Chapter 2

Gravitational Waves: Sources, Detectors and Analysis

As we have seen in the previous chapter, GW are ripples in space-time, and due to the fall-off of their amplitude with distance, they exhibit minuscule amplitudes. After a century since the formulation of GW formalism, and nearly four decades of technological advancements dedicated to constructing GW detectors, scientists successfully measured this space-time deformation in 2015. Nowadays, GW detectors have discovered over 90 GW signals, probing down to the densest and most energetic regions of cosmic objects, which were hidden from astronomers' sight up until now [11, 12, 13]. Furthermore, novel astronomical detections are expected with the upgrade of second-generation detectors, as well as the construction of third-generation detectors, such as the Laser Interferometer Space Antenna (LISA), Einstein Telescope and Cosmic Explorer [14, 15, 16].

In this chapter, we present an overview of GW sources detectable by current and/or future ground-based detectors. We also introduce the current state-of-the-art of GW detectors and their noise sources, as well as introducing basic data analysis techniques.

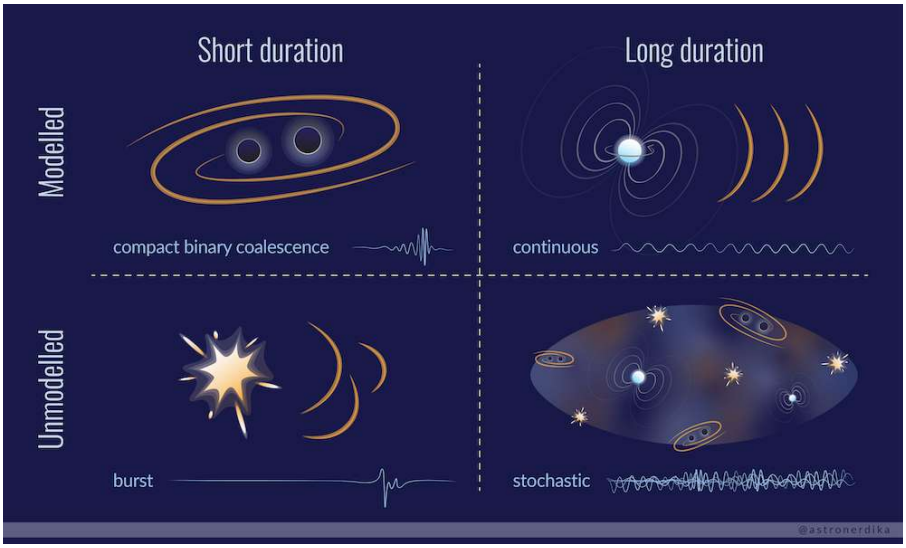

Figure 2-1: The different types of GW sources can be differentiated by their duration (short/long) and their associated astronomical models (modelled/unmodelled). Further details of the sources are described in the main text. Credits: Shanika Galaudage.

2.1 Gravitational-wave sources

In Fig. 2-1 we can observe an overview of GW sources classified according to their duration (short/long), and if astrophysical models are associated with the source (modelled/unmodelled). In the next subsections, we provide an overview of sources of interest for ground-based detectors.

2.1.1 Compact binary coalescence

The main type of short duration and modelled GW sources are known as Compact Binary Coalescence (CBC). These coalescing systems are composed of two compact bodies, either binary black holes (BBH), binary neutron stars (BNS) or one neutron star and one black hole (NSBH). In this context, "compactness" is the ratio between the mass and the radius of the object, which is related to the strength of the GW emission, and as a consequence, it provides insights into the dynamics and astrophysical properties of the binary system.

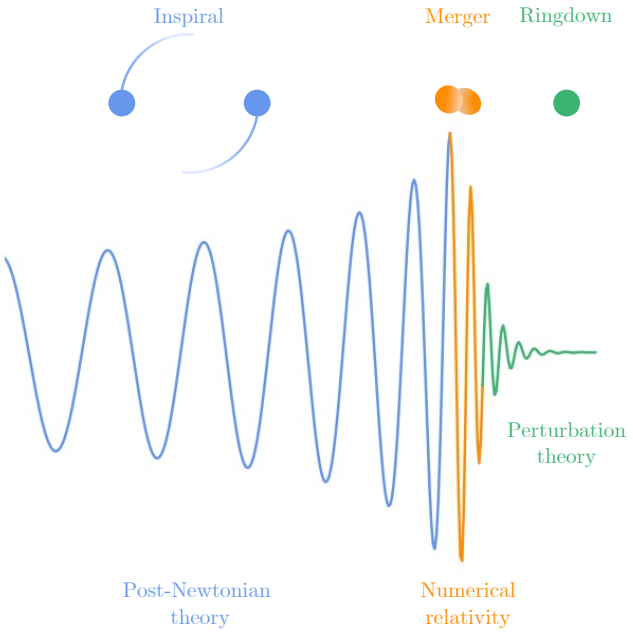

FIGURE 2-2: Temporal evolution of a binary system with component masses $m_1 = m_2 = 20M_\odot$. We colour in blue the inspiral, in orange the merger, and in green the ringdown.

As these two bodies orbit each other, they emit GW radiation while losing orbital energy and angular momentum, shrinking their orbit (*inspiral phase*). Eventually, an astronomical cataclysm occurs with their abrupt collision (*merger phase*). The remnant of this coalescence will quickly return to ground-state (*ringdown phase*). As we show in Fig. 2-2, the inspiral phase is modelled using post-Newtonian expansion (see Section 1.5). The merger phase is modelled with numerical relativity since in general relativity the two-body problem is not analytically solvable. The ringdown phase is modelled using perturbation theory, where the resulting compact object from the coalescence, known as *remnant* returns to ground-state "ringing" like a bell so that the resulting GW is a superposition of damped sinusoids, known as quasi-normal modes.

Since this GW sources are the most understood, we can use modelled algorithms like matched filtering techniques to detect them (see 3.2.1 for details). However, we can also use weakly modelled or model-free algorithms, as in the case of intermediate-mass black holes [17].

2.1.2 Transient bursts sources

Short and unmodelled GW sources are known as bursts. Burst can be short, up to a few seconds duration, or long, up to $\sim 10^3$ s duration. In this work, we will focus on short-duration GW transients, which include but are not limited to, core-collapse supernovae [18] and cosmic strings [19]. These sources are generally unmodelled, due to either unknown theoretical background and/or complex dynamics of the system. Since burst searches are meant to detect the unexpected, the unmodelled search algorithms employed use minimal (targeted search) or

no assumptions (generic search) about the source GW signal [20, 21, 22]. As we have seen in the previous Section, model-free algorithms can also be sensitive to CBC.

2.1.3 Continuous wave sources

Persistent (long duration) and modelled GW are known as continuous waves, which are quasi-monochromatic signals with roughly constant frequency and amplitude compared to the observation time. Sources of continuous GW are single rapidly rotating non-axisymmetric massive objects, such as neutron stars with either a deformation on the surface of the star or due to some fundamental oscillation mode [23]. Continuous waves could also be produced by exotic objects, such as the annihilation of ultra-light boson clouds around spinning black holes [24, 25].

2.1.4 Stochastic background

Persistent (long duration) and unmodelled GW is known as the Stochastic background of GW. This background could be the result of the superposition of incoherent GW signals. It could arise from cosmological sources, such as the inflationary epoch, first-order phase transitions in the early universe or cosmic strings; or from astrophysical sources, such as supernovae or the inspiral and merger of CBC over the history of the universe (see [26] for a comprehensive review). LIGO and Virgo have placed upper bounds on the energy density of the stochastic background in the range $[20, 10^3]$ Hz, by calculating the cross-correlation between pairs of detectors in the search of an excess in the distribution [27].

