Extracting astrophysics from long-lived gravitational wave signals with the Laser Interferometer Space Antenna

Christian Edward Anthony Chapman-Bird

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Physics and Astronomy
College of Science and Engineering
University of Glasgow

June 2024

Abstract

The Laser Interferometer Space Antenna (LISA) mission is a space-based gravitational wave (GW) detector that will operate in the mHz frequency band, and is expected to observe a rich variety of astrophysical GW sources including Galactic binaries (GBs), massive black hole binaries (MBHBs) and extreme-mass-ratio inspirals (EMRIs). While the analysis and astrophysical interpretation of LISA data has been growing as a field over the last few decades, the adoption of LISA by the European Space Agency (ESA) in January 2024 places well-defined goalposts (the launch of the mission in a decade's time) that present an urgent need for rapid development of waveform models and data analysis methods. This thesis addresses this need, describing tools for the accurate and efficient analysis of LISA data.

The enormous astrophysical potential of LISA observations is only accessible if techniques developed for its extraction can be constructed and extensively validated. In order to ensure that analysis pipelines will perform optimally on the real LISA dataset once the mission flies, they must be tested on realistic simulations of the instrument. However, the infrastructure required for these simulations has only recently been developed. Part of this thesis focuses on the validation of these simulation tools, assessing the accuracy with which injected GW signals are recovered with statistical techniques. The essential capability of time delay interferometry (TDI) to suppress instrumental noise sources by orders of magnitude without degrading GW signals in the data stream is demonstrated for realistic LISA simulations. The results of this work, which were presented to ESA as part of the successful LISA adoption effort, lay the necessary groundwork for future investigations of the performance of the LISA instrument and the algorithms constructed to analyse its observations.

While LISA is expected to measure the parameters of astrophysical GW sources with great precision, this in turn requires waveform models for these signals to be highly accurate. However, these accurate models must also be computationally inexpensive if LISA data analysis is to be a feasible prospect. This is particularly pertinent for EMRIs, which produce complicated and long-lived waveforms that must be carefully modelled to avoid systematic biases. In this thesis, EMRI waveform models are augmented with machine learning (ML) techniques to improve their efficiency. By substituting inefficient operations in the waveform model with a neural network, a significant reduction in computational cost is attained without loss of accuracy. The resulting streamlined EMRI waveform model is an order of magnitude faster than existing techniques

and can be used to rapidly infer the parameters of EMRI signals. As EMRI waveform models improve, their complexity will grow; this is reflected in the techniques presented in this thesis, which can be readily extended as EMRI waveform models become increasingly sophisticated.

Similarly to the analysis of individual GW sources, population studies with LISA observations will be computationally expensive. Accurate modelling of selection effects is necessary for unbiased population inferences, but the high computational cost of waveforms for LISA sources makes selection biases prohibitively expensive to correct for with standard techniques. The latter part of this thesis focuses on addressing this limitation by modelling the detectability of GW sources with ML methods. Investigating EMRI population inference as a representative example, selection effects are accurately modelled by neural networks that are orders of magnitude more efficient than standard techniques. Embedding this method in a Bayesian hierarchical inference framework, we rapidly constrain EMRI population parameters from simulated detection catalogues and demonstrate that the resulting inferences are free of systematic bias. The methodology we develop is completely agnostic to the waveform and population models used, and is sufficiently flexible to be applied to the complicated population inference problems that will be encountered in the hierarchical analysis of LISA observations.

Contents

List of Figures

List of Tables

Acknowledgements

Working towards (and writing) this thesis has been a challenging yet rewarding experience. It is without a doubt the hardest thing I have accomplished in my life so far, and I am certain that without the support of those around me that I would not have succeeded. I would therefore like to take this opportunity to thank the people that offered me their time, motivation and assistance throughout my PhD.

