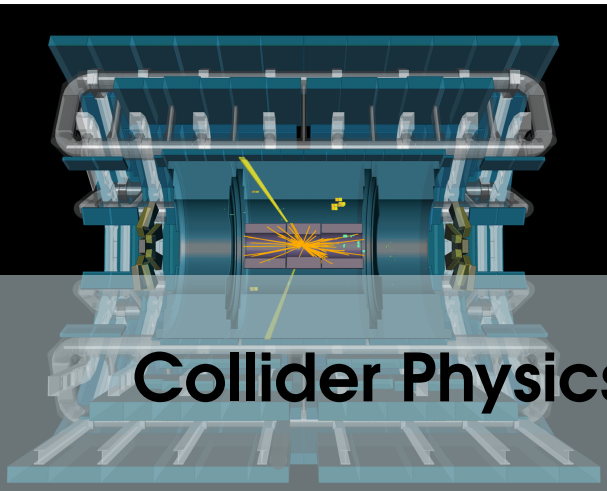
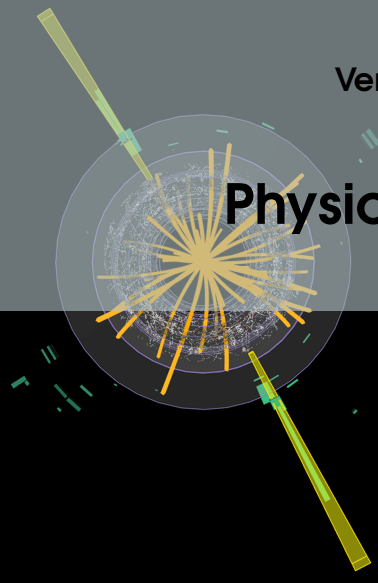
UNIVERSITY
OF SUSSEX



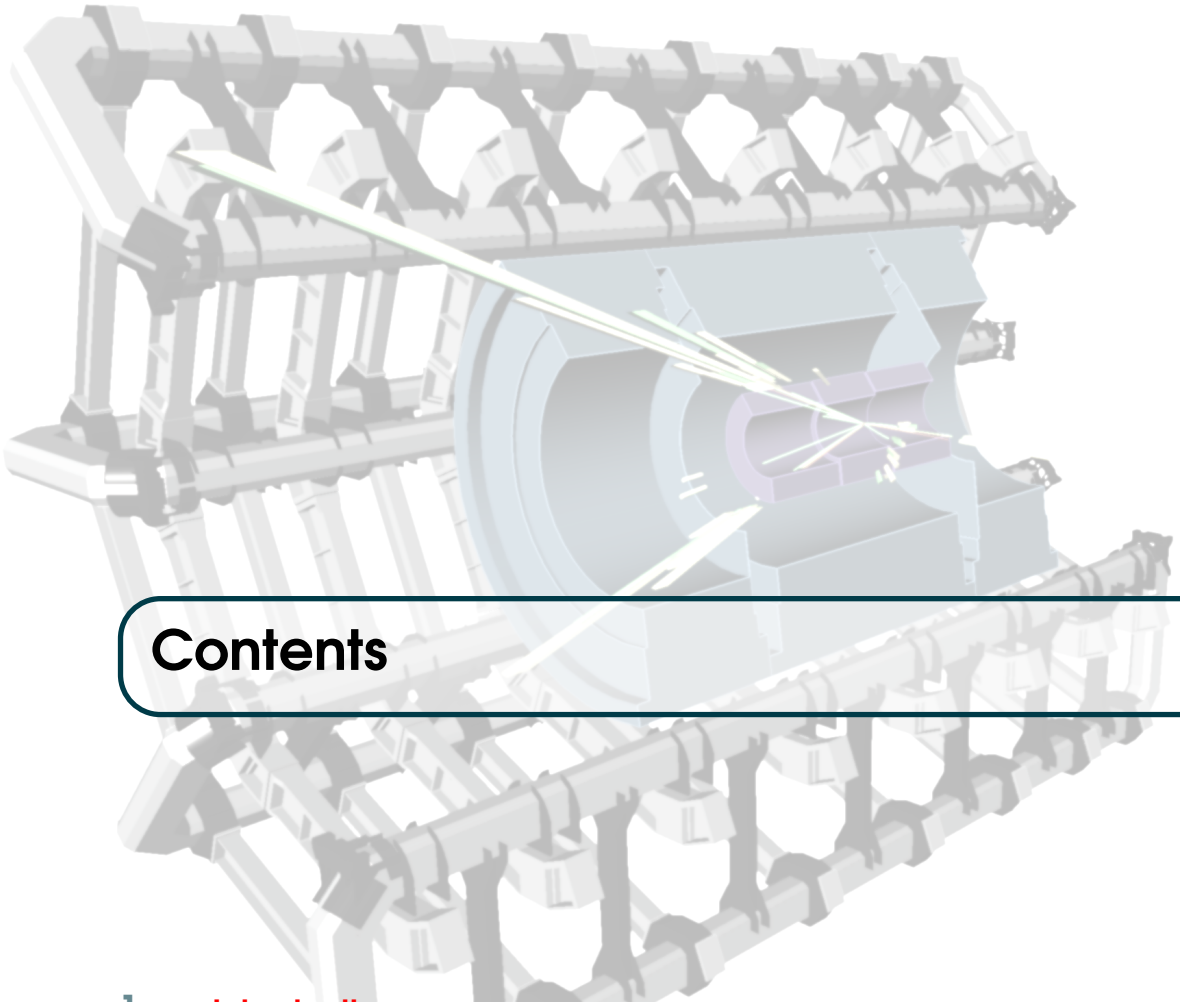
Collider Physics using ATLAS Open Data

Advanced Physics Laboratory

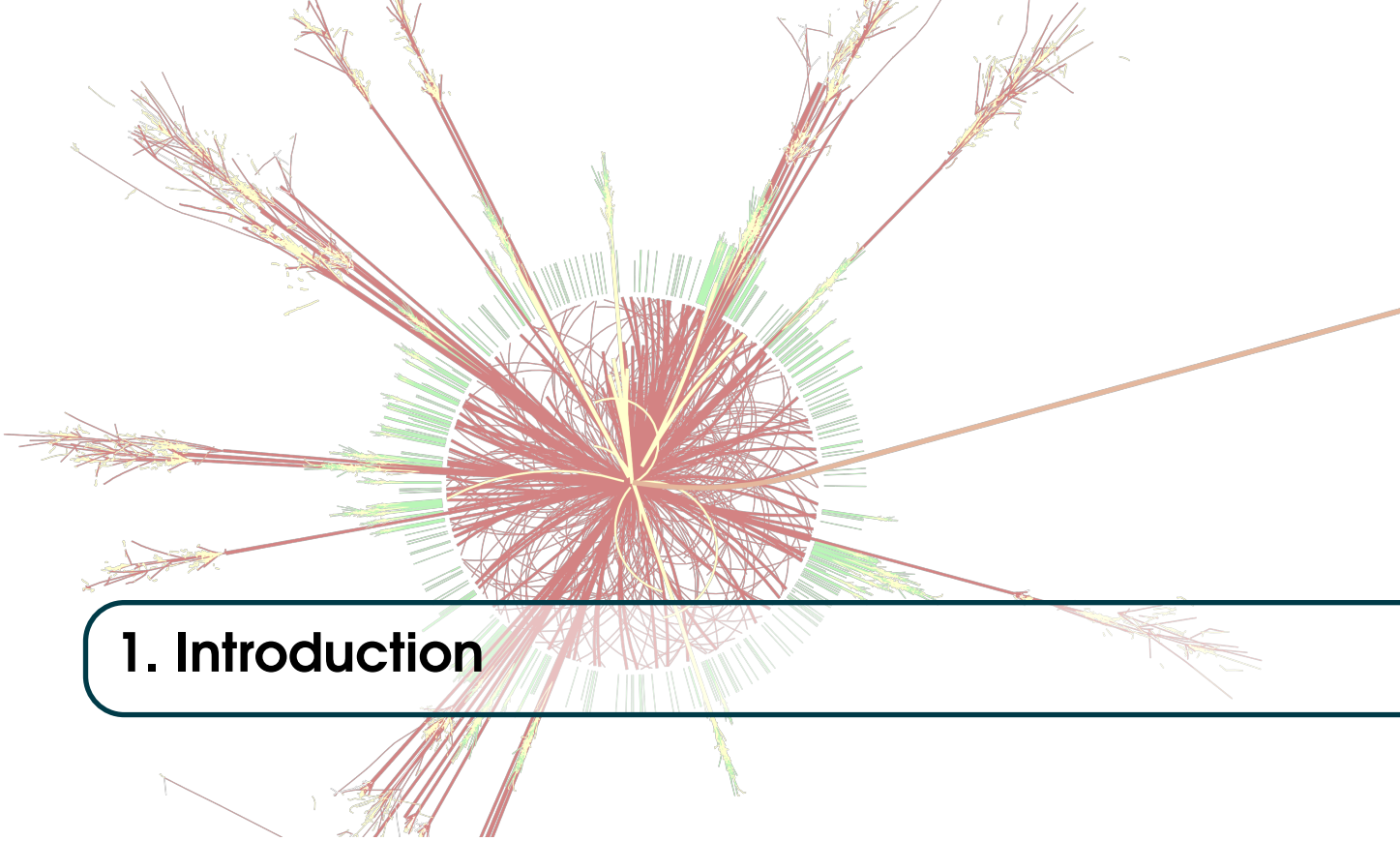
Version: 1.1 Contact: I.Vivarelli@sussex.ac.uk



Physics and Astronomy Department



Contents



1. Introduction

This Advanced Physics Laboratory experience will guide you through the main steps that a physicist working in a large particle physics collaboration (like ATLAS, CMS, but also DUNE, NOVA, etc.) has to to do look at the experiment data and draw conclusions about the underlying physics. The approach to data analysis is by the way not specific to particle physics: it is here presented in a particle physics context, but large collaborations in astrophysics and cosmology would use very similar techniques.

Specifically, you will be using simulated and real ATLAS proton-proton collisions at $\sqrt{s} =$ 13 TeV to do a series of particle physics measurements. Given the structure of the course you are following with us, your particle physics knowledge at the time of this experience will be limited: the first part of this document is a very quick and dirty introduction to the particle physics you need to know to be able to understand what you are doing when running the experiments. You will also need a review (or introduction, depending on previous experience) of some software tools that will be useful for this experience, but also for your future career, so long as it involves any software development and data analysis.