2.2 Detectors through history

In the 1960s, Joseph Weber began experiments to detect GW with his resonant mass detectors, which measure the oscillations of a bar caused by a passing GW. Weber's detector reached a sensitivity of 10^{-16} m, achieving an important milestone towards GW detection [28]. In 1969 he claimed to have observed signals from GW, but his results remained unreproducible [29].

The concept of interferometric detectors emerged in the early 1960s and 1970s, with the fundamental design resembling that of a Michelson-Morley interferometer, which was initially conceived in 1887 to prove the existence of *luminiferous ether*: a hypothetical medium for the propagation of light waves. Their experiment measured the relative motion between Earth and such medium, not only finding null results for the existence of *luminiferous ether* but also suggesting that the speed of light is constant and independent of the observer's motion, playing a pivotal role in the development of special relativity. Furthermore, since the Michelson-Morley interferometer was designed to measure the relative length changes of two perpen-

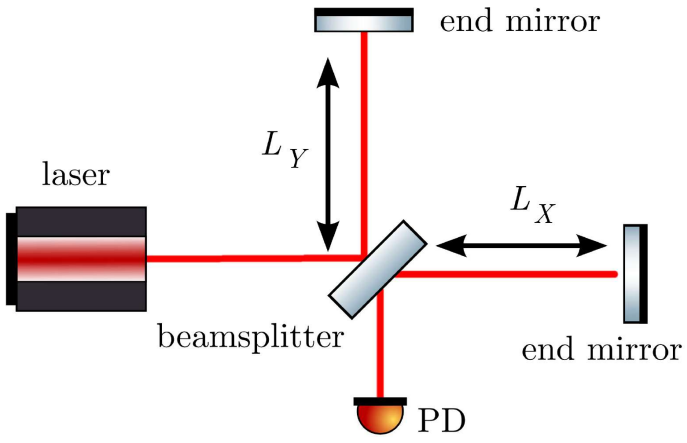

Figure 2-3: A schematic setup of a Michelson-Morley interferometer retrieved from [30].

dicular directions, it can serve as a GW detector by measuring the deformation of the test masses (see Section 1.4) [31, 30].

The individual elements composing modern GW interferometric detectors, such as lasers and mirrors, are individually well described by mostly classical physics and the complexity in the detector's behaviour arises from the combination of these elements into an optical cavity to enhance the interaction of light waves. As the description of a GW interferometer is a complex topic, the interested reader can refer to [30], but in this work, we provide a basic overview of the detector. In Fig. 2-3 we present a schematic illustration of a Michelson-Morley interferometer, composed of a laser, a beam splitter (a 50% reflecting mirror), two highly reflecting end mirrors and a photodetector (PD). The three mirrors can be considered free-falling objects in the horizontal direction, as they are suspended as pendulums, only allowed to swing freely in that direction. As we can observe in Fig. 2-3, the reflecting mirrors are placed equidistant from the beamsplitter, at a distance L_X and L_Y for the x- and y-axis respectively, but in orthogonal directions. The light beam input from the laser incides in the beam splitter, where it splits it into two beams. These two beams travel through the detector's arms until they reach the highly reflecting end mirrors, being redirected to the beamsplitter, where they are recombined and detected at the PD. When the beam recombines, it will interfere constructively or destructively if the lengths of the two arms differ by an even or odd number of wavelengths.

$$\Delta L \equiv L_X - L_Y = \begin{cases} n\lambda, & \text{(constructive interference)} \\ \left(n + \frac{1}{2}\right)\lambda, & \text{(destructive interference)} \end{cases} \tag{2.1}$$

where $n \in [0, 1, \ldots, N]$. As we saw in Section 1.3, in the proper detector frame the passage of a GW along the z-axis causes a displacement of the test masses in the $x - y$ plane from their original position. This introduces a relative temporary change in the light path on the x-axis with respect to the y-axis, that can be measured with interferometric GW detectors. As photons travel through null geodesics, which implies that the interval $ds^2 = 0$ (see Eq. 1.32), this relative change can be expressed as,

$$dx = \frac{dt}{\sqrt{1 + h_+ \cos\left[w(t-z)\right]}} \approx \left(1 - \frac{h_+ \cos\left[w(t-z)\right]}{2}\right) dt = L_x dt, \tag{2.2}$$

$$dy = \frac{dt}{\sqrt{1 - h_+ \cos\left[w(t-z)\right]}} \approx \left(1 + \frac{h_+ \cos\left[w(t-z)\right]}{2}\right) dt = L_y dt \tag{2.3}$$

where we have considered only a +-polatized GW and performed a Taylor expansion around $h = 0$. Therefore, we can compute the difference in path length between x and y arms,

$$\Delta L = L_x - L_y = hdt = Lh \implies h(t) = \frac{\Delta L}{L} \tag{2.4}$$

where $L = dt$ is the unperturbed path length. From Eq. 2.4 we can observe that a passing GW produces a fractional change in distance in the detector, generating an output called *strain*. On the other hand, Eq. 2.1 indicates such strain will cause a phase shift detectable by the PD.

Current GW interferometric detectors are modified Michelson-Morley interferometers, that form a global infrastructure for the discovery and study of GW. Nowadays, the network is formed by two Advanced LIGO [5], one located in Hanford, Washington (USA) and another one in Livingston, Louisiana (USA), Advanced Virgo [6], located in Cascina (Italy), GEO 600 [32], located in Hanover (Germany) and Kamioka Gravitational Wave Detector or KAGRA [33], located in the Gifu-prefecture (Japan). In ~~the next~~ two decades, we expect to improve the current advanced detectors at A$^\sharp$ sensitivity, as well as the addition of LIGO Aundha (India) [34, 35, 36]. Furthermore, we also expect the launch of LISA [14], as well as the construction

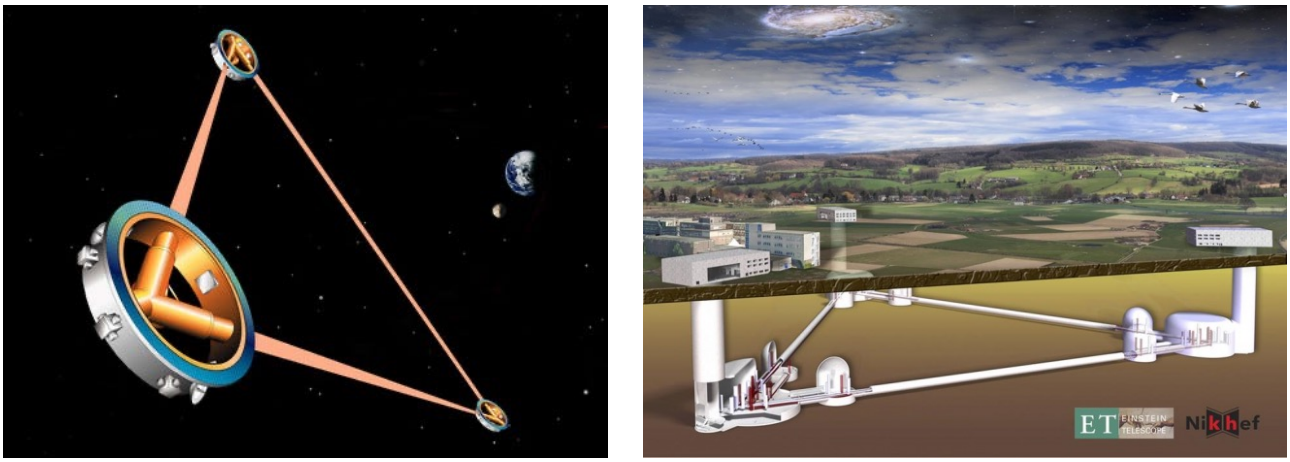

Figure 2-4: (Left) LISA orbiting around the Sun. (Right) Einstein Telescope underground. Artists impressions.

of third-generation detectors such as Einstein Telescope [15] and Cosmic Explorer [16] (see Fig. 2-4 for a visualization). In Fig. 2-5 we show the location of the second-generation of GW interferometers.

