeV'. In: *JHEP* 08 (2018), p. 152. DOI: 10.1007/JHEP08(2018)152. arXiv: 1806.03548 [hep-ex].

[176] J. Alwall et al. 'The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations'. In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[177] Benoit Hespel, David Lopez-Val and Eleni Vryonidou. 'Higgs pair production via gluon fusion in the Two-Higgs-Doublet Model'. In: *JHEP* 09 (2014), p. 124. DOI: 10.1007/JHEP09(2014)124. arXiv: 1407.0281 [hep-ph].

[178] NNPDF Collaboration. 'Parton distributions for the LHC Run II'. In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[179] Sebastien Wertz and Vincent Lemaitre. 'Search for Higgs boson pair production in the $b\bar{b}\ell\nu\ell\nu$ final state with the CMS detector'. 2018. URL: http://cds.cern.ch/record/2632195.

[180] Fabian Pedregosa et al. 'Scikit-learn: Machine learning in Python'. In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[181] CMS Collaboration. 'Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV'. In: *JHEP* 07 (2018), p. 161. DOI: 10.1007/JHEP07(2018)161. arXiv: 1802.02613 [hep-ex].

[182] CMS Collaboration. *CMS Luminosity Measurements for the 2016 Data Taking Period*. Tech. rep. CMS-PAS-LUM-17-001. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2257069.

[183] Jon Butterworth et al. 'PDF4LHC recommendations for LHC Run II'. In: *J. Phys.* G43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865 [hep-ph].

[184] The ATLAS Collaboration, The CMS Collaboration, The LHC Higgs Combination Group. *Procedure for the LHC Higgs boson search combination in Summer 2011*. Tech. rep. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11. Geneva: CERN, Aug. 2011. URL: https://cds.cern.ch/record/1379837.

[185] CMS Collaboration. *Combination of searches for Higgs boson pair production in proton-proton collisions at* $\sqrt{s} = 13$ TeV. Tech. rep. CMS-PAS-HIG-17-030. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2628486.

[186] Pablo De Castro and Tommaso Dorigo. 'INFERNO: Inference-Aware Neural Optimisation'. In: (2018). arXiv: 1806.04743 [stat.ML].

[187] Simon N. Wood. 'Statistical inference for noisy nonlinear ecological dynamic systems'. In: *Nature* 466.7310 (2010), p. 1102.

[188] Kyle Cranmer, Juan Pavez and Gilles Louppe. 'Approximating likelihood ratios with calibrated discriminative classifiers'. In: *arXiv preprint arXiv:1506.02169* (2015).

[189] Claire Adam-Bourdarios et al. 'The Higgs boson machine learning challenge'. In: *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*. Ed. by Glen Cowan et al. Vol. 42. Proceedings of Machine Learning Research. Montreal, Canada: PMLR, 13 Dec 2015, pp. 19–55. URL: http://proceedings.mlr.press/v42/cowa14.html.

[190] D Basu. 'On partial sufficiency: A review'. In: *Selected Works of Debabrata Basu*. Springer, 2011, pp. 291–303.

[191] David A Sprott. 'Marginal and conditional sufficiency'. In: *Biometrika* 62.3 (1975), pp. 599–605.

[192] Dustin Tran et al. 'Edward: A library for probabilistic modeling, inference, and criticism'. In: *arXiv preprint arXiv:1610.09787* (2016).

[193] A Hocker et al. 'TMVA—Toolkit for Multivariate Data Analysis, in proceedings of 11th International Workshop on Advanced Computing and Analysis Techniques in Physics Research'. In: *Amsterdam, The Netherlands* (2007).

[194] Pierre Baldi, Peter Sadowski and Daniel Whiteson. 'Searching for exotic particles in high-energy physics with deep learning'. In: *Nature communications* 5 (2014), p. 4308.

[195] Radford M. Neal. 'Computing likelihood functions for High-energy Physics Experiments when Distributions are defined by Simulators with Nuisance Parameters'. In: *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*. CERN, 2007, pp. 111–118. URL: https://cds.cern.ch/record/1099977.

[196] Johann Brehmer et al. 'Mining gold from implicit models to improve likelihood-free inference'. In: (2018). arXiv: 1805.12244 [stat.ML].

[197] Johann Brehmer et al. 'Constraining Effective Field Theories with Machine Learning'. In: *arXiv preprint arXiv:1805.00013* (2018).

[198] Johann Brehmer et al. 'A Guide to Constraining Effective Field Theories with Machine Learning'. In: *arXiv preprint arXiv:1805.00020* (2018).

[199] Bai Jiang et al. 'Learning summary statistic for approximate Bayesian computation via deep neural network'. In: *arXiv preprint arXiv:1510.02175* (2015).

[200] Gilles Louppe, Michael Kagan and Kyle Cranmer. 'Learning to Pivot with Adversarial Networks'. In: *Advances in Neural Information Processing Systems*. 2017, pp. 982–991.

[201] Pablo de Castro. *Code and manuscript for the paper "INFERNO: Inference-Aware Neural Optimisation"*. https://github.com/pablodecm/paper-inferno. 2018.

[202] Joshua V Dillon et al. 'TensorFlow Distributions'. In: (2017). arXiv: 1711.10604 [cs.LG].

[203] Roger Barlow. 'Extended maximum likelihood'. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 297.3 (1990), pp. 496–506.