I owe a great deal to my colleagues at the Institute for Gravitational Research (IGR), whom I interacted with on a daily basis and had to endure my frequent complaints regarding the thesis writing process. It has been my great fortune to enjoy the warm and collegiate atmosphere of the IGR during my years as a postgraduate researcher, and despite my relief in completing this work I am sad to see this period of my life conclude. In particular, I would like to thank my office mates Jack Callaghan, Jose Gonzalez, Stuart Hill, Graeme McGhee, Rhona McTeague and Thomas Wallace for the many everyday conversations we shared, and for putting up with my distracting habits (...props to `jetpunk.com` for consuming more hours of my life than I'd care to admit). I would like to thank Alasdair Taylor, Daniel Williams and Michael Williams (no relation) for their sound advice regarding both my research and how better to distance myself from it for my sanity's sake. Thank you to my supervisors, Graham Woan and Christopher Berry, for your patience and enthusiasm in keeping me from straying too far off-track during my research (and I know I made that difficult at times!). I am especially grateful to Christopher for his assistance in improving the quality of my writing, and for his careful proof-reading of my work (which I hope became less painful as time progressed).

Outside of the IGR, I have been fortunate to meet some excellent people in my academic travels that I have enjoyed many interesting conversations with; in particular I would like to thank Ollie Burke, Michael Katz, Konstantin Leyde, Lorenzo Speri, Rodrigo Tenorio and Xueting Zhang for their words of encouragement and lively chatter. There are far more people who deserve a mention here, but I'm going to be diplomatic and leave everyone out rather than trying to recall and list dozens of names. Thank you to the attendees and contributors of the CAPRA conference for the many lively discussions about extreme mass ratio inspirals I have enjoyed (both as a participant and as a witness from a distance). I'd also like to thank Simon "Silky Si McFly" Dick for putting on excellent pub quizzes every Tuesday night, which served to break up stressful weeks and reminded me that there is always more to learn (however useless that

knowledge may be).

It would be remiss of me not to thank my parents, Josephine Chapman and Christopher Bird, for the unconditional love, support and encouragement they have given to me over the course of my PhD (and for quite a while before that as well!). The knowledge that the two of you were rooting for me at every step gave me the motivation I needed to keep going and it is hard for me to express my gratitude for this with words alone. My mum's advice regarding how one should eat an elephant (one bite at a time) echoed in my mind as I tackled the writing of this ponderous tome head on, and I can confirm that it is a sound strategy. I should highlight that no elephants were harmed in the writing of this thesis.

Last, I would like to thank Phoebe Utting for her care, support and patience as I went through the many emotions that one goes through when writing a PhD thesis. Also, while he is not a person, I would like to thank my cat Tycho for his continued support (although I suspect this is in fact due to his reliance on me for sustenance and entertainment, and not a reflection of his interest in my work).

Declaration

The contents of this thesis are my own work, undertaken under the supervision of Dr Christopher Berry and Prof. Graham Woan at the University of Glasgow.

- Chapter 1 consists entirely of review material that has been referenced where appropriate.

- The work presented in Chapter 2 was undertaken in collaboration with (in addition to my PhD supervision): Dr Jean-Baptiste Bayle of the University of Glasgow; Dr Aurelien Hees and Dr Marc Lilley of Observatoire de Paris; and Dr Olaf Hartwig of the Albert Einstein Institute. These collaborators constructed the LISA simulation and data pre-processing software described in Section 2.1, which I assisted with in its later stages when I became involved with the project. The development of the analysis and statistical inference framework for use with this LISA simulation software was my own work. The applications of this framework presented in Sections 2.2 to 2.4 are my own work. Any additional software tools used during this work (such as waveform models or stochastic sampling techniques) are referenced where appropriate.

- Chapter 3 consists of my own work, developing from the existing and open-source software package FastEMRIWaveforms (FEW). Section 3.1 consists of review material that has been referenced where appropriate. The codes for the KerrEqEcc and PN5 waveform models introduced in that section (and used throughout the rest of the chapter) are under development by the FEW development team [1] as of the writing of this thesis. The theoretical basis of these waveform models is published material [2, 3] (and is referenced as such) and the in-development code is publicly available. I have been involved in this development work, but this is not the focus of the work presented in this chapter and these waveform models should not be considered as products of my PhD research in their own right. Rather, they serve as representative examples of more complex waveform models for which the methods introduced and developed in this chapter may be applied. The Teukolsky flux and mode amplitude datasets used in the KerrEqEcc waveform model were provided by Prof. Scott Hughes of the Massachusetts Institute of Technology. The methodology I develop in Section 3.2 and apply in waveform characterisation and data analysis in Sections 3.3 and 3.4 is my own work.