This handbook is organised as follows: this chapter lists the objectives of this experience. Chapter 2 is a review of some of the particle physics you need to know. By no means this can be a substitute of a real particle physics course, but it should be enough to run these experiments. Section 3 describes a bit more in detail a few specific aspects of particle detection. It also introduces you to some analysis techniques of a modern collider physics experiment. Section 4 describes the practical aspects to set you up to run the experiments. It also discusses how the code should be written to take the most out of this experience. Finally, Section 5 lists three experiments that should be fun to do and suitable for your level.

Feedback on the handbook and on the specific experience is more than welcome, please send it to `i.vivarelli@sussex.ac.uk`.

1.1 Objectives

The main aim of the experience connected with this handbook is to get a glimpse of frontier particle physics. You will be *seeing* the basic bricks with which our universe is composed. You will be

observing *Higgs* bosons, and top quarks and W and Z bosons. These particles form the foundations of our understanding of the electroweak interaction, mass, and stability of our universe. You will be asked to verify some of the properties of these particles, and hopefully by the end of the experience you will have a better understanding of the work that collider physicists do.

1.1.1 Learning Outcomes

By the end of this experience:

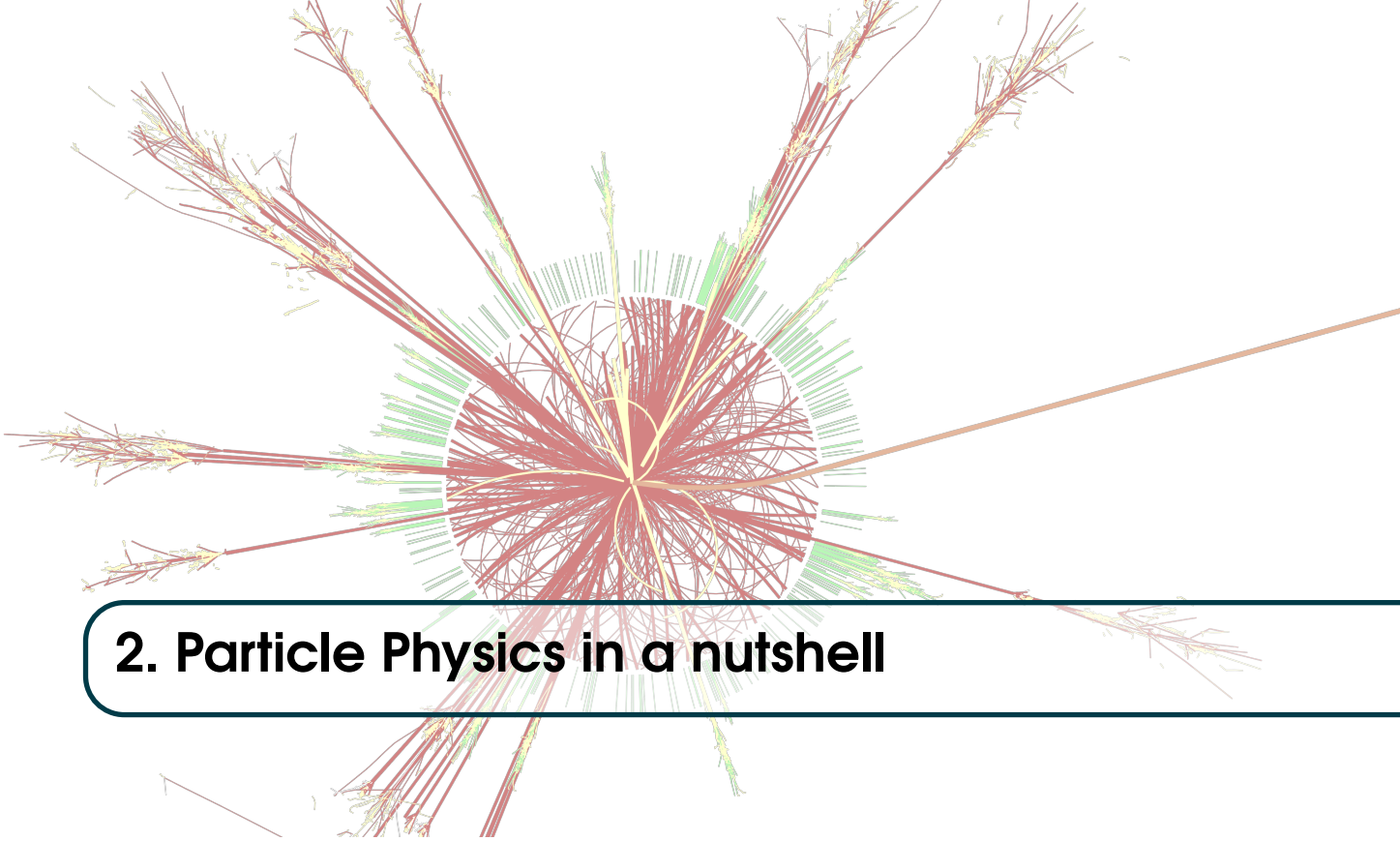
- You will have developed a good understanding of the experimental techniques used at a modern particle physics experiment through your preparation work.
- You will have performed a cutting-edge physics experiments, and applied advanced analysis techniques when running the practical execution of the experiments.
- You will have assessed the risks associated with a hypothetical running of ATLAS from its control room.

1.1.2 What you should deliver

It is expected that the preparation and findings of your experience will be summarised in a logbook. The final logbook should be delivered in pdf. It does not matter whether you prepare it in word, latex, or any other text editor, but the final result should be a pdf file. Investing in learning (or improving) your latex skills is probably a good idea in view of your final year report: if you are a newbie, why don't you have a look at Overleaf?

The logbook must contain the following entries:

- **Preparation:** Before even starting to write a single line of code, you should dedicate an extensive amount of time to prepare. The preparation includes a number of different steps:
  1. To improve your understanding of particle physics, you should read carefully Chapter 2 and answer all exercises in Chapters 2, 3 and 4. For each exercise, your logbook should contain details about your reasoning. If something is not clear, you should check the references provided. If things are still not clear, you should discuss with your tutor.
  2. You should expand on the contents of Chapter 3, and give a better description of the ATLAS detector. Chapter 1 of Ref. [1] can help a lot.
  3. You should have a risk assessment about a hypothetical data taking shift that you would take from the ATLAS Control Room. You can take some inspiration from the blog available at this link. Remember that ATLAS is buried 100 m into the ground below the control room.
- **Experiments:** You should run the experiments in Chapter 5, document what you have done, and answer all questions as quantitatively as possible. The three experiments are in order of increasing difficulty. It is better to deliver the first two experiments well done and neglect the third one, rather than delivering three rushed experiments.



2. Particle Physics in a nutshell

Particle physics is an area of physics that deals with the behaviour of the most basic constituents of matter and their mutual interaction. This section wants to serve as a **very** compact summary of the main aspects of particle physics that you need to know to be able to run this experiment. The book which is used in the particle physics module here at the University of Sussex is the one listed in Ref. [4]: you may want to have a look there for more details.

Particles are very small, and, for what concern this experiment, very fast as well. A mathematically sound physical theory of particle physics needs to be compliant with the laws of quantum mechanics and (special) relativity. The standard formulation of particle physics is done in terms of a Quantum Field Theory. We will avoid almost any reference to QFT: this will imply that in some places you will have to "believe" some of the results that will be presented.

Depending of the property one is interested in, particles can be grouped in a few different ways:

- If one is interested in their spin properties, particles are divided into **fermions**, whose spin is fractional (like $1/2$, $3/2$, etc.), and **bosons**, whose spin is integer. All elementary particles that constitute ordinary matter (the up and down quarks, the electrons) are fermions. All particles that mediate an interaction (the photon, the Higgs boson, etc.) are bosons.
- If one is interested in their spacial size, particles are divided into **elementary** (more on them later), that should be imagined essentially as a geometrical point in space, with no physical size, and **composite**, that is, bound states of other particles. Examples of elementary particles are the electron and the quarks. Composite particle examples are the proton and the neutron.
- If one is interested in the interactions they feel, particles are divided into **leptons**, that interact through gravity and the electroweak force, and **hadrons** which interact through gravity, the electroweak force and the strong nuclear force. The electron and neutrinos are leptons, the quarks are hadrons.

Other groupings are possible: we will introduce them if and when they will be needed. In the following, the word "particle" will be used with the meaning of "elementary particle", unless stated otherwise.

Particles are objects with a mass of the order of that of the proton. There are large variations on this order of magnitude: the heaviest known elementary particle is the top quark, whose mass

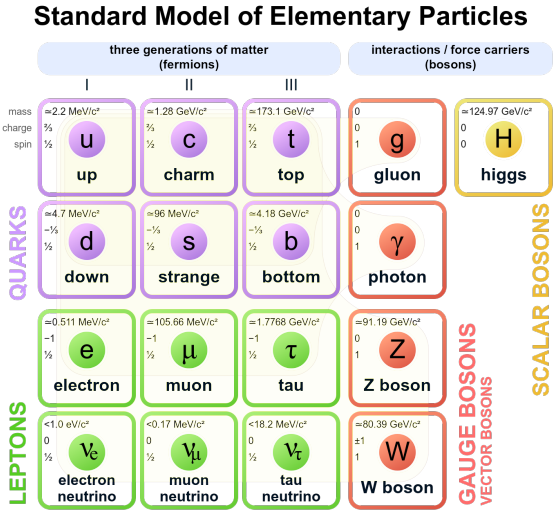


Figure 2.1: Particle content of the Standard Model of Particle Physics. For each particle, the mass, the charge and the spin are reported. Please bear in mind that the mass of light quarks (u, d, s) is not straightforward to define and estimate, given that no free quark can be observed. Taken from Ref. [7].

is about 172 times that of the proton. The electron is the lightest massive particle, with a mass of about half of a thousandth of a proton. Massless particles exist as well (the photon mass is zero, that of the neutrinos is very small although not zero).

2.1 The Standard Model of Particle Physics

The best theory of particle physics we have available is the so–called Standard Model of Particle Physics (SM in the following). It specifies the full list of particles, their properties and their interactions. Table 2.1 summarises the particle content of the SM. There are three leptonic families (in green). In each leptonic family, there is a charged lepton and a neutrino of the same flavour. There are also three quark families (in purple). In each quark family there is a quark with charge $+\frac{2}{3}$ and one with charge $-\frac{1}{3}$. Each quark defines its own flavour. Not shown in the table: for each particle family there is an identical family composed by anti-particles. So, for the family composed by the electron and its neutrino, there is a family composed by a positively charged electron (the positron) and an antineutrino, and same for the quarks.

