Università degli Studi di Padova

Dipartimento di Fisica e Astronomia “Galileo Galilei”
Area of Experimental Particle Physics

Master of Science in Physics

Final dissertation

Autoencoder-based characterization of QCD multijet background at the LHC

Author:
Javier Mariño Villadamigo

Supervisor:
Dr. Tommaso Dorigo
(Istituto Nazionale di Fisica Nucleare)

Co-supervisor:
Dr. John Alison
(Carnegie Mellon University)

Academic Year 2022-2023

UNIVERSITÀ DEGLI STUDI DI PADOVA

Abstract

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Area of Experimental Particle Physics

Master of Science in Physics

Autoencoder-based characterization of QCD multijet background at the LHC

by Javier MARIÑO VILLADAMIGO

A proof of principle for the application of autoencoders in encoding high-dimensional multijet data is presented. A simulation with events containing four b-quark QCD jets is used to train the autoencoder. The reconstruction of events after a reduced dimension step is attempted. We also demonstrate the capabilities of this autoencoder to generate new artificial events.

The characterization of QCD multijet events plays a crucial role in estimating background samples for processes involving $4b$-jets final states, including the production of Higgs boson pairs. Such a process is accessible at the LHC and may be observed in a combined search by the end of the high-luminosity run of the LHC. The $4b$ channel is expected to contribute significantly to this combined search, and it is the goal of this work to showcase the ability of autoencoders to model such multijet events in a reduced-dimension space. This approach has the potential to yield embedded metrics for effective background-signal discrimination. A deep learning architecture, largely inspired by an ongoing analysis within the CMS collaboration, is built as the autoencoder skeleton. After the description of the architecture, loss function, and training schedule, reconstruction of events is done successfully.

To quantify the accuracy of this reconstruction, we compute the Wasserstein distance between several kinematic variables of interest, along with a figure of merit to measure the similarity between the reconstructed dijet and quadjet invariant masses. Furthermore, our approach demonstrates the capability to generate an arbitrarily large number of events from the encoded space, showing promising agreement with the reconstructed samples. This study not only underscores the applicability of autoencoders in high-energy physics but also offers insights into their potential contributions to future experimental analyses within the field.

Acknowledgements

To my mother and sister, who paved the way and supported me through each and every obstacle. If I have made it this far is because you have had my back all along.

To my friends in Italy, who made me feel at home right from the beginning of this journey. A special thanks to you, Vanessa, for being there in difficult times.

To my friends back in Spain, your unwavering support transcends the kilometers that separate us.

Lastly, heartfelt gratitude to my brilliant supervisor, Tommaso Dorigo, for his invaluable knowledge and life advice. Working alongside you has been an exceptionally rewarding experience. I extend my sincere gratitude to John Alison for his co-supervision of this work, and to Patrick Bryant, whose assistance has been immensely helpful.

Contents

Abstract **iii**

Acknowledgements **v**

1 Introduction **1**

1.1 The Standard Model 4

1.2 Higgs boson pair production at the LHC 9

1.3 CMS results 11

1.4 ATLAS results 13

2 The LHC and the CMS detector **17**

2.1 The LHC complex 17

2.2 The CMS detector 18

3 Determination of the kinematics of the QCD background **31**

3.1 Hemisphere mixing 31

3.2 Kinematic reweighting 34

4 Machine learning: the autoencoder **37**

4.1 The autoencoder 38

4.2 Data preparation 39

4.3 Embedding and ghost batch normalization 41

4.4 Autoencoder architecture outline 44

4.5 Background sample, training, and loss 49

5 Results **51**

5.1 General results and Wasserstein distance 51

5.2 Decoded dataset 53

5.3 Generated dataset 57

5.4 Comprehensive results 63

6 Conclusions **65**

A Additional plots **69**

References **79**

List of Figures

1.1 Particle content of the Standard Model. 6
1.2 Complex scalar potential. 7
1.3 Higgs potential and its stability. 9
1.4 Higgs boson pair production at the LHC. 10
1.5 Higgs boson pair production cross section. 11
1.6 CMS limits on the Higgs boson pair production cross section. 12
1.7 CMS limits on the Higgs boson self-coupling and di-vector di-Higgs coupling. 13
1.8 ATLAS limits on the Higgs boson pair production cross section. 14
1.9 ATLAS limits on the Higgs boson self-coupling. 14

2.1 CERN accelerator complex. 18
2.2 CMS detector diagram. 19
2.3 CMS coordinate system. 19
2.4 CMS tracking system. 21
2.5 CMS electromagnetic calorimeter. 22
2.6 CMS muon system. 24
2.7 The anti-k_t algorithm. 28
2.8 Sketch of a b-jet. 29

3.1 Illustration of the hemisphere mixing procedure. 33
3.2 Percentage bias on the estimated signal fraction with the hemisphere mixing procedure. 34
3.3 Example of events and background estimation at the Higgs boson pair production search. 35
3.4 LO partial amplitudes contributing to multijet background in double Higgs search. 36

4.1 General scheme of an autoencoder. 39
4.2 High level sketch of the input data preparation. 40
4.3 Input embedding block. 42
4.4 Encoder architecture. 45
4.5 Decoder architecture. 47

5.1 Autoencoder training and validation losses. 52
5.2 Marginal one-dimensional distributions of the activations in the bottleneck. . 52

5.3 One-dimensional reconstructed distributions. 54
5.4 Two-dimensional $m_{2j}(m_{4j})$ reconstructed distributions. 56
5.5 Gaussian Mixture Models applied to the bottleneck activations. 58
5.6 Kernel Density Estimation applied to the bottleneck activations. 60
5.7 One-dimensional generated distributions. 61
5.8 Two-dimensional $m_{2j}(m_{4j})$ generated distributions. 62

A.1 One-dimensional reconstructed distributions of $\{p_T, \eta, \phi, m\}$ values. 69
A.2 One-dimensional generated distributions of $\{p_T, \eta, \phi, m\}$ values. 70
A.3 Two examples of η-ϕ and p_x-p_y reconstructed jets coordinates. 71
A.4 Activations marginal one-dimensional distributions for each of the three models offsets. 72
A.5 One-dimensional distributions of kinematic variables monitored during training, epoch 5. 73
A.6 One-dimensional distributions of kinematic variables monitored during training, epoch 20. 74
A.7 Distributions of residuals monitored during training, epoch 5. 75
A.8 Distributions of residuals monitored during training, epoch 20. 76
A.9 One-dimensional distributions of kinematic variables monitored during training **with *symmetrization* applied**, epoch 20. 77
A.10 Distributions of residuals monitored during training, epoch 20. 78

List of Tables

1.1 Higgs boson pair branching fractions. 11

5.1 Comprehensive results for the similarity between distributions. 63

Chapter 1

Introduction

Our current understanding of physics at the subnuclear level is essentially described by the Standard Model (SM) of particle physics. It is, as of today, the most successful (and experimentally consistent) theory able to describe particle properties, interactions, and high-level observables in experiments. It is a theoretical construction based on the $SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$ gauge group, describing quantum chromodynamics (QCD), the chiral $SU(2)_L$ electroweak (EW) sector and the hypercharge $U(1)_Y$ sector in which quantum electrodynamics (QED) is embedded. The SM describes three of the four types of forces that are known in nature: strong, weak, and electromagnetic interactions. In the case of non-abelian groups, such as $SU(3)_c$ and $SU(2)_L$, also the self-interactions that can take place between force carriers of the same type are described. The messengers of these interactions, known as gauge bosons, have been identified in the very recent history of particle physics, primarily through discoveries made using particle accelerators. Among these bosons, the photon was the first one to be postulated, though it is unclear whether we can assign a concrete date to its first inception. Albert Einstein talked, as early as 1905, about "discrete packages" of energy in the transmission of light during the photoelectric effect. On the other hand, the charged weak boson $W^{\pm}$ was discovered in January of 1983 by the UA1 and the UA2 experiments [1, 2]. The observation of its neutral counterpart, the Z boson, followed a few months later, in May 1983. These discoveries promptly led to the award of the Nobel Prize in physics to Carlo Rubbia and Simon van der Meer in the same year.

The SM, first formulated in the 1960s, also includes a scalar field that enables the breaking of the EW symmetry, which in turn allows for the existence of non-zero mass terms for a triplet of weak bosons in the Lagrangian density [3]. The Higgs boson, which is a scalar particle, is the remaining Goldstone boson from such field after the EW symmetry breaking. The search for this particle, whose mass is a free parameter of the SM, lasted for decades: over twenty years passed since its formulation before it was possible to narrow the investigation. This was achieved with the Large Electron Positron (LEP) collider [4] at CERN and later with the Tevatron [5] at Fermilab. Even though the Higgs boson was not discovered at the LEP, which was running at a $\sqrt{s}_{\text{max}} \simeq 210\,\text{GeV}$, the mere missed discovery implied a lower bound on the Higgs boson mass of $M_H \geq 114\,\text{GeV}$. The Large Hadron Collider (LHC) [6] was therefore devised, in good part, as a Higgs discovery machine, working at

$\sqrt{s}$ = 7-8 TeV in its first run. In 2010, when the LHC started taking data, the achieved energy of the collisions was finally large enough to probe the existence of the Higgs boson.

Thus, in July 2012, CMS [7] and ATLAS [8] experiments reported the observation of a particle of 125 GeV of mass [9, 10]. Subsequent results from both experiments, summarized in refs. [11–15], established that all the properties of the discovered particle, including its spin, CP properties, and coupling strengths to the SM particles, were indeed consistent within the uncertainties with the SM Higgs boson. After the discovery of the Higgs boson, a new era in understanding the EW symmetry breaking, completing the SM, and setting constraints on New Physics (NP) phenomena has opened. Regarding the Higgs boson, specifically, one of the most pressing matters is the precise measurement of its self-couplings. The self-couplings of the Higgs boson determine the shape of the EW potential and dictate the absolute stability of the EW vacuum, which is connected to the phase transition of the early universe from the unbroken to the broken EW symmetry. These self-couplings are also very loosely constrained from EW precision measurements, and can therefore be highly sensitive to NP phenomena and Beyond Standard Model (BSM) effects [16, 17].

The Higgs boson trilinear coupling can be probed essentially in two manners: with the so-called *direct* method, which involves measuring the direct effects the coupling has on the production of multi-Higgs boson final states via an intermediate Higgs boson; and the *indirect* method, based on measuring the loop corrections the coupling induces on the production of single-Higgs boson final states. The quartic self-coupling, being further suppressed with respect to the trilinear coupling, is not (and will not be) accessible at the LHC. This work will focus solely on the *direct* process, which has the perk of being theoretically "cleaner" with respect to the *indirect* method (where the higher order effects of the trilinear coupling are harder to disentangle) and the disadvantage of having a lower cross section. The *direct* mechanism leads, as mentioned earlier, to multi-Higgs boson final states, most commonly di-Higgs final states. The searches for these processes are enormously challenging from the perspective of the substantial background that arises with the same final states. It is therefore of the utmost importance to characterize these background distributions, mainly arising from QCD processes, as precisely as possible, in order to perform an accurate signal versus background discrimination.

This is where deep learning and neural networks come into play. Machine-learning tools have seen extraordinary growth in recent years, opening up more and more possibilities in high-energy physics, including tasks that were previously considered impractical. In particular, autoencoders, which take advantage of an encoding-decoding structure, are being used for projects that include anomaly detection, dimensionality reduction, and metrics studies. One of their main advantages lies in the inclusion of a reduced-dimension space between the encoding and decoding halves. As a result, when the autoencoder is forced to reproduce the same objects it receives as input in the output, it learns to capture the distinctive features of the input objects and embed them in the encoded space. This work is centered on the use of such a tool to achieve a faithful reconstruction of the QCD background distributions present in searches for Higgs boson pair production. In particular, it exploits the use of an autoencoder to reproduce the main features of 4 b-jets background, aiming to reconstruct in output the principal kinematic distributions of the jets given in input. This is particularly meaningful, as will be described in the corresponding section, when the embedded space dimension is highly reduced, allowing for the efficient compression of the relevant information in the phase space of the input jets. Some studies have been carried out thus far exploring the metrics between these events [18, 19], and some of them have even been applied to the background of double Higgs boson production [20]. In this work, one of these metrics, the Wasserstein distance [21, 22], will be used to assess the similarity between the true and the reconstructed background samples, which is an indirect measurement of the quality of the encoded space distribution.

In Section 1.1, a brief introduction to the Standard Model of particle physics is presented, focusing on the spontaneous symmetry breaking in Section 1.1.1, and on the importance of the measurement of the Higgs boson self-coupling in Section 1.1.2.

In Section 1.2, the Higgs boson pair production at the LHC is addressed, presenting current estimates for the most contributing process to the cross section. Sections 1.3 and 1.4 show the latest results obtained for the measurement of the cross section for Higgs boson pair production at CMS and ATLAS experiments, respectively.

In Chapter 2, a brief description of the Large Hadron Collider complex is presented, along with a detailed explanation of the structure and functioning of the CMS detector.

In Chapter 3, we outline the challenging difficulties in estimating the QCD background dominant in the Higgs boson pair production searches, and we describe two state-of-the-art methods to do so in Sections 3.1 and 3.2.

Chapter 4 serves as a description of the material and methods used for this work, with a brief introduction to basic elements of machine learning and the general autoencoder skeleton in Section 4.1, data preparation in Section 4.2, working principles and the employed architecture in Sections 4.3 and 4.4. Lastly, Section 4.5 is dedicated to the training dataset, loss, and training setup.

Results are thoroughly described in Chapter 5. We provide some general aspects of those in Section 5.1, while the results for decoded and generated datasets are presented separately (although analyzed similarly) in Sections 5.2 and 5.3, respectively.

Lastly, conclusions are drawn in Chapter 6, where we also outline the future prospects that this proof of concept is likely to undergo.

1.1 The Standard Model

The Standard Model is arguably the only quantum field theory that is able to describe the observable universe, that is, the particles we can observe and their interactions. It is, however, not complete, since it does not include a description of one of the four elementary forces known in nature: gravity. The Standard Model fails to describe other aspects of nature: it cannot provide an explanation for the hierarchy of fermion masses, for the non-zero mass of the neutrinos, or for the matter-antimatter asymmetry in the universe. This should not prevent us from using it to make predictions, since as an effective field theory, it predicts with astonishing precision a large percentage of experimental results. Nonetheless, it is worth mentioning that the SM could fail at large energy scales, i.e. $\sim \mathcal{O}(\text{TeV})$, where heavier degrees of freedom could come into play and become an essential ingredient in the description of the subnuclear processes. Furthermore, there is an expectation (or hope) that these hypothetical degrees of freedom could aid in explaining the phenomena that the SM currently cannot account for [23].

The particle content of the SM is formed by 12 fermions, 4 vector gauge bosons and 1 scalar Higgs boson. Fermions differ from bosons in their spin numbers, a quantum property that is conserved under rotations of a given symmetry group. In particular, fermions are characterized by having half-integer spin numbers, whereas bosons possess integer spin numbers. Fermions can be further divided in:

- Leptons: which interact through weak forces and, in case they are charged under $U(1)_{\text{em}}$, also through the electromagnetic force. Charged leptons include the *electron*, *muon* and *tau*, each with a charge of -1 in electron charge units. Neutral leptons are called *neutrinos* and each one of the charged leptons has its neutrino counterpart.
- Quarks: they differ from leptons in that they interact not only through the weak and electromagnetic forces but also through the strong force. They thus possess an additional quantum number with respect to leptons, known as the color charge: a conserved quantity under $SU(3)_c$ transformations. They can, just like leptons, be divided into 3 generations, approximately according to their mass scale: *up* and *down* quarks belong to the first generation, *charm* and *strange* quarks to the second, and *top* and *bottom* quarks to the third.

Fermions can be grouped in $SU(2)$ doublets, according to their generation or, equivalently, their mass scale:

$$\underbrace{\begin{pmatrix} e \\ \nu_e \end{pmatrix}, \begin{pmatrix} \mu \\ \nu_\mu \end{pmatrix}, \begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix};}_{\text{Leptons}} \quad \underbrace{\begin{pmatrix} u \\ d \end{pmatrix}, \begin{pmatrix} c \\ s \end{pmatrix}, \begin{pmatrix} t \\ b \end{pmatrix}}_{\text{Quarks}}. \tag{1.1}$$

The main difference between the strong and the weak/electromagnetic forces is their behavior concerning the energy scale (or distance) of the interaction. While the weak and electromagnetic forces become less intense when the particles are further apart, the strong force behaves in the opposite way. This has profound implications for the way strong interactions occur: when a quark is produced, for example, in a particle accelerator, the interaction with other quarks quickly becomes non-perturbative ($\alpha_s \gg 1$) as the distance between them is large enough. This energy gain results in the creation of quark-antiquark pairs, which can form bound states or undergo successive creation of more of such pairs. Such a process is akin to an avalanche, building up until it eventually creates a cone (conventionally called *jet*) of particles originating from a single quark. A more technical discussion about jets will be developed in Sections 2.2.7 and 2.2.8.

Regarding bosons, they can also be divided in two categories:

- Gauge bosons: consisting of spin-1 force mediators. The gluon and the photon are massless gauge bosons that mediate the strong and electromagnetic forces, respectively. The $W^{\pm}$ and Z bosons are massive, with a mass of around 80 GeV and 91 GeV, respectively. They are responsible for mediating the weak force.
- Scalar (Higgs) boson: the Higgs boson, on the other hand, is the only scalar fundamental particle of the SM, or at least that is what the experimental results show (see for example [14]). Unlike gauge bosons, the Higgs boson does not transmit electroweak or strong forces; it is the remnant of a scalar field that confers mass to other particles.

It is also important to note that particles that are electromagnetically charged possess an anti-particle counterpart, with opposing charge. In the case of neutral bosons, like Z, the photon or the Higgs boson, they can be thought of as their own antiparticle. In the case of neutrinos, though, the discussion is more complex, since neutrinos and antineutrinos present manifest differences in their interactions and decay modes.

The particle content of the SM and the principle particle properties are detailed in Figure 1.1.

1.1.1 Spontaneous Symmetry Breaking (SSB): the Higgs mechanism

Before considering the particular case of the EW symmetry breaking, let us first introduce the spontaneous breaking of a global symmetry. The Lagrangian density for a complex scalar $\phi(x)$ can be written as

$$\mathcal{L} = T - V = \frac{1}{2}(\partial_\mu \phi)^2 - \mu^2 \phi^\dagger \phi - \lambda(\phi^\dagger \phi)^2, \tag{1.2}$$

where $\lambda > 0$ and $\mathcal{L}$ is invariant under $U(1) : \phi \to \phi' = e^{i\alpha}\phi$ transformations. The potential can acquire two distinctive shapes, as shown in Figure 1.2, that have different physical meanings.

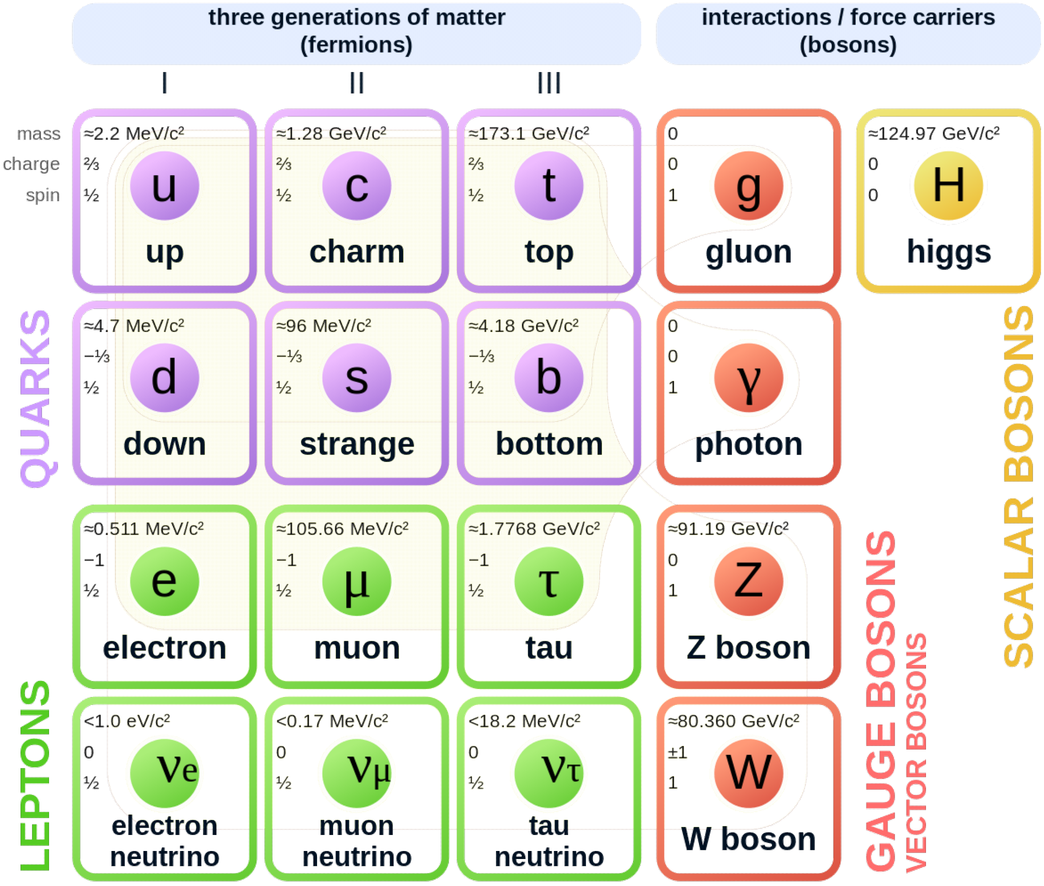

Figure 1.1: From [24]. The particle content of the Standard Model of particle physics.

1. if $\mu^2 > 0$, the potential presents only the trivial minimum, corresponding to $\phi = 0$, describing a scalar field with mass μ and quartic coupling λ (Figure 1.2 left).
2. if $\mu^2 < 0$, the potential presents a degenerate ring of minima, defined by the condition

$$|\phi_v| = \sqrt{\frac{-\mu^2}{2\lambda}} \equiv \frac{v}{\sqrt{2}}, \tag{1.3}$$

   where ϕ_v is the vacuum expectation value (v.e.v.) of the ϕ field (Figure 1.2 right).

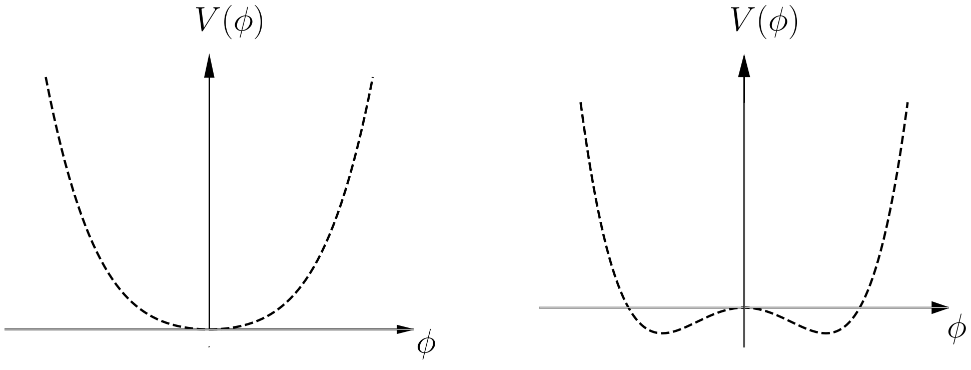

Figure 1.2: Scalar potential from Equation 1.2 for the cases where $\mu^2 > 0$ (left) and $\mu^2 < 0$ (right).

It is now clear that, for a specific ground state, the original $U(1)$ symmetry gets spontaneously broken. This is apparent when the scalar field is parameterized as:

$$\phi(x) = v + \frac{1}{\sqrt{2}}\left[\phi_1(x) + i\phi_2(x)\right]. \tag{1.4}$$

The potential then takes the form:

$$V(\phi) = V(\phi_v) - \mu^2\phi_1^2 + \lambda v\phi_1(\phi_1^2 + \phi_2^2) + \frac{\lambda}{4}(\phi_1^2 + \phi_2^2). \tag{1.5}$$

The real part of the field, ϕ_1, describes thus a scalar particle with mass $m_{\phi_1} = \sqrt{2\lambda v^2} = -2\mu^2$, whereas ϕ_2 describes a massless state. This is nothing but a consequence of the Goldstone theorem [25], which states that in a Lagrangian that is invariant under a group of symmetry $\mathcal{G}$, where $\dim(\mathcal{G}) = N$, if $M < N$ generators of the group are spontaneously broken (i.e. their currents are conserved, but the ground state is not invariant under the action of the corresponding charges), then there will be M (one for each broken generator) massless particles, named Nambu-Goldstone bosons, that do not preserve the ground state.

The Brout-Englert-Higgs (BEH) mechanism [3, 26–30] is the particular case of the spontaneous breaking of the EW symmetry $SU(2)_L \otimes U(1)_Y$. In this scenario, the Goldstone bosons resulting from the symmetry breaking would be the longitudinal polarizations of $W^{\pm}$ and Z bosons which, however, are not directly observable. Since this symmetry is gauged, the three would-be Goldstone bosons are absorbed by the three gauge bosons. This gives them a mass and the associated polarization third degree of freedom.

To make a similar discussion for the case of the SM, a $SU(2)$ doublet of complex scalar fields should be introduced:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{1.6}$$

The potential can be instead written as

$$V(\phi) = \mu^2(\phi^\dagger\phi) + \lambda(\phi^\dagger\phi)^2, \tag{1.7}$$

where λ is assumed to be positive (otherwise the potential would be unbounded from below). In the unitary gauge, the scalar potential ground state can be chosen to be:

$$\phi_1 = \phi_2 = \phi_4 = 0, \quad \phi_3 = \sqrt{\frac{-\mu^2}{\lambda}} = v \tag{1.8}$$

The ϕ field can consequently be expanded around the vacuum as a function of a perturbation $H(x)$:

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}. \tag{1.9}$$

The perturbation $H(x)$ is in reality the physical Higgs scalar field. We now have all the ingredients to write the scalar Lagrangian including the physical gauge fields:

$$\begin{aligned}\mathcal{L}_{\text{Higgs}} =& \frac{1}{2}\partial_\mu H \partial^\mu H + \mu^2 H^2 + \frac{g^2}{4}(v+H)^2 \left(W_\mu^+ W^{\mu-} + \frac{1}{2\cos^2\theta_W} Z_\mu Z^\mu \right) - \\ & - \lambda v H^3 - \frac{\lambda}{4} H^4,\end{aligned} \tag{1.10}$$

where θ_W is the mixing angle (see for example the introduction of [31] and [32]). The first two terms of the Lagrangian represent the kinetic and the mass terms of the Higgs scalar field, respectively; whereas the next term is composed of different summands dependent on the W and Z gauge fields: their mass terms and the summands describing the vertices of the interaction with the Higgs boson. The last two terms, on the other hand, contain the trilinear and the quartic couplings of the Higgs boson. Some physical quantities that can be derived from the previous expression are:

$$M_H^2 = 2\lambda v^2, \quad M_W^2 = \frac{g^2 v^2}{4}, \quad M_Z^2 = \frac{g^2 v^2}{4\cos^2\theta_W}, \quad v = \sqrt{\frac{1}{\sqrt{2}G_F}} \simeq 246 \text{ GeV}; \tag{1.11}$$

where M_H, M_W, M_Z are the masses of the Higgs and weak gauge bosons, respectively; and v is the v.e.v. of the Higgs scalar field.

1.1.2 The Higgs boson self-coupling

The last two terms in Equation 1.10 describe the Higgs self-interactions. However, if we consider the SM an effective theory, λ stands for two otherwise free parameters, namely the trilinear (λ_{HHH}) and the quartic (λ_{HHHH}) couplings, respectively [33, 34]:

$$\lambda_{HHH} \text{ (or } \lambda_3) = \frac{3M_H^2}{v}, \qquad \lambda_{HHHH} \text{ (or } \lambda_4) = \frac{3M_H^2}{v^2}. \tag{1.12}$$

The Higgs self-coupling determines, as can be deduced from Equation 1.10, the shape of the Higgs potential for large values of the field i.e. $V(H) \sim \lambda H^4/4$. It is therefore unclear whether the EW vacuum that has been measured up to this day is the true vacuum of the field, or if otherwise there could be a more stable vacuum, to which the field could eventually quantum-tunnel within a finite time scale (see Figure 1.3).

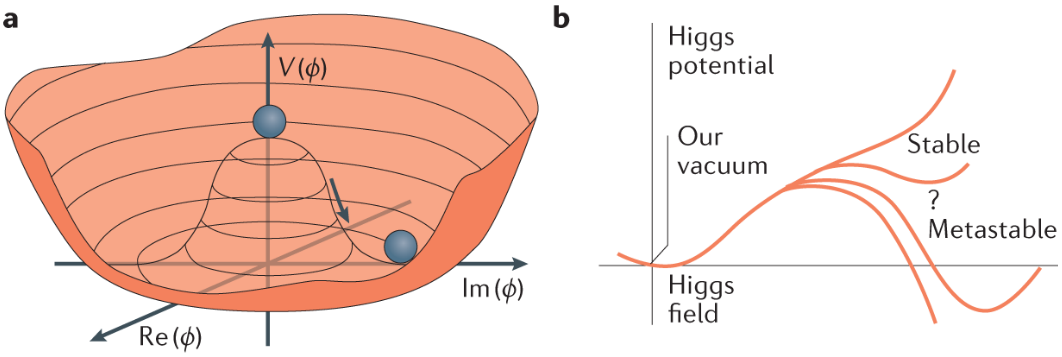

Figure 1.3: From [35]. Left: The Higgs potential for $\mu^2 < 0$. Choosing any of the points at the bottom of the potential spontaneously breaks the rotational $U(1)$ symmetry. Right: Quantum corrections can change the shape of the Higgs potential and its stability.

Large deviations of the trilinear and quartic couplings from the nominal values are possible in BSM scenarios. As an example, in the Two-Higgs-Doublets Model (THDM), deviations of the trilinear coupling can be of the order of 100%, even when all the couplings of the lightest Higgs boson with gauge bosons and fermions are consistent with SM values. At linear colliders, such a large difference in the HHH coupling may be detected [36]. Anomalous Higgs boson self-couplings also appear in other BSM scenarios, like theories with a composite Higgs (see for example [37] and references therein) or Little Higgs models [38]. As previously stated, the trilinear Higgs boson self-coupling can be probed directly in searches for multi-Higgs final states and indirectly via its effects on precision observables or loop corrections to single-Higgs production. In contrast, the quartic coupling is not accessible at the LHC [33].

1.2 Higgs boson pair production at the LHC

The direct manner to probe the Higgs boson trilinear coupling is by producing pairs of Higgs bosons. At hadron colliders, Higgs pairs are dominantly produced via gluon-gluon fusion (ggF), vector boson fusion (VBF) associated production of Higgs boson pairs with a vector boson (VHH), and associated production of top quark pairs with Higgs boson pairs ($t\bar{t}HH$). The dominant mode at the LHC is ggF, which for $\sqrt{s} = 13$ TeV, $M_H = 125$ GeV and $M_t = 173$ GeV, yields the current estimation for the cross section [39]:

$$\sigma^{\text{ggF}}_{pp\to HH} = 31.05 \pm 3.0\%(\text{PDF} + \alpha_S)\ ^{+6\%}_{-23\%}(\text{scale} + M_t \text{ unc.})\ \text{fb}, \tag{1.13}$$

which is almost 20 times larger than the second most dominant mode, vector boson fusion, and a factor 1000 times smaller than the single Higgs production cross section [39]. The uncertainties on Equation 1.13 make reference, in the case of "PDF + α_S", to the uncertainties of the PDFs combined with that of α_S computation; and "scale + M_t unc." refers to the uncertainty in QCD renormalization and factorization scale combined with the uncertainty arising from missing finite top quark mass effects.