Figure 2-5: Location of the current network of second-generation interferometric detectors.

2.3 Antenna pattern of interferometers

Interferometric detectors are omnidirectional antennas and have a good sensitivity over a large fraction of the sky. From Eq. 2.4, the output of the detector will be $h(t)$, which will respond to a passing GW as $h_{ij}(t, \mathbf{x})$. The general transfer function for GW detectors is

$$h(t, \mathbf{x}) = D^{ij} h_{ij}(t, \mathbf{x}), \tag{2.5}$$

where D_{ij} is the detector tensor that depends on its geometry. For a detector which is sensitive only to GW with a reduced wavelength $\lambda\!\!\!^{-}$ much larger than its size, we can neglect the spatial dependence of the GW signal $h_{ij}(t, \mathbf{x})$, such that

$$h_{ij}(t) = \sum_{A=+,\times} e_{ij}^{A}(\hat{\mathbf{n}}) h_A(t). \tag{2.6}$$

The direction of propagation of the wave is $\hat{\mathbf{n}}$, and e_{ij}^{A} the polarization tensor defined as

$$e_{ij}^{A}(\hat{\mathbf{n}}) = \begin{cases} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_j - \hat{\mathbf{v}}_i \hat{\mathbf{v}}_j & \text{for } A = + \\ \hat{\mathbf{u}}_i \hat{\mathbf{v}}_j + \hat{\mathbf{v}}_i \hat{\mathbf{u}}_j & \text{for } A = \times \end{cases} \tag{2.7}$$

where $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are unit vectors orthogonal to $\hat{\mathbf{n}}$. Thus, Eq. 2.5 can be expressed as,

$$h(t) = \sum_{A=+,\times} D^{ij} e_{ij}^{A}(\hat{\mathbf{n}}) h_A(t) = \sum_{A=+,\times} F_A(\hat{\mathbf{n}}) h_A(t), \qquad \text{where } F_A(\hat{\mathbf{n}}) = D^{ij} e_{ij}^{A}(\hat{\mathbf{n}}). \tag{2.8}$$

We have conveniently defined $F_A(\hat{\mathbf{n}})$ as the detector pattern functions, which depend on the direction of propagation of the wave $\hat{\mathbf{n}} = (\theta, \phi)$. Hence, the output of the detector yields

$$\begin{aligned} h(t) = h_+(t) F_+(\theta, \phi, \psi) + h_\times(t) F_\times(\theta, \phi, \psi), \qquad & \text{where} \\ F_+(\theta, \phi, \psi) &= -\frac{1}{2}(1 + \cos^2\theta)\cos 2\phi \cos 2\psi - \cos\theta \sin 2\phi \sin 2\psi, \\ F_\times(\theta, \phi, \psi) &= \frac{1}{2}(1 + \cos^2\theta)\cos 2\phi \sin 2\psi - \cos\theta \sin 2\phi \cos 2\psi, \end{aligned} \tag{2.9}$$

where ψ is the so-called polarization angle.

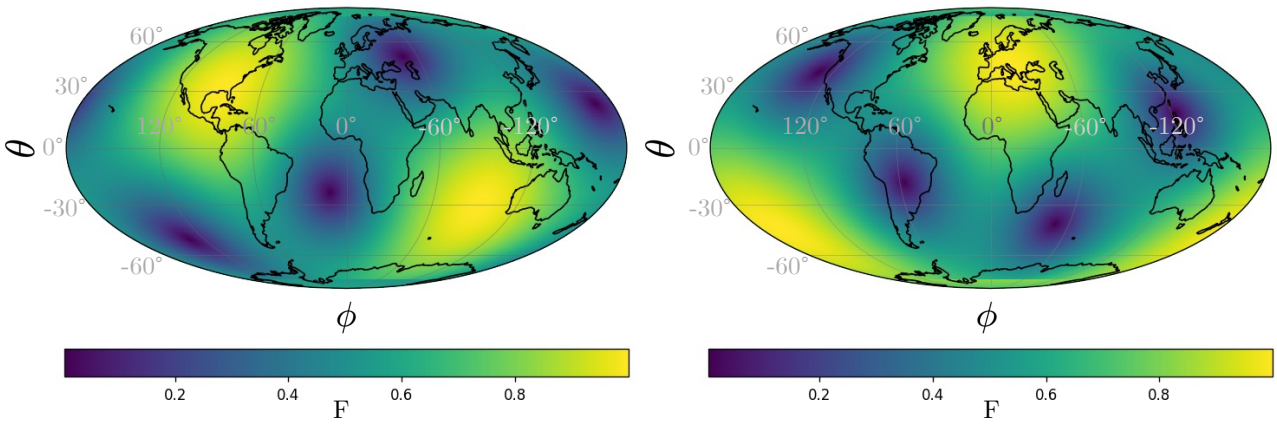

FIGURE 2-6: Variation of $F = \sqrt{F_+^2 + F_\times^2}$ for $\psi = 0$ as a function of the longitude ϕ and the latitude θ for LIGO Livingston (left) and Virgo (right).

In Fig. 2-6 we represent $F = \sqrt{F_+^2 + F_\times^2}$ as a function of the longitude ϕ and the latitude θ for LIGO Livingston in Florida (USA) and Virgo in Pisa (Italy). Large values of F imply a better sensitivity which is dependent on the orientation of the detectors. With respect to the plane defined by the arms L_X and L_Y, the most sensitive directions are orthonormal, and less sensitive directions are bisectors, such that $\cos\theta = \cos 2\phi = 0$.

2.4 Sources of noise

In the previous Section, we have seen the expression of the detector response, Eq. 2.5, in an ideal setting where there are no external perturbances. Nonetheless, the real world is full of imperfections causing undesired noise in the detector strain. Mathematically, the main strain of the detector measures

$$s(t) = n(t) + h(t), \tag{2.10}$$

where $n(t)$ represents the combination of all noise sources. Many textbooks treat the $n(t)$ as Gaussian and stationary, which is a poor approximation to the interferometers' data. The understanding and unbiased modelling of the different sources of noise is fundamental to infer the significance of GW signals and their astrophysical properties. Tasks to understand and mitigate noise sources both from the instrument and the data analysis side, also known as *detector characterization* tasks, are a significant portion of LIGO-Virgo-KAGRA collaboration's work [37, 38, 39, 40]. Following detector characterization guidelines, we can classify noise sources by dividing them into three different categories:

- *Fundamental noises:* They cannot be reduced without a major instrument upgrade, such as the installation of a new laser. An example of fundamental noise is thermal noise , associated with sources of energy dissipation, and quantum noise, related to the quantum nature of photons due to the Heisenberg uncertainty principle and quantum fluctuations.
- *Technical noises:* they arise from electronics or dust in the mirrors, and they can be reduced once identified and carefully studied.
- *Environmental noises:* include seismic-motion, acoustic and magnetic noises [41].