There are also bosons, that mediate the interactions between the particles in the leptonic and quark families (in red). These are:

- The photon, that mediates the electromagnetic interaction between charged particles.
- The W^+, W^- and Z bosons, which mediate the weak interaction between leptons and between quarks.
- The gluons, that mediate the strong nuclear interaction between quarks.

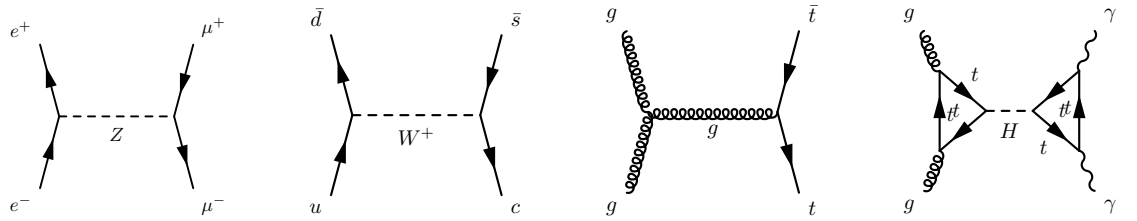
Finally, the Higgs boson is shown in yellow: it interacts with all particles with an intensity that is proportional to the particle mass.

We will use a helpful graphical tool to represent interactions between particles. These are the Feynman diagrams. Please bear in mind that we will use them only as a graphical tool, but they are actually a representation of quantitative mathematical expressions that allow to precisely calculate how likely a given process is to occur. In these graphs, we will always represent incoming particles on the left and outgoing particles on the right.

There are important rules that have to be respected when drawing possible Feynman diagrams:

1. Electric charge, momentum, energy and angular momentum are always conserved at a vertex.
2. Fermions are represented with arrows. An incoming (in a vertex) fermion is represented with an incoming arrow, while an outgoing fermion is represented with an outgoing arrow. For anti-fermions, rules are reversed: an incoming anti-fermion is represented with an outgoing arrow, while an outgoing anti-fermion is represented with an incoming arrow. See examples in Figure 2.2.
3. The total number of leptons has to be conserved. Each lepton contributes to the number of leptons with an additive $+1$, while each anti-lepton with an additive -1.
4. Likewise, the total number of quarks has to be conserved. Each quark contributes with an additive $+\frac{1}{3}$, while each anti-quark with an additive $-\frac{1}{3}$.
5. As a consequence of rule 1 for angular momentum, a three-line vertex can only be: two fermions and a boson; three bosons.
6. Gluons interact only with quarks and other gluons.
7. The Z and γ bosons always conserve the flavour. So, Ze^+e^-, $Zu\bar{u}$, $\gamma c\bar{c}$ are perfectly legal vertices, $Ze^+\mu^-$, $Zc\bar{s}$ and $\gamma e^-\mu^+$ are not.
8. The Z boson does not interact with the photon.
9. The W boson interaction conserves the leptonic flavour. So $We\bar{\nu}_e$ is a legal vertex, but $We\bar{\nu}_\mu$ is not.

Figure 2.2 [1] shows examples of allowed Feynman diagrams, while Figure 2.3 shows some processes which are not allowed.



(a) Production of a Z boson decaying into a $\mu\bar{\mu}$ pair at an electron-positron collider.

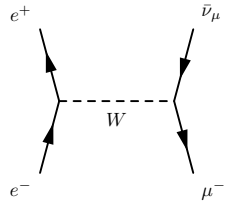
(b) A diagram for W production at a hadron collider, followed by the W decay into two quarks.

(c) A diagram for $t\bar{t}$ production at a hadron collider.

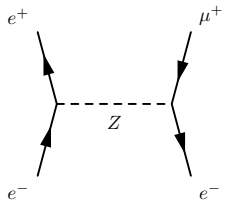
(d) Higgs boson production via the so-called gluon fusion process, followed by $H \to \gamma\gamma$, at a hadron collider. Both the gluons and the photons are massless particles, the Higgs boson couples to them through intermediate particles.

Figure 2.2: Examples of valid Feynman diagrams. For each of them, the fermion lines do not stop within the diagram, rather they enter and exit the diagram fully. This is a consequence of angular momentum and charge conservation.

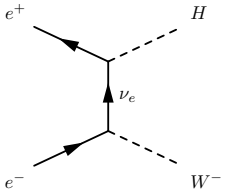
[1] In this handbook, we will always deal with so-called *leading-order* Feynman diagrams. If you want to know more about *higher order* Feynman diagrams, please consult Ref. [4]. We will also neglect possible leading order Feynman diagram vertices that include more than three lines.



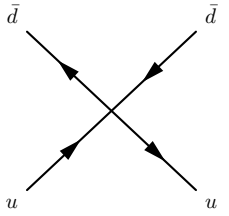
(a) Violation of charge conservation at the W production vertex.



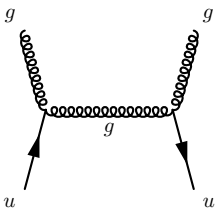
(b) Flavour violation at the Z decay vertex. The Z boson only couples to leptons and quarks of the same flavour.



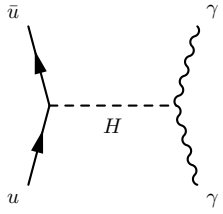
(c) Charge conservation violation at the Higgs boson production vertex.



(d) Four-fermion interaction vertices do not exist in the Standard Model.



(e) There is a broken fermion line at both vertices (violation of charge, angular momentum, baryon number conservation).



(f) The Higgs boson does not couple directly with particles with no mass.

Figure 2.3: Examples of processes which are NOT allowed in the Standard Model, with a quick explanation of why they are not allowed.

2.2 Special relativity and units

In particle physics it is customary to choose a system of units in which the value of the Planck constant $\hbar$ and the speed of light c are both set to unity: $\hbar = c = 1$. Energy and momentum are therefore measured with the same units. Often the electronvolt (eV) is the energy unit chosen. For the applications of this experiment, we will be using MeV, GeV and TeV. Because of the Heisenberg uncertainty principle, the unit for space is also fixed:

$$\Delta p \Delta x \sim \hbar = 1 \implies [x] = \frac{1}{[p]} = \frac{1}{\text{MeV}} \tag{2.1}$$

Likewise, from $\Delta E \Delta t \sim \hbar = 1$, it follows that time has the same units of space. From $E = mc^2$, it follows that mass and energy have the same units.

The particle free motion is described in terms of their energy E and momentum $\mathbf{p}$. The three momentum components and the energy form a so-called 4-vector. You already have the knowledge of what a 4-vector is, but probably nobody has used this name just yet. Let's start from a standard position vector in space $\mathbf{x}$, with components (x, y, z) in a given reference frame S. Let's name this a 3-vector (and you can probably already guess where I am going...). You know well that in special relatively the spacial coordinates and the time of a given event cannot be treated independently from each other, simply because they mix if one describes the same event from a different reference frame in uniform motion with respect to the original one. For a change of reference frame from a frame S to one S' moving with speed v along $\hat{x}$ with respect to S, the 4-vector transforms into $(\mathbf{x}', t')$ such that:

$$
\begin{aligned}
x' &= \gamma(x+\beta t) \\
y' &= y \\
z' &= z \\
t' &= \gamma(t+\beta x)
\end{aligned} \tag{2.2}
$$

where we are using units in which the speed of light c is equal to 1. x' is a function of both x and t, and the same is true for t'. Perfect! Let's then define a 4-vector as $x_\mu = (x,y,z,t)$. The components of a 4-vector transform by Lorentz boost as in Eq. 2.2.

The symbols γ and β are defined (for $c = 1$) by:

$$
\gamma = \frac{1}{\sqrt{1-v^2}} \qquad \beta = \frac{v}{c} = v \tag{2.3}
$$

The transformations in Eq. 2.2 are such that the product

$$
s = t^2 - \mathbf{x}\cdot\mathbf{x} \tag{2.4}
$$

is invariant for Lorentz transformations, that is, its value is the same in the reference frame S or in any other boosted frame S'.

Exercise 2.1 Prove it! Prove that $t^2 - \mathbf{x}\cdot\mathbf{x} = t'^2 - \mathbf{x}'\cdot\mathbf{x}'$, where the primed quantities are connected to the non-primed ones by a Lorentz transformation like in eq. 2.2 ∎

So, momentum and energy form another 4-vector, $p_\mu = (p_x, p_y, p_z, E)$. That means that if I go from one reference frame to another in relative motion along $\hat{x}$,

$$
\begin{aligned}
p'_x &= \gamma(p_x+\beta E) \\
p'_y &= p_y \\
p'_z &= p_z \\
E' &= \gamma(E+\beta p_x)
\end{aligned} \tag{2.5}
$$

It also means that $E^2 - \mathbf{p}\cdot\mathbf{p}$ is an invariant.....

2.2.1 Invariant Mass

The fact that for a 4-vector of type $(\mathbf{m}_x, m_t)$ the product $s_m = m_t^2 - \mathbf{m}_x\cdot\mathbf{m}_x$ is Lorentz-invariant (that is, it is the same in every reference frame, regardless of its boost) is general. The number $\sqrt{s_m} = \sqrt{m_t^2 - \mathbf{m}_x\cdot\mathbf{m}_x}$ is the *magnitude* of the 4-vector. As pointed out earlier, the energy and momentum form a 4-vector $(\mathbf{p}, E)$. Therefore the combination $s = E^2 - p^2$ (where p indicates the magnitude of $\mathbf{p}$) is Lorentz-invariant. What is the value of this Lorentz-invariant quantity? Let's consider a particle of mass m. In its own rest frame:

$$E = m \tag{2.6}$$

$$\mathbf{p} = 0 \tag{2.7}$$

$$s = E^2 - p^2 = m^2 \tag{2.8}$$

But since s is Lorentz-invariant, *this mathematical combination of the momentum and the energy of the particle will yield the particle mass regardless of the rest frame where it is computed!*. We will use this fact repeatedly.