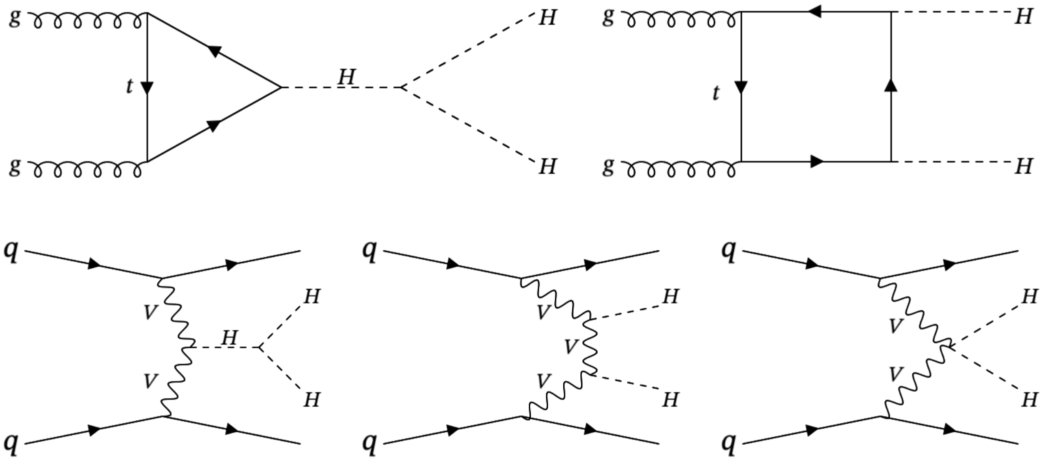

Figure 1.4: Main processes contributing the Higgs boson pair production at the LHC. Top row: ggF processes. Bottom row: VBF processes. On both rows, the left-most diagram indicates the process through which the trilinear coupling contributes.

In Figure 1.4, one can find the diagrams contributing to the ggF production mode (top row) and VBF production mode (bottom row), where only the left-most diagram in both rows is dependent on the trilinear self-coupling. While searches in the ggF production mode are more sensitive to deviations in the Higgs self-interactions, the VBF production mode is particularly sensitive to c_{2V}, i.e. the coupling between two Higgs bosons and two vector bosons ($HHVV$). Figure 1.5 shows the current total cross sections for Higgs pair production at a proton-proton collider inluding higher-order corrections.

Final states for the study of the Higgs boson pair production (like most other processes) are typically chosen as a balance between a sufficiently large branching fraction and a sufficiently background-clean final state [43]. As a result, one Higgs boson is typically required to decay to a pair of b quarks, while the other can be observed in a leptonic final state i.e. decaying to a pair of τ leptons ($b\bar{b}\tau\bar{\tau}$), two photons ($b\bar{b}\gamma\gamma$), etc. The second Higgs boson can also be required to decay to an additional pair of b quarks ($b\bar{b}b\bar{b}$), taking advantage of

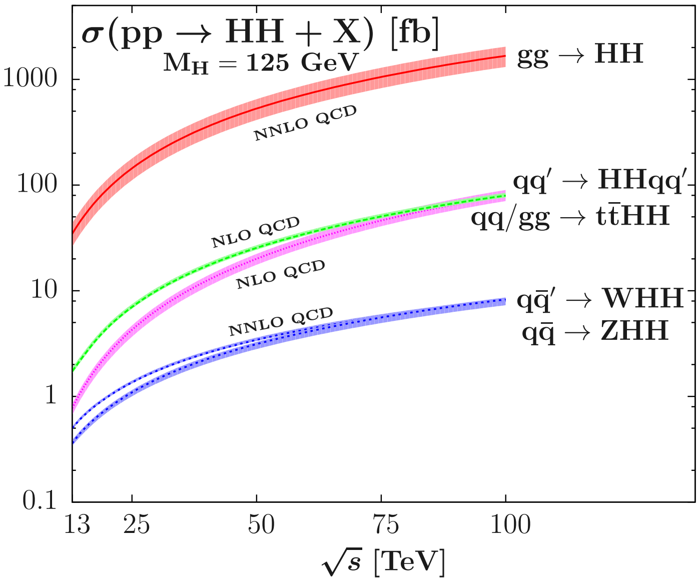

Figure 1.5: From [40]. The total cross sections for Higgs boson pair production at a proton-proton collider, including higher-order corrections discussed the indicated corrections, in the main production channels as a function of the center-of-mass energy with $M_H = 125$ GeV. The MSTW2008 [41, 42] PDF set has been used and theoretical uncertainties are included as corresponding bands around the central values.

its larger branching fraction. Six of the largest branching fractions for the Higgs boson pair system decay are shown in Table 1.1.

Decay channel	Branching fraction
$b\bar{b}b\bar{b}$	$3.37 \cdot 10^{-1}$
$b\bar{b}W^+W^-$	$2.50 \cdot 10^{-1}$
$b\bar{b}\tau^+\tau^-$	$7.27 \cdot 10^{-2}$
$b\bar{b}\gamma\gamma$	$2.64 \cdot 10^{-3}$
$W^+W^-\gamma\gamma$	$9.77 \cdot 10^{-4}$
$W^+W^-W^+W^-$	$4.63 \cdot 10^{-2}$

Table 1.1: From [34]. HH branching fractions for a Higgs boson of mass $M_H = 125.09$ GeV.

1.3 CMS results

The latest results of the CMS collaboration on the HH production cross section, as well as the Higgs boson self-coupling, were produced in 2023, encompassing measurements in different decay channels and production modes [44]. Comprehensively, results shown in Figure 1.6 use datasets corresponding to an integrated luminosity ($\mathcal{L}$) up to 138 fb^{-1}, collected by CMS in 2016-2018 at a center of mass energy of $\sqrt{s} = 13$ TeV.

The cross section for Higgs boson pair production is extremely small, thus escaping detection at the LHC so far. The results are therefore expressed as an upper limit on the production cross section. Figure 1.6 (left) shows the expected and observed limits on Higgs boson pair production, expressed as ratios with respect to the SM expectation, in searches using different final states and their combination. The current upper value for the HH production cross section corresponds to 3.4 times the SM expected value at 95% CL. Figure 1.6 (right) shows the evolution of the limits for the cross section for the three most sensitive channels, and also a projection of the expected upper value after the High-Luminosity run of the LHC (HL-LHC), which is expected to collect data up to 3000 fb^{-1} of integrated luminosity [45].

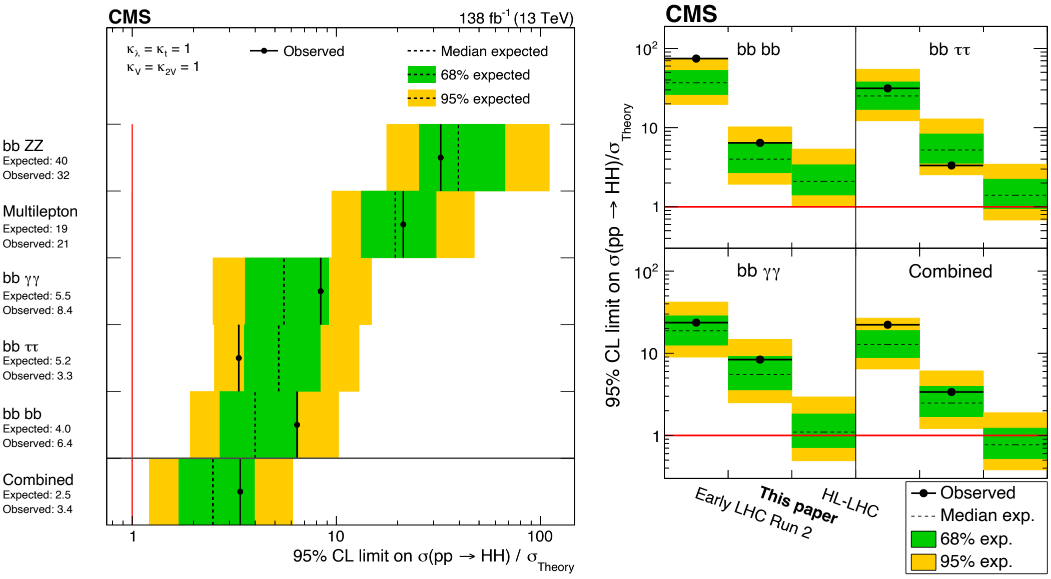

Figure 1.6: From [44]. Limits on the production of Higgs boson pairs and their time evolution. Left: expected and observed limits on the ratio of experimentally estimated production cross section and the expectation from the SM (σ_{Theory}) in searches using different final states and their combination. The search modes are ordered, from upper to lower, by their expected sensitivities from the least to the most sensitive. The overall combination of all searches is shown by the lowest entry. Right: expected and observed limits on HH production in different datasets: early LHC Run 2 data (35.9 fb^{-1}), present results using full LHC Run 2 data (138 fb^{-1}), and projections for the HL-LHC (3000 fb^{-1}).

Figure 1.7 displays the expected and observed upper limits on the value for the Higgs boson self-coupling modifier $\kappa_\lambda \equiv \lambda/\lambda_{\text{SM}}$ (left) and for the coupling between two vector bosons and two Higgs bosons ($VVHH$) $\kappa_{2V} \equiv g^{2V}/g^{2V}_{\text{SM}}$. The red lines illustrate the predicted cross sections as a function of the coupling modifiers, which exhibit a characteristic dip in the vicinity of the SM value ($\kappa = 1$) due to the destructive interference of contributing production amplitudes. The experimental limits on the Higgs boson pair production cross section (black lines) also exhibit a strong dependence on the coupling modifiers, due to not only the changes in the cross sections that the couplings introduce, but also to the effects this has on the efficiency for detecting such signal events, characterized by different kinematic properties.

With the current dataset, the values for the coupling modifiers at a 95% CL are

$$-1.24 < \kappa_\lambda < 6.49, \qquad 0.67 < \kappa_{2V} < 1.38. \tag{1.14}$$

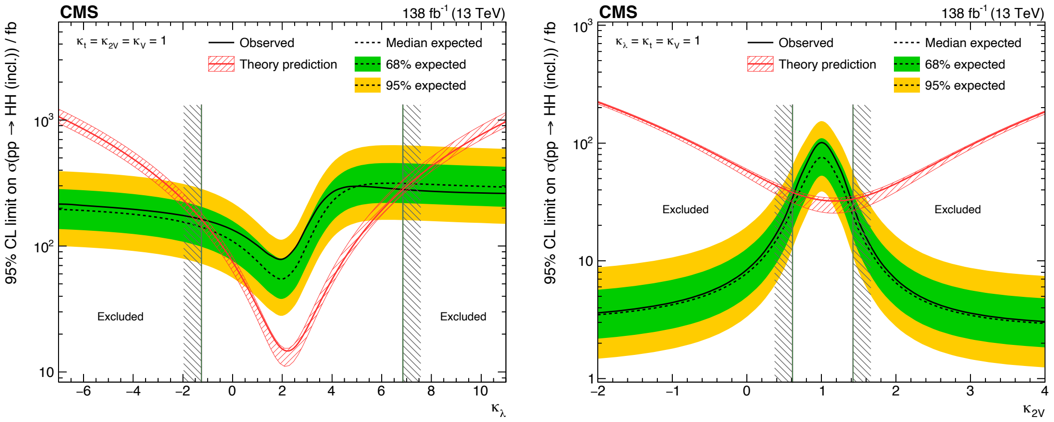

Figure 1.7: From [44]. Limits on the Higgs boson self-interaction (left) and coupling between two Higgs bosons and two vector bosons (right). Combined expected and observed 95% CL upper limits on the HH production cross section for different values of κ_λ (left) and κ_{2V} (right), assuming the SM values for the modifiers of Higgs boson couplings to top quarks and vector bosons. The green and yellow bands represent, respectively, the 1 and 2 s.d. extensions beyond the expected limit; the red solid line (band) shows the theoretical prediction for the HH production cross section (its 1 s.d. uncertainty). The areas to the left and to the right of the hatched regions are excluded at 95% CL.

It is interesting to note that $\kappa_{2V} = 0$ is excluded with a significance of 6.6 s.d., establishing the existence of such coupling, as depicted in Figure 1.4 (bottom row, right).

1.4 ATLAS results

The most recent results for HH production from ATLAS collaboration, however, still only include a fraction of the integrated luminosity of Run 2, which are reported in [46]. The results are extracted from data consisting of 36.1 fb^{-1} of integrated luminosity of proton-proton collisions at a center of mass energy $\sqrt{s} = 13$ TeV. The combination is done using six analyses searching for Higgs boson pairs decaying into $b\bar{b}b\bar{b}$, $b\bar{b}W^+W^-$, $b\bar{b}\tau^+\tau^-$, $W^+W^-W^+W^-$, $b\bar{b}\gamma\gamma$ and $W^+W^-\gamma\gamma$ final states. Although the article provides results for the resonant Higgs boson pair production (with no discovery reported), these are beyond the objective of the present paper, and we will only focus on the non-resonant analyses.

The upper limits at 95% CL on the cross section of the ggF Higgs boson pair production normalized to the SM value are shown in Figure 1.8 for the mentioned final states and their combination. The combined observed (expected) upper limit on the SM HH production cross section is 6.9 (10) times the SM predicted value.

Regarding the Higgs boson self-coupling modifier, on the other hand, expected (black dashed line) and observed (black solid line) upper limits are shown in Figure 1.9. Also the individual observed limits for the three most sensitive channels are shown as solid lines. In the case of the expected combined upper limit, the ±1 and ±2 s.d. bands are also shown as shaded green and yellow regions, respectively.

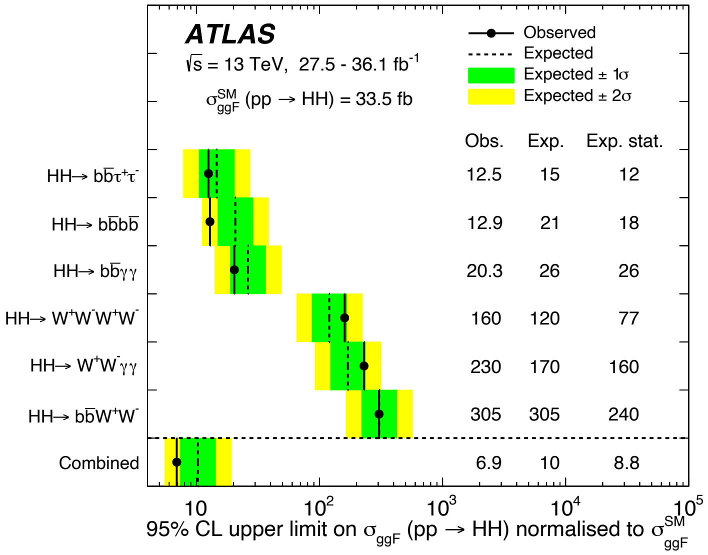

Figure 1.8: From [46]. Upper limits at 95% CL on the cross section of the ggF SM HH production normalized to its SM expectation $\sigma^{\text{SM}}_{\text{ggF}}(pp \to HH)$ from the $b\bar{b}\tau^+\tau^-$, $b\bar{b}b\bar{b}$, $b\bar{b}\gamma\gamma$, $W^+W^-W^+W^-$, $W^+W^-\gamma\gamma$ and $b\bar{b}W^+W^-$ searches, and their statistical combination. The column "Obs." lists the observed limits, "Exp." the expected limits with all statistical and systematic uncertainties, and "Exp. stat." the expected limits obtained including only statistical uncertainties in the fit.

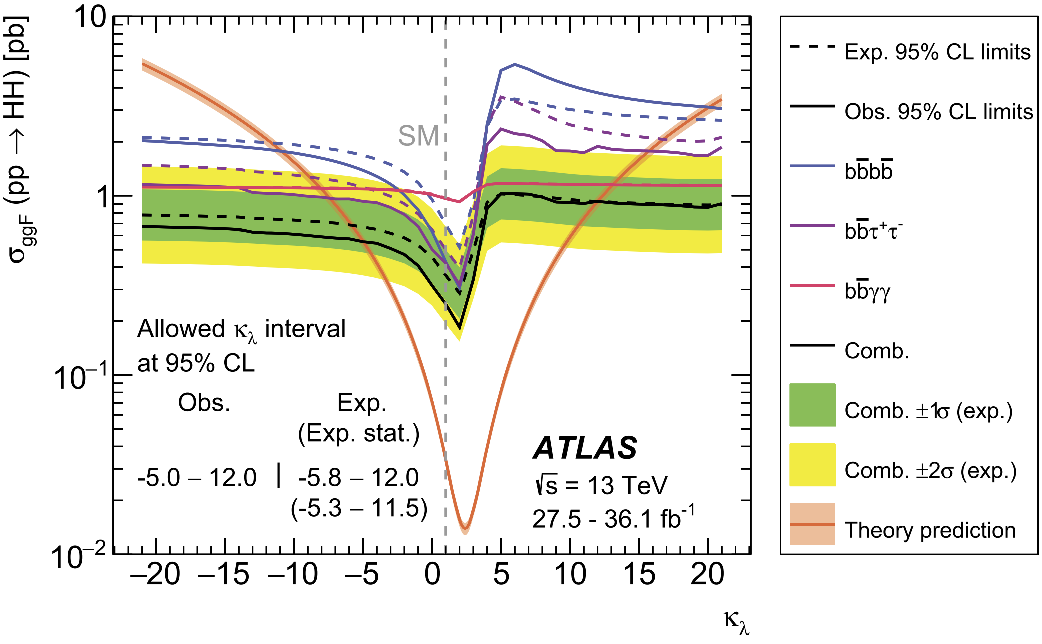

Figure 1.9: From [46]. Upper limits at 95% CL on the cross section of the ggF non-resonant SM HH production as a function of κ_λ. The observed (expected) limits are shown as solid (dashed) lines. In the $b\bar{b}\gamma\gamma$ final state, the observed and expected limits coincide. The ±1 and ±2 s.d. bands are only shown for the combined expected limit. The theoretical prediction of the cross section as a function of κ_λ is also shown. The effect of non-SM Higgs decay branching fractions due to κ_λ variations is not taken into account, which impacts the κ_λ intervals by no more than 7%.

The resulting observed (expected) confidence interval at 95% CL for κ_λ is:

$$-5.0\ (-5.8) < \kappa_\lambda < 12.0\ (12.0). \tag{1.15}$$

Chapter 2

The LHC and the CMS detector

This chapter provides a concise overview of the Large Hadron Collider and the Compact Muon Solenoid detector. In particular, Section 2.1 offers a general description of the LHC, focusing on the primary experiments it hosts, with specific emphasis on the CMS experiment in Section 2.2. The CMS detector's structure is described from the inside out, starting with the depiction of the inner tracking system in Section 2.2.1, followed by a comprehensive account of the calorimetry structure in Sections 2.2.2 and 2.2.3. Additionally, a brief overview of the muon system is provided in Section 2.2.4. To manage the remarkably high collision rate, the CMS detector is equipped with an advanced trigger system, detailed in Section 2.2.5. The event reconstruction algorithm used to obtain physics objects is explained in Section 2.2.6. Lastly, Section 2.2.7 discusses the reconstruction of jets in the CMS detector, dedicating Section 2.2.8 to the identification of jets originating from b quarks.

2.1 The LHC complex

The Large Hadron Collider (LHC) [47] is a proton-proton and heavy-ion collider operating at CERN since 2009. It is situated in the same ring tunnel that previously hosted the Large Electron-Positron collider (LEP) [4] from August 1989 to November 2000. This collider is designed with two accelerating rings, each featuring superconducting magnets.

The LHC injection chain is composed of several accelerators [48]. It all begins at Linac4 (a small linear accelerator), where negative hydrogen ions H^- are accelerated up to 160 MeV. Subsequently, the ions are then stripped of their two electrons during injection from Linac4 into the Proton Synchrotron Booster (PSB), leaving only protons. These are accelerated to 2 GeV before being injected into the Proton Synchrotron (PS), which further boosts the beam to 26 GeV. Protons are then sent to the Super Proton Synchrotron (SPS), where they are accelerated up to 450 GeV.

Lastly, protons will be transferred to the two beam pipes of the LHC, where one of the beams will circulate clockwise while the other will do it anti-clockwise. After around 20 minutes, both beams will attain their final collision energy of 6.5 TeV, culminating in a total center of mass collision energy of 13 TeV. These beams will intersect and collide inside four detectors

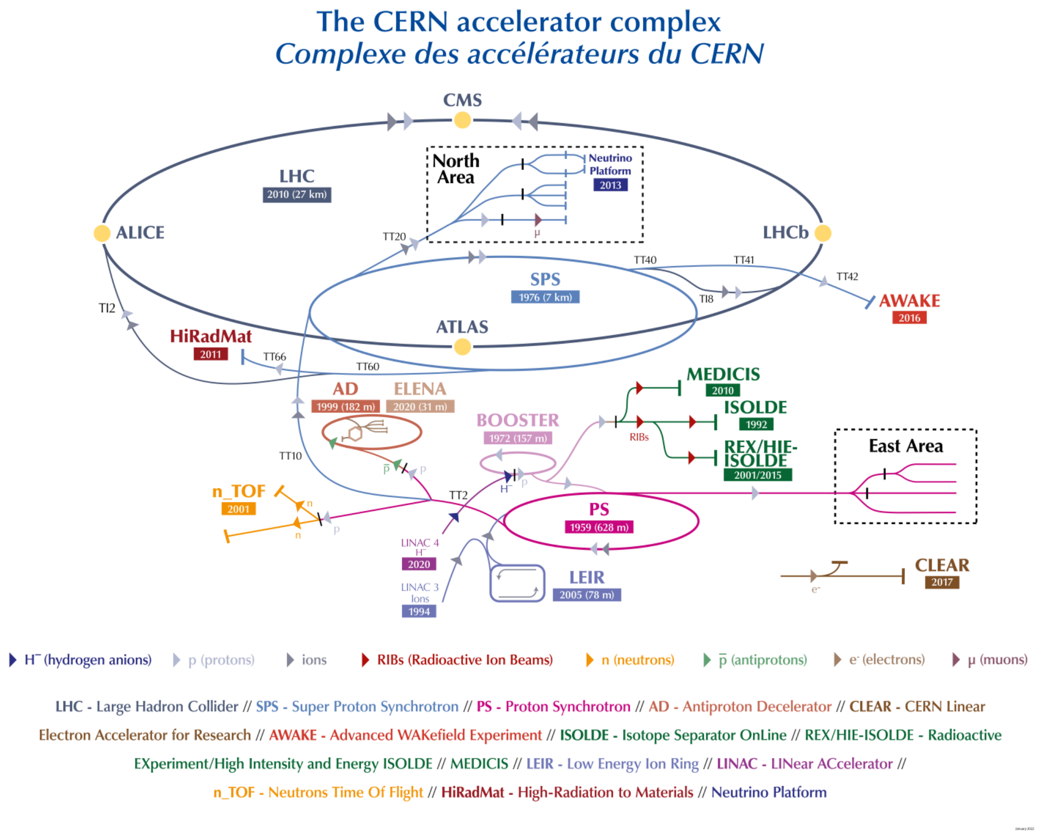
The CERN accelerator complex
Complexe des accélérateurs du CERN
CMS
LHC
2010 (27 km)
North Area
Neutrino Platform
2013
ALICE
LHCb
TT20
TT40
TT41
TT42
AWAKE
2016
SPS
1976 (7 km)
TI8
TI2
ATLAS
HiRadMat
2011
TT66
TT60
AD
1999 (182 m)
ELENA
2020 (31 m)
MEDICIS
2010
ISOLDE
1992
BOOSTER
1972 (157 m)
RIBs
REX/HIE-ISOLDE
2001/2015
East Area
TT10
n_TOF
2001
TT2
PS
1959 (628 m)
LINAC 4
H⁻
2020
LINAC 3
Ions
1994
LEIR
2005 (78 m)
CLEAR
2017
H⁻ (hydrogen anions)
p (protons)
ions
RIBs (Radioactive Ion Beams)
n (neutrons)
p̄ (antiprotons)
e⁻ (electrons)
μ (muons)
LHC - Large Hadron Collider // SPS - Super Proton Synchrotron // PS - Proton Synchrotron // AD - Antiproton Decelerator // CLEAR - CERN Linear Electron Accelerator for Research // AWAKE - Advanced WAKefield Experiment // ISOLDE - Isotope Separator OnLine // REX/HIE-ISOLDE - Radioactive EXperiment/High Intensity and Energy ISOLDE // MEDICIS // LEIR - Low Energy Ion Ring // LINAC - LINear ACcelerator // n_TOF - Neutrons Time Of Flight // HiRadMat - High-Radiation to Materials // Neutrino Platform

Figure 2.1: From [48]. The CERN accelerator complex, layout in January 2022.

at the experiment site: the Compact Muon Solenoid (CMS) [7] and the A Toroidal LHC ApparatuS (ATLAS) [8] are multipurpose detectors designed to provide sensitivity to SM processes (including the Higgs boson), extra dimensions and particles that could make up dark matter; the LHC beauty experiment (LHCb) [49] is dedicated to heavy flavor physics, designed to look for pieces of evidence of CP-violation and rare decays of b and c quarks; and A Large Ion Collider Experiment (ALICE) [50] focuses on the study of quark-gluon plasma produced in heavy-ion collisions and strongly interacting matter. These are however not the only experiments conducted at CERN, but are the biggest four located around the two large rings of the LHC. A schematic view of the CERN accelerator complex is shown in Figure 2.1, where the accelerating structures, along with other experiments, are also shown.

2.2 The CMS detector

The CMS experiment, depicted in Figure 2.2, is a 21 meters long, 15 meters wide, and 15 meters high general-purpose detector built around a superconducting magnet [51]. Noteworthy aspects of its design include the capability for precise measurement of the muon momentum in the muon system, an excellent energy determination for electrons and photons in the electromagnetic calorimeter, and a state-of-the-art tracking system enabling accurate measurements of the transverse momentum and impact parameter of charged particles.

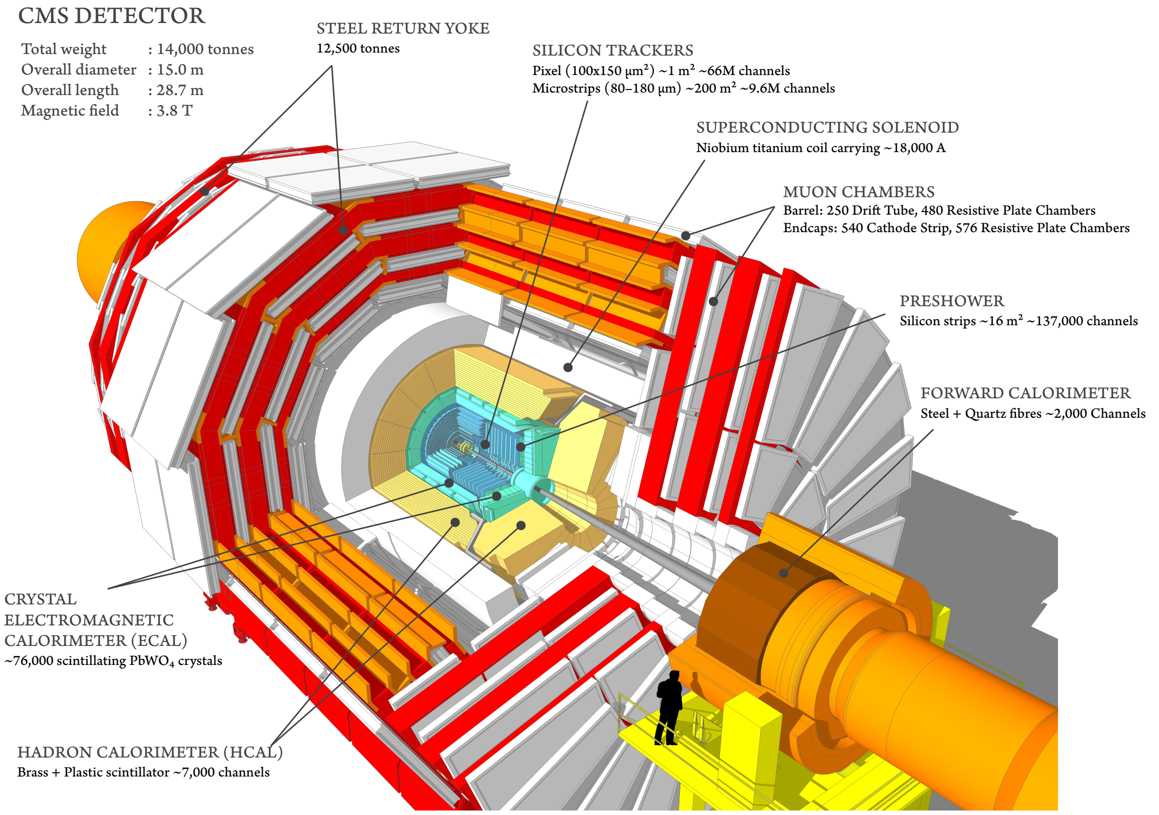
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm²) ~1 m² ~66M channels
Microstrips (80–180 μm) ~200 m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000 A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16 m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 2.2: From [52]. Cutaway diagram of CMS detector.

The CMS magnet is the largest and most powerful among the 4 largest experiments at the LHC. Its design permits the tracker (pixel and strips), the electromagnetic (ECAL), and hadronic (HCAL) calorimeters to be positioned inside the superconducting coil. The high current provides a homogeneous magnetic field of about 3.8 T [53]. In contrast, ATLAS opted to embed only the tracker inside its 2 T solenoid, with the calorimeters and muon system located outside of the magnet in two additional toroidal magnetic fields.

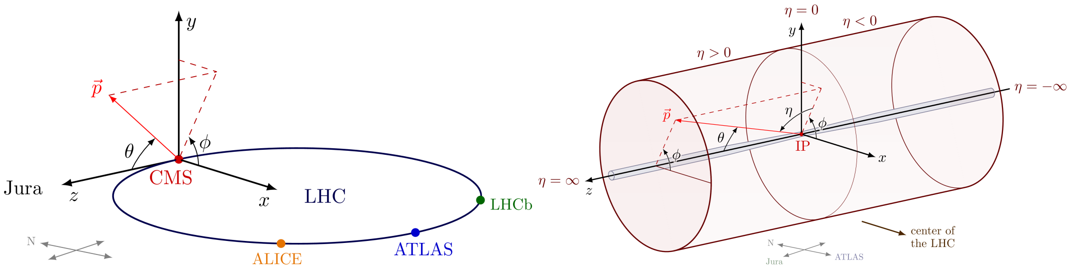

Figure 2.3: From [54]. Left: CMS coordinate system, including the LHC and a compass (the z axis points to the Jura). Right: CMS coordinate system with the cylindrical detector.

The CMS coordinate system, detailed in Figure 2.3, is oriented such that the x axis points toward the center of the LHC. Perpendicular to the x axis, the y axis points upward from the LHC plane. The z axis aligns with the direction of the anti-clockwise circulating beam, pointing toward the Jura mountains, with the origin of coordinates located at the interaction point. The azimuthal angle ϕ is measured from the x axis to the y axis in the x-y plane, extending from 0 to 2π. On the other hand, the polar angle θ is defined in the plane formed by the z axis and the projection of a specific momentum vector onto the x-y plane. It extends from $\theta = 0$ at the z axis to $\theta = \pi/2$ at the x-y plane, with a range spanning from 0 to π. The pseudorapidity η is defined as $\eta = -\log\left[\tan\left(\theta/2\right)\right]$. For particles moving perpendicular to the beam, this leads to $\eta = 0$, and for particles moving parallel (or anti-parallel) $\eta \to \pm\infty$. The rapidity is another related quantity, defined as $y = \log\left[\left(E + p_z\right) / \left(E - p_z\right)\right] / 2$. Differences in rapidity possess the advantageous property of being invariant under boosts along the beam axis. However, in practice, rapidity can be a difficult quantity to measure; one needs both the energy and the total momentum and for highly relativistic particles, where the z component of the momentum is large, this might be challenging as the beam pipe can be in the way of measuring it precisely. Pseudorapidity is thus highly useful, since for highly relativistic (or equivalently, massless) particles, both quantities are equivalent, which is likely to be the case in the LHC experiments. The momentum component transverse to the beam direction, denoted p_T, is known as transverse momentum and is computed from the x and y components of the total momentum vector. Similarly, several transverse quantities can be defined, including the transverse energy $E_T = E\sin\theta$ or the transverse mass $M_T^2 = E^2 - p_z^2$. Differences both in η ($\Delta\eta$) and ϕ ($\Delta\phi$) are Lorentz-invariant quantities. Consequently, the distance between two particles can be measured by a third Lorentz-invariant variable known as ΔR, defined as $\Delta R = [(\Delta\eta)^2 + (\Delta\phi)^2]^{1/2}$.

