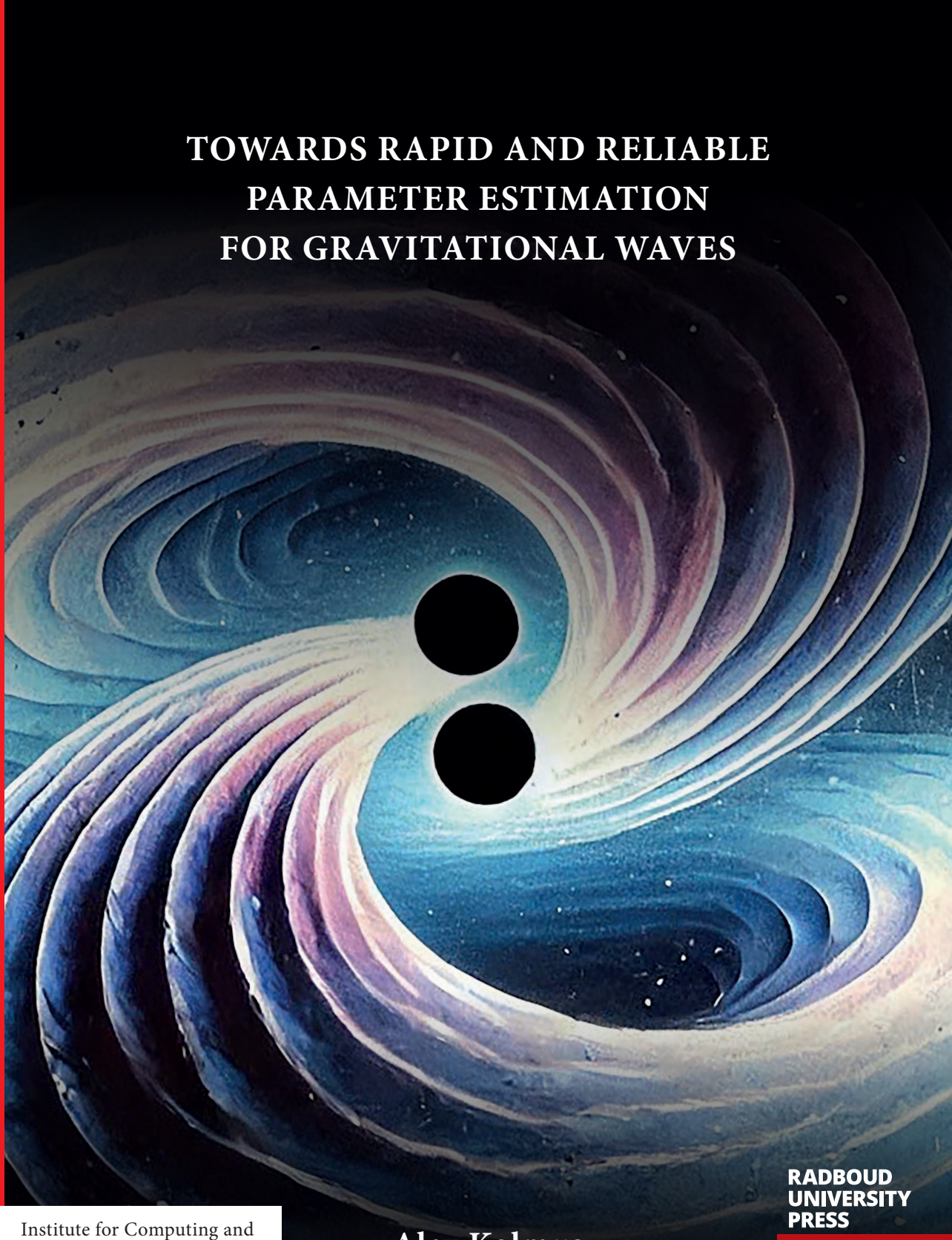
TOWARDS RAPID AND RELIABLE PARAMETER ESTIMATION FOR GRAVITATIONAL WAVES



Alex Kolmus

Institute for Computing and Information Sciences

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

TOWARDS RAPID AND RELIABLE PARAMETER ESTIMATION FOR GRAVITATIONAL WAVES

Alex Kolmus

Alex Kolmus
Towards Rapid and Reliable Parameter Estimation for Gravitational Waves

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Alex Kolmus
Cover image: DALL-E & Proefschrift AIO/Guus Gijben
Printing: DPN Rikken/Pumbo

ISBN: 9789465150871
DOI: 10.54195/9789465150871
Free download at: https://doi.org/10.54195/9789465150871

RADBOUD
UNIVERSITY
PRESS

TOWARDS RAPID AND RELIABLE PARAMETER ESTIMATION FOR GRAVITATIONAL WAVES

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen

op dinsdag 20 mei 2025
om 14:30 uur precies

door

Alex Kolmus

geboren op 23 maart 1993
te Gouda

Promotor
Prof. dr. T.M. Heskes

Copromotor
Dr. T.M. van Laarhoven

Manuscriptcommissie
Prof. dr. B. Krishnan
Dr. S.M. Nissanke (Universiteit van Amsterdam)
Dr. M. Fays (Université de Liège, België)

Contents

Chapter 1

Introduction

Humanity has always been fascinated with its place in the Universe. Ancient civilizations observed the celestial bodies and distilled meaning from their configuration. To support these interpretations, they built cosmological models [1]. For instance, the ancient Greeks formulated the geocentric model, positioning Earth at the center of the Universe with celestial bodies orbiting it in perfect spheres. Similarly, the cosmological model in ancient China portrayed Earth as a flat square, with the heavens represented by an encompassing (hemi-)sphere. Although now known to be incorrect, these models were the prevailing beliefs for the majority of human history.

The Copernican revolution of the 16th century marked a pivotal moment in the history of science, challenging centuries of cosmological and religious beliefs. In 1543 Nicolaus Copernicus published *De revolutionibus orbium coelestium* in which he argued in favor of a heliocentric cosmological model: the planets revolve around the Sun rather than around the Earth. This was a radical departure from the geocentric worldview, especially considering that at the time a planet was nothing more than a 'wandering' star. Despite its vast implications, the heliocentric cosmological model was met with mild interest, possibly due to Copernicus's untimely death not so long after the publication [2].

Galileo Galilei's observations – and controversy – brought the heliocentric model to the forefront of the public debate. In 1610, using the just-invented refracting telescope, Galilei made many discoveries: the phases of Venus, multiple moons orbiting Jupiter, sunspots, countless unknown stars, the existence of the Milky Way, and the surprising roughness of the Moon's surface. These discoveries supported a heliocentric cosmological model and sparked controversy due to potential conflicts with the Bible. The Church eventually ordered Galilei's house arrest and banished all his written work[1]. Around the same time, Johannes Kepler,

[1]Not until 1992 did the Roman Catholic Church officially exonerate Galilei [3].

through meticulous mathematical analysis and the careful observations of his employer Tycho Brahe, established the laws of planetary motion that now bear his name. These laws were consistent with the heliocentric model and provided further support. More support for the heliocentric model came from Isaac Newton in 1687 when he published *Philosophiae Naturalis Principia Mathematica*, in which he presented his law of universal gravitation and the laws of motion. These laws offered a comprehensive explanation for the observed motions of celestial bodies within the heliocentric model. By the end of the 17th century, the heliocentric model was firmly established as the prevailing cosmological model [4].

Crucial in this intellectual revolution was the technological innovation of the refracting telescope. The invention of the telescope transformed astronomy from a speculative endeavor into a rigorous empirical science, providing astronomers with the capabilities to verify hypotheses and build upon them. In the 400 years since its invention, telescopes have been continuously refined and augmented, enabling astronomers to peer deeper into the cosmos than ever before. This evolution encompassed several pivotal breakthroughs: from Galileo's pioneering observations of Jupiter's moons to William Herschel's discovery of Uranus in 1781, to Edwin Hubble's use of the 100-inch Hooker telescope in 1923 to prove the existence of galaxies beyond our own [5]. The development of radio telescopes in the 1930s opened up new electromagnetic wavelengths for observation, leading to the discovery of cosmic phenomena like quasars and pulsars [6, 7]. Most recently, the James Webb Space Telescope has provided unprecedented views of the early Universe through its Deep Field images [8], continuing this legacy of ever-expanding cosmic exploration.

However, telescopes rely on light, also known as an electromagnetic wave, which originates from and interacts only with matter that has a charge. While light-based observations have provided invaluable insights into the composition and behavior of the Universe, they only offer a partial view[2]. To expand this view, scientists have explored various forms of cosmic messengers beyond electromagnetic waves, including neutrinos, cosmic rays, and gravitational waves. Gravitational waves, in particular,

[2]It is estimated that 95% of the Universe consists of dark matter and dark energy, which cannot directly interact with electromagnetic waves [9, 10].

have opened a new window into the Universe. Unlike light, gravitational waves can be emitted from matter regardless of its charge. This unique characteristic enables gravitational wave detectors to observe cosmic phenomena that are often difficult or impossible to detect through electromagnetic radiation alone, such as binary black hole mergers, while also complementing electromagnetic observations of events like neutron star mergers. The combination of data from both electromagnetic and gravitational waves promises to provide a more comprehensive understanding of the Universe.

The story of gravitational waves begins with Einstein's revolutionary theory of general relativity, published in 1915 [11, 12]. In this theory, Einstein postulates that mass and energy bend the fabric of spacetime and that what we perceive to be the gravitational force is nothing more than objects moving along a straight line in curved spacetime. This dynamic of mass dictating the curvature of spacetime and spacetime dictating the path of the mass is captured in the Einstein field equations. In 1916, Einstein showed that these equations share a parallel with Maxwell's equations, which describe the interaction between electric and magnetic fields and their relationship with charge and current. Just as Maxwell's equations led to the deduction of electromagnetic radiation traveling at the speed of light, Einstein's work revealed the existence of gravitational radiation [13, 14]. Analogous to how accelerating a charge produces electromagnetic waves, accelerating a mass results in the emission of gravitational waves. These waves propagate outward from the accelerating mass at the speed of light. The amplitude of a gravitational wave is proportional to the acceleration and mass of the source object and inversely proportional to the distance between object and observer. The stiffness of space[3] combined with the enormous distances between Earth and stellar objects makes the direct detection of gravitational waves incredibly difficult. With current and near-future detectors, we can primarily measure events that involve the densest objects accelerating towards relativistic velocities: mergers between black holes, neutron stars, or a combination of these two, also known as compact binary coalescences. Future detectors, such as Einstein Telescope, may expand this range to include other sources like supernovae. Given the

[3]Space is 10^{20} times stiffer than steel [15].

faintness of these signals, highly sensitive and specialized technology is required for their detection.

Since the 1970s, the Michelson interferometer has been a leading contender for detecting gravitational waves [16, 17]. Comprising two perpendicular arms connected at one joint, each outfitted with mirrors at their opposite ends, the Michelson interferometer operates by dividing a laser beam at the shared joint and directing the split components along the arms to the opposite ends where a mirror reflects them back to the shared joint. There the two components interfere with one another. If the arms are equally long, the two components cancel each other out and no light is transmitted. Conversely, if the arms are not equally long, constructive interference takes place and the resulting light beam is captured by a sensor. This setup allows for the detection of incredibly small changes in the relative lengths of the interferometer arms, see also Figure 1.1. While these changes could be induced by passing gravitational waves, the extreme sensitivity of the interferometer also means it detects numerous sources of noise that can mask the gravitational wave signal. These include seismic activity and thermal fluctuations in the mirror coatings. Consequently, significant effort has gone into characterizing and mitigating these noise sources [18–20]. For example, the initial Laser Interferometer Gravitational-Wave Observatory (LIGO) – operational between 2002 and 2005 – used a single pendulum design to isolate the mirrors from seismic noise. In contrast, the current version employs a sophisticated quadruple pendulum design along with an active damping system [21]. These improvements, among many others, have resulted in a tenfold increase in sensitivity [22].

On 14 September 2015, just shy of a century after Einstein's prediction, LIGO directly detected a gravitational wave for the first time in human history [25]. Using two interferometers, one in Livingston, Louisiana, and the other in Hanford, Washington, they observed the waves emitted by the merger of two black holes. The signal was named GW150914[4] and was observed for approximately 0.2 seconds. This seemingly short blip was the

[4]All detections are named after their detection date in the year-month-day format. Since 2019, due to the increased frequency of detections, this notation is followed by: _hour-minute-second.

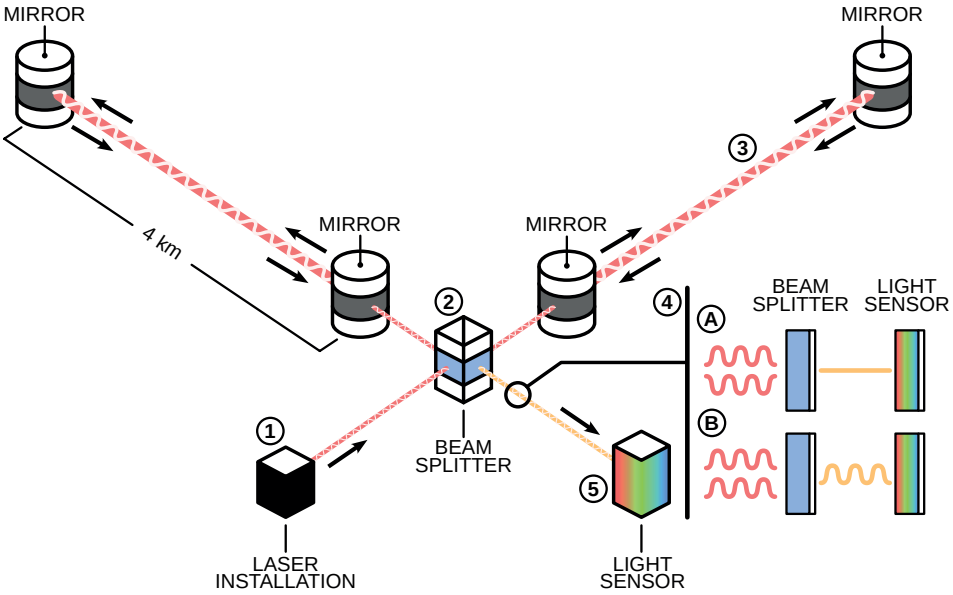


Figure 1.1: The interferometer setup at LIGO consists of the following components: **1.** A laser installation generates a 1064 nm beam with exceptional purity and a power output of 200W. This laser beam is directed towards a beam splitter. **2.** The beam splitter divides the incoming laser beam into two beams, each directed into a 4 km long arm. **3.** The arms of the interferometer function as Fabry-Perot cavities [23], which allow the laser beam to bounce back and forth multiple times, effectively increasing the arms' length and enhancing the detector's sensitivity. When the beams recombine at the beam splitter, the pattern of interference between the recombined beams reveals any relative change in the arms' lengths. **4A.** When the arms are precisely equal in length, the two beams nullify each other through destructive interference, resulting in no power transmission. **4B.** However, discrepancies in arm lengths, such as those caused by gravitational waves, lead to constructive interference, allowing power to be transmitted. **5.** The transmitted power is measured by a sensor, which allows for the determination of the disparity in arm lengths. This figure is strongly inspired by work from Johan Jarnestad [24].

start of gravitational-wave astronomy and its detection was key in Rainer Weiss, Barry Barish, and Kip Thorne being awarded the Nobel Prize in physics in 2017 [26].

In the almost nine years since GW150914, gravitational-wave astronomy has made huge strides through the collaborative efforts of a global detector network. This network includes LIGO in the United States, Virgo in Europe [27], and more recently, KAGRA in Japan [28]. Together, they have observed and analyzed a total of ninety gravitational wave events [29], among them two binary neutron star mergers [30, 31]. These observations have provided bounds on the plausibility of alternative theories of general relativity [32] and given us a glimpse into the black hole make-up of the Universe [33]. The combination of multiple detectors vastly improves sensitivity and enhances our ability to localize gravitational wave sources. This precise localization enables rapid follow-up observations with optical telescopes, potentially capturing electromagnetic counterparts to these events. The joint observation of electromagnetic and gravitational waves from compact object mergers involving neutron stars, whether in binary neutron star or neutron star-black hole systems, allows us to delve into the process of heavy element generation [34], probe the limits of general relativity further [35], and gain unique insights into the composition and structure of neutron stars [36].

A critical part of solving all of these science cases is the analysis of the events. When analyzing a gravitational wave event, our primary goal is to characterize the astrophysical source that produced the signal. This involves determining properties such as the masses and spins of the compact objects, their orbital parameters, and the distance to the source. Thus, the analysis process constitutes an inverse problem: we work backward from the observations to infer the parameters that best explain the detected signal. Essential in solving an inverse problem is a forward model that can map given parameters to a signal, also called a simulation model. Such a forward model allows us to find parameters that give rise to outputs similar to the observation, and thus characterize the source accurately. For gravitational waves coming from compact binary coalescences, multiple simulation models are available. The most accurate of these models are numerical relativity simulations [37–39], but

these are prohibitively expensive[5]. Therefore, multiple families of simpler simulation models are used, see subsection 2.1.2, each with their own set of assumptions to make the computational load manageable.

Noise is ever-present in gravitational wave observations, and plays an important role in solving inverse problems as it introduces uncertainty into observations. Rather than having a deterministic relationship between observations and parameters, the presence of noise implies that multiple parameter combinations can explain the observation well. In a Bayesian framework, this uncertainty is elegantly handled by treating parameters as random variables and expressing beliefs about them in terms of probability distributions. Bayesian approaches rely on Bayes' theorem, which states that the posterior – the probability that parameters θ are responsible for observation D – is proportional to the likelihood times the prior:

$$\underbrace{P(\theta|D)}_{\text{posterior}} \propto \underbrace{P(D|\theta)}_{\text{likelihood}} \times \underbrace{P(\theta)}_{\text{prior}}.$$

The likelihood function $P(D|\theta)$ expresses the probability of observing D assuming that the given parameters θ and the simulation model are correct; it quantifies how well the parameters explain the observed data. For noisy inverse problems, the likelihood function is defined by the noise distribution, as it measures how well the observation minus the simulation matches the noise distribution. The prior distribution $P(\theta)$ captures our initial assumptions about the parameters. In cases where no prior knowledge is available, one might assign a uniform distribution over the possible values of θ. Once the likelihood function is defined and the prior distribution is chosen, Bayesian inference methods allow us to compute the posterior distribution.

In gravitational-wave astronomy, a popular Bayesian inference method is nested sampling [41]. This method divides the prior into nested shells of ascending likelihood and constructs a posterior distribution from them. The process begins by randomly sampling a set of points from the prior distribution. Each of these points represents a possible set of parameters for the model. These points are then evaluated based on their likelihood,

[5] A single numerical-relativity simulation of a binary black hole merger can take months to complete on a supercomputer [40].

and the one with the lowest likelihood is identified. This lowest likelihood point defines the outer boundary of the first shell. Next, this point is replaced with a new point drawn from the prior, but constrained to have a higher likelihood than the current boundary. This iterative process of identifying and replacing the lowest likelihood point continues, each time shrinking the shell and increasing the overall likelihood threshold. By systematically refining the parameter space in this manner, nested sampling effectively handles multi-modal distributions and, given a sufficiently large set of points, is guaranteed to accurately determine the posterior distribution. This robustness and efficiency make nested sampling particularly well-suited for the multi-modal and high-dimensional parameter spaces encountered in gravitational-wave data analysis[6].

Nested sampling offers several advantages, including effective handling of multi-modalities, the generation of independent posterior samples, and it can easily be parallelized. However, its computational demands and time-consuming nature are notable challenges; the analysis of a short gravitational wave event can require multiple hours, even on powerful hardware [43]. At present, given the current observation rates and technological capabilities, these challenges are manageable. Nonetheless, as detectors undergo upgrades and new instruments such as Cosmic Explorer or Einstein Telescope[7] are built, the overall sensitivity is expected to increase significantly, see the left side of Figure 1.2. This improved sensitivity will enable exploration of much larger volumes of space, and thereby increase the rate of binary merger detections as is illustrated in the right side of Figure 1.2. To demonstrate this (anticipated) increase in detection frequency, we provide the (expected) event rate of binary black hole mergers for several dates: during the initial run in 2015, the rate was 8.5 events per year [44]; by the third run in 2020, this rate had escalated to 81 events per year [29]. Projections for 2025 estimate a rate ranging between 10^3 and 10^4 events per year, while by 2030, it is anticipated to reach between 10^4 and 10^5 events per year [45]. In addition to the exponential growth of the number of events, their life time in the detectors' sensitive band will also drastically increase. The duration of an event is

[6]The level of robustness differs between implementations of nested sampling [42].

[7]These future detectors are part of the third generation (3G) detectors.

strongly dependent on the lowest sensitive frequency of the detectors; current analysis often uses a value of 20 Hz. Halving this sensitive frequency roughly sextuples the event duration. Realistically, it is anticipated to reach 10 Hz after current detectors receive upgrades and at most 5 Hz when Cosmic Explorer and Einstein Telescope become operational. Even when ignoring more compounding factors, a conservative back-of-the-envelope calculation suggests that the analysis of all black hole merger events in 2030 would require a computer cluster to continuously run for 45 years. In reality, this number should be even higher, considering the number of observable binary neutron stars will also skyrocket. These systems, due to their lower masses, produce significantly longer signals than binary black holes, drastically increasing their analysis time. Additionally, the necessity for repeated inference runs with different simulation models to test scientific hypotheses further compounds the computational challenge. In summary, nested sampling is well-suited for the current gravitational wave event rate, however, the expected exponential surge in gravitational events in the coming years poses a difficult challenge.

In preparation for this significant challenge, an increasing number of studies are focused on speeding up gravitational wave inference. Various solutions are being explored, including alternative simulation models [48, 49], efforts to utilize accelerator hardware [50, 51], interpolation strategies to ease the computational load [52–55], and the development of machine learning methods to replace Bayesian inference [56–58]. The latter, often called likelihood-free or simulation-based inference, aims to construct the posterior distribution without direct reliance on the likelihood function. Within simulation-based inference, strategies can broadly be categorized into two approaches. The first approach involves constructing a metric to compare simulations with observations. However, this is less relevant for gravitational wave inference since the majority of the computational load is the simulation itself. The second approach focuses on inferring a generic relationship between simulations and the posterior distribution, enabling the prediction of posterior distributions for observations without the need for simulations during inference. The second strategy is particularly desirable for gravitational waves, as it eliminates the need for real-time simulations during inference. Sadly, for gravitational waves – and most

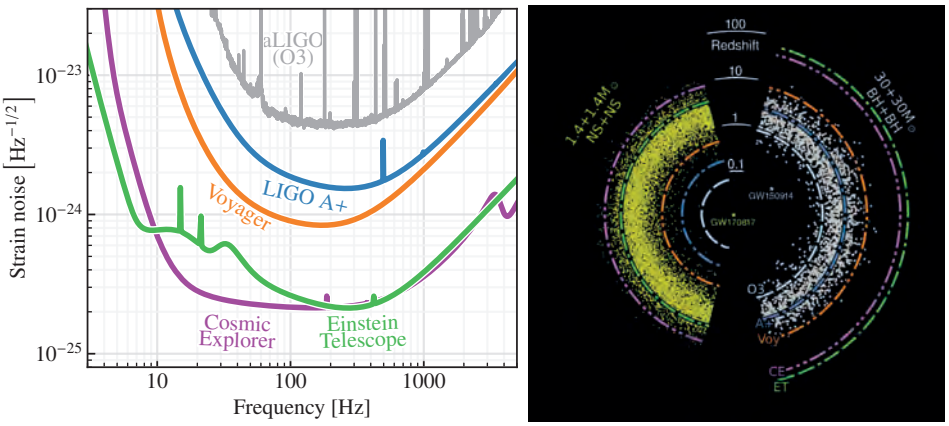


Figure 1.2: Comparison of detection capabilities between current and future detectors. **Left:** The (expected) noise amplitudes of various detectors are depicted as a function of frequency, with lower values indicating greater sensitivity. The gray line represents the measured noise strength of the LIGO detector during the O3 run (2019-2020), while the thick-colored lines represent either upgrades (LIGO A+ and Voyager) or new detectors (Cosmic Explorer and Einstein Telescope). The figure is taken from [46]. **Right:** The detection range of these upgraded and future detectors is shown for binary neutron star and binary black hole mergers, along with the corresponding number of events it represents. The yellow and white dots represent populations of binary neutron star and binary black hole mergers respectively. These upgrades and new detectors promise access to a significantly larger volume of space, thereby drastically increasing the number of detections. The figure is adapted from [47].

other inverse problems – establishing the relationship between simulations or observations and their posterior distributions is an immensely challenging task, hindered by the highly non-linear and high-dimensional nature of the relationship.

Fortunately, neural networks are a powerful tool for modeling highly complex relationships. Since the deep learning revolution in 2012 [59, 60], they have emerged as the go-to approach for constructing the relationships between high-dimensional inputs, for example, images, and corresponding observables, such as the object in the image. Consequently, neural networks have gained considerable prominence in the field of simulation-based inference. Prominent neural methods mimic a part of the Bayesian inference toolbox: the likelihood function [61], the likelihood-ratio [62], or the posterior distribution [63]. The last method, often referred to as neural posterior estimation, can generate an estimate of the posterior distribution within seconds. If the estimates are accurate, it could enable the analysis of gravitational waves well into the future.

To explain how neural posterior estimation works in practice, let us consider the process of training such a model to infer the parameters of a binary black hole merger. This method requires the following components:

Prior distribution: a chosen distribution spanning the parameter space.

Simulation model: generates gravitational wave data for specified parameters.

Neural network: maps gravitational wave inputs to a defined set of variables.

Flexible distribution: represents the posterior distribution, with the neural network's outputs determining its shape and coverage across the parameter space.

To train the model, synthetic training data is created by sampling true values from the prior distribution and generating corresponding gravitational wave signals using the simulation model. By embedding the signal in simulated noise, we approximate real-world observational conditions. During training, the neural network learns to map these synthetic observations to

a distribution. The quality of this prediction is evaluated by the likelihood of the true parameters in the predicted distribution. This likelihood is then used to define the loss function for optimizing the neural network. With each prediction iteration, the network adjusts its internal parameters to minimize the loss function. After millions of iterations, the neural network learns to approximate the actual posterior distributions with its predicted distributions. Once trained, this model can rapidly estimate posterior distributions for new gravitational wave observations, potentially enabling real-time parameter inference for future high-rate detections.

Neural posterior estimation and its counterparts in neural simulation-based inference hold considerable promise, generating posterior distributions in a fraction of the time required by traditional Bayesian inference methods. However, these predictions are made by neural networks, which are essentially black-box models. They excel at learning mappings between given inputs and outputs, but their predictions lack interpretability because the reasoning is hidden within a complex, non-linear computation. This opacity can undermine trust in the model's prediction. Moreover, neural networks are known to suffer from catastrophic forgetting [64] and miscalibration [65, 66] even when trained on millions to billions of samples. They can also exhibit unexpected behavior when given inputs outside their training data. For these simulation-based inference methods to serve as viable alternatives, they must match the reliability and accuracy of traditional Bayesian inference while retaining their rapid inference speed.

This thesis aims to improve the reliability and capabilities of the neural posterior estimation methods, with a specific focus on their application to gravitational waves. **Chapter 2** provides the necessary background to understand the remaining chapters. In **Chapter 3**, we demonstrate that one can use importance sampling to verify and improve probabilistic skymaps produced by Von-Mises distributions parameterized by neural networks. Then, in **Chapter 4**, we demonstrate the ability of continuous normalizing flows to estimate the posterior distribution of overlapping signals well even in regimes where traditional methods fail. In **Chapter 5**, we address the specific challenges posed by high-variance regions of the parameter space for neural posterior estimation and propose the use of

so-called effective priors and a fine-tuning scheme. We demonstrate the improvements of our proposed solutions on simulated low-mass binary black hole signals. Finally, in **Chapter 6**, we summarize our results and sketch our vision of the future gravitational-wave analysis pipelines.

Chapter 2

Background

This chapter provides an overview of the essential concepts in gravitational wave physics and machine learning. We start with linearized gravity and its prediction of Gravitational Waves (GWs) and their properties. The discussion then moves to the characterization and simulation of GWs from binary systems. Finally, we explore parameter estimation techniques, including traditional Bayesian methods and novel neural approaches. While this chapter does not present new ideas, it aims to give the reader the necessary information to understand the remainder of this thesis. Readers already familiar with these topics may choose to skip sections.

2.1 Gravitational Waves

Einstein's theory of gravity, known as General Relativity (GR), is the leading explanation for gravity and is widely regarded as one of the most thoroughly tested and respected theories in physics. One of its predictions is the existence of GWs in spacetime. Here, we will establish the theoretical foundation of GWs and discuss their relationship with binary systems using linearized gravity.

The Einstein Field Equations (EFE) form the foundation of GR. They are a set of mathematical expressions that describe the relationship between spacetime curvature and the distribution of energy-momentum[1]:

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu}. \tag{2.1}$$

In this equation, $T_{\mu\nu}$ is the symmetric energy-momentum tensor, which describes the density and flux of energy and momentum. The constants G

[1] A note on notation: we will use Greek letters, particularly μ and ν, to denote one of the four spacetime components, where 0 represents time, and $\{1, 2, 3\}$ represent the spatial dimensions.

and c represent the gravitational constant and speed of light, respectively. $G_{\mu\nu}$ is the symmetric Einstein tensor, which characterizes the curvature of spacetime. The Ricci tensor $R_{\mu\nu}$ and the Ricci scalar R describe particular aspects of this curvature, while the metric tensor $g_{\mu\nu}$ defines the geometric properties of spacetime.

The EFE do not fully constrain spacetime curvature, and thus the mass distribution across space does not uniquely define the spacetime curvature. Moreover, hidden inside equation 2.1 are ten coupled non-linear partial differential equations. Consequently, deriving exact solutions for the EFE is generally infeasible. Solutions are typically found only under idealized conditions, constrained by symmetries or boundary conditions, as shown by the Schwarzschild [67] and Kerr [68] solutions.

One common idealized condition is the absence of strong gravitational sources. Under this assumption, the spacetime metric $g_{\mu\nu}$ can be modeled as a sum of the flat background metric $\eta_{\mu\nu}$ and a small perturbation $h_{\mu\nu}$:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \quad \text{with} \quad |h_{\mu\nu}| \ll 1. \tag{2.2}$$

Given the small magnitude of $h_{\mu\nu}$, any terms beyond linear order can be neglected. This approximation is known as linearized gravity. The harmonic gauge simplifies the equations significantly by imposing a condition on the coordinates, reducing the degrees of freedom. By applying this gauge transformation and several lines of calculus[2], the EFE reduce to:

$$\left(-\frac{\partial^2}{c^2\partial t^2} + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\bar{h}_{\mu\nu} = -\frac{16\pi G}{c^4}T_{\mu\nu}. \tag{2.3}$$

In vacuum, where there is no matter or energy ($T_{\mu\nu} = 0$), the wave equation reduces to a homogeneous form, describing waves propagating at the speed of light through spacetime – these are GWs. While the harmonic gauge simplifies the equations, it leaves six degrees of freedom. By further imposing the transverse-traceless gauge, we can reduce these to just two degrees of freedom, exhausting the remaining gauge freedom. Under these

[2]For the full derivation, we refer the reader to chapter 1 of [69].

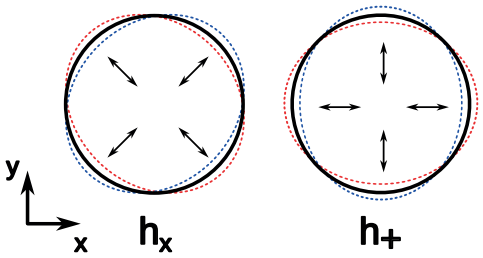


Figure 2.1: Illustration of the two polarization states on a ring of particles. The black ring represents the particles at rest, while the blue and red dashed rings represent the particles when the GW has phases equal to $\pi/2$ and $3\pi/2$, respectively.

conditions, the expression for a GW traveling along the z-axis becomes:

$$\bar{h}^{TT}_{\mu\nu} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & h_+ & h_\times & 0 \\ 0 & h_\times & -h_+ & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \cos\left(\omega\left(t - \frac{z}{c}\right)\right). \tag{2.4}$$

A wave with angular frequency ω, and the two degrees of freedom h_+ and $h_\times$ represent the two polarization states of the GW, often referred to as the 'plus' and 'cross' polarizations respectively. A visualization is shown in Figure 2.1.