- Chapter 4 is my own work, developed in collaboration with my supervisors. A significant fraction of this work was published in [4]. I have expanded upon that which was published in Section 4.A, which consists of additional investigations that were not included in the published document. These subsequent investigations are my own work.

List of Acronyms

FEW FastEMRIWaveforms. xii, 64, 66, 92–95, 98, 120, 134, 139, 145, 159, 168, 169

AAK augmented analytic kludge. 95, 96, 145

AGN active galactic nuclei. 16, 17, 19–22, 132, 152

AM CVn AM Canum Venaticorum. 14, 56

AMP automatic mixed-precision. 104

ASD amplitude spectral density. 13, 45–47, 58, 61, 66, 74

BBH binary black hole. 2, 5, 6, 14, 20, 60, 62, 141, 152

BH black hole. 1, 5, 6, 12–14, 16–22, 54, 152

BHPT black-hole perturbation theory. 86

CBC compact-binary coalescence. 2, 5, 15, 84, 87, 90

CDF cumulative density function. 70, 71, 105–108, 141, 143, 149, 155, 157

CI credible interval. 70, 71, 156

CO compact object. 1, 14, 18–22, 83, 85–90, 94, 115, 116, 130, 137, 144, 150, 152, 153, 157, 163

CPU central processing unit. 62, 74, 98, 100, 104, 115–123, 130, 134

CV cataclysmic variable. 14, 22

DNS double neutron star. 14

DWD double white-dwarf. 14, 22, 35

EGB extra-Galactic binary. 14, 15

EM electromagnetic. 1, 2, 5, 13, 15–17, 20–22, 57, 60

EMR extreme-mass-ratio. 84, 89

EMRI extreme-mass-ratio inspiral. i, ii, vii, 13, 18–22, 29, 36, 38–40, 55, 64, 66, 67, 69, 73, 80, 83, 84, 86, 88–91, 93–100, 102–105, 107, 109, 110, 113, 115, 124, 126, 128, 129, 132–135, 137–153, 155, 157–165, 168–171

ESA the European Space Agency. i, 7, 42, 44, 79, 168

FFT fast Fourier transform. 50, 56, 62, 98, 116

FIM Fisher information matrix. 146, 147, 159–165, 170

GB Galactic binary. i, 13–16, 22, 35, 36, 38, 39, 43, 51–62, 73–76, 78–80, 168

GPU graphics processing unit. 26, 31, 39, 62, 63, 66, 92, 98, 100, 102, 104, 107, 109, 115, 117–125, 138, 151

GR general relativity. 1–3, 5, 14, 17, 21, 22, 52, 86, 104

GSF gravitational self-force. 83, 85–90, 95, 132, 168

GW gravitational wave. i, ii, 1–24, 28–30, 34–36, 38–45, 47, 48, 52–56, 60, 62, 64, 66, 68, 73, 76, 79, 80, 84, 87–90, 97, 104, 126, 127, 142, 167, 168, 171

IGR Institute for Gravitational Research. x

K-S Kolmogorov–Smirnov. 70, 71, 77, 78, 157

KL Kullback–Leibler. 161–163, 165

LDC LISA data challenge. 37, 39, 43, 50

LIGO Laser Interferometer Gravitational-Wave Observatory. 4, 5

LISA Laser Interferometer Space Antenna. i, ii, vii, xii, 7–23, 28, 29, 34–47, 49, 51–64, 66–68, 73–81, 83, 84, 97, 124, 126, 130, 137, 139, 142, 145, 146, 151–153, 157, 158, 167–171

LPF LISA Pathfinder. 11, 12, 37

LSA linear signal approximation. 146, 147, 161, 163, 165, 170

LTT light travel time. 10, 44, 47, 48

LVK LIGO–Virgo–KAGRA. 5, 6, 15, 20

MBH massive black hole. 6, 12, 16–22, 35, 36, 60, 62, 64, 83–90, 93, 94, 96, 97, 107, 109, 110, 115, 126, 134, 135, 137, 141, 144, 145, 151–153, 157, 158, 169, 170