In the subsequent sections, various specific aspects of the CMS detector capabilities are described. However, for a more detailed description, reference [7] should be consulted.

2.2.1 Tracker

The CMS tracking system [55] is the innermost detector subsystem. It is composed of essentially two parts: the Pixel Tracker and the Strip Tracker, covering a pseudorapidity range up to $|\eta| = 2.5$. Both parts possess different characteristics:

- The Pixel Tracker is the detector that sits the closest to the interaction point. It comprises approximately 66 million silicon pixel cells with dimensions $100 \times 150\ \mu\text{m}^2$, covering a total area of about 1.6 m^2. Pixel detectors allow a spatial resolution of 10 μm in the r-ϕ plane and around 15 μm in the z direction. This resolution is crucial not only for an accurate track reconstruction but also for the determination of primary and secondary interaction vertices.

- The Strip Tracker constitutes the outer layers of the tracking system. It is composed of silicon strip tracker modules, each one of them bearing one or two silicon sensors. In the barrel region, the Strip Tracker modules are arranged in ten layers, extending up to a radius of 1.1 m (see Figure 2.4). These are further divided into an inner part called Tracker Inner Barrel (TIB), consisting of string modules, and an outer part, named Tracker Outer Barrel (TOB), which consists of rod modules. In the endcap region ($|\eta| >$

1.6), the silicon Strip Tracker is formed by two blocks of disks. Three of them belong to the Tracker Inner Disks (TIDs) substructure, while nine disks are part of the Tracker EndCaps (TECs). In the barrel, the strips are oriented along the z direction, while in the endcaps they are oriented along the r direction. All four regions (TIB, TOB, TID, TEC) are equipped with both single- and double-sided microstrip modules. This detector is designed to provide a spatial resolution of approximately 20-50 μm in the r-ϕ plane, and between 200-500 μm along the z direction.

For a more detailed description of the tracking system, and also for the upgrades planned for Phase-2 of the LHC, the reader is referred to [55, 56].

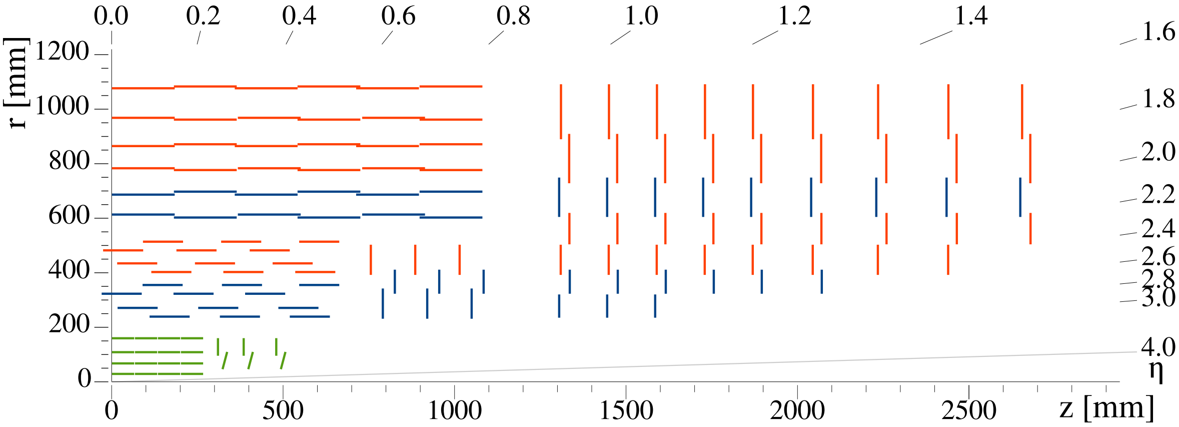

Figure 2.4: From [56]. Sketch of one quarter of the Phase-1 CMS tracking system in r-z plane view. The pixel detector is shown in green, while single-sided and double-sided strip modules are depicted as red and blue segments, respectively.

2.2.2 Electromagnetic calorimeter

The CMS Electromagnetic CALorimeter (ECAL) [57] is a crystal calorimeter that measures the energy of electrons and photons, located just outside the tracker system, but still inside the solenoid. It provides excellent energy resolution in the harsh radiation environment of the LHC, being able to achieve, in particular, a 1% mass resolution for the Higgs boson in the $\gamma\gamma$ decay channel [58]. The calorimeter offers a coverage up to $|\eta| = 3$, and can be split into barrel and endcap regions as shown in Figure 2.5. The barrel region is composed of 61200 crystals across 36 supermodules and uses avalanche photodiodes (APDs) and photodetectors, whereas the endcap region contains 14648 crystals in four half-disk dees and uses vacuum phototriodes. These crystals are composed of lead tungstate ($PbWO_4$), a transparent material denser (8.3 g/cm^3) than iron, with a radiation length of $X_0 = 0.89$ cm, a Molière radius of $R_M = 2.19$ cm, and with fast response (80% of light is emitted within 25 ns). The barrel crystals have a front area of 2.2×2.2 cm^2, a length of 23 cm ($25.8X_0$) and are positioned at $r = 1.29$ m. On the contrary, the endcap crystals are built with a 2.47×2.47 cm^2 front area, a length of 22 cm ($24.7X_0$) and are positioned at $z = \pm 3.17$ m. The barrel covers a region $|\eta| < 1.48$ while the endcap extends within $1.48 < |\eta| < 3$ [59].

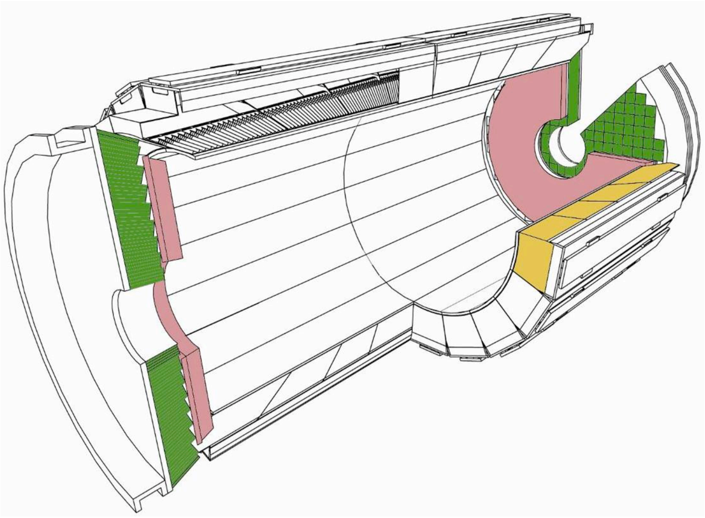

Figure 2.5: From [59]. Layout of the CMS ECAL. One of the 36 barrel supermodules is highlighted in yellow, and the endcaps are highlighted in green.

The ECAL energy resolution can be parameterized by three different contributions:

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} \otimes \frac{b}{E} \otimes c, \tag{2.1}$$

where the first term represents statistical fluctuations associated with the showering process and amplification through photodiodes. The second term accounts for electronic noise contributions and pileup effects and the last term is related to calibration offset.

2.2.3 Hadronic calorimeter

The CMS Hadronic CALorimeter (HCAL) [60] is a sampling calorimeter that plays an important role in the reconstruction of missing energy. It is located immediately outside the ECAL and inside the magnetic coil. Its active elements are brass absorber plates, selected for their short interaction length (providing great capability for containing hadron showers) and non-magnetic nature. The hadron calorimeter can be divided in two segments: the central calorimeter ($|\eta| < 3.0$) and the forward/backward calorimeter ($3.0 < |\eta| < 5.0$).

The central calorimeter comprises the Hadron Barrel (HB) and Hadron Endcap (HE) calorimeters, both positioned inside the magnetic cryostat. To ensure the complete containment of high-energy jets, an Outer Calorimeter (HO) is situated in the barrel and endcap regions. The thickness of the absorber layers composing the HCAL ranges from 60 mm in the barrel to 80 mm in the endcaps.

In the barrel region, the total thickness along a particle trajectory ranges from 5.46 interaction lengths at $\eta \approx 0$ to 10.8 at $\eta \approx 1.3$, while in the endcaps the average is of 11 interaction lengths [61].

The photodetection readout of the HCAL is based on multi-channel hybrid photodiodes, providing an amplified response proportional to the original signal across a wide range of particle energies. The energy resolution of the HCAL can be parameterized as:

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} \otimes b \tag{2.2}$$

where a accounts for statistical fluctuations, analogously to Equation 2.1, and varies along the η range, whilst b is related to calibration and remains relatively independent over the entire pseudorapidity range.

The main feature of the HCAL is its hermeticity, which is crucial for precise energy measurements. To ensure comprehensive energy containment, the central calorimeter is complemented with the forward/backward calorimeter, situated outside the magnet return yokes, 11 meters away from the interaction point.

2.2.4 Muon system

At the time of the construction of the LHC, it was well known that electrons and muons would play a fundamental role in the studies of any physics sector: from the Higgs search (where the $H \rightarrow llll$ are the so-called "golden channels") to "new physics" phenomena, and from electroweak precision measurements to b and t physics [62]. For these reasons, the CMS muon system has been designed and built to have excellent trigger, identification and reconstruction performances with these particles. The LHC physics signatures, along with their separation from the expected background, impose several requirements that must be –and are– fulfilled by the muon system:

- Muon trigger: excellent trigger performances are required on single and multi-muon events. An unambiguous identification of the bunch crossing is obtained by combining fast dedicated trigger detectors, known as Resistive Plate Chambers (RPCs), with detectors that possess precise spatial resolution, such as Drift Tube (DT) and Cathode Strip Chambers (CSCs).

- Redundancy in both trigger and reconstruction is obtained using three technologies. This approach ensures the presence of two independent angular systems that cover the entire angular region.

- CMS provides the capability to measure muons using both the tracker system and the muon spectrometer, enabling accurate measurements of both muon momentum and charge throughout the whole η region, spanning from a few GeV up to a TeV.

- Muon identification: in order to achieve a very high efficieency up to $\eta = 2.4$, at least 16 interaction lengths of material are required, along with a very sophisticated set of algorithms.

- Robustness: detectors are able to cope with large magnetic fields and high radiation.

To meet these requirements, the muon system is located inside the magnet return yoke and consists of three types of gaseous detectors to account for different radiation environments: DT chambers in the barrel and CSCs in the endcaps, both complemented by a RPC system (see Figure 2.6). The system is equipped with four layers of muon chambers in the barrel and four in each endcap region, with each layer providing track segments reconstructed from distributed hits. These tracks are combined with information from the tracker to form complete muon tracks. The geometric arrangement of the system enables the coverage of a pseudorapidity range of up to 2.4. The high return yoke field ensures excellent momentum resolution and charge identification, even in the absence of inner tracker information.

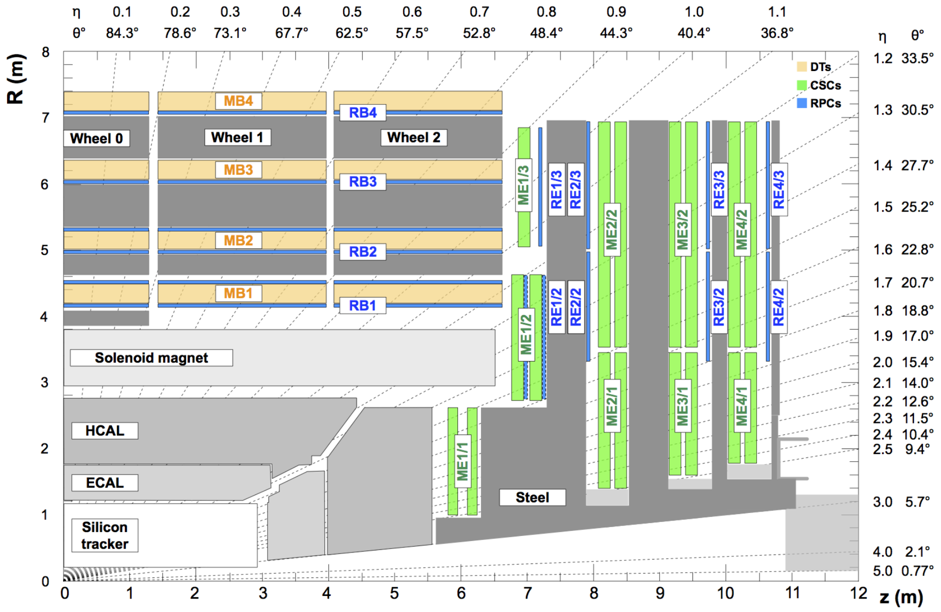

Figure 2.6: From [63]. An R-z cross section of a quadrant of the CMS detector with the axis parallel to the beam (z) running horizontally and the radius (R) increasing upward. The interaction point is at the lower left corner. The locations of the various muon stations and the steel flux-return disks (dark areas) are shown. The Drift Tube (DT) chambers are labeled MB ("Muon Barrel") and the Cathode Strip Chambers (CSCs) are labeled ME ("Muon Endcap"). Resistive Plate Chambers (RPCs) are mounted in both the barrel and endcaps of CMS, where they are labeled RB and RE, respectively.

Given the uniform magnetic field, the low muon rate and the neutron-induced background events, DTs are the choice for the muon detectors in the barrel region within $|\eta| < 1.2$. DTs are long aluminum cells operated with a mixture of gases and an anode in the center to collect ionization charges. CSCs are instead used in the endcap regions due to the expected high number of background of events, the high muon flux and a large non-uniform magnetic field. These cover the pseudorapidity range $0.9 < |\eta| < 2.4$. CSCs are trapezoidal multiwire proportional chambers filled with different gases, allowing for charge collection in the event of gas ionization. Lastly, the RPC system, which consists of 6 layers of RPCs, complements the measurements up to $|\eta| = 1.6$. Although the spatial resolution is poor, the RPCs compensate for it with an excellent time resolution of about 1 ns, making them ideal for identification of bunch crossings and triggering purposes. The RPCs consist of double-gap bakelite chambers filled with several gases, sharing similar functionality with DTs and CSCs.

2.2.5 CMS trigger system

The opposing beams circulating around the LHC cross each other once every 25 ns, resulting in a crossing frequency of 40 MHz. Depending on the luminosity, multiple collisions occur at each bunch crossing, translating into a data flux of $\geq$ 40 MHz. However, due to technical limitations, the rate at which events can be recorded is restricted to a few hundred per second, as the average disk space required for a single event is typically a few MBytes. In addition, if all events were to be recorded, the physical space needed for data storage would quickly become unmanageable. Therefore, during data-taking (online), events must undergo selection criteria, significantly reducing the storage rate while maintaining a high efficiency for potentially interesting events.

The system responsible for this task is the *trigger*, while the data acquisition system (DAQ) handles the transfer of data from the subdetectors to the storage structures. The CMS trigger is implemented at two levels:

- The Level-1 (L1) trigger [64] is the first selection that the events have to pass in order to be recorded. The L1 trigger serves as a rapid and coarse selection, aimed at discriminating as many non-interesting events as possible. It reduces the rate of events from the order of 40 MHz (with tens of event for each bunch crossing) down to around 100 kHz in less than 3 μs. The L1 trigger makes use of calorimetric measurements and the muon system without considering any information from the tracker. The trigger decision is based on what is known as "trigger primitives", which entail the presence and number of objects, such as electrons, photons, muons, jets, τ-jets, and E_T^{miss} with transverse energy or p_T above a certain threshold.
- The High Level Trigger (HLT) [65] is a software system deployed on a filter farm of approximately a thousand commercial processors. The HLT has full access to the complete readout data and can execute more complex calculations than the L1 trigger, similar to those performed by offline analysis software. By doing this, the HLT reduces the rate of events from approximately 100 kHz down to about 300 Hz and, subsequently, these events are pipelined toward data storage.

2.2.6 Event reconstruction at CMS

The CMS detector has cylindrical symmetry along the beam axis, and a particle that emerges from the interaction point crosses several subsystems which can, in turn, provide information about its nature and kinematic properties. The global event reconstruction at CMS strongly relies on this combination of information between the several subdetectors that compose it. This approach is called particle-flow (PF) algorithm [66, 67].

The energy of photons is directly obtained from an energy cluster measured in the ECAL, without any association to a track in the tracker system. Electrons, on the other hand, being charged particles, are identified as clusters in ECAL linked with a compatible track in the tracker system. Their energy is derived from a combination of the measurement of the electron momentum at the primary interaction vertex, the energy of the ECAL cluster, and the energy sum from all the bremsstrahlung photons that are spatially compatible with originating from the electron track. The momentum of muons is obtained from the curvature of the corresponding track. Charged hadrons are identified as energy clusters in the HCAL and the ECAL, accompanied by a compatible track in the tracker system. Their energy is inferred from the combination of measurements of their momenta inside the tracker and the energy deposits inside the ECAL and HCAL. Lastly, neutral hadrons are identified as clusters in the HCAL and ECAL with no track associated, or as energy excesses in ECAL clusters that are not compatible with energy depositions of the corresponding electromagnetic signature.

The PF algorithm is an iterative process that attempts to link all these signals from individual subdetectors. Initially, the algorithm selects a global muon, giving rise to a "particle-flow muon". If the combined momentum of a "particle-flow muon" aligns within three standard deviations with that determined solely from the tracker, it is removed from the pool of physical signatures, or "block". The identification and reconstruction of electrons follow a similar procedure, employing the ECAL clusters and compatible tracks within the tracker system. If the link of the physical signatures is of sufficient quality, then the elements of the event are removed from the block. Similar procedures are followed for the rest of the tracks and clusters, in order to identify charged and neutral hadrons, along with photons, until no objects are left in the block. For a more thorough description of the particle-flow algorithm and global event reconstruction at CMS, the reader is referred to [66, 67].

2.2.7 Jet reconstruction

In this text, mainly final states with four b quarks will be analyzed. As already mentioned in Section 1.1, b quarks undergo hadronization shortly after they are produced, due to the nature of strong interactions. This leads to the impossibility of detecting them directly; instead, the hadronization process produces several particles that leave traces in the calorimeters of the experiment. Generally, these particles emerge along the momentum direction of the initial quark, and their distribution can serve as an indicator of the initial quark's momentum boost. Therefore, it is of the utmost importance to discuss the procedure for jet reconstruction in this section. In Section 2.2.8, also flavor tagging will be addressed.

CMS reconstructs jets (PFJets) starting from all the PF candidates in the event: muons, electrons, photons, neutral hadrons, and charged hadrons. As explained in the previous section, clusters in the ECAL (HCAL) that are linked with compatible tracks in the tracker system are associated with electrons (charged hadrons). If instead an excess of calorimetric energy deposition is found with respect to the momentum as measured from the track reconstruction, or the calorimetric deposits are not linked with any tracks, then the energy is identified to be coming from a photon or neutral hadron. Linked objects, like jets, are built starting from the list of particle candidates in the event.

Clustering algorithms are responsible for tentatively merging together these particle candidates into jets. The clustering algorithm most widely used at CMS is the anti-k_t algorithm [68], which successively merges pairs of particle candidates in order of increasing relative transverse momentum. Firstly, one can define the distance d_{ij} between entities (particles, pseudojets) i and j:

$$d_{ij} = \min(p_{Ti}^{2m},\, p_{Tj}^{2m}) \frac{\Delta_{ij}^2}{R^2}, \tag{2.3}$$

where $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and p_{Ti}, y_i and ϕ_i are respectively the transverse momentum, rapidity and azimuth angle of particle i. In addition to the parameter R (which is typically 0.4 in most reconstructions), the parameter m governs the relative power of the energy versus geometrical (Δ_{ij}) scales. The $m = 1$ choice recovers the inclusive k_t algorithm, while the case $m = 0$ corresponds to the inclusive Cambridge/Aachen algorithm. The anti-k_t algorithm employs a value of $m = -1$. This is particularly convenient, since in the case where an event presents few well-separated hard particles and many soft particles, the distance between a hard particle 1 and a soft particle i,

$$d_{1i} = \min(1/p_{T1}^2,\, 1/p_{Ti}^2) \frac{\Delta_{1i}^2}{R^2}, \tag{2.4}$$

is entirely determined by the transverse momentum of the hard particle and the Δ_{1i} separation. The d_{ij} between similarly separated soft particles will instead be much larger. As a consequence, soft particles will tend to cluster with hard ones long before they cluster with other soft particles. The key feature of this algorithm is that soft particles do not modify the shape of the jet, making it inherently resilient to soft radiation. This feature is known as *infrared safety*. The anti-k_t is also *collinear safe*, which means that the splitting of a hard particle in two or more collinear softer candidates does not change the result of the clustering. A schematic depiction of the anti-k_t clustering algorithm compared to other algorithms is shown in Figure 2.7.

Once the clustering ends, the raw jet momentum is determined as the vectorial sum of the momenta of all the particles present in the jet. However, this energy measurement is biased [69], and several corrections need to be applied to obtain the energy of the initial parton. The jet energy resolution typically amounts to 15% at 10 GeV, 8% at 100 GeV and 4% at 1 TeV.

2.2.8 B-tagging at CMS

The ability to identify jets containing B hadrons is important to many of the physics analyses developed at the LHC: $b\bar{b}$ and $t\bar{t}$ production, Higgs bosons with $b\bar{b}$ final states, and essentially any process that involves the production of b quarks [70].

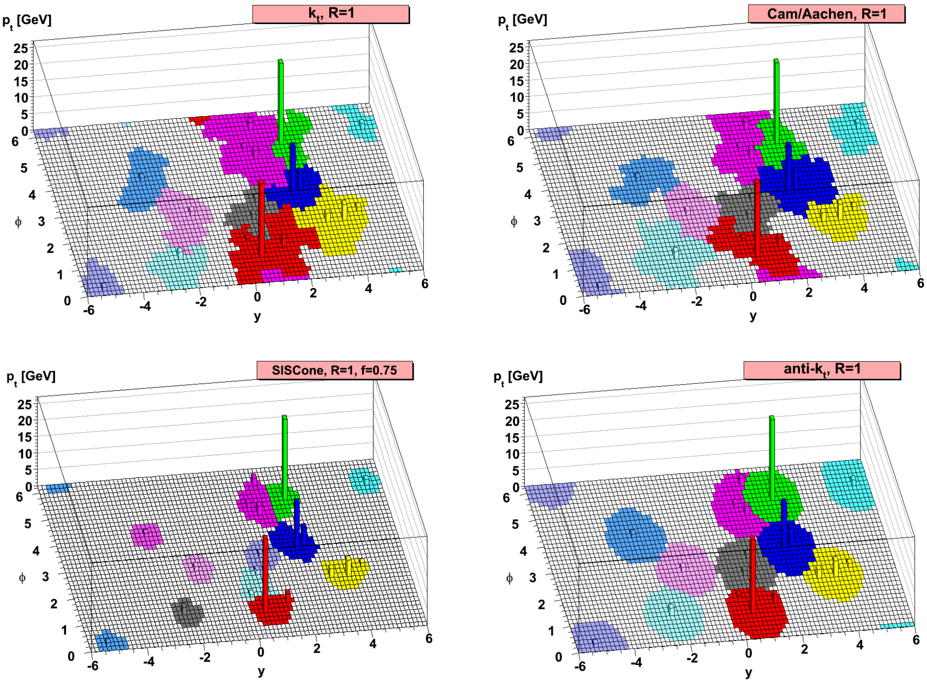

Figure 2.7: From [68]. A Monte Carlo sample parton-level event, together with many random soft particle candidates, clustered with four different jets algorithms, illustrating the "active" catchment areas of the resulting hard jets.

The simplest *lifetime-based* b-tags are based on the 3D signed impact parameter d_0. This is the signed distance of closest approach of a track to the primary vertex, and indicates that the track is oriented in the direction of the jet if it is positive, while a negative value is a result of a track originating away from the jet. Quality cuts are applied in this algorithm to reject tracks that are badly reconstructed or originating from γ conversions, etc. By analyzing the d_0 values of tracks within a jet, the algorithm can identify jets that are likely to be originating from b quarks.

Track-counting b-tagging method, on the other hand, orders the tracks in a jet by decreasing impact parameter significance d_0/σ, which serves as the discriminator, and it typically presents large, positive tails for b-jets. The algorithm is based on the simple requirement of a minimum number of good quality tracks with an impact parameter significance above a certain threshold.

The slightly more sophisticated *probability* b-tag calculates a confidence level for each track in the jet, indicating how compatible it is with being originated from the primary vertex. These individual track confidence levels are then combined to form a single pseudo-confidence level, representing the probability that all the tracks in the jet originated from the primary vertex. This pseudo-confidence level serves as the b-tag discriminator.

As an even more complex alternative, one can also tag b-jets by searching for the electronic and muonic signatures arising at semi-leptonic B decays. Several variables relating to the lepton are fed and combined inside a neural network. This is however less efficient than the lifetime tags, due to the small B semi-leptonic decay branching fraction.

Lastly, the most sophisticated algorithm is the *Combined Secondary Vertex* (CSV) b-tag. This attempts to reconstruct a secondary vertex of the weakly decaying B hadron and employs variables related to such decay to calculate its b-tag discriminator. The combined algorithm merges the information about track impact parameters and secondary vertices, providing information even when no secondary vertices are found. A likelihood discriminator is built and trained on categories ordered by the characteristics of the reconstructed jets, such as the number of tracks in the jet, the secondary vertex mass, etc (see Reference [71] for a thorough description).

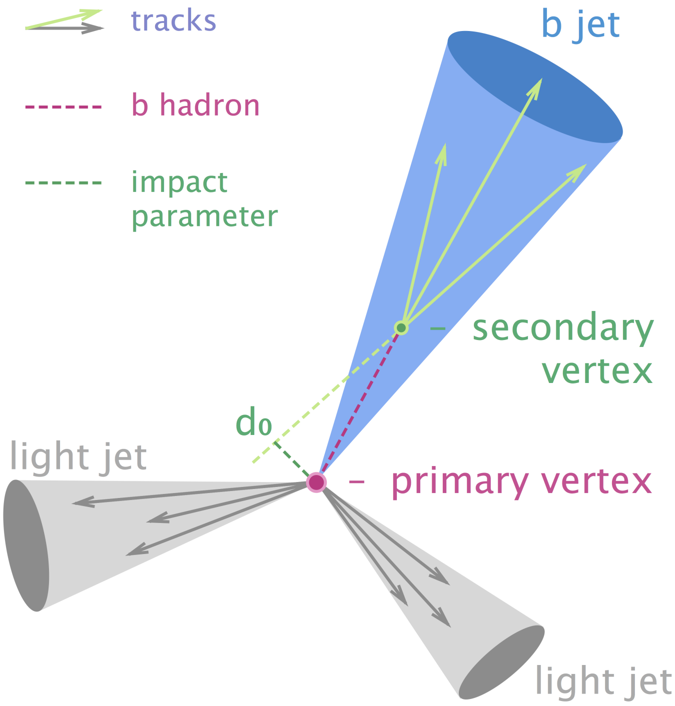

Figure 2.8: From [72]. Sketch of a b-jet (blue cone) originating from the secondary vertex (green point) where B hadron has decayed a distance d_0 apart from the primary vertex of the interaction.

The efficiency of the b-tagging and the rate of the misidentification of non-b-jets depends on the cut of whichever discriminant is used for each algorithm. Typically, three working points are defined depending on the efficiency values: the tight (T) working point, for a considerably low misidentification rate of non-b-jets (0.1%), which in turn provides a not-so-high b-tag efficiency; the loose (L) working point, in which the b-tag efficiency is high, but also the misidentification rate is increased (10%); and the medium (M) working point, which is a compromise between both (1% of misidentification rate). The performance of tagging algorithms is measured on independent samples of multijet events, which produce multiplicative scale factors to account for differences between data and simulation.

Chapter 3

Determination of the kinematics of the QCD background

As already mentioned in Chapter 1, the search for Higgs boson pair production decaying to a final state containing four b quarks poses significant challenges. However, as shown in the CMS [44] and ATLAS [46] di-Higgs combination results, this channel, along with $b\bar{b}\tau^+\tau^-$, has the potential to exceed the clean but statistically limited $b\bar{b}\gamma\gamma$ sensitivity. Doing so will require high-dimensional background models to facilitate the multivariate signal extraction required for optimal sensitivity. To avoid being dominated by systematic uncertainties, these models will have to be validated at sub-percent level precision. In this chapter we present, in Sections 3.1 and 3.2, two novel methods that have been used to determine the QCD background probability density function (PDF). The approach described in Section 3.2, in particular, is used in an ongoing analysis that takes advantage of a complex neural network to estimate this PDF, and serves as a stepping stone toward the construction of the autoencoder architecture, which will be described in Chapter 4.

Here we shall indicate, nonetheless, that the determination of the PDF of QCD background at the LHC is an intricate problem. In particular cases, physics backgrounds from QCD processes may also often prove impossible to precisely model with simulated data, due to the very large cross section of the reactions to be considered and the consequently unmanageable demands posed on computing time. The LHC experiments obtain, in those cases, estimates of background shapes and normalization by using data-driven methods in a variety of ways that depend on the specific features of the final states in examination. Two of such methods are the hemisphere mixing and the kinematic reweighting techniques, which can also be used complementarily for cross-validation.

3.1 Hemisphere mixing

A previous CMS $HH \to b\bar{b}b\bar{b}$ analysis [73] estimated the $4b$ background using a hemisphere mixing technique [74]. Hemisphere mixing forms a synthetic dataset by splitting individual events into two "hemispheres" and then combines hemispheres from different events. The main idea behind this approach is that the mixed dataset will be similar to the expected $4b$ background because the b-jets primarily arise from gluon splitting from an underlying $2 \to 2$ scattering process. For searches with a sufficiently low signal-to-background ratio, as it is the case for HH in the $4b$ channel, the hemisphere mixing produces synthetic datasets that are essentially signal-depleted. The assumption under which this mechanism works is that the relevant event-level correlations are captured at the level of hemispheres, and not between hemispheres' substructures. The substructures between hemispheres are assumed to be uncorrelated for background events, whereas for HH events, both hemispheres will have a similar substructure from the underlying 125 GeV resonances. The way this method is applied is the following: let us consider a dataset of QCD multijet events, either collected at an LHC experiment or simulated my an MC program. For each event an axis may be constructed on the plane transverse to the beams, the "transverse thrust axis", defined as the azimuthal angle ϕ_T which maximizes the transverse thrust quantity T:

$$T = \sum_j p_{T,j} |\sin(\phi_j - \phi_T)|, \tag{3.1}$$

where $p_{T,j}$ is the transverse momentum of the jet j. Once ϕ_T is known, one can also define the related quantity T_a:

$$T_a = \sum_j p_{T,j} |\sin(\phi_j - \phi_T)|. \tag{3.2}$$

The transverse thrust axis defines a plane orthogonal to it which divides the event in two separate hemispheres. Each hemisphere is characterized by its number of jets N_j, its number of b-tagged jets N_b, the sum of the projections of the p_T along the thrust axis T, the combined mass of the jets M, the variable T_a, and the sum of the jets p_z components, P_z. If we start with N events, we can form a library of $2N$ hemispheres. With this library, artificial replicas of the original events can be formed by exchanging hemispheres of events with similar kinematics. For each original event, composed of hemispheres h_1 and h_2, we can look in the library for the two hemispheres h_p and h_q that are the most similar to h_1 and h_2, in the sense that they have the same exact value of N_j and N_b, and the smallest distances $D(1,p)$ and $D(2,q)$, defined as below:

$$D(1,p)^2 = \frac{\left[T(h_1)-T(h_p)\right]^2}{V_T} + \frac{\left[M(h_1)-M(h_p)\right]^2}{V_M} + \frac{\left[T_a(h_1)-T_a(h_p)\right]^2}{V_{T_a}} + \frac{\left[|P_z(h_1)|-|P_z(h_p)|\right]^2}{V_{P_z}}, \tag{3.3}$$

$$D(2,q)^2 = \frac{\left[T(h_2)-T(h_q)\right]^2}{V_T} + \frac{\left[M(h_2)-M(h_q)\right]^2}{V_M} + \frac{\left[T_a(h_2)-T_a(h_q)\right]^2}{V_{T_a}} + \frac{\left[|P_z(h_2)|-|P_z(h_q)|\right]^2}{V_{P_z}}. \tag{3.4}$$

In the equations above, the denominators contain the variances of the considered variables. The identified pair of hemispheres h_p, h_q constitutes an entirely new event, as they are prevented from being equal to h_1, h_2, respectively. Lastly, h_p and h_q are rotated along the azimuthal direction to match the original thrust axis of the modeled event. This procedure, also depicted in Figure 3.1, allows to create a synthetic dataset with the same number of events as the original sample. When the original sample contains a small fraction (say, a few percent) of events originated by a heavy particle decay, for example, di-Higgs boson decay, the mixing procedure smears out the features of the minority component. This happens naturally as the probability of the mixing procedure of combining two hemispheres both proceeding from a signal event scales as the square of the signal fraction. The remaining dataset is therefore a faithful model of the dominant process, in this case, QCD background.