While the above scenario provides a foundation for understanding GWs in the absence of strong sources, our primary interest lies in the GWs emitted by compact binary systems, such as pairs of orbiting neutron stars or black holes. Compact binaries distinguish themselves as remarkable GW emitters due to two key factors: (1) the continuous acceleration of their orbiting masses and (2) their capacity to orbit at extraordinarily close distances. The extreme density of these objects allows them to endure strong tidal forces that would pull apart normal stars, producing more intense gravitational fields and significantly stronger GWs. As the binary components draw closer, they reach relativistic velocities, further amplifying the intensity of the emitted GWs. To model GW emission from these systems, we solve equation 2.3 for two point masses m_1 and m_2. This solution assumes a distant observer and employs the weak-field, slow-

motion limit ($v << c$) of GR[3]. The first-order expression for the emitted GWs by a point mass binary system is:

$$h_+(t) = \frac{4}{r}\left(\frac{G\mathcal{M}_c}{c^2}\right)^{\frac{5}{3}}\left(\frac{\omega_{orb}}{c}\right)^{\frac{2}{3}}\left(\frac{1+\cos^2\iota}{2}\right)\cos\left(2\omega_{orb}\left(t-\frac{r}{c}\right)\right), \quad (2.5)$$

$$h_\times(t) = \frac{4}{r}\left(\frac{G\mathcal{M}_c}{c^2}\right)^{\frac{5}{3}}\left(\frac{\omega_{orb}}{c}\right)^{\frac{2}{3}}\cos(\iota)\cos\left(2\omega_{orb}\left(t-\frac{r}{c}\right)\right), \quad (2.6)$$

where r is the distance to the observer, ω_{orb} the orbital angular frequency, $\mathcal{M}_c$ is the chirp mass of the binary system, and ι is the inclination angle of the orbital plane relative to the line of sight, see also Figure 2.2. The chirp mass is expressed as:

$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}, \quad (2.7)$$

which is the dominant factor in the amplitude and frequency evolution.

While the expressions derived so far represent a stable orbit, in reality, the emission of GWs causes the binary system to lose energy, resulting in orbital decay. This leads to a characteristic 'chirp' signal, where both the frequency and amplitude of the GWs increase over time as the objects spiral closer together. The rate of this orbital decay is primarily determined by the system's chirp mass. For a circular orbit, we can quantify this evolution. The orbital frequency change due to GW emission can be approximated to leading order as [69]:

$$\frac{df_{GW}}{dt} = \frac{96\pi^{8/3}}{5}\left(\frac{G\mathcal{M}_c}{c^3}\right)^{5/3} f_{GW}^{11/3}\left(t-\frac{r}{c}\right), \quad (2.8)$$

where f_{GW} is the GW frequency, which is twice the orbital frequency ω_{orb}. This equation demonstrates how the frequency increases more rapidly for systems with larger chirp masses, leading to a faster inspiral and shorter signal durations.

Beyond these considerations of orbital dynamics, a more accurate model of GWs must incorporate additional complexities. Moving beyond

[3]For a full derivation, we refer the reader to chapters 3 and 4 of [69].

the quadrupole approximation, we can express the GW as a sum of spin-weighted spherical harmonics [70]:

$$h(t,\theta,\phi) = \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} h_{\ell m}(t)\,{}_{-2}Y^{\ell m}(\theta,\phi). \tag{2.9}$$

Here, $h_{\ell m}(t)$ are the mode amplitudes and ${}_{-2}Y^{\ell m}(\theta,\phi)$ are the spin-weighted spherical harmonics, (θ, ϕ) defining the sky location with respect to the detector. While the $(\ell = 2, m = \pm 2)$ modes correspond to the quadrupole radiation we have discussed earlier, higher values of ℓ and m represent higher-order or sub-dominant modes. These higher-order modes, though typically less prominent, play a crucial role in certain scenarios. They become increasingly significant for systems with highly uneven mass pairings, large total masses, or significant inclination. By carrying additional information about the source, these modes can be vital for accurate parameter estimation. The importance of these higher-order modes has been demonstrated in recent analyses by the LIGO-Virgo-KAGRA collaboration. For instance, in the analysis of GW190814, a binary system with a highly asymmetric mass ratio, the inclusion of higher-order modes significantly improved the precision of the source parameter estimates [71].

In this section, we have established the basic theory of GWs and explored their relationship with binary systems using linearized gravity. While these simplified expressions are useful for conveying core concepts, they fall short of describing real GW signals. To bridge this gap, we must first explore the key parameters and characteristics that shape GWs and affect their detection, which is the focus of our next subsection. Following this, we will introduce advanced simulation models for GWs. These sophisticated models, though more intricate, provide a substantially more accurate representation of GW signals throughout the entire merger process.

2.1.1 Characterization

In Chapter 1, we gave a brief description of ground-based interferometers for GW detection[4]. Here, we will take the next step and detail the interaction between GWs and the detector to understand how to translate GWs into measurable signals. Subsequently, we will discuss all the parameters required to fully characterize the detected GW.

As illustrated in Figure 2.1, a GW affects its surroundings by periodically stretching spacetime in one direction while simultaneously squeezing it in the perpendicular direction, and vice versa. Interferometric detectors are designed to measure these minute distortions in spacetime. The L-shaped configuration, as shown in Figure 1.1, represents the most straightforward geometry to implement for GW detection, effectively capturing the differential changes in perpendicular directions. A passing GW alters the proper distance between the interferometer's mirrors, effectively changing the length of the interferometer arms. This change in arm length affects the path of the laser light traveling within the interferometer. The resulting change in power output from the interferometer is proportional to the fractional difference in the arm lengths [72]. This fractional difference is quantified by the strain

$$h(t) = \frac{\Delta L(t)}{L}, \tag{2.10}$$

where $\Delta L(t)$ is the change in length of the interferometer arm caused by the passing GW, and L is the original length of the arm.

To accurately interpret the signals detected by these interferometers, we must consider the antenna pattern functions, F_+ and $F_\times$. These functions are required to translate a GW signal from the source frame to the detector frame. In the source frame, we define the GW propagation direction as the z-axis. However, this source frame z-axis generally does not align with the z-axis of the detector frame. The antenna pattern functions account for this misalignment, allowing us to correctly project the incoming GW signal onto the detector's reference frame. In Figure 2.2, the frames and the relevant angles are shown. The expressions for the antenna pattern

[4]For a more extensive description, we refer the reader to chapter 9 of [69] or to watch [72].

functions can be found in [73]. The final expression for the strain due to a passing GW reads

$$h(t) = F_+(\theta, \phi, \psi)h_+(t) + F_\times(\theta, \phi, \psi)h_\times(t). \tag{2.11}$$

Note that sky angles θ and ϕ are fixed in the detector frame, which itself moves as the Earth rotates. Astronomers therefore often transform the sky angles to the equatorial angles, right ascension (RA), and declination (DEC), which do not depend on Earth's rotation.

To characterize a GW means to assign its measurement with a physical description. This description is often a set of parameters. In GW science, these parameters are categorized into two sets: intrinsic and extrinsic. Intrinsic parameters concern the properties inherent to the source of the GW, while extrinsic parameters relate to all remaining descriptors. So far, we have already introduced quite a few of the GW parameters. Now, we will introduce the remaining ones.

Starting with the intrinsic parameters, the no-hair conjecture states that a stationary black hole[5] is fully described by only three quantities: its mass, electric charge, and angular momentum [75]. However, in the context of Binary Black Holes (BBHs), the electric charge is typically ignored. This is because the expected charge of astrophysical black holes is negligible, and the distance between the two black holes in a binary system is so large that the Coulomb force is insignificant. This leaves mass and angular momentum (or spin) as the primary intrinsic parameters that define the properties and behavior of a BBH.

In a binary black hole system, we have two masses and two spin vectors. The black holes are indexed by their mass, with the more massive black hole assigned index 1 and the less massive one assigned index 2. A popular alternative mass parameterization is to use the chirp mass $\mathcal{M}_c$ and the mass ratio $q = m_2/m_1$, since the chirp mass is a strong descriptor of the GW polarizations, see also equation 2.6.

The spin vectors of the black holes, denoted as $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, can influence the GW signal to varying degrees, depending on their magnitude and

[5]Other binary systems, such as neutron star binaries, have additional parameters like tidal deformability, which describes the stars' deformation in response to gravitational fields.

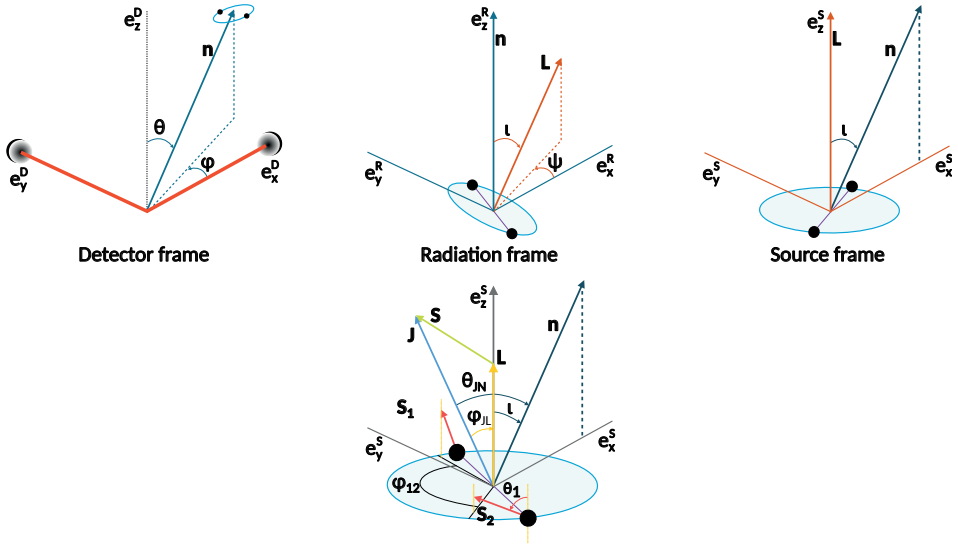


Figure 2.2: Three coordinate systems for GW calculations, with an additional panel for precessing binaries. **Top left:** Detector frame, the interferometer's orthogonal arms define the x and y axes. Angles θ and ϕ define the line of sight $\boldsymbol{n}$ between the detector and binary. **Top middle:** Radiation frame, $\boldsymbol{n}$ forms the z-axis. ι (inclination) and ψ (polarization) are the polar and azimuth angles of orbital angular momentum $\boldsymbol{L}$. **Top right:** Source frame, $\boldsymbol{L}$ defines the z-axis. $\boldsymbol{n}$'s projection onto the binary plane forms the x-axis. **Bottom:** Precessing binary configuration. $\boldsymbol{J}$ is the total angular momentum, $\boldsymbol{S}$ is the total spin (sum of spin components $\boldsymbol{S}_{1,2}$). The spin configuration can also be described by the spin component magnitudes, spin difference angle ϕ_{12}, angle between $\boldsymbol{J}$ and $\boldsymbol{L}$ (ϕ_{JL}), and tilt angles $\theta_{1,2}$ (θ_2 omitted for figure clarity). For precessing systems, the inclination angle is θ_{JN}. Figure adapted from [74].

orientation. In some configurations, particularly when misaligned with the orbital angular momentum, spins can induce precession effects that modulate the waveform. While challenging to measure precisely, spin information can potentially offer insights into the binary's formation history [76].

Spins can be parameterized using a Cartesian representation, but due to their changing orientation during the merger, a spherical parameterization is more commonly used. In the spherical parameterization, the spin magnitude is often parameterized by:

$$\chi_{1,2} = \frac{c|\boldsymbol{S}_{1,2}|}{Gm_{1,2}} \tag{2.12}$$

or its dimensionless counterpart $a_{1,2} = \chi_{1,2}/m_{1,2}$.

The orientation of the spin vectors is then given by the tilt angles $\theta_{1,2}$, which are the angles between $\boldsymbol{S}_{1,2}$ and the orbital angular momentum $\boldsymbol{L}$, and the azimuthal angles ϕ_{12} and ϕ_{JL} which are all depicted in Figure 2.2. These parameters fully describe the spin configuration, but their importance depends on whether the system is aligned or precessing:

> Aligned spin systems: in these systems, both spin vectors are parallel or anti-parallel to $\boldsymbol{L}$. The spins can be fully characterized by projecting their magnitudes onto the z-axis (direction of $\boldsymbol{L}$). These systems do not exhibit precession and produce simpler gravitational waveforms.
>
> Precessing spin systems: in these more complex systems, at least one spin vector is not aligned with $\boldsymbol{L}$. All spin parameters (magnitudes and orientations) are needed to describe the system. The misalignment causes the orbital plane and spins to precess around the total angular momentum vector, leading to modulations in the GW signal.

Aligned spin systems are simpler to model but may miss important physical effects, while precessing systems capture the full complexity of the binary's dynamics but require more sophisticated analytical and computational techniques.

Most of the extrinsic parameters have been introduced already, but we will cover them here briefly again. The relative position of the binary with respect to the detector is given by the distance r between the two and the azimuth angles θ and ϕ. These are, however, not the most convenient parameterization; instead, the luminosity distance[6] D_L and the equatorial angles RA and DEC are used. Another extrinsic parameter is the inclination angle. For non-precessing systems, this is denoted as ι, which represents the angle between the line of sight and the orbital angular momentum $\boldsymbol{L}$. However, in precessing systems, where the orbital plane itself evolves, we instead use θ_{JN}, which is the angle between the line of sight and the total angular momentum $\boldsymbol{J}$. This distinction is necessary because $\boldsymbol{J}$ remains approximately constant in precessing systems, while $\boldsymbol{L}$ does not. The polarization angle ψ is required to translate the polarization from the source frame to the detector frame.

The two extrinsic parameters we have not covered yet are the coalescence time t_c and coalescence phase ϕ_c. The coalescence time is the exact time at which the two black holes merge. The coalescence phase is the phase of the GW at the time of merger. Both parameters are critical for aligning the waveform with the observed signal. In practice, the time at which the GW signal reaches the Earth's center, the geocentric arrival time, t_{geo}, is used instead of t_c.

We have now discussed all the parameters required to characterize a measured GW, summarized in Table 2.1. In the next section, we will discuss how to simulate a GW given these parameters.

2.1.2 Simulations

We have now discussed all the parameters required to characterize a BBH merger. In the preceding sections, we introduced the concepts necessary for understanding GWs and the parameters essential for their characterization. Now, we shift our focus to the methods used to simulate these GWs, particularly from binary systems. Simulations play a vital role in GW astronomy, serving as the cornerstone for both data analysis and

[6]The luminosity distance is a concept in cosmology that represents the distance an astronomical object would be at if the universe were not expanding.

Category	Parameter	Symbol
Intrinsic	Chirp mass	$\mathcal{M}_c$
	Mass ratio	q
	Spin magnitudes	a_1, a_2
	Tilt angles	θ_1, θ_2
	Azimuthal angles	ϕ_{12}, ϕ_{JL}
Extrinsic	Luminosity distance	D_L
	Right ascension	RA
	Declination	DEC
	Inclination angle	ι or θ_{JN}
	Polarization angle	ψ
	Geocentric arrival time	t_{geo}
	Coalescence phase	ϕ_c

Table 2.1: Summary of parameters used to characterize a GW signal from a BBH merger. Parameters are categorized as intrinsic or extrinsic. Note that for aligned-spin systems, only the z-components of the spin magnitudes (χ_{1z}, χ_{2z}) are used instead of the full set of spin parameters. For precessing systems, θ_{JN} is used instead of ι for the inclination angle. Alternative mass parameterizations, e.g. the individual masses m_1, m_2, are not shown but can be derived from the listed parameters.

theoretical predictions. Here, we will give an overview of the numerous methods that exist to simulate GWs and provide context concerning their use cases.

Previously, we examined the expressions for GWs emitted by binary systems under the assumption of stable orbits and non-relativistic approximations. While these approximations are useful, they fall short in accurately describing the highly dynamical and relativistic regime of black hole mergers. As the two black holes spiral closer and eventually merge, their velocities approach a significant fraction of the speed of light, and the gravitational fields become extremely strong. In such scenarios, the analytical solutions we discussed earlier are no longer sufficient, necessitating more sophisticated approaches.

To understand why more advanced methods are needed, it is important to consider the different phases of black hole mergers: inspiral, merger, and ringdown, see also Figure 2.3. During the inspiral phase, the black holes gradually lose energy through GW emission and spiral inward in a manner that can be accurately described by analytical methods. The merger phase involves the highly non-linear dynamics of the black holes colliding, producing the most intense GWs. Finally, the ringdown phase occurs as the newly formed single black hole, resulting from the merger, settles into a stable state, emitting GWs characterized by a series of damped oscillations. Each of these phases presents unique challenges and requires different modeling techniques to accurately describe the GWs produced.

To obtain complete and accurate solutions for the GWs emitted by merging black holes, we must numerically solve the EFE with initial and boundary conditions tailored to describe a BBH system. This approach, known as Numerical Relativity (NR), involves discretizing spacetime into a large, finite grid and evolving this grid over time according to the EFE. These problems are highly non-trivial, requiring over a decade of research to progress from early simulations of head-on colliding black holes [77] to simulations of non-spinning, equal-mass BBH mergers [78]. Subsequently, [79, 80] enabled simulations of spinning, unequal mass BBHs, expanding the scope of NR to more realistic astrophysical scenarios. Since then, NR has made significant strides. There are now multiple open GW catalogs, such as those in [81–83]. Despite these advancements and the

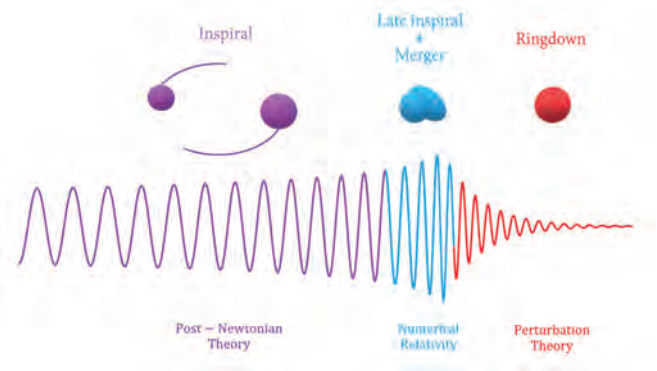


Figure 2.3: Color-coded representation of a GW's evolution. Inspiral (purple) modeled by Post-Newtonian theory shows increasing frequency. Late inspiral and merger (blue) require NR, capturing the peak emission. Ringdown (red), described by Perturbation Theory, exhibits decaying oscillations. Each phase demands distinct theoretical approaches for accurate modeling. Figure adapted from [85].

use of supercomputers, these three cited catalogs hold roughly 4500 waveforms. However, it is crucial to understand the limitations of this dataset, as they directly impact our ability to study the full range of astrophysical scenarios. First, many simulations concentrate on the merger's final stages, omitting earlier parts of the inspiral. Second, the precision of these waveforms is not uniform, with varying levels of accuracy across different simulations. Third, the waveforms are not sampled uniformly across the parameter space. This uneven coverage means that some regions of the parameter space, such as extreme mass ratios or high spin configurations, are underrepresented [81, 84]. These limitations underscore the ongoing challenges in NR, despite its significant progress.

While NR provides highly accurate waveforms, its computational demands make it impractical for large-scale analyses. Instead, researchers often rely on waveform approximants, which are simplified models that can be computed more efficiently while still capturing the essential physics of gravitational waveforms. Below, we discuss the main categories of these approximants in detail, highlighting their pros and cons and providing examples:

Post-Newtonian (PN) expansions describe the inspiral phase of a merging BBH by expanding the phase and time[7] evolution as a Taylor series with powers of v/c. They are particularly effective when velocities are relatively low and relativistic effects are weak. This makes PN expansions suited for the early inspiral phase and computationally efficient. However, as the black holes get closer and their velocities increase, the accuracy of PN expansions diminishes, limiting their applicability to later stages of the merger. A well-known example of a PN waveform is TaylorF2 [86–88].

Effective One-Body (EOB) framework reinterprets the two-body problem in general relativity as an effective one-body problem. It constructs a Hamiltonian that governs the motion of a single particle in an effective potential, accurately representing the original two-body dynamics. EOB models are suitable for modeling all phases of BBH mergers, including the inspiral, merger, and ringdown, as they combine elements of PN theory with additional relativistic corrections. However, they are more complex and computationally intensive than PN expansions and require careful calibration with NR results to ensure accuracy. The SEOBNRv5PHM waveform [89] is one of the most accurate approximants available.

Surrogate models use interpolation and decomposition techniques to fit the space between precomputed NR waveforms. These models excel in two key aspects: speed and accuracy. They can generate waveforms rapidly, while achieving accuracy comparable to NR simulations in well-sampled regions of the parameter space. This makes them valuable for in-depth analyses of specific events of interest. However, their accuracy is limited by the range and density of NR simulations used in training. In regions with fewer samples, the quality of surrogate models significantly decreases. Since the number of NR waveforms is low, parts of parameter space are poorly covered. A notable example is the NRSur7dq2 waveform model [90].

[7] Some variants expand in frequency instead of time.

Phenomological models combine multiple approximation methods to create efficient and accurate hybrid waveforms. These models use PN theory as a base for the inspiral, supplemented with fitted correction terms. They model the ringdown using perturbation theory and fit the merger to NR waveforms using nontrivial factorization schemes. A key feature of these models is how they stitch together these different regions to create a coherent waveform. This approach allows phenomenological models to balance computational efficiency with accuracy across a wide parameter range. Their speed and accuracy make them popular choices for GW analysis. A notable example is the IMRPhenomXPHM waveform model [91].

While improved waveform models enhance our theoretical understanding of GWs, their true value lies in application to observational data. The refinements in modeling precession, higher-order modes, and the merger-ringdown phase allow for more precise comparisons with detected signals. The next section focuses on parameter estimation – the process of inferring the physical properties of GW sources by comparing these models with detector data.

2.2 Parameter Estimation

This section explores the landscape of parameter estimation techniques in GW science. We start with Bayesian inference and traditional sampling methods that have been the workhorse of GW analysis since the first detection. We then progress to neural density estimators that could potentially address the computational challenges posed by future detectors like Cosmic Explorer and Einstein Telescope.

2.2.1 Bayesian Inference

The interplay between theoretical modeling and observational inference is fundamental to GW astronomy. The waveform models discussed earlier are essential tools in parameter estimation. This process requires us to invert our approach: instead of predicting signals from known parameters,

we must infer parameters from observed signals. This inverse problem is central to GW astronomy. Building on Chapter 1, we now present a more formal description of the statistical framework used to address this challenge. At the core of this framework is Bayes' theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{d},\mathcal{H}) = \frac{p(\boldsymbol{d}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(\boldsymbol{d}|\mathcal{H})}. \tag{2.13}$$

In this equation, $\mathcal{H}$ represents our hypothesis, realized through the chosen waveform model. The vector $\boldsymbol{\theta}$ contains the parameters we aim to infer, and $\boldsymbol{d}$ represents the observed detector data. The four terms in Bayes' theorem are referred to as the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d},\mathcal{H})$, the likelihood function $p(\boldsymbol{d}|\boldsymbol{\theta},\mathcal{H})$, the prior distribution $p(\boldsymbol{\theta}|\mathcal{H})$, and the evidence or marginal likelihood $p(\boldsymbol{d}|\mathcal{H})$. It is worth noting that the hypothesis $\mathcal{H}$ is often omitted from the notation for brevity, especially when dealing with a single model or hypothesis.

Bayes' theorem enables inference about unknown parameters $\boldsymbol{\theta}$ by relating the posterior probability to the product of the prior probability and the likelihood, scaled by the marginal likelihood of the data. The power of Bayes' theorem lies in its ability to reverse conditional probabilities, allowing us to infer causes (parameters) from effects (observed data). This provides a formal mechanism for combining prior knowledge with observed data to draw conclusions about model parameters or hypotheses. In GW science, this approach is crucial for extracting information about GW sources from detector data, enabling us to estimate properties of astrophysical systems from the signals they produce.

Two quantities are of particular interest in scientific analysis: the posterior distribution and the evidence. Bayes' theorem, as shown in equation 2.13, provides an expression for the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d},\mathcal{H})$, which quantifies the probability of parameter values given the data and hypothesis. The evidence $p(\boldsymbol{d}|\mathcal{H})$ allows us to compare different hypotheses quantitatively and is obtained by integrating the likelihood function over all possible parameter values, weighted by their prior probabilities:

$$p(\boldsymbol{d}|\mathcal{H}) = \int p(\boldsymbol{d}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})d\boldsymbol{\theta}. \tag{2.14}$$

This integration effectively marginalizes out the model parameters, providing a single number that represents the probability of observing the data

$\boldsymbol{d}$ under the hypothesis $\mathcal{H}$. The ratio between evidences of two competing hypotheses, known as the Bayes factor, can be used to quantitatively compare their relative plausibility, providing a principled approach to hypothesis testing.

To obtain the posterior distribution or the evidence, a likelihood function needs to be known. The likelihood function, $p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{H})$, quantifies the probability of observing the data $\boldsymbol{d}$ given the model parameters $\boldsymbol{\theta}$ and the hypothesis $\mathcal{H}$. In GW science, analysis often occurs in the frequency domain to easily incorporate the detector noise characteristics into the likelihood function:

$$p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{H}) \propto \exp\left(-\frac{\langle \boldsymbol{d} - h(\boldsymbol{\theta})|\boldsymbol{d} - h(\boldsymbol{\theta})\rangle}{2}\right), \tag{2.15}$$

where the inner product is defined as:

$$\langle a|b\rangle = 4\mathfrak{R}\int_0^\infty \frac{a^*(f)b(f)}{S_n(f)}df. \tag{2.16}$$

Here, $h(\boldsymbol{\theta})$ is the model waveform, $S_n(f)$ is the one-sided power spectral density (PSD) of the detector noise, and $\mathfrak{R}$ denotes the real part. This likelihood function measures the agreement between the observed data and the model predictions in the frequency domain, weighted by the noise characteristics of the detector, and is known as the Whittle likelihood [92].

For complex models such as those used in GW analysis, computing analytical solutions for the posterior distribution or evidence is generally intractable. Consequently, GW science relies on advanced sampling techniques to efficiently explore these vast parameter spaces and estimate posterior distributions and evidence. These sophisticated techniques, crucial for extracting physical information from GW signals, will be discussed in detail in the following subsection.

2.2.2 Classical Sampling Methods

Classical sampling methods are fundamental to Bayesian inference, enabling the estimation of posterior distributions and evidences for complicated, high-dimensional parameter spaces. While these methods can

handle both discrete and continuous variables, this subsection focuses on their application to continuous parameter spaces, which are typical in GW data analysis. This subsection examines four key techniques: rejection sampling, importance sampling, Markov Chain Monte Carlo (MCMC), and nested sampling. We will explore their principles, strengths, and limitations.

Rejection sampling is a simple method for drawing samples from a target probability distribution. In Bayesian inference, this target is typically the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{H})$. Importantly, we only need to know this distribution up to a normalization constant, allowing us to work directly with the product of the likelihood and prior. We indicate the unnormalized posterior distribution by $\hat{p}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{H})$. The algorithm proceeds as follows:

1. Pick a proposal distribution $q(\boldsymbol{\theta})$ that is non-zero wherever $\hat{p}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{H})$ is non-zero, and find a constant M such that $Mq(\boldsymbol{\theta}) \geq \hat{p}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{H})$ for all $\boldsymbol{\theta}$.

2. Draw a sample $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta})$.

3. Generate a uniform random number u from the interval $[0, 1]$.

4. Accept $\boldsymbol{\theta}'$ if $u \leq \frac{\hat{p}(\boldsymbol{\theta}'|\boldsymbol{d}, \mathcal{H})}{(Mq(\boldsymbol{\theta}'))}$; otherwise, reject it and return to step 2.

Figure 2.4A illustrates this process, showing how samples are accepted or rejected based on their position relative to the target distribution. Although conceptually straightforward, rejection sampling becomes highly inefficient in high-dimensional spaces. This is because the ratio of the volume of the target distribution to that of the proposal distribution typically shrinks exponentially with each added dimension, a phenomenon known as the curse of dimensionality. In GW parameter estimation, where we might be dealing with 15 parameters, the acceptance rate becomes prohibitively low, rendering it impractical and necessitating more sophisticated sampling techniques.

Importance sampling offers an improvement over rejection sampling by reducing sample waste. Instead of rejecting samples, it assigns weights to all samples drawn from a proposal distribution $q(\boldsymbol{\theta})$. The weight for a

sample $\boldsymbol{\theta}'$ is given by:

$$w(\boldsymbol{\theta}') = \frac{\hat{p}(\boldsymbol{\theta}'|\boldsymbol{d}, \mathcal{H})}{q(\boldsymbol{\theta}')}. \tag{2.17}$$

These weighted samples can be used to construct an approximation of the posterior distribution. Figure 2.4B depicts this concept, showing how samples are weighted according to their importance relative to the target distribution. The posterior probability of any region A in the parameter space can be estimated as:

$$P(\boldsymbol{\theta} \in A|\boldsymbol{d}, \mathcal{H}) \approx \frac{\sum_{i=1}^{N} w(\boldsymbol{\theta}_i)\mathbb{I}_A(\boldsymbol{\theta}_i)}{\sum_{i=1}^{N} w(\boldsymbol{\theta}_i)}, \tag{2.18}$$

where $\mathbb{I}_A(\boldsymbol{\theta})$ is the indicator function for region A. This allows us to approximate the entire posterior distribution from the weighted samples. The choice of proposal distribution $q(\boldsymbol{\theta})$ is crucial for the efficiency of importance sampling. Ideally, $q(\boldsymbol{\theta})$ should be as close as possible to the target distribution $\hat{p}(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{H})$. In practice, it is often beneficial to have a $q(\boldsymbol{\theta})$ that slightly overcovers the target distribution to ensure adequate sampling of the entire parameter space. A poor choice of $q(\boldsymbol{\theta})$ can lead to high variance in the weights, reducing the effective sample size[8] and compromising the accuracy of the posterior estimates. Despite its improvements over rejection sampling, importance sampling can still struggle in high-dimensional spaces. As the number of dimensions increases, it becomes increasingly challenging to design a proposal distribution that efficiently the target distribution, especially if it is a complicated target distribution.

MCMC methods can handle complicated posterior distributions much better. Like rejection sampling, MCMC involves proposing and accepting or rejecting samples, but it does so in a way that forms a Markov chain. A Markov chain is a sequence of samples, in which the probability of a sample is only dependent on its predecessor. It allows the algorithm to efficiently explore the parameter space by making local moves, gradually moving towards regions of high probability. Figure 2.4C provides a visual

[8]Effective sample size reflects the number of independent samples the weighted samples are equivalent to. A low effective sample size indicates that a few samples dominate the estimate, potentially leading to inaccurate results.

representation of an MCMC chain, demonstrating how the algorithm explores the parameter space through these local moves. This approach is more effective than independent sampling methods, like rejection sampling, because it can adapt to the shape of the target distribution, spending more time in high-probability regions while still occasionally exploring less likely areas. A well-known method to construct such a Markov chain is the Metropolis-Hastings algorithm [93]:

1. Start with an initial parameter value $\boldsymbol{\theta}_0$.
2. Propose a new value $\boldsymbol{\theta}'$ from a proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_0)$.
3. Calculate the acceptance ratio $\alpha = \min(1, \frac{\hat{p}(\boldsymbol{\theta}'|\boldsymbol{d},\mathcal{H})q(\boldsymbol{\theta}_0|\boldsymbol{\theta}')}{\hat{p}(\boldsymbol{\theta}_0|\boldsymbol{d},\mathcal{H})q(\boldsymbol{\theta}'|\boldsymbol{\theta}_0)})$.
4. Accept $\boldsymbol{\theta}'$ with probability α; if accepted, set $\boldsymbol{\theta}_1 = \boldsymbol{\theta}'$, otherwise set $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$.
5. Repeat steps 2-4, using the most recent accepted value as the starting point.

Provided that the Markov chain is ergodic[9], posterior samples can be extracted by discarding the first n samples, known as the "burn-in" period, and thinning the chain by selecting only every k-th sample to lessen autocorrelation and mimic independent sampling. It is common to run multiple independent chains to handle multimodality in the target distribution, as different chains might settle in different modes, guaranteeing a more exhaustive examination of the parameter space. While this description covers the basics of MCMC, many advanced MCMC methods have been developed to improve efficiency and handle complicated posterior distributions; for a comprehensive review see [94, 95].

MCMC methods, while excellent for sampling from the posterior distribution, face challenges when it comes to estimating the evidence. The evidence, as defined in equation 2.14, requires integration over the entire parameter space, which MCMC does not naturally perform. In complex parameter spaces, especially those with multi-modal or highly skewed

[9] A Markov chain is ergodic if it can explore the entire parameter space regardless of its starting point.

distributions, MCMC chains may struggle to efficiently explore all regions. If the chain is not sufficiently ergodic, it can become trapped in certain regions, oversampling them at the expense of others. This leads to a situation where sampling proportions within explored regions are accurate, but the relative exploration between different regions is poor due to infrequent transitions. Consequently, MCMC methods may miss significant contributions to the evidence integral from inadequately sampled regions, potentially leading to biased evidence estimates. Due to these difficulties in evidence estimation, nested sampling is the preferred method in most GW analyses.

Nested sampling [41], introduced by Skilling in 2004, offers a solution to the evidence estimation problem while simultaneously providing posterior samples. The algorithm operates by maintaining a set of N "live points" drawn from the prior distribution, progressively replacing the point with the lowest likelihood until the contribution to the evidence from unexplored regions becomes negligibly small. This process effectively samples from nested contours of increasing likelihood in the parameter space. Figure 2.4D illustrates this nested sampling process, showing how the algorithm progressively samples from these nested contours. The nested sampling algorithm proceeds as follows.

1. Initialize a set of N live points $\{\boldsymbol{\theta}_i\}$ drawn from the prior distribution $p(\boldsymbol{\theta}|\mathcal{H})$.

2. Initialize an empty set to store the dead points.

3. At each iteration j:

   (a) Identify the live point $\boldsymbol{\theta}_{low}$ with the lowest likelihood.