MBHB massive black hole binary. i, 6, 13, 16–21, 35–39, 55, 60, 62–66, 73, 80, 168

MCMC Markov chain Monte Carlo. 26–28

ML machine learning. i, ii, 26, 29–31, 34, 39, 40, 83, 99, 168–171

MLP multi-layer perceptron. 30–34, 100, 102–105, 109, 132, 149, 151, 153, 157, 169, 170

MPR measured pseudorange. 10, 46–48, 79

MS main sequence. 14

NR numerical relativity. 84

NS neutron star. 1, 5, 6, 12, 14, 27, 28

ODE ordinary differential equation. 88, 89, 92, 93, 120, 125, 142

P–P probability–probability. 68, 70–80, 153, 155–158, 164, 168, 170, 171

PDF probability density function. 24–27, 58, 61, 70, 71, 141

PE parameter estimation. 24, 25, 27–29, 34–36, 38, 41–44, 47–50, 53–65, 67–80, 83, 89, 94, 98, 102, 115, 118, 119, 122–130, 132–135, 139, 159, 161–163, 165, 168–171

PN post-Newtonian. 55, 86, 95, 134, 145

PSD power spectral density. 12, 28, 34, 37, 41, 45, 50, 52, 53, 55, 57, 62, 79, 115, 146

PTA pulsar timing array. 6, 12, 40

RAM random-access memory. 54, 73, 104

RJMCMC reversible jump Markov chain Monte Carlo. 36, 38

SiLU sigmoid linear unit. 103, 149

SNR signal-to-noise ratio. 13, 15–18, 21, 29, 38, 52, 56–66, 74–80, 90, 94, 95, 130, 133, 135, 138, 139, 143–151, 157–159, 163, 167, 168, 170

SSB Solar–System barycentre. 44, 96, 97

SVD singular value decomposition. 147

TCB Barycentric Coordinate Time. 10, 42, 47

TDI time delay interferometry. i, 9–12, 37, 42, 44, 47–50, 52, 56–58, 60, 62, 65, 68, 73, 74, 79, 80, 97, 168

TPR true positive rate. 105, 106

TT transverse-traceless. 1, 52, 126

WD white dwarf. 12, 14

Chapter 1

Introduction

1.1 Gravitational waves

GWs are perturbations in spacetime produced by the acceleration of massive bodies. They are a prediction of general relativity (GR) (and any other special relativistic theory of gravity), which posits that these ripples in spacetime propagate at the speed of light [5]. The frequency of GWs is related to the motion of the source (at lowest order, the quadrupole of the mass distribution) which produced them [6]. For two bodies in a circular orbit, the GW frequency is therefore equal to twice the orbital frequency of the binary. The information presented in this introductory section is largely drawn from [7], which provides a comprehensive derivation of GWs and their properties from GR.

Any GW may be described as a superposition of plane waves. The passage of GWs deforms spacetime in an oscillatory fashion, with the magnitude of the deformation characterised by the strain amplitude of the gravitational radiation. This describes the shift in the relative positions of a ring of test particles from an initially circular configuration due to geodesic deviation [7]; GWs are transverse, so this motion occurs only in the plane perpendicular to the direction of GW propagation. Also predicted by GR is the residual deformation of spacetime which persists after a GW has passed, known as the gravitational memory effect [8]. Similarly to electromagnetic (EM) waves (i.e., light), GWs are polarised according to the orientation of the source with respect to the detector, which describes the shape of the periodic deformation of the ring of test particles. In GR, the only GW polarisations are the transverse-traceless (TT) tensor polarisations (described as the plus and cross polarisation states). The effect of a passing GW on a ring of test particles is shown in Figure 1.1.