2.3 Proton-proton Collisions - Collider Physics Variables

In this experience, we will study proton-proton collisions produced by the LHC and recorded by ATLAS. The LHC collides two beams of protons head-on. The energy of the protons in each beam is $E_{\text{beam}} = 6.5$ TeV.

As said earlier, protons are not elementary particles. They are made by three *valence* quarks (two u and one d) and a *sea* of other quark-antiquark pairs and gluons, which are created and destroyed continuously according to quantum mechanics. All these particles (the valence quarks and the sea quarks and gluons) take the generic name of *partons*.

The scale for the proton binding energy is given by its own mass (~ 1 GeV). With respect to the typical energy transfer in a LHC collision of interest, with energy exchanged of hundreds of GeV, the binding energy of the proton is small. Therefore the partons in the proton in a LHC collision **can be considered as free**: the LHC is indeed a parton collider.

Partons inside a proton however do not have the same energy/momentum as the proton of course: it is the sum of the energies and momenta of all partons that will yield the energy and momentum of the proton. Let's say that each parton i in the collision carries a fraction x_i of the momentum and energy of the proton. The 4-vectors of the protons and colliding partons are (neglecting all masses, and setting the $\hat{z}$ as beam axis)

$$\text{Proton1} : (0,0,E_{\text{beam}},E_{\text{beam}}) \tag{2.9}$$
$$\text{Proton2} : (0,0,-E_{\text{beam}},E_{\text{beam}}) \tag{2.10}$$
$$\text{Parton1} : (0,0,x_1E_{\text{beam}},x_1E_{\text{beam}}) \tag{2.11}$$
$$\text{Parton2} : (0,0,-x_2E_{\text{beam}},x_2E_{\text{beam}}) \tag{2.12}$$
$$\tag{2.13}$$

The magnitude of the 4-vector corresponding to the sum of the two proton 4-vectors is

$$\sqrt{s} = \sqrt{(2E_{\text{beam}})^2 - (E_{\text{beam}} - E_{\text{beam}})^2} = \sqrt{4E_{\text{beam}}^2} = 2E_{\text{beam}} = 13 \text{ TeV} \tag{2.14}$$

which is known as the centre-of-mass energy of the LHC. However, this is **not** the centre of mass energy of the parton-parton collisions! The actual particles colliding are the partons. Their own centre-of mass energy is:

$$\sqrt{\hat{s}} = \sqrt{(x_1+x_2)^2 E_{\text{beam}}^2 - (x_1-x_2)^2 E_{\text{beam}}^2}$$

Exercise 2.2 Prove that this implies $\sqrt{\hat{s}} = \sqrt{x_1x_2}\sqrt{s}$ ∎

By the way: the frame in which the proton-proton collision happens at rest (the laboratory frame) does not coincide with that in which the parton collision happens at rest. With respect to the laboratory frame, the parton-parton collision happens in a system which is boosted along the beam axis by $p_{\text{boost}} = |x_1 - x_2|E_{\text{beam}}$.

Let's recap: on an **event-by-event basis**, the parton-parton collision happens in a frame with a different boost with respect to the laboratory frame, and the collision energy is smaller than the nominal proton-proton centre-of-mass energy.

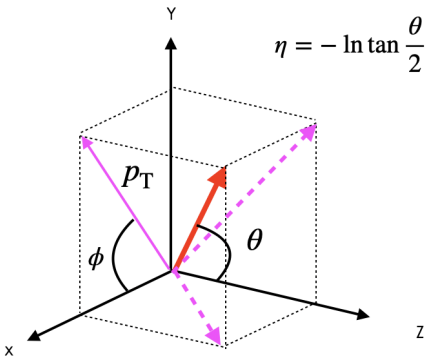


Figure 2.4: Diagram of the kinematic variables used at a hadron collider. The z axis coincides with the beam axis.

Exercise 2.3 The mass of the top quark is 172 GeV. The SppS, a collider colliding protons and antiprotons at CERN, worked up to energies of $\sqrt{s} = 900$ GeV. Yet, we had to wait for the Tevatron ($\sqrt{s} = 1.8$ TeV) for the top quark discovery. Why? ■

Exercise 2.4 Draw one of the most relevant Feynman diagram for the production of the following particles at the LHC. If quarks are involved, specify if they are likely to be valence or sea quarks: Z boson production, W boson production, $t\bar{t}$ production. ■

Exercise 2.5 The mass of the W boson is $m = 80$ GeV. In a given collision pp collision at $\sqrt{s} = 13$ TeV, the W boson is produced by a collision of a valence quark with $x = 0.3$, and a sea quark. Find x of the sea quark and compute p_{boost} for this collision. ■

So, in practice what happens is that in each event, the LHC experiments look at a collision happening with an unknown boost along the $\hat{z}$ axis. There is one important implication: if we measure any variable which is not Lorentz-invariant, we would make a big confusion when measuring it in many events, because we would mix up values belonging to different reference frames! This implies we have to carefully choose the quantities that we measure at a hadron collider like the LHC to characterise the particles' final states: we should avoid using any variable relying on *longitudinal* quantities, that is, quantities relying on variables measured along the $\hat{z}$ axis. The reason Therefore, the variables we use at a hadron collider are (see also Figure 2.4)

- The momentum in a plane transverse to the beam, p_T.
- The angle ϕ in the transverse plane with respect to the direction pointing towards the LHC centre.
- The mass.

And now we have a problem, because three variables do not determine a 4-vector, of course. We need somewhat to provide the direction with respect to the beam axis. This is done with the rapidity y variable, or with its approximated variable (valid for massless particles) the pseudorapidity η.

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

2.4 Cross-section and luminosity

Imagine you have a beam of particles impinging on a slab of targets, and that the targets are small enough that you can neglect shielding effects of one target with respect to another. The total number of interactions that you will observe per unit time will be determined by:

- characteristics of the beam of particles (how many particles per unit surface per unit time you are shooting, that is, the incoming flux of particles),
- the target density,
- the transverse area that the targets offer to the beam, that is, the cross-section of the individual targets.

Inheriting this language, we define the cross-section and the luminosity of a given process at a hadron collider such that the number of occurrences of that process that we observe per unit time is given by:

$$\frac{dN}{dt} = \sigma \times \mathscr{L} \tag{2.15}$$

In the example above, the luminosity is something connected with how many protons per beam we have, how many beams are circling the LHC, etc., while the cross-section is intrinsically connected with the physical interaction yielding a given process. The cross-section units are those of a surface (cm^2, for example[2]), while those of the luminosity are

$$[\mathscr{L}] = [\text{surface}]^{-1} \cdot [\text{time}]^{-1}. \tag{2.16}$$

A typical number $\mathscr{L}$ during the LHC Run 2 is $\mathscr{L} \sim 2 \times 10^{34}\ \mathrm{cm}^{-2}\ \mathrm{s}^{-1}$. The total number of events we will count for a given process in a given period of time is of course given by

$$N = \int_{\Delta T} \frac{dN}{dt} dt = \int_{\Delta T} \sigma \times \mathscr{L}\, dt = \sigma \int_{\Delta T} \mathscr{L}\, dt = \sigma \mathscr{L}_{\text{int}}, \tag{2.17}$$
$$\mathscr{L}_{\text{int}} = \int_{\Delta T} \mathscr{L}\, dt. \tag{2.18}$$

The data collected by a given experiment over a period of time are typically expressed with the corresponding *integrated luminosity* $\mathscr{L}_{\text{int}}$. Let's make an example. The ATLAS experiment has collected a number of proton-proton collisions corresponding to $\mathscr{L}_{\text{int}} = 139\ \mathrm{fb}^{-1}$ during the so-called Run 2 at $\sqrt{s} = 13$ TeV. The production cross-section for top pairs is predicted to be $\sigma_{t\bar{t}} = 831$ pb. Let's compute how many top pair production events we expect to have collected during the full Run 2:

$$N_{t\bar{t}} = \sigma_{t\bar{t}} \times \mathscr{L}_{\text{int}} = 831 \times 10^{-12}\ \text{barns} \times 139 \times \left(10^{-15}\ \text{barns}\right)^{-1} \tag{2.19}$$
$$N_{t\bar{t}} = 1.16 \times 10^{8} \tag{2.20}$$