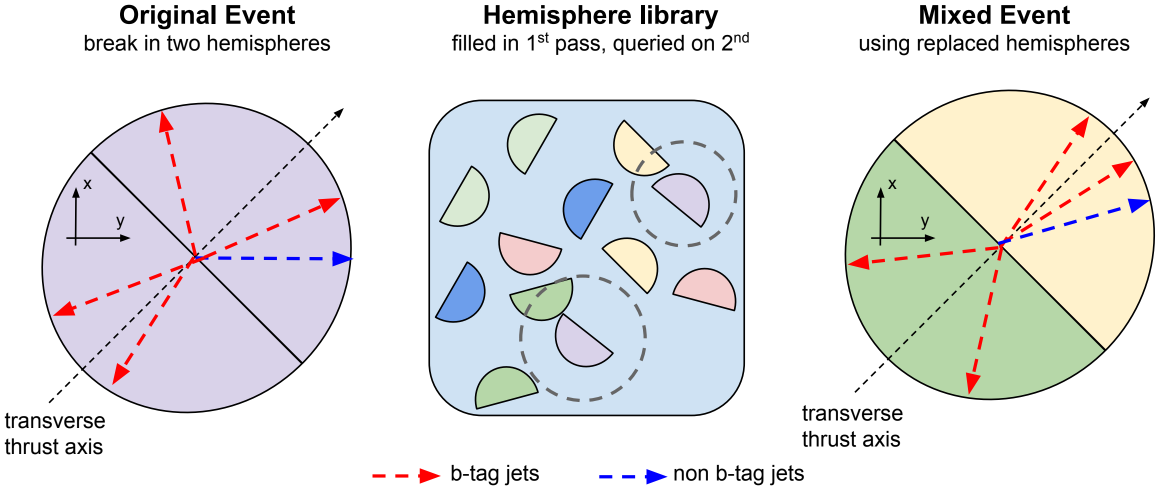

Figure 3.1: From [73]. An illustration of the hemisphere mixing procedure. The transverse thrust axis is defined as the axis on which the sum of the absolute values of the projections of the p_T of the jets is maximal. Once the thrust axis is defined, the event is divided into two halves by cutting along the axis perpendicular to the transverse thrust axis. One such half is called a hemisphere. In a preliminary step, each event in the original N-event dataset is split into two hemispheres that are collected in a library of $2N$ hemispheres. Once the library is created, each event is used as a basis for creating artificial events. These are constructed by picking two hemispheres from the library that are similar to the two hemispheres that make up the original event.

Hemisphere mixing has, however, several potential drawbacks. First, because the $4b$ dataset is used to derive the background, the statistical uncertainty in the predicted background is the same as for that observed in the Signal Region (SR)[1]. A second concern is that mixing may not fully suppress the event-level signal correlations, especially when the hemisphere matching is likely to form pairs of hemispheres both belonging to signal events. If this happens, then the background model will effectively be biased by signal contamination. In Figure 3.2, the percentage of bias on the signal fraction estimated in [74] from fits to the reconstructed Higgs boson mass distributions is drawn as a function of the true signal fraction. For signal fractions above 5%, the signal contamination is seen to affect the corresponding artificial sample.

There is generally a trade off in choosing the matching criteria: the mixed events can be more similar to the ones from the unmixed dataset, but at the cost of preserving event-level correlations, which are more likely, in turn, to introduce signal contamination biases. Some studies have employed this method to determine the shape and scale of the $4b$ background

[1] To be fully precise, one can actually create more than one pair starting from the same event, so the statistical uncertainty could thus be largely reduced. This technique has been employed in [74] to evaluate the systematic uncertainties.

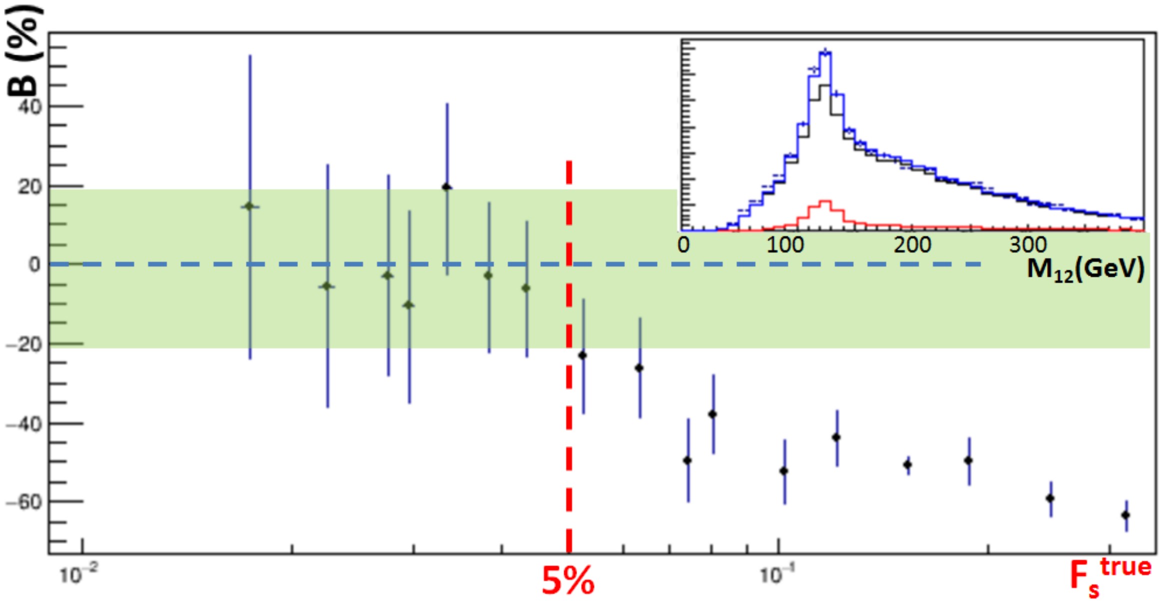

Figure 3.2: From [74]. The percentage bias B (black points) on the estimated signal fraction extracted from fits to the reconstructed Higgs boson mass distributions is drawn as a function of the true signal fraction F_s^{true}. The green band shows the level of bias considered acceptable for searches of new signals in hadron collider data. The upper right inset shows the distribution of the leading jet pair mass m_{12} of the QCD (black) and signal (red) components, their sum (blue), and the fit result (points with uncertainty bars) for a 5% signal contamination.

templates (see for example [73]), but others use it as a cross-validation method, to assess systematic uncertainties of alternative approaches, such as kinematic reweighting.

3.2 Kinematic reweighting

Kinematic reweighting is a general procedure with a major use-case for particle physics in modifying the output of MC simulations to reduce the disagreement with real data collected at colliders [75]. The idea behind kinematic reweighting is that with the help of a Boosted Decision Tree (BDT) or more advanced techniques, such as MultiVariate Analysis (MVA) [76], one is able to identify regions in the phase space that are the most sensitive to differences between MC simulations and real data, and thus compute weights for the events in those regions to account for such differences. An example of application of this method is the study carried out by CMS [77]. In that work, the multijet background arising from QCD and $t\bar{t}$ hadronic processes is estimated from data using background-dominated regions. Analysis signal (A_{SR}) and control (A_{CR}) regions are defined by requiring $\chi < 25$ GeV and 25 GeV $< \chi <$ 50 GeV, respectively, where χ is the distance from the expected peak position of the two Higgs boson candidates' invariant masses and is defined as:

$$\chi = \sqrt{(m_{H_1} - 125\,\text{GeV})^2 + (m_{H_2} - 120\,\text{GeV})^2}. \tag{3.5}$$

The center of the expected peak differs from (125, 125) GeV because of the residual momentum dependence of multivariate energy regression that more strongly impacts the softer H candidate. Both A_{SR} and A_{CR} are divided into a four b-jet ($4b$) and three b-jet ($3b$) region by requiring the b-jet candidate with the lowest value of the b-tagging discriminant to either satisfy or fail the medium working point of such discriminant. An example of the mentioned control and signal regions in the $4b$ split of the dataset is shown in Figure 3.3 (left).

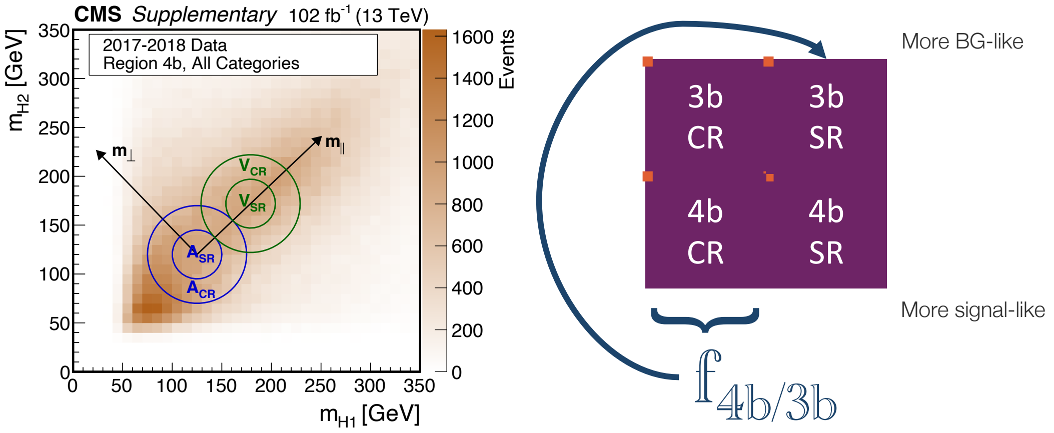

Figure 3.3: Left: from [77]. Distribution of data events selected in the four b-tagged category for events recorded in 2017-2018 as a function of the masses of the two Higgs boson candidates. The blue and green circles correspond to the analysis and validation regions, respectively, denoted by the letters A and V. The inner circle corresponds to the signal region while the ring around it defines the control region used for the background modeling. The figure also shows the definition of the variables $m_{\|}$ and $m_{\perp}$ that are used in the analysis. For a full description of the analysis method, we invite the reader to consult [77]. Right: depiction of the computation of the normalization factor to $3b$ signal region events to model the background in the $4b$ signal region.

Background events in the $A_{\mathrm{SR}}^{4\mathrm{b}}$ region are modeled from events in the $A_{\mathrm{SR}}^{3\mathrm{b}}$ region. The former is the most sensitive region to signal contamination of the analysis, while the latter provides a sample enriched in multijet background events. The normalization is determined by scaling the observed number of events in $A_{\mathrm{SR}}^{3\mathrm{b}}$ by a transfer factor computed as the ratio between the number of events in $A_{\mathrm{CR}}^{4\mathrm{b}}$ and $A_{\mathrm{CR}}^{3\mathrm{b}}$ regions (called $f_{4\mathrm{b}/3\mathrm{b}}$ in Figure 3.3, right). Differences in the distributions of several variables between the $3b$ and the $4b$ regions are then addressed with the BDT-reweighting method, which uses a dedicated metric to identify the regions in the phase space with the largest differences in the distributions, and computes an event weight to correct for them. In this approach, the space of variables is split into a few large regions, and a decision tree is used to separate these regions based on fulfilling simple conditions. The regions (associated with leaves of the tree) that are suitable for reweighting are found by optimizing the χ^2:

$$\chi^2 = \sum_{\mathrm{leaf}} \frac{(\omega_{\mathrm{leaf,\,3b}} - \omega_{\mathrm{leaf,\,4b}})^2}{\omega_{\mathrm{leaf,\,3b}} + \omega_{\mathrm{leaf,\,4b}}}. \tag{3.6}$$

This procedure is validated by applying it to a signal-depleted region, defined by shifting the signal and control regions according to the definition of χ, using $(179,\ 172)$ GeV as the center position in the candidates' invariant masses distribution.

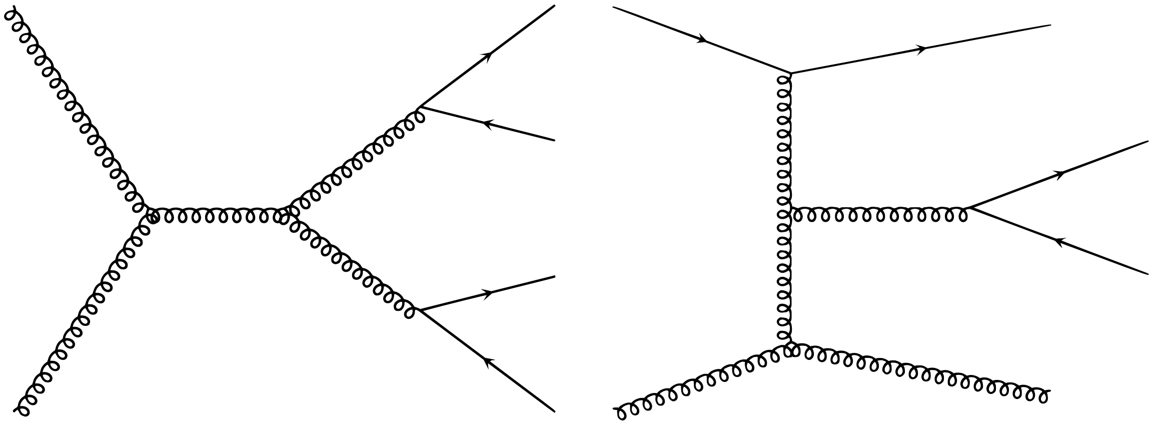

Figure 3.4: From [78]. Two LO partial amplitudes contributing to the multijet background. The diagram on the left represents two-to-two gluon scattering which is one of the dominant terms contributing to the $b\bar{b}b\bar{b}$ and $b\bar{b}c\bar{c}$ processes. This process produces two low-mass pairs of b-jets due to the off-shell gluon propagators. The diagram on the right is heavily suppressed in the four-tag selection where you only get one $b\bar{b}$ pair from the gluon splitting. In this case the $\Delta R(j,j)$ distribution of the two candidate jets which are not the closest is expected to be larger.

An example of the variables that are especially sensitive to the differences in the distributions is the angular separation between pairs of dijets. In particular, the $4b$ sample is dominated by two-to-two gluon scattering where each outgoing gluon splits to $b\bar{b}$ (Figure 3.4, left). This produces a topology where the two tagged b dijets are produced with low $\Delta R(j,j)$. The $3b$ sample also contains this process, in addition to a mixture of processes where the untagged candidate jet can be produced without gluon splitting (Figure 3.4, right). A recent analysis [78], currently ongoing, exploits this kind of kinematic differences with a hierarchical residual neural network and attention blocks to perform $4b$ versus $3b$ separation. This network is called FvT (four versus three-tag), and will be discussed further in Chapter 4, as it will serve as the building skeleton for our autoencoder architecture.

Chapter 4

Machine learning: the autoencoder

In recent years, Machine Learning (ML) has brought about a revolutionary shift in computation capabilities. It has allowed computers to learn from data, detect patterns, and make decisions independently, eliminating the need for explicit programming. This transformative potential has found applications across various disciplines, propelling advancements in numerous fields. High-Energy Physics (HEP) has not been an exception, since it is a field where large amounts of data, complex data structures, and hidden (or, at least, not apparent) correlations between variables are ubiquitous. These are all characteristics that make the discipline ideal for the application of ML algorithms, and their success has done nothing but incite their use in an increasing number of tasks. The current most frequently used ML algorithms in HEP are Boosted Decision Trees (BDT) and Neural Networks (NN) [79].

Decision trees are hierarchical structures that make decisions based on a series of rules or conditions. They split data into subsets based on feature values and make predictions at the leaves of the tree. Decision trees are known for their ability to handle complex relationships in data. Neural networks, on the other hand, consist of layers of interconnected nodes, also known as neurons. These neurons process and transmit information in a way that allows the network to learn the main features of the data.

In ML applications to HEP, physical variables are typically selected and introduced to a model. This model is then trained for classification tasks (such as distinguishing between the signal and background nature of an event) or regression tasks (to capture the most important features within a dataset). During training, supervised models will use the input variables to produce output results, which are then compared to the expected output. In the classification signal-versus-background example, this is done by labeling each event as signal or background, depending on its origin. Then, during training, events are input into the network, which employs adjustable weights and biases to produce the probability that an event belongs to either class. This phase is referred to as the *forward pass*. The calculated probability is then compared to the event's true label, typically represented as 0 or 1. The disparity between the output probability and the actual label contributes to a quantified loss. However, a fundamental characteristic of machine learning algorithms is the differentiability of this loss. This means that its derivatives with respect to the weights and biases of the network can be computed, and after each training cycle, the network will automatically adjust its weights to decrease the loss, thus improving the output for the next round. This process, in which the model adjusts its weights in reverse from the output, is referred to as *backpropagation*.

Neural networks have been used in HEP for some time. Yet, advancements in training algorithms and increased computing power have precipitated the widespread adoption of what are termed Deep Learning (DL) algorithms [80, 81]. These algorithms comprise a distinct subset of machine learning (ML) techniques, utilizing neural networks with multiple layers–hence the term "deep". Deep learning models are particularly well-suited for tasks involving large and complex datasets, such as image and speech recognition, natural language processing, and many more. The deep architecture allows these models to automatically learn hierarchies of features from the data, which can lead to highly accurate and sophisticated predictions. Some DNNs used in HEP are: fully-connected (FCN), convolutional (CNN) and recurrent (RNN) networks. Additionally, NNs are used in the context of generative models, forming part of Variational AutoEncoders (VAE) and more recent Generative Adversarial Networks (GAN).

In the following sections, we explain the main characteristics of the autoencoder structure, its functioning and the implemented model in this work.

4.1 The autoencoder

An autoencoder is a type of artificial neural network used to learn meaningful representations of unlabeled data. Autoencoders belong to the category of unsupervised learning tools because they derive knowledge not from external, human-generated labels, but by aiming to replicate their input as output. Their utility stems from their characteristic "bottleneck" structure: in essence, the input information undergoes a dimension-reduction phase, filtering out irrelevant or less significant features before attempting reconstruction based on the retained features. An autoencoder is typically divided into two sections: the portion preceding the bottleneck is termed the *encoder*, while the subsequent part is referred to as the *decoder*. A generic autoencoder arrangement is illustrated in Figure 4.1.

Given the encoder and decoder, their outputs can be respectively written as:

$$\mathbf{z} = g(\mathbf{x},\, \mathcal{W}_g), \qquad \mathbf{y} = f(\mathbf{z},\, \mathcal{W}_f); \tag{4.1}$$

where g and f are the –generally– non-linear transformations and $\mathcal{W}_g$ and $\mathcal{W}_f$ are the sets of weights of the encoder and the decoder, respectively. As already mentioned, the autoencoder is an unsupervised learning tool, which means that it learns from comparing the input with its own output. The loss function that is typically used with such a structure

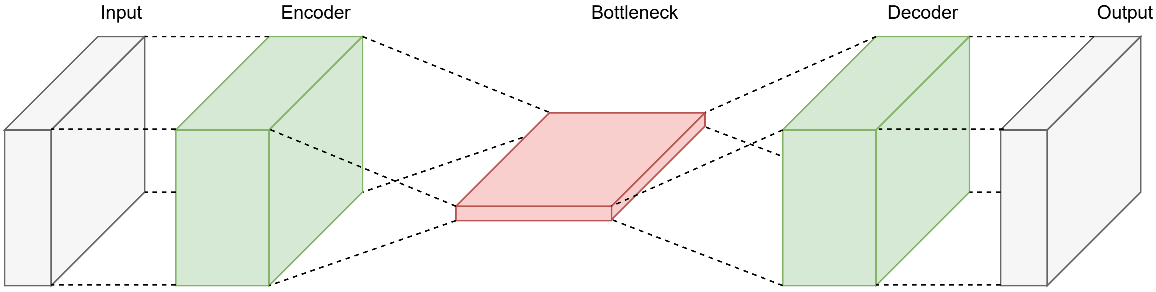

Figure 4.1: General scheme of an autoencoder. The input features are translated by the encoder into a dimension-reduced latent space in the bottleneck, which is later used by the decoder to reproduce the input.

is the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{MSE}}(\mathbf{x}, \mathcal{W}_g, \mathcal{W}_f) = ||\mathbf{x} - \mathbf{y}||_2^2 = ||\mathbf{x} - f(g(\mathbf{x}, \mathcal{W}_g), \mathcal{W}_f)||_2^2, \tag{4.2}$$

where $||\mathbf{x} - \mathbf{y}||_2$ is the L_2-norm or the Euclidean distance [82] between the n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, defined as:

$$||\mathbf{x} - \mathbf{y}||_2 = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}. \tag{4.3}$$

The Equation 4.2 is thus used as an estimator of the true accuracy of an autoencoder in reconstructing its input, as it measures the squared deviations from the true data. It is, however, very important that the input features all have a similar magnitude because the loss will be directly proportional to their difference. Larger quantities, even if reproduced accurately in proportion, will present a much larger loss than, say, quantities smaller than 1 that are more poorly reproduced. The MSE loss also has other issues, as it tends to produce outputs centered around the mean of the input features, and the variance of such distributions has to be carefully examined in order to choose an appropriate architecture to improve the expressiveness of the output. This will be more deeply discussed in Section 4.5.

4.2 Data preparation

An important process in building machine learning models is data preparation. Typically, data that will be fed to a DL architecture present hidden symmetries, redundancies, or non-linear topologies. An example is the azimuthal angle ϕ of jet objects at the LHC. When using Equation 4.2 to compute the loss between ϕ values, one has to be extremely careful with the domain where they are defined. If these range in $[-\pi, \pi]$, as it is the case, we cannot simply impose the loss between the output and input values, as a computed output of 2π for an input of 0, for example, will have a MSE loss of $(2\pi)^2$, where in fact input and output correspond to the same point in ϕ space, and should therefore have a loss of zero.

In our case, the autoencoder takes as input the jet four-vectors in $\{p_T, \eta, \phi, m\}$ coordinates.

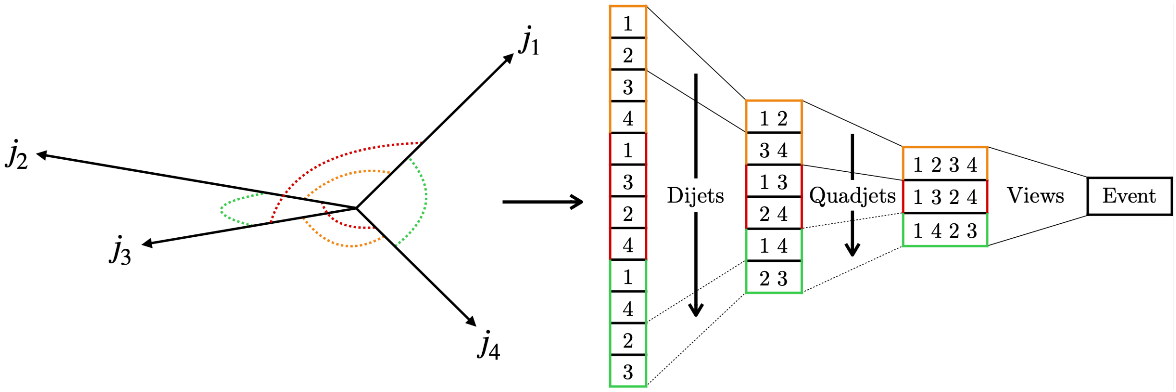

Figure 4.2: From [78]. High level sketch of the input data preparation.

Its task is to reproduce in output the kinematic features of 4 of such input jets corresponding to a background event. The preparation of input data is depicted in Figure 4.2 and done in the following way:

1. Jets coordinates are momentarily transformed from $\{p_T,\ \eta,\ \phi,\ m\}$ to $\{p_x,\ p_y,\ p_z,\ E\}$ representation.
2. All the six possible pairs of dijets are formed, by adding the single jets four-vectors. Pairings are 1-2, 3-4, 1-3, 2-4, 1-4 and 2-3.
3. From the constructed dijets, quadjets are formed in a similar way, adding together the dijet four-vectors with indices 1-2, 3-4 and 5-6.
4. Lastly, the logarithm is taken on the p_T and mass variables. This is done in order to compress the distributions, which present long tails, with the objective of easing the training of the network.

There is an important note, however, that should be made on the *symmetrization* of the input. While the results that will be presented in Chapter 5 have been obtained with an input modified only with the prescription detailed above, several tests have been carried out with such symmetrization, which is explained down below.

It is important to note, beforehand, that reproducing the exact values of $\{p_T,\ \eta,\ \phi,\ m\}$ for all four jets is not necessary for a comprehensive event characterization: one only needs to retain the same physical information, which remains invariant if the event is rotated in all three spatial directions. Although this rotation may result in entirely different $\{p_T,\ \eta,\ \phi,\ m\}$ values, the physical meaning remains unchanged. This symmetrization can be achieved by rotating all events in input, in a way that these rotational symmetries are eliminated, and the space of possibilities of $\{p_T,\ \eta,\ \phi,\ m\}$ representing the same event reduces to a single combination:

- To make the autoencoder output invariant under $\eta \to -\eta$ rotations, all four-vectors can be flipped to the frame where the leading, i.e. the one with highest p_T, jet η is positive.

- Further degeneracies under ϕ rotations are addressed by setting the leading jet ϕ to 0 (and rotating the other three jets accordingly) and setting the subleading jet ϕ positive. This ensures that the leading jet has no p_y component (therefore all the transverse momentum is in the x direction) and its p_z component is positive defined. It is not difficult to see that once the leading jet four-momentum vector is fixed, one can still rotate the event within the plane defined perpendicularly to such four-momentum vector. By rotating the three remaining jets' ϕ by an angle of π in the cases where the subleading jet ϕ is negative, one achieves the desired results.
- While the transformations above are maintained in the input features during the computation of the loss (which is what the autoencoder is enforced to learn) one can also make the architecture intrinsically invariant under arbitrary ϕ rotations. To do this, one can replace all azimuthal information with relative angular information: the jet ϕ coordinates are replaced with $\Delta\phi(j_m, d_{m,n})$ i.e. the relative angle with the dijet to which the jet belongs, dijet ϕ coordinates are replaced with $\Delta\phi(d_m, q_{m,n})$ i.e. the relative angle with the quadjet to which the dijet belongs, and the remaining global ϕ information from the quadjet four-vector can be removed.

The *symmetrization* was originally carried out with the expectation that it would substantially facilitate the training process. However, our findings revealed that the autoencoder exhibited significantly better performance when the elimination of symmetries was not implemented; removing global η sign information and performing the aforementioned ϕ rotations is a strong inductive bias[1] motivated by the symmetry of the detector and colliding beams, but when the task at hand is encoding and decoding input, these transformations should be avoided. The underlying reason is that the absence of a particular feature in the input data poses a significant challenge for the autoencoder to accurately reproduce it in the output. The *symmetrization* procedure entails transformations that effectively lead to such removal of information (individual jet ϕ coordinates, for instance) which, in turn, hinders the autoencoder's capacity for precise reconstruction.

4.3 Embedding and ghost batch normalization

The autoencoder employs convolutional neural networks [84], which owe their success to two fundamental properties: they implement intrinsically the property that at each layer only local features are relevant and they use the same set of weights on each local group of pixels. The network is thus forced to learn such features in a hierarchical way, being able to generalize from the most local features to the most general ones, as the layers of convolutions are applied. In particular, layers near the bottleneck become localized representations of complex non-local features of the input.

[1] Inductive bias refers to the set of assumptions or prior knowledge that a machine learning algorithm incorporates to guide its learning process. It reduces overfitting and helps the algorithm make predictions or generalizations from a limited amount of data [83].

The first part of the autoencoder architecture can be defined as an *embedding*. Embedding is a procedure where physical values of the jets are translated into *pixels*, rather abstract objects that contain the physical information in a non-trivial manner, but which are more practical for the model's functioning. To perform this embedding, three one-dimensional (1D) embedding convolutions of stride and kernel[2] one are applied to the jet, dijet, and quadjet input features to project them all into $\mathbb{R}^d$ (red triangles in Figure 4.3). The dimension d has no physical meaning and can be freely chosen. In our case $d = 16$ was set, keeping the input dimensions intact throughout the whole encoding process, to finally step it down to a dimension of 6 in the bottleneck. Bear in mind that this is a large attempted compression, since the input information dimension is 16. To complete the input feature embedding, a non-linear activation, the sigmoid-weighted linear unit (SiLU) [85], is applied, followed by a final set of three single pixel convolutions, completing the embedding. The input embedding block is shown in Figure 4.3.

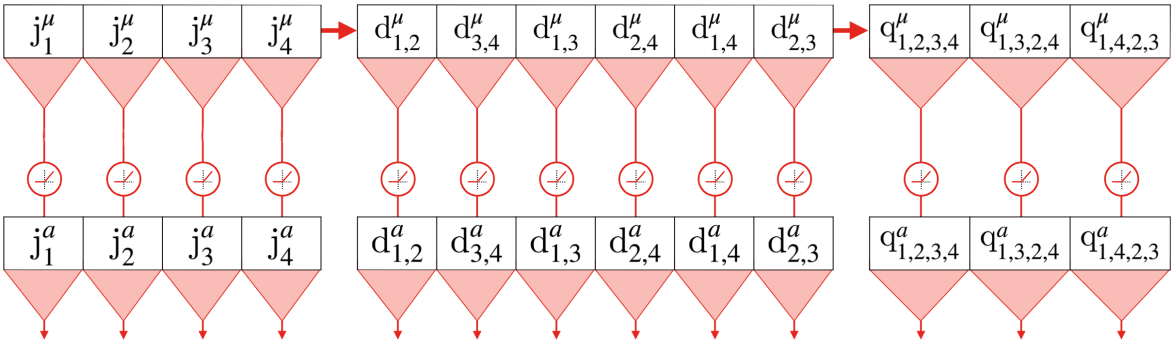

Figure 4.3: From [78]. Input embedding block. Red triangles represent 1D convolutions. Circles with the kinked line inside represent the application of SiLU activations.

Typically, immediately before applying convolutional layers, one has to normalize the input to have zero mean and unit variance. This normalization is done with each subset of the events on which the model is training at each time, which is called a *batch*. Batch Normalization (BN) has always been found to improve stability of training, accelerating convergence and reducing the covariate shift [86]. In this work, however, we do not perform simple batch normalization but, instead, a variation (see Algorithm 1) of the Ghost Batch Normalization (GBN) originally proposed in [87] is implemented. GBN works by splitting the training batches into smaller *ghost batches* and applying BN to them. This allows training with very large batches, while also keeping at each layer the regularization that BN provides. We start with batches of 1024 events that are subdivided into 64 ghost batches of 16 events each. The batch size is increased by a factor 2, and the number of ghost batches is divided by 4 after epochs 1, 3, 6, and 10.