In the following subsections, we will provide details on the sensitivity of the detectors, while a summary of the most dominant sources of noise is provided in Appendix A.1.

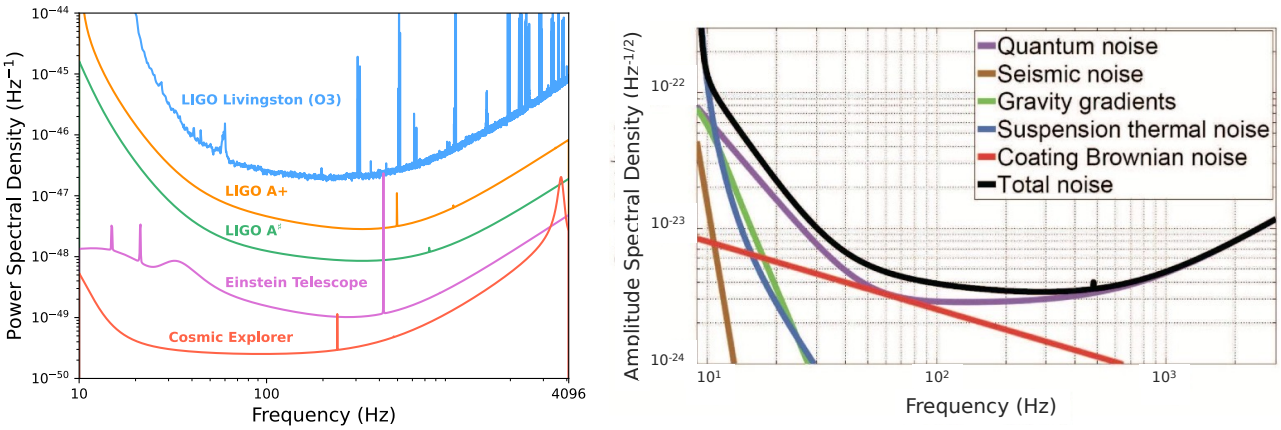

Figure 2-7: (Left) PSDs of LIGO Livingston during the third observing run (O3) (blue), LIGO A+ (orange), LIGO A♯ (green), Einstein Telescope (pink) and Cosmic Explorer (red). (Right) Design sensitivity of Advanced LIGO (black), and noise sources from seismic noise (brown), suspension thermal noise (blue), coating thermal noise (red), gravity gradient (green) and quantum noise (purple), retrieved from [42].

2.4.1 Power spectral density

If the noise is non-stationary, then the different components of the noise are uncorrelated, and therefore the ensemble average of the Fourier components of the noise is of the form

$$\left\langle \tilde{n}^*(f)\tilde{n}(f')\right\rangle = \delta(f-f')\frac{1}{2}S_n(f), \tag{2.11}$$

where $\tilde{\cdot}$ denotes the Fourier transform, * the complex conjugate and, as in the previous chapter, $\left\langle \dots \right\rangle$ represents the average. As we assume $n(t)$ to be dimensionless, $S_n(f)$ has dimensions Hz^{-1}. We can also assume without loss of generality that $\left\langle n(t)\right\rangle = 0$. In Eq. 2.11 the right hand side diverges for $f = f'$, but in any experiment we measure $\tilde{n}(f)$ in a finite time interval T, so for $f = f'$ Eq. 2.11 takes the form:

$$\left\langle |\tilde{n}|^2\right\rangle = \frac{1}{2}S_n(f)T \;\rightarrow\; \frac{1}{2}S_n(f) = \left\langle |\tilde{n}|^2\right\rangle \Delta f \qquad \text{for } f = f', \text{ and } \Delta f = \frac{1}{T} \tag{2.12}$$

where Δf is the resolution of the measurement in frequency. Note that the factor $1/2$ is conventionally inserted, such that $S_n(f)$ is obtained integrating over physical frequencies $f > 0$:

$$\left\langle |\tilde{n}|^2\right\rangle = \int_{-\infty}^{\infty} df\, df' \left\langle n^*(f)n(f')\right\rangle = \frac{1}{2}\int_{-\infty}^{\infty} df\, S_n(f) = \int_0^{\infty} df\, S_n(f) \tag{2.13}$$

The function $S_n(f)$ is known as the power spectral density (PSD). Moreover, the noise of the detector can also be characterized by $\sqrt{S_n(f)}$, known as amplitude spectral density ASD with dimensions $\mathrm{Hz}^{-1/2}$. In Fig. 2-7 left panel we present the PSD of various detectors PSD. While we show the average PSD of LIGO Livingston during the third observing run (O3), we also show the design sensitivity of LIGO A+ and LIGO A$^\sharp$, which will be beyond the fifth observing run. Furthermore, we present the PSD of Einstein Telescope (Europe-based) and Cosmic Explorer (USA-based), which will collaborate to detect GW signals during the third-generation era. In Fig. 2-7 right panel we present the design ASD of Advanced LIGO, as well as its noise budget with different sources of noise, namely, seismic noise (brown); suspension thermal noise (blue), which is dominant at $f \lesssim 10\,\mathrm{Hz}$; coating thermal noise (red); gravity gradient (green); and quantum noise (purple), dominant at $f \gtrsim 10\,\mathrm{Hz}$ [42]. The details of these sources are elaborated in Appendix A.1, except for coating thermal noise, which interested readers can refer to [43].

Characterizing the noise of the detector with the PSD is fundamental for GW data analysis, as it provides key information about the frequency content of the data at hand. Nonetheless, the PSD is not known a priori and it needs to be properly estimated. There are two commonly used methods to compute these estimates: "off-source" and "on-source" estimate [44]. The "off-source" estimation assumes that the PSD does not vary over the duration being averaged and that there are no non-Gaussian features in the data. The "on-source" estimation, in contrast, uses the commonly adopted method of simultaneously fitting the signal and power spectral density of the detector noise. While the second method is less affected by noise artifacts, it is computationally more intensive.

The "off-source" estimation is usually preferred for GW searches. The simplest estimate is Welch's method (see Appendix A.3 for its mathematical formalism), which assumes Gaussian and stationary data. To overcome its limitations, other authors have proposed more sophisticated methods for the PSD estimation [45, 46, 47].

2.4.2 Noise lines

GW interferometers are complex experiments with many sub-systems that couple to the main detector strain $h(t)$, causing large narrow-band contributions to the PSD. As an example, we can see in Fig. 2-7 in blue the PSD of LIGO Livingston during O3 presenting several narrow-band peaks, known as *noise lines*. Most lines in the detector data are stationary, but some of them have time-varying behaviour, degrading the detector sensitivity over a larger frequency.

Such behaviour hampers continuous GW searches as these artifacts can lead to spurious outliers which require laborious follow-up.
Some noise lines occur in a "comb" pattern, and the frequency peak of their nth "teeth" follows

$$\omega_N = f_o + n\delta f, \tag{2.14}$$

where f_o is the offset and δf the spacing between the teeth. Combs are associated with linear or non-linear couplings of non-sinusoidal sources, or with non-linear couplings of sinusoidal sources. Lines and combs can have time-dependent behaviours as the configuration of the detector changes: as interferometers undergo enhancements and upgrades, and issues related to coupling with various sub-systems are resolved, but the addition of new hardware, while inevitable, introduces further challenges related to data quality [39].