   (b) Add $\boldsymbol{\theta}_{low}$ to the set of dead points.

   (c) Draw a new point $\boldsymbol{\theta}_{new}$ from the prior distribution, subject to the constraint $p(\boldsymbol{d}|\boldsymbol{\theta}_{new}, \mathcal{H}) > p(\boldsymbol{d}|\boldsymbol{\theta}_{low}, \mathcal{H})$.

   (d) Replace $\boldsymbol{\theta}_{low}$ with $\boldsymbol{\theta}_{new}$ in the set of live points.

4. Repeat step 2 until a stopping criterion is met.

5. Add all remaining live points to the set of dead points.

While conceptually straightforward, the implementation of nested sampling faces a significant challenge in step 3(c): efficiently drawing samples from the constrained prior region. Simple rejection sampling in this step would suffer from the same poor scaling in high-dimensional spaces as discussed earlier. To address this issue, more sophisticated methods have been developed to sample from this constrained region efficiently. These include region-based sampling, slice sampling, constrained Hamiltonian Monte Carlo, and diffusive nested sampling [96–100]. These advanced sampling techniques are crucial for the practical implementation of nested sampling, especially in high-dimensional parameter spaces typical of GW analysis.

As the algorithm progresses, it generates a sequence of discarded points, often called "dead points", each associated with a likelihood value and an estimate of the prior volume it represents. These dead points are key to the algorithm's dual capability of evidence estimation and posterior sampling. The evidence can be estimated by summing the contributions from each dead point:

$$p(\boldsymbol{d}|\mathcal{H}) \approx \sum_j p(\boldsymbol{\theta}_j|d, H)\Delta X_j, \tag{2.19}$$

where $p(\boldsymbol{\theta}_j|d, H)$ is the likelihood of the dead point at iteration j, and ΔX_j is an estimate of the decrease in prior volume represented by that point. This decrease is estimated using the statistical properties of the sampling process. At each iteration, the remaining prior volume is expected to shrink by a factor related to the number of live points, allowing for a probabilistic estimate of ΔX_j. Specifically, $\Delta X_j \approx X_j(1 - \exp(-1/N))$ with X_j being the prior volume at iteration j and N the number of live points. Moreover, posterior samples can be extracted from the dead points by assigning each point a weight proportional to $p(\boldsymbol{\theta}_j|d, H)\Delta X_j$ and resampling. To enhance sampling efficiency and posterior resolution, dynamic nested sampling [101] can be implemented. This method adjusts the number of live points during the run, concentrating more samples in the posterior bulk. These capabilities – accurate evidence estimation, detailed posterior reconstruction, and adaptive sampling – make nested sampling a preferred choice for many GW analyses.

The sampling methods discussed in this section have been implemented

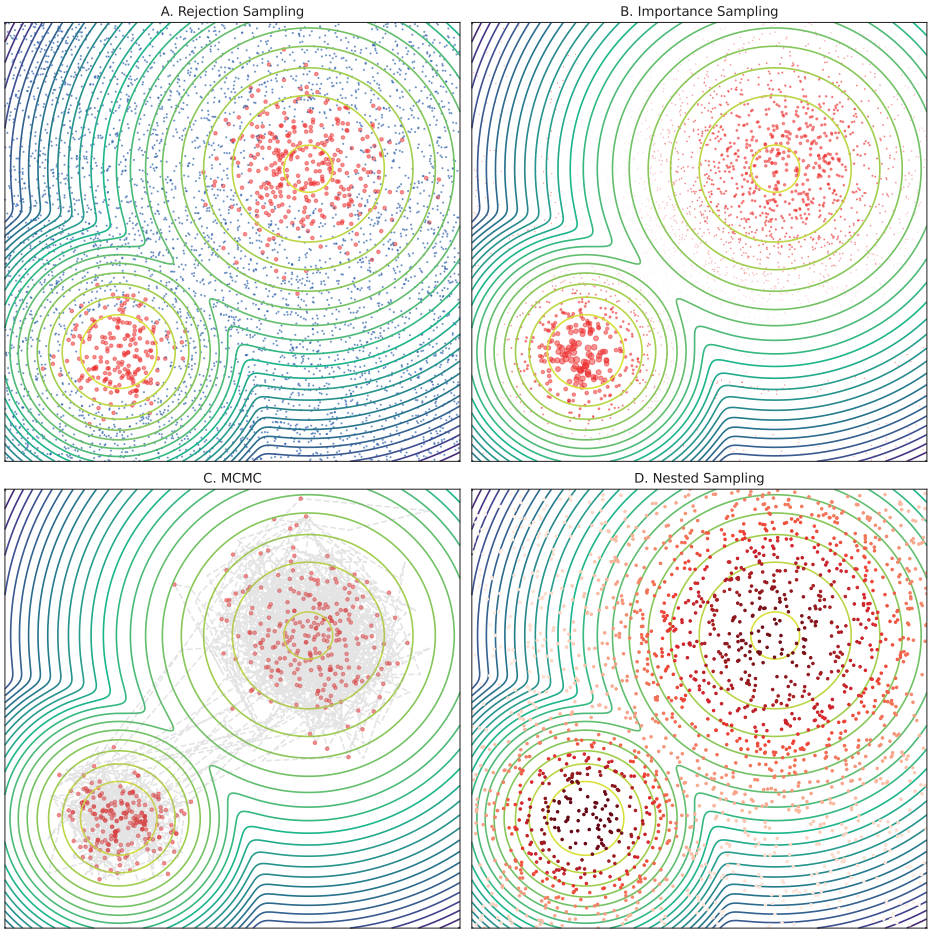


Figure 2.4: Comparison of sampling methods for parameter estimation. Each panel shows a different technique applied to a bimodal target distribution (contour lines). (A) Rejection sampling: red points accepted, blue rejected. The proposal distribution is a uniform distribution. (B) Importance sampling: point size represents importance weights. The proposal distribution is a standard normal distribution. (C) MCMC: the gray line shows the full path of the Markov chain exploring the parameter space, with every 10th sample highlighted in red to illustrate thinning. (D) Nested sampling: color gradient from light to dark red shows progression from prior exploration to high-likelihood concentration, efficiently estimating both posterior distribution and evidence.

in various software packages specifically designed for GW analysis. Two prominent examples are LALInference [102] and bilby [103]. LALInference has its own MCMC and nested sampling implementation and has been used extensively, for example in the analysis of GWTC-2 [104]. bilby uses off-the-shelf implementations of nested samplers, with the default implementation being dynesty [105]. bilby was one of the analysis pipelines used for the analysis of GWTC-3 [29] and is often used in analysis studies such as [106–108].

These classical sampling methods have been invaluable in GW analysis, providing robust tools for parameter estimation and model comparison. However, they all share a common limitation: they require starting from the prior distribution and evaluating millions of likelihood samples to accurately infer the posterior distribution and evidence for each new observation. This process is computationally intensive and time-consuming, especially for complex waveform models. As the detection rate of GW events is expected to increase dramatically with future detectors like Cosmic Explorer and Einstein Telescope, the computational demands of these methods may become prohibitive. These limitations motivate the exploration of alternative approaches, such as the neural methods we will discuss in the next section, which aims to provide rapid parameter estimation while maintaining the accuracy of traditional Bayesian techniques.

2.2.3 Neural Density Estimators

Recent advances in deep learning have enabled novel approaches to parameter estimation. Neural density estimators approximate the posterior distribution with Neural Networks (NNs), offering the potential for rapid inference once trained. This subsection assumes a basic understanding of NNs; readers seeking a comprehensive overview or wanting to update their knowledge of NNs are directed to [60, 109].

One straightforward approach to neural density estimation is to use a NN to parameterize a known probability distribution [110, 111]. Consider a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ as the chosen distribution, where both the mean vector μ and covariance matrix Σ are estimated by a NN. For a given observation $\boldsymbol{x}$, a neural density estimator $q_\phi(\boldsymbol{\theta}|\boldsymbol{x})$ with parameters

ϕ is then defined as:

$$q_\phi(\boldsymbol{\theta}|\boldsymbol{x}) = \mathcal{N}([f(\boldsymbol{x}, \phi)]_0, [f(\boldsymbol{x}, \phi)]_1) \tag{2.20}$$

Here, $f(\cdot, \phi)$ is a NN that outputs the parameters of the normal distribution. The first output, $[f(\boldsymbol{x}, \phi)]_0$, represents μ and the second output, $[f(\boldsymbol{x}, \phi)]_1$, represents Σ. To obtain appropriate values for μ and Σ, ϕ needs to be optimized. Training typically involves maximizing the expected log-likelihood of the true parameters given the data:

$$\mathcal{L}_{\text{ELL}}(\phi) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{\theta})\sim\mathcal{D}}[\log q_\phi(\boldsymbol{\theta}|\boldsymbol{x})], \tag{2.21}$$

where $\mathcal{D}$ represents a dataset of observations or simulations $\boldsymbol{x}$ and their corresponding parameters $\boldsymbol{\theta}$. This approach can be viewed as a form of amortized Bayesian inference, where the NN learns to directly approximate the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$. By optimizing ϕ, the model captures the relationship between observations and parameters across a wide range of scenarios. Once trained, this neural density estimator can efficiently generate posterior estimates for new observations, circumventing the need for costly per-instance inference runs typically associated with traditional Bayesian methods.

While the approach of directly parameterizing a distribution is straightforward, it may not always capture complicated, multi-modal distributions effectively. To address this limitation, more sophisticated models have been developed, notably Normalizing Flows (NFs) [112–115]. An NF model consists of a sequence of invertible transformations that map a simple base distribution to a target distribution. This construction establishes a bijective relationship between the distributions, preserving total probability mass while allowing individual sample probabilities to change. A significant feature of NFs is their potential for universal approximation: under certain conditions, such as using sufficiently expressive bijective functions (e.g., splines or polynomials [116, 117]) and enough transformations, NFs can theoretically transform any continuous probability distribution into any other continuous distribution [118]. This universality, coupled with the bijective nature of the transformations, allows NFs to model complex, multi-modal distributions effectively. The subsequent discussion will

examine the mathematical foundations, implementation strategies, and practical considerations of NFs.

To understand how NFs achieve the necessary flexibility, we need to examine the change of variables rule. Let T be an invertible transformation such that $x = T(z)$, where z is drawn from a simple base distribution $\pi(z)$ and x represents a sample from our target distribution $p(x)$. The change of variables rule allows us to compute the density of the target distribution given the density of the base distribution and this transformation:

$$p(\boldsymbol{x}) = \pi(\boldsymbol{z}) \left| \det \left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} \right) \right| = \pi(T^{-1}(\boldsymbol{x})) \left| \det \left(J_{T^{-1}}(\boldsymbol{x}) \right) \right| \tag{2.22}$$

where $J_{T^{-1}}(x)$ is the Jacobian matrix of T^{-1} evaluated at x. For computational efficiency and stability, the log-probability form is often preferred:

$$\log p(\boldsymbol{x}) = \log \pi(T^{-1}(\boldsymbol{x})) + \log \left| \det \left(J_{T^{-1}}(\boldsymbol{x}) \right) \right| \tag{2.23}$$

These equations relate the density $p(x)$ of the target distribution to the density $\pi(z)$ of the base distribution through the differentiable and invertible transformation T. The Jacobian determinant term accounts for the volume change induced by the transformation, ensuring the resulting distribution is properly normalized.

NFs enhance their flexibility by chaining multiple invertible transformations, see for example Figure 2.5. Given a sequence of K invertible transformations $T_1, T_2, ..., T_K$, the complete transformation T is their composition:

$$T = T_K \circ T_{K-1} \circ ... \circ T_2 \circ T_1 \tag{2.24}$$

This composition enables increasingly complex mappings between the base and target distributions. The change-of-variable formula extends to this composed transformation, with the Jacobian determinant becoming a product:

$$\left| \det \left(J_T(\boldsymbol{x}) \right) \right| = \prod_{k=1}^{K} \left| \det \left(J_{T_k}(T_{k-1} \circ ... \circ T_1(\boldsymbol{x})) \right) \right| \tag{2.25}$$

The primary computational challenge in implementing NFs lies in the calculation of the Jacobian determinant. For high-dimensional problems, this calculation can become costly. This computational burden arises

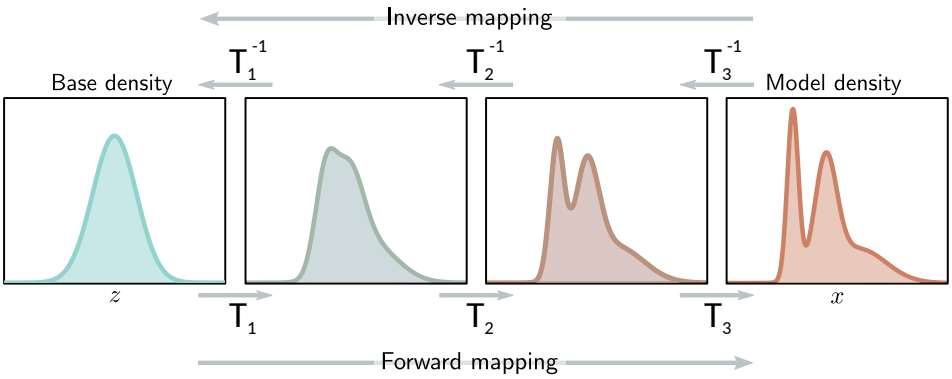


Figure 2.5: Illustration of the sequence of invertible transformations in an NF model. Each layer applies a transformation to the input, incrementally mapping a simple base distribution to a complex target distribution in the forward direction, and vice versa in the inverse direction. Figure adapted from [109].

because a naive implementation would require $\mathcal{O}(d^3)$ operations for a d-dimensional parameter space, making it impractical for many real-world applications. The need to compute this determinant for every transformation in the NF model in forward and backward passes further exacerbates the issue, leading to prohibitively long training times and high memory requirements.

The key to overcoming the computational limitations of NFs lies in architectural designs that force the Jacobian matrix to be triangular. A triangular Jacobian matrix is computationally advantageous because its determinant is the product of its diagonal elements, calculable in $\mathcal{O}(d)$ time instead of $\mathcal{O}(d^3)$. The triangular structure is achieved by designing transformations where each output component depends on a specific subset of input components, creating a form of ordered dependence. Despite this constraint, these architectures can represent complex transformations. The two approaches that implement this triangular Jacobian matrix strategy are autoregressive models and coupling layers, both of which have made NFs practical for a wide range of high-dimensional problems [115, 119, 120].

Autoregressive layers achieve a triangular Jacobian matrix by imposing

an ordered dependency in the transformation [114, 115, 121]. The i-th dimension of the output is computed with only the first i dimensions of the input:

$$z_i = g_i(y_i), \quad \text{with} \quad y_i = f_i(\boldsymbol{x}_{1:i}, \phi) \tag{2.26}$$

where g_i is a differentiable, bijective function parameterized by the output of a NN f_i. This structure results in a lower triangular Jacobian matrix, as $\frac{\partial z_i}{\partial x_j} = 0$ for $j > i$. While this approach allows for highly flexible transformations, it introduces an asymmetry between forward and inverse computations. The forward pass can be parallelized, but the inverse transformation (crucial for sampling) requires d sequential steps for d-dimensional data[10]. This trade-off between expressiveness and sampling speed is a key consideration when using autoregressive layers in practice.

Enforcing the ordered dependency in autoregressive layers presents a challenge in implementation. A popular solution to this challenge is the use of masking techniques in the NN f_i, as introduced by the Masked Autoencoder for Distribution Estimation (MADE) model [122]. In MADE, each layer of the NN is equipped with a specific binary mask that controls the connections between its units and those of the previous layer. These masks are designed to ensure that the autoregressive property is maintained throughout the network. Specifically, the mask for output unit y_i allows connections only from input units x_1 to x_i. This approach not only preserves the desired dependency structure but also enhances computational efficiency, making autoregressive models feasible for high-dimensional data.

In contrast to autoregressive layers, coupling layers provide computational symmetry in the forward and backward pass[11], allowing for efficient parallel processing in both directions [113]. This is achieved by partitioning the input $\boldsymbol{x}$ into $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and applying the following transformations:

$$\boldsymbol{z}_1 = \boldsymbol{x}_1, \tag{2.27}$$

$$\boldsymbol{z}_2 = g(\boldsymbol{x}_2; y), \quad \text{with} \quad y = f(\boldsymbol{x}_1, \phi). \tag{2.28}$$

[10] One can also opt for a sequential forward pass and a parallelized inverse transformation, as done in [114].

[11] Technically, coupling layers are also autoregressive but have only two partitions, making them a special case with a simpler dependency structure. This simplification allows for parallel computation in both directions.

Here, g is a differentiable, bijective function parameterized by the output of a NN f. This structure guarantees a block triangular Jacobian matrix:

$$J = \begin{bmatrix} I & 0 \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} \end{bmatrix}. \tag{2.29}$$

The Jacobian determinant is simply $|\det(\partial z_2/\partial \boldsymbol{x}_2)|$. Coupling layers face an inherent trade-off between expressiveness and computational efficiency. Their partitioning strategy, while enabling fast bidirectional transformations, restricts the model's capacity to capture complex dependencies in a single layer. As a result, coupling layer architectures typically need to stack more layers to achieve comparable modeling power. Nevertheless, their symmetric computational properties often prove advantageous in practice, especially when both generation and density estimation tasks are of equal importance.

Coupling and autoregressive layers have addressed the challenge of tractable Jacobian determinant computation by imposing specific dependency constraints. The introduction of Neural Ordinary Differential Equations (Neural ODEs) [123] in 2018, opened up new possibilities for NFs. It makes Continuous Normalizing Flows (CNFs), also known as infinitesimal flows, possible. Figure 2.6 visualizes the continuous transformation in CNFs. These flows handle the Jacobian determinant computation in a fundamentally different way: instead of designing architectures to ensure a tractable Jacobian determinant for discrete transformations, CNFs reformulate the problem in continuous time. Here, 'time' refers to a fictitious dimension along which the transformation evolves, not physical time. This allows the log-density change to be computed using the trace of the Jacobian matrix, which can be efficiently estimated without explicitly constructing the full Jacobian matrix. Unlike discrete normalizing flows, CNFs allow all parameters to depend on all other parameters. This increased flexibility stems from the continuous-time formulation where only the trace of the Jacobian matters. In the limit of infinitesimal steps, off-diagonal Jacobian elements do not contribute to the Jacobian determinant, allowing for more complex interdependencies in the transformation function. As a result, CNFs can use more flexible architectures for their transformations while still maintaining computational feasibility.

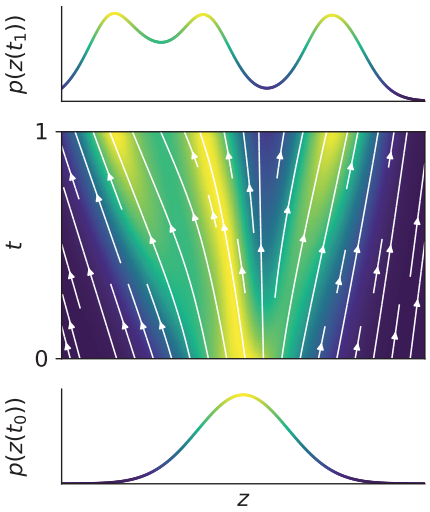


Figure 2.6: Visualization of a CNF transforming a simple distribution into a more complex one. **Bottom**: Initial 1D Gaussian distribution $p(z(t_0))$ along coordinate z. **Middle**: Transformation process of the CNF, depicting the flow direction and density changes along the continuous dimension t. **Top**: Final learned density $p(z(t_1))$ after the CNF transformation. Figure adapted from [124].

Neural ODEs extend the concept of NNs to continuous-time dynamics. Unlike traditional NNs, which apply a fixed number of discrete, layer-by-layer transformations, Neural ODEs define a continuous transformation of their inputs. To model this continuous transformation, Neural ODEs introduce the concept of a hidden state h(t). This hidden state relates to our previous notation, but with important distinctions. In the context of normalizing flows, we can think of $h(t_0)$ as being initialized with a sample z from the base distribution, and $h(t_1)$ as representing the transformed sample x in the target distribution. The evolution of this hidden state is achieved by parameterizing its derivative with respect to time using a NN f_ϕ:

$$\frac{d\boldsymbol{h}(t)}{dt} = f_\phi(\boldsymbol{h}(t), t), \tag{2.30}$$

where $\boldsymbol{h}(t)$ represents the hidden state at time t. The evolution of this system from an initial state $\boldsymbol{h}(t_0)$ to a final state $\boldsymbol{h}(t_1)$ is then computed

using an ODE solver:

$$\boldsymbol{h}(t_1) = \boldsymbol{h}(t_0) + \int_{t_0}^{t_1} f_\phi(\boldsymbol{h}(t), t)dt. \tag{2.31}$$

This formulation allows for adaptive computation: the ODE solver can automatically adjust its step size to maintain a desired level of accuracy. Moreover, the use of ODE solvers extends to the backward pass, where the adjoint sensitivity method enables efficient gradient computation without storing intermediate activations [123, 125]. This approach not only provides a continuous analogue to residual networks[12] but also opens up new possibilities for modeling dynamical systems and, as we will see, for constructing flexible NFs.

The evolution of the log-probability density in a CNF can be understood by considering an infinitesimal step in the flow. The instantaneous change in the log-probability density is given by:

$$\frac{\partial \log p(\boldsymbol{z}(t))}{\partial t} = -\mathrm{Tr}\left(\frac{\partial f}{\partial \boldsymbol{z}(t)}\right). \tag{2.32}$$

This equation describes how the log-density changes during this infinitesimal step. The trace term represents the instantaneous relative change in volume of an infinitesimal region around $\boldsymbol{z}(t)$. As the ODE solver integrates the trajectory from t_0 to t_1, it accumulates these infinitesimal changes in log-density. The total change in log-density over the entire transformation is thus given by:

$$\log p(\boldsymbol{z}(t_1)) - \log p(\boldsymbol{z}(t_0)) = -\int_{t_0}^{t_1} \mathrm{Tr}\left(\frac{\partial f}{\partial \boldsymbol{z}}(\boldsymbol{z}(t), t)\right) dt. \tag{2.33}$$

This formulation allows us to compute the change in log-density along the entire flow without explicitly constructing large Jacobian matrices at discrete steps, as is typically done in traditional NFs. A notable implementation of CNFs is the Free-Form Jacobian of Reversible Dynamics (FFJORD)

[12]Residual networks are deep neural architectures that use skip connections, allowing for easier training of very deep models. These skip connections $h_{t+1} = h_t + f_\phi(h_t)$ mirror the continuous trajectory of the hidden state in Neural ODEs, where the evolution is governed by a differential equation rather than discrete layers.

model [124], which leverages these concepts to create highly flexible and scalable NFs.

Having explored these advanced neural density estimation techniques, we can now appreciate their potential for parameter estimation. These methods offer the flexibility to model the complicated posterior distributions. By learning to directly approximate the posterior $p(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{H})$, these models can potentially provide rapid parameter estimates for new GW observations, addressing the computational challenges posed by the increasing detection rates expected from future detectors. However, it is important to note that while these methods offer promising speed advantages, their accuracy and reliability in the context of GW analysis remain active areas of research.

Chapter 3

Fast Sky Localization of Gravitational Waves using Deep Learning Seeded Importance Sampling

Fast, highly accurate, and reliable inference of the sky origin of gravitational waves would enable real-time multi-messenger astronomy. Current Bayesian inference methodologies, although highly accurate and reliable, are slow. Deep learning models have shown themselves to be accurate and extremely fast for inference tasks on gravitational waves, but their output is inherently questionable due to the black-box nature of neural networks. In this work, we merge Bayesian inference and deep learning by applying importance sampling on an approximate posterior generated by a multi-headed convolutional neural network. The neural network parametrizes Von Mises-Fisher and Gaussian distributions for the sky coordinates and two masses for given simulated gravitational wave injections in the LIGO and Virgo detectors. We generate skymaps for unseen gravitational-wave events that highly resemble predictions generated using Bayesian inference in a few minutes. Furthermore, we can detect poor predictions from the neural network, and quickly flag them.

Based on: Kolmus, A., Baltus, G., Janquart, J., van Laarhoven, T., Caudill, S., Heskes, T., "Fast sky localization of gravitational waves using deep learning seeded importance sampling". Physical Review D **106**, 023032 (2022).

3.1 Introduction

Gravitational waves (GWs) have immensely advanced our understanding of physics and astronomy since 2015 [126–129]. These GWs are observed by the Hanford (H) and Livingston (L) interferometers of the Laser Interferometer Gravitational Wave Observatory (LIGO) [130] and the Advanced Virgo (V) interferometer [27]. The collaboration between these three detectors has enabled triple-detector observations of GWs [127], making it possible to do proper sky localization of their astrophysical sources. This additional detector changes the sky distribution from a broad band to a more narrow distribution [127].

Better early sky localization capabilities would allow for real-time multi-messenger astronomy (MMA), observing astrophysical events via multiple channels – electromagnetic transients, cosmic rays, neutrinos – only seconds after the GW is detected. MMA is limited to GWs originating from binary neutron star (BNS) and neutron star-black hole mergers. According to current literature, it is unlikely that binary black holes (BBHs) emit an electromagnetic counterpart during their merger [131, 132]. Currently, astrophysicists try to collect the non-GW channels in the weeks after the event. A notable example is GW170817 [30, 133]. This process takes an enormous amount of effort, while the obtained data quality is often sub-optimal. Having all channels observed for the full duration of the event would be a major leap forward. Real-time MMA would enable a plethora of new science, e.g. unraveling the nucleosynthesis of heavy elements using r- and s-processes, more accurate and novel tests of general relativity, and a deeper understanding of the cosmological evolution [134–136]. As aforementioned, real-time MMA relies on the generation of a skymap and it imposes two limits on the methodology used to obtain one. First, it needs to be swift to allow observatories to turn towards an event's origin, preferably only seconds after its observation. Second, the skymap needs to be as accurate as possible since telescopes have a limited area they can observe. Below we present current approaches for generating sky maps for GW events.

Most GW software libraries [102, 137] use Bayesian inference methods – Markov chain Monte Carlo (MCMC) and nested sampling [41] – to con-

struct the posterior over all GW parameters. These methods asymptotically approach the true distribution given a sufficient number of samples [138]. Although theoretically optimal, a chain with around 10^6 to 10^8 samples is required [102] to closely approximate the true posterior distribution for a GW event. Even when using Bilby [103] – a modern Bayesian inference library made for GW astronomy – to perform the inference for a single BBH event, takes hours to produce [139]; BNS events take even longer. Bayesian inference is the most accurate method available for GW posterior estimation, but its run-time is prohibitively long when it comes to MMA.

To overcome the speed limitations of the Bayesian approaches, Singer and Price developed BAYESTAR in 2016 [140], an algorithm that can output a robust skymap for a GW event within a minute. BAYESTAR realizes this speedup in two ways. First, it exploits the information provided by the matched filtering pipeline used in the detection of GWs. The inner product between time strain and matched filters contains nearly all of the information regarding arrival times, amplitudes, and phases, which are critical for skymap estimation. Second, Singer and Price derive a likelihood function that is semi-independent from the mass estimation and does not rely on direct computation of GW waveforms, allowing for massive speedups and parallelization. Although BAYESTAR is fast, its predictions tend to be broader and less precise than those made by Bilby [1].

Deep learning (DL) algorithms have shown themselves to be exceptionally quick and powerful when handling high-dimensional data [141, 142]. Therefore, they are an interesting alternative to the Bayesian methods. Several papers have proposed methods to estimate the GW posterior, including the skymap, using DL algorithms. Examples of such algorithms are Delaunoy et al. [143] and Green and Gair [144]. Delaunoy et al. [143] use a convolutional neural network (CNN) to model the likelihood-to-evidence ratio when given a strain-parameter pair. By evaluating a large amount of parameter options in parallel, they can generate confidence intervals within a minute. The reported confidence intervals are slightly wider than those made by Bilby. A completely different approach was

[1]The GWTC-2 catalog [104] data release provides skymaps made using Bayesian inference methods for recent events. Comparison with the skymaps made by BAYESTAR can be made by looking at skymaps on https://gracedb.ligo.org/latest/.

taken by Green and Gair [144]. They showcase a complete 15-parameter posterior estimate for GW150914 using normalizing flows. They apply a sequence of invertible functions to transform an elementary distribution into a complex distribution [145] which, in this case, is a BBH posterior. Within a single second, their method can generate 5,000 independent posterior samples that are in agreement with the reference posterior[2]. A Kolmogorov-Smirnov test confirms that these samples very closely resemble the samples that are drawn from the exact posterior. Both DL methods are fast and seem to be accurate for the 100 - 1000 simulated GW events they have been evaluated on. However, these methods have a few issues: (1) they are both susceptible to changes in the power spectral density (PSD) and signal-to-noise ratio (SNR), (2) both are close in performance to Bilby but do not match it, (3) they can act unpredictably outside of the trained strain-parameters pairs and, even within this space, they can act unpredictably due to the black box nature of neural networks (NNs). Issues (1) and (2) have been addressed for the normalizing flow algorithm in a recent paper by Dax et al. [57], however, the robustness guarantees remain behind those of traditional Bayesian inference.

Our method tries to bridge the gap between Bayesian inference and DL methods, allowing for fast inference while still guaranteeing optimal accuracy. It is to be noted that combining Bayesian inference and DL methods has recently gained traction in the GW community, see for example reference [146]. The goal of our algorithm is to restrict the parameter space such that, via sampling, one can quickly obtain an accurate sky map. We use a multi-headed CNN to parameterize an independent sky and mass distribution for a given BBH event. The model is trained on simulated precessing quasi-circular BBH signals resembling the ones observed by the HLV detectors. The parameterized sky and mass distributions are Gaussian-like and are assumed to approximate the sky and mass distributions generated by Bayesian inference. Using the parameterized sky and mass distributions, we construct a proposal posterior in which all other BBH parameters are uniformly distributed. By using importance sampling we can then sample from the exact reference posterior. This implies that

[2]Throughout this chapter, reference posterior is used to imply a posterior that is generated using Bayesian inference.

we effectively match the performance of Bayesian inference in a short time span, without exploring the entire parameter space. We stress that this work is a proof of concept to show the promises of combining NNs and Bayesian inference. More flexible DL models and BNS events will be considered in future studies.

This chapter is organized as follows. Section 2 discusses the model architecture and the importance sampling scheme. Section 3 details the performed experiments, including the model training. Section 4 covers the results of these experiments and subsequently assesses the performance of the model and importance sampling scheme by comparing it with skymaps generated using Bilby for a non-spinning BBH system. Conclusions and future endeavors are specified in Section 5.

3.2 Methodology

Our inference setup is a two-step method. In the initial step, we infer simple distributions for the sky localization and the masses of the BBH by using a neural network. Subsequently, we apply importance sampling to these simple distributions to compute a more accurate posterior. The first subsection describes the role and implementation of importance sampling. The second subsection discusses the neural network setup and our method for distribution estimation.

3.2.1 Importance sampling

High-dimensional distributions in which the majority of the probability density is confined to a small volume of the space are hard to sample from, which results in long run times to get proper estimates when using MCMC methods. A well-known method to cope with this problem is importance sampling. By using a proposal distribution q that covers this high probability density region of the complex distribution p one can quickly obtain useful samples. There are two requirements when using importance sampling. First, the desired distribution p needs to be known up to the normalization constant Z: $p(\lambda) = \frac{1}{Z}\theta(\lambda)$, where $\theta(\lambda)$ is the non-normalized $p(\lambda)$. Second, the proposal distribution q needs to be non-zero

for all λ where p is non-zero. Importance sampling can be understood as compensating for the difference between the distributions p and q by assigning an importance weight $w(\lambda)$ to each sample λ,

$$w(\lambda) = \frac{\theta(\lambda)}{q(\lambda)}, \tag{3.1}$$

where the fraction is the likelihood ratio between the – not-normalized – p and q. The distribution created by the reweighted samples will converge to the p distribution given enough samples [147].

Generating accurate posteriors for GW observations using MCMC is very time-consuming, and thus importance sampling is an interesting alternative. Importance sampling requires us to have a viable proposal distribution. Published posteriors for known gravitational waves show that the probability density in the posterior is relatively well confined for both the sky location and the two masses [104]. A Von Mises Fisher (VMF) and Multi-Variate Gaussian (MVG) distribution are good first-order approximations of the sky and mass distribution respectively, and thus suitable to use as a proposal distribution for importance sampling. We propose to construct this proposal distribution by assuming a uniform distribution over all non-spinning BBH parameters, except for the sky angles which will be represented by a VMF and a MVG distribution for the masses. Assuming that the BBH parameters, sky angles, and masses are independent, our proposal distribution becomes the product of these two distributions. In the next subsection, we discuss how we create this proposal distribution using a neural network.

Importance sampling demands a likelihood function for the proposal distribution and the desired distribution. In the previous paragraph we have discussed how we want to create a proposal distribution, we will now focus on the desired distribution p. For the likelihood function of the GW posterior $p(s|\lambda)$ we take the definition given by Canizares et al. [148]:

$$p(s|\lambda) \propto \theta(s|\lambda) = \exp\left(-\frac{\langle s - h(\lambda)|s - h(\lambda)\rangle}{2}\right), \tag{3.2}$$

where s is the observed strain, $h(\lambda)$ is the GW template defined by parameters λ. The inner product is weighted by the PSD of the detector's noise. In

practice, we use the likelihood implementation provided by Bilby named *GravitationalWaveTransient*.