[2] The kernel is the number of pixels in an image that are convolved in a single application of the convolution. The stride, on the other hand, is the number of pixels by which the convolution filter is moved after each application.

Algorithm 1: From [78]. Implemented version of Ghost Batch Normalization (GBN) [87], which is applied to the input of each convolutional layer. During training, batches are subdivided into smaller batches, called *ghost batches*, before shifting and normalizing them to have zero mean and unit variance.
The number of ghost batches N_g must evenly divide the size of the training batch. The convolution kernel and stride are assumed to have the same value N_s. The stride must also divide the number of pixels N_p in each input image into the number of kernel applications per event $N_k \equiv N_p/N_s$. The running mean μ and standard deviation σ for use during inference are updated at each training step with the average over the ghost batch means and standard deviations using momentum η.
By default $\eta = 0.9$ and the trainings are started with batches of 1024 events. These are split into $N_g = 64$ ghost batches of 16 events.

```
Input: $X[\text{event, feature, pixel}]$
$S_X,\ N_f,\ N_p = X.\text{shape}$
$N_k = N_p/N_s$
$X \leftarrow X.\text{transpose}(1,\ 2)$          /* Transpose pixel and feature indices */
if $training$ and $N_g\ != 0$ then
    /* Split the training batch into ghost batches.  The number of ghost
     batches is $S_X/N_g$ times the number of kernel applications $N_k$.  The
     factor $N_k$ ensures $\mu$ and $\sigma$ are computed over all objects rather than
     separately for each object in the event                              */
    $S_g = N_k \times (S_X/N_g)$
    $X \leftarrow X.\text{view}(N_g,\ S_g,\ N_s,\ N_f)$
    /* Compute the means and standard deviations of each ghost batch      */
    $\mu_g = X.\text{mean}(\text{dim} = 1)$
    $\sigma_g = X.\text{std}(\text{dim} = 1)$
    $X \leftarrow (X - \mu_g)/\sigma_g$
    /* Update running $\mu$ and $\sigma$ with their means over the ghost batches.  It
     is critical to detach $\mu_g$ and $\sigma_g$ from the gradient backpropagation
     calculation at this stage to prevent the gradient of this batch from
     depending on the gradient of all previous batches                    */
    $\mu_X = \mu_g.\text{detach}().\text{mean}(\text{dim} = 0)$
    $\sigma_X = \sigma_g.\text{detach}().\text{mean}(\text{dim} = 0)$
    $\mu \leftarrow \eta\mu + (1-\eta)\mu_X$
    $\sigma \leftarrow \eta\sigma + (1-\eta)\sigma_X$
else
    $X \leftarrow X.\text{view}(S_X,\ N_k,\ N_s,\ N_f)$
    $X \leftarrow (X - \mu)/\sigma$
/* Rearrange $X$ for application of the convolution                       */
$X \leftarrow X.\text{view}(S_X,\ N_p,\ N_f)$
$X \leftarrow X.\text{transpose}(1,\ 2)$
$X \leftarrow \text{Convolution}(X)$
```

4.4 Autoencoder architecture outline

Deep networks have been studied for a few years now, yet they have seen massive development in recent times. In particular, one of the main issues they presented is that the deeper the network was, the lower the performance and generalization capabilities, and the longer the training times [88]. However, in 2016, it was shown that learning *residual* features at each layer rather than direct convolutions dramatically reduced network training times, even allowing networks of more than hundreds of layers to quickly converge [89]. The idea is that instead of learning a non-linear function $x \to f(x)$, one could instead learn non-linear residual features added to the input $x \to x + f(x)$. This can be seen as learning a hierarchy of *deformations*, rather than a hierarchy of features. These *residual networks* or ResNets have been employed in the autoencoder architecture for such reason, and justify the way the embedding block works. In particular, one may notice that the three quadjets obtained in the embedding procedure, as detailed in Section 4.3, all represent the same physical four-vector. This seems obvious from the point of view of the commutativity of the sum of the individual jets' four-vectors. However, the differences among them arise from the order in which the underlying dijets and jets have been paired. This is particularly useful, since the network is therefore able to learn which pairings are more physically relevant and assigns proper deformations to account for these characteristics.

We use residual learning to convolve the embedded jets, dijets, and quadjets down to an encoded feature space, and then to upsample from the encoded feature space to a full reconstructed event. We present, in the following sections, the encoder and the decoder architectures.

4.4.1 Encoder

The encoder is the first part that processes the input and encodes the jets' features into the encoded space, which we will call z. In this work, z has been chosen to be six-dimensional as a trade off between a number of meaningful dimensions and a reasonably accurate reproduction of the input features. The architecture of the encoder is detailed in Figure 4.4 and can be described as the following encoding path:

1. First, the physical jets $\mathrm{j}^{\mu}_{1,2,3,4}$ are embedded as described in Section 4.3 to produce 12, 6, and 3 pixels of jets, dijets, and quadjets, respectively; all of them containing 16 features for each pixel.

2. We define two GBN convolutions of kernel 2 and stride 2, highlighted in blue in Figure 4.4: *jets_to_dijets* and *dijets_to_quadjets*. The convolution *jets_to_dijets* is applied to the 12 pixels of the embedded jets, and produces a pixel for each pair of input pixels, creating a feature map of 6 pixels, analogous to that of the corresponding embedded dijets. After applying the SiLU activation, they are summed to the original embedded dijet features to produce the feature map defined as the yellow rectangle "d" in the

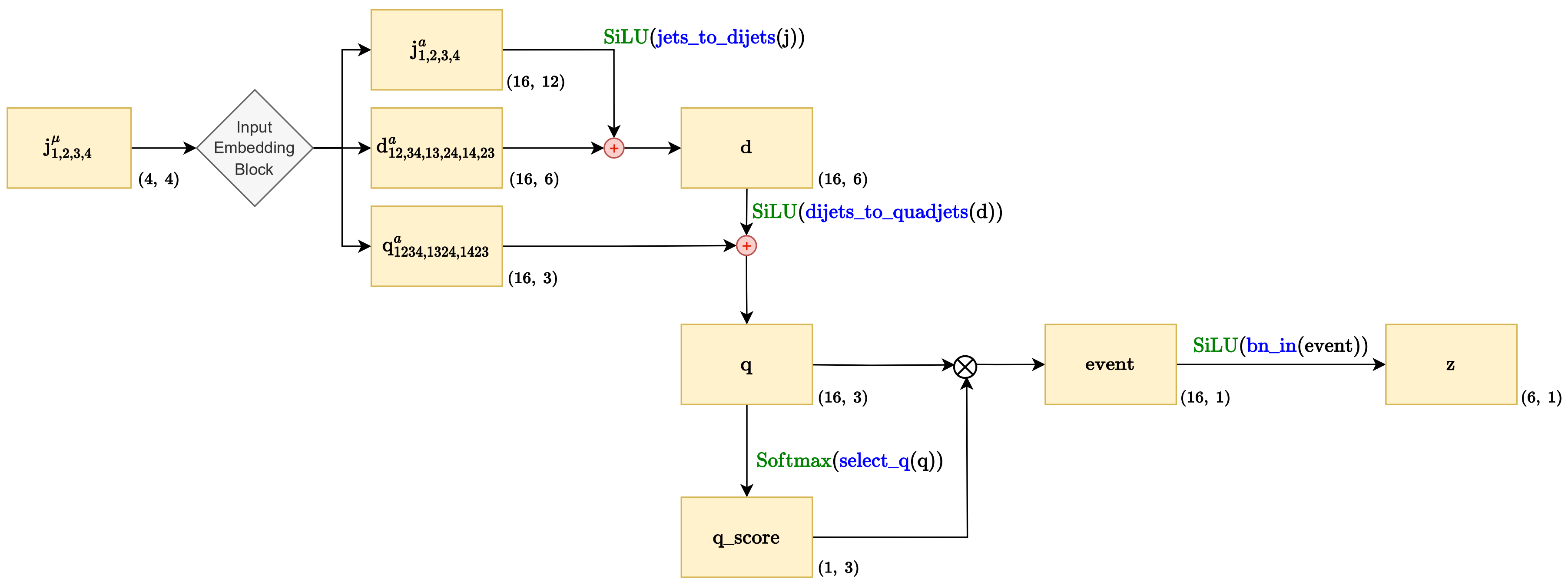

Figure 4.4: Encoder architecture. The yellow rectangles represent feature maps in the successive layers of the encoder. The gray square represents the input embedding block depicted in Figure 4.3, while the blue text accompanying the directional arrows stand for GBN convolutions. Green text is used for representing non-linear activations. The red circles next to the embedded features' rectangles are to indicate element-wise sum of feature maps, while the direct product symbol joining "q" and "q_score" stands for matrix multiplication. Additionally, the dimensions of each feature map (with the exception of the batch size) have been added by each feature rectangle, in order to give a taste of how the information is progressively reduced.

architecture scheme. The convolution *dijets_to_quadjets* works similarly, computing the “deformations” for the “d” feature map and being then added to the original embedded quadjet features, also after the activation is applied.

3. The previous step produces a map of 16 features of quadjet-level information, each one with 3 pixels corresponding to each of the possible pairing histories. As mentioned in the introduction, this is crucial, as it allows the network to keep track of the pairing combinations along the encoding procedure. The residual feature learning in this step also allows the network to encode the information in a layered manner, adding the proper deformations to each dijet first, and then to each quadjet pixel. Once “q” is obtained, one can perform another GBN convolution *select_q*, which essentially *selects* one of the feature maps from the 16 present in q. Successively, the *softmax* function

$$\text{Softmax}(x_i) = \frac{\exp{(x_i)}}{\sum_j \exp{(x_j)}} \tag{4.4}$$

is applied to produce what is called “q_score”, effectively producing a feature that gives a *score* to each pixel of the selected q feature, with all the three scores summing 1.

4. The feature map “q_score” is then transposed and the matrix product “event” of “q” and “q_score” is obtained. This is nothing but the sum of the three quadjet pixels weighted by the score of the selected feature, across all 16 features. The feature map “event” becomes then one-dimensional for each feature.

5. Lastly, a final convolution “bn_in” and a SiLU activation is applied to the event-level features pixel. This finally encodes the information into the bottleneck in a six-dimensional vector z.

4.4.2 Decoder

The decoder is the second half of the autoencoder and takes the encoded features vector z as input to produce in output the reconstructed jets four-vectors. The decoding path aims at being as symmetric as possible with respect to the encoding, although it is impossible to obtain a complete symmetry given the lack of initial information stored in z. In particular, one could try to obtain embedded dijets and quadjets from the pseudo-event, but doing so directly will result in convolutions promptly passing from a (16, 1) to (16, 12)-dimensional vector, resulting in a large number of dimensions being empty or not containing much information. The architecture of the decoder is shown in Figure 4.5 and can be described as the following decoding path:

1. Starting from the encoded vector z, a GBN convolution of kernel and stride 1 *bn_out* and a SiLU activation are applied. This exits the bottleneck space and produces a feature map “pseudo-event” that aims at containing the same kind of information present at the “event” feature map in the encoding path.

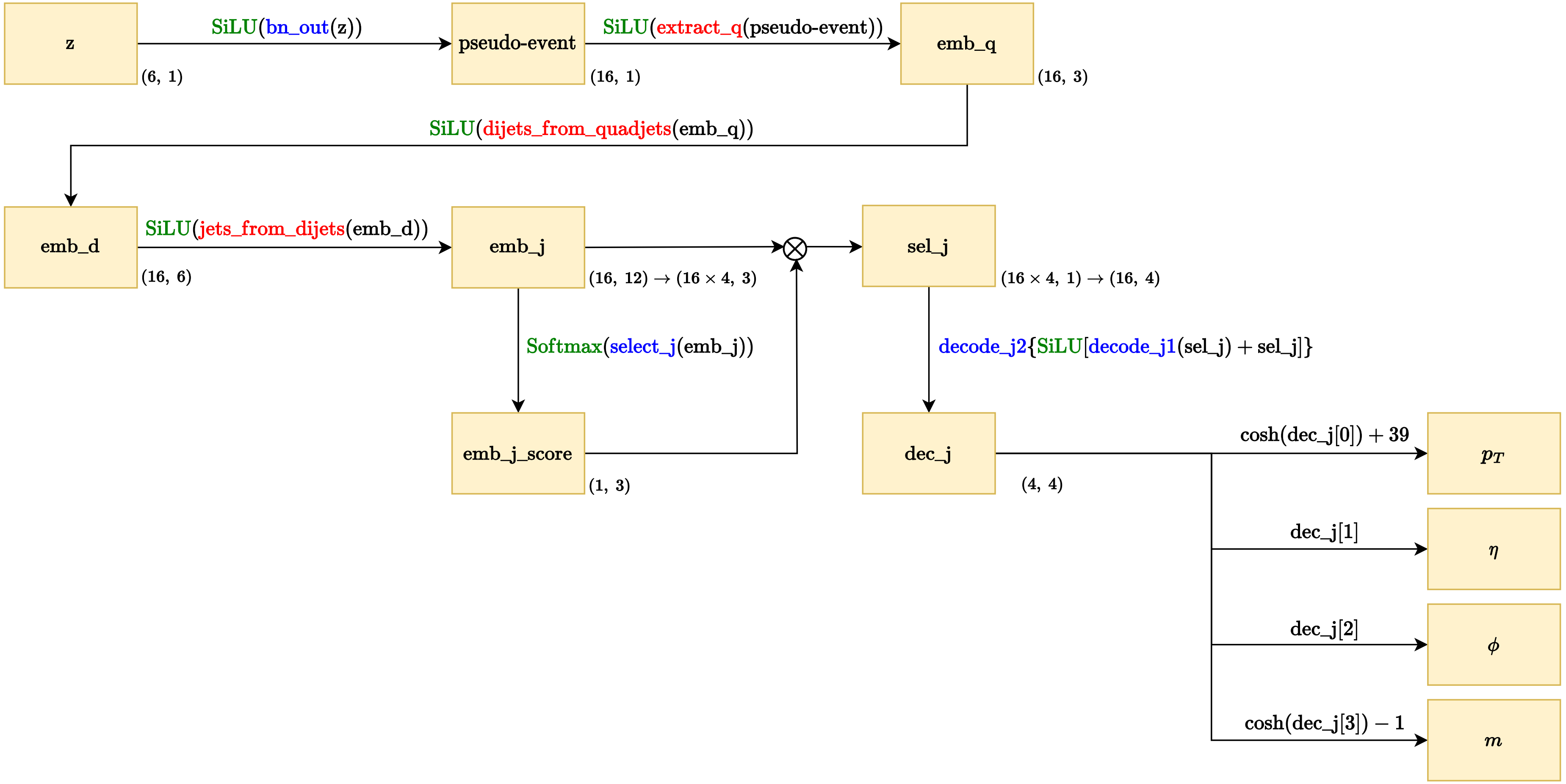

Figure 4.5: Decoder architecture. The yellow rectangles represent feature maps in the successive layers of the decoder. The blue text stands for GBN 1D convolutions with kernel and stride 1, while the red text signifies GBN transposed 1D convolutions. In particular, *extract_q* is a kernel and stride 3 transposed 1D convolution, while *dijets_from_quadjets* and *jets_from_dijets* are kernel and stride 2 transposed 1D convolutions. Green text is used for representing non-linear activations. The direct product symbol joining "emb_j" and "emb_j_score" stands for matrix multiplication. Additionally, the dimensions of each feature map (except for the batch size) have been added by each feature rectangle, to give a taste of how the information is progressively augmented. When there is an arrow pointing to a different dimension, it is to indicate that the feature map is promptly reshaped after being obtained.

2. From the pseudo-event pixel, a transposed convolution[3] with kernel and stride 3 obtains 3 embedded quadjet pixels: "emb_q". These are intended to contain the quadjet pairing history information, just as "q" in the encoding path.

3. Subsequently, we obtain "emb_d" and "emb_j" with two kernel and stride 2 transposed convolutions: *dijets_from_quadjets* and *jets_from_dijets*. The obtained feature map is "emb_j", which contains 12 pixels for each of the 16 features.

4. The feature map obtained in the previous step is promptly reshaped to a $(16 \times 4,\ 3)$ shape. This means that each jet, for each of the 3 quadjet pairings, contains 16 pixels of information. This is not completely accurate, as the information between jets is now shared across the same dimension, but the idea is that these pixels represent a set of complete information on the 4 jets, and the fact that they are separated into the 3 possible pairing histories allows us to compare which are the most significant for the network.

5. This comparison is done in a similar way to how we computed the score "q_score" in the encoding. First, we apply the GBN convolution *select_j*, which effectively selects one of the 64 features of 3 pixels, and then apply a softmax activation to turn it into a score. This feature contains 3 numbers that are to be thought of as the scores of each of the three pairing histories. These are multiplied by the 64 features contained in "emb_j", to produce a feature map that is $(16 \times 4,\ 3) \otimes (3,\ 1) = (16 \times 4,\ 1)$-dimensional.

6. The obtained feature map is called "sel_j", and from the way it was constructed, it contains 64 numbers that can be thought of as a set of 16 numbers corresponding to each of the 4 jets. It can be thus reshaped into a (16, 4)-dimensional vector. Once this is done, two GBN convolutions of stride and kernel 1 are applied on top of each other with an intermediate SiLU activation and a residual addition to obtain "dec_j", the final decoded jets.

7. The kinematic features of the jets are the rows of "dec_j", and to improve the expressiveness of the network, hyperbolic cosines are taken on the first row (p_T) and last (m). Additionally, 39 GeV are summed to the p_T to ensure that it is larger than 40 GeV ($\cosh(x) \geq 1\ \forall x$), and 1 GeV is subtracted from the mass for it to be strictly positive defined.

Before defining the final layout presented in Figures 4.4 and 4.5, this network has withstood a large number of modifications throughout this work, as we experimented with different numbers of layers, types of activations, training times, and many other features. The choice of decoding the $\{p_T,\ \eta,\ \phi,\ m\}$ representation is based on practical motivations since the

[3] Transposed convolutions are standard convolutions with a modified input feature map, with the idea of carrying out a trainable up-sampling. Transposed convolutions use padding (filling the input borders with zeros) to obtain larger dimensions in output. For a guide on how to visualize these arithmetic operations, we invite the reader to consult [90].

decoding of $\{p_x,\ p_y,\ p_z,\ E\}$ has seemed to be less straightforward for the autoencoder. We believe that this could be induced by the difference in order of magnitude between p_x, p_y ($\sim$ 300 GeV) and p_z, E ($\sim$ 1000 GeV). Also, different activations should be applied if one was to produce these values in output since they are substantially larger than the general weights of the network, and the general architecture should also be tuned in such a case. In the following section, we present the general characteristics of the training and hyperparameters (which are the user-imposed, i.e. non-learnable, parameters) of the network.

4.5 Background sample, training, and loss

The network is trained on a background sample of 4 b-tagged QCD jets. The events have been generated with the MadGraph software [91]. The measured jet energies are smeared according to the CMS calorimeters' energy resolution, and the jets' masses are set to 0 for simplicity. This approach is adopted due to the absence of a detailed simulation of quark showering, hadronization processes inside the calorimeters, and detector response. The simplest solution is then to set the masses to 0, rather than trying to emulate the true mass distribution. For more details about how this background sample has been obtained, we refer the reader to [20, Section 6.1]. The sample contains a total of 2,202,261 events with four b-jets, which is split into three samples with 734,087 events each. Two of the three samples are used for training with the other one reserved for validation. We cross-validate the results by training three times changing which third is used for validation in a process known as k-folding. The event number i modulo 3 specifies the validation set for each k-fold using an offset in {0, 1, 2}:

$$\text{event}_i \in \begin{cases} \text{Validation Set} & \text{offset} \equiv i \mod 3 \\ \text{Training Set} & \text{offset} \not\equiv i \mod 3 \end{cases} \tag{4.5}$$

The result for event i is therefore obtained from the model with offset $\equiv i$ mod 3. This way, the final distributions are obtained without biases from the training set.

The training consists of 25 epochs for each of the three models and its schedule is inspired by [92], which suggests an inverse relationship between batch size and learning rate. The initial learning rate is set at 0.01 and is divided by 4 in epoch 10. After epoch 15, it is divided by 4 after every epoch until the end of the training. The training batch size is initially set at 1024 events and is doubled at epochs 1, 3, 6, and 10. The inference batch size is set at 16,384 and does not change throughout training. The model is implemented in PyTorch v2.0.1 [93] and trained with Adam [94] optimizer (other parameters left at default values).

The implemented loss function is the MSE loss between $\{p_x,\ p_y,\ p_z,\ E\}$ representations of input and output jet features. A loss weighting factor of 0.3 is used to downsize the losses on p_z and E since they are significantly larger than p_x and p_y. The final loss for N events is obtained as the weighted sum of the square root of MSE losses for each event:

$$\text{Loss} = \frac{\sum_n^N \omega_n \cdot \sqrt{\sum_p^{p_x,p_y,p_z,E} \sum_j^{4\text{ jets}} \mathcal{L}_{n,p,j}^{\text{MSE}}(\text{input jet, output jet})}}{\sum_n^N \omega_n} \tag{4.6}$$

The reason why the $\{p_x,\ p_y,\ p_z,\ E\}$ representation is used for the computation of the loss lies in the shape of the distributions. While we are not interested in reconstructing the mass since it is also not given in the input, the reconstruction of η and ϕ carries problems on its own. In particular, one has to be very careful with the topology of the ϕ output values and perform corrections to address the azimuthal symmetries the autoencoder is not aware of (see the first paragraph of Section 4.2). Additionally, the η distribution ends sharply at $(-2.5,\ 2.5)$ due to the lack of angular coverage of the CMS detector. This is an artificial effect, of course, since the angular distribution of jets produced does not simply falls to 0 at these values. As such, the autoencoder has trouble acknowledging it, and values of η produced in a smooth way beyond these limits are highly penalized by the MSE loss, which consequently incentives the autoencoder to over-centralize the distribution. The $\{p_x,\ p_y,\ p_z,\ E\}$ representation does not present these issues, and output values that are Gaussian-distributed around the input quantities represent more accurately the true distributions.

Chapter 5

Results

In this chapter we present, at last, the results obtained with our autoencoder approach. In the following Section 5.1, we present the general results of the autoencoder training, and the reconstructed kinematic features of the input jets. We also remark on some important aspects of the results, and why they are useful to guide the following steps in the investigation of autoencoders applied to background modeling.

In Sections 5.2 and 5.3, we outline the main results obtained for the decoding and the generation of an artificial dataset, respectively. We use a well-known metric to measure the agreement between one-dimensional distributions, while a dedicated figure of merit is built to address the compatibility of two-dimensional invariant mass distributions. The summary of results is also gathered in the final Section 5.4.

5.1 General results and Wasserstein distance

For completeness, we present in Figure 5.1, the training and validation losses obtained according to Equation 4.6, and whose units are GeV. Each of the three colors represents the three different models trained, each one withholding one different third of the dataset for validation. We observe no significant differences in the training among the three models, with the models with offset 0 and 2 obtaining slightly lower losses than the model for offset 1. The solid lines indicate the trend of the training loss, whilst the dashed line does it for the validation loss. It is interesting to note that throughout the training of the three models, these two are practically identical. This is expected from the point of view that the only difference between the training and validation splits is the event indices and so the performance should be similar for both.

Once the training is done, we have access to the value of the activations in the bottleneck, which are essentially 6 numbers that encode the information of the 16 (12 if we do not count the jets' masses) initial physical features. The cumulative distributions of such activations are depicted in Figure 5.2 with the form of the six one-dimensional marginal distributions. It is important to note that each of these distributions accumulates the activations across

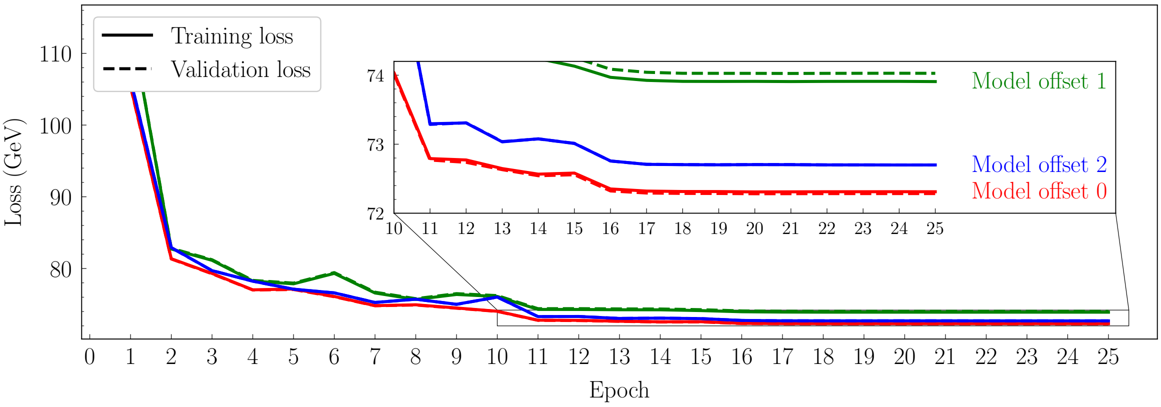

Figure 5.1: Training (solid lines) and validation (dashed lines) losses for the models with offset 0 (red), offset 1 (green) and offset 2 (blue). The inset shows a zoom of the region from epoch 10 up to epoch 25.

the three offsets, and each of the three distributions for each marginal can be profoundly different, as can be observed in Figure A.4 in the Appendix A.

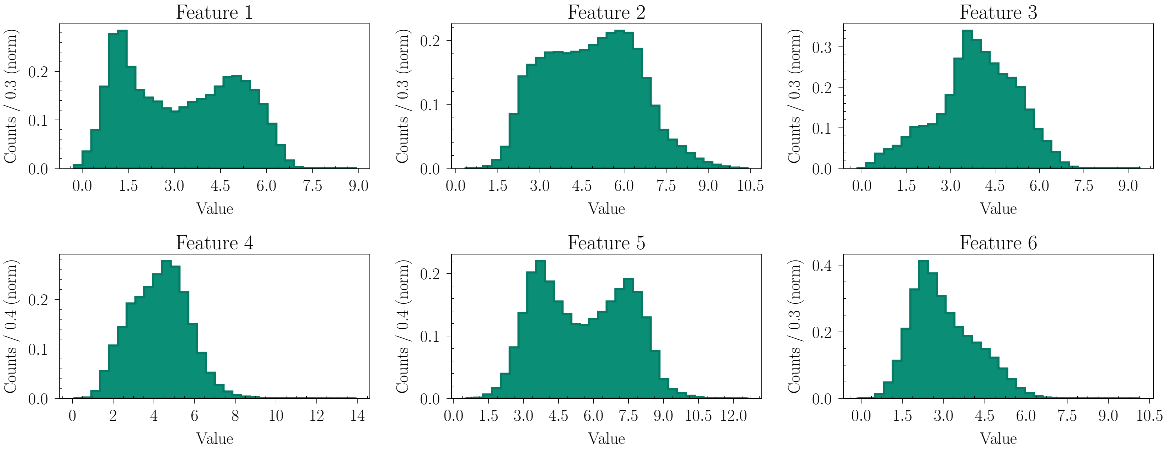

Figure 5.2: Marginal one-dimensional distributions of the activations in the bottleneck, cumulative for all the three models.

After the training is done, one can obtain the decoded dataset, which is the procedure of feeding the trained decoder with the obtained activations. If the training is reasonably good, the decoder output should be a similar dataset to that of the input. In order to quantify the similarity between the output and the input variables, we will use a metric known as the Wasserstein distance [21, 22]. The Wasserstein metric measures the difference between two distributions by the optimal cost of rearranging one distribution into the other. The mathematical definition of the Wasserstein 1-distance between two distributions $f : X \to \mathbb{R}$, $g : Y \to \mathbb{R}$ can be written as:

$$W_1(f, g) = \inf_{\pi \in \mathcal{M}(f, g)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| \, d\pi(x, y). \tag{5.1}$$

In the equation above, $\mathcal{M}(f, g)$ is the complete set of distributions on $\mathbb{R} \times \mathbb{R}$ whose marginals are f and g. In other words $\mathcal{M}$ holds the set of all maps that rearrange f into g. The Wasserstein distance is therefore the distance between f and g computed within the map that ensures W_1 is minimal. In the following Sections 5.2.1 and 5.3.1, we will provide the Wasserstein distance between the obtained one-dimensional distributions to quantify the similarity between them. Naturally, the Wasserstein distance between p_x, p_y p_z and E distributions has units of energy, and can be directly thought of as an energy cost for transforming one set of physical events into the other.

5.2 Decoded dataset

The decoded dataset is obtained in inference by decoding the bottleneck features in Figure 5.2. The result is a set of 2,202,261 (same size as the input) reconstructed events with four b-jets. In Section 5.2.1 we present the comparison between the obtained one-dimensional distributions, whereas Section 5.2.2 depicts the two-dimensional histogram between the invariant masses of reconstructed quadjets and dijets, and its comparison with the true distribution. This comparison is useful as it conveys the accuracy of the reconstruction of invariant masses, which are, in turn, key variables to discriminate background events from Higgs boson candidate events. A good reproduction of these distributions is strictly linked to a good reproduction of background templates.

5.2.1 One-dimensional distributions

The reconstructed distributions for p_x, p_y, p_z and E are shown in Figure 5.3. The true (or input) distributions are drawn as a solid red histogram, while the reconstructed events are shown as a dotted blue histogram. The p_x and p_y distributions have a distinctive bimodal shape, peaking at approximately ±40 GeV, as a result of the p_T threshold required in the event simulation. The reconstructed events, remarkably enough, are able to capture this feature and reproduce it quite accurately after the bottleneck. It is also worth noting that the correspondence between reconstructed and true distributions stays reasonably high through the entire range for all of the 4 kinematic quantities.