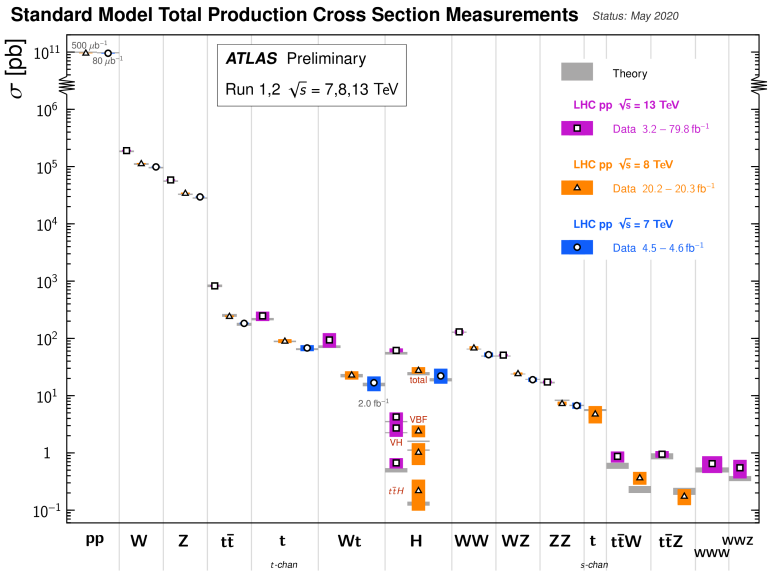


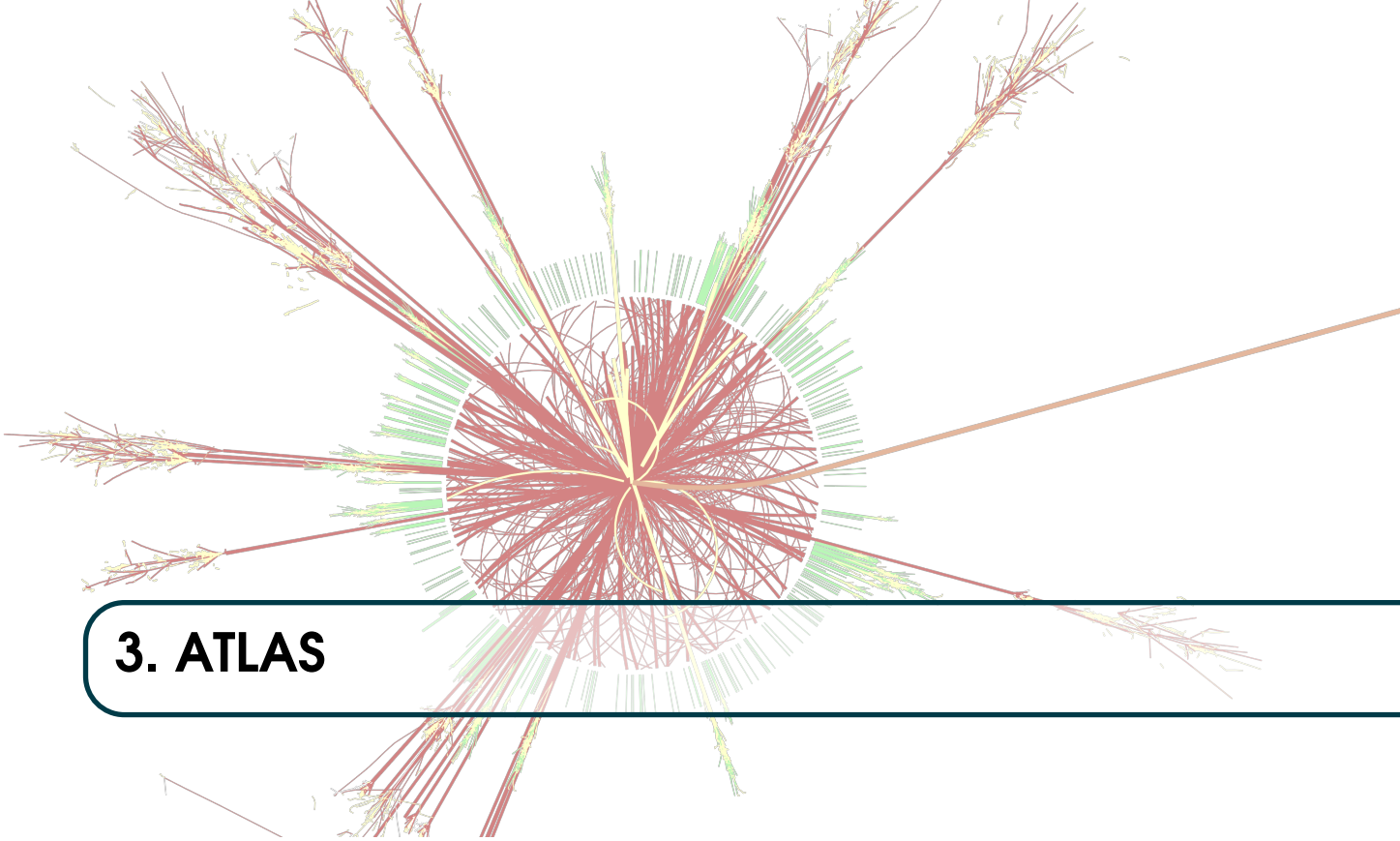
Figure 2.5: Predictions and ATLAS measurements for the production cross sections of several Standard Model processes. Taken from Ref. [5].

So, we expect something like 116 million top pairs to have been produced during Run 2 in ATLAS.

Now, let's take a close look at Figure 2.5. It shows the cross-sections for several SM processes at the LHC, for different proton-proton centre-of-mass energies. The proton-proton inelastic cross-section at $\sqrt{s} = 13$ TeV is about 80 mb. Let's consider one random process as an example, say WW production. The plot tells us that the production cross section is between 100 and 200 pb - let's take 160 pb. So, from the ratio of these numbers, we learn that only one in 500,000 proton-proton collisions will produce WW.

Exercise 2.6 From Figure 2.5, the Higgs boson production cross-section is $\sigma = 80$ pb. Compute how many Higgs bosons LHC has produced during Run 2, and using the typical value for $\mathscr{L}$ given above, compute how many Higgs bosons per minute the LHC is producing when running at that luminosity. ■

[2]Particle physics cross sections are normally expressed in barns, where 1 barn $= 10^{-24}$ cm^2



3. ATLAS

We will now go though the main aspects of how different particles are reconstructed by ATLAS. This discussion will be crucial to understand the variables that are available for this experiment.

3.1 The ATLAS detector

ATLAS is a general-purpose particle detector operating at a hadron collider. The structure of the detector is sometimes referred to as onion-like, because different layers of detectors are present when going from the collision point outwards. A detailed description of the ATLAS detector and the technologies used for particle detection goes well below the scope of this document: if you are interested please refer to Ref. [1]. What we will do here is to summarise the main principles behind particle detection.

Figure 3.1 shows the various components of ATLAS. Starting from the collision point and going outward:

- **Inner Detector (ID):** Starting at about 3 cm away from the collision point, the inner detector is optimised to measure the position of charged particles. It is immersed in a magnetic field directed along the z axis, such that charged particles trajectories are bent. From the curvature radius it is possible to measure the particle momentum transverse to the magnetic field.

  Exercise 3.1 Remembering that $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$ is acting as centripetal force, estimate the curvature radius of a particle with unit charge and momentum $p = 10$ GeV if $B = 2$ T. ■

- **Electromagnetic (EM) calorimeter:** It stops electrons and photons and measures their energy. It also participate in the energy measurement of hadrons.
- **Hadronic (HAD) calorimeter:** Together with the electromagnetic calorimeter, it measures the energy of neutral and charged hadrons.
- **Muon spectrometer (MS):** Muons and neutrinos are the only particles in the SM that are able to exit the ATLAS calorimeters. The principle of the muon spectrometer is similar to that of the inner detector: the position of muons is measured several times while they are bending in a magnetic field generated by the air-core toroids. From that, a second measurement of the

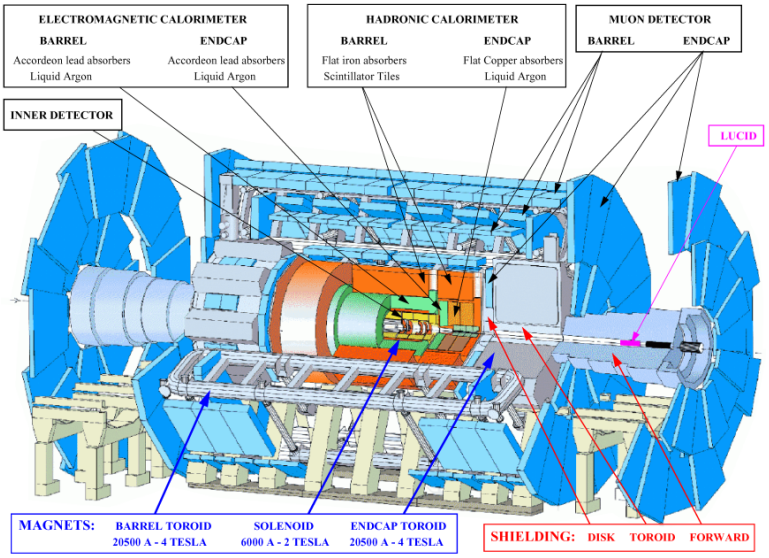


Figure 3.1: Overview of the ATLAS detector. Taken from Ref. [6]

muon momentum is performed, which is typically combined with the first one taken by the ID.

The subdetectors outlined above are used in combination to reconstruct a number of different final state "objects". The number of types of objects which is normally used in an analysis is actually relatively limited. The main features of each of these objects is reviewed below.

- Electrons: they are identified by a cluster of energy in the EM calorimeter matched with a track in the ID. *Shower shape variables* (that is, the shape of the cluster in the calorimeter), the track quality, and the quality of the matching between the cluster and the track (are the track momentum and cluster energy compatible?) determine the identification.
- Photons: similar to electrons for the cluster in the EM calorimeter, but with no associated track in the ID.
- Muons: a track in the MS is almost uniquely identifying a muon (no other charged particle pass through the calorimeters). Typically the MS track is matched to one in the ID.
- Jets: Quarks and gluons cannot emerge from a collision as such. They go through a process of fragmentation and hadronisation [3], resulting in a jet of collimated hadrons, whose total momentum and direction resembles that of the originating quark/gluon. Such jets are reconstructed mainly from the HAD and EM calorimeter information, but they use some information from the ID as well.
- b-jets: b-quarks have a longer lifetime than the other quarks. They travel of the order of a cm or less before decaying. The presence of such secondary vertex (that is, a vertex that does not coincide with the primary one from the pp collision in the transverse plane) can be exploited to "tag" the jet as originating from a b.
- τ leptons: The lifetime of the τ is extremely short. For all practical purposes, it decays before producing any visible effect in the detector. There are two main categories of τ decays: the leptonic ones and the hadronic ones. If a τ decays to an electron and muon (through $\tau \to \ell \bar{\nu}_\ell \nu_\tau$ to conserve the lepton number and flavour), it will be seen as a lepton in the detector. If instead a tau decays to hadrons, then it looks like a very narrow jet with few tracks attached, and it can be identified based on these characteristics.
- Invisible particles: If an invisible particle is produced (neutrinos, in the standard model), their presence is inferred from the measurement of the total momentum transverse to the beam.

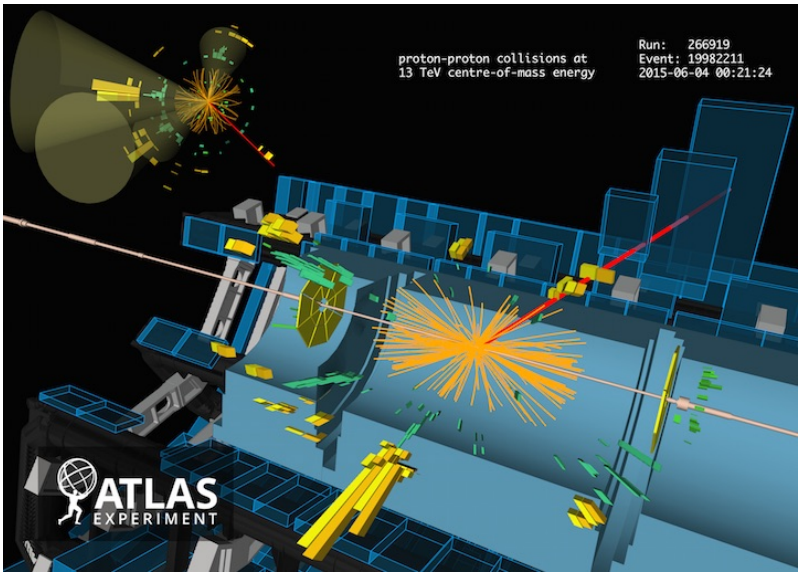


Figure 3.2: Display of a $t\bar{t}$ candidate event from proton-proton collisions recorded by ATLAS at a collision energy of 13 TeV. The red line shows the path of a muon with transverse momentum around 35 GeV through the detector. The green and yellow bars indicate energy deposits in the EM and HAD. From close-by deposits in these calorimeters, four jets are identified with transverse momenta between 25 and 80 GeV. Taken from Ref. [2]

Given that the momentum in the plane transverse to the beam is zero before the proton-proton collision, it has to be zero after the collision. The negative vectorial sum of the transverse momenta of all object in the events is called the *missing transverse momentum*, and is an estimate of the total transverse momentum of invisible particles.