We now have all the parts needed to discuss how we utilize importance sampling for a given strain s. A trained neural network parameterizes the proposal distribution q for the given strain. The proposal distribution generates n samples, these samples represent possible GW parameter configurations. For each sample, we calculate the logarithm of the importance weight,

$$\log w(\lambda) = \log \theta(s|\lambda) - \log q(\lambda) + C, \tag{3.3}$$

instead of the importance weight $w(\lambda)$ itself to prevent numeric under- and overflow. The constant C is added to set the highest $\log w(\lambda)$ to zero, to prevent very large negative values from becoming zero when we calculate the associated likelihood. Since we normalize the weights afterwards the correct importance weights are still obtained. The reweighted samples represent the desired distribution p.

If the proposal distribution does not cover the true distribution well enough, the importance samples will be dominated by only a single to a few weights if we restrict the run-time. We can use this as a gauge to check if the skymap produced by the neural network and importance sampling is to be trusted.

3.2.2 Model

Previous work done by George et al. [149] shows that convolutional neural networks (CNN) can extract the masses from a BBH event just as well as the currently-in-use matched filtering. Furthermore, work done by Fan et al. [150] indicates that 1D CNNs can locate GW origins. We therefore chose to use a 1D CNN to model both the distribution across the sky for the origin of the GWs and a multivariate normal distribution for the two masses of the BBH system.

The network architecture of this 1D CNN is presented in Figure 3.1 and consists of four parts: a convolutional feature extractor and three neural network heads. These heads are used to specify the two distributions. The following properties were tested or tuned for optimal performance: number of convolutional layers, kernel size, dilation, batch normalization,

and dropout. The model shown in Figure 3.1 produced the best result on a validation set.

The convolutional feature extractor generates a set of features that characterize a given GW. This set of features is passed on to the neural heads. Each head is specialized to model a specific GW parameter. The first head determines the sky distribution, the second head the masses, and the third head the uncertainty over the two masses. Below we will elaborate on each of these heads and how they characterize these distributions.

The first head specifies the distribution of the GW origin. Since the sky is described by the surface of a 3D sphere, a 2D Gaussian distribution is an ill fit. A suitable alternative is the Von Mises-Fisher (VMF) distribution [151] which is the equivalent of a Gaussian distribution on the surface of a sphere. The probability density function and the associated negative log-likelihood (NLL) of the VMF distribution:

$$p(x|\mu,\kappa) = \frac{\kappa}{4\pi \sinh(\kappa)} \exp\left(\kappa x^T \mu\right) \tag{3.4}$$

$$NLL_{\text{VMF}}(x,\mu,\kappa) = -\log(\kappa) - \log(1-\exp(-2\kappa)) - \kappa - \log(2\pi) + \kappa x^T \mu, \tag{3.5}$$

where x and μ are normalized vectors in $\mathbb{R}^3$, with the former being the true direction and the latter being the predicted direction. κ is the concentration parameter, which determines the width of the distribution. It plays the same role as the inverse of the variance for a Gaussian distribution. We use this distribution by letting the first head output a three-dimensional vector $D = (D_x, D_y, D_z)$. The norm of D specifies the concentration parameter κ, and its projection onto the unit sphere gives the mean μ, $\kappa = |D|$, and $\mu = D/|D|$. These values together with the true direction x are used to calculate the negative log-likelihood, which is used as the loss function of the first head.

The second and third neural heads specify a 2D multivariate Gaussian (MVG), which describes the possible configurations of the masses. The means ν of the MVG are given by the second head and the covariance matrix Σ is specified by the third head. Given the true values of the masses $y = (m_1, m_2)$ the probability density function and associated negative

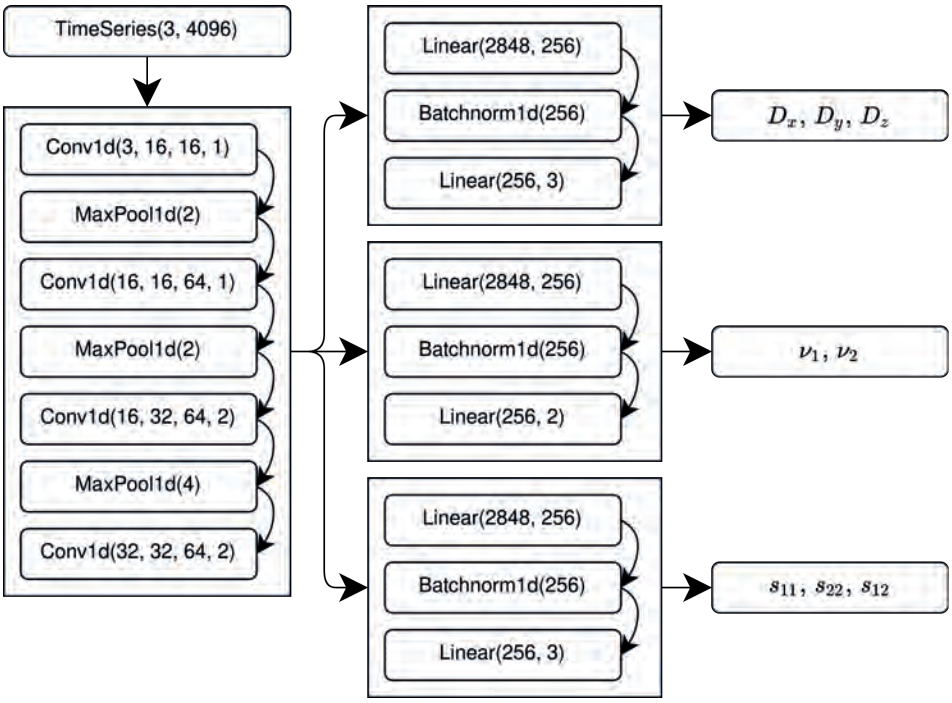


Figure 3.1: A graphical depiction of the convolutional neural network used in this work. After each MaxPool1d and Batchnorm1d layer, a leaky ReLU activation function with an $\alpha = 0.1$ is applied. The convolutional part is shown on the left and takes as input a time series of 4096 elements with 3 channels. Conv1D(i, o, k , d) denotes a 1D convolution with i input channels, o output channels, kernel size k and dilation factor d. MaxPool1d(k) denotes a 1D max pooling layer with kernel size k. The output of the convolutions is given to three independent neural network heads. The first head predicts the sky location parameterized as $D = (D_x, D_y, D_z)$, the second head predicts the mean of the masses of the two black holes, and the last head predicts the uncertainty elements of the covariance matrix over the two masses. Linear(i, o) denotes a linear transformation with i input features and o output features. Lastly, Batchnorm1d(i) denotes a 1D batch normalization layer with i input features.

log-likelihood of the MVG are:

$$p(y|\nu,\Sigma) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left(-\frac{1}{2}(y-\nu)^T\Sigma(y-\nu)\right) \tag{3.6}$$

$$NLL_{\text{MVG}}(y,\nu,\Sigma) = \frac{1}{2}(y-\nu)^T\Sigma^{-1}(y-\nu) + \frac{1}{2}\log\left(|\Sigma|\right) + \log\left(2\pi\right). \tag{3.7}$$

The inverse covariance term in the negative log-likelihood can contain imaginary numbers if the covariance matrix is not positive-definite. To ensure that the covariance matrix Σ remains positive-definite, it is parameterized through:

$$\Sigma_{11} = \exp(s_{11}) \tag{3.8}$$

$$\Sigma_{22} = \exp(s_{22}) \tag{3.9}$$

$$\Sigma_{21} = \Sigma_{12} = \tanh(s_{12})\sqrt{\Sigma_{11}\Sigma_{22}}\,. \tag{3.10}$$

The three variables s_{11}, s_{22}, s_{12} are predicted by the third neural head and define the covariance matrix completing the MVG prediction of the masses. The parametrization and implementation of the MVG are based on the work of Russell et al. [110].

By further assuming that the sky distribution is independent of the mass distribution, we obtain a first approximation of the posterior distribution, thereby satisfying the requirements for importance sampling.

3.3 Experiments

Experiments were performed on two different fronts: (1) training the neural network followed by the empirical evaluation of its performances on unseen test data, and (2) comparing the neural network model, importance sampling scheme, and Bilby based on several metrics and skymaps. Below we describe the experimental details and justify the decisions we made. All experiments were performed on a computer with a 16-core AMD Ryzen 5950X CPU, NVIDIA 3090 RTX GPU, and 64 GB of RAM.

3.3.1 Training and evaluating the neural model

To obtain strain-parameter pairs for training and validation, we sampled parameters from a BBH parameter prior (see Table 3.1) and generated the associated waveforms using the *IMRPhenomPv2* waveform model [152]. The waveforms were generated in the frequency domain in the frequency band of 20 to 2048 Hz. The duration of the signal is 2 seconds. Subsequently, these waveforms were projected onto the HLV interferometers. We sampled the SNR from a scaled and shifted Beta distribution with its peak set to 15 (see Figure 3.2). The luminosity distance in the prior was set to 1000 Mpc and scaled afterward to match the desired SNR. We generated Gaussian noise from the design sensitivity PSD for each detector. Finally, the signal was injected into the noise and an inverse Fourier transform was applied to obtain the strains as time series. This setup allowed us to generate an arbitrary amount of unique strain-parameter pairs, which resulted in every training epoch having a unique dataset.

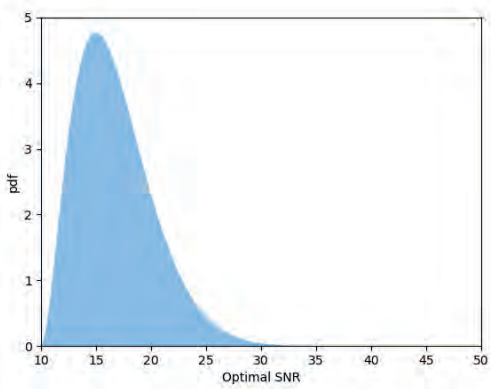


Figure 3.2: Scaled and shifted Beta distribution that acts as the SNR sampling distribution during training and validation. The vertical axis represents the probability density function of this Beta distribution, the horizontal axis represents the SNR value.

We applied three preprocessing steps to the data. All time series were whitened with the aforementioned PSDs. Next, the time series were normalized. A normalizer was calculated such that noise-only strains have

Table 3.1: The priors used for the data generation. The luminosity distance in the prior was set to a 1000 Mpc and scaled afterwards to match the desired SNR.

Parameter	Prior	Minimum	Maximum	Unit
Masses (constraint)	-	20	80	$M_\odot$
Chirp mass	Uniform	10	100	$M_\odot$
Mass ratio	Uniform	0.25	1.0	-
Spin magnitudes	Uniform	0	0.95	-
Spin polar angles	Sine	0	π	rad
Spin azimutal angles	Uniform	0	2π	rad
Right ascension	Uniform	0	2π	rad
Declination	Cosine	-0.5π	0.5π	rad
Binary inclination angle	Sine	0	π	rad
Coalescence phase angle	Uniform	0	2π	rad
Polarization angle	Uniform	0	π	rad
Time Shift	Uniform	-0.1	0.1	s
Luminosity distance	-	1000	1000	Mpc

a mean of zero and a standard deviation of one. We found empirically that calculating a normalizer for the noise instead of noise plus signal allowed the neural network to converge faster and achieve lower losses. Lastly, to make the mass distribution easier to learn we calculated a shift and scaling factor for the target masses such that all target masses were between -1 and +1. The shifting and scaling were applied inversely to the neural network output during importance sampling to get the correct masses.

The model was trained for 300 epochs with a batch size of 128. During each epoch, we drew 500 000 strain-parameter pairs for training and 100 000 strain-parameter pairs for validation. The Adam optimizer [153] was used to optimize the weights of the model in conjunction with a cosine annealing scheme with warm restarts [154]. The learning rate oscillated between 10^{-3} and 10^{-5} with a period of 20 epochs; weight decay was set to 10^{-6}. Multiple hyperparameter configurations were tested; this configuration obtained the best performance.

To benchmark the trained model, an unseen test set was generated of 100 000 strain-parameter pairs at specific SNR values. The model was evaluated using the mean absolute angular error (maae) and the average 90% confidence area of the predicted VMF distributions.

3.3.2 Applying and evaluating importance sampling

To evaluate the importance sampling procedure, we constructed a slightly simpler test set in which we restricted the maximum spin magnitude to zero. This was done to limit the Bilby run-time. The importance sampling procedure discussed in Section 2.2 was applied to the first 100 strain-parameter pairs of this test set at three different optimal SNR values: 10, 15, and 20. For each strain-parameter pair, we generated 200 000 importance samples. To simulate multiple independent runs at various time points for the same strain-parameter pair, we subsampled from these 200 000 importance samples during the experiments.

We ran two experiments to test the convergence of the importance sampling method. In the first experiment, we used the importance sampling scheme as a maximum likelihood estimator. For a given set of importance samples, we chose the sample with the highest likelihood and calculated the angle between this sample and the true sky coordinates. In the second experiment, we represented the probability density function of the importance samples by a kernel density estimator and tested how well the resulting density covered the true right ascension. Specifically, we used a Gaussian kernel density estimator[3] to fit the right ascension distribution proposed by the importance samples. The log-likelihood of the actual right ascension was used to measure the quality of the estimated density. We removed a few outliers from the second experiment, by restricting ourselves to only the right ascension the number of outliers was reduced. These outliers had densities that did not cover the true right ascension at all, resulting in extreme negative log-likelihoods which dominate the average log-likelihood. For both experiments we expect the metric to improve as the number of importance samples increases, and to level after a significant number of importance samples indicating convergence.

[3]The *gaussian_kde* from the *scipy* python package.

3.3.3 Generating skymaps

We use Bilby as a benchmark to generate skymaps for the first ten strain-parameter pairs of the test set and for each create a version at an SNR of 10, 15, and 20. To make a fair comparison, the prior given to the Bilby sampler has its spin components set to zero. Moreover, the posterior inference was performed with standard settings, and each run took between 2.5 and 7 hours to complete. During these runs, the live points of the sampler were saved every 5 seconds and labeled by the total number of sampled points. These saved points were used to run the two importance sampling experiments for Bilby.

3.4 Results

In this section, we first discuss the performance of the CNN. Then, the importance sampling scheme is evaluated using the experimental setup discussed in the previous section. Lastly, we compare sky maps generated using only the neural network, importance sampling, and Bilby.

3.4.1 CNN

In Figure 3.3 we summarize the results for the first experiment: the left panel gives the mean absolute angular error (maae) in the sky location and the right panel we plot the 90% confidence area of the VMF distribution. As expected, as the SNR increases the prediction error in the sky location decreases and the 90% confidence area becomes smaller. The error in the mass prediction is similar to those of other CNN approaches [149], see Figure 3.4, indicating that the setup works well. We do note that the error in the sky location seems to be quite high for SNR < 10 and that it does not converge to zero for high SNR. We can think of two possible explanations for the poor performance at low SNR. First, the detection rate using either CNNs or matched filtering pipelines at an SNR of 5 is less than 40% [149, 155]. At such a low SNR, it is difficult for the model to discern the differences in arrival time at each detector, which explains the slightly better than random predictions for SNR < 7. When we compare

our angular error with other CNN approaches [150, 156], the average error seems to be similar. Furthermore, Chua and Vallisneri [157] reported that Gaussian approximations are only accurate for high SNR (SNR > 8), and even then multimodality might arise. Second, the sky distribution can be multimodal. This multi-modality is either due to strong noise or can be due to a sky reflection [102]. For three detectors, there are two viable solutions to the triangulation problem: the true sky location and its reflection. In most cases, the amplitude information is sufficient to break the degeneracy between the location and its reflection. However, at certain sky angles, this amplitude information does not lift the degeneracy, and a multimodal distribution is required. For these angles, the model has a 50% chance of guessing the wrong mode and thus has an average angular error of 90°.

3.4.2 Importance sampling

The results of the importance sampling experiments are shown in Figure 3.5. The left panel shows the maae as the number of importance samples increases. The right panel shows the log-likelihood of the true right ascension given by kernel density based on a varying number of importance samples. Most maae convergence occurs within the first 30 ,000 samples. The slow convergence mostly stems from strains with wide predicted sky distributions.. When we compare this to the results of Bilby, we see that the maae of the highest likelihood sample for all SNR is always between 1 and 8 degrees. Importance sampling is competitive for an SNR of 20 and is close for an SNR of 15, especially when we consider that in both cases 2 out of the 100 sky distributions were parameterized as the sky reflection.

However, importance sampling is not competitive with Bilby in the second experiment. For all SNR values Bilby reports log-likelihoods between 2 and 3, see the left side of Figure 3.6, and importance sampling does not reach these values. If we consider runs that show good convergence, i.e. where 90% of the importance weight is not determined by less than ten importance samples, importance sampling also reports log-likelihoods between 2 and 3. On the right side of Figure 3.6 we have repeated the kernel density experiment, but only for the well-converged runs. These runs represent 30% of all runs, and almost no SNR < 10 runs.

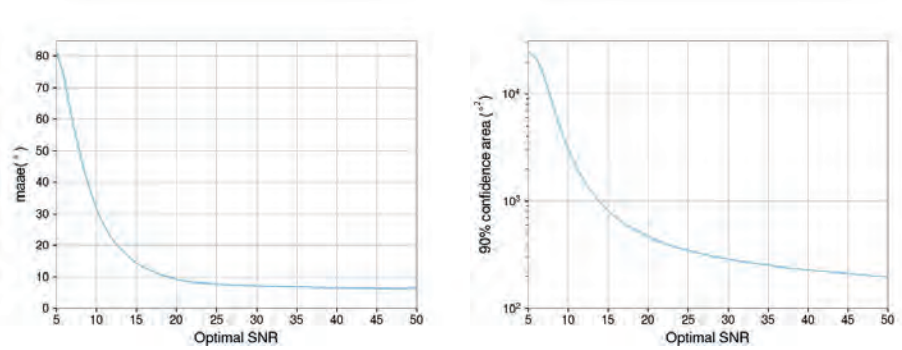


Figure 3.3: Characterization of the neural network in terms of accuracy and certainty over the test. Left: the maae (mean absolute angular error) between the sky angle predicted by the model and the actual sky location as a function of the SNR. Right: the average size of the 90% confidence area, expressed in degrees squared, of the predicted VMF distributions as a function of the SNR.

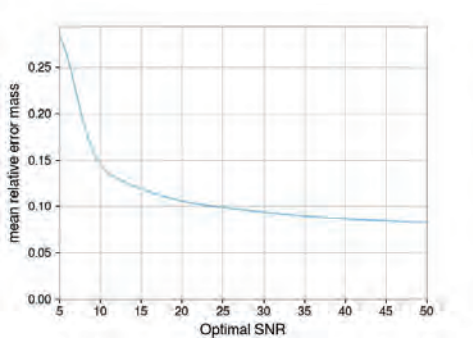


Figure 3.4: The mean relative error of the estimated masses by the neural network on the test set as a function of the optimal SNR. It is almost identical to Figure 5 in [149].

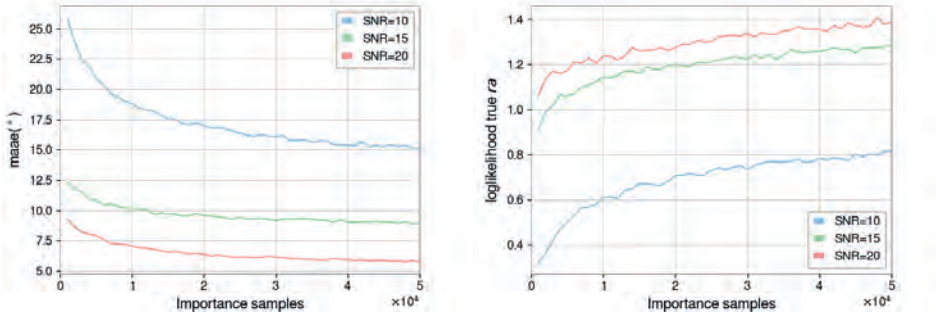


Figure 3.5: Characterization of the importance sampling, with the number of importance samples ranging from 1 000 to 50 000. The colors represent different SNR values with blue, green, and red being 10, 15, and 20 respectively. Left: the maae of the importance sample with the highest likelihood as a function of the sample size. Right: the log-likelihood of the true right ascension according to the kernel density estimator created by importance samples as a function of sample size.

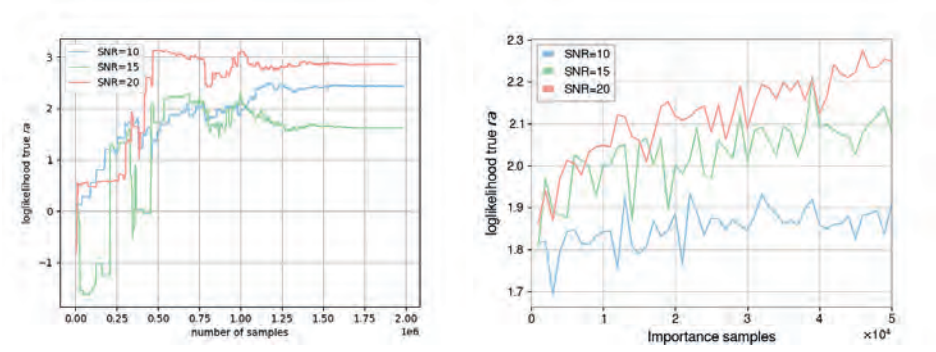


Figure 3.6: Left: The loglikelihood of the true right ascension according to the kernel density estimator created by the Bilby samples. The vertical axis represents how many samples Bilby has generated (live plus dead samples). Right: The log-likelihood of the true right ascension according to the kernel density estimator using only the importance samples of well converged runs. These values are more in line with those of Bilby.

3.4.3 Generating skymaps

As a final test, we generated skymaps using the neural network, importance sampling, and Bilby on the same signals. Three representative skymaps are shown in Figure 3.7. The skymaps generated by the neural network are significantly more spread out than those generated by importance sampling and Bilby. As we explained in the previous sections, this might be due to the neural network overestimating the uncertainty and having difficulty extracting the exact signal from the detector noise.

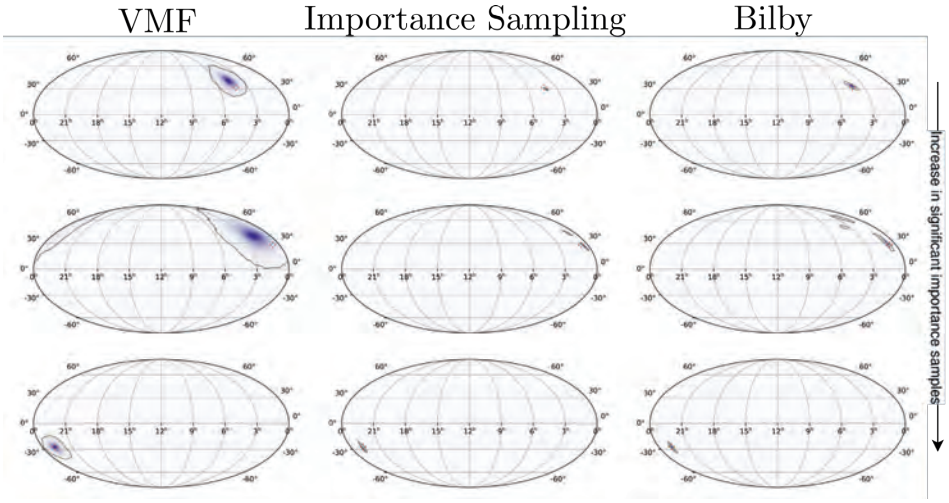


Figure 3.7: Examples of predicted skymaps by our neural network (left), importance sampling after 100 000 steps or roughly 5 minutes of computing time (middle), Bilby at convergence (right). The Bilby runs took at least 3 hours to complete. The true sky location is indicated in red. The shown skymaps were generated for signals with an SNR of 15. The number of significant importance samples, and hence the quality of the sky maps, increases as we go from the top row to the bottom row.

The skymaps generated by importance sampling and Bilby resemble each other quite a lot, their peak intensities are in the same position and the sky distributions occupy roughly in the same area. However, the importance sampling skymaps are grainy and sometimes do not cover the complete area that Bilby does. As can be seen in the bottom row of Figure

3.7, when the predicted VMF distribution has its peak intensity in the correct position the importance sampling creates better-looking sky maps. This improvement is due to the increased number of significant importance samples. These results indicate that a larger number of significant importance samples is needed, which is to be expected with only 5 minutes of run-time. Within only 1-4% of the Bilby run-time, we are already able to recover the essentials of the skymaps.

3.5 Conclusion

In this chapter, we produced skymaps for simulated BBH events using an importance sampling scheme that turns an approximate skymap made by a neural network into a skymap that represents the exact Bayesian posterior distribution. Experiments show that our method is competitive with Bilby and can produce the essentials of the skymap within 4% of the Bilby run-time. However, in some cases, the proposal distributions made by the neural network are too crude, which hampers the efficiency of the importance sampling scheme. If the sampling efficiency is improved further, importance sampling could be used as a quick alternative to Bilby or LALInference for inferring the GW posterior. Currently, the DL model has only been trained and tested on simulated noise with a given PSD. We expect that providing the DL model with various PSD representations as input into the model during training, as was done in [57], should allow the model to interpret the real-world signals correctly regardless of the noise profile. In future work, we will also consider more advanced deep learning models such as normalizing flows to infer more accurate posterior distributions and apply the model to real measurements.

Chapter 4

Normalizing Flows as an Avenue to Study Overlapping Gravitational Wave Signals

Due to its speed after training, machine learning is often envisaged as a solution to a manifold of the issues faced in gravitational-wave astronomy. Demonstrations have been given for various applications in gravitational-wave data analysis. In this chapter, we focus on a challenging problem faced by third-generation detectors: parameter inference for overlapping signals. Due to the high detection rate and increased duration of the signals, they will start to overlap, possibly making traditional parameter inference techniques difficult to use. Here, we show a proof-of-concept application of normalizing flows to perform parameter estimation on overlapped binary black hole systems.

Based on: Langendorff, J., Kolmus, A., Janquart, J., Van Den Broeck, C., "Normalizing flows as an avenue to studying overlapping gravitational wave signals". Physical Review Letters **130**, 171402 (2023).

4.1 Introduction

Over the last few years, the improved sensitivity of the LIGO [130] and Virgo [27] detectors has made the detection of gravitational waves (GWs) originating from compact binary coalescences (CBCs) more and more common, with over 90 detections reported after the third observation run [29]. Soon, the upgrade of the current detectors and the addition of KAGRA [28, 158–160] and LIGO India [161] to the network of ground-based interferometers will lead to even more detections. In addition, the passage from second-generation (2G) to third-generation (3G) detectors (Einstein Telescope (ET) [162, 163] and Cosmic Explorer (CE) [164–166]) will lead to an important increase in the number of observed CBCs. These detectors are also projected to have a reduced lower frequency cutoff [167], leading to longer signal durations. Therefore, CBC signals will overlap in 3G detectors [168–172].

Analyzing one of the overlapping signals without accounting for the presence of the other can lead to biases in the recovered posteriors, especially when the merger times of the two events are close [169–173]. These could impact any direct science case for CBCs (e.g. tests of general relativity [174]), but also indirectly related ones such as the hunt for primordial black holes [175–180]. In Ref. [181], the authors demonstrate on two overlapped binary black holes (BBHs) how adapted Bayesian inference can help reduce the biases. In particular, they perform joint parameter estimation, where the two signals are analyzed jointly. While accounting for all the noise characteristics, their analysis also suffers from some instabilities, and further upgrades are needed for it to be entirely reliable. An issue also mentioned in this work is the computational time. With hundreds of thousands of CBC mergers expected in the 3G era [169], analyses taking several weeks are not a realistic alternative.

Even if traditional methods can be sped-up [52–54, 182], or quantum computing [183] could potentially be used in the future, the development of frameworks capable of doing complete analyses in short timescales is crucial for the development of 3G detectors. Therefore, in this work, we propose the first step in that direction, showing how overlapping BBHs can be analyzed with a normalizing flow (NF) approach [112, 114, 115].

4.2 Machine learning for overlapping gravitational waves

The use of machine learning (ML) in GW data analysis has been growing over the last years, having a wide range of applications [184]. A subset of these methods fall under the umbrella of simulation-based inference [185], and are being developed to perform parameter estimation for CBCs [57, 139, 143, 144, 146, 186–188]. Refs. [57, 187, 188] use NFs to get posterior distributions for BBH parameters, obtaining results close to those from traditional Bayesian methods. Our approach is somewhat similar to theirs, with some notable differences explained below.

Our approach uses *continuous conditional NFs* [145, 189] (CCNFs), a variant of NFs suited for probabilistic modeling and Bayesian inference. Due to the recursive and continuous nature of these models, their memory footprint can be quite small [123], allowing for extensive training on home-grade GPUs while retaining the ability to capture complex distributions.

NFs are a method in ML through which a neural network can learn the mapping from some simple base distribution $p_u(\boldsymbol{u})$ to a more complex final distribution $q(\boldsymbol{\theta})$. This is done through a series of invertible and differentiable transformations, summarized by a function $g(\boldsymbol{\theta})$. However, in our case, the final distribution we seek depends on the GW data to analyze. Therefore, we use *conditional NFs* [190], where the transformation functions are dependent on the data $\boldsymbol{d}$ (hence, $g = g(\boldsymbol{\theta}, \boldsymbol{d})$). A major difference with [190] is that our base distributions are kept static. Thus our model $g(\boldsymbol{\theta}, \boldsymbol{d})$ is a trainable conditional bijective function transforming a simple 30-D Gaussian into a 30-D complex distribution. The bijectivity allows us to express and sample $q(\boldsymbol{\theta}|\boldsymbol{d})$ in terms of $g(\boldsymbol{\theta}, \boldsymbol{d})$ and $p_u(\boldsymbol{u})$ via:

$$q(\boldsymbol{\theta}|\boldsymbol{d}) = \left|\det(\mathrm{J}_{\mathrm{g}^{-1}}(\boldsymbol{\theta}, \boldsymbol{d}))\right| \mathrm{p}_{\mathrm{u}}(\mathrm{g}^{-1}(\boldsymbol{\theta}, \boldsymbol{d}))\,, \tag{4.1}$$

where $\det(\mathrm{J}_{\mathrm{g}^{-1}}(\boldsymbol{\theta}, \boldsymbol{d}))$ is the determinant of the Jacobian $J_{g^{-1}}(\boldsymbol{\theta}, \boldsymbol{d})$ of the transformation. For training, we minimize the forward KL-divergence, which is equivalent to maximum likelihood estimation [63, 115]. As noted by [188], $q(\boldsymbol{\theta}|\boldsymbol{d})$ should cover the actual (Bayesian) posterior $p(\boldsymbol{\theta}|\boldsymbol{d})$, and asymptotically approach it as training progresses due to the mode-covering nature of the forward KL divergence.

A distinctive choice of our method is the continuous nature of the flow, which is linked to the transformation function itself. Neural ordinary differential equations (neural ODEs) [123] are the foundation of continuous NFs; they are not represented by a stack of discrete layers but by a hypernetwork [191]. Hypernetworks can be understood as regular networks where 'external' inputs such as a time or depth variable smoothly change the output of the network for identical inputs. They can thus represent multiple transformations. In [123], hypernetworks are used to represent ODEs and are trained by using ODE-solvers and clever use of the adjoint sensitivity method. A continuous NF uses neural ODEs as its transformations.

We will now explain the training of a continuous flow. For clarity, we will use h to refer to a continuous transformation and g for a discrete one. If $\boldsymbol{\theta}(t)$ represents the samples from the distribution at a given time t, when going from t_1 to t_2, the continuous NF obeys

$$\frac{\mathrm{d}\boldsymbol{\theta}(\mathrm{t})}{\mathrm{dt}} = h(t, \boldsymbol{\theta}(t)) \,. \tag{4.2}$$

The change in likelihood associated with this 'step' differs slightly from Eq. (4.1) due to the continuous nature of the flow:

$$\log(p(\boldsymbol{\theta}(t_1))) = \log(p(\boldsymbol{\theta}(t_0))) - \int_{t_0}^{t_1} \mathrm{Tr}\left[\mathrm{J_g}(\boldsymbol{\theta}(\mathrm{t}))\right] . \tag{4.3}$$

Assuming a non-stiff ODE the integration can be performed rapidly with state-of-the-art ODE-solvers, MALI [192] in our case. In addition, we have to calculate a trace instead of a determinant, speeding up the computation which reduces the complexity, going from $\mathcal{O}(D^3)$ to at most $\mathcal{O}(D^2)$ with D being the dimensionality of posterior space, speeding-up the computation [124]. Moreover, using continuous NFs removes the need to use coupling layers between transformations, instead, all parameter dimensions can be dependent on each other throughout the flow. Combining the continuous and conditional flows leads to CCNFs, where the conditional consists of the GW $\boldsymbol{d}$ and the time t.