We observe, nonetheless, some systematic effects in all four distributions from Figure 5.3 that must be commented on. In particular, we observe an excess of reconstructed events around the bin containing the highest number of true events. The residuals, defined as the number of reconstructed events minus the number of true events, all divided by the number of true events, show thus positive values for p_x, p_y close to ±40 GeV, p_z close to 0 and E close to 100 GeV. This is a consequence of the construction of the MSE loss, that naturally tends to produce values that are Gaussian-distributed: when the autoencoder lacks information, an output set of features produced close to the mode is in average less penalized than if it is produced in the tail of the distribution. This accumulation of events is the cause, in turn, of residuals lower than 0 across a large range of the distributions. The Wasserstein distance

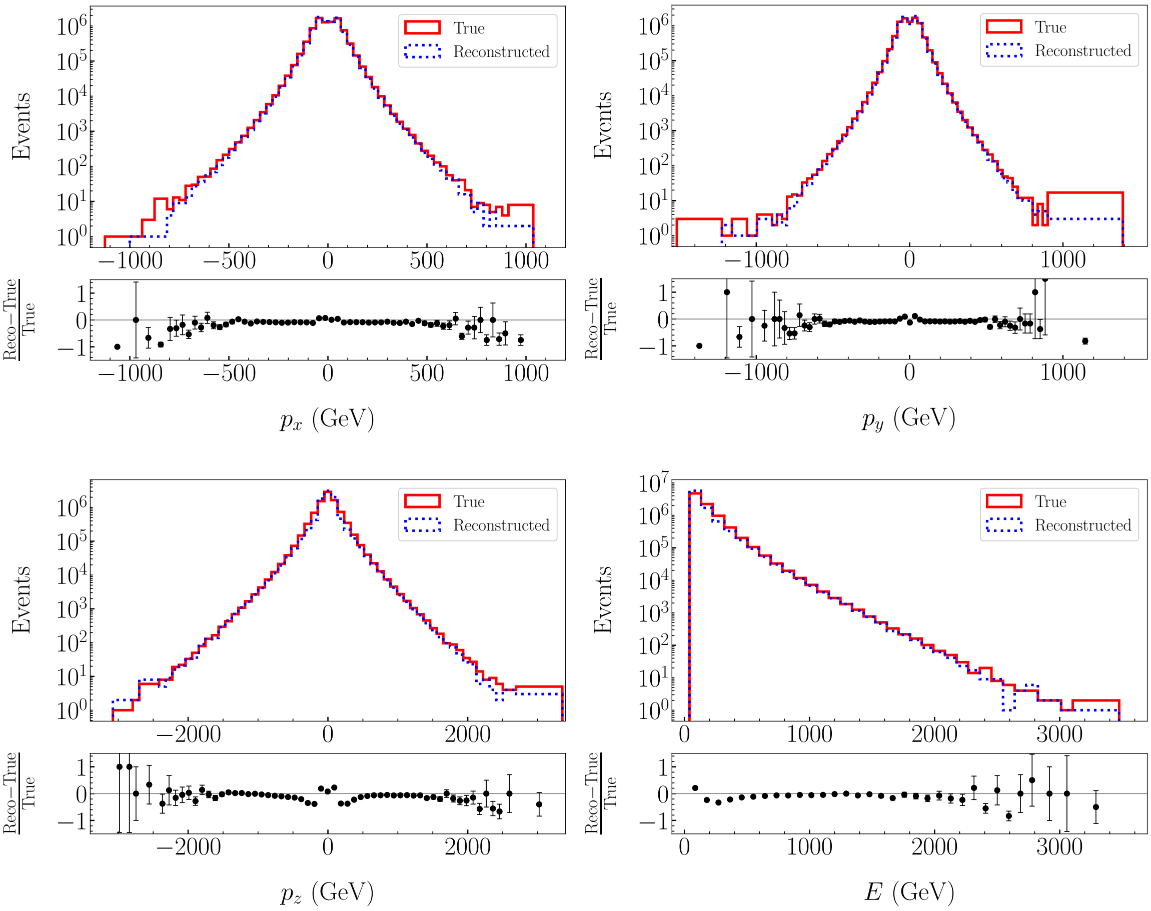

Figure 5.3: True (solid red histogram) and reconstructed (dotted blue histogram) distributions for p_x (top left), p_y (top right), p_z (bottom left) and E (bottom right) of the 4 b-jets. Below every distribution comparison, the ratio of reconstructed minus true events divided by true events is shown.

between each pair of distributions, true and reconstructed, is given in Equation 5.2:

$$\begin{aligned} W_1^{\mathrm{reco}}(p_x) \approx 2.31\,\mathrm{GeV}, \quad W_1^{\mathrm{reco}}(p_y) \approx 2.13\,\mathrm{GeV}, \\ W_1^{\mathrm{reco}}(p_z) \approx 18.70\,\mathrm{GeV}, \quad W_1^{\mathrm{reco}}(E) \approx 18.82\,\mathrm{GeV}. \end{aligned} \tag{5.2}$$

As already mentioned, by the defintion of the Wasserstein metric, one may interpret this result as the amount of energy one should "put" into one of the distributions to obtain the other. If we recall that these distributions contain more than two million events, the amount of energy one should shift in a reconstructed distribution is, in the four cases, less than 10 keV per event.

5.2.2 $m_{2\mathrm{j}}$ vs $m_{4\mathrm{j}}$

The two-dimensional distributions $m_{2\mathrm{j}}(m_{4\mathrm{j}})$ convey a large amount of information in searches for Higgs boson pair production. For a final state with four jets, in particular, they show the invariant masses for each of the 6 possible dijet pairings as a function of the quadjet invariant mass. This is useful as the Higgs boson candidates should peak at a dijet invariant mass of approximately 125 GeV, while the other 4 dijet pairings would blend into background. The background is, unfortunately, also maximal in these regions, making it difficult to distinguish the signal from the background. In Figure 5.4 (top left), the true (or input) distribution is shown, where each bin is colored as a function of the number of events contained in it. The shaded region marks the area with no physical meaning, since there $m_{2\mathrm{j}} > m_{4\mathrm{j}}$. In Figure 5.4 (top right), we show the equivalent distribution for the reconstructed events, plotted with the exact same binning.

Since the binning is exactly the same in both figures, one can produce a third histogram shown in Figure 5.4 (bottom left), where each bin content is the difference between the number of events in the reconstructed distribution (top right) and the true distribution (top left). Blue bins indicate regions in which the true distribution contains more events, whereas red bins indicate the opposite, being the magnitude of the difference encoded as color intensity. It is clear that also here some systematic effects are apparent. In particular, the reconstructed distribution seems to be biased toward the region where $m_{2\mathrm{j}} \sim m_{4\mathrm{j}}$, which means, in turn, that the true distribution has a larger presence of jets produced back-to-back than it is being reconstructed. In order to better visualize the systematic differences between the two distributions, one can construct the bin-by-bin percent difference:

$$\Delta_{ij} \equiv \frac{h_{ij}^{\mathrm{reco}} - h_{ij}^{\mathrm{true}}}{h_{ij}^{\mathrm{reco}} + h_{ij}^{\mathrm{true}}}. \tag{5.3}$$

In the equation above, the figure of merit Δ_{ij} is the difference in events in reconstructed and true distributions, divided by their sum for the given ij-th bin. This distribution is shown in Figure 5.4 (bottom right) where the differences for the bulk of the data ($m_{2\mathrm{j}}$, $m_{4\mathrm{j}} < 1000$ GeV)

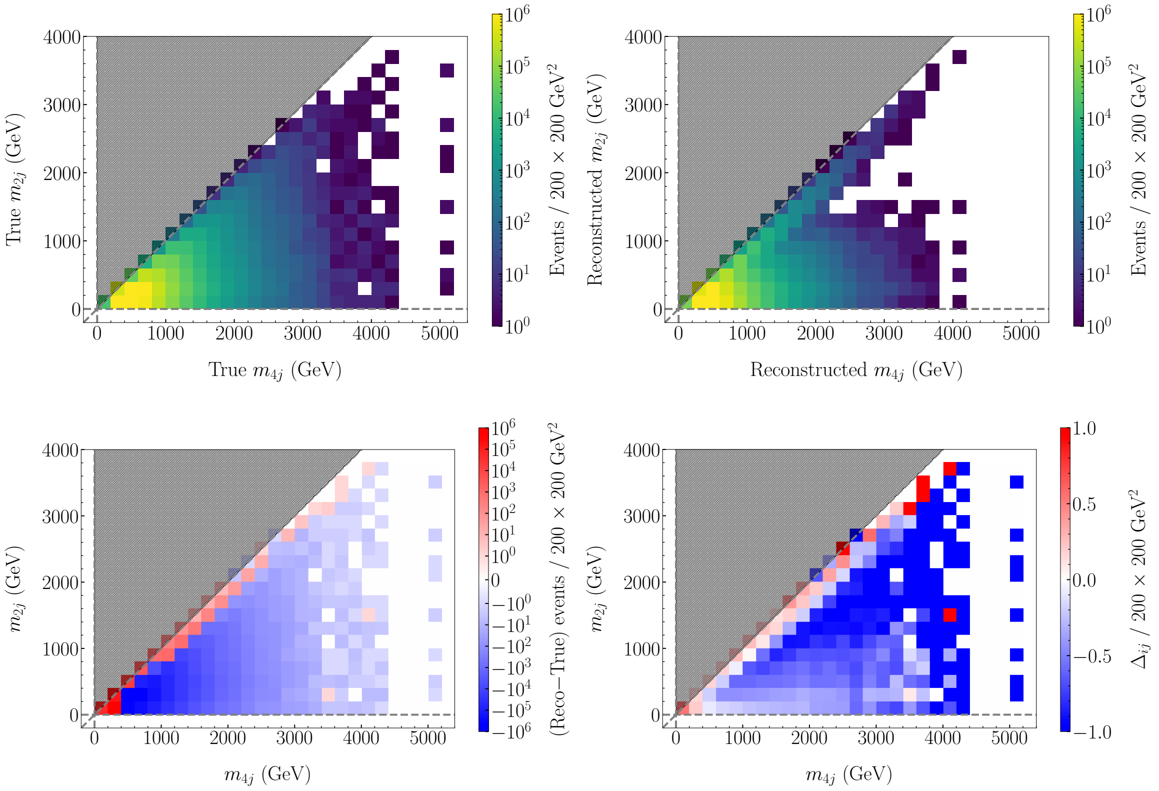

Figure 5.4: Two-dimensional distributions of true (top left) and reconstructed (top right) dijet invariant masses as a function of the quadjet invariant mass. The bin-by-bin difference between the reconstructed and true distributions is shown in the bottom left, whereas the bin-by-bin percent difference defined in Equation 5.3 is shown in the bottom right.

are smeared and it is the tails where we find the largest percentage differences, as one could expect.

One could also calculate the total sum of these differences, i.e. $\sum_{ij}^{N\text{ bins}} \Delta_{ij}$, and divide it by the number of bins, i.e. $N = 47$ in our case. This provides the result:

$$\Delta_N \equiv \frac{\sum_{ij}^{N\text{ bins}} \Delta_{ij}}{N} \approx -2.80, \tag{5.4}$$

which could be interpreted as the average deviation per bin. The result cited above indicates a prominence of true events in more bins than those in which reconstructed events dominate. Once again, this is a systematic effect of events being reconstructed in very localized regions of the phase space, giving rise to few overcompensations in exchange for a general underestimation of the rest of the regions.

5.3 Generated dataset

One great advantage of the autoencoder is that it can be used to generate data once the encoded space is a good representation of the input data. By randomly sampling the distribution in the encoded space, one could decode an entirely new dataset, with similar features to that of the input. The first attempt to perform this sampling on the encoded space was carried out by means of a Gaussian Mixture Model (GMM) [95]. The GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities, created to represent a complicated –or unknown– distribution. The GMM-estimated probability can be expressed as:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} \omega_i \cdot g(\mathbf{x}|\mu_i, \sigma_i). \tag{5.5}$$

In the equation above, $\mathbf{x}$ is a one-dimensional vector representing the encoded space features' values, ω_i is the weight for the i-th component in the mixture and $g(\mathbf{x}|\mu_i, \sigma_i)$ is the i-th component Gaussian density. The parameter λ is, on the other hand, the set of parameters that completely defines the mixture model: $\{\omega_i, \mu_i, \sigma_i\} \quad \forall i = 1, \ldots, M$. For our modeling, p has been computed for a total number of components $M \in \{1, 2, \ldots, 6\}$ for each of the 6 encoded features. For each M, the Bayesian Information Criterion (BIC) [96]:

$$BIC = -2\log\hat{L} + d \cdot \log N \tag{5.6}$$

is computed. Above, $\hat{L}$ is the maximum likelihood of the model, d is the number of parameters and N is the number of samples. The optimal number of GMM components for each of the 6 features is chosen as the one with the lowest BIC. The optimal GMMs obtained for each of the 6 bottleneck dimensions are shown in Figure 5.5.

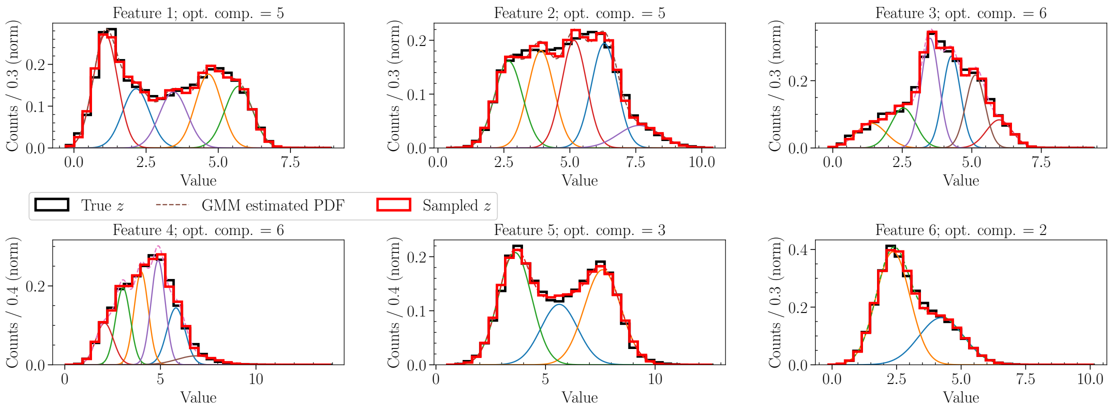

Figure 5.5: Gaussian Mixture Models (GMM) obtained for each of the 6 bottleneck encoded features. The true encoded features are shown as black histograms. For each feature, the optimal components number for the GMM is obtained as the one with lowest BIC (Equation 5.6), and the components are drawn as solid colored lines. The dashed lines are the GMM estimated probability density functions, and the red histograms are the result of sampling such PDFs.

Even though the agreement between true and sampled marginal distributions is apparent, this one-dimensional GMM recipe manifests some flaws. The most important are:

1. The procedure described above was applied to the full cumulative encoded distributions, i.e. across all three offsets. This is non-ideal as the sampling assumes similar distributions for all three offsets which is, as explained in Section 5.1, far from being the case.

2. The algorithm is considerably slow, as one has to generate 6 GMMs for each of the 6 distributions, which gives a total of 36 GMMs. The solution to the problem noted above, which is to perform independent sampling for each offset, results in a total of 108 GMMs.

3. Even overcoming the first issue and bearing with the extended computing times, the procedure has a fundamental flaw: sampling the one-dimensional marginal distributions is not enough to obtain a truthful six-dimensional representation of the original encoded distribution. This is natural, as sampling 6 dimensions independently does not in general provide a similar six-dimensional distribution, given the fact that correlations between features are completely lost.

We solve all the previous drawbacks by implementing the Kernel Density Estimation (KDE) [97] via the available method *KernelDensity* within the Python module SCIKIT-LEARN [98]. The KDE is the application of kernel smoothing for probability density estimation. It is a non-parametric method that uses kernels as weights in order to estimate the probability density function of a random variable in a given number of dimensions. Mathematically, a kernel is a positive function $K(x;\ h)$ that is controlled by the bandwidth parameter h. The density estimate at a point y within a group of points x_i, $i = 1, \ldots, N$; is given by:

$$\rho_K(y) = \frac{1}{N} \sum_{i=1}^{N} K(y - x_i;\, h). \tag{5.7}$$

In our case, the applied kernel is Gaussian, which means that

$$K(y - x_i;\, h) \propto \exp\left[-\frac{(y - x_i)^2}{2h^2}\right]. \tag{5.8}$$

Additionally, the bandwidth parameter for the density estimation in Equation 5.7 has been set to the default value $h = 1$. The sampled distributions using KDE are shown as red dashed histograms in Figure 5.6. It is important to note that these are not samples of one-dimensional marginal densities, but of the estimated six-dimensional probability density function. Naturally enough, the produced marginal distributions are in great agreement with the original ones.

5.3.1 One-dimensional distributions

Once the sampling has been executed, one can use the artificial samples to generate a new dataset and perform a completely analogous analysis as the one done in Section 5.2. The generated dataset should be rather similar to the one decoded, as we are using the exact same decoder –with the same weights– to obtain physical values from the encoded distribution. The trick is that the points in the six-dimensional bottleneck space are no longer the same as before, and can thus give rise to a complete new set of data. In this section and, for direct comparison, we chose to generate the same number of events as in the decoded dataset, but this could be extended to an arbitrarily large number. In analogy with Section 5.2.1 the p_x, p_y, p_z and E one-dimensional distributions are shown in Figure 5.7. The true (or input) distributions are drawn as a solid red histogram, while the generated events are shown as a dotted blue histogram. The generated events, like the decoded ones, are able to capture distinctive features, such as the bimodality of p_x and p_y, and reproduce them in output. Similarly to the decoded events, also here a reasonably good agreement between true and generated distributions throughout the entire range can be observed for all 4 kinematic quantities.

The same systematic effects present in the decoded dataset (Figure 5.3) are nevertheless present also in the generated one. In particular, we observe a similar overestimation in the bins containing the highest number of events, leading to an underestimation of the remaining bins. The Wasserstein distance is also computed for this generated dataset, similarly to Equation 5.2:

$$\begin{aligned} W_1^{\text{gen}}(p_x) &\approx 2.15\,\text{GeV}, & W_1^{\text{gen}}(p_y) &\approx 1.98\,\text{GeV}, \\ W_1^{\text{gen}}(p_z) &\approx 17.75\,\text{GeV}, & W_1^{\text{gen}}(E) &\approx 17.89\,\text{GeV}. \end{aligned} \tag{5.9}$$

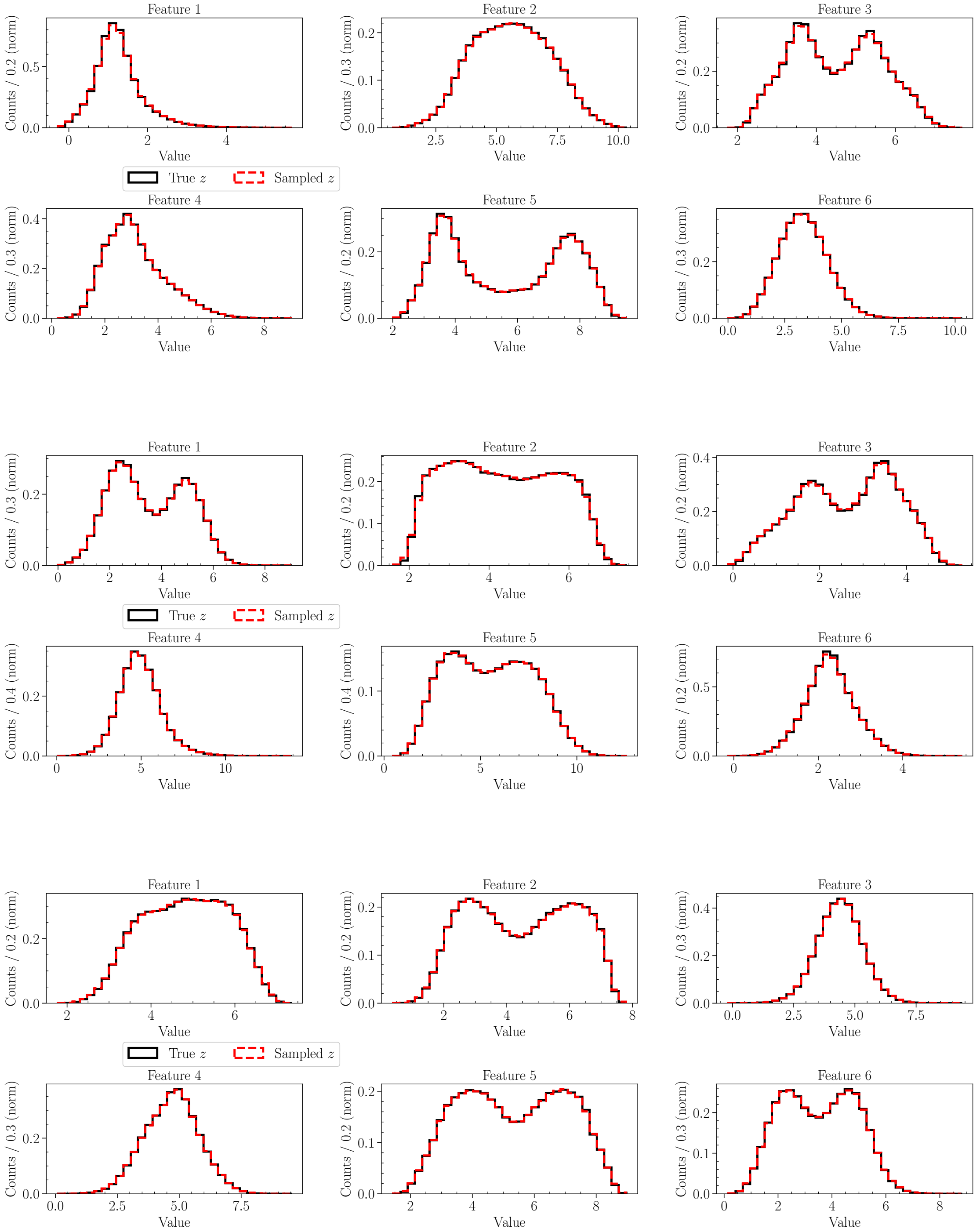

Figure 5.6: Ordered from top to bottom, the encoded features distributions (in solid black histograms) for models offsets 0, 1, and 2, respectively. The sampled distribution using KDE (Equation 5.7) is shown as red dashed histograms for each dimension.

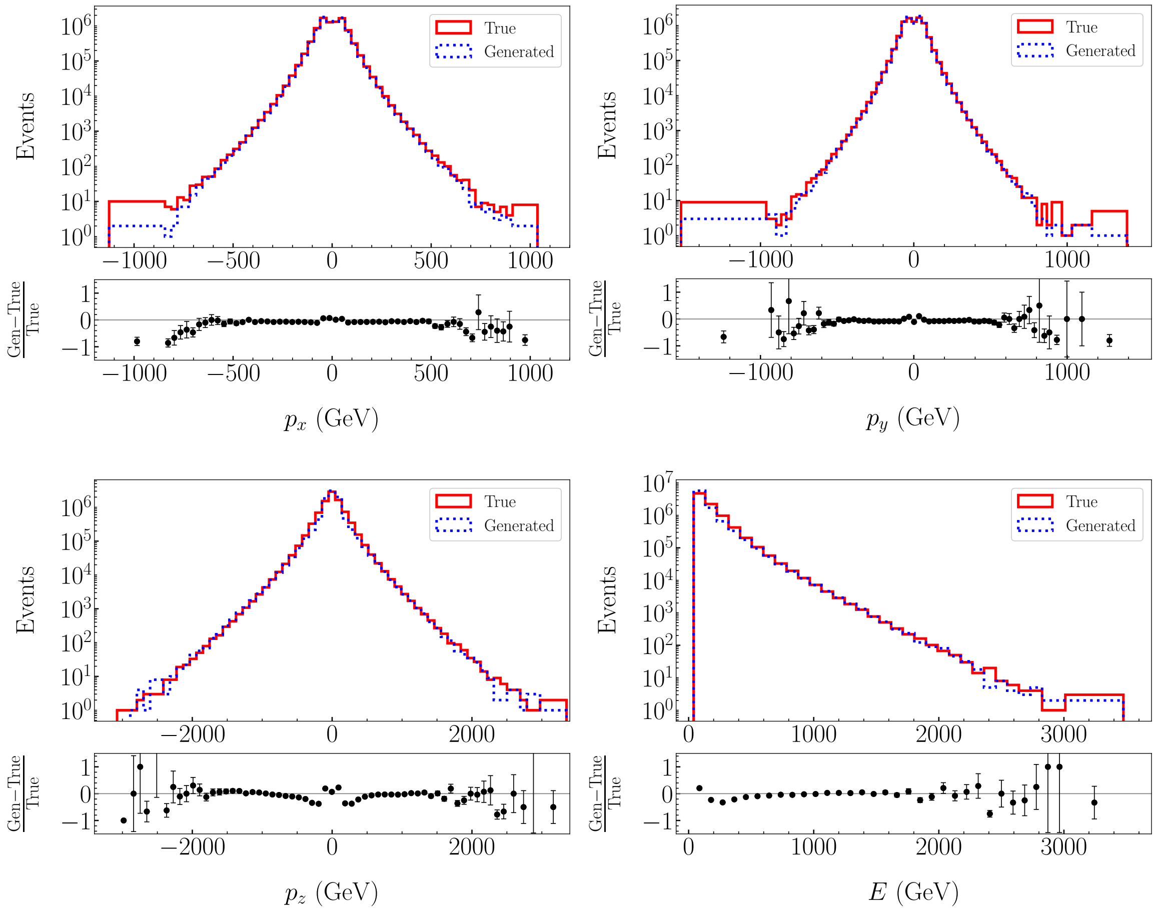

Figure 5.7: True (solid red histogram) and generated (dotted blue histogram) distributions for p_x (top left), p_y (top right), p_z (bottom left) and E (bottom right) of the 4 b-jets. Below every distribution comparison, the ratio of generated minus true events divided by true events is shown.

Remarkably enough, the obtained Wasserstein distances between true and generated distributions are all slightly lower than those between true and reconstructed distributions. It is not reasonable to draw any conclusions, nonetheless, as we show these numbers as a quantitative comparison of the similarity between distributions, rather than a comprehensive study of their compatibility. This is also the reason why we are not taking account here of systematic effects in the distributions, nor assigning them uncertainties.

5.3.2 $m_{2\mathrm{j}}$ vs $m_{4\mathrm{j}}$

To follow up and analogously to Section 5.2.2, we outline here the two-dimensional distributions $m_{2\mathrm{j}}(m_{4\mathrm{j}})$. In Figure 5.8 (top left), the true (or input) distribution is shown, where each bin is colored as a function of the number of events contained in it. The shaded region marks the area with no physical meaning, since there $m_{2\mathrm{j}} > m_{4\mathrm{j}}$. In Figure 5.8 (top right), we show the equivalent distribution for the generated events, plotted with the exact same binning.

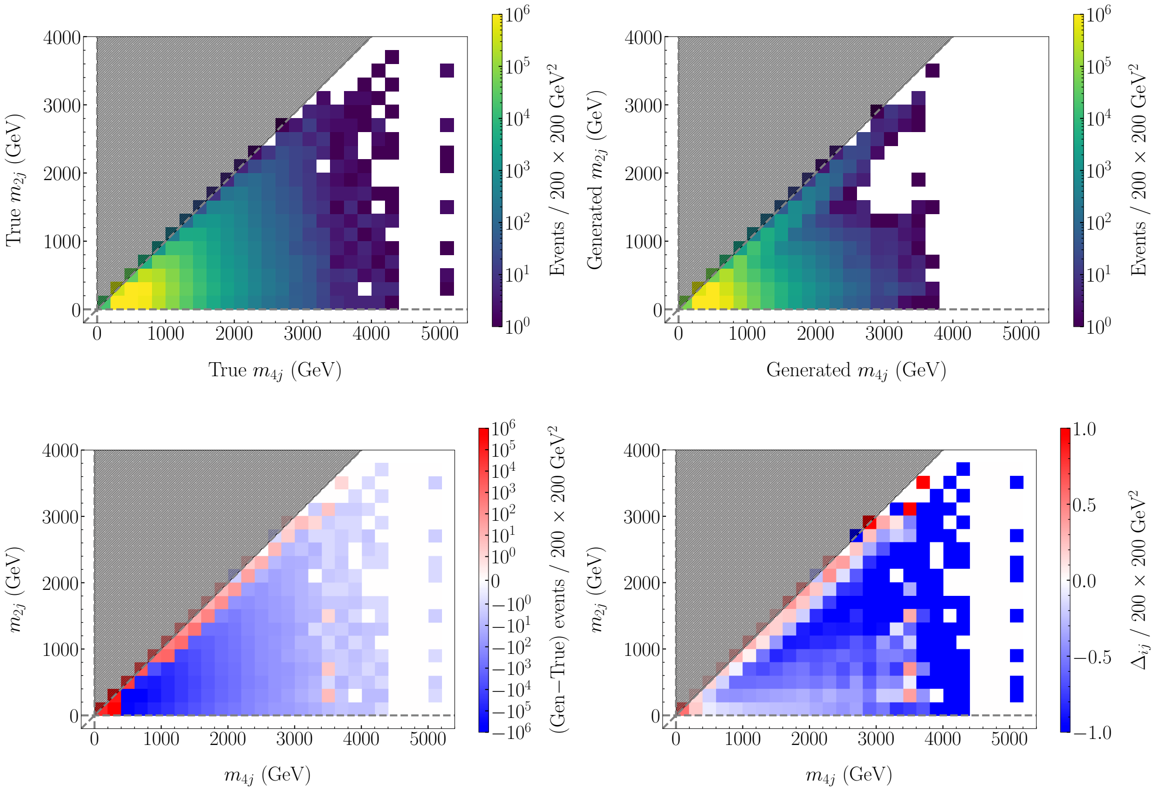

Figure 5.8: Two-dimensional distributions of true (top left) and generated (top right) dijet invariant masses as a function of the quadjet invariant mass. The bin-by-bin difference between the generated and true distributions is shown in the bottom left, whereas the bin-by-bin percent difference defined in Equation 5.10 is shown in the bottom right.

One can, once again, produce a third histogram shown in Figure 5.8 (bottom left), where each bin content is the difference between the number of events in the generated distribution (top right) and the true distribution (top left). Blue bins indicate regions in which the true distribution contains more events, whereas red bins indicate the opposite, being the magnitude of the difference encoded as color intensity. It is clear that similar systematic effects as the ones present in Figure 5.4 (bottom left) are also apparent here. The generated distribution, just like the reconstructed one, also seems to be biased toward the region where $m_{2\mathrm{j}} \sim m_{4\mathrm{j}}$. In order to better visualize the systematic differences between the two distributions, one can construct the bin-by-bin percent difference for the generated dataset, analogous to the one defined in Equation 5.3:

$$\Delta_{ij} \equiv \frac{h_{ij}^{\mathrm{gen}} - h_{ij}^{\mathrm{true}}}{h_{ij}^{\mathrm{gen}} + h_{ij}^{\mathrm{true}}}. \tag{5.10}$$

In the equation above, the figure of merit Δ_{ij} is the difference in events in generated and true distributions, divided by their sum for the given ij-th bin. This distribution is shown in Figure 5.8 (bottom right), displaying a very similar result to the one depicted in Figure 5.4 for the decoded dataset.

Also for the generated dataset, one could calculate the total sum of these differences, i.e. $\sum_{ij}^{N\ \mathrm{bins}} \Delta_{ij}$, and divide it by the number of bins, i.e. $N = 47$ in our case. This provides the result:

$$\Delta_N \equiv \frac{\sum_{ij}^{N\ \mathrm{bins}} \Delta_{ij}}{N} \approx -2.83, \tag{5.11}$$

which is approximately equal to the one obtained for the decoded dataset, $\Delta_N \approx -2.80$. We remind here that obtaining a negative value is a sign of events being systematically reconstructed in very localized regions of the phase space, giving rise to the already mentioned few overcompensations in detriment of a general, more subtle, underestimation.

5.4 Comprehensive results

As the final section for this part, we gather here the numeric estimators obtained through the course of the results chapter. In Table 5.1, these values are listed as quantifiers of the agreement between the one-dimensional kinematic distributions and the two-dimensional mass distributions.

	Wasserstein 1-distance				$m_{2\mathrm{j}}(m_{4\mathrm{j}})$ percent diff.
	p_x	p_y	p_z	E	Δ_N
Reconstructed	2.31 GeV	2.13 GeV	18.70 GeV	18.82 GeV	−2.80
Generated	2.15 GeV	1.98 GeV	17.75 GeV	17.89 GeV	−2.83

Table 5.1: Quantitative estimators of the similarities between true distributions and reconstructed (first row) or generated (second row) distributions. The first 4 columns show the Wasserstein 1-distance computed as described in Equation 5.1. The last column shows the sum of the bin-by-bin percent differences divided by the total number of bins, defined as in Equation 5.3 for the reconstructed dataset and as in Equation 5.10 for the generated dataset.

In case of the former, the Wasserstein 1-distance has been used to address the cost of transforming one distribution into the other, giving a sense of the energy "wrongly-allocated" in the reconstructed/generated dataset. For the latter, on the other hand, a figure of merit has been built to plot the bin-by-bin percent differences between the true and reconstructed/generated $m_{2\mathrm{j}}(m_{4\mathrm{j}})$ distributions. We believe that this, rather than the simple differences in events per bin, reflects the performance of the autoencoder in a more meaningful way.