While GWs are in principle produced by any accelerating mass distribution, the strain amplitude of the radiation is only significant when both the total mass of the system and its acceleration is sufficiently large. Such systems are some of the most extreme found anywhere in the Universe, including binary coalescences of compact objects (COs) (such as black holes (BHs) or neutron stars (NSs)), supernova explosions and rapidly-rotating NSs [7]. However, the number

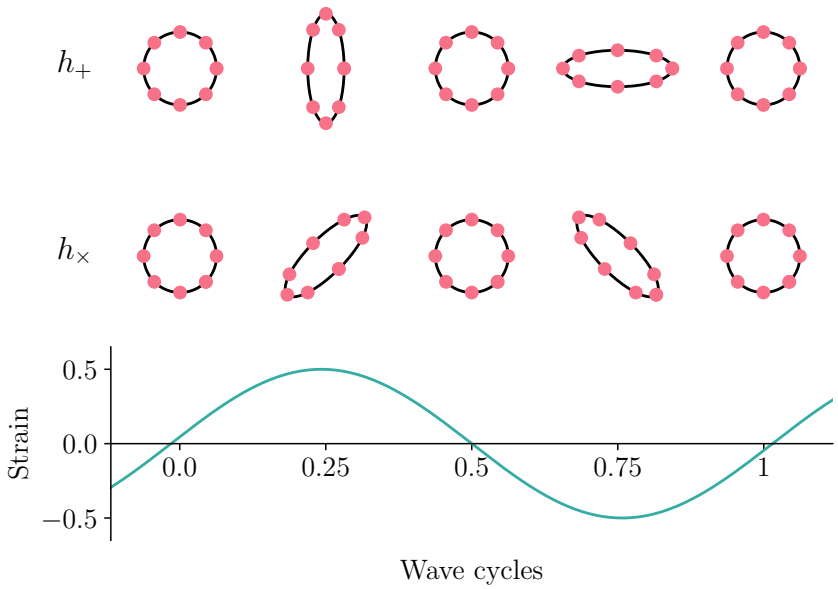


Figure 1.1: The plus (h_+) and cross ($h_\times$) GW polarisations, which are the only polarisation states of GWs in GR. A ring of test particles is deformed by a passing GW (incident to the page). The full GW strain is a weighted sum of h_+ and $h_\times$ according to the orientation of the source in the reference frame of the observer. The strain amplitude of the GW in this example is 0.5; the typical strain amplitude of loud astrophysical GW signals is $\lesssim 10^{-20}$.

density of these systems in the Universe is sufficiently low that sources are typically very distant (for compact-binary coalescences (CBCs), often in excess of 1 Gpc). The amplitude of GWs scales inversely with the luminosity distance d_L to the source, and is typically $\lesssim 10^{-20}$ by the time it reaches Earth. Compared to the size of the Earth, the deformation imparted by a GW of this amplitude is of the same order as the radius of a proton. Detecting the passage of GWs from their effect on matter is therefore extremely difficult.

Overcoming this technological challenge and observing GW signals is a research area of great interest to the scientific community, with new techniques for probing different areas of the GW spectrum under continuous development [9]. As a gravitational phenomenon, detecting and measuring GWs enables the underlying theory of gravity to be tested directly [10]. The dynamical evolution of a GW source is encoded in their gravitational radiation emission; solving this inverse problem by modelling the evolution of systems under GR and comparing with observation provides a new means of observing these systems that is completely independent of their brightness in the EM spectrum. Some GW sources such as binary black holes (BBHs) are typically invisible in EM (unless they occur in gas-rich environments), so GWs are the only way to observe their strong-field dynamics. Supernovae [11] and rotating neutron stars (often observed in the form of pulsars [12], but also as isolated systems via X-ray emission [13]) are EM-bright; in such cases, GW observations complement other measurement techniques to provide a more complete picture of the dynamics of these systems (multi-messenger astronomy). For instance, the mass dynamics within a supernova cannot be directly observed with EM be-

cause the collapsing star is optically thick, but the corresponding GW emission passes through the star almost unchanged and can be measured [7].

1.1.1 Indirect evidence for GWs

The first observational evidence for the existence of gravitational radiation was not based in GW detection, emerging instead from radio observations of the binary pulsar system PSR B1913+16 (which is more commonly known as the Hulse-Taylor pulsar, named after those who discovered it). First observed in 1974, pulsar timing revealed that the pulse arrival frequency of $\sim$17 Hz was modulated by an additional component of frequency $\sim$36 μHz (corresponding to a period of ~ 7.75 hr) [14]. This periodic variation in the pulse arrival time occurs because the pulsar is rapidly orbiting a compact companion; the radial velocity of the pulsar cycles each orbit, which couples to the pulse arrival rate by the Doppler effect. This effect is detectable due to the compact nature of the binary orbit, which has a projected semi-major axis (accounting for orbital inclination) of $(1.009\,122\,9 \pm 0.000\,000\,9)\,R_\odot$ [14, 15]. Observations of this binary system therefore provided one of the earliest views into the two-body problem of GR in the strong field, enabling relativistic effects on the binary's orbital evolution to be measured directly.