We also need a better data representation than the raw strain to train and analyze the data. Therefore, we follow a similar approach as in [57, 144,

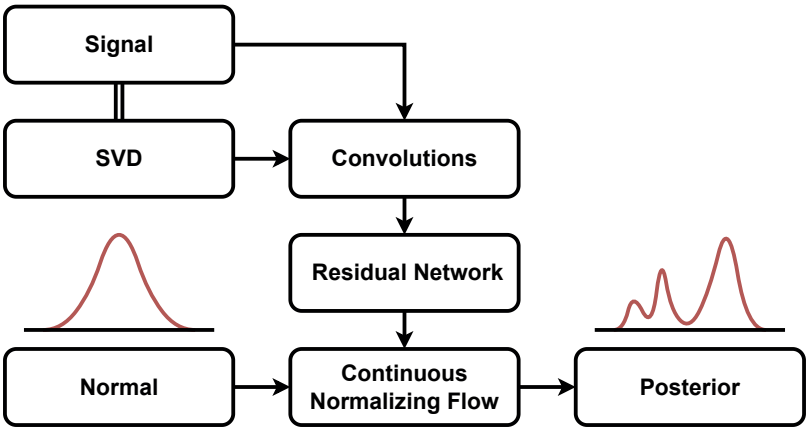


Figure 4.1: Representation of our analysis framework. It is made of a pre-processing part where we build an SVD basis to filter the data, followed by a normalizing-flow-based neural network.

186–188], using a singular value decomposition (SVD) [193] as summary statistics, reducing the dimension and the noise content of the data while retaining at least 99% of the original signal. Each of the 256 generated basis vectors is used as a kernel in 1D convolutions used as an initial layer in a ten-layer residual convolutional neural network (CNN), enabling one to capture the time variance of the signal. Therefore, we do not need to use a Gibbs sampler to estimate the time of the signal as done in [57, 187, 188], and can sample over time like any other variable. The CCNF itself is represented by two multi-layer perceptrons with 3 hidden layers of 512 units. Furthermore, we use a different representation for the angles. Instead of directly using their values, we project them onto a sphere for the sky location and onto a circle for the other angles. This makes for a better-posed domain for these angles, and plays on the strong interpolation capacities of the network, making the training step easier.

In the end, our framework combines data representation as a hybrid between SVD and CNN, followed by the CCNF network. A representation of our analysis framework is given in Fig. 4.1. Our entire framework is relatively small compared to the ones presented in [187], both the residual network and CNF network. Therefore, it can run on lower-end GPUs, but could also be limited in its capacity to model the problem.

4.3 Data and setup

To test our framework, we start with a simplified setup, considering a network made of the two LIGO detectors and the Virgo detector, at design sensitivity [27, 194], and with a lower sensitive frequency of 20Hz. We generate stationary Gaussian noise from their power spectral density (PSD) and inject two precessing BBH mergers using the IMRPhenomPv2 waveform [195]. Our data frames have an 8 seconds duration and are whitened after the signals are injected. The chirp mass ($\mathcal{M}_c = (m_1 + m_2)^{3/5}/(m_1 m_2)^{1/5}$) and mass ratio ($q = m_2/m_1$) are sampled from uniform distributions, between $10M_\odot$ and $100M_\odot$ and 0.125 and 1, respectively. The individual component masses are constrained between $5M_\odot$ and $100M_\odot$. During the data generation, the luminosity distance is kept fixed. It is then rescaled to result in a network signal-to-noise ratio value taken randomly between 10 and 50 from a beta distribution with a central value of 20. The coalescence time for the two events is set randomly around a time of reference, with $t_c \in [t_{\rm ref} - 0.05, t_{\rm ref} + 0.05]$s, ensuring that the two BBH merge in the high bias regime [170]. The other parameters are drawn from their usual domain. Table 4.1 gives an overview of the parameters and the function from which they are sampled.

During the training, we continuously generate data by sampling the prior distributions for the events and making a new noise realization for each frame. The training is stopped when convergence is reached and before over-fitting occurs. Our model was trained for about 12 days on a single *Nvidia GeForce GTX 1080*.

4.4 Results

To demonstrate the method's reliability, a P-P plot for the recovered parameters is shown in Fig. 4.2. It is constructed by sampling the posteriors of 1000 overlapped events[1] with parameters drawn from the distributions detailed in Table 4.1. Since the cumulative density aligns along the diagonal, our network is reliable. Comparing this to the results given in [187]

[1]We refer the reader to Fig. 1 in Ref. [181] for an illustration of overlapping BBH signals.

Parameter	Function
Chirp mass ($\mathcal{M}$)	$\mathcal{U}(10, 100)M_\odot$
Mass ratio (q)	$\mathcal{U}(0.125, 1)$
Component masses ($m_{1,2}$)	Constrained in $[5, 100]M_\odot$
Luminosity distance (D_L)	Rescaled to follow SNR
SNR	$\mathcal{B}(10, 50)$
Coalescence time (t_c)	$\mathcal{U}(t_{\rm ref} - 0.05, t_{\rm ref} + 0.05)$
Spin amplitudes ($a_{1,2}$)	$\mathcal{U}(0, 1)$
Spin tilt angles ($\theta_{1,2}$)	Uniform in sine
Spin vector azimuthal angle (ϕ_{jl})	$\mathcal{U}(0, 2\pi)$
Spin precession angle (ϕ_{12})	$\mathcal{U}(0, 2\pi)$
Inclination angle (θ_{jn})	Uniform in sine
Wave polarization (ψ)	$\mathcal{U}(0, \pi)$
Phase of coalescence (ϕ)	$\mathcal{U}(0, 2\pi)$
Right ascension (RA)	$\mathcal{U}(0, 2\pi)$
Declination (DEC)	Uniform in cosine

Table 4.1: Summary of the parameters considered and the function used to generate the BBHs.

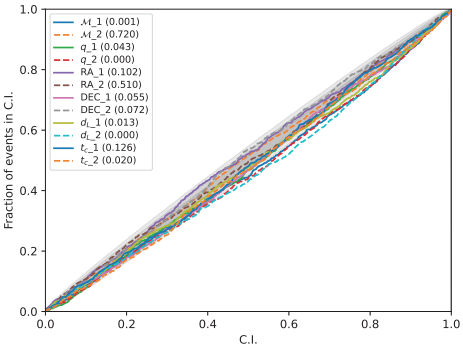


Figure 4.2: P-P plots for a subset of the recovered parameters for the two events in the data. The parameters shown are representative of all the BBH parameters for the two events. In both cases, the lines align along the diagonal, showing that our method can be trusted. The legend indicates which line corresponds to which parameters. The parameters for event 1 (resp. 2) are noted P_1 (resp. P_2), where P are the usual parameter symbols as presented in Table 4.1. The values between the brackets are the KS test statistic.

for single signals, there is a broadening of the shell around the diagonal, showing more variability in signal recovery, meaning our inference is less accurate than for single signals. Possible origins are the degenerate posteriors, increased complexity of the problem, and the reduced size of our network. This increased variability when going from single to joint parameter estimation has also been noted in Bayesian approaches [181].

While Bayesian methods have been developed in [181], they are not yet fully stable and take a long time to analyze a BBH system. Therefore, making a statistically significant study comparing the two approaches seems a bit premature at this stage. However, to have some sense of the performances of our network compared to traditional methods, we make 15 injections complying with our network's setup and analyze them with the framework presented in [181]. Using these analyses, we can already identify some trends between the two pipelines. The first is that our ML pipeline typically has broader posteriors than the Bayesian approach. As mentioned in Ref. [181], the classical joint parameter estimation approach

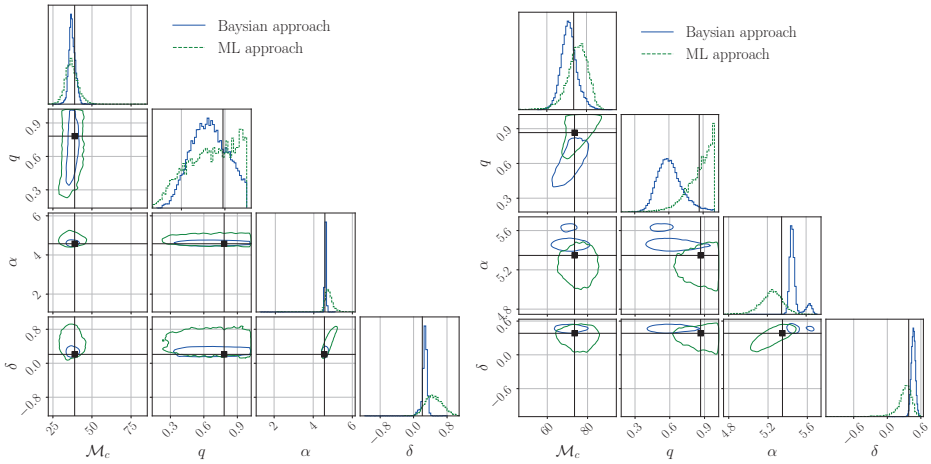


Figure 4.3: Comparison between our approach and the one from [181] for two separate events and for the chirp mass, mass ratio, right ascension, and declination. The injected values are given by the black lines. For the left event the true value is encapsulated by the posteriors of both methods, for the right event this is only the case for our method. Our posteriors are generally broader but include the injected value within the 90% confidence interval. This could be corrected by applying importance sampling on the output samples.

can sometimes get overconfident –see Ref. [181] for a discussion on the Bayesian algorithm–, where the recovered injected value lies outside of the 90% confidence interval. Our method is not confronted with this bottleneck as the broader posterior encapsulates the injected value. Fig. 4.3 illustrates the two representative situations: one where the Bayesian approach finds the event correctly, and one where we see that our ML approach covers the injected values while it does not for the classical approach. Bias in the posterior, similar to the one noted in Ref. [181], can exist in our method and would not be seen because of the broad posteriors. However, because we are using the forward KL divergence, we expect the posteriors to have some support for the injected values. The origin of the larger posterior, which is not observed in the single parameter estimation machine learning-based methods, is probably due to the increased complexity of the problem combined with the small residual and CNF network sizes. One possible avenue is applying importance sampling after the normalizing flow as shown in Chapter 3 or Ref. [188]. However, such methods can be tricky, and additional modifications to our network could be needed.

Finally, an important advantage of our method is its speed. After being trained, it can analyze two overlapping BBH signals in about a second, to compare with $\mathcal{O}(20days)$ reported in [181]. While it is difficult to estimate the time gain for other CBC signals, we can expect the inference time after training not to be significantly larger than for BBHs. Since computational time is a crucial aspect of studies in the 3G era, ML approaches seem to be more suited to study realistic scenarios for these detectors.

4.5 Conclusions and Perspectives

In this work, we have presented a proof-of-concept machine learning-based method to analyze overlapping BBH signals. We focused on a 2G detector scenario with the two LIGO, and the Virgo detectors at design sensitivity, with a lower frequency cutoff of 20Hz. Our approach is based on continuous normalizing flows.

While also using normalizing flows, as in [57, 144, 186–188], we bring extra modifications that seem to help in the inference task. We represent the data through a mixture of SVD and convolutions, enabling us

to sample directly over the events' arrival time, retaining the ability to access the likelihood of a sample. We also move to continuous conditional normalizing flows, reducing the computational cost of the method as we need to solve a trace instead of a determinant when going from one step to the other in the transformation. Finally, we also use a particular representation of the angles, projecting them onto circles (for the phase, the polarization, …) and spheres (for the sky location). We believe that these modifications make our network more flexible, enabling it to deal with overlapping signals even in a reduced form.

With this simplified setup, we have shown that our approach is reliable, with posteriors consistent with the injected values. Our method takes about one week to train on a single GPU. After that, it only takes about a second to analyze two overlapped BBHs. While, in reality, other types of CBC mergers can happen, their inference after training should not be significantly longer than for BBHs. We also compared our machine learning method with classical Bayesian methods for overlapping signals. While our scheme leads to wider posteriors, it can correctly recover the injected values, even when the Bayesian approach gets overconfident and misses the injection. A possibility to correct for the widened posteriors is to use importance sampling.

Our method's combined reliability and speed show that machine learning is a viable approach to analyzing CBC mergers in the 3G era. More interestingly, it would even be possible without needing to account for the development of more powerful computational means and could enable some science-case studies for ET and CE soon. For example, once trained for all possible BBH systems, it could help study the BBH mass function in the 3G era.

Still, one should note that extra improvements are needed before using our method in realistic 3G scenarios. One would first need to change our setup to the 3G detectors, where a lower frequency cutoff and extreme SNRs could be encountered. In addition, a wider range of objects should be accounted for. One should include higher-order modes and eccentricity as they could play a crucial role in the 3G era. Other modifications could also be implemented. Additionally, we need to account for the change in noise realization from one event to the other. Some of these steps, like

changing the detector configuration, should be relatively easy. Others are more complex, as it is hard to perform parameter inference for long-lasting mergers due to the computational burden. So, extra developments in parameter estimation using machine learning would be required to get to the realistic 3G scenario. For overlapping signals, one would also benefit from developments in the classical study of the 3G scenario, such as how to deal with the noise characterization or the types of other events that could come into the data.

In the end, there is still work to be done before machine learning can be used in realistic 3G scenarios. However, we believe that this work shows it is an interesting avenue and could be practical on a relatively short time scale.

Chapter 5

Tuning Neural Posterior Estimation for Gravitational Wave Inference

Modern simulation-based inference techniques use neural networks to solve inverse problems efficiently. One notable strategy is neural posterior estimation (NPE), wherein a neural network parameterizes a distribution to approximate the posterior. This approach is particularly advantageous for tackling low-latency or high-volume inverse problems. However, the accuracy of NPE varies significantly within the learned parameter space. This variability is observed even in seemingly straightforward systems like coupled-harmonic oscillators. This chapter emphasizes the critical role of prior selection in ensuring the consistency of NPE outcomes. Our findings indicate a clear relationship between NPE performance across the parameter space and the number of similar samples trained on by the model. Thus, the prior should match the sample diversity across the parameter space to promote strong, uniform performance. Furthermore, we introduce a novel procedure, in which amortized and sequential NPE are combined to swiftly refine NPE predictions for individual events. This method substantially improves sample efficiency, on average from nearly 0% to 10-80% within ten minutes. Notably, our research demonstrates its real-world applicability by achieving a significant milestone: accurate and swift inference of posterior distributions for low-mass binary black hole (BBH) events with NPE.

Based on: Kolmus, A., Janquart, J., Baka, T., van Laarhoven, T., Van Den Broeck, C., & Heskes, T., "Tuning neural posterior estimation for gravitational wave inference". arXiv preprint arXiv:2403.02443 (2024).

5.1 Introduction

Inverse problems encompass the challenging task of deducing the underlying causal factors behind observed phenomena in various scientific domains [196–199]. A specific example of such a phenomenon is a gravitational wave (GW) – coherent, tiny ripples in space-time generated by the acceleration of massive celestial objects such as black holes or neutron stars [200]. The observatories of the LIGO-Virgo-KAGRA collaboration [27, 28, 130] regularly observe these GW events [29]. The insights derived from analyzing these events have a huge impact on the field of astronomy [201–204]. To continue progressing, it is crucial to infer the properties of new GW events accurately, and in a timely manner, especially since the computational demands continue to grow as the detectors improve [205]. In this introduction, we will give a brief overview of traditional and neural methods for solving inverse problems, focusing on their applicability in GW astronomy.

How does one find the causal factors explaining an observation x_{obs}? Traditionally, tackling complex inverse problems involves three components. First, a simulation model is needed to translate event parameters θ into synthesized observations x. Next, a likelihood function $p(x|\theta)$ is determined, and finally, Bayesian inference methods construct a posterior distribution over the parameters θ given by Bayes' theorem:

$$p(\theta|x_{obs}) = \frac{p(x_{obs}|\theta)p(\theta)}{p(x_{obs})}, \tag{5.1}$$

where $p(\theta)$ is the prior distribution and $p(x_{obs})$ is the evidence. These Bayesian methods often evaluate millions to billions of potential event parameters before converging to the true posterior distribution. Therefore, quick evaluation of the likelihood function is a necessity. However, obtaining such a practical likelihood function $p(x|\theta)$ can be challenging due to mathematical or computational complexity.

Current GW pipelines built on this traditional framework take a lot of time to run, ranging from hours to a full month depending on the event properties and desired accuracy [102, 103, 206]. The primary factor contributing to the runtime is the evaluation of the likelihood, which

requires simulating a GW. Simulating a GW can take anywhere from tens of milliseconds to several seconds [207], dependent on variables like sampling frequency, signal duration, and the chosen simulation algorithm. With the anticipated construction of third-generation detectors [162], alongside the planned upgrades to existing observatories such as LIGO and Virgo [208], the computational demands are expected to surge. Consequently, the accurate inference of posterior distributions for future GW observations without substantial enhancements poses a growing challenge. As a result, there is a growing interest in alternative methods for GW inference [52, 53, 56, 58, 139, 146, 157, 188, 209–212].

Simulation-based inference (SBI) methods [185] offer potential alternatives for solving inverse problems in a more computationally efficient manner. These methods approximate the posterior distribution and need only a simulation model. In recent years, neural networks (NNs) have gained considerable prominence in the SBI domain [213]. Due to their expressivity and capacity, NNs can mimic essential components of the Bayesian inference framework: the likelihood [61], the likelihood-ratio [214], and the posterior itself [112]. The neural likelihood ratio and neural posterior methods can be trained either for a single event or for any possible event from the prior distribution; these modes are respectively referred to as non-amortized and amortized inference. The latter takes longer to train and is potentially less accurate but the computational burden is paid in advance and only once. Consequently, amortized inference is preferred when faced with low-latency or high-volume challenges. Of special interest is amortized neural posterior estimation (NPE) [112], where one trains a conditional neural density estimator to transform a simple, well-understood distribution into an approximate posterior $Q(\theta|x)$. To our knowledge, this is the only neural SBI method that does not require any subsequent Bayesian or variational inference steps to construct an approximate posterior and thus allows for sub-second inference [144].

In NPE, an NN parameterizes an approximate posterior distribution over the event parameters. The NN is trained by feeding it simulated observations x and iteratively increasing the likelihood of the true parameters θ in the predicted distribution. While mixture density networks [215] and normalizing flow (NF) models [112] are both commonly used in NPE,

our focus in this work is exclusively on NF models due to their high expressivity. NFs consist of a sequence of differentiable, bijective functions with parameters defined by NNs. These functions can transform a simple distribution into a complex one while accurately tracking the likelihood via the change of variables theorem. The loss function commonly used in NPE is the forward KL divergence, which is equivalent to maximum likelihood estimation for NFs [145]. Due to the mode-covering property of the forward Kullback-Leibler (KL) divergence, the approximate posterior should always cover the true posterior[216]. This property enables using importance sampling to converge to the true posterior when a known likelihood function is available. As we will see, NPE followed by importance sampling produces similar results as traditional methods for significantly reduced computational costs [188, 217].

While NPE holds promise as an alternative for full GW inference, certain challenges need to be addressed. First, NPE can struggle with generalizing across the entire parameter space. As we shall demonstrate in section 5.3, even for simple problems, NPE can have poor sample efficiency for specific subsets of the data. We hypothesize and characterize a correlation between the performance of an NPE model for a specific event and the number of similar samples it has been trained on. The reasoning behind this hypothesis is that NNs learn from examples. Effective training thus demands a prior that exposes the NN to diverse samples, which often does not correspond to the uninformative prior, but an effective one. Second, NPE models struggle to be competitive with Bayesian inference when they need to learn large numbers of high-dimensional observations, producing posteriors that appear correct but are wider than their Bayesian counterparts. Extended training can compensate to some extent, but does not scale well. In section 5.4 we propose fine-tuning of trained NPE models for single instances of the problem. This procedure optimizes the NPE model by self-sampling and correcting these samples with an importance-weighted loss function. By switching from learning all possible events to only a single instance, the problem becomes a lot easier to optimize for. To demonstrate the improvements offered by switching to effective priors and fine-tuning, section 5.5 shows that we can infer

previously inaccessible low-mass parameter ranges[1] for binary-black hole (BBH) mergers observed with GWs. To our knowledge, the inference of posterior distributions for low-mass BBH events remains beyond the reach of existing SBI algorithms [56, 143, 188].

5.2 Experimental setup

To investigate the behavior of NPE, we first start with a simple toy problem. This section will first describe our toy problem: coupled-harmonic oscillators, an ideal toy problem for three reasons: (1) they are computationally inexpensive to generate, (2) there is a known and cheap likelihood function, which makes importance sampling straightforward, and (3) like gravitational wave observations, they are a time series and have correlated channels. This section will end with a description of our NPE model and training setup.

5.2.1 Toy problem description

We study a linear chain containing four oscillators moving along a single axis, as illustrated in Figure 5.1. Each oscillator has a mass m and is connected to its neighbors by springs with a spring constant k. The first and last oscillators in the chain are attached to rigid walls by springs on their left and right sides. The system's dynamics can be expressed in terms of normal modes v. This expression derived in [218] where the displacement over time $x_u(t)$ of oscillator u is expressed as a sum over the four normal modes with known amplitudes a_v and phases ϕ_v:

$$x_u(t) = \sum_{v=1}^{4} |a_v| \sin\left(\frac{v}{5} u\pi\right) \cos\left(2\sqrt{\frac{k}{m}} |\sin\left(\frac{v\pi}{10}\right) |t + \phi_v\right) \tag{5.2}$$

Conversely, given the displacements over time, the amplitudes and phases of the normal modes can be determined. The described toy problem has a 10-dimensional parameter space: mass m, spring constant k, and for each normal mode an amplitude a_v and phase ϕ_v.

[1] Down to a chirp mass of 5 solar masses.

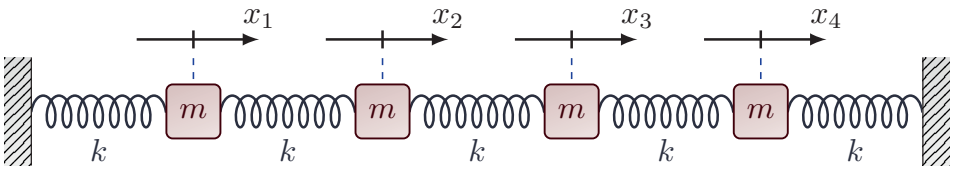


Figure 5.1: An illustration of the coupled-harmonic oscillators used in the toy problem. There are four oscillators, each has a mass m, and they are connected via springs with spring constant k. Their displacement from rest position $x_u(t)$ is measured along the horizontal plane.

To introduce uncertainty into the inverse problem, we incorporate white noise into the observed displacements $x_{obs}(t)$, and discretize these into $\boldsymbol{x}_{obs}$[2] depending on the sampling frequency. This allows us to use the simplest version of the Whittle likelihood function [92], given by:

$$p(\boldsymbol{x}_{obs}|\boldsymbol{\theta}) \propto \exp\left(-\sum_{t,u}\frac{(\boldsymbol{x}_{obs,u}-\boldsymbol{s}_u(\boldsymbol{\theta}))^2}{2}\right), \tag{5.3}$$

where $\boldsymbol{\theta}$ represents the inferred mass, spring constant, amplitudes, and phases, and $\boldsymbol{s}_u(\boldsymbol{\theta})$ denotes a clean simulated signal parameterized by $\boldsymbol{\theta}$. This likelihood function measures how well the residuals, the observation minus the simulation, match a standard normal distribution. A set of $\{\boldsymbol{\theta}_i\}$ for which the residuals resemble white noise should explain the observation well. To evaluate the performance of an NPE model, ideally, we would calculate the KL divergence between the exact Bayesian posterior and the posterior predicted by the NPE model. However, the computational costs would be excessively high for all the experiments in this chapter. Instead, the sample efficiency η is used to quantify the performance of the NPE model. For n drawn samples from the NPE model, the sample efficiency is defined as:

$$\eta = \frac{\left(\sum_{i=0}^{n} w_i\right)^2}{n\sum_{i=0}^{n} w_i^2} = \frac{n_{\text{eff}}}{n}, \tag{5.4}$$

where w_i represents the ratio between the Whittle likelihood and the likelihood given by the NPE model for the ith sample. And n_{eff} is the Kish

[2]Bold symbols indicate vectors.

effective sample size [219]. If n is sufficiently large and the support of our approximate distribution covers the support of the true distribution, we can interpret the sample efficiency as a quality measure of the approximate distribution. NF models trained with the forward KL-divergence are in general mode-covering and can generate thousands of posterior samples within a second, satisfying these requirements.

5.2.2 NPE model specification and training

As can be seen in Figure 5.2, the NPE model is a combination of two models:

(1) The context model transforms the time series into a neural representation. It begins with a linear transformation, followed by three residual blocks, and ends with another linear transformation to produce the neural representation. Although it has a consistent structure across experiments, the dimensions of the linear layers can change to accommodate longer or more complex time series. The specific dimensions of the context network for each experiment can be found in the at the end of this section.

(2) The NF model transforms a base distribution Q_b into a complicated distribution Q_z. The NF model builds this transformation via a series of coupling layers [113]. A coupling layer with index l divides the input into two halves: a dynamic $\boldsymbol{b}_i^l$ and static $\boldsymbol{b}_j^l$, where the static half acts as a condition for the transformation of the dynamic half. The transformation is a function f, which has to be differentiable and bijective in its first parameter. It is typically a monotonically increasing polynomial whose coefficients β are generated by an NN g with parameters $\boldsymbol{\tau}_l$. The input to g is the static half, and possibly a context vector c. The output of such a coupling layer is

$$\boldsymbol{b}_i^{l+1} = f(\boldsymbol{b}_i^l, g(\boldsymbol{b}_j^l; \boldsymbol{\tau}_l)) = f(\boldsymbol{b}_i^l, \beta^l) \tag{5.5}$$

$$\boldsymbol{b}_j^{l+1} = \boldsymbol{b}_j^l. \tag{5.6}$$

By alternating which dimensions are dynamic and which are static in consecutive coupling layers, the model can represent a flexible distribution over the parameters. The entire series of coupling layers is denoted S and the corresponding set of parameters is denoted $\boldsymbol{\psi}$. The principle of a normalizing flow is based on the equivalence relation between Q_z and Q_b

via the change of variables theorem:

$$Q_z(\boldsymbol{z}|\boldsymbol{\psi}) = Q_b(\boldsymbol{b}) \,|\, \det \mathbf{J}_S(\boldsymbol{b};\boldsymbol{\psi}) \,|^{-1} \text{ where } \boldsymbol{z} = S(\boldsymbol{b}). \tag{5.7}$$

Here, $\mathbf{J}_S(\boldsymbol{b};\boldsymbol{\psi})$ represents the Jacobian of the transformation function. One can optimize Q_z to approximate an (unnormalized) target distribution $P(\boldsymbol{z})$ with $Q_z(\boldsymbol{z}|\boldsymbol{\psi})$ by minimizing the forward KL-divergence $D_{KL}(P(\boldsymbol{z})|Q_z(\boldsymbol{z}|\boldsymbol{\psi}))$, which for NF models is equivalent to fitting $Q_z(\boldsymbol{z}|\boldsymbol{\psi})$ by maximum likelihood estimation [145]. The loss function for a single sample reads:

$$L(\boldsymbol{z}|\boldsymbol{\psi}) = -\log(Q_z(\boldsymbol{z}|\boldsymbol{\psi})). \tag{5.8}$$

The base distribution of our NF model is a truncated standard normal distribution. We went with a truncated distribution since they match naturally with the boundaries of the parameter space, for example, the phase is bounded between 0 and 2π. For our transformation function, we choose Bernstein polynomials [117], which are both highly expressive and robust, regardless of noise or polynomial order. These qualities allow us to build a relatively shallow, yet highly expressive NF model. It is also faster to train and has a smaller memory footprint, compared to the conventional RQ-spline NF models [116]. As we shall see in section 5.4, a fast NF model is very beneficial if low latency is desired. For all of the experiments, the Bernstein polynomials are parameterized by a shallow multi-layer perceptron (MLP).

Table 5.1: The priors used for the data generation of the coupled-harmonic oscillators. The prior for the mass is a power law prior whose coefficient was either -3.0, -1.5, or 0.0, depending on the experiment.

Parameter	Prior	Min	Max	Unit
m	Power law (-3.0, -1.5, 0.0)	0.1	10	kg
k	Uniform	10	100	N/m
$a_{0,1,2,3}$	Uniform	0.5	5.0	m
$\phi_{0,1,2,3}$	Uniform	0	2π	rad

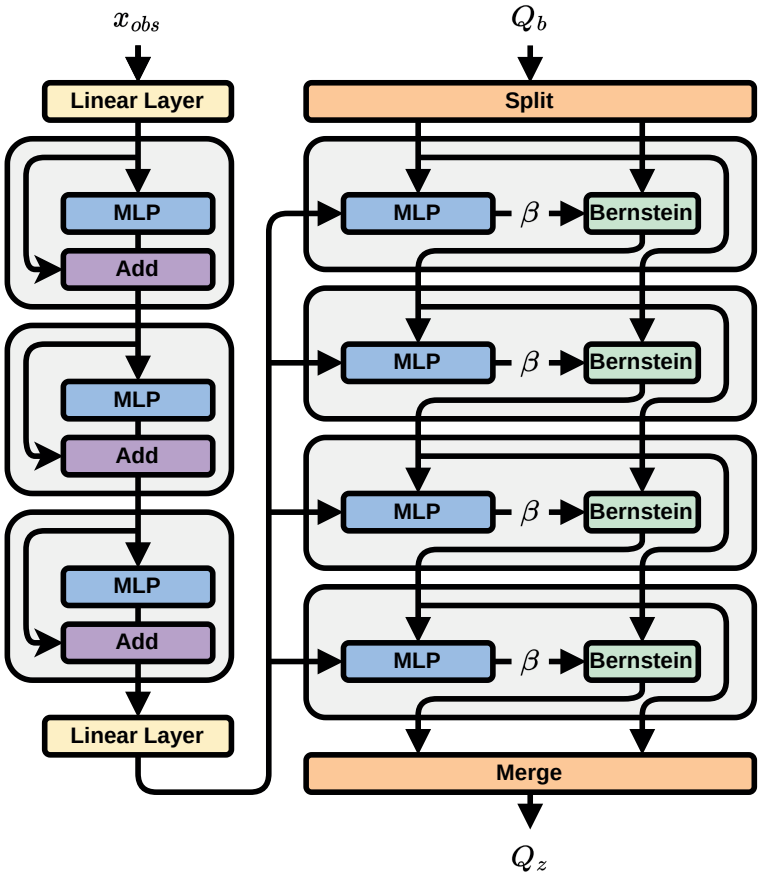


Figure 5.2: A schematic of our NPE model. The flow through the schematic is made explicit by arrows. The left side of the schematic shows the context network, which is a residual network consisting of three residual blocks. The right side shows the NF model which transforms a simple distribution Q_b into an approximate posterior distribution Q_z. The NF model consists of four coupling layers, each conditioned by the output of the context network. The method of conditioning is discussed in more detail in Section 5.4. Each coupling layer has two inputs, a dynamic half and a static half. The static half is used as input into an MLP which produces the β coefficients for the Bernstein polynomial, which transforms the dynamic half. At the end of the coupling layer the dynamic and static halves trade places for the consecutive coupling layer.

Table 5.2: The priors used for the data generation of the GWs. Instead of luminosity distance the optimal SNR of the signal is sampled. During the generation of the waveform the luminosity distance is set to 1000 MPC and after the generation, the waveform and luminosity distance are scaled to match the desired SNR.

Parameter	Name	Prior	Min	Max	Unit
m_1, m_2	Component mass	Constraint	3	150	$M_\odot$
$\mathcal{M}_c$	Chirp mass	Power law (-3.0)	5	100	$M_\odot$
q	Mass ratio	Power law (-1.5)	0.2	1.0	-
$\|\chi_1\|$, $\|\chi_2\|$	Spin amplitudes	Uniform	0	0.9	-
THETA	Sky coordinate 1	Uniform	0	2π	rad
PHI	Sky coordinate 2	Cosine	0	π	rad
t_c	Coalescence time	Uniform	-0.1	0.1	s
ϕ_c	Coalescence phase	Uniform	0	2π	rad
ι	Inclination angle	Sine	0	π	rad
ψ	Polarization angle	Uniform	0	π	rad
SNR	Signal-to-noise ratio	Uniform	10	30	-

Coupled-harmonic oscillators setup

The context network consisted of a linear layer, three residual blocks, and a linear layer. The initial linear layer reduced the dimension from number of oscillators × duration × sampling frequency down to 512. The residual blocks contained an MLP following the pre-activation format suggested in reference [220]. Specifically, the MLP was defined by the following sequence a GELU activation function [221], a LayerNorm [222], a linear layer with an output dimension of 512, followed by a GELU activation, a LayerNorm, and linear layer with an output dimension of 512. The final linear layer reduced the dimension from 512 to 128. The weight vectors of the linear layers were reparamertized following the weight normalization paper [223], which significantly improved convergence rates.

Each coupling layer in the NF model had its own MLP to parameterize its 128-degree Bernstein polynomial. These MLPs consisted of a linear layer with an output dimension of 256, a GELU activation function, a

LayerNorm, and another linear layer with an output dimension of 128. It was observed that the use of weight normalization did not improve convergence in this context, so it was not used.