To further understand the noise of the detector, its status is continuously monitored through a large set of data streams at various sampling rates, outputting $\sim 10^6$ time series from instrumental and environmental sensors. These auxiliary channels can be divided into safe (insensitive to GW) and unsafe (sensitive to GW). Some subset of these channels may serve as "witnesses" to narrow-band couplings in the detector, but they can also "witness" the production of non-gaussian transient burst noise, as we will see in the next subsection.

An example of identified narrow-band couplings is the power lines caused at 50 Hz in Virgo and 60 Hz in LIGO, as well as their respective harmonics, caused by the main power supplies. Other examples are also mechanical resonances of mirror suspension, known as "violin modes", and simulated GW signals known as "hardware injections" [48, 49].
The standard process for mitigating lines or combs is often iterative and experimental, but their main steps are:

1. Identification of noise in the main detector strain.
2. Determination of the properties of the noise, such as duration, associated frequencies and possible channel "witnesses".
3. On-site investigations or interventions.

Work on site is constrained by time availability and the risk of creating novel sources of noise. Hence, the mitigation of noise sources is usually prioritized by their strength, the number of frequency bins contaminated, and the ease of addressing their cause. Lines which are not well-understood are catalogued afterwards, helping GW searches on cleaning the data and rejecting outliers. Mitigation efforts are challenging as they can take order of days or weeks to determine if these methods have contributed significantly to data quality. Furthermore, configuration changes in the detector that lead to line generation can also take time to appear and be mitigated. We recommend that interested readers refer to [50] and references therein for an in-depth description of the methodologies developed to address these issues.

2.4.3 Glitches

Noise couplings can also cause a transient non-astrophysical burst of non-Gaussian noise, which are colloquially known as *glitches* [51, 52]. Glitches may be caused by the environment (e.g., earthquakes, wind, anthropogenic noise) or couplings with instruments (e.g., control systems, electronic components [53]), though in many cases their causes remain unknown [54]. They come in a large variety of time-frequency morphologies, have a typical duration of between sub-seconds and seconds, and have a high rate of occurrence ($\sim$ 1 per minute during the first half of the third observing run, O3a [12]).

Glitches are problematic due to their large abundance and capability of hampering GW data analysis. They can reduce the amount of analyzable data increase the noise floor, produce false positives in GW data, affect the estimation of the detector power spectral density and reduce candidate significance in searches for short- and long-lived GW signals [55, 56, 57, 27, 58]. Glitches can also bias astrophysical parameter estimation, making it difficult to determine which part of the signal corresponds to a glitch and which part to the actual GW event [59, 60, 61]. Additionally, glitches can impact line-cleaning procedures in GW searches, which rely on replacing disturbed frequency bins with artificially generated data, consistent with their neighbors [62, 50, 58]. If the surrounding data contains elevated noise floors, the efficacy of mitigation methods will be reduced.

Glitch identification and characterization is a crucial first step towards their mitigation, but due to their overwhelming amount, their characterization by hand is unfeasible. [63, 39, 64]. A promising option is then to construct machine learning (ML) algorithms for their identification. Most of the current approaches to glitch characterization with ML utilize supervised classification algorithms, where models learn to identify glitches through labelled data representations of GW strain data $h(t)$ [65, 66, 67, 68, 69, 70, 71]. In practice, glitches are visualized in time-frequency representations, which involves a modification of the standard short-time Fourier transform parameterized by a quality factor Q [72, 73, 74]. For a discussion on time-frequency representations, the interested reader can refer to Section 3.3.2.

Nonetheless, this procedure of glicht identification presents several limitations. Firstly, generating labelled data is an expensive task, since ML methods need a lot of examples for training, and experts must vet the labelling procedure. Secondly, glitch classes are highly unbalanced, biasing the models towards the most common classes. Moreover, supervised learning needs fixed class definitions that are not exhaustive nor representative of all glitch morphologies, as there could be many possible sub-classes to discover [67]. Furthermore, as GW detectors are improved, novel glitch morphologies could arise [75].

Despite these challenges, these methods have been instrumental in GW detector characterization and data analysis. In the following, we describe the different classes defined by one of the most well-known supervised ML algorithms, `GravitySpy` [65, 68], whose morphologies can be visualized in Fig. 2-8.

- *1080Lines:* these glitches manifest as brief, recurrent spikes occurring $\sim 0.1\,\mathrm{s}$ at $\sim 1080\,\mathrm{Hz}$, and additionally accompanied by noise $< 64\,\mathrm{Hz}$. These disturbances were notably widespread in LIGO Hanford during the early stages of the second observing run (O2) but saw a reduction after that by improvements in the output mode cleaner [68].

- *1400Ripples:* they exhibit a short timespan ($\leq 0.5\,\mathrm{s}$) with a wavy morphology at $\sim 1400\,\mathrm{Hz}$.

- *Air_Compressor:* they appears as a broad, horizontal line at $\sim 50\,\mathrm{Hz}$. Investigations in LIGO Hanford concluded that these glitches were linked to air compressor motors at the end stations. They were mitigated by replacing the vibration isolators [68].

- *Blip:* these glitches have a characteristic morphology of a symmetric "teardrop" shape in time-frequency in the range $[30, 250]\,\mathrm{Hz}$ with short-durations, $\sim 0.04\,\mathrm{s}$. They appear in both LIGO Livingston and LIGO Hanford, as well as Virgo and GEO 600 [54]. Due to their abundance and form, these glitches hinder both the unmodeled burst and modelled CBC can searches, with particular emphasis on compact binaries with large total mass, highly asymmetric component masses, and spins anti-aligned with the orbital angular momentum [55, 52]. Moreover, since there is no clear correlation to the auxiliary channels, they cannot be removed from astrophysical searches yet.

- *Blip_Low_Frequency:* they have a similar shape to *Blips*, but occur at lower frequencies, with peak $\sim 10-50\,\text{Hz}$. As they fall in the expected frequency band for high-mass CBC, they hinder their detection. Note that this class was added during O3 [75].

- *Chirp:* they are GW signals from CBC artificially added in the detector data via created by hardware injections, i.e. by physically displacing the detectors' test masses. Note that these signals do not accurately reflect our present understanding of CBC populations [48].

- *Extremely_Loud:* their main characteristic is an exceptionally high signal-to-noise-ratio[1] (SNR), saturating their time-frequency representation. Typically, loud glitches result from significant disruptions in the detector, adversely affecting its sensitivity [39].

- *Fast_Scattering:* they appear as short-duration arches ($\sim 0.2-0.3\,\text{s}$) in the frequency range $[20-60]\,\text{Hz}$. These glitches are strongly correlated with ground motion in range $[0.1-0.3]\,\text{Hz}$ and $[1-6]\,\text{Hz}$, which in turn is associated with thunderstorms and human activity near the detector. Note that this class was added during O3, and it is more abundant in LIGO Livingston than LIGO Hanford, due to differences in ground motion and detector sensitivity [75].

- *Helix:* these glitches are usually grouped in sets of $2-3$ separated by $\sim 0.1\,\text{s}$ in the frequency band $16-512\,\text{Hz}$. Investigations point out that they might be related to the auxiliary lasers used to calibrate the detectors [68].

- *Koi_Fish:* Their naming comes from their imaginative frontal resemblance to koi fishes. They are also similar to *Blips* but typically feature high-SNR, spanning the frequency range of $\sim 20-1000\,\text{Hz}$.

- *Light_Modulation:* These glitches are in the frequency range $16-128\,\text{Hz}$, often displaying broad-band spikes. They exhibit high-SNR and stem from fluctuations in the amplitude of the control signal for the optical sidebands, which are responsible for adjusting the length and alignment of optical cavities [68].