One such effect, which is small but accumulates over time, is the decay of the binary separation over time (and therefore a shortening of the orbital period) due to GW emission [16]. In studies of the original timing observations, the orbital period as a function of time was found to agree precisely with the predictions of GR over a 7 yr period [17]; these observations therefore served as indirect evidence for the existence of gravitational radiation. Further studies have continued to monitor the orbital decay of PSR B1913+16 over longer time periods. Analysis of timing observations spanning 30 yr confirmed that the orbital evolution was consistent with decay due to gravitational radiation at a level of $\sim 0.2\%$ [18]. This result was refined to a value of $\sim 0.1\%$ with the inclusion of measurements spanning a further 8 yr [15]. Other pulsars in compact binary systems have been observed to follow similar trends, including PSR J0737 − 3039 (the first double-pulsar system observed) [19] and the binary pulsar system PSR B1534 + 12 [20]. While the existence of GWs can be strongly inferred from these observations, the information contained in GW signals can only be exploited by measuring them directly. This important step from indirect to direct detection gives rise to the field of GW astronomy, in which the properties of GW sources are inferred through the analysis of their GW emission.

1.1.2 Ground-based laser interferometric GW detection

Detecting and measuring a GW signal is a monumental technological challenge. GWs interact with matter weakly; the passage of GWs can instead be inferred by their effect on spacetime (as outlined in Section 1.1). However, due to the large distances to most astrophysical sources of GWs, the strain amplitude of their emission is typically $\lesssim 10^{-20}$ upon reaching Earth. Measur-

ing such a small perturbation to background spacetime requires the use of extremely sensitive detection techniques, which includes the suppression of a large variety of instrumental and terrestrial noise sources [21, 22].

The first set of GW detectors to achieve this feat were the constituents of the Laser Interferometer Gravitational-Wave Observatory (LIGO) detector network. The LIGO detectors are laser interferometers in a Michelson configuration [21], which we will now describe. First, a laser beam is passed into a beam-splitter. The two split beams are then directed along two orthogonal detector arms of equal length before reflecting from mirrored test masses positioned at the ends of each arm. The reflected beams return to the beam-splitter, recombine and are measured by a photo-detector. In the absence of any perturbation which acts differentially on the length of the detector arms (i.e., simultaneously extending the length of one arm and reducing the length of the other), the phase difference between the split beams upon recombination remains constant. The plus- and cross-polarised GWs (see Section 1.1) incident orthogonal to the detector lead to differential arm motion, modifying the path length of the laser beams; this leads to a phase variation in the recombined beam that can be measured (and the encoded GW signal extracted) [23]. The detector configuration is tuned such that the split beams destructively interfere upon recombination to decouple variations in the output due to oscillations in phase difference from those due to fluctuations in laser power [23]. This is known as dark fringe operation (referring to the interference pattern of the recombined laser beams).

The phase difference between the laser beams accumulates due to a change in the effective arm length on the timescale of the beam's round-trip time (up and down the detector arm). The sensitivity of an interferometric GW detector therefore varies as a function of GW frequency. For a given frequency f, the path length of the split beams which optimises sensitivity is [23]

$$L \sim 750\,\mathrm{km}\left(\frac{100\,\mathrm{Hz}}{f}\right). \tag{1.1}$$

Above this length, the phase difference imparted by the GW begins to cancel out and the detector's sensitivity decreases. The frequency band targeted by the LIGO detectors is $\mathcal{O}(100\,\mathrm{Hz})$, which would require impractical 750 km arms. To achieve sensitivity at these frequencies with 4 km, an additional mirror is added to each arm after the beam-splitter to form Fabry–Pérot cavities in each arm [21]. Once laser light enters these cavities, it proceeds to reflect between the two mirrors a large number of times, significantly increasing the optical path length of the laser beam. This has the additional benefit of significantly increasing the laser power circulating in the laser arms. This mitigates the amplitude of photon shot-noise (which scales inversely with the laser power [23]) sufficiently for GW detection at high frequencies. Fabry–Pérot cavities are therefore an essential component of ground-based interferometric GW detectors, facilitating sensitivity at high GW frequencies with a detector of reasonable size.