For each training step, we generated a new data batch, by sampling the parameters $\boldsymbol{\theta}$ from the prior specified in Table 5.1 and running the simulations to generate the corresponding oscillations and finally adding white noise. Each batch consisted of 1024 parameters-observation pairs and each epoch had 5000 training steps. The NPE model was optimized using Adam [153], with a learning rate of 0.01 for the first 90% of the epochs, and 0.001 for the last 10% of the epochs. Lowering the learning rate further gave minimal improvements so the cut-off was set at a learning rate of 0.001.

Gravitational waves setup

The context network is identical in setup compared to the one made for the coupled-harmonic oscillator, except that the dimensions are bigger. The all linear layer, except the last linear layer, has an output dimension of 4096. The last linear layer has an output dimension of 512. The setup of the NF model is the same for the coupled-harmonic oscillator.

For each training step, we generated a new data batch, by sampling the parameters $\boldsymbol{\theta}$ from the prior specified in Table 5.2. To improve the generation speed we generate only 64 waveforms, and each waveform is copied 16 times and gets new sky coordinates, polarization angle, SNR, and arrival time which are used to scale, and subsequently, project the waveform onto the HLV detectors. The waveforms are then whitened after which we add white noise.

Each batch consisted of 1024 parameters-observation pairs and each epoch was 5000 training steps. The NPE model was optimized using Adam [153], with a learning rate of 0.01 for the first 450 epochs, and 0.001 for the last 50 epochs. Lowering the learning rate further gave minimal improvements so the cut-off was set at a learning rate of 0.001.

For the sky coordinates we use the polar coordinates over the celestial coordinates removing the implicit dependence on Greenwich Mean Sidereal Time. Since we already have several theta's and phi's as notation we opted to use THETA and PHI as notation for these polar / sky coordinates.

5.3 Effective priors for NPE

The relationship between the size of the training dataset and NN performance remains a topic of ongoing research [224–226]. However, the general sentiment is that increasing the size of the dataset improves performance. Conversely, NNs do not perform well at inference time for input which it has not been sufficiently trained on. Neural simulation-based inference relies on training the model with simulated data originating from a chosen prior. Conventionally, one uses an uninformative prior to mirror the Bayesian inference framework. In this section, we argue that to train a robust and accurate NPE model one has to choose the prior such that the model trains on an as diverse set of samples as possible. In other words, the prior should be effective in training the NN.

We used the toy problem for all the experiments in this section. We simulated observations from the coupled-harmonic oscillator of two seconds at a sampling frequency of 128 Hz. The simulated time series is a mix of four sinusoids, whose frequencies f_v are proportional to $\sqrt{k/m}$. A change in mass does not translate into a linear response in frequency. It implies that with a uniform prior on m, there is more data with low frequencies. Ideally, for any sample drawn from the prior, the number of similar samples is roughly equal.

To quantify the similarity between time series we use the cosine similarity, also known as the match in GW astronomy. By keeping one sample as a constant argument and drawing the other from a chosen prior, we estimate the NN's exposure to the reference sample. Here, the analysis is limited to mass because the similarity between samples changes the most across this dimension, and is, therefore, the most troublesome to infer correctly. To determine the number of similar signals that the model sees as a function of the prior $p(m)$ we define the sample exposure as

$$\xi(m, \boldsymbol{\theta}_{rest}) = \mathbb{E}_{m' \sim p}\left[\frac{\boldsymbol{s}(m, \boldsymbol{\theta}_{rest})^T \boldsymbol{s}(m', \boldsymbol{\theta}_{rest})}{||\boldsymbol{s}(m, \boldsymbol{\theta}_{rest})||\,||\boldsymbol{s}(m', \boldsymbol{\theta}_{rest})||}\right], \tag{5.9}$$

where $\boldsymbol{\theta}_{rest}$ are the all parameters except the mass m. We approximated $\xi(m, \boldsymbol{\theta}_{rest})$ by sampling 1000 equidistant points from the inverse cumulative density function of $p(m)$. On the left side of Figure 5.3, we show the result

of this calculation for three different priors: a power law[3] with $\alpha = -3$, $\alpha = -1.5$, $\alpha = 0$. The right side shows the sample efficiency of the corresponding NPE models. The sample exposure seems to align well with the sample efficiency. It seems that NPE models demonstrate strong performance only when they have been exposed to a sufficient number of (similar) observations. If one desires a stable performance over the entire parameter space, it is critical that one chooses a prior that gives uniform sample exposure. If we switch from the power law prior back to a uniform prior during evaluation of the trained NPE models, these conclusions remain true, see Figure 5.4.

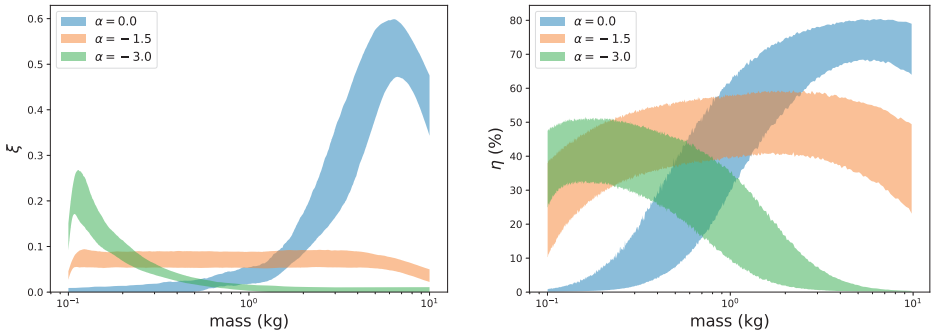


Figure 5.3: A comparison between the sample exposure for different priors and the sample efficiency of the corresponding NPE models. **Left.** The sample exposure at a specific mass for three different priors. The priors consist of a uniform prior (blue), a power law with an exponent of -1.5 (orange), and a power law with an exponent of -3.0. To cover the influence of the other parameters, we compute the sample exposure across the mass with 1000 different instances of $\boldsymbol{\theta}_{rest}$. The band shows the central 50% of computed sample exposures. **Right.** The sample efficiency for three NPE models trained with the three different priors, the shown band covers the central 50%. Although the sample exposure and sample efficiency do not match exactly, there is a clear correspondence between them.

[3]A power law distribution with power α is defined as $p_\alpha(x) = x^\alpha / A$ where A is a normalization constant.

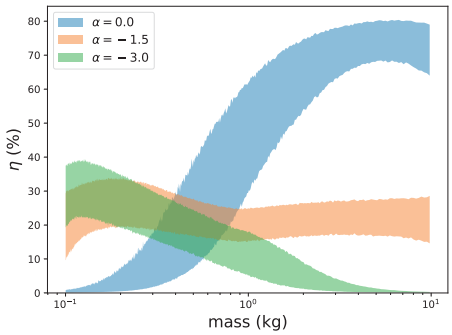


Figure 5.4: Sample efficiency of NPE models trained with different priors, but evaluated with posterior probabilities divided by prior probabilities (mimicking uniform prior). Colors indicate training priors: uniform (blue), power law -1.5 (orange), and power law -3.0. The NPE model trained with power law (-1.5) still recovers the posterior across the entire mass range.

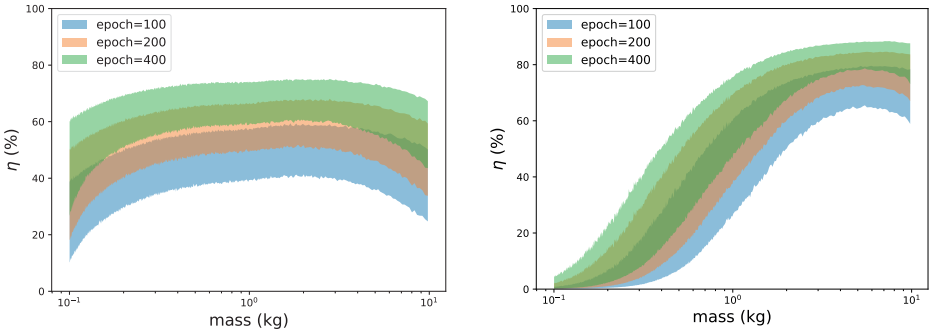


Figure 5.5: The sample efficiency for different training durations: 100 (blue), 200 (orange), and 400 (green) epochs. The shown bands cover the central 50%. **Left.** An NPE model trained with power law distribution ($\alpha = -1.5$) as mass prior. **Right.** An NPE model trained with uniform distribution as mass prior. Training for longer improves sample efficiency regardless of prior. Despite the improvements in sample efficiency, longer training does not give satisfactory performance for small masses when using a uniform prior. Thousands of epochs are probably needed to guarantee sufficient sample efficiency across the mass range. Only the power law distribution shows stable performance over the entire mass range regardless of training iteration.

To further validate the hypothesis that the model's performance improves with increased exposure to similar samples, we conducted a second experiment. Two NPE models, one with a uniform mass prior and the other with a power law prior of $\alpha = -1.5$, were trained for 100, 200, and 400 epochs[4]. Figure 5.5 presents the sampling efficiencies obtained from this extended training. The results demonstrate a significant enhancement in performance for both $\alpha = -1.5$ and $\alpha = 0$ as the training duration increases. This finding further supports the notion that the NPE model functions optimally when it has encountered a sufficient number of similar observations. However, it also indicates that training for longer has diminishing returns. In the next section, we propose a scheme to overcome this problem.

5.4 Fine-tuning neural posterior estimation

As demonstrated in the previous section, exposure to a diverse set of samples is necessary to ensure a strong NPE model. However, when dealing with an enormous parameter space, obtaining adequate exposure can require billions of samples. To accurately store the massive volume of information, the NPE model must become bigger and thus will be slower to train, requiring more sophisticated hardware. At a certain point, NPE will no longer be viable due to the training requirements. In this section, we address the complications arising from large parameter spaces and present a fine-tuning procedure designed to maintain high sample efficiency, regardless of the parameter space's scale.

5.4.1 Challenges in large parameter spaces

To illustrate the challenges posed by larger parameter spaces, we repeat the experiment of the previous section with a longer duration signal. The duration was changed from two to twenty seconds. This adjustment decreased the sample exposure by a factor of ten. To counteract the decrease in exposure, the NPE model needs to be trained ten times longer. Moreover, the NPE model was given a context network that was a factor ten wider than

[4]A single epoch is 5000 updates with a batch of 1024.

the original one. Despite these changes, the sample efficiency was only 0.02% after 100 epochs, and 0.07% after 1000 epochs. This is significantly worse than the 20–30% sample efficiency achieved in Section 5.3.

While the NPE model was still able to approximate the posterior distribution for the extended twenty-second signal (see Figure 5.6) its predictions were significantly wider than its importance-sampled counterpart. This widening suggests that the NPE model was not able to extract all the information from the signal, despite being trained on roughly 5 billion unique samples. Naturally, a signal with a longer duration contains more information, and therefore a tighter posterior distribution. This is reflected in the decreased sample exposure, but cannot account for the significant drop in performance. We attribute the lower performance to the inherent difficulty of accurately storing more and much higher-dimensional time series. Traditional methods can still find the correct posterior by running for longer. Altering amortized NPE such that iterative improvements post-prediction are possible might be the solution for large parameter space problems.

5.4.2 Fine-tuning procedure

As is evident from the last subsection, learning the posterior distributions for all possible events becomes increasingly harder as the parameter space grows or the sample exposure decreases. To circumvent these difficulties we propose switching back to a non-amortized setting after training the NPE model. From now on, we will refer to the optimization of a trained, amortized NPE model for a single observation $\boldsymbol{x}_{obs}$ as fine-tuning. Fine-tuning makes learning the posterior distribution more manageable for two reasons. First, the NPE model only needs to train on parameters that produce simulations resembling $\boldsymbol{x}_{obs}$. These can be sampled from the amortized model. Second, the NF only needs to learn a single posterior distribution it already roughly approximates. In summary, fine-tuning enables the NPE model to quickly learn the posterior distribution by being more sample-efficient and simplifying the objective.

We will now discuss the implementation of the fine-tuning, outlined in Algorithm 1. To switch from an amortized setting to a non-amortized

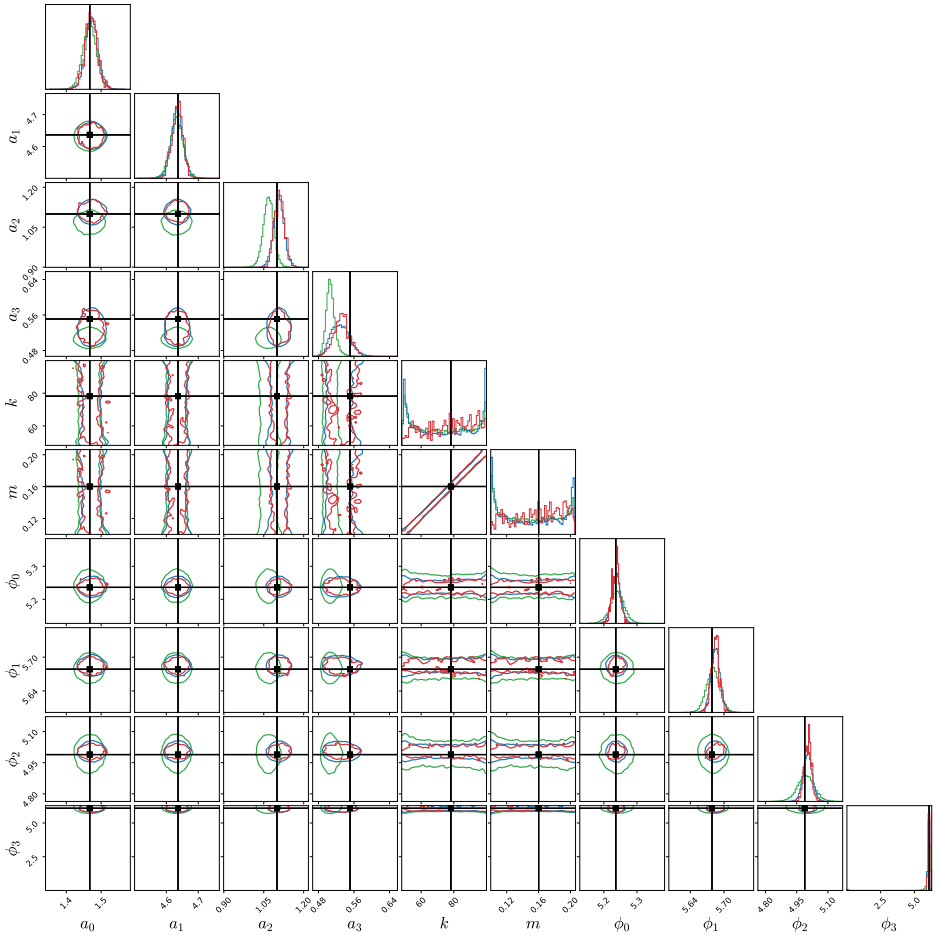


Figure 5.6: The predicted posterior distribution for 20 seconds long signal, shown as 1D histograms and 2D contour plots. There are three posterior predictions, the original NPE model (green), the fine-tuned NPE model (blue), and the importance-sampled posterior (red). The contours represent the 90% confidence area. The sample efficiency of the green posterior is 0.1%, and of the blue posterior it is 3.5%. The improvement was achieved in five seconds.

Algorithm 1 Fine-tune NPE model

Require: observation $\boldsymbol{x}_{obs}$, context model M, parameters of pre-trained NPE model $\boldsymbol{\psi}$

$\boldsymbol{c} \leftarrow M(\boldsymbol{x}_{obs})$ ▷ Generate the context vector

$\boldsymbol{\psi}' \leftarrow \boldsymbol{\psi}$

for i in 1..cycles **do**

$\boldsymbol{\theta}_i \sim R(\boldsymbol{\theta}|\boldsymbol{\psi}', \boldsymbol{c})$

$p_i \leftarrow p(\boldsymbol{x}_{obs}|\boldsymbol{\theta}_i)\, p(\boldsymbol{\theta}_i)$

for j in 1..10 **do**

$w_i \leftarrow p_i / R(\boldsymbol{\theta}_i|\boldsymbol{\psi}', \boldsymbol{c})$ ▷ No gradients are calculated

$L \leftarrow -w_i^2 \log(R(\boldsymbol{\theta}_i|\boldsymbol{\psi}', \boldsymbol{c}))$

$\boldsymbol{\psi}' \leftarrow \text{update}(L, \boldsymbol{\psi}')$ ▷ Update $\boldsymbol{\psi}'$ with Adam using gradient $\partial L / \partial \boldsymbol{\psi}'$

end for

end for

setting, the context vector $\boldsymbol{c}$ is calculated by passing observation $\boldsymbol{x}_{obs}$ through the context model and using it as a static condition for the NF model. For clarity, we define a new NPE model $R(\boldsymbol{\theta}|\boldsymbol{\psi}', \boldsymbol{c})$ whose parameters $\boldsymbol{\psi}'$ are initialized with the parameters $\boldsymbol{\psi}$ of $Q(\boldsymbol{c})$. The remainder of the fine-tuning procedure operates in three steps:

1. Generate samples $\boldsymbol{\theta}_i$ from distribution $R(\boldsymbol{\theta}|\boldsymbol{\psi}', \boldsymbol{c})$ and calculate the true posterior probability $p(\boldsymbol{\theta}_i|\boldsymbol{x}_{obs})$ by multiplying the Whittle likelihood $p(\boldsymbol{x}_{obs}|\boldsymbol{\theta}_i)$ and the prior $p(\boldsymbol{\theta}_i)$.

2. Calculate posterior probability of $\boldsymbol{\theta}_i$ under the NF model $R(\boldsymbol{\theta}_i|\boldsymbol{\psi}', \boldsymbol{c})$.

3. Update $\boldsymbol{\psi}'$ with the χ^2-divergence as loss function:

$$L(\boldsymbol{\theta}_i; \boldsymbol{x}_{obs}, \boldsymbol{\psi}') = -\left(\frac{p(\boldsymbol{\theta}_i|\boldsymbol{x}_{obs})}{R(\boldsymbol{\theta}_i|\boldsymbol{\psi}', \boldsymbol{c})}\right)^2 \log(R(\boldsymbol{\theta}_i|\boldsymbol{\psi}', \boldsymbol{c})) \tag{5.10}$$

The loss function, as introduced in reference [227], uses the square of the importance weight rather than the regular importance weight. This approach serves to minimize the variance of importance weights and

discourages the importance weights from becoming too big, leading to improved convergence and sample efficiency. During fine-tuning most of the time is consumed by running simulations to calculate the Whittle likelihood. By repeating steps (ii) and (iii) for the same samples generated in step (i), we can cut down on simulation time and still improve our model. In our experiments, we could repeat steps (ii) and (iii) at least ten times while still having a similar loss progression as without any repeated steps. We will refer to completing steps (i), (ii), and (iii) – including repetitions – as a cycle. For harder problems, more cycles, and samples, are needed to reliably converge to the correct posterior. Increasing the number of samples generated in step (i) has been sufficient to always find the correct posterior distribution, regardless of multi-modality or the quality of the initial posterior prediction. However, this does increase the time needed to fine-tune the model. As we will see in section 5.5 we can mitigate most issues with multimodalities by redefining R, saving a lot of time.

Fine-tuning has a close resemblance to sequential NPE methods [63, 228–230]. Both use self-sampling to generate samples and a (pseudo-)importance weight to update the model. However, sequential NPE models seem to shun the use of amortized models as initial priors and use their own likelihood estimates as a replacement for the true likelihood. The importance ratio is then calculated between sequential iterations of the model, potentially requiring many rounds to converge. Moreover, without using an amortized model, the initial sample quality can be poor, potentially missing part of the posterior due to strong non-convex likelihood landscapes. Or requiring long run-times to explore the parameter space. All of these issues are mitigated by using fine-tuning. To our knowledge, this is the first time amortized and non-amortized SBI have been combined.

The results of our fine-tuning procedure, depicted in Figure 5.6, underline its ability to improve the sample efficiency of NPE models. The green area represents the posterior predicted by the original NPE model, while the blue area represents those predicted by the fine-tuned NPE model, and the red area depicts the true posterior, derived through importance sampling of the fine-tuned distribution. The fine-tuning was performed for 10 cycles, with a batch size of 10240, 10 repetitions, and finished within five seconds. Fine-tuning brought the sample efficiency from 0.1% to 3.5%,

sufficient to extract the true posterior with importance sampling. To reach higher sample efficiencies we need to add more NF layers to the model, as we will see in the next subsection.

5.4.3 Optimizations for Fine-tuning Performance

Our investigation into improving the fine-tuning procedure led us to explore two key aspects of the NPE model architecture: the method of conditioning coupling layers and the addition of extra normalizing flow layers.

Conditioning Methods

There are several straightforward ways to condition the coupling layers on the output of the context network. Perhaps the easiest is to concatenate the context vector and the static half $\boldsymbol{b}_j^l$ and feed the new vector to the MLP of the coupling layer. Slightly more involved methods transform the context vector, via a linear transformation, into a bias vector, a scaling vector, or a vector followed by a sigmoid function. To evaluate these methods, we conducted experiments using our oscillator toy model with a duration of 2 seconds, training for 5 epochs. Table 5.3 summarizes the performance of these different conditioning methods.

Table 5.3: The performance of NPE models with different forms of conditioning. They were trained for 5 epochs on the toy problem, with a duration of 2 seconds. The percentages represent sampling efficiency before and after fine-tuning.

Conditioning	Training loss	Pre fine-tuning (%)	Post fine-tuning (%)
Concatenate	-11.5	0.76	39.1
Bias	-11.7	0.76	38.9
Scale	-11.2	0.60	47.5
Sigmoid	-11.5	0.76	41.2

As we can see in Table 5.3, while concatenation works well for training, it is not optimal for fine-tuning. Of the conditioning methods tested, the scaling vector proved to be the most effective for fine-tuning performance.

Consequently, we adopted the scale method of conditioning throughout this chapter.

Additional Normalizing Flow Layers

Fine-tuning allows us to add NF layers after the amortized training. These additional layers do not need conditioning and can improve the flexibility of the NPE model. Since these NF layers are initialized to approximate the identity function, they should not significantly alter the initial output.

To test the impact of additional layers, we used our NPE model trained for 100 epochs on coupled-harmonic oscillators with signals of 20 seconds. Table 5.4 demonstrates the effect of increasing the number of coupling layers on sample efficiency after fine-tuning.

Table 5.4: The results of fine-tuning for with additional layers.

Additional coupling layer	After fine-tuning (%)
0	2.1
1	4.4
2	7.9
4	8.3
8	7.9

As evident from Table 5.4, adding coupling layers significantly improves the sample efficiency, with the best performance achieved with 4 additional layers, increasing efficiency to 8.3% after fine-tuning. This approach allows us to enhance the model's capacity without requiring retraining of the entire network.

These architectural optimizations play a crucial role in improving the performance of our fine-tuning procedure, enabling more efficient and accurate posterior estimation. By carefully considering both the conditioning method and the number of additional flow layers, we were able to substantially enhance the capabilities of our NPE model in the context of fine-tuning.

5.5 Gravitational waves

Posterior inference for GW events via NPE is possible for BBH events with chirp masses above 15 solar masses [188]. However, extending SBI models to low-mass BBH events has proven challenging. Here, we will show that by adapting an effective prior and by fine-tuning the NPE model for given events, it becomes possible to accurately infer posteriors for BBH events with chirp masses between 5 and 100 solar masses. This section is structured as follows: first, we discuss the choice of prior in gravitational wave inference. Second, the data generation and preprocessing steps are discussed. Third, we discuss the incorporation of symmetry relations into the fine-tuning procedure to ensure all modes of the posterior are found. Finally, we present and discuss the inference results for simulated, non-precessing BBH GW events with a chirp mass between 5 and 100 solar masses.

5.5.1 Effective priors for gravitational waves

Previous works in machine learning for gravitational wave inference commonly adopt either uniform priors for chirp mass and mass ratio or uniform priors for the component masses [56, 57]. As shown in section 5.3, NPE model performance matches the sample exposure caused by the choice of prior. In the left graph of Figure 5.7, we show the sample exposure as a function of chirp mass. By switching from a uniform prior to a power law with $\alpha = -3$, the sample exposure is evenly distributed across the chirp mass range. To put the difference in GW similarity into context: the average match between gravitational waves with chirp masses of 5.000 and 5.025 equals the average match between gravitational waves with chirp masses of 60 and 90. A similar analysis can be performed for the mass ratio; the results are shown in the right graph of Figure 5.7. While a power law as prior may not result in a uniform sample exposure, significant improvement is achieved by choosing a power law with $\alpha = -1.5$. To improve sample diversity during training, we selected power laws with $\alpha = -3$ for chirp mass and $\alpha = -1.5$ for mass ratio.

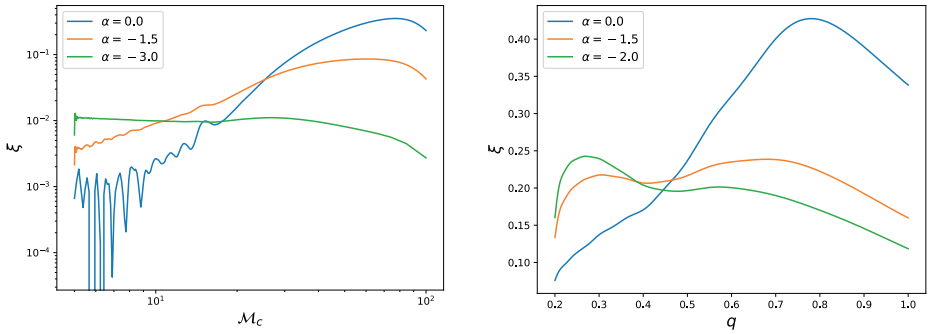


Figure 5.7: The mean sample exposure as a function of the chirp mass $\mathcal{M}_c$ and mass ratio q. Switching from a uniform prior to a power law distribution improves sample exposure for both the chirp mass and the mass ratio.

5.5.2 Data generation

Table 5.2 specifies the full prior used for parameter sampling. For waveform generation, we used the IMRPhenomXAS waveform model [231] provided by the ripple library [50], enabling GPU-based waveform generation. The training waveforms were generated in the frequency domain between 20 and 256 Hz and with a duration of 24 seconds. The selected frequency range was chosen to optimize data generation and reduce memory burden during training. This range, while not covering the entire frequency span of low-mass BBH mergers, suffices for training the NPE model to capture rough posteriors. During the fine-tuning procedure, we generate waveforms in the 20 to 2048 Hz frequency band to ensure that the model converges to the correct posterior. To speed up data generation further, we use each generated waveform eight times, with each use featuring a new sky position, signal-to-noise ratio (SNR), and polarization angle. To ensure that detectable signals are given to the NPE model, we sample the optimal SNR from a uniform distribution between 10 and 30 and scale the luminosity distance to match the sampled SNR. All waveforms were whitened with the design sensitivity power spectral densities of the HLV detectors [27, 28, 130]. These steps allowed us to quickly and continuously generate parameter-strain pairs during training to prevent overfitting.

5.5.3 Reduced-order basis

The NPE model trained on the generated data remains the same as specified in Section 5.2. However, its input is not the raw frequency series, but the frequency series projected on a reduced-order basis (ROB). This approach creates a lower-dimensional approximation of the high-dimensional gravitational waveform data, significantly reducing computational complexity while preserving essential features of the signal.

While building an ROB for GWs is regularly performed with singular value decomposition (SVD) [232], our approach utilizes the covariance matrix and its eigendecomposition. This choice was made to accommodate the large number of samples that are required to guarantee strong coverage. Computing the covariance matrix and its eigendecomposition consumes constant memory with respect to the number of samples. Consequently, the ROB can be constructed with as many samples as necessary to achieve sufficient coverage.

Our ROB was made by calculating, per detector, the eigenbasis of the covariance matrix over five million simulated strains and taking the first 768 eigenvectors, which were necessary to reach a minimal match of 0.95 when tested on a million samples. This dimensionality reduction serves two important purposes. First, it provides rudimentary denoising of the frequency series; the ROB preserves at least 95% of the signal content while significantly reducing dimensionality, effectively increasing the signal-to-noise ratio as the noise becomes distributed across fewer dimensions. Second, the dominant eigenvectors likely correlate with high-impact parameters such as chirp mass, making these crucial features more accessible to the NPE model. This not only increases the convergence speed of the NPE model but also reduces the computational resources needed, as the network no longer needs to learn these features from raw data.

5.5.4 Fine-tuning for gravitational wave inference

The fine-tuning procedure is a slightly augmented version of Algorithm 1 – we redefine our $R(\boldsymbol{\theta}|\boldsymbol{\psi}', \boldsymbol{c})$ to incorporate the potential symmetries in the polarization-phase (ψ-ϕ_c) plane. For each sample drawn from $R(\boldsymbol{\theta}|\boldsymbol{\psi}', \boldsymbol{c})$, three additional copies are introduced, each shifted by π in ϕ_c and/or 0.5π

in ψ to encompass all four potential modalities. To reflect this symmetry in R, we average the likelihood of the four samples, assigning this average likelihood to all four instances. This approach safeguards against missing modes due to unfortunate sampling or inaccuracies in predictions from the amortized NPE model.

As already shown in Section 5.4.3 adding additional flow layers before fine-tuning improves the performance of the model. From our experience, the effect is not as pronounced for GWs, however it is still a positive effect.

We fine-tune the NPE model for 20 cycles. In each of the first 10 cycles, we generate 100000 strains with a frequency range spanning from 20 to 256 Hz. In these cycles, the initial rough posterior concentrates its probability mass in the correct parts of the parameter space but does not necessarily match the true posterior perfectly. In each of the remaining cycles, we generate 50000 strains with a frequency range spanning from 20 to 2048 Hz. In this second phase, the likelihood contributions of the higher frequencies correct the posterior prediction where needed. Afterward, the NPE model generates samples, which are importance-weighted, until 100000 samples are generated or the effective sample size reaches 5000. The entire fine-tuning procedure takes 10 minutes at most on an NVIDIA GeForce RTX 3090, including model loading, JAX compilation, and the importance sampling after fine-tuning.

The fine-tuned posteriors often closely match the importance-sampled ones. In Figure 5.8, we can see that the fine-tuned posterior (blue) closely aligns with the importance-sampled posterior (red). Despite the challenging characteristics of this event—featuring a low chirp mass, mass ratio, and high multimodality—the fine-tuned NPE model accurately captures the posterior distribution. Notably, just 10 minutes of fine-tuning results in a significant increase in sample efficiency, increasing from 0.00249% to 51.2%. We see similar performance across the entire parameter space. An example of a low-sample efficiency posterior is shown in Figure 5.9.

To compare with Bayesian inference methods a posterior distribution was inferred with nested sampling. For a fair comparison, the nested sampling algorithm was implemented in JAX to have access to GPU waveform generation. The implementation is based on RADFRIENDS [42], due to the ease of implementation and robustness of its results. The nested sampling

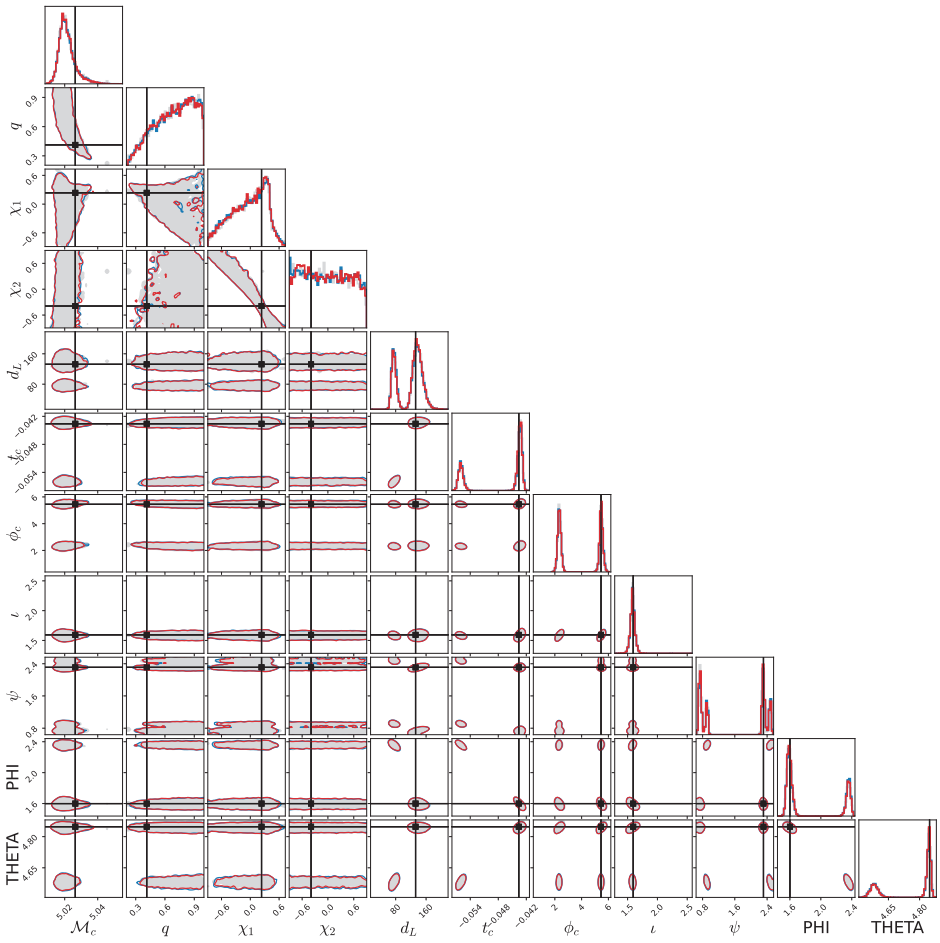


Figure 5.8: The predicted posterior distribution for a low-mass O3 signal, shown as 1D histograms and 2D contour plots. The contours in blue depict the predictions of the fine-tuned NPE model and the contours in red the importance-sampled posterior distribution. The grey mass is the posterior distribution obtained via nested sampling, for easy comparison we choose to use a filled contour. The NPE posterior matches the nested sampling posterior quite well. Moreover, despite the many modes in the posterior distribution, the fine-tuning procedure is still able to find all of them. The sample efficiency of amortized NPE and after fine-tuning NPE for this event differs by a factor of 20000 (0.0025% vs 51.2%).