- *Low_Frequency_Burst:* they are short-duration glitches ($\sim 0.25\,\text{s}$) in the frequency range $[10-20]$ Hz with a distinctive blob shape. These occurrences were prevalent in LIGO Livingston data during O1 and LIGO Hanford data in O3a.

- *Low_Frequency_Lines:* these glitches have a flat-line-like morphology with durations $\sim 1.5-2\,\text{s}$ and usually at frequencies $< 20\,\text{Hz}$.

- *Power_Line:* They are narrow lines, typically lasting $\sim 0.2-0.5\,\text{s}$ near the frequency of the power grid in the United States (or harmonics of this frequency). These glitches can be attributed to various equipment dependent on this power supply.

- *Repeating_Blips:* these glitches consist on multiple *Blip* glitches, often repeating every $\sim 0.25-0.50$ s.

- *Scattered_Light:* also known as Slow Scattering, these glitches have longer duration harmonics ($\sim 2.0-4.0\,\text{s}$), and in the time-frequency domain, they appear as arches often stacked on top of each other. These glitches are quite problematic since their frequency content lies in the band of interest of GW astrophysical events. In O3, they were found to be coupled with the relative motion between the optical suspension system's end test-mass chain and the reaction-mass chain [75].

[1] It must be noted that this definition of SNR is different to the one presented in Section 3.2.1, as it is defined in the context of Omicron [73].

- *Scratchy:* These glitches manifest as a sequence of distinct sharp peaks, primarily at intermediate frequencies in range $\sim 60 - 250\,\mathrm{Hz}$. These peaks can occur at a rate of $\sim 10 - 30\,\mathrm{s}^{-1}$, and are associated with light scattering from the Swiss cheese baffles [68].

- *Tomte:* these glitches are also short-duration ($\sim 0.25\,\mathrm{s}$) with characteristic triangular morphology. Since there is no clear correlation to the auxiliary channels, they cannot be removed from astrophysical searches.

- *Wandering_Line:* They are long-duration glitches with a distinctive undulating morphology. They can span a broad spectrum of frequencies, often displaying multiple lines simultaneously at various frequencies, but typically they spam in frequencies $> 256\,\mathrm{Hz}$.

- *Whistle:* These glitches have a characteristic V, U or W shape at higher frequencies ($\gtrsim$ 128 Hz) with typical durations $\sim 0.25\,\mathrm{s}$. They are caused when radio-frequency signals beat with the voltage-controlled oscillators [37].

As we mentioned before, persistent glitches at specific frequencies critical to data analysis constitute a problem for GW detection and parameter estimation. Hence, it is essential to eliminate the harmful influence of these glitches on the searches for GW signals. Discerning the background of glitches is indeed a challenging problem. One major difficulty is the lack of glitch simulations, making it challenging to transition from perfect simulated noise to real data. Additionally, the unknown real population of glitches complicates the evaluation and understanding of the performance of ML classifiers. Moreover, while glitches are observed in the strain data $h(t)$ they are produced in the different subsystems of the interferometers, so understanding their formation is challenging by solely utilizing $h(t)$ data.

To tackle some of these challenges, in Chapter 6, we simulate *Blip* glitches—one of the main classes of glitches that hinder transient GW searches—using ML techniques. Additionally, we propose several applications and provide a practical example to evaluate the performance of `Gravity Spy`. In Chapter 6, we found that our ML algorithm could learn anomalous morphologies due to the lack of ground truth in the real *Blip* population. To address this fundamental issue, in Chapter 7, we construct an unsupervised ML method to explore the auxiliary channel data—time series originating from the monitors of different subsystems—and learn the underlying distribution of the data, uncovering unknown glitch morphologies and overlaps.

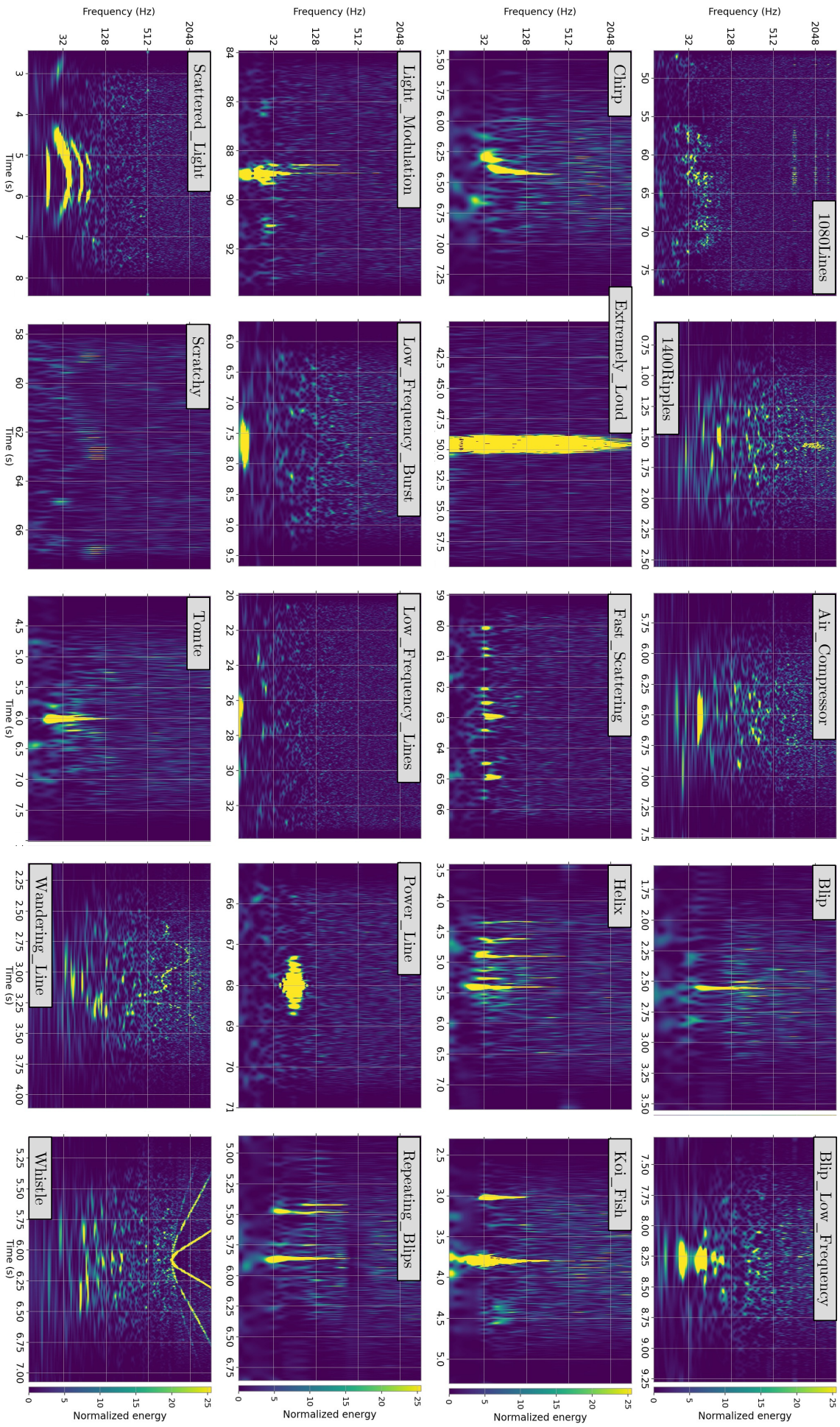

Figure 2-8: Time-frequency representations of different glitch classes in LIGO during O3.