In the lower frequency band of 10 Hz to 100 Hz, the effects of seismic displacement noise are

mitigated by suspending test masses with pendulum-based isolation systems [21, 22]. Seismic noise is attenuated passively because pendulums are harmonic oscillators; above the resonant frequencies of the system, noise power is quadratically suppressed as a function of frequency. The LIGO detectors employ four pendulum stages, multiplying this effect [22]. Other noise sources limit performance at lower frequencies, including radiation pressure noise and thermal fluctuations coupling to both the suspension pendulum fibres and test-mass coatings [22, 23].

The original LIGO detectors operated from 2002–2010, detecting no GWs [22], before being dismantled and replaced with newer technology. In 2015, shortly after completion and commissioning of the Advanced LIGO upgrades to both detectors [22], the LIGO detectors made the first direct observation of a GW signal [24]. The GW event, denoted GW150914 (the date of the event), was emitted by the coalescence of two BHs of masses $\sim 30\,\mathrm{M}_\odot$ at a distance of 470^{+160}_{-180} Mpc [25]. This detection provided direct evidence for the existence of BBH systems and enabled constraints to be placed on their merger rates [26]. The observed waveform matched the predictions of GR within measurement uncertainties, constraining theories of gravity in the strong-field regime [27]. The LIGO network would go on to make two more confident BBH detections in this observing run [28].

With the inclusion of the Advanced Virgo interferometer [29] in the detector network, another landmark detection was achieved in the second observing run in 2017: the coalescence of two NSs [30]. The merger of GW170817, observed with GWs, was accompanied by the detection of a near-simultaneous short-duration gamma-ray burst by the Fermi space telescope [31, 32]. GW events can be localised on the sky from analysis of the time delays between the signals observed in different detectors due to the finite speed of gravity, which provides an estimated three-dimensional localisation volume when combined with a luminosity distance measurement [33]. By combining this localisation with the pointing information from Fermi, the location of the source was determined with sufficient precision that EM follow-up successfully identified the merger's kilonova remnant in NGC 4993 [34]. This multi-messenger observation of a binary NS coalescence provided the opportunity for a plethora of astrophysical investigations that would not be possible with any of the constituent observations in isolation. New upper limits were placed on the speed of gravity [32]; the distance–redshift degeneracy was broken, enabling a measurement of the Hubble constant of $(68.6 \pm 11.3)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ after peculiar velocity corrections [35, 36]; and such prompt discovery of the kilonova remnant allowed for its evolution over a timescale of months to years to be examined across the EM spectrum [34], providing a complete picture of the complicated processes at work in binary NS mergers. The wide range of discoveries and scientific return from GW170817 underscores the importance of GWs as part of the observational toolkit of multi-messenger astronomy.

As of the writing of this thesis, the LIGO–Virgo–KAGRA (LVK) detector network has completed three observing runs with a fourth underway [37, 38]. The public catalogue of probable CBC detections currently numbers 90 events, comprised of 84 BBH mergers, 2 binary NS merg-

ers and 3 NS-BH mergers (with one further candidate that may either be a BBH or a NS-BH merger) [38]. The ongoing success of the ground-based GW detectors has fuelled the development of next-generation GW detectors, which hope to improve upon existing detector sensitivity by more than an order of magnitude [39, 40]. The field of GW astronomy is rapidly growing to form a foundational pillar of observational astrophysics [9].

1.1.3 Pulsar timing arrays

In addition to the low-mass sources accessible to the LVK detector network, a significant number of slowly-evolving MBHBs are expected to be present across cosmic time due to the mergers of their host galaxies [41]. Despite being far from merger, the large total mass of these binaries results in a sufficiently-high GW amplitude that these signals are expected to form a measurable astrophysical stochastic background in the nHz regime [42]. This frequency band may also contain GW emission due to rapid inflation and/or phase transitions in the early Universe [42].