Figure 3.1 shows a candidate event for $t\bar{t} \to \mu\nu_\mu + 4$ jets (where no distinction is made between particle and antiparticle for the muon and neutrino, given that either the top or the anti-top quark could be decaying into a muon). See the caption for further details.

3.2 Monte Carlo and real events

In this experiment you will deal with two types of events: simulated and real ones. The simulated ones are often referred to as Monte Carlo, or simply MC, events. MC events are the result of a computer simulation, that emulates at best the theory we have for the proton-proton collision. They simulate a specific process (like Z production, or H production), they save the so called truth-level information (that is, the particles that were produced in the collision and that enter the detector) and they also simulate the response of the detector (reco-level). The MC represents our theoretical prediction.

4. Open data: practical guide

The ATLAS open data consist in a set of MC and real events that the ATLAS collaboration has released for open use. The main web page where they are introduced is `http://atlas.cern/resources/opendata`. Please spend some time on this page, explore it, play with it.

The next step is to actually do something. In the "Online Open Data Analysis" section of the page, click on "Analyse ATLAS Open Data with Jupyter Notebooks". You should get to this page. Now select `13-TeV-examples`, then `uproot_python`. Each notebook corresponds to a different analysis. Please open `HZZAnalysis.ipynb`, and read it through. The notebook should work (if it does not, please get in touch with one of the tutors).

Let's dissect what this notebook does.

4.1 Review of the jupyter notebook step by step

This notebook runs over the input data and MC to produce a plot of the four-lepton invariant mass starting from the information about the lepton 4-vectors.

- The first cell is a technical one, and it takes care of installing a number of modules that may not be available on the machine. In particular note the module `uproot`, which is the one that will allow us to decouple completely from root, despite the open data being physically saved as `root` files.
- The second cell imports a number of modules which will be necessary for the analysis. You should be more or less familiar with everything, with the possible exception of `uproot`. Google is your friend, please have a look at any module you are not specifically familiar with.
- The third cell contains a few important things. First of all, it specifies two float numbers (`lumi` and `fraction`). `lumi` specifies what total $\mathscr{L}_{\text{int}}$ the data that we will be using correspond to (10 fb^{-1} in this case). Running on all data can take time. In case you only want to quickly check whether your code is working, you can only run on a `fraction` of all available data. Finally (very important!), `tuple_path` specifies where the data and simulated samples are. This example uses the input samples from a web location. Depending on the quality of your connection and the storage capabilities of your machine, you may find faster to download the

samples to your machine and run locally on them.
The location specified is actually browsable. You can copy `https://atlas-opendata.web.cern.ch/atlas-opendata/samples/2020/4lep/` in a new browser tab. As the link suggests, this folder contains events containing four leptons. There are two subfolders, `Data` and `MC`. I am sure you guessed already what they contain. Let's explore the `Data` one first. This is not very interesting: the available data are split in four files. A lot more interesting is the `MC` one. It contains simulated collisions yielding specific processes. A word on the naming convention:
  - `mc` at the beginning specifies these are MC samples.
  - the number that follows is a unique numerical identifier of the sample, called *dataset identifier*, or DSID.
  - then you have a string that characterises the process. For example, `WH125_ZZ4lep` means this is a process where WH is produced, the Higgs boson decays into two Z bosons, which in turn produce four leptons.
  - `4lep` means.... well, you know.
  - the `root` extension specifies the type of file.

  In related folders, you also have other types of events. If you go to `https://atlas-opendata.web.cern.ch/atlas-opendata/samples/2020/` you will get an overview of what type of events are available. We will be using those with two and three leptons in our bonus exercise.
- The fourth cell defines a dictionary for the samples we will be using, exploiting one of the field of the MC naming convention discussed above. This analysis will look for $pp \to H \to ZZ \to 4\ell$. In this context, "signal" is any process that contains a Higgs boson production giving four leptons in the final state. The "background" is any other SM process that would yield four leptons in the final state. The main contribution to the background is the production of two Z bosons, and that is why this process is singled out. Other processes that could give four leptons are listed separately.

  You may wonder how, for example, `Zee` (that is Z boson production followed by $Z \to e^+e^-$) can give four leptons in the final state. That is a very good question, and it gives the chance to introduce the concept of *reducible* and *irreducible* background. $pp \to ZZ \to 4\ell$ is an irreducible background to our Higgs analysis, because it gives a final state with the same particle content of our signal. $Z \to e^+e^-$ is a reducible background. That means it normally does not produce a final state with the same particle content as the signal, but because of rare features (either physical, or from the detector), it may actually yield the same final state. In this case, in a small fraction of events one of the electrons in the final state may radiate a photon that could convert in a second electron pair.

- The fifth cell creates a function to loop over the data and MC samples, relying on a `read_file()` function that will be defined later.
- The next two cells define the `calc_weight` and `get_xsec_weight` functions. The MC events need a series of corrections to truthfully represent the data. For example, the electron identification efficiency predicted by the plain MC might be a bit optimistic or pessimistic with respect to that that we have with the real detector. Therefore the MC is corrected with a factor `scaleFactor_ELE` to take this into account. The other factors appearing in the `calc_weight` have similar explanations. `xsec_weight` deserves a dedicated explanation

  Definition 4.1.1 Cross-section weight: Suppose that I simulate N events of a process whose cross-section is σ. I know that the data I will be looking at correspond to an integrated luminosity L. To help understanding, let's use some numbers: $\sigma = 10$ fb,

$L = 10\ \text{fb}^{-1}$, $N = 100000$. I now want to use my N MC events to predict the yields in data. Clearly I need to rescale my MC events in some way. I will need to apply a multiplicative weight w to each of my N events such that they will eventually represent the number of events I expect in data, that is (see Section 2.4) $N_{\text{data}} = \sigma \times L$

$$wN = N_{\text{data}} = \sigma \times L$$
$$w = \frac{\sigma \times L}{N}$$

With our specific numbers above: we expect $N_{\text{data}} = 100$ events in data, and we have $N = 100000$ to represent them. We should therefore multiply our N MC events by $w = N_{\text{data}}/N = 10^{-3}$.

That is in essence what is in these cells, with just slight differences.

- The next cell simply defines to compute the invariant mass of the leading four leptons in the event.

Exercise 4.1 From the definition of invariant mass in Section 2.2.1 and the definition of η, ϕ, p_{T} in Section 2.3, show that the calculation in this cell is correct. ∎

- The next cell defines two boolean functions (that is, functions whose outcome is either True or False) that will later be used to filter the events.
- The next cell does most of the work using the function defined earlier. This defines the function `read_file()` mentioned in the sixth cell. The inline comments present there should help understanding what is going on. In case you still have questions, please get in touch with one of the tutors.
- The last cell shows a possible way of plotting the results.

The jupyter notebook examples just discussed can be found in the git repository at this link (it can also be reached by clicking on the top-right corner button `Visit Repo`). You should be able to download and execute them. A number of other examples (in the form of python modules and scripts) can be found at this other git repository. This portfolio of examples should give a wide range of answers to possible questions that you might have when trying to do the experiments proposed in Section 5.

By the way, it is worth to spend the time to digest the result of the analysis outlined above: the plot (repeated here in Figure 4.1) shows the invariant mass of the four leptons in events with four leptons. The red and purple histograms represent the prediction from the MC for the irreducible and reducible backgrounds, respectively. One notices a peak at 90 GeV, nicely predicted by the ZZ MC: 90 GeV is the mass of the Z boson. These events correspond to cases where the four leptons all come from one Z boson. A second feature of the ZZ prediction is a step at about 180 GeV. That is the mass of two Z bosons: most events where the Z bosons are emitted on shell are forced to yield an invariant mass above two times the Z boson mass. The irreducible background is *non-resonant*, that is, does not have a peak at any specific value of the invariant mass.

A small peak at 125 GeV in the data can be observed. Congratulations: you have just re-discovered the Higgs boson in its decay into four leptons!

4.2 ATLAS Open Data: Getting Started

Although the jupyter notebook example is a good starting point to understand the mechanics of what we are going to do, we will need to have more freedom in what we do. Anaconda is a great framework to do this. You can start by cloning the git repository at this link (there is a button

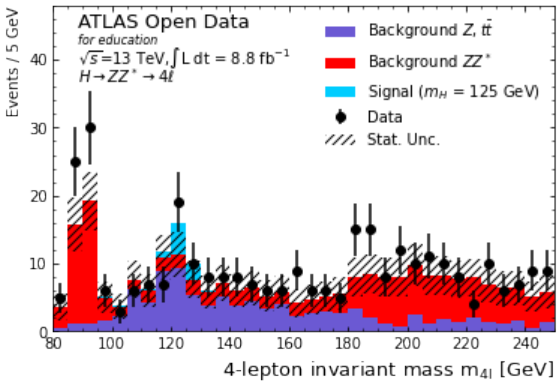


Figure 4.1: Result of the example analysis described in the text. A Higgs boson peak in the four-lepton invariant mass is visible at $m \sim 125$ GeV.