Chapter 6

Conclusions

In this chapter we draw together the conclusions of this work, centered on the use of an autoencoder structure for characterizing the QCD multijet background at the LHC. This text cannot be understood without the context it is embedded in, which is the search of the Higgs boson pair production carried out at the CMS and ATLAS experiments, at CERN. In Chapter 1, we have outlined the main theoretical details behind this search, and its powerful reach in uncovering the value of the Higgs boson self-coupling. This quantity, not predicted by the Standard Model, stands as one of the most decisive numbers of the effective theory, so much so that it governs the (meta-)stability of the electroweak vacuum. This has profound implications in our understanding of many physical processes, as they may be the meta-stable manifestation of fields that could, eventually, transition to the true stable state within a finite time scale. Additionally, large deviations of the Higgs boson self-coupling are expected in minimally modified Beyond Standard Model scenarios, which turns them into an excellent probe into these new physics processes. The LHC hosts two experimental facilities to test out these hypotheses, which are the CMS experiment, largely described in Chapter 2, and the ATLAS experiment. One of the best channels for these collaborations to observe the production of pairs of Higgs bosons is, along with $b\bar{b}\tau^+\tau^-$ and $b\bar{b}\gamma\gamma$, the $b\bar{b}b\bar{b}$ channel. This channel is characterized by the presence of at least 4 b-jets in the final state and faces a variety of challenges. First of all, a large number of background processes lead to the same final state and are considerably more dominant (signal-to-background ratio is expected to be around 3% in highest purity bins after multivariate signal extraction). Secondly, the correct flavor tagging of these jets adds an extra layer of difficulty, as already outlined in Chapter 2, with the best efficiencies in identifying b-jets being around 70%.

These difficulties have been traditionally addressed with quite advanced methods, such as hemisphere mixing or kinematic reweighting, both explained in Chapter 3. Such tools aim at extracting a signal-depleted background template, modeled solely from data and relying on simple assumptions, such as the type of correlations expected between Higgs boson candidates, or the region where the signal is predicted to appear in the phase space.

The increasing capabilities of machine learning tools, however, have opened the door for more applications in high-energy physics, with background modeling currently undergoing extensive exploration. Many advanced classifiers, for example, are now being used in flavor tagging, outperforming the traditional multivariate analyzers. Another such tool, known as the autoencoder, has been employed in this work for the characterization of QCD multijet background in the di-Higgs search. Autoencoders have a great advantage in working with very few assumptions regarding the physics of the processes they are presented with. When the autoencoder is fed a physics dataset, it is trained by trying to reproduce in the output the features given in input, which is a non-trivial task if the dimension in the middle of the autoencoder is drastically reduced. The autoencoder is thus able to learn the most prominent features of the input, embed them in a useful way in the bottleneck, and then upsample from it. This is part of the reason why they have been recently used for anomaly detection in particle physics [99, 100], training them on SM processes and analyzing the loss on anomalies introduced in the simulations.

The purpose of this work has been to successfully encode the 16 kinematic variables corresponding to the $\{p_T,\ \eta,\ \phi,\ m\}$ values for each of four b-jets. In the utilized simulation sample, the jets' masses were all zero for simplicity, so one can think of a reduction from 12 dimensions in input down to a six-dimensional bottleneck, and then up again to an output with 16 dimensions corresponding to the $\{p_x,\ p_y,\ p_z,\ E\}$ values. The primary effort presented here involved the identification and design of a model capable of reconstructing the features in input with reasonable fidelity. This was a non-trivial assignment since the architecture needed to both reflect the internal symmetries of the dataset and encode it in a way such that reconstructing it was a feasible task. In particular, once the obtained embedded features were a meaningful representation of the input dataset, which was not straightforward, the design of the decoder architecture posed an even larger challenge. The purpose was to construct a reasonably symmetric architecture with respect to the encoder, but the lack of dimensions in the encoded space was prohibitive to such a goal. After many tests with high dimensions, we found that convoluting from a pseudo-event feature to an embedded quadjet space and then applying transposed convolutions to obtain dijets and jets was a coherent way to expand the physical dimensions. Several convolutional layers were first applied at that stage, but after carefully refining the final activations, the expressiveness of the network became good enough to retain only the final *decode_j1* and *decode_j2* layers, as shown in Chapter 4. Several tests were also carried out with the way the autoencoder learns. For example, computing the loss on the $\{p_T,\ \eta,\ \phi,\ m\}$ values was attempted, only to find out that the autoencoder struggled much more with such differently shaped distributions. In the end, we found that the best way for the autoencoder to learn the physical features was to implement the loss between $\{p_x,\ p_y,\ p_z,\ E\}$ representations.

To summarize, three equal autoencoder models have been trained on a sample of two-thirds of 2,202,261 background events with four b-jets. Cross-validation has been done by evaluating each model on the third of the dataset it has not been trained on. Results have been described in Chapter 5, being the decoded dataset analyzed in Section 5.2. For a large dimensionality reduction in the bottleneck, a reasonable agreement between one-dimensional distributions has been found (Section 5.2.1). To quantify such agreement, the Wasserstein 1-distance has been calculated for the reconstructed three components of momentum and the energy of the 4 jets. The two-dimensional distributions of the dijet invariant mass versus the quadjet invariant mass have been built for the true and reconstructed datasets, observing general agreement in the bulk region, although revealing several systematic effects. A figure of merit has been built in order to quantify the agreement in these two-dimensional distributions, constructed as the sum of the percent bin-by-bin deviations divided by the total number of bins. A value of approximately -2.80 has been obtained, indicating a systematic effect of underestimation in the reconstructed mass bins.

Autoencoders can also be used for data generation (more often variational autoencoders) when the model has produced a useful encoded representation. Therefore, the decoder has been utilized to upsample a random six-dimensional distribution, producing new artificial samples. In this work, the first attempt to sample the encoded distribution was done by using Gaussian Mixture Models. This did not provide the desired results, as the correlations between sampled one-dimensional marginal distributions were lost. The Kernel Density Estimation of the six-dimensional distribution, on the other hand, provided an accurate sampling and thus faithfully reproduced the one-dimensional marginals. Once the artificial dataset was generated, analogous studies to those carried out with the decoded dataset were performed, obtaining very similar results. In particular, the generated distributions reproduce with similar accuracy the same features as the decoded dataset, conserving rather alike systematic effects.

This work, as a whole, functions as a proof of concept for its primary aim: demonstrating that significantly reduced encoded features of an input event can hold a large amount of the initial information, allowing for efficient compression. However, this is not all; this encoded space could potentially incorporate a metric capable of effectively distinguishing between background and signal events, thereby achieving a clear separation within the six-dimensional domain. The natural next step would involve utilizing these encoded features for extrapolating background from background-populated regions to the signal region. This extrapolation would allow for accurate modeling of the background in, for example, Higgs boson pair production searches. However, the predominant concern remains the sizable systematic effects observed in the autoencoder's replication of the input dataset. As a future perspective, we estimate that tackling and quantifying these effects should be the subsequent steps in achieving a more accurate output. As a general recipe, one could extract the ratio between generated and reconstructed bins in each distribution. For example, for the reconstructed variable x_{reco}, define $f(x_{\text{reco}}) \equiv x_{\text{reco}}/x_{\text{gen}}$. Then, generate a different distribution from the encoded features, and multiply it by $f(x_{\text{reco}})$. This way, one can obtain a reconstructed distribution x'_{reco} that is ideally corrected by the generated distribution. Once this is done, one can extract the distribution of differences $x_{\text{true}} - x'_{\text{reco}}$. The root mean square (RMS) of this distribution can be then used as an estimate of the systematic errors in the modeling of the reconstructed features.

Appendix A

Additional plots

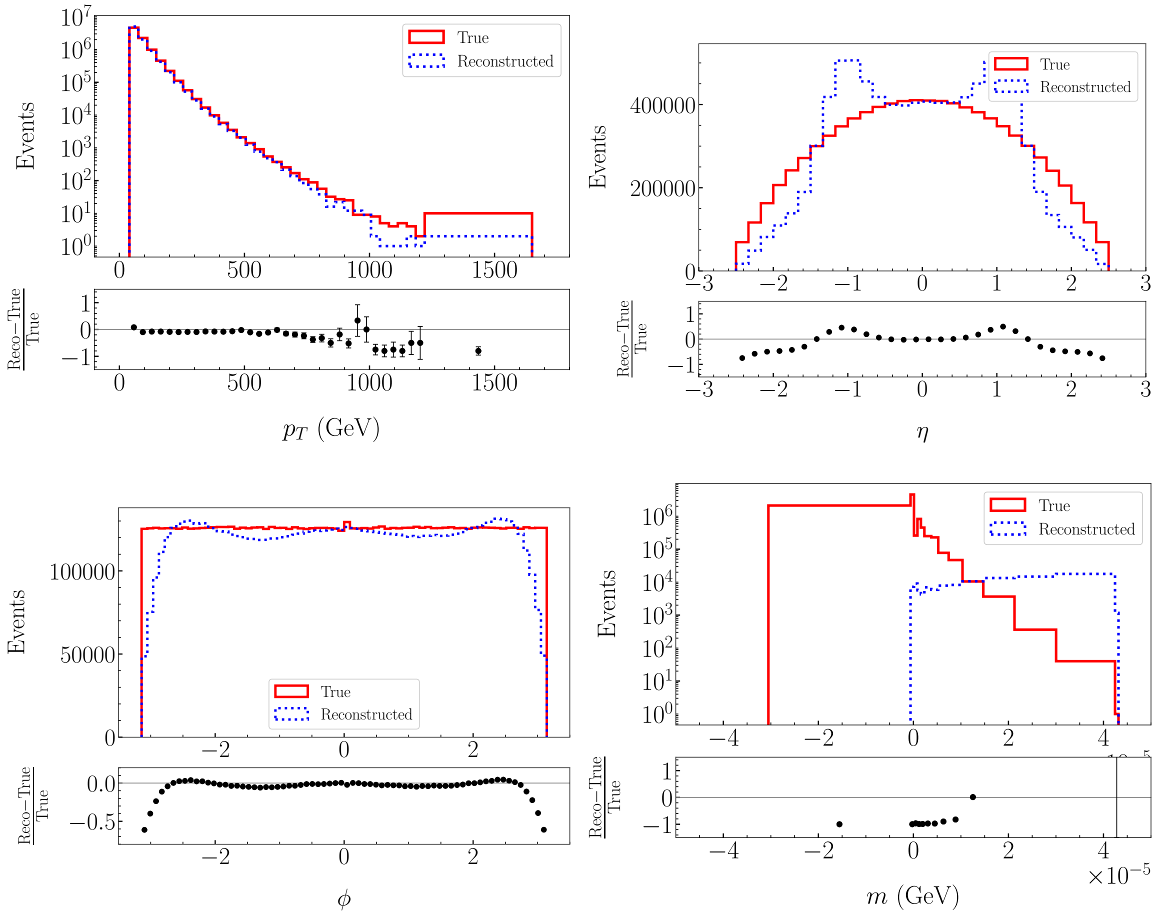

Figure A.1: True (solid red histogram) and reconstructed (dotted blue histogram) distributions for p_T (top left), η (top right), ϕ (bottom left) and m (bottom right) of the 4 b-jets. Below every distribution comparison, the ratio of reconstructed minus true events divided by true events is shown.

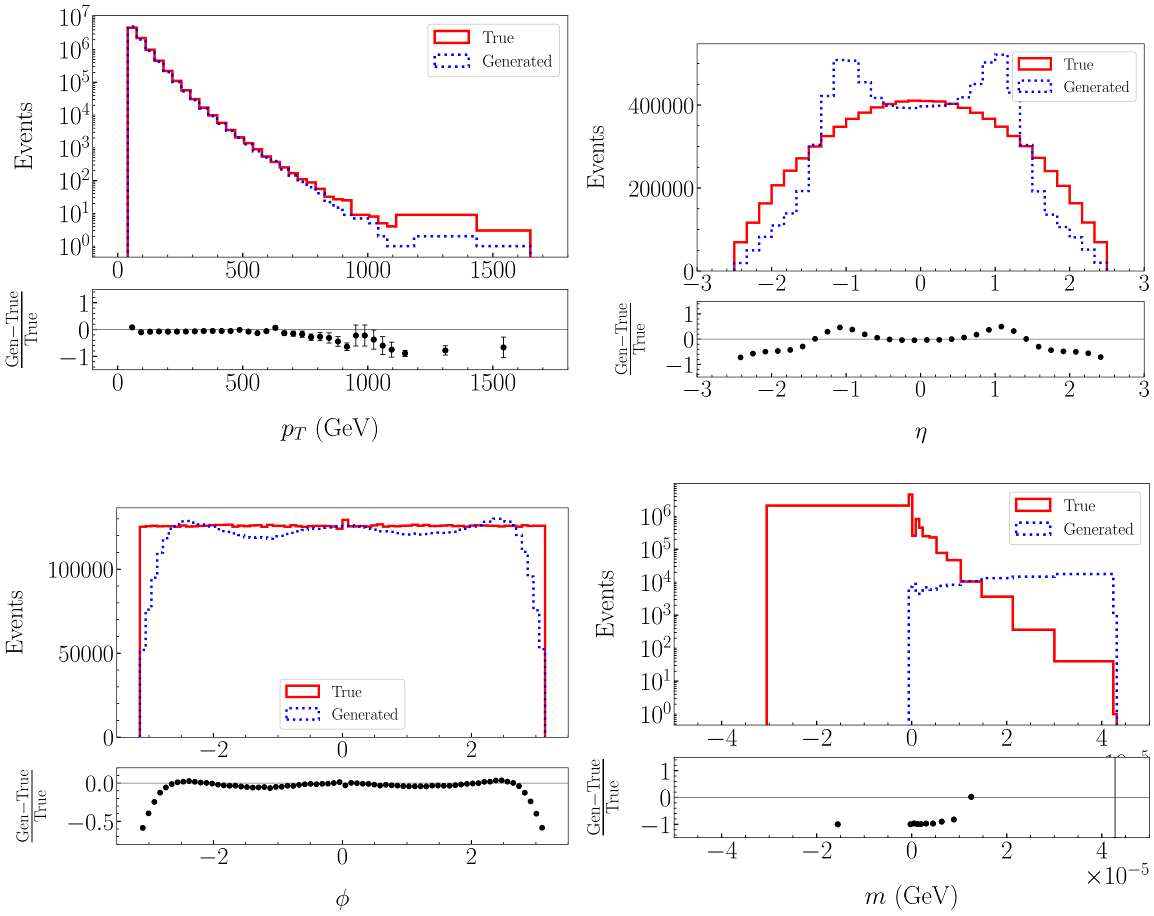

Figure A.2: True (solid red histogram) and generated (dotted blue histogram) distributions for p_T (top left), η (top right), ϕ (bottom left) and m (bottom right) of the 4 b-jets. Below every distribution comparison, the ratio of generated minus true events divided by true events is shown.

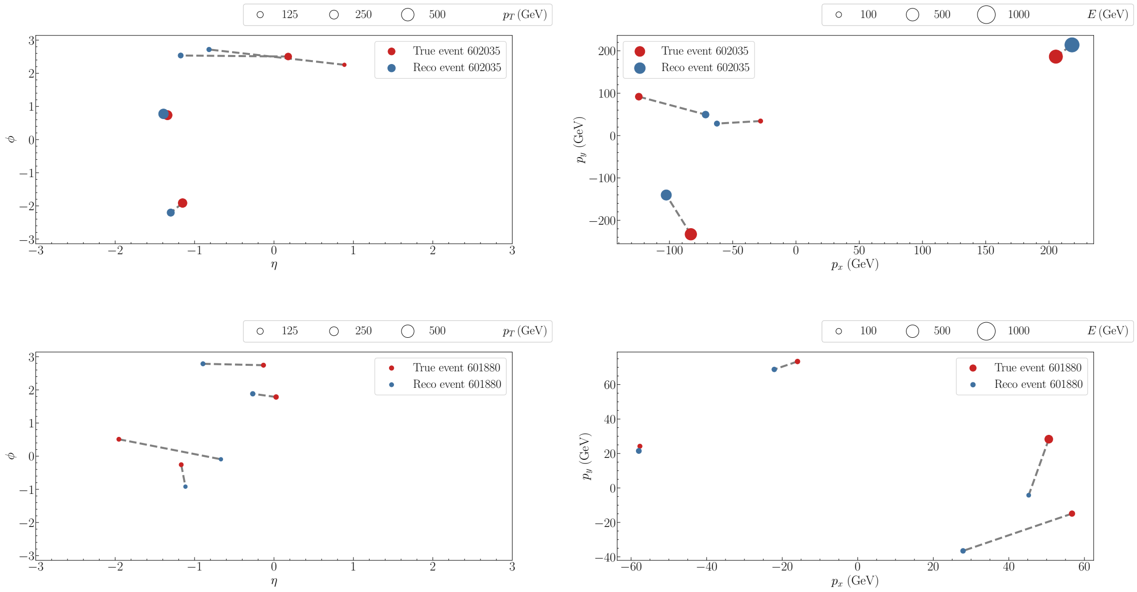

Figure A.3: Two examples of η-ϕ (left column) and p_x-p_y (right column) reconstructed pairs for event number 602035 (upper row) and 601880 (bottom row). Red (blue) points represent the input (reconstructed) jets coordinates. The size of the marker scales with the p_T for the η-ϕ pairs, and with the energy of the jet E for the p_x-p_y coordinates. Reasonably good agreement is obtained in all cases, being slightly better for higher p_T jets.

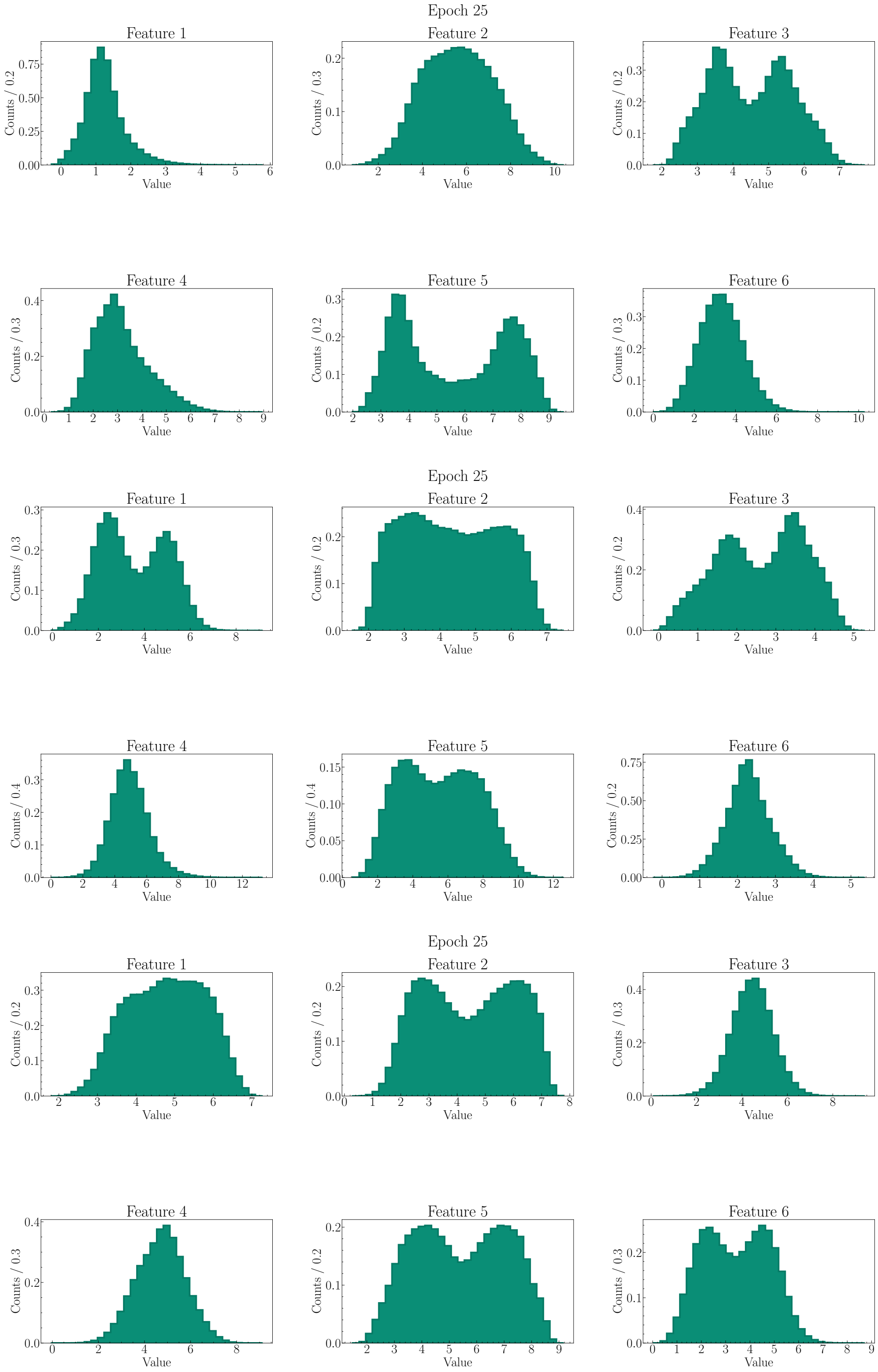

Figure A.4: Ordered from top to bottom, marginal one-dimensional distributions of the activations for models offsets 0, 1, and 2, respectively. It is evident here the large differences that arise between models, presumably due to random initialized weights for each of the three trainings.

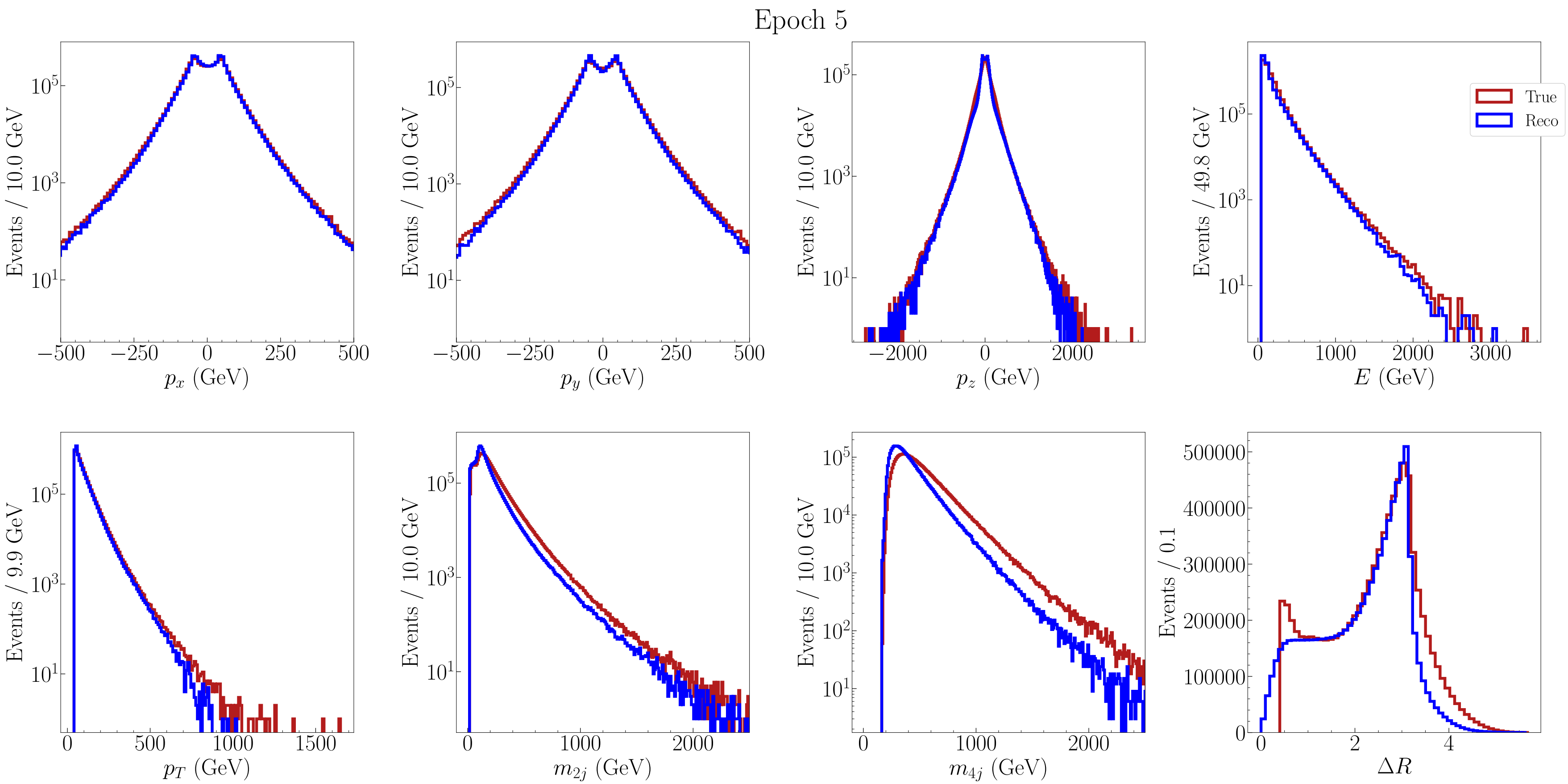

Figure A.5: One-dimensional distributions of kinematic variables monitored during training, epoch 5. The plotted distributions correspond to the training set containing events with offset 1 and 2 (i.e. the model that would later be evaluated on the events with offset 0). From left to right, and upper row to bottom row: p_x, p_y, p_z, E, p_T, m_{2j}, m_{4j}, and $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$. It can be seen that the autoencoder is not aware that jets are reconstructed with a $\Delta R = 0.4$ radius, so there should not be angular separations below that threshold. This effect helps to explain (in part) the bias shifting m_{4j} toward lower values.

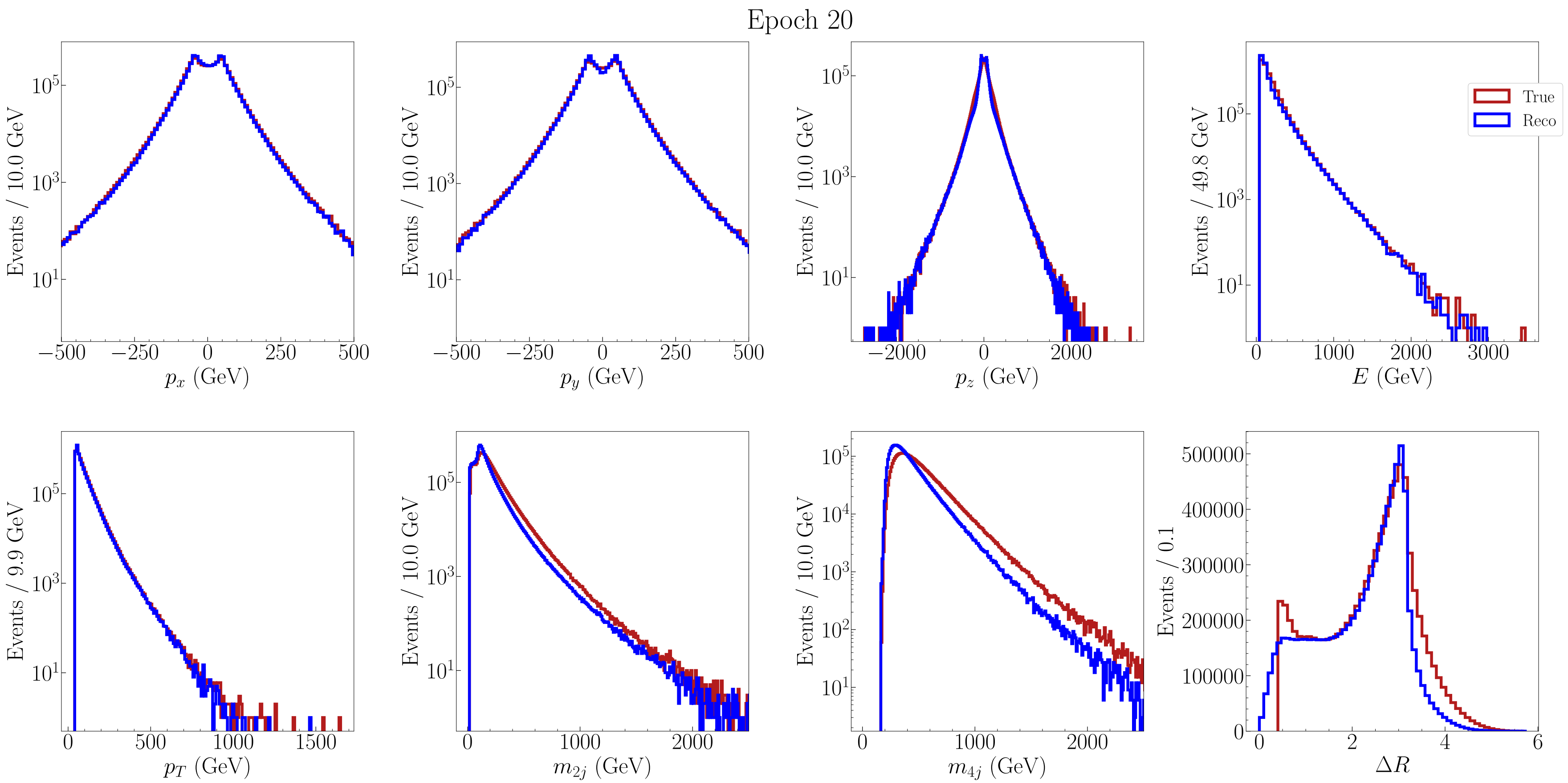

Figure A.6: One-dimensional distributions of kinematic variables monitored during training, epoch 20. The plotted distributions correspond to the training set containing events with offset 1 and 2 (i.e. the model that would later be evaluated on the events with offset 0). From left to right, and upper row to bottom row: p_x, p_y, p_z, E, p_T, $m_{2\mathrm{j}}$, $m_{4\mathrm{j}}$, and $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$. It can be seen that the autoencoder is not aware that jets are reconstructed with a $\Delta R = 0.4$ radius, so there should not be angular separations between jets below that threshold. This effect helps to explain (in part) the bias shifting $m_{4\mathrm{j}}$ toward lower values.