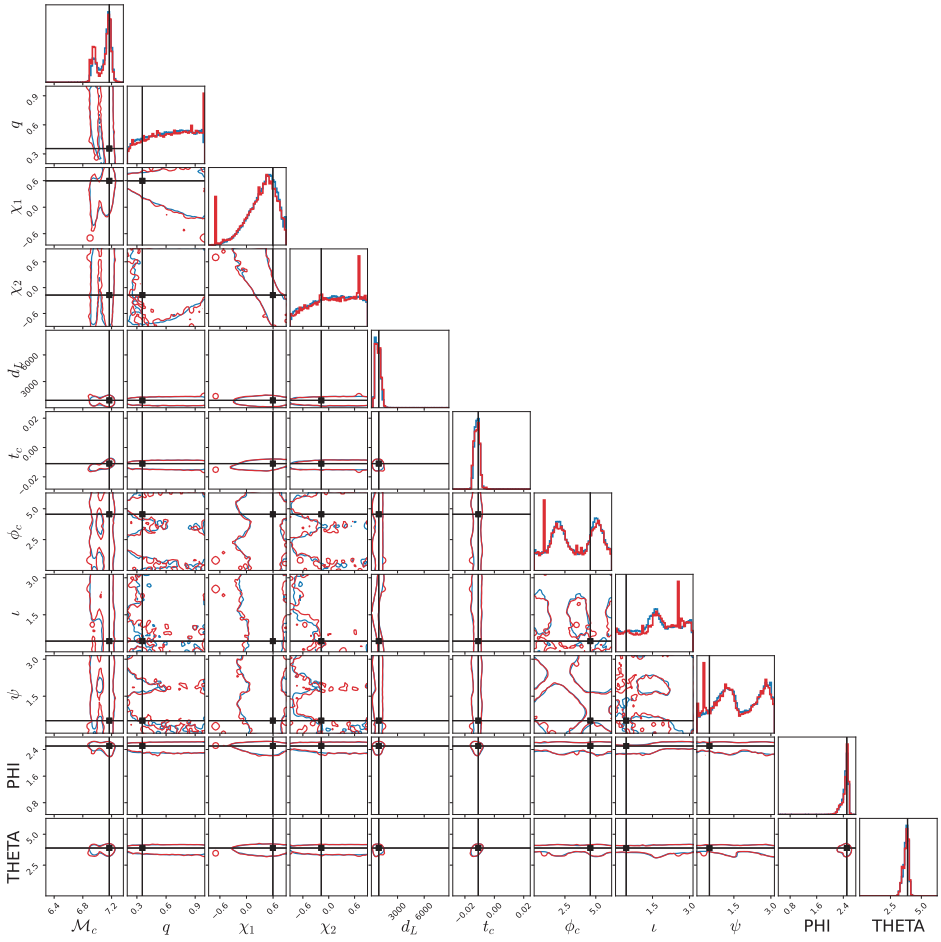


Figure 5.9: The predicted posterior distribution for a low-mass O3 signal, shown as 1D histograms and 2D contour plots. The contours in blue depict the predictions of the fine-tuned NPE model and the contours in red the importance-sampled posterior distribution. The sample efficiency of the fine-tuned posterior is 0.79%. Although it seems to have found the posterior distribution, for a single sample the ratio between actual and assigned likelihood is massive, resulting in a low sample efficiency. This sample shows up in the 1D histograms as a sharp peak in red.

posterior, shown as a filled grey contour in Figure 5.8, closely aligns with the importance sampled posterior. However, the time to compute the posterior distribution with nested sampling is more than 3 days, 400 times longer than our fine-tuning algorithm. To be fair, RADFRIENDS is not the most time-efficient nested sampling implementation and we expect more sophisticated implementations can complete the posterior inference within a day.

It is important to acknowledge that all inference methods face inherent challenges in complex parameter spaces. The accuracy of nested sampling depends on appropriate prior selection and sufficient sampling density, especially for multimodal distributions like those in GW inference. Similarly, as we have shown in Sections 5.3 and 5.4, our NPE approach requires careful consideration of effective priors and fine-tuning parameters. Nevertheless, our comparison demonstrates that the fine-tuned NPE approach offers a favorable balance between accuracy and computational efficiency for gravitational wave inference.

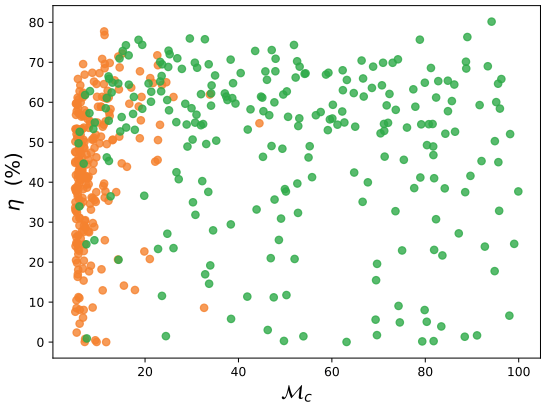


Figure 5.10: The sample efficiency after fine-tuning and chirp mass over 500 simulated GW events. Half of these events were sampled from a power law (orange), and the other 250 events were sampled from a uniform distribution (green). Across the entire chirp mass range we achieve strong performance. The near-zero sampling efficiencies are due to large importance weights that dominate the sampling efficiency or not-yet fully converged posteriors.

To quantify the NPE model performance after fine-tuning, we simulated 500 GW events and predicted their posterior distributions by fine-tuning the NPE model for the events. The corresponding sampling efficiencies and chirp masses are depicted in Figure 5.10. For half of the 500 GW events, the chirp mass was drawn from a power-law distribution (shown in orange), while for the remaining 250 events, the chirp mass was drawn from a uniform distribution (depicted in green). Across the 500 events, 14 events exhibited a sample efficiency below 5%. For 11 out of the 14 events, the low sample efficiency can be attributed to the NPE model assigning a low likelihood to a high likelihood sample, reducing the sampling efficiency. The loss of the remaining 3 events did not converge within the two rounds and required an additional ten cycles for convergence. These events can easily be identified by a high percentage of near-zero importance weights.

5.6 Conclusion

In conclusion, our investigation into GW inference using NPE models has yielded promising results that advance the capabilities of SBI methods. The performance of NPE models appears to align closely with the sample exposure, stressing the importance of prior selection. Moreover, our fine-tuning approach proves pivotal in overcoming the inherent limitations of amortized NPE models, providing a pathway to accurate inference for low-mass BBH posteriors. While acknowledging that all inference methods involve inherent trade-offs between computational efficiency, accuracy, and robustness, our approach offers a balanced solution that addresses many limitations of traditional methods. Although our primary focus is on GW inference, we believe that our findings may prove fruitful in other research areas.

Looking ahead, we see many avenues for further improvement in fine-tuning for GW events. The implementation of adaptive stopping mechanisms holds promise to enhance convergence speed, allowing us to monitor loss or adjust frequency ranges based on initial chirp mass estimates. Differentiable waveforms enable us to use score matching to reduce the number of cycles required during fine-tuning, saving even more valuable time. Additionally, the likelihood function can be chosen

at the start of fine-tuning, removing problems of unseen power spectral densities, or changing different waveform models after convergence. As part of future work, we also aim to explore even longer signal durations and to go to even lower chirp masses by considering NSBH or BNS events.

Chapter 6

Conclusion

GW astronomy has made remarkable strides since the first detection in 2015. However, as detector sensitivity improves and more events are observed, the field faces significant challenges. The computational demands for analyzing GW signals are increasing dramatically due to longer in-band durations and higher event rates. Traditional Bayesian inference methods, while accurate, are becoming prohibitively slow for the volume of data expected from future detectors. This thesis aimed to address these challenges by developing rapid and reliable parameter estimation methods for GWs using machine learning techniques.

Key Findings and Contributions

In Chapter 3, we introduced a novel approach to GW sky localization that combines deep learning with importance sampling. The key innovation lies in using a multi-headed convolutional neural network to parameterize simple distributions for the sky location and masses of binary black hole systems. These distributions serve as proposal distributions for importance sampling, providing a good initial estimate of the sky position. The importance sampling step then refines this estimate, converging to the true Bayesian posterior. This method combines the speed of neural networks with the potential to achieve the accuracy of traditional Bayesian methods. By using importance sampling, we can also quantify the reliability of the predictions of the NN, flagging cases where the network might be underperforming. Our approach demonstrated the ability to generate sky maps in minutes rather than hours, while generating sky maps that resembled those generated by bilby. While the method showed promise on simulated data, its performance was limited by the relatively inflexible neural architecture used. The impact of this work lies in its potential to

enable rapid sky localization for multi-messenger astronomy, facilitating prompt follow-up observations of GW events.

Chapter 4 addressed the challenge of analyzing overlapping GW signals, a scenario expected to become common with next-generation detectors. Our key contribution was demonstrating that posterior inference for overlapping GWs is possible using CCNFs. The main results of this work are twofold. First, we showed that CCNFs can successfully analyze two overlapping binary black hole signals within seconds, a task that takes weeks with traditional Bayesian methods. Second, and perhaps more importantly, we found that our method produces well-calibrated posteriors, avoiding some of the biases observed in regular Bayesian inference when dealing with overlapping signals. The impact of this work lies in its potential application to future detectors with higher event rates, where overlapping signals will be commonplace. While our demonstration was limited to simulated data and specific scenarios – namely, two overlapping high-mass binary black hole signals – it opens up new avenues for tackling the challenge of overlapping signals in GW astronomy. The speed and scalability of our approach suggest that it could be extended to handle multiple overlapping signals of various types, a crucial capability for future GW data analysis.

In Chapter 5, we tackled the challenge of improving the performance of NPE for GW analysis, with a focus on low-mass binary black hole systems. Our key innovations were twofold: the use of effective priors to improve training efficiency, and a fine-tuning procedure to enhance performance for individual events. Recognizing that neural networks learn by example, we developed the concept of effective priors to tackle the crucial challenge of achieving sufficiently high sample efficiency in NPE models. We demonstrated that the performance of NPE models correlates strongly with the sample exposure during training, which is directly influenced by the choice of prior distribution. Traditional uninformative priors, while theoretically sound, can lead to suboptimal performance in practice due to uneven sampling across the parameter space. We introduced the notion of sample exposure, a metric quantifying how often the model encounters similar samples during training. By carefully choosing priors that provide uniform sample exposure across the parameter space, we significantly

improved the model's ability to estimate posteriors accurately for a wide range of signals. This approach is particularly beneficial for GW analysis, where the parameter space is vast and signals can vary greatly in their characteristics. For instance, we found that using a power-law prior for the chirp mass, rather than a uniform prior, led to more consistent performance across the mass range. This insight not only improved our model's performance but also provides a general principle for training NPE models in other domains with complex parameter spaces. The effective prior approach ensures that the model receives adequate training across all regions of the parameter space, leading to more robust and reliable posterior estimates.

To complement the effective priors approach, we developed a novel fine-tuning procedure that allows for rapid optimization of a pre-trained NPE model for a specific GW event. This method addresses a fundamental challenge in amortized inference: the difficulty of achieving high precision across the entire parameter space. Amortized models, tasked with learning to generate posteriors for all possible events, often produce overly conservative estimates as a compromise. Our fine-tuning procedure overcomes this limitation by allowing the model to adapt to the specific characteristics of a given event. The procedure works by using the pre-trained model's output as a proposal distribution for importance sampling, then iteratively refining the model based on the importance weights. This approach combines the advantages of amortized inference (rapid initial estimates) with the precision of event-specific optimization, all while maintaining computational efficiency. The main result of this work is a dramatic improvement in the analysis of low-mass binary black hole events, a regime that has been challenging for previous machine learning approaches due to the longer duration and higher complexity of these signals. We achieved sample efficiencies of up to 80% for events with chirp masses as low as 5 solar masses, representing a significant advance in the application of machine learning techniques to GW parameter estimation. This high sample efficiency translates to more accurate and reliable posterior distributions, crucial for precise astrophysical inference.

The fine-tuning procedure we introduced extends the model's capabilities, allowing it to accommodate specific noise patterns and evolving

detector characteristics. This adaptability is a key strength, positioning the method well for future real-world applications. While our current results are based on simulated data, the demonstrated rapid adaptation suggests significant potential for analyzing real GW signals. The possible reduction in parameter estimation time from days to minutes could transform our approach to GW astronomy, enabling swift responses to detected events and accelerating scientific discovery.

Final Thoughts

As we look to the future of GW astronomy, it is clear that data analysis techniques must evolve in tandem with instrumental advances. Future GW analysis will likely rely on neural models to obtain initial posterior estimates. However, these will be initial estimates only, as neural networks are unlikely to learn the full parameter space of future signals to sufficient accuracy, especially for overlapping signals or events with unusual characteristics.

To refine initial estimates into final, accurate posteriors, techniques like importance sampling or fine-tuning will be crucial. These methods can combine the speed of neural networks for rapid initial estimates with the precision required for rigorous scientific inference. As the field progresses, we anticipate a hybrid approach where machine learning and traditional methods complement each other. This synergy will be essential in handling the increased complexity and volume of data from future detectors, enabling real-time analysis for multi-messenger astronomy, and facilitating comprehensive population studies.

Summary

This thesis addresses a critical challenge in gravitational wave astronomy: developing efficient methods to analyze the rapidly increasing volume of data from gravitational wave detectors. As detector sensitivity improves and future observatories come online, traditional analysis methods will struggle to keep pace with the volume and complexity of observations. The research presented here explores machine learning approaches to significantly accelerate gravitational wave parameter estimation while maintaining the reliability of traditional Bayesian methods.

Gravitational waves, ripples in spacetime predicted by Einstein's theory of general relativity, were first directly detected in 2015 by the LIGO-Virgo collaboration. This landmark discovery opened a new window to observe our universe, enabling us to study binary black hole and neutron star mergers. Analysis of these signals provides insights into astrophysics, fundamental physics, and cosmology. However, each merger event requires intensive computational analysis to extract the physical parameters of the source, such as masses, spins, and sky location.

The current analysis pipeline relies on Bayesian inference methods that, while accurate, can take hours to weeks to process a single event. This approach becomes unsustainable with the projected detection rates of at least thousands of events per year from future observatories. This thesis tackles this computational bottleneck through three innovative approaches using machine learning.

First, we developed a method for rapid sky localization of gravitational wave sources by combining deep learning with importance sampling. Using a convolutional neural network to generate approximate distributions of sky location and source masses, followed by importance sampling to refine these estimates, we demonstrated the ability to produce accurate sky maps within minutes rather than hours.

Second, we addressed the challenge of overlapping gravitational wave

signals, which will become common in future detectors. Using continuous conditional normalizing flows, we created a framework that can analyze two overlapping binary black hole signals in seconds rather than weeks. This method not only offers tremendous speed improvements but also produces well-calibrated posteriors that avoid some of the biases observed in traditional methods when signals overlap.

Third, we tackled the fundamental limitations of neural posterior estimation through two key innovations: effective priors and a fine-tuning procedure. By recognizing that neural networks learn from examples, we showed that carefully choosing training priors to provide uniform sample exposure across the parameter space significantly improves performance. We then developed a novel fine-tuning procedure that rapidly optimizes a pre-trained model for specific events, achieving high sample efficiencies for low-mass binary black hole systems — a regime previously challenging for machine learning approaches.

The methods developed in this thesis demonstrate that machine learning can dramatically accelerate gravitational wave parameter estimation without sacrificing reliability. By reducing analysis times from days to minutes, these approaches could transform how we respond to gravitational wave events and analyze large populations of sources. As detector sensitivity improves and event rates increase, we envision a hybrid approach where neural networks provide rapid initial estimates, refined through techniques like importance sampling or fine-tuning to achieve the precision required for scientific inference.

This research represents a significant step toward addressing the computational challenges of next-generation gravitational wave astronomy, helping to ensure that our analysis capabilities keep pace with the remarkable advances in detector technology. The methods developed here show promise not only for gravitational wave astronomy but potentially for other fields facing similar challenges in inverse problems and high-dimensional parameter estimation.

Samenvatting

Dit proefschrift behandelt een cruciale uitdaging in de zwaartekrachtsgolfastronomie: het ontwikkelen van efficiënte methoden om de snel toenemende hoeveelheid data van zwaartekrachtsgolfdetectoren te kunnen analyseren. Naarmate de gevoeligheid van detectoren verbetert en toekomstige observatoria operationeel worden, zullen traditionele analysemethoden moeite hebben om gelijke tred te houden met het volume en de complexiteit van de waarnemingen. Het hier gepresenteerde onderzoek verkent *machine learning*-benaderingen om de parameterschatting van zwaartekrachtsgolven aanzienlijk te versnellen, terwijl de betrouwbaarheid van traditionele Bayesiaanse methoden behouden blijft.

Zwaartekrachtsgolven, rimpelingen in de ruimtetijd die Einstein met zijn algemene relativiteitstheorie voorspelde, werden in 2015 voor het eerst rechtstreeks gedetecteerd door de LIGO-Virgo-samenwerking. Deze baanbrekende ontdekking bood nieuwe mogelijkheden om ons universum te observeren: we kunnen bijvoorbeeld nu versmeltingen van binaire zwarte gaten en neutronensterren waarnemen. Analyse van deze signalen biedt nieuwe inzichten in de astrofysica, fundamentele natuurkunde, en kosmologie. Elke observatie vereist echter gigantisch veel computationele rekenkracht om de fysieke parameters van de bron te extraheren, zoals massa's, spins en de positie aan de hemel.

De huidige analyses zijn afhankelijk van Bayesiaanse inferentiemethoden die, hoewel ze nauwkeurig zijn, uren tot weken nodig hebben om een enkele waarneming te verwerken. Deze aanpakken zijn niet meer werkbaar als er duizenden zwaartekrachtsgolven per jaar waargenomen gaan worden. Dit proefschrift pakt dit computationele knelpunt aan en presenteert hiervoor drie innovatieve machine learning-benaderingen.

De eerste is een methode voor snelle lokalisatie van zwaartekrachtsgolfbronnen die *deep learning* combineert met *importance sampling*. Door gebruik te maken van een convolutioneel neuraal netwerk dat distributies

van de hemellocatie en bronmassa's genereert, gevolgd door importance sampling om deze schattingen te verfijnen, hebben we aangetoond dat nauwkeurige hemelkaarten binnen minuten in plaats van uren kunnen worden geproduceerd.

Vervolgens gingen we aan de slag met de uitdaging overlappende zwaartekrachtgolfsignalen te kunnen analyseren, die in de toekomst steeds vaker waargenomen zullen worden. Met behulp van *continuous conditional normalizing flows* creëerden we een methode die twee overlappende binaire zwarte gat-signalen in seconden in plaats van weken kan analyseren. Deze methode produceert daarnaast ook goed gekalibreerde posterior-distributies die sommige vertekeningen die in traditionele methoden worden waargenomen vermijdt.

Ten derde tackelden we de fundamentele beperkingen van *neural posterior estimation* door twee belangrijke innovaties: *effective priors* en een finetuning-procedure. We toonden aan dat zorgvuldige keuze van training priors de analyses aanzienlijk verbetert. Vervolgens ontwikkelden we een nieuwe finetuning-procedure die een getraind model snel optimaliseert voor specifieke waarnemingen, waardoor de parameterschatting van binaire zwarte gat-systemen (met een lage massa) snel, accuruut, en betrouwbaar zijn.

De methoden die in dit proefschrift zijn ontwikkeld, tonen aan dat machine learning de parameterschatting van zwaartekrachtsgolven drastisch kan versnellen zonder aan betrouwbaarheid in te boeten. We presenteren een hybride aanpak waarbij neurale netwerken snelle initiële schattingen leveren, die door technieken zoals importance sampling of finetuning verfijnd worden om de precisie te bereiken die nodig is voor wetenschappelijke inferentie.

Dit onderzoek vertegenwoordigt een belangrijke stap in het aanpakken van de computationele uitdagingen van de volgende generatie zwaartekrachtsgolfastronomie, waardoor we kunnen garanderen dat onze analysecapaciteiten gelijke tred houden met de grote vooruitgang in detectortechnologie. De hier ontwikkelde methoden zijn niet alleen veelbelovend voor zwaartekrachtsgolfastronomie, maar potentieel ook voor andere vakgebieden die voor vergelijkbare uitdagingen met inverse problemen en hoogdimensionale parameterschatting staan.

Contributions

The following paragraphs outline my specific contributions to each of the main research chapters, as well as additional publications resulting from collaborative work during my studies.

Chapter 3
For this study, I developed the multi-headed convolutional neural network architecture and implemented the importance sampling scheme. I conducted all the experiments, including the comparison with Bilby, and wrote the majority of the manuscript with input and edits from co-authors.

Chapter 4
For this study, I developed the majority of the codebase. For the experiments, I assisted with the training and evaluation of the model, and helped write and edit the manuscript.

Chapter 5
I introduced the idea of using effective priors for neural posterior estimation, developed and implemented the fine-tuning procedure, and conducted all experiments. I wrote the majority of the manuscript with feedback and edits from co-authors.

Other Publications
Throughout my studies, I also contributed to several collaborative research projects, resulting in the following publications:

1. Straalen, W., Kolmus, A., Janquart, J., Van Den Broeck, C., "Pre-Merger Detection and Characterization of Inspiraling Binary Neutron Stars Derived from Neural Posterior Estimation". Under review at Physical Review D (2024).

2. Vlijmen, D., Kolmus, A., Liu, Z., Zhao, Z., Larson, M., “Generative Poisoning Using Random Discriminators”. Accepted at the Responsible Computer Vision Workshop, ECCV (2022).

3. Liu, Z., Zhao, Z., Kolmus, A., Berns, T., van Laarhoven, T., Heskes, T., Larson, M., “Going Grayscale: The Road to Understanding and Improving Unlearnable Examples”. arXiv preprint arXiv:2111.13244 (2021).

Research Data Management

This thesis research has been carried out under the research data management policy of the Institute for Computing and Information Science of Radboud University, The Netherlands.

The following research datasets have been produced during this PhD research:

- The code for Chapter 3 is available at:
  https://gitlab.science.ru.nl/akolmus/swiftsky
- The code for Chapter 4 is available at:
  https://gitlab.science.ru.nl/akolmus/overlapping
- The code for Chapter 5 is available at:
  https://gitlab.science.ru.nl/akolmus/tuning

Acknowledgements

Time seemingly weaves ever onward, and so I have arrived at the last part of this journey. One I could not have completed – nor enjoyed as much – without the support of many.

I would first like to express my gratitude to my supervisors Tom and Twan. Tom, dankjewel voor je vertrouwen en geduld, of het nu ging om een nieuw onderzoeksidee, vastgelopen schrijfwerk, of het simpelweg ordenen van gedachtes. Je scherpzinnigheid, humor, en kalmte hebben deze reis niet alleen mogelijk maar ook aangenamer gemaakt. Twan, dankjewel voor je altijd openstaande deur – het liefst wel later op de dag – , je vermogen om door mijn soms cryptische gedachtegangen heen te prikken en er meteen op voort te borduren, en voor de rustgevende, authentieke sfeer die je met je meebrengt.

A significant part of my research journey was shaped by my collaboration with the Gravitational Wave group in Utrecht. Their guidance provided the foundation of my gravitational wave knowledge, and our weekly meetings became a cornerstone of my PhD experience. I am particularly grateful to Justin, whose vast knowledge of astrophysics and boundless enthusiasm have been incredibly valuable. Justin, je onbegrensde energie voor wetenschap was aanstekelijk, en onze gesprekken die vaak van onderzoek naar het leven zwenkten behoorden tot de hoogtepunten van deze jaren. De kruisbestuiving van onze ideeën heeft niet alleen tot nieuwe inzichten geleid, maar heeft het onderzoek ook ferm plezanter gemaakt. Tomek, thank you for your impressive dedication to work-life balance and dry wit. As you approach the finish line of your own PhD journey, I wish you all the best for a successful completion – you are almost there! Luca, our mad dash through Glasgow airport is permanently etched in my memory, and I wish you the best in your PhD journey. My gratitude extends to Greg, Melissa, Thibeau, Peter, Sarah, and Chris for their support and contributions to my research and PhD experience.

My homebase, the Data Science research department at iCIS, has been a place where I found both academic stimulation and a warm, open atmosphere. Despite the pandemic limiting my physical presence, and my tendency to work from home most days, whenever I did come in, I would inevitably end up chatting for hours and leaving much later than planned. I am particularly grateful to Charlotte, Gabriel, Iordan, Jelle, Mirthe, Nik, Olivier, Norman, Parisa, Paulus, Roel, and Zhuoran for countless conversations that ranged from technical problem-solving to life's broader questions. A special mention goes to our Temptation Island reading group (whose members shall remain anonymous) which provided levity and a delightful amount of deep discussion. I also cherish the DnD sessions with Erkan, Gabriel, Jelle, Janneke, and Kasper. To everyone in the department who contributed to creating such a welcoming and stimulating environment – thank you.

Zhuoran and Martha, thank you for introducing me to unlearnable examples and for welcoming me into your research adventure. What started as casual discussions evolved into co-supervising students, papers, and most importantly, an enjoyable research collaboration. Your openness to exploring these questions together provided an interesting reprieve from my main research.

In late 2023, I had the privilege of interning at Alliander, which offered a refreshing perspective outside academia. Sander, bedankt voor je vertrouwen en de fijne werksfeer die je creëerde – ik kijk ernaar uit om onze samenwerking voort te zetten. Ferran, bedankt voor onze gezellige samenwerking en de openheid waarmee je met me meedacht tijdens onze data science avonturen. Evelyn, thank you for connecting the dots that made this opportunity possible! I'm grateful to the entire team for welcoming me into their energy grid world and sharing their insights so generously.

Teaching and supervising students has been one of the most rewarding aspects of my PhD journey. I would like to thank Mark, Mario, Tijn, Glenn, Bram, Chris, Dirren, Jurriaan, and Wouter for their trust and enthusiasm. To these students: watching you work through challenges and develop your research skills has been genuinely fulfilling and taught me just as much in return. I was particularly honored that several of you chose to

work with me – and other supervisors – for both your bachelor and master theses.

Although it preceded my PhD journey, I think it deserves special mention since it most definitely influenced it. Aan mijn GraphKite-kompanen – Jelle, Roeland, Simon – hartelijk dank voor alle avonturen, persoonlijke groei, en gedeelde ambities. Onze startup-dagen hebben een blijvende impact op mijn ontwikkeling gehad, en de herinneringen toveren nog steeds een glimlach op mijn gezicht.

En natuurlijk zijn er ook de vrienden die buiten mijn academische wereld om van onschatbare waarde zijn geweest. Arjen, Benoît, Esther & Frank, Jennifer & Mike, Jeroen & Larissa, Laura, Rik, Steven, en Tim – bedankt voor jullie vriendschap, voor het luisteren naar mijn (academische) gezeik, en voor alle momenten van ontspanning die jullie me hebben geboden tijdens deze jaren. En aan mijn vrienden van de Noordkantine: bedankt voor de onvergetelijke momenten die mijn studentenjaren hebben gekleurd.

Pap, mam, jullie hebben het fundament gelegd waarop ik heb kunnen bouwen. Toen ik jong was had ik veel "hobbies" – geen interesse was jullie te gek, alles werd gestimuleerd. Jullie hebben mijn nieuwsgierigheid altijd gevoed en mij tegelijkertijd de discipline bijgebracht om door te zetten. De typische "Kolmusheid" die ik van jullie heb geleerd en jullie subtiele sturing, zelfs als ik die op dat moment niet waardeerde, hebben me vaak geholpen. Bedankt voor jullie onwrikbare vertrouwen dat ik er uiteindelijk wel zou komen.

Bart en Marlien, ondanks dat we alle drie behoorlijk verschillend zijn, waardeer ik onze band enorm. Als we elkaar zien voelt het als thuis komen, zelfs na maanden van radiostilte. Bedankt dat jullie er altijd zijn.

Opa, je had dit moment graag meegemaakt. Je levenslust, je wereldse verhalen, en je rotsvaste overtuiging hebben me aangemoedigd om altijd mijn eigen route uit te stippelen.

Lieve Gretha, vijf jaar lang heb je deze reis van dichtbij meegemaakt. Je hebt me aangemoedigd tijdens de hoogtepunten, opgevangen tijdens de dipjes, en me de ruimte gegeven wanneer deadlines naderden. Maar bovenal zorgde je ervoor dat er naast werk ook volop plezier, gekkigheid en warmte was. Jouw aanwezigheid maakt elke dag mooier.

Curriculum vitea

Alex Kolmus was born on March 23, 1993, in Gouda, the Netherlands. He completed his Bachelor's degrees in Chemistry (2014) and Physics (2017), followed by a Master's degree in Physics (2019) with distinction from Radboud University.

During his studies in 2017, he participated in the Volkswagen Deep Learning and Robotics Challenge, where he and three teammates won the Jury Prize and achieved second place overall.

While pursuing his Master's degree, Alex co-founded GraphKite (2017-2020) with three friends. At GraphKite, he developed machine learning solutions for the insurance sector and provided technical expertise to help startups implement data-driven approaches.

From 2019 to 2024, Alex conducted his PhD research at Radboud University, focusing on the work presented in this thesis. During his doctoral studies, he completed a data science internship at Alliander (2023-2024).

Currently, he works as a data scientist at Alliander, where he focuses on solving inverse problems, simulating grid infrastructure, and addressing network congestion in energy grids.

Bibliography

[1]H. Thurston, *Early Astronomy* (Springer Science & Business Media, 1996).

[2]E. Rosen, "Was Copernicus' *Revolutions* approved by the Pope?", Journal of the History of Ideas **36**, 531–542 (1975).

[3]M. Segre, "Light on the Galileo Case?", Isis **88**, 484–504 (1997).

[4]T. S. Kuhn, *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought* (Harvard University Press, 1992).

[5]E. Hubble, "NGC 6822, A Remote Stellar System", Astrophysical Journal **62**, 409–433 (1925).

[6]M. Schmidt, "3C 273 : A Star-Like Object with Large Red-Shift", Nature **197**, 1040 (1963).

[7]A. Hewish, S. Bell, J. Pilkington, P. Scott, and R. Collins, "Observation of a Rapidly Pulsating Radio Source", Nature **217**, 709–713 (1968).

[8]M. J. Rieke et al., "JADES Initial Data Release for the Hubble Ultra Deep Field: Revealing the Faint Infrared Sky with Deep JWST NIRCam Imaging", The Astrophysical Journal Supplement Series **269**, 16 (2023).

[9]G. Hinshaw et al., "Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results", The Astrophysical Journal Supplement Series **208**, 19 (2013).

[10]N. Aghanim et al., "Planck 2018 results-VI. Cosmological parameters", Astronomy & Astrophysics **641**, A6 (2020).

[11]A. Einstein, "Die Feldgleichungen der Gravitation", Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften, 844–847 (1915).

[12]A. Einstein, "Die Grundlage der allgemeinen Relativitätstheorie", Annalen der Physik, 769–822 (1916).

[13]A. Einstein, "Näherungsweise Integration der Feldgleichungen der Gravitation", Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften, 688–696 (1916).

[14]A. Einstein, "Über Gravitationswellen", Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, 154–167 (1918).

[15]R. Weiss, "Nobel Lecture: LIGO and the discovery of gravitational waves I", Reviews of Modern Physics **90**, 040501 (2018).

[16]M. Gerstenshtein and V. Pustovoit, "On the Detection of Low Frequency Gravitational Waves", Soviet Physics-JETP **16**, 433–435 (1963).

[17]R. Weiss, "Electromagnetically coupled broadband gravitational antenna", Quarterly Progress Report **16**, 54–76 (1972).

[18]T. T. Fricke et al., "DC readout experiment in Enhanced LIGO", Classical and Quantum Gravity **29**, 065005 (2012).

[19]B. K. Berger, "Identification and mitigation of Advanced LIGO noise sources", in Journal of Physics: Conference Series, Vol. 957 (IOP Publishing, 2018), p. 012004.

[20]D. Davis et al., "Improving the Sensitivity of Advanced LIGO Using Noise Subtraction", Classical and Quantum Gravity **36**, 055011 (2019).

[21]S. M. Aston et al., "Update on quadruple suspension design for Advanced LIGO", Classical and Quantum Gravity **29**, 235004 (2012).

[22]LIGO, *Our Evolving Detectors*, (Apr. 2024) https://www.ligo.caltech.edu/page/ligo-evol.

[23]J.-Y. Vinet, B. Meers, C. N. Man, and A. Brillet, "Optimization of long-baseline optical interferometers for gravitational-wave detection", Physical Review D **38**, 433 (1988).