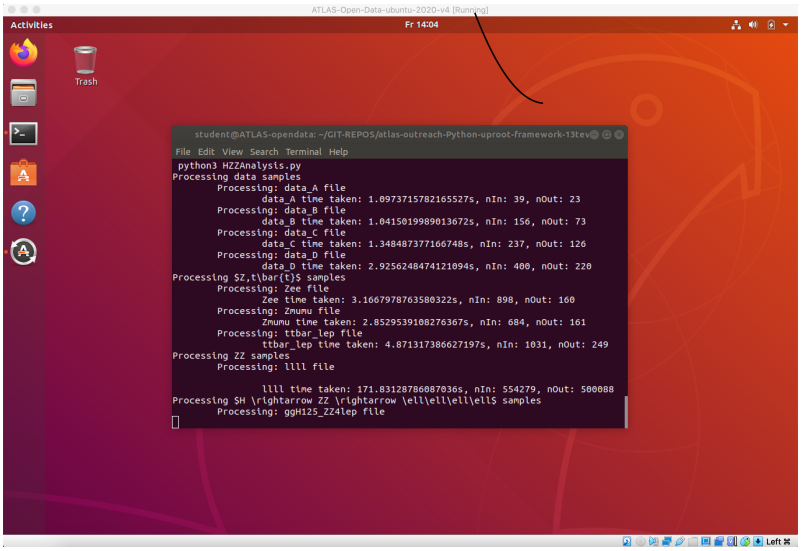


Figure 4.2: Running HZZAnalysis.py from the ATLAS Virtual Machine terminal.

labelled "Code" with a download symbol beside the label. Click and select “Download ZIP”)[1]. From the terminal window, or Anaconda prompt can now execute HZZAnalysis.py. For example

```
python3  HZZAnalysis.py
```

Now be patient. The running may take few minutes. If everything is smooth, you should see something like Figure 4.2. Of course the numbers corresponding to the time taken to run on a given sample may change depending on your machine and connection. For example: on my machine it took few seconds to run on data, but some minutes to run on `llll`.

At the end of the running, you should have a file called `HZZ_mlll.pdf`. You can see it by typing `evince  HZZ_mlll.pdf` in the virtual machine terminal. It will be very similar to what the notebook above has produced.

You can edit this file, or (better) make a copy and then edit. The full list of variables available in the open data files is available at this link.. For most of them, the variable description should

[1] You might actually take the chance to learn to use a modern software management tool, create an account on GitHub and see what you can do with it.

help to understand what they are. Some others will not be used in this experiment. Finally, the meaning of some variables will become clear later.

You are now ready to start with your experiments!

4.2.1 ATLAS Virtual Machine

In case you have problems getting this working on your laptop, you can try the ATLAS virtual machine, with instructions below. Full instructions on how to install the ATLAS virtual machine can be found on the ATLAS open data web site[2]. Full instructions on how to install the ATLAS virtual machine can be found on the ATLAS open data web site.

ATLAS Virtual Machine: Please follow the video about the `VirtualBox` installation (if you do not have it already), then follow the instruction specific to the installation of the 13 TeV ATLAS Open Data virtual machine installation.

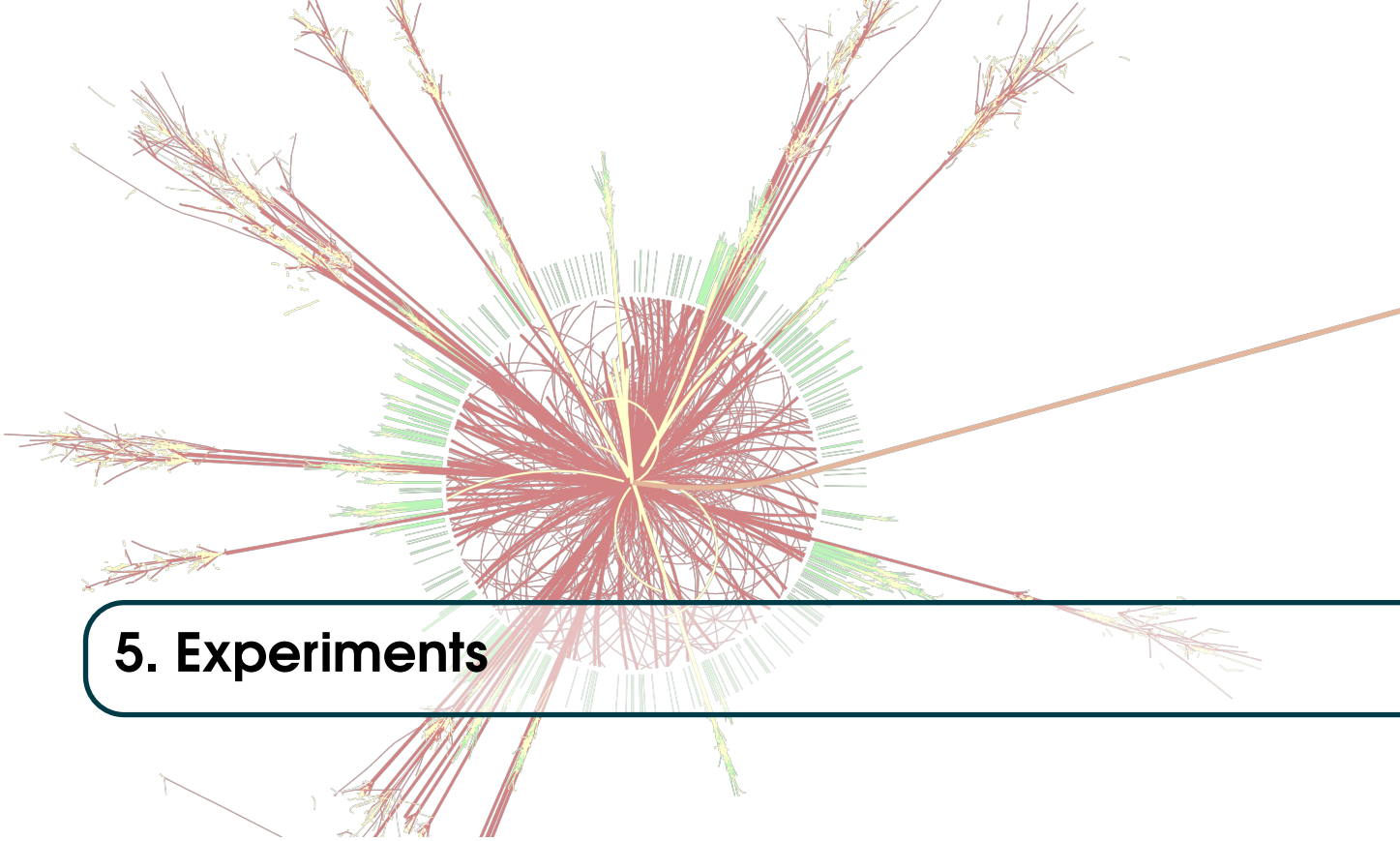
As the videos describe, after having started the ATLAS virtual machine, you will be able to type `localhost:8888` in your browser.

That will show you the directories and files which are there in the virtual machine. If you now go back to the ATLAS virtual machine window, and there you open a terminal, and type `ls`, you will see the same view. Please go to

`GIT-REPOS/GIT-REPOS/atlas-outreach-Python-uproot-framework-13tev`

If the directory is not there, you will need to clone the git repository from this link or download the ZIP as described above.

[2] If you end up with an error "VT -x Is disabled in teh BIOS for all CPU modes", you will need to enable virtualisation. You can find out how to do that by googleing "[your machine] enable virtualisation" (unfortunately instructions are laptop-specific).



5. Experiments

5.1 Experiment 1: Understanding $H \to 4\ell$

In this experiment we will dig deeper into the $H \to ZZ$ analysis that we have seen earlier. Please document everything you do, including the answers to the questions below, in your logbook. It is suggested you make a copy of the HZZAnalysis.py before you start to modify it.

1. Edit the file to print the histogram integrals of the different processes: $ZZ \to 4\ell$ (red histogram), Z and $t\bar{t}$ production (shown together in purple), $H \to ZZ$ (cyan), and, finally, the data. From the discussion about MC weights in Section 4.1, it should be clear that the integral of the histograms and the actual number of events are different for MC (because of all the applied weights), while they will coincide for the data.
2. Validate these printouts: try to have a rough estimate of those integrals by looking at the output histogram of the code. Make sure that the rough estimate and the exact printouts are in the same ballpark.
3. Compute the total expected background B by summing the integrals of the ZZ, Z and $t\bar{t}$ processes.

It is important at this point to clarify a couple of things.

- The histogram displays the number of events in a bin with width 5 GeV. The units displayed on the y-axis are "Events/5 GeV". Therefore, the area of a rectangle representing, for example, a yield of 20 Events/5 GeV is:

$$A = 20 \text{ Events}/5 \text{ GeV} \times 5 \text{ GeV} = 20 \text{ Events}. \tag{5.1}$$

  Therefore, the integral corresponds simply to the sum of the entries of all histogram bins.
- Depending on where you print the yields in the code, you may notice that the entries of the red histogram would correspond to the sum of the ZZ, Z and $t\bar{t}$ processes. This is essentially a graphical trick: to prepare the stack histogram, and display the ZZ contribution correctly, a histogram corresponding to the sum of the three processes is prepared and plot, and then a second histogram, corresponding to the sum of Z and $t\bar{t}$ is overlaid. The result is that the "uncovered" bit of the red histograms correctly displays the ZZ contribution only. The bottom line is that

the yields that you obtain from the code should be validated against the plot. Do they match?

4. Now compute the background and the observation only in the invariant mass range 120 GeV $< m_{4\ell} <$ 130 GeV. We will now compute its agreement with the observed data yield O:

Exercise 5.1

- The total background B represents the number of events that one *expects*. Suppose I expect 100 events: the outcome of the actual experiment in general will not be exactly 100. *Assuming that the background expectation is correct*, the probability of a given outcome will be described by a Poisson probability function. The probability of obtaining x counts in an experiment where B are expected is:

$$P(x;B) = \frac{B^x}{x!}e^{-B}. \tag{5.2}$$

- Please review the properties of a Poisson probability function, $P(x;\mu)$. What is the RMS of a Poisson distribution of mean value μ?
- Evaluate the level of compatibility of your data yield O by computing the probability to have observed a result equal or worse (that is, further away from the expectation) $p(x \geq O) = \sum_{x=O}^{\infty} \frac{B^x}{x!}e^{-B}$. Such probability is known as a p-value. What is the p-value? Small values of the p-value indicate low level of compatibility of the hypothesis tested (in this case that the observed yield comes from an expected background level B).
- Physicists like to convert p-values. This conversion can be done for example using this web site. A high z score, or significance, corresponds to a low level compatibility of the hypothesis tested (in this case that the observed yield comes from an expected background level B). What is the significance in this case?