2.5 Data conditioning

From the previous Sections, we have learned that the main strain of the detector $h(t)$ is impacted by fundamental noises, that can only be reduced with a major instrument upgrade, technical noises, that arise from the different subsystems within the detector, and environmental noises. As GW have minuscule amplitude and are buried in the detector noise, it is crucial to develop robust data analysis techniques that enhance these signals. In this section, we provide details about the most basic data conditioning methods that have led to GW discovery.

2.5.1 Sampling of continuous-time signals

The signal $h(t)$ is continuous, but the data measured in the detector is a representation of this sampled signal. Typically, we can obtain the discrete-time representation $x_s(t) = x[n]$ of a continuous-time signal x_c through periodic sampling as

$$x[n] = x_c(nT), \quad -\infty < n < \infty \tag{2.15}$$

where T is the sampling period, or time resolution, and $f_s = 1/T$ is the sampling frequency. It is important to note that the sampling operation is not invertible in general since many continuous-time signals can produce the same output sequence of samples. This inherent ambiguity is a fundamental issue in signal processing, but it is possible to restrict it by controlling and narrowing down the range of input signals fed into the sampling system [76]. Now, we derive the frequency domain relation between $x_c(t)$ and $x_s(t)$, so we consider that their relation is modulated by a periodic impulse train $s(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT)$ such that

$$x_s(t) = x_c(t)s(t) = x_c(t) \sum_{n=-\infty}^{\infty} \delta(t - nT) = \sum_{n=-\infty}^{\infty} x_c(nT)\delta(t - nT) \tag{2.16}$$

Then, its Fourier transform, denoted as $\tilde{\cdot}$, will be

$$\tilde{x}_s(f) = \frac{1}{T} \sum_{k=-\infty}^{\infty} \tilde{x}_c(\omega - k\omega_s), \quad \text{where } \omega_s = \frac{2\pi}{T} \tag{2.17}$$

Here, ω is the angular frequency, ω_s is the sampling angular frequency, which are continuous variables. Also, k is an integer number. From Eq. 2.17 we can observe that $x_s(t)$ consists of periodically repeated copies of $\tilde{x}_c$, which are shifted by integer multiples of ω_s and then superimposed to produce the periodic Fourier transform. According to the Nyquist sampling theorem, if $x_c(t)$ is band-limited, it is uniquely determined by

$$x[n] = x_c(nT) \quad \text{with } n \in \mathbb{Z} \quad \text{if } \omega_s = \frac{2\pi}{T} \geq \omega_N \tag{2.18}$$

where ω_N is known as Nyquist frequency. In this way the replicas do not overlap $\omega_s > 2\omega_N$ [77]. Otherwise, the frequency would components overlap, resulting in *aliasing*.

We express the discrete frequency $\Omega = \omega/T$ in radians/sample and, therefore, dimensionless. Hence, we can define the discrete Fourier transform of the sequence $x[n]$, often expressed in signal processing as $\tilde{x}(e^{i\Omega})$, is

$$\tilde{x}(e^{i\Omega}) = \frac{1}{T} \sum_{k=-\infty}^{\infty} \tilde{x}_c(\omega - k\omega_s), \tag{2.19}$$

where we have used Eq. 2.17 in the last term. Equivalently,

$$\tilde{x}(e^{i\Omega}) = \frac{1}{T}\sum_{k=-\infty}^{\infty} \tilde{x}_c(\frac{\Omega}{T} - \frac{2\pi}{T}k). \tag{2.20}$$

2.5.2 Low-pass, high-pass and band-pass filters

A way to avoid aliasing is to restrict the frequencies of the signal such that $f_s > 2\omega_N$. For this, we can use a low-pass filter which allows frequencies below a given cut-off frequency. Another possibility is to employ a band-pass filter, which allows frequencies in a pre-defined range. In practice, a band-pass filter can be constructed with a low-pass filter and a filter which allows frequencies below a cut-off frequency, known as a high-pass filter. For illustration, in Fig. 2-9 we provide an example of these filters.

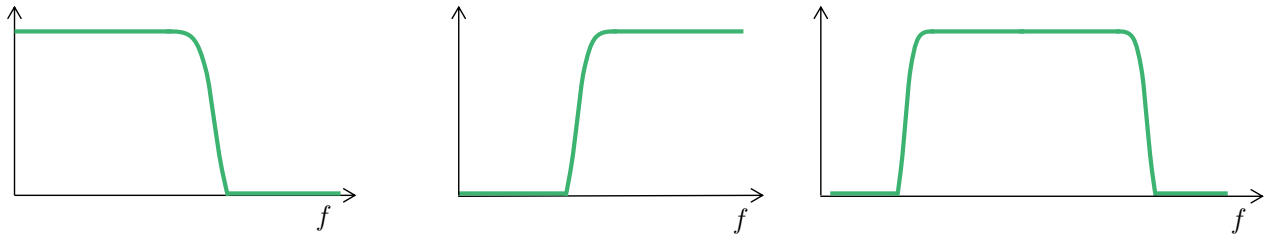

FIGURE 2-9: Schematic low-pass (left), high-pass (middle) and band-pass filters (right).

These filters are relevant in GW data analysis as the output of the detector spans from 10 Hz since the detector is not calibrated at lower frequencies, to $\sim$ 10 kHz. Limiting the frequency range of our analysis does not only avoid aliasing (see discussion in the previous section), but might simplify the data, which can lead to more accurate and reliable results. Depending on the nature of the target source, we would be inclined to perform a narrow band search, as in the cause of continuous waves with quasi-monochromatic signals (see Section 2.1.3), or even a wide range search, as in the case of transient Burst searches (see Section 2.1.2).

2.5.3 Resampling

The main strain of the detector $h(t)$ is a time series sampled at 16384 Hz for LIGO and 20 kHz for Virgo, but these data are only calibrated $>$ 10 Hz for the A+ detectors. Depending on the frequency content of the targeted signal, we might want to reduce the sampling rate of the data by defining a new time series x_d such that

$$x_d[n] = x[nM] = x_c(nMT), \tag{2.21}$$

where M is a reduction factor and T is the sampling period. $x_d[n]$ is identical to the time series that would be obtained from the continuous signal $x_c(t)$ sampled with a period $T' = MT$. The sampling rate can be reduced by a factor M without aliasing if the new time series is band-limited to $\omega < \omega_N$, and then $x_d[n]$ would be an exact representation of $x_c(t)$ if $\pi/T' = \pi/MT \geq \omega_N$. This operation is called *downsampling*.

Similarly to Eq. 2.20, the discrete-time Fourier transform of $x_d[n] = x[nM] = x_c(nT')$ is

$$\tilde{x}_d(e^{i\omega}) = \frac{1}{T'}\sum_{r=-\infty}^{\infty} \tilde{x}_c\left(\frac{\omega}{T'} - \frac{2\pi r}{T'}\right) = \frac{1}{MT}\sum_{r=-\infty}^{\infty} \tilde{x}_c\left(\frac{\omega}{MT} - \frac{2\pi r}{MT}\right). \tag{2.22}$$

Relating Eq. 2.20 and Eq. 2.22, we can express the summation index r from Eq. 2.22 as $r = m + kM$ for $k \in (-\infty, \infty)$ and $m \in [0, M-1]$. Thus, we can rewrite Eq. 2.22 as

$$\tilde{x}_d(e^{i\omega}) = \frac{1}{M}\sum_{m=0}^{M-1}\left[\frac{1}{T}\sum_{k=-\infty}^{\infty}\tilde{x}_c\left(\frac{\omega}{MT} - \frac{2\pi k}{MT} - \frac{2\pi m}{MT}\right)\right] = \frac{1}{M}\sum_{m=0}^{M-1}\tilde{x}e^{i(\omega/M - 2\pi m/M)}. \tag{2.23}$$

In this way we can interpret $\tilde{x}_d(e^{i\omega})$ as composed of M copies of the periodic Fourier transform $\tilde{x}(e^{i\omega})$, frequency scaled by M and shifted by integer multiples of 2π. From this we can understand that $\tilde{x}_d(e^{i\omega})$ is periodic with period 2π and that aliasing can be avoided by ensuring that $\tilde{x}(e^{i\omega})$ is band-limited. Therefore, it is important to apply a low-pass or anti-aliasing filter before downsampling. Such procedure is known as *decimation*.