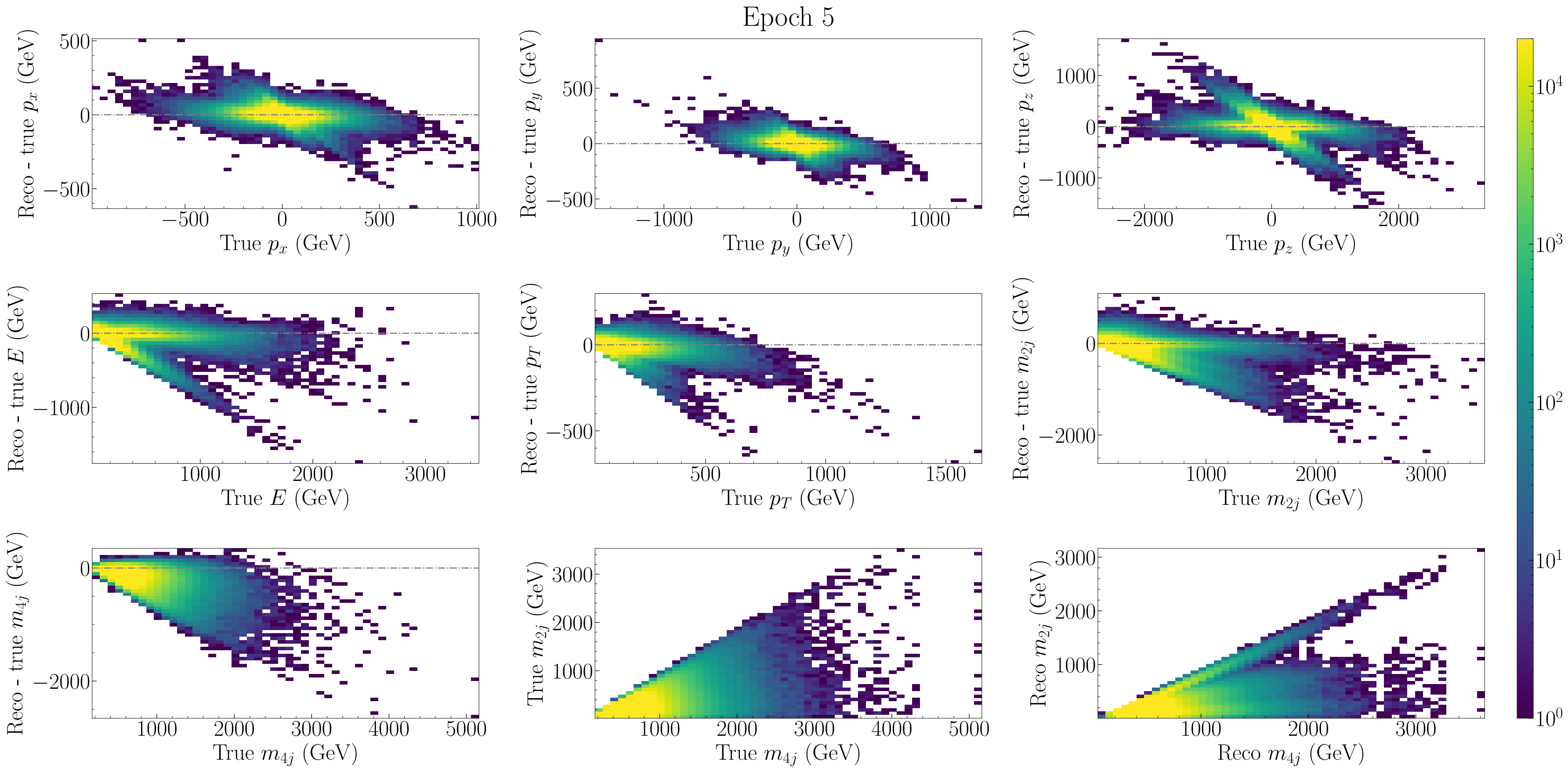

Figure A.7: Distributions of residuals monitored during training, epoch 5. The plotted residuals correspond to the training set containing events with offset 1 and 2 (i.e. the model that would later be evaluated on the events with offset 0). From upper to bottom row and left to right: p_x, p_y, p_z, E, p_T, m_{2j}, m_{4j}, true $m_{2j}(m_{4j})$ two-dimensional distribution, reconstructed $m_{2j}(m_{4j})$ two-dimensional distribution.

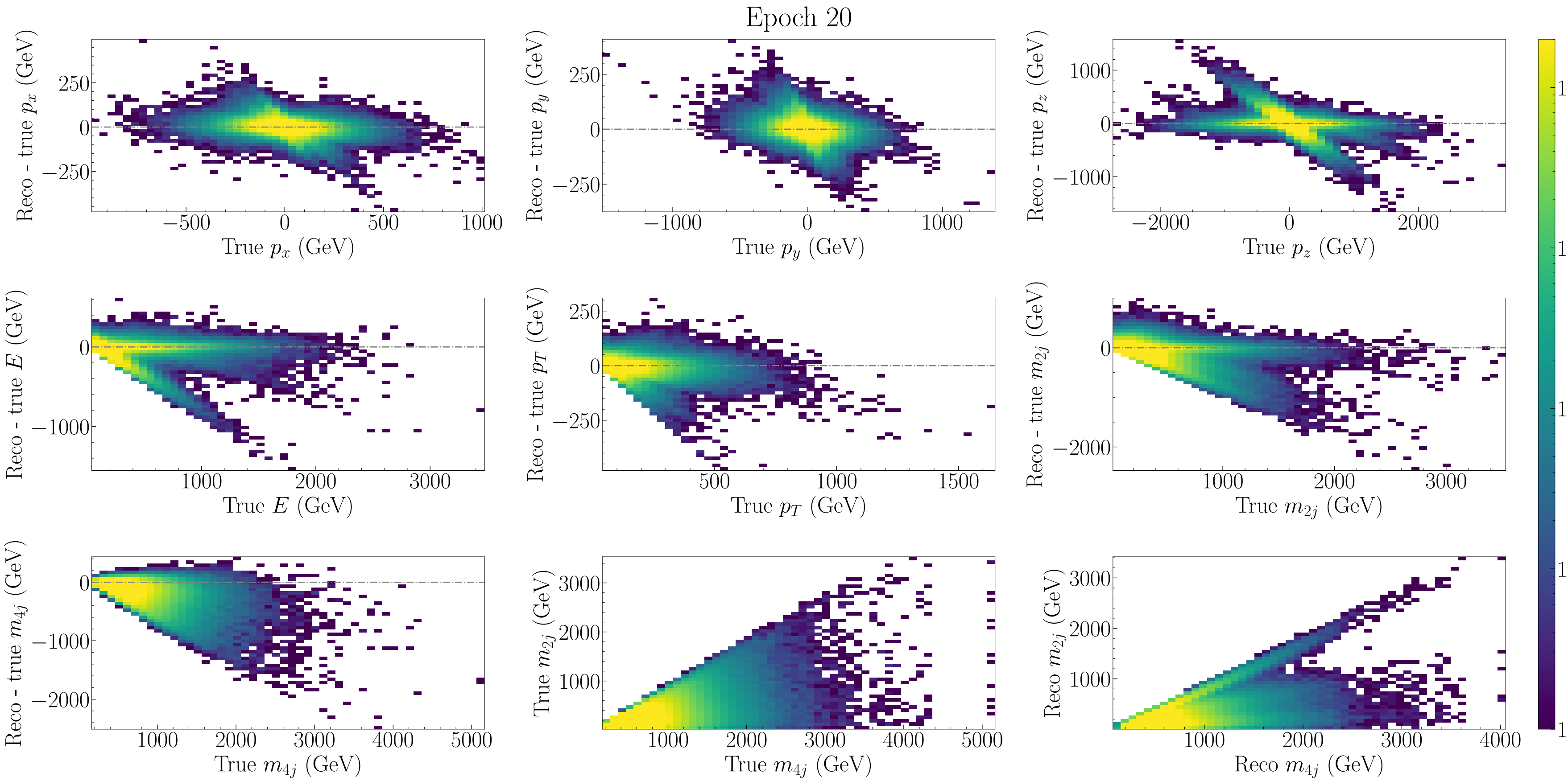

Figure A.8: Distributions of residuals monitored during training, epoch 20. The plotted residuals correspond to the training set containing events with offset 1 and 2 (i.e. the model that would later be evaluated on the events with offset 0). From upper to bottom row and left to right: p_x, p_y, p_z, E, p_T, $m_{2\text{j}}$, $m_{4\text{j}}$, true $m_{2\text{j}}(m_{4\text{j}})$ two-dimensional distribution, reconstructed $m_{2\text{j}}(m_{4\text{j}})$ two-dimensional distribution.

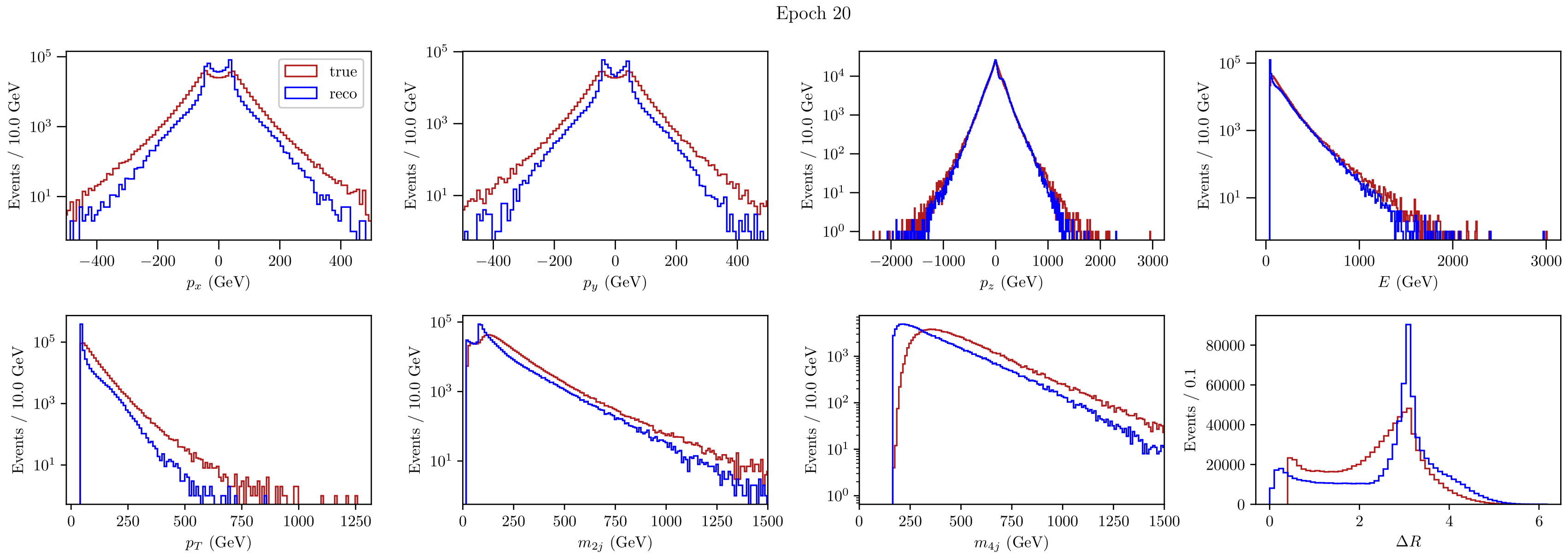

Figure A.9: One-dimensional distributions of kinematic variables monitored during training **with *symmetrization* applied**, epoch 20. The plotted distributions correspond to the training set containing events with offset 1 and 2 (i.e. the model that would later be evaluated on the events with offset 0). Here, the model is equipped with a 16-dimensional bottleneck, so the expected performance should be much better than the 6-dimensional autoencoder described in the work. From left to right, and upper row to bottom row: p_x, p_y, p_z, E, p_T, m_{2j}, m_{4j}, and $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$. The reconstruction of p_x and p_y distributions is largely worsened, mostly due to the failure in reconstructing ϕ information, as explained in Section 4.2.

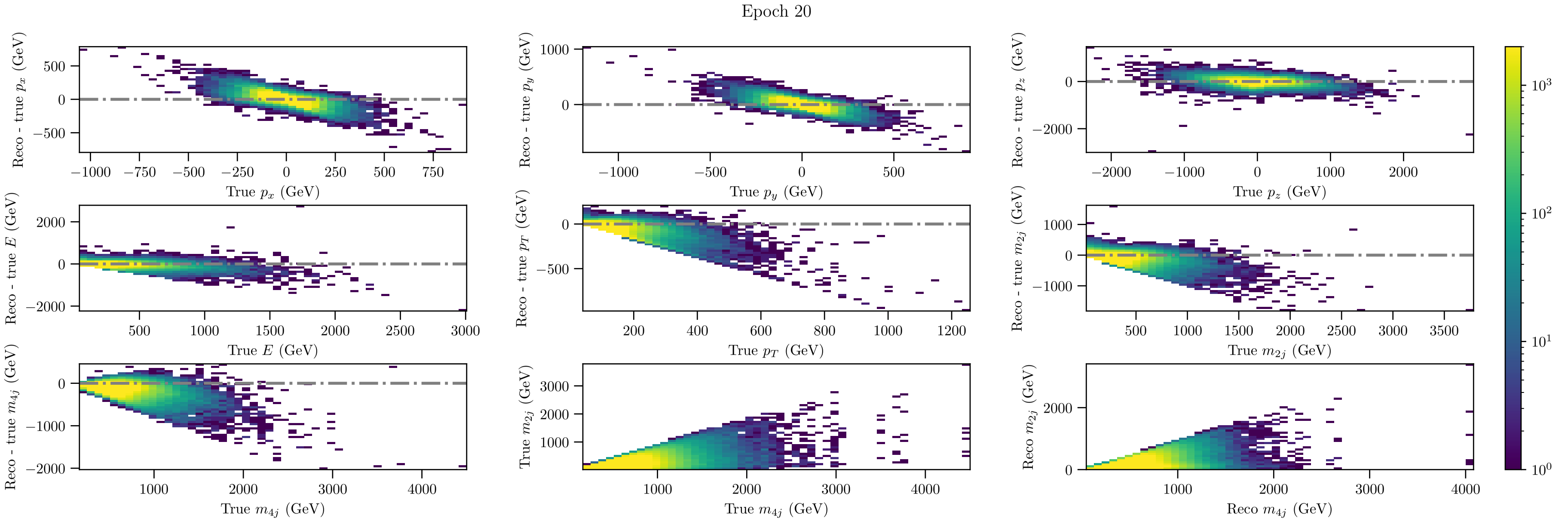

Figure A.10: Distributions of residuals monitored during training **with *symmetrization* applied**, epoch 20. The plotted residuals correspond to the training set containing events with offset 1 and 2 (i.e. the model that would later be evaluated on the events with offset 0). From upper to bottom row and left to right: p_x, p_y, p_z, E, p_T, m_{2j}, m_{4j}, true $m_{2j}(m_{4j})$ two-dimensional distribution, reconstructed $m_{2j}(m_{4j})$ two-dimensional distribution.

References

[1] G. Arnison et al., “Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s} = 540$ GeV”, Physics Letters B **122**, 103–116 (1983), DOI: 10.1016/0370-2693(83)91177-2.

[2] M. Banner et al., “Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider”, Physics Letters B **122**, 476–485 (1983), DOI: 10.1016/0370-2693(83)91177-2.

[3] P. W. Higgs, “Broken symmetries and the masses of gauge bosons”, Physical review letters **13**, 508 (1964), DOI: 10.1103/PhysRevLett.13.508.

[4] S. Myers and E. Picasso, “The design, construction and commissioning of the CERN Large Electron-Positron collider”, Contemporary Physics **31**, 387–403 (1990), DOI: 10.1080/00107519008213789.

[5] G. Dugan, “Tevatron collider: Status and prospects”, `https://doi.org/10.48550/arXiv.0910.3612` (1989), DOI: 10.48550/arXiv.0910.3612.

[6] L. R. Evans, “The Large Hadron Collider project”, in Proceedings of the sixteenth international cryogenic engineering conference/international cryogenic materials conference (Elsevier, 1997), pp. 45–52, DOI: 10.1016/B978-008042688-4/50012-0.

[7] CMS Collaboration, “The CMS experiment at the CERN LHC”, Journal of Instrumentation **3**, S08004 (2008), DOI: 10.1088/1748-0221/3/08/S08004.

[8] ATLAS Collaboration, “The ATLAS Experiment at the CERN Large Hadron Collider”, Journal of Instrumentation **3**, S08003 (2008), DOI: 10.1088/1748-0221/3/08/S08003.

[9] CMS Collaboration, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”, arXiv preprint arXiv:1207.7235, `https://doi.org/10.1016/j.physletb.2012.08.021` (2012), DOI: 10.1016/j.physletb.2012.08.021.

[10] G. Aad et al., “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”, Physics Letters B **716**, 1–29 (2012), DOI: 10.1016/j.physletb.2012.08.020.

[11] G. Aad et al., “Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment”, Eur. Phys. J. C **76**, 6 (2016), DOI: 10.1140/epjc/s10052-015-3769-y.

[12] V. Khachatryan et al., “Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton

collisions at 7 and 8 TeV", Eur. Phys. J. C **75**, 212 (2015), DOI: 10.1140/epjc/s10052-015-3351-7.

[13] S. Chatrchyan et al., "Study of the mass and spin-parity of the Higgs boson candidate via its decays to Z boson pairs", Phys. Rev. Lett. **110**, 081803 (2013), DOI: 10.1103/PhysRevLett.110.081803.

[14] G. Aad et al., "Evidence for the spin-0 nature of the Higgs boson using ATLAS data", Phys. Lett. B **726**, 120–144 (2013), DOI: 10.1016/j.physletb.2013.08.026.

[15] V. Khachatryan et al., "Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV", Phys. Rev. D **92**, 012004 (2015), DOI: 10.1103/PhysRevD.92.012004.

[16] F. Bezrukov et al., "Living beyond the edge: Higgs inflation and vacuum metastability", Physical Review D **92**, 083512 (2015), DOI: 10.1103/PhysRevD.92.083512.

[17] A. Bednyakov, "An advanced precision analysis of the SM vacuum stability", Physics of Particles and Nuclei **48**, 698–703 (2017), DOI: 10.1134/S1063779617050057.

[18] P. T. Komiske et al., "Metric space of collider events", Physical review letters **123**, 041801 (2019), DOI: 10.1103/PhysRevLett.123.041801.

[19] T. Cai et al., "Which metric on the space of collider events?", Physical Review D **105**, 076003 (2022), DOI: 10.1103/PhysRevD.105.076003.

[20] T. Manole et al., "Background Modeling for Double Higgs Boson Production: Density Ratios and Optimal Transport", arXiv preprint arXiv:2208.02807, `https://doi.org/10.48550/arXiv.2208.02807` (2022), DOI: 10.48550/arXiv.2208.02807.

[21] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure", Probability theory and related fields **162**, 707–738 (2015), DOI: 10.1007/s00440-014-0583-7.

[22] A. Ramdas et al., *On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests*, 2015, DOI: 10.48550/arXiv.1509.02237.

[23] A. David and G. Passarino, "Through precision straits to next Standard Model heights", Reviews in Physics **1**, 13–28 (2016), DOI: 10.1016/j.revip.2016.01.001.

[24] Wikipedia contributors, *The Standard Model*, [Online; accessed 17/05/2023].

[25] J. Goldstone et al., "Broken symmetries", Physical Review **127**, 965 (1962), DOI: 10.1103/PhysRev.127.965.

[26] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons", Physical review letters **13**, 321 (1964), DOI: 10.1103/PhysRevLett.13.321.

[27] G. S. Guralnik et al., "Global conservation laws and massless particles", Physical Review Letters **13**, 585 (1964), DOI: 10.1103/PhysRevLett.13.585.

[28] P. W. Higgs, "Spontaneous symmetry breakdown without massless bosons", Physical review **145**, 10.1103/PhysRev.145.1156, 1156 (1966).

[29] T. W. Kibble, "Symmetry breaking in non-abelian gauge theories", Physical Review **155**, 1554 (1967), DOI: 10.1103/PhysRev.155.1554.

[30] S. Weinberg, "A model of leptons", Physical review letters **19**, 1264 (1967), DOI: 10.1103/PhysRevLett.19.1264.

[31] K. Kumar et al., “Low-energy measurements of the weak mixing angle”, Annual Review of Nuclear and Particle Science **63**, 237–267 (2013), DOI: 10.1146/annurev-nucl-102212-170556.

[32] S. L. Glashow, “Partial-symmetries of weak interactions”, Nuclear physics **22**, 579–588 (1961), DOI: 10.1016/0029-5582(61)90469-2.

[33] T. Plehn and M. Rauch, “Quartic Higgs coupling at hadron colliders”, Physical Review D **72**, 053008 (2005), DOI: 10.1103/PhysRevD.72.053008.

[34] E. Rossi, “Measurement of Higgs-boson self-coupling with single-Higgs and double-Higgs production channels”, arXiv preprint arXiv:2010.05252, `https://doi.org/10.48550/arXiv.2010.05252` (2020), DOI: 10.48550/arXiv.2010.05252.

[35] S. D. Bass et al., “The Higgs boson implications and prospects for future discoveries”, Nature Reviews Physics **3**, 608–624 (2021), DOI: 10.1038/s42254-021-00341-2.

[36] S. Kanemura et al., “New physics effect on the Higgs self-coupling”, Physics Letters B **558**, 157–164 (2003), DOI: 10.1016/S0370-2693(03)00268-5.

[37] K. Agashe et al., “LHC signals for warped electroweak neutral gauge bosons 2007”, Phys. Rev. D **76**, 0709–0007, DOI: 10.1103/PhysRevD.76.115015.

[38] N. Arkani-Hamed et al., “Electroweak symmetry breaking from dimensional deconstruction”, Physics Letters B **513**, 232–240 (2001), DOI: 10.1016/S0370-2693(01)00741-9.

[39] M. Grazzini et al., “Higgs boson pair production at NNLO with top quark mass effects”, Journal of High Energy Physics **2018**, 1–21 (2018), DOI: 10.1007/JHEP05(2018)059.

[40] J. Baglio et al., “Prospects for Higgs physics at energies up to 100 TeV”, Reports on Progress in Physics **79**, 116201 (2016), DOI: 10.1088/0034-4885/79/11/116201.

[41] A. D. Martin et al., “Parton distributions for the LHC”, The European Physical Journal C **63**, 189–285 (2009), DOI: 10.1140/epjc/s10052-009-1072-5.

[42] A. Martin et al., “Uncertainties on α_S in global PDF analyses and implications for predicted hadronic cross sections”, The European Physical Journal C **64**, 653–680 (2009), DOI: 10.1140/epjc/s10052-009-1164-2.

[43] B. Di Micco et al., “Higgs boson potential at colliders: Status and perspectives”, Reviews in Physics **5**, 100045 (2020), DOI: 10.1016/j.revip.2020.100045.

[44] CMS Collaboration, “A portrait of the Higgs boson by the CMS experiment ten years after the discovery”, Nature **607**, 60–68 (2022), DOI: 10.1038/s41586-022-04892-x.

[45] M. Cepeda et al., “Higgs physics at the HL-LHC and HE-LHC”, arXiv preprint arXiv:1902.00134, `https://doi.org/10.48550/arXiv.1902.00134` (2019), DOI: 10.48550/arXiv.1902.00134.

[46] G. Aad et al., “Combination of searches for Higgs boson pairs in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”, Physics Letters B **800**, 135103 (2020), DOI: 10.1016/j.physletb.2019.135103.

[47] M. Benedikt et al., “LHC Design Report”, CERN, Geneva, `10.5170/CERN-2004-003-V-1` (2004), DOI: 10.5170/CERN-2004-003-V-1.

[48] E. Lopienska, *The CERN accelerator complex, layout in 2022*, tech. rep., https://cds.cern.ch/record/2800984 (2022), CERN-GRAPHICS-2022-001.

[49] LHCb Collaboration, "The LHCb Detector at the LHC", Journal of Instrumentation **3**, S08005 (2008), DOI: 10.1088/1748-0221/3/08/S08005.

[50] ALICE Collaboration, "The ALICE experiment at the CERN LHC", Journal of Instrumentation **3**, S08002 (2008), DOI: 10.1088/1748-0221/3/08/S08002.

[51] *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*, Technical design report. CMS, https://cds.cern.ch/record/922757 (CERN, Geneva, 2006), "There is an error on cover due to a technical problem for some items".

[52] T. Sakuma and T. McCauley, "Detector and Event Visualization with SketchUp at the CMS Experiment", J. Phys. Conf. Ser. **513**, edited by D. L. Groep and D. Bonacorsi, 022032 (2014), DOI: 10.1088/1742-6596/513/2/022032.

[53] CMS Collaboration, "Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays", Journal of Instrumentation **5**, T03021 (2010), DOI: 10.1088/1748-0221/5/03/T03021.

[54] Neutelings, Izaak, *CMS coordinate system*, [Online; accessed 22/05/2023].

[55] V. Karimäki et al., *The CMS tracker system project: Technical Design Report*, Technical design report. CMS, https://cds.cern.ch/record/368412 (CERN, Geneva, 1997), CERN-LHCC-98-006.

[56] A. Tumasyan et al., "The Phase-2 Upgrade of the CMS Tracker", edited by K. Klein, `10.17181/CERN.QZ28.FLHW` (2017), DOI: 10.17181/CERN.QZ28.FLHW.

[57] CMS Collaboration and others, "The CMS electromagnetic calorimeter project: technical design report", Technical Design Report CMS. CERN, Geneva **47**, https://cds.cern.ch/record/349375 (1997), CERN-LHCC-97-033.

[58] M. Cipriani, "Photon detection with the CMS ECAL in the present and at the HL-LHC and its impact on Higgs-Boson measurement", https://cds.cern.ch/record/2708020, 150–155 (2019).

[59] C. Cooke, "Upgrade of the CMS Barrel Electromagnetic Calorimeter for the High Luminosity LHC", Instruments **6**, 29 (2022), DOI: https://doi.org/10.3390/instruments6030029.

[60] CMS collaboration and others, "The CMS hadron calorimeter project: technical design report", Technical Design Report CMS. CERN, Geneva **49**, https://cds.cern.ch/record/357153 (1997), CERN-LHCC-97-031.

[61] A. Zucchetta, "Searches for signatures of an extended Higgs sector in final states with leptons and Higgs to bb decays at CMS", https://cds.cern.ch/record/2004824 (2015), CERN-THESIS-2015-025.

[62] P. Paolucci and C. muon collaboration., "The CMS muon system", in *Astroparticle, particle and space physics, detectors and medical physics applications* (World Scientific, 2006), pp. 605–615, DOI: 10.1142/9789812773678_0096.

[63] A. Sirunyan et al., “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV”, 10.1088/1748-0221/13/06/P06015 (2018), DOI: 10.1088/1748-0221/13/06/P06015.

[64] CMS collaboration and others, “CMS technical design report for the Level-1 trigger upgrade”, CMS Technical Design Report, https://cds.cern.ch/record/1556311 (2013), CERN-LHCC-2013-011.

[65] A. Rácz and P. Sphicas, *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger*, tech. rep., http://cds.cern.ch/record/578006 (CMS-TDR-006, 2002), CERN-LHCC-2002-026.

[66] CMS collaboration and others, *Particle-flow event reconstruction in CMS and performance for jets, taus and MET*, https://cds.cern.ch/record/1194487, Geneva, 2009.

[67] CMS collaboration and others, *Commissioning of the particle-flow event reconstruction with the first LHC collisions recorded in the CMS detector*, https://cds.cern.ch/record/1247373, 2010, CMS-PAS-PFT-10-001.

[68] M. Cacciari et al., “The anti-k_t jet clustering algorithm”, Journal of High Energy Physics **2008**, 063 (2008), DOI: 10.1088/1126-6708/2008/04/063.

[69] CMS collaboration and others, “Determination of jet energy calibration and transverse momentum resolution in CMS”, Journal of Instrumentation **6**, P11002 (2011), DOI: 10.1088/1748-0221/6/11/P11002.

[70] I. Tomalin, “B-tagging in CMS”, in Journal of physics: conference series, Vol. 110, 9 (IOP Publishing, 2008), p. 092033, DOI: 10.1088/1742-6596/110/9/092033.

[71] C. Weiser, *A combined secondary vertex based B-tagging algorithm in CMS*, tech. rep., https://cds.cern.ch/record/927399 (CERN-CMS-NOTE-2006-014, 2006), CMS-NOTE-2006-014.

[72] Tsai, Jui-Fa, *"B-tagging commissioning"*, https://github.com/alphatsai/BTaggingCommission/blob/master/README.md, [Online; accessed 24/05/2023].

[73] A. M. Sirunyan et al., “Search for nonresonant Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state at $\sqrt{s} = 13$ TeV”, Journal of high energy physics **2019**, 1–49 (2019), DOI: 10.1007/JHEP04(2019)112.

[74] P. Manzano et al., “Hemisphere mixing: a fully data-driven model of QCD multijet backgrounds for LHC searches”, arXiv preprint arXiv:1712.02538, https://doi.org/10.48550/arXiv.1712.02538 (2017), DOI: https://doi.org/10.48550/arXiv.1712.02538.

[75] A. Rogozhnikov, “Reweighting with boosted decision trees”, in Journal of physics: conference series, Vol. 762, 1 (IOP Publishing, 2016), p. 012036, DOI: 10.1088/1742-6596/762/1/012036.

[76] D. Martschei et al., “Advanced event reweighting using multivariate analysis”, in Journal of physics: conference series, Vol. 368, 1 (IOP Publishing, 2012), p. 012028, DOI: 10.1088/1742-6596/368/1/012028.

[77] A. Tumasyan et al., “Search for Higgs boson pair production in the four b quark final state in proton-proton collisions at s= 13 TeV”, Physical review letters **129**, 081802 (2022), DOI: 10.1103/PhysRevLett.129.081802.

[78] J. Alison et al., “Search for ZZ and ZH production in the four b-jet final state”, (unpublished), CMS AN-19-254.

[79] K. Albertsson et al., “Machine learning in high energy physics community white paper”, in Journal of physics: conference series, Vol. 1085 (IOP Publishing, 2018), p. 022008, DOI: 10.1088/1742-6596/1085/2/022008.

[80] I. Goodfellow et al., *Deep learning*, http://www.deeplearningbook.org (MIT press, 2016).

[81] Y. Bengio et al., *Deep learning*, Vol. 1 (MIT press Cambridge, MA, USA, 2017).

[82] E. Phaisangittisagul, “An Analysis of the Regularization Between L2 and Dropout in Single Hidden Layer Neural Network”, in 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS) (2016), pp. 174–179, DOI: 10.1109/ISMS.2016.14.

[83] J. Baxter, “A model of inductive bias learning”, CoRR **abs/1106.0245**, `https://doi.org/10.48550/arXiv.1106.0245` (2011), DOI: 10.48550/arXiv.1106.0245.

[84] S. Albawi et al., “Understanding of a convolutional neural network”, in 2017 international conference on engineering and technology (icet) (Ieee, 2017), pp. 1–6, DOI: 10.1109/ICEngTechnol.2017.8308186.

[85] S. Elfwing et al., “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning”, Neural networks **107**, 3–11 (2018), DOI: 10.1016/j.neunet.2017.12.012.

[86] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, in (2015), DOI: 10.48550/arXiv.1502.03167.

[87] E. Hoffer et al., “Train longer, generalize better: closing the generalization gap in large batch training of neural networks”, `https://doi.org/10.48550/arXiv.1705.08741` (2018), DOI: 10.48550/arXiv.1705.08741.

[88] R. K. Srivastava et al., “Training very deep networks”, **28**, `https://doi.org/10.48550/arXiv.1507.06228` (2015), DOI: 10.48550/arXiv.1507.06228.

[89] K. He et al., “Deep residual learning for image recognition”, `https://doi.org/10.48550/arXiv.1512.03385` (2016), DOI: 10.48550/arXiv.1512.03385.

[90] V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning”, arXiv, `https://doi.org/10.48550/arXiv.1603.07285` (2016), DOI: 10.48550/arXiv.1603.07285.

[91] J. Alwall et al., “Madgraph 5: going beyond”, Journal of High Energy Physics **2011**, 1–40 (2011), DOI: 10.1007/JHEP06(2011)128.

[92] S. L. Smith et al., “Don’t Decay the Learning Rate, Increase the Batch Size”, **abs/1711.00489**, `https://doi.org/10.48550/arXiv.1711.00489` (2017), DOI: 10.48550/arXiv.1711.00489.

[93] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library”, in *Advances in neural information processing systems 32* (Curran Associates, Inc., 2019), pp. 8024–8035.

[94] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization”, arXiv preprint arXiv:1412.6980, `https://doi.org/10.48550/arXiv.1412.6980` (2014), DOI: 10.48550/arXiv.1412.6980.

[95] D. A. Reynolds et al., “Gaussian mixture models.”, Encyclopedia of biometrics **741**, `https://doi.org/10.1007/978-0-387-73003-5_196` (2009), DOI: 10.1007/978-0-387-73003-5_196.

[96] A. A. Neath and J. E. Cavanaugh, “The Bayesian information criterion: background, derivation, and applications”, Wiley Interdisciplinary Reviews: Computational Statistics **4**, 199–203 (2012), DOI: 10.1002/wics.199.

[97] Y.-C. Chen, “A tutorial on kernel density estimation and recent advances”, Biostatistics & Epidemiology **1**, 161–187 (2017), DOI: 10.1080/24709360.2017.1396742.

[98] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python”, Journal of Machine Learning Research **12**, 2825–2830 (2011).

[99] P. Jawahar et al., “Improving variational autoencoders for new physics detection at the LHC with normalizing flows”, Frontiers in big Data **5**, 803685 (2022), DOI: 10.3389/fdata.2022.803685.

[100] B. Ostdiek, “Deep Set Auto Encoders for Anomaly Detection in Particle Physics”, SciPost Physics **12**, DOI: 10.21468/SciPostPhys.12.1.045, 045 (2022).