■

5. Assume that the difference between the background and observation in the range 120 GeV $< m_{4\ell} <$ 130 GeV is now a yield due to $H \to 4\ell$. Knowing that the fraction of Higgs boson production that you select with this analysis is $\varepsilon = 2 \times 10^{-5}$ (coming from $BR(H \to ZZ) = 2.6 \times 10^{-2}$, $BR(Z \to \ell\ell) \sim 6\%$, where $\ell = e, \mu$, plus some inefficiency in reconstructing leptons), what is your estimate for the production cross-section of the Higgs boson? Do your best to associate an error to it coming from the number of observed events, and compare to the standard model best prediction of $\sigma(H) = 55$ pb at $\sqrt{s} = 13$ TeV, with an uncertainty of 5%.
6. Assuming your estimate for the signal and background, how much integrated luminosity would you need to declare a 5σ discovery?

5.2 Experiment 2: Plot the di-electron and di-muon invariant mass in data

In this experiment we will try to use data only to plot the di-lepton invariant mass in events with two leptons, and we will se what the meaning of the cuts on isolation on lines 27 and 31 of `HZZCuts.py` actually do. Finally, we will try to understand the mass resolution for the measurement of $Z \to e^+e^-$ and $Z \to \mu^+\mu^-$. We will start from a copy of `HZZAnalysis.py` and associated modules and heavily modify them. To start with, copy `HZZAnalysis.py`, `HZZCuts.py`, `HZZHistograms.py`, `HZZSamples.py`, into new files `ZeeAnalysis.py`, `ZeeCuts.py`, etc.

TIP: When modifying and playing with the code, reduce `fraction` to something like 0.1 or even lower, to speed up the code while debugging.

1. Edit the ntuple path to point to the 2lep events rather than the 4lep ones.
2. Change the Histograms to show only one variable, m_{ee} and plot it from 0 GeVto 200 GeVin log scale.
3. Change the cuts file to have only one cut, where you reject events if the leading two leptons are not electrons of opposite charge.
4. Change the samples to have only the data ones.
5. Adapt the plotting code to plot only the data.
6. Add a variable to the DataFrame (m_{ee}), similar to what was done with $m_{4\ell}$ in the `HZZAnalysis.py` example.
7. Plot the m_{ee}.

The histogram should display a peak at about 90 GeV, with a lot of events at different masses. The peak at 90 GeV is clearly due to $Z \to e^+e^-$, but.... what are all other events?

8. Try to add cuts that select on the lepton isolation variables, looking at the examples in `HZZCuts.py`. What you want is to reject events where either the first or the second leading lepton is not isolated.
9. Plot again the invariant mass. Comment on differences that you see, especially at small values of m_{ee}.

There are two main categories of leptons that the detector sees: genuine leptons and fake leptons. Genuine leptons are those where reconstruction correctly identifies a lepton as such. Fake leptons are those where the reconstruction thinks there is a lepton, but in fact there wasn't.

Genuine leptons, in turn, are divided into two categories: prompt leptons arise from an on shell W or Z going to leptons (like in $pp \to Z \to e^+e^-$, or $t\bar{t} \to WWbb \to \mu\nu_\mu jjbb$). These leptons are *isolated*, that is, normally do not have other particles nearby. Non-prompt leptons are emitted in hadrons decay, like $B_0 \to e^-\nu K^+$, and they tend to be non-isolated. Isolation is a powerful variable to reject fake and non-prompt leptons.

10. Fit the Z peak in data with a gaussian. Now repeat the analysis selecting muons instead of electrons, and fit the $Z \to \mu\mu$ peak with a gaussian. Compare the widths of the two gaussians. Which one is larger? Can you try to guess why?

5.2.1 Experiment 3: Understanding the events outside the mass window

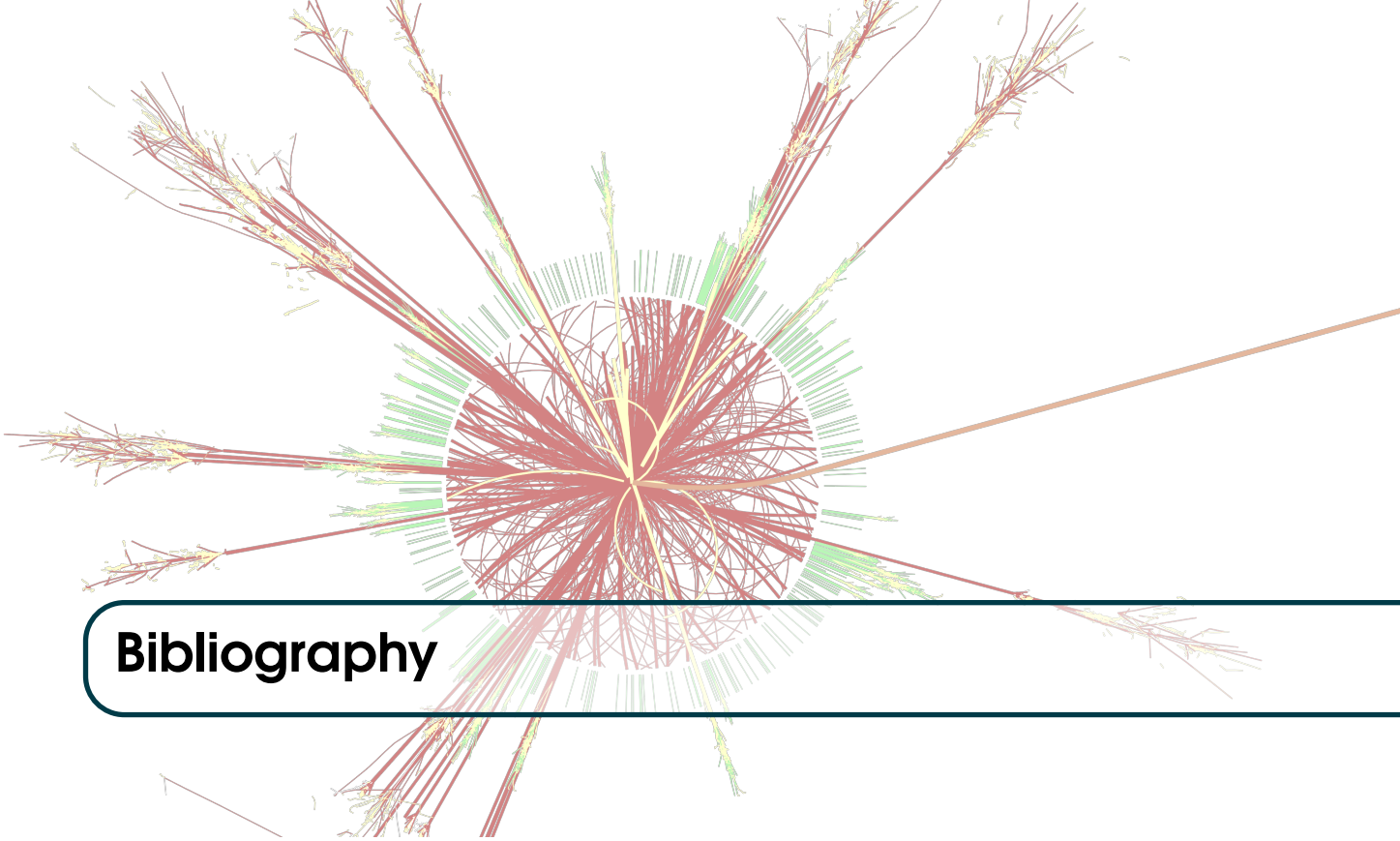
In this experiment we will start from the previous one and understand what the other non-Z events in the di-lepton sample actually are. Let's focus on the sample with two opposite charge electrons in the final state.

TIP: When modifying and playing with the code, reduce `fraction` to something like 0.1 or even lower, to speed up the code while debugging.

1. Change the Histograms file to add a missing et plot. The range will need to be something like 0 GeVto 300 GeVin log scale.
2. Change the samples list to include $t\bar{t}$, and diboson (WW, WZ and ZZ, with DSID in the range 363356 to 363493 - select those that contain two leptons) production.
3. Modify the plotting code to include the MC, similarly to what you had for `HZZAnalysis.py`
4. Plot the m_{ee} and E_T^{miss}.
5. Add a cut to reject events with m_{ee} on the Z peak. Plot again the E_T^{miss}.

6. Comment about the results. What are the events which are not Z in this data sample? Why do they tend to have a larger missing transverse momentum than Z events? Plot the expected composition of the data (what fraction is Z, what fraction is the rest) for $E_T^{miss} > X$ as a function of X.

Bonus task: DSIDs from 392501 to 392521 contain events from physics processes which are not foreseen by the Standard Model, but by some extension of it. Do you want to try and see whether you will be able to discover new physics using the ATLAS Data? Those new processes are $pp \to \tilde{\chi}_1^+ \tilde{\chi}_1^- \to 2\tilde{\chi}_1^0 2\nu\ell^+\ell'^-$. All you need to know is that the new particle $\tilde{\chi}_1^0$ behaves like a neutrino in the detector, and that ℓ and ℓ' may be of the same flavour or not. Can you design a selection where events would be mostly populated by these events?



Bibliography

[1] ATLAS Collaboration. “The ATLAS Experiment at the CERN Large Hadron Collider”. In: *JINST* 3 (2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003 (cited on pages 6, 17).

[2] *ATLAS top pair production candidate event display*. (accessed July 25, 2020). URL: https://twiki.cern.ch/twiki/pub/AtlasPublic/EventDisplayRun2Collisions/ATLAS_ttbar_candidate_13TeV_VP1_run266919_evt19982211_2015-06-04T00-21-24.png (cited on page 19).

[3] *Fragmentation and Hadronization*. Accessed September 20, 2020. URL: https://www.slac.stanford.edu/econf/C990809/docs/webber.pdf (cited on page 18).

[4] Brian R. Martin and Graham Shaw. *Particle Physics*. 4th edition. 10. Wiley, July 1993. ISBN: 978-1-118-91190-7 (cited on pages 7, 9).

[5] *Standard Model Summary Plot, May 2020*. (accessed July 25, 2020). URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2020-010/fig_01.png (cited on page 15).

[6] *The ATLAS detector*. Accessed July 25, 2020. URL: http://hedberg.web.cern.ch/hedberg/home/atlas/atlas.html (cited on page 18).

[7] *Wikipedia - Standard Model*. Accessed July 25, 2020. URL: https://en.wikipedia.org/wiki/Standard_Model (cited on page 8).