For completeness, but not utilized in this work, we could be interested in increasing the sampling rate by a factor of L, such that the new sequence will be defined as

$$x_e[n] = x[n/L] = x_c(nT/L), \qquad \text{for } n = 0, \pm L, \pm 2L, \ldots \tag{2.24}$$

This operation is known as *upsampling*, and it can be mathematically expressed as

$$x_e[n] = \sum_{k=-\infty}^{\infty} x[k]\delta[n - kL], \tag{2.25}$$

where its Fourier transform

$$\tilde{x}_e(e^{i\omega}) = \sum_{n=-\infty}^{\infty}\left(\sum_{k=-\infty}^{\infty} x[k]\delta[n-kL]\right)e^{-i\omega n} = \sum_{k=-\infty}^{\infty} x[k]e^{-i\omega Lk} = X(e^{i\omega L}), \tag{2.26}$$

such that the Fourier transform of the upsampled output is a frequency-scaled version of the Fourier transform of the input. To avoid artifacts, the resulting upsampled sequence needs to be low-passed.

2.5.4 Whitening

As we have seen in Section 2.4 the noise from the detector is composed of couplings of different sub-systems, as well as its surrounding environment. In GW data analysis it is key to estimate the PSD of the data to understand the sources of noise. As we can see from the top panel of Fig. 2-7, where we present the PSDs of current and future detectors, and 2-10, where we present the raw coloured time-series data, the noise of the detector is dominated by low and high frequencies. GW signals are buried within coloured detector noise, so a common practice in GW data analysis is to make the data delta-correlated or Gaussian-like with uniform variance by removing all the correlation of the noise. In practice, the resulting PSD has equal amplitude fluctuations at all frequencies, allowing for easy comparisons. The process of transforming coloured noise to Gaussian-like noise is known as *whitening* [78], and mathematically

$$\tilde{d_w}(f) = \tilde{d}(f)/S_n^{-1/2}(f) \tag{2.27}$$

where $\tilde{d}$ and $\tilde{d_w}(f)$ are the Fourier transform of the coloured data and whitened data, respectively. To obtain the time series of the whitened data we can simply apply the inverse Fourier transform in $\tilde{d_w}(f)$. It is important to note that during this process artifacts due to aliasing can occur, so the whitened data should be cropped to avoid biasing in the subsequent analysis. In the second and fourth panels of Fig. 2-10 we show the whitened time-series data and the PSD, respectively. We can see by eye the presence of GW150914, but to highlight it we band-pass between $30 - 250$ Hz, as we can observe in the third panel, and its respective PSD in the fourth panel of Fig. 2-10.

2.6 Data quality and vetoes

To enhance the sensitivity of GW searches and reduce the number of false alarms, multiple types of data quality products are used to indicate the state of the detector during the data analysis. One such product is *data quality flags*, which indicates the suitability of collected data for analysis [39, 40].

Category 1: These flags denote that the detector noise has been severely impacted, rendering it unsuitable for astrophysical analysis. They could indicate significant changes in the properties of the noise of the detector or incorrect calibration, which might arise from incorrect detector settings or on-site maintenance work, among others. In O2 and O3 data flagged in this category represented $< 2.0\%$ of the total data.

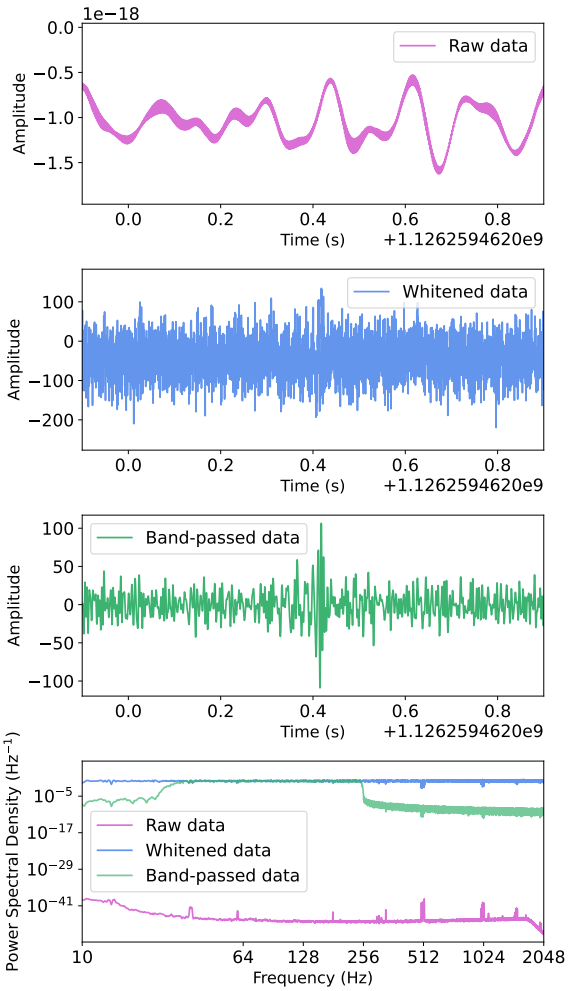

FIGURE 2-10: (First panel) Raw data from the detector $s(t)$, dominated by low and high frequencies. (Second panel) Whitened data, (Third panel) Whitened and band-passed data. (Third panel) PSDs of coloured data (pink), white data (blue), white and band-passed data (green).

Category 2: These flags indicate periods of time where the data is impacted by excess noise and should be treated with caution, as investigations have demonstrated a firm correlation between auxiliary channels and $h(t)$. Search algorithms are recommended not to consider potential candidates during these times, as they are more likely to have been caused by instrumental or environmental couplings. As Category 2 flags reduce the amount of analyzable data, they can potentially jeopardize the number of detectable gravitational waves (GWs) if the amount of time removed is not minimized. Consequently, these vetoes are generally not utilized in CBC searches unless it has been demonstrated that the flagged data significantly impairs the search. While CBC searches rely on priors of the shape of the GW, unmodelled GW have minimal to no assumptions. Thus, burst searches must add further restrictions to data quality flags, increasing the removed data. On average, over O2 and O3, CBC searches removed approximately $\sim 0.20\%$ of the data using Category 2 flags, while burst searches removed approximately $\sim 0.52\%$.

Category 3: These flags indicate periods that correlate with auxiliary channels, though the exact nature of these correlations is not yet fully understood. Most of these flags are generated with the Hveto algorithm [79] by correlating safe auxiliary channels, i.e. channels insensitive to GW, with the main strain of the detector $h(t)$. Such flags remove $\sim 10\%$ of the data and are used only by some burst searches.

The data quality flags and their details are available via Gravitational-Wave Open Science Center (GWOSC) [80, 81].