[24]The Royal Swedish Academy of Sciences, *The Nobel Prize in Physics 2017*, Press release, Oct. 2017.

[25]B. P. Abbott et al., "Observation of Gravitational Waves from a Binary Black Hole Merger", Physical Review Letters **116**, 061102 (2016).

[26]The Nobel Prize organisation, *Nobel Prizes 2017*, (June 2024) https://www.nobelprize.org/all-nobel-prizes-2017/.

[27]F. Acernese et al., "Advanced Virgo: a second-generation interferometric gravitational wave detector", Classical and Quantum Gravity **32**, 024001 (2014).

[28]T. Akutsu et al., "Overview of KAGRA: Detector design and construction history", Progress of Theoretical and Experimental Physics **2021**, 05A101 (2021).

[29]R. Abbott et al., "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run", Physical Review X **13**, 041039 (2023).

[30]B. P. Abbott et al., "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral", Physical Review Letters **119**, 161101 (2017).

[31]B. Abbott et al., "GW190425: Observation of a Compact Binary Coalescence with Total Mass $\sim 3.4\ M_\odot$", The Astrophysical Journal **892**, L3 (2020).

[32]R. Abbott et al., "Tests of General Relativity with GWTC-3", arXiv preprint arXiv:2112.06861 (2021).

[33]R. Abbott et al., "Population of Merging Compact Binaries Inferred Using Gravitational Waves through GWTC-3", Physical Review X **13**, 011048 (2023).

[34]D. Kasen, B. Metzger, J. Barnes, E. Quataert, and E. Ramirez-Ruiz, "Origin of the heavy elements in binary neutron-star mergers from a gravitational-wave event", Nature **551**, 80–84 (2017).

[35]H. O. Silva, A. M. Holgado, A. Cárdenas-Avendaño, and N. Yunes, "Astrophysical and theoretical physics implications from multimessenger neutron star observations", Physical Review Letters **126**, 181101 (2021).

[36]N. Yunes, M. C. Miller, and K. Yagi, "Gravitational-Wave and X-ray Probes of the Neutron Star Equation of State", Nature Reviews Physics **4**, 237–246 (2022).

[37]T. Andrade et al., "GRChombo: An adaptable numerical relativity code for fundamental physics", Journal of Open Source Software **6**, 3703 (2021).

[38]F. Löffler et al., "The Einstein Toolkit: A Community Computational Infrastructure for Relativistic Astrophysics", Classical and Quantum Gravity **29**, 115001 (2012).

[39]I. Ruchlin, Z. B. Etienne, and T. W. Baumgarte, "SENR/NRPy+: Numerical Relativity in Singular Curvilinear Coordinate Systems", Physical Review D **97**, 064036 (2018).

[40]I. Hinder et al., "Error-analysis and comparison to analytical models of numerical waveforms produced by the NRAR Collaboration", Classical and Quantum Gravity **31**, 025012 (2013).

[41]J. Skilling, "Nested Sampling for General Bayesian Computation", Bayesian Analysis **1**, 833–859 (2006).

[42]J. Buchner, "A statistical test for nested sampling algorithms", Statistics and Computing **26**, 383–392 (2016).

[43]R. J. Smith, G. Ashton, A. Vajpeyi, and C. Talbot, "Massively parallel Bayesian inference for transient gravitational-wave astronomy", Monthly Notices of the Royal Astronomical Society **498**, 4492–4502 (2020).

[44]B. P. Abbott et al., "GWTC-1: a Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs", Physical Review X **9**, 031040 (2019).

[45]S. E. Perkins, N. Yunes, and E. Berti, "Probing Fundamental Physics with Gravitational Waves: The Next Generation", Physical Review D **103**, 044024 (2021).

[46]S. E. Gossan, E. D. Hall, and S. M. Nissanke, "Optimizing the Third Generation of Gravitational-wave Observatories for Galactic Astrophysics", The Astrophysical Journal **926**, 231 (2022).

[47]M. Evans et al., "A Horizon Study for Cosmic Explorer: Science, Observatories, and Community", arXiv preprint arXiv:2109.09882 (2021).

[48]L. M. Thomas, G. Pratten, and P. Schmidt, "Accelerating multimodal gravitational waveforms from precessing compact binaries with artificial neural networks", Physical Review D **106**, 104029 (2022).

[49]S. Schmidt et al., "Machine Learning Gravitational Waves from Binary Black Hole Mergers", Physical Review D **103**, 043020 (2021).

[50]T. D. Edwards et al., "RIPPLE: Differentiable and Hardware-Accelerated Waveforms for Gravitational Wave Data Analysis", arXiv preprint arXiv:2302.05329 (2023).

[51]F. Iacovelli, M. Mancarella, S. Foffa, and M. Maggiore, "GWFAST: A Fisher Information Matrix Python Code for Third-generation Gravitational-wave Detectors", The Astrophysical Journal Supplement Series **263**, 2 (2022).

[52]S. Morisaki, "Accelerating parameter estimation of gravitational waves from compact binary coalescence using adaptive frequency resolutions", Physical Review D **104**, 044062 (2021).

[53]B. Zackay, L. Dai, and T. Venumadhav, "Relative Binning and Fast Likelihood Evaluation for Gravitational Wave Parameter Estimation", arXiv preprint arXiv:1806.08792 (2018).

[54]N. Leslie, L. Dai, and G. Pratten, "Mode-by-mode Relative Binning: Fast Likelihood Estimation for Gravitational Waveforms with Spin-Orbit Precession and Multiple Harmonics", Physical Review D **104**, 123030 (2021).

[55]H. Narola, J. Janquart, Q. Meijer, K. Haris, and C. V. D. Broeck, "Relative binning for complete gravitational-wave parameter estimation with higher-order modes and precession, and applications to lensing and third-generation detectors", arXiv preprint arXiv:2308.12140 (2023).

[56]U. Bhardwaj, J. Alvey, B. K. Miller, S. Nissanke, and C. Weniger, "Sequential simulation-based inference for gravitational wave signals", Physical Review D **108**, 042004 (2023).

[57]M. Dax et al., "Real-Time Gravitational Wave Science with Neural Posterior Estimation", Physical Review Letters **127**, 241103 (2021).

[58]J. Lange, R. O'Shaughnessy, and M. Rizzo, "Rapid and accurate parameter inference for coalescing, precessing compact binaries", arXiv preprint arXiv:1805.10457 (2018).

[59]A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems **25** (2012).

[60]C. M. Bishop and H. Bishop, *Deep learning: foundations and concepts* (Springer Nature, 2023).

[61]G. Papamakarios, D. Sterratt, and I. Murray, "Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows", in The 22nd International Conference on Artificial Intelligence and Statistics (PMLR, 2019), pp. 837–848.

[62]K. Cranmer, J. Pavez, and G. Louppe, "Approximating Likelihood Ratios with Calibrated Discriminative Classifiers", arXiv preprint arXiv:1506.02169 (2015).

[63]G. Papamakarios and I. Murray, "Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation", Advances in Neural Information Processing Systems **29** (2016).

[64]R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan, "Measuring Catastrophic Forgetting in Neural Networks", in Proceedings of the thirty-second aaai conference on artificial intelligence, Vol. 32, 1 (2018).

[65]C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks", in Proceedings of the 34th international conference on machine learning (PMLR, 2017), pp. 1321–1330.

[66]M. Minderer et al., "Revisiting the Calibration of Modern Neural Networks", Advances in Neural Information Processing Systems **34**, 15682–15694 (2021).

[67]K. Schwarzschild, "Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie", Sitzungsberichte der königlich preussischen Akademie der Wissenschaften, 189–196 (1916).

[68]R. P. Kerr, "Gravitational field of a spinning mass as an example of algebraically special metrics", Physical Review Letters **11**, 237 (1963).

[69]M. Maggiore, *Gravitational waves: volume 1: theory and experiments* (Oxford University Publising, 2008).

[70]K. Arun, A. Buonanno, G. Faye, and E. Ochsner, "Higher-order spin effects in the amplitude and phase of gravitational waveforms emitted by inspiraling compact binaries: Ready-to-use gravitational waveforms", Physical Review D **79**, 104023 (2009).

[71]R. Abbott et al., "Gw190814: gravitational waves from the coalescence of a 23 solar mass black hole with a 2.6 solar mass compact object", The Astrophysical Journal Letters **896**, L44 (2020).

[72]C. Cahillane, *How does LIGO detect gravitational waves?*, (June 2024) https://www.youtube.com/watch?v=X7RJHxeCulY.

[73]B. F. Schutz, "Networks of gravitational wave detectors and three figures of merit", Classical and Quantum Gravity **28**, 125023 (2011).

[74]V. Varma et al., "Gravitational-wave observations of binary black holes: Effect of nonquadrupole modes", Physical Review D **90**, 124004 (2014).

[75]K. S. Thorne, C. W. Misner, and J. A. Wheeler, *Gravitation* (Freeman San Francisco, 2000).

[76]D. Gerosa et al., "Spin orientations of merging black holes formed from the evolution of stellar binaries", Physical Review D **98**, 084036 (2018).

[77]S. L. Shapiro and S. A. Teukolsky, "Collisions of relativistic clusters and the formation of black holes", Physical Review D **45**, 2739 (1992).

[78]F. Pretorius, "Evolution of binary black-hole spacetimes", Physical Review Letters **95**, 121101 (2005).

[79]J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, "Gravitational-wave extraction from an inspiraling configuration of merging black holes", Physical Review Letters **96**, 111102 (2006).

[80]M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower, "Accurate evolutions of orbiting black-hole binaries without excision", Physical Review Letters **96**, 111101 (2006).

[81]M. Boyle et al., "The SXS Collaboration catalog of binary black hole simulations", Classical and Quantum Gravity **36**, 195006 (2019).

[82]J. Healy, C. O. Lousto, Y. Zlochower, and M. Campanelli, "The RIT binary black hole simulations catalog", Classical and Quantum Gravity **34**, 224001 (2017).

[83]K. Jani et al., "Georgia Tech Catalog of Gravitational Waveforms", Classical and Quantum Gravity **33**, 204001 (2016).

[84]J. Healy and C. O. Lousto, "Fourth RIT binary black hole simulations catalog: Extension to eccentric orbits", Physical Review D **105**, 124010 (2022).

[85]F. Cattorini and B. Giacomazzo, "GRMHD study of accreting massive black hole binaries in astrophysical environment: a review", Astroparticle Physics, 102892 (2023).

[86]T. Damour, B. R. Iyer, and B. S. Sathyaprakash, "Comparison of search templates for gravitational waves from binary inspiral", Physical Review D **63**, 044023 (2001).

[87]T. Damour, B. R. Iyer, and B. S. Sathyaprakash, "Comparison of search templates for gravitational waves from binary inspiral: 3.5 PN update", Physical Review D **66**, 027502 (2002).

[88]A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, "Comparison of post-Newtonian templates for compact binary inspiral signals in gravitational-wave detectors", Physical Review D **80**, 084043 (2009).

[89]A. Ramos-Buades et al., "Next generation of accurate and efficient multipolar precessing-spin effective-one-body waveforms for binary black holes", Physical Review D **108**, 124037 (2023).

[90]J. Blackman et al., "Numerical Relativity Waveform Surrogate Model for Generically Precessing Binary Black Hole Mergers", Physical Review D **96**, 024058 (2017).

[91]G. Pratten et al., "Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes", Physical Review D **103**, 104056 (2021).

[92]P. Whittle, "Estimation and information in stationary time series", Arkiv för matematik **2**, 423–434 (1953).

[93]W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications", Biometrika **57**, 97–109 (1970).

[94]D. W. Hogg and D. Foreman-Mackey, "Data Analysis Recipes: Using Markov Chain Monte Carlo", The Astrophysical Journal Supplement Series **236**, 11 (2018).

[95]D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä, "A survey of Monte Carlo methods for parameter estimation", EURASIP Journal on Advances in Signal Processing **2020**, 1–62 (2020).

[96]R. M. Neal, "Slice sampling", The Annals of Statistics **31**, 705–767 (2003).

[97]M. Betancourt, "Nested Sampling with Constrained Hamiltonian Monte Carlo", in AIP Conference Proceedings, Vol. 1305, 1 (American Institute of Physics, 2011), pp. 165–172.

[98]J. Shaw, M. Bridges, and M. Hobson, "Efficient Bayesian inference for multimodal problems in cosmology", Monthly Notices of the Royal Astronomical Society **378**, 1365–1370 (2007).

[99]B. J. Brewer, L. B. Pártay, and G. Csányi, "Diffusive nested sampling", Statistics and Computing **21**, 649–656 (2011).

[100]W. Handley, M. Hobson, and A. Lasenby, "POLYCHORD: next-generation nested sampling", Monthly Notices of the Royal Astronomical Society **453**, 4384–4398 (2015).

[101]E. Higson, W. Handley, M. Hobson, and A. Lasenby, "Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation", Statistics and Computing **29**, 891–913 (2019).

[102]J. Veitch et al., "Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library", Physical Review D **91**, 042003 (2015).

[103]G. Ashton et al., "BILBY: A User-friendly Bayesian Inference Library for Gravitational-wave Astronomy", The Astrophysical Journal Supplement Series **241**, 27 (2019).

[104]R. Abbott et al., "GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo during the First Half of the Third Observing Run", Physical Review X **11**, 021053 (2021).

[105]J. S. Speagle, "Dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences", Monthly Notices of the Royal Astronomical Society **493**, 3132–3158 (2020).

[106]J. Janquart, O. A. Hannuksela, K. Haris, and C. Van Den Broeck, "A fast and precise methodology to search for and analyse strongly lensed gravitational-wave events", Monthly Notices of the Royal Astronomical Society **506**, 5430–5438 (2021).

[107]C. Talbot and E. Thrane, "Gravitational-wave astronomy with an uncertain noise power spectral density", Physical Review Research **2**, 043298 (2020).

[108]Q. Wu, T. Zhu, R. Niu, W. Zhao, and A. Wang, "Constraints on the Nieh-Yan modified teleparallel gravity with gravitational waves", Physical Review D **105**, 024035 (2022).

[109]S. J. D. Prince, *"Understanding Deep Learning"* (The MIT Press, 2023).

[110]R. L. Russell and C. Reale, "Multivariate Uncertainty in Deep Learning", IEEE Transactions on Neural Networks and Learning Systems **33**, 7937–7943 (2021).

[111]L. Sluijterman, E. Cator, and T. Heskes, "Optimal training of Mean Variance Estimation neural networks", Neurocomputing, 127929 (2024).

[112]D. Rezende and S. Mohamed, "Variational Inference with Normalizing Flows", in Proceedings of the 32nd International Conference on Machine Learning (PMLR, 2015), pp. 1530–1538.

[113]L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear Independent Components Estimation", in Proceedings of the Third International Conference on Learning Representations, Workshop Track (2015).

[114]D. P. Kingma et al., "Improved Variational Inference with Inverse Autoregressive Flow", Advances in Neural Information Processing Systems **29** (2016).

[115]G. Papamakarios, T. Pavlakou, and I. Murray, "Masked Autoregressive Flow for Density Estimation", Advances in Neural Information Processing Systems **30** (2017).

[116]C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, "Neural Spline Flows", Advances in Neural Information Processing Systems **32** (2019).

[117]S. Ramasinghe, K. Fernando, S. Khan, and N. Barnes, "Robust Normalizing Flows using Bernstein-type Polynomials", arXiv preprint arXiv:2102.03509 (2021).

[118]F. Draxler, S. Wahl, C. Schnörr, and U. Köthe, "On the Universality of Volume-Preserving and Coupling-Based Normalizing Flows", arXiv preprint arXiv:2402.06578 (2024).

[119]L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP", in Proceedings of the Fifth International Conference on Learning Representations (2017).

[120]D. P. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions", Advances in Neural Information Processing Systems **31** (2018).

[121]B. Uria, I. Murray, and H. Larochelle, "RNADE: the real-valued neural autoregressive density-estimator", Advances in Neural Information Processing Systems **26** (2013).

[122]M. Germain, K. Gregor, I. Murray, and H. Larochelle, "MADE: Masked Autoencoder for Distribution Estimation", in Proceedings of the 32nd International Conference on Machine Learning (PMLR, 2015), pp. 881–889.

[123]R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations", Advances in Neural Information Processing Systems **31** (2018).

[124]W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, "FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models", in Proceedings of the Sixth International Conference on Learning Representations (2018).

[125]L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, "The Mathematical Theory of Optimal Processes", Interscience (1962).

[126]B. P. Abbott et al., "Tests of General Relativity with GW150914", Physical Review Letters **116** (2016).

[127]B. P. Abbott et al., "GW170814: A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence", Physical Review Letters **119**, 141101 (2017).

[128]K. Yamada, T. Narikawa, and T. Tanaka, "Testing massive-field modifications of gravity via gravitational waves", Progress of Theoretical and Experimental Physics **2019**, 103E01 (2019).

[129]L. Baiotti, "Gravitational waves from neutron star mergers and their relation to the nuclear equation of state", Progress in Particle and Nuclear Physics **109**, 103714 (2019).

[130]J. Aasi et al., "Advanced LIGO", Classical and Quantum Gravity **32**, 074001 (2015).

[131]Z. Doctor et al., "A Search for Optical Emission from Binary Black Hole Merger GW170814 with the Dark Energy Camera", The Astrophysical Journal Letters **873**, L24 (2019).

[132]R. Perna, D. Lazzati, and W. Farr, "Limits on Electromagnetic Counterparts of Gravitational-wave-detected Binary Black Hole Mergers", The Astrophysical Journal **875**, 49 (2019).

[133]P. S. Cowperthwaite et al., "The Electromagnetic Counterpart of the Binary Neutron Star Merger LIGO/Virgo GW170817. II. UV, Optical, and Near-infrared Light Curves and Comparison to Kilonova Models", The Astrophysical Journal Letters **848**, L17 (2017).

[134]E. Berti, K. Yagi, and N. Yunes, "Extreme Gravity Tests with Gravitational Waves from Compact Binary Coalescences:(I) Inspiral–Merger", General Relativity and Gravitation **50**, 1–45 (2018).

[135]L. Barack et al., "Black holes, gravitational waves and fundamental physics: a roadmap", Classical and Quantum Gravity **36**, 143001 (2019).

[136]M. Fishbach et al., "A Standard Siren Measurement of the Hubble Constant from GW170817 without the Electromagnetic Counterpart", The Astrophysical Journal Letters **871**, L13 (2019).

[137]C. M. Biwer et al., "PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals", Publications of the Astronomical Society of the Pacific **131**, 024503 (2019).

[138]K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).

[139]H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, "Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy", Nature Physics **18**, 112–117 (2022).

[140]L. P. Singer and L. R. Price, "Rapid bayesian position reconstruction for gravitational-wave transients", Physical Review D **93**, 024013 (2016).

[141]Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", Nature **521**, 436–444 (2015).

[142]J. Schmidhuber, "Deep learning in neural networks: an overview", Neural Networks **61**, 85–117 (2015).

[143]A. Delaunoy et al., "Lightning-Fast Gravitational Wave Parameter Inference through Neural Amortization", in Machine Learning and the Physical Sciences. Workshop at the 34th Conference on Neural Information Processing Systems (2020).

[144]S. R. Green and J. Gair, "Complete parameter inference for GW150914 using deep learning", Machine Learning: Science and Technology **2**, 03LT01 (2021).

[145]G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for Probabilistic Modeling and Inference", Journal of Machine Learning Research **22**, 1–64 (2021).

[146]M. J. Williams, J. Veitch, and C. Messenger, "Nested Sampling with Normalizing Flows for Gravitational-Wave Inference", Physical Review D **103**, 103006 (2021).

[147]C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).

[148]P. Canizares, S. E. Field, J. R. Gair, and M. Tiglio, "Gravitational wave parameter estimation with compressed likelihood evaluations", Physical Review D **87**, 124005 (2013).

[149]D. George and E. Huerta, "Deep Neural Networks to Enable Real-Time Multimessenger Astrophysics", Physical Review D **97**, 044039 (2018).

[150]X. Fan, J. Li, X. Li, Y. Zhong, and J. Cao, "Applying deep neural networks to the detection and space parameter estimation of compact binary coalescence with a network of gravitational wave detectors", Science China Physics, Mechanics & Astronomy **62**, 1–8 (2019).

[151]N. I. Fisher, T. Lewis, and B. J. Embleton, *Statistical analysis of spherical data* (Cambridge University Press, 1993).

[152]M. Hannam et al., "Simple model of complete precessing black-hole-binary gravitational waveforms", Physical Review Letters **113**, 151101 (2014).

[153]D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", in Proceedings of the Third International Conference on Learning Representations (2015).

[154]I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts", in Proceedings of the Fifth International Conference on Learning Representations (2017).

[155]T. D. Gebhard, N. Kilbertus, I. Harry, and B. Schölkopf, "Convolutional neural networks: A magic bullet for gravitational-wave detection?", Physical Review D **100**, 063015 (2019).

[156]C. Chatterjee, L. Wen, K. Vinsen, M. Kovalam, and A. Datta, "Using Deep Learning to Localize Gravitational Wave Sources", Physical Review D **100**, 103025 (2019).

[157]A. J. Chua and M. Vallisneri, "Learning Bayesian posteriors with neural networks for gravitational-wave inference", Physical Review Letters **124**, 041102 (2020).

[158]K. Somiya and K. Collaboration), "Detector configuration of KAGRA–the Japanese cryogenic gravitational-wave detector", Classical and Quantum Gravity **29**, 124007 (2012).

[159]Y. Aso et al., "Interferometer design of the KAGRA gravitational wave detector", Physical Review D **88**, 043007 (2013).

[160]T. Akutsu et al., "KAGRA: 2.5 Generation Interferometric Gravitational Wave Detector", Nature Astronomy **3**, 35–40 (2019).

[161]B. Iyer et al., *LIGO-India Tech. rep.* https://dcc.ligo.org/LIGO-M1100296/public, 2011.

[162]M. Punturo et al., "The Einstein Telescope: A third-generation gravitational wave observatory", Classical and Quantum Gravity **27**, 194002 (2010).

[163]S. Hild et al., "Sensitivity Studies for Third-Generation Gravitational Wave Observatories", Classical and Quantum Gravity **28**, 094013 (2011).

[164]D. Reitze et al., "Cosmic Explorer: The US Contribution to Gravitational-Wave Astronomy beyond LIGO", Bulletin of the American Astronomical Society **51**, 035 (2019).

[165]B. P. Abbott et al., "Exploring the Sensitivity of Next Generation Gravitational Wave Detectors", Classical and Quantum Gravity **34**, 044001 (2017).

[166]T. Regimbau et al., "Digging Deeper: Observing Primordial Gravitational Waves Below the Binary-Black-Hole-Produced Stochastic Background", Physical Review Letters **118**, 151105 (2017).

[167]B. Sathyaprakash et al., "Scientific objectives of Einstein Telescope", Classical and Quantum Gravity **29**, 124013 (2012).

[168]T. Regimbau and S. A. Hughes, "Gravitational-wave confusion background from cosmological compact binaries: Implications for future terrestrial detectors", Physical Review D **79**, 062002 (2009).

[169]A. Samajdar, J. Janquart, C. Van Den Broeck, and T. Dietrich, "Biases in parameter estimation from overlapping gravitational-wave signals in the third-generation detector era", Physical Review D **104**, 044003 (2021).

[170]E. Pizzati, S. Sachdev, A. Gupta, and B. Sathyaprakash, "Toward inference of overlapping gravitational-wave signals", Physical Review D **105**, 104016 (2022).

[171]P. Relton and V. Raymond, "Parameter estimation bias from overlapping binary black hole events in second generation interferometers", Physical Review D **104**, 084039 (2021).

[172]Y. Himemoto, A. Nishizawa, and A. Taruya, "Impacts of overlapping gravitational-wave signals on the parameter estimation: Toward the search for cosmological backgrounds", Physical Review D **104**, 044010 (2021).

[173]A. Antonelli, O. Burke, and J. R. Gair, "Noisy neighbours: inference biases from overlapping gravitational-wave signals", Monthly Notices of the Royal Astronomical Society **507**, 5069–5086 (2021).

[174]S. Wu and A. H. Nitz, "Mock data study for next-generation ground-based detectors: The performance loss of matched filtering due to correlated confusion noise", Physical Review D **107**, 063022 (2023).

[175]S. Sachdev, T. Regimbau, and B. Sathyaprakash, "Subtracting compact binary foreground sources to reveal primordial gravitational-wave backgrounds", Physical Review D **102**, 024051 (2020).

[176]A. Sharma and J. Harms, "Searching for cosmological gravitational-wave backgrounds with third-generation detectors in the presence of an astrophysical foreground", Physical Review D **102**, 063009 (2020).

[177]S. Biscoveanu, C. Talbot, E. Thrane, and R. Smith, "Measuring the Primordial Gravitational-Wave Background in the Presence of Astrophysical Foregrounds", Physical Review Letters **125**, 241101 (2020).

[178]B. Zhou et al., "Subtracting compact binary foregrounds to search for subdominant gravitational-wave backgrounds in next-generation ground-based observatories", Physical Review D **108**, 064040 (2023).

[179]B. Zhou et al., "Compact Binary Foreground Subtraction in Next-Generation Ground-Based Observatories", arXiv preprint arXiv:2209.01221 (2022).

[180]L. Reali et al., "The impact of confusion noise on golden binary neutron-star events in next-generation terrestrial observatories", arXiv preprint arXiv:2209.13452 (2022).

[181]J. Janquart, T. Baka, A. Samajdar, T. Dietrich, and C. Van Den Broeck, "Analyses of overlapping gravitational wave signals using hierarchical subtraction and joint parameter estimation", Monthly Notices of the Royal Astronomical Society **523**, 1699–1710 (2023).

[182]L. Dai, T. Venumadhav, and B. Zackay, "Parameter Estimation for GW170817 using Relative Binning", arXiv preprint arXiv:1806.08793 (2018).

[183]S. Gao, F. Hayes, S. Croke, C. Messenger, and J. Veitch, "Quantum algorithm for gravitational-wave matched filtering", Physical Review Research **4**, 023006 (2022).

[184]E. Cuoco et al., "Enhancing gravitational-wave science with machine learning", Machine Learning: Science and Technology **2**, 011002 (2020).

[185]K. Cranmer, J. Brehmer, and G. Louppe, "The frontier of simulation-based inference", Proceedings of the National Academy of Sciences **117**, 30055–30062 (2020).

[186]S. R. Green, C. Simpson, and J. Gair, "Gravitational-wave parameter estimation with autoregressive neural network flows", Physical Review D **102**, 104057 (2020).

[187]M. Dax et al., "Group equivariant neural posterior estimation", in Proceedings of the Tenth International Conference on Learning Representations (2022).

[188]M. Dax et al., "Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference", Physical Review Letters **130**, 171403 (2023).

[189]I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods", IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 3964–3979 (2020).

[190]C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling, "Learning Likelihoods with Conditional Normalizing Flows", arXiv preprint arXiv:1912.00042 (2019).

[191]D. Ha, A. M. Dai, and Q. V. Le, "Hypernetworks", in Proceedings of the Fifth International Conference on Learning Representations (2017).

[192]J. Zhuang, N. C. Dvornek, S. Tatikonda, and J. S. Duncan, "MALI: a memory efficient and reverse accurate integrator for neural odes", in Proceedings of the Ninth International Conference on Learning Representations (2021).

[193]N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions", SIAM Review **53**, 217–288 (2011).

[194]L. Barsotti, P. Fritschel, M. Evans, and S. Gras, *Advanced LIGO anticipated sensitivity curves*, https://dcc.ligo.org/LIGO-T1800044/public, 2021.

[195]S. Khan et al., "Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era", Physical Review D **93**, 044007 (2016).

[196]K. Akiyama et al., "First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole", The Astrophysical Journal Letters **875**, L4 (2019).

[197]W. W. Symes, "The seismic reflection inverse problem", Inverse Problems **25**, 123008 (2009).

[198]S. R. Arridge and J. C. Schotland, "Optical tomography: forward and inverse problems", Inverse Problems **25**, 123010 (2009).

[199]S. Baillet, J. C. Mosher, and R. M. Leahy, "Electromagnetic brain mapping", IEEE Signal Processing Magazine **18**, 14–30 (2001).

[200]B. S. Sathyaprakash and B. F. Schutz, "Physics, Astrophysics and Cosmology with Gravitational Waves", Living Reviews in Relativity **12**, 1–141 (2009).

[201]S. Vitale, "The first 5 years of gravitational-wave astrophysics", Science **372**, eabc7397 (2021).

[202]B. P. Abbott et al., "Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1", Physical Review D **100**, 104036 (2019).

[203]R. Abbott et al., "Population properties of Compact Objects from the Second LIGO–Virgo Gravitational-Wave Transient Catalog", The Astrophysical Journal Letters **913**, L7 (2021).

[204]A. Finke, S. Foffa, F. Iacovelli, M. Maggiore, and M. Mancarella, "Cosmology with LIGO/Virgo dark sirens: Hubble parameter and modified gravitational wave propagation", Journal of Cosmology and Astroparticle Physics **2021**, 026 (2021).

[205]B. P. Abbott et al., "Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA", Living Reviews in Relativity **23**, 1–69 (2020).

[206]C. P. L. Berry et al., "Parameter estimation for binary neutron-star coalescences with realistic noise during the advanced ligo era", The Astrophysical Journal **804**, 114 (2015).

[207]H. Estellés et al., "Time-domain phenomenological model of gravitational-wave subdominant harmonics for quasicircular non-precessing binary black hole coalescences", Physical Review D **105**, 084039 (2022).

[208]R. W. Kiendrebeogo et al., "Updated Observing Scenarios and Multimessenger Implications for the International Gravitational-Wave Networks O4 and O5", The Astrophysical Journal **958**, 158 (2023).

[209]K. W. Wong, M. Isi, and T. D. Edwards, "Fast Gravitational-wave Parameter Estimation without Compromises", The Astrophysical Journal **958**, 129 (2023).

[210]S. Fairhurst, C. Hoy, R. Green, C. Mills, and S. A. Usman, "Simple parameter estimation using observable features of gravitational-wave signals", Physical Review D **108**, 082006 (2023).

[211]V. Tiwari, C. Hoy, S. Fairhurst, and D. MacLeod, "Fast non-Markovian sampler for estimating gravitational-wave posteriors", Physical Review D **108**, 023001 (2023).

[212]L. Pathak, A. Reza, and A. S. Sengupta, "Fast likelihood evaluation using meshfree approximations for reconstructing compact binary sources", Physical Review D **108**, 064055 (2023).

[213]J. Lueckmann, J. Boelts, D. Greenberg, P. J. Goncalves, and J. H. Macke, "Benchmarking Simulation-Based Inference", in Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (PMLR, 2021), pp. 343–351.

[214]J. Hermans, V. Begy, and G. Louppe, "Likelihood-free MCMC with Amortized Approximate Ratio Estimators", in Proceedings of the 37th International Conference on Machine Learning, Vol. 119 (2020), pp. 4239–4248.

[215]C. M. Bishop, *Mixture density networks* (Aston University, 1994).

[216]T. Minka et al., *Divergence measures and message passing*, tech. rep. (Technical report, Microsoft Research, 2005).

[217]L. I. Midgley, V. Stimper, G. N. Simm, B. Schölkopf, and J. M. Hernández-Lobato, "Flow annealed importance sampling bootstrap", in Proceedings of the Eleventh International Conference on Learning Representations (2023).

[218]C. G. Torre, *Foundations of wave phenomena*, 2012.

[219]L. Kish, *Survey sampling* (Wiley, 1965).

[220]K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks", in Computer Vision – ECCV (Springer, 2016), pp. 630–645.

[221]D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)", arXiv preprint arXiv:1606.08415 (2016).

[222]J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization", arXiv preprint arXiv:1607.06450 (2016).

[223]T. Salimans and D. P. Kingma, "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks", Advances in Neural Information Processing Systems **29** (2016).

[224]C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization", Communications of the ACM **64**, 107–115 (2021).

[225]J. Kaplan et al., "Scaling Laws for Neural Language Models", arXiv preprint arXiv:2001.08361 (2020).

[226]A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets", arXiv preprint arXiv:2201.02177 (2022).

[227]T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák, "Neural Importance Sampling", ACM Transactions on Graphics **38**, 1–19 (2019).

[228]D. Greenberg, M. Nonnenmacher, and J. H. Macke, "Automatic Posterior Transformation for Likelihood-Free Inference", in Proceedings of the 36th International Conference on Machine Learning (PMLR, 2019), pp. 2404–2414.

[229]J. Lueckmann et al., "Flexible statistical inference for mechanistic models of neural dynamics", Advances in Neural Information Processing Systems **30** (2017).

[230]M. Deistler, P. J. Goncalves, and J. H. Macke, "Truncated proposals for scalable and hassle-free simulation-based inference", Advances in Neural Information Processing Systems **35**, 23135–23149 (2022).

[231]G. Pratten et al., "Setting the cornerstone for a family of models for gravitational waves from compact binaries: the dominant harmonic for nonprecessing quasicircular black holes", Physical Review D **102**, 064001 (2020).

[232]K. Cannon et al., "Singular value decomposition applied to compact binary coalescence gravitational-wave signals", Physical Review D **82**, 044025 (2010).

9 789465 150871 >

Radboud Universiteit