Given the approximate scaling described by Equation (1.1), it is clear that this frequency band is out of reach of interferometry techniques, as this would require effective arm-lengths of $\mathcal{O}(1\,\mathrm{ly})$. This region of the GW spectrum may instead be probed with radio observations of millisecond pulsars. The stability of these pulsars (which rivals that of atomic clocks [43]) is so high that the subtle variation in the pulse arrival time due to passing GWs in the nHz regime (which is $\mathcal{O}(50\,\mathrm{ns})$) can be detected, given sufficient observation time. Similarly to the interferometry case, one encounters a variety of noise sources. The shifting interstellar medium between the telescope and the pulsar alters the dispersion measure [12], and therefore the pulse arrival time. This effect cannot be corrected for directly: instead, one may time many millisecond pulsars simultaneously in a pulsar timing array (PTA) and measure the correlations between them in a pairwise fashion [44]. The effects of a passing GW are common between the pulsars, whereas dispersion-related shifts are independent to each pulsar. The GW signal encoded in these pairwise correlations can be extracted to directly constrain the amplitude of the nHz stochastic GW background [45]. This offers a unique perspective of the massive black hole (MBH) population inaccessible to other techniques, and is therefore a useful input to models of these populations [46].

Multiple independent large-scale analyses of long-running PTA data-sets have produced results which match the expected form of a stochastic GW signal (the Hellings–Downs curve) to a compelling degree, with a statistical significance of $2\,\sigma$ to $4.6\,\sigma$ [47–50]. It is expected that a combination of these analyses will further improve the statistical significance of the correlation observed [48, 51], and may even enable close MBHB sources to be identified [52].

1.2 The Laser Interferometer Space Antenna

Many GW sources are expected to exist in the mHz band. These are described in Section 1.3; in this section, we will first describe how GWs may be observed in this frequency band. The principles of interferometric GW detection are still applicable, but ground-based laser interferometric GW detectors are completely infeasible for two principal reasons:

- From Equation (1.1), optimal sensitivity at e.g. 10 mHz will require optical path lengths of $\mathcal{O}(10^6\,\mathrm{km})$. Even with Fabry–Pérot cavities, this is far from achievable on Earth.
- Terrestrial seismic noise sources are prohibitively loud at mHz frequencies.

One must therefore consider the alternative of a space-based GW detector, where space is ample and seismic noise is absent. This is the motivation behind the LISA mission, which aims to produce and operate such an instrument.

In this section, we will describe how LISA will measure GWs. We first outline the mission profile: how the instrument is structured and operated. This is followed by a description of the procedure for detecting and characterising GWs with LISA. Last, we identify the limiting noise sources of the instrument, along with some sub-dominant noise sources that must be carefully controlled to maximise sensitivity.

1.2.1 Mission profile

LISA is a large-class ESA mission adopted as part of its Cosmic Vision science program, with a planned launch date in the mid-2030s [53]. It comprises three spacecraft in heliocentric orbits of radius 1 AU, arranged in a triangular constellation (of side-lengths 2.5 Gm) that leads Earth's orbit by $\sim 20°$. The constellation rotates as the spacecraft move along their orbital trajectories. Over a nominal mission duration of 4.5 yr (over which $\sim$4 yr of data will be produced), LISA aims to observe GWs in the 10^{-4} Hz to 1 Hz frequency band.

LISA is an interferometric GW detector, measuring GW strain by monitoring the relative separations of the three spacecraft using laser links. However, such a measurement must be free of non-gravitational forces such as drag from the solar wind [54]. To meet this requirement, each spacecraft houses two 1.92 kg test masses which act as reference points for separation measurements [53]. By monitoring the position of these test masses through capacitive sensing, the enclosing spacecraft (which is acted upon by drag forces) adjusts its trajectory with high-precision thrusters such that both test-masses remain in free-fall. The test masses are kept in vacuum chambers to minimise collisions with particles produced by spacecraft outgassing, and are constructed out of gold–platinum alloy to minimise magnetic interaction with the interplanetary magnetic field. Charge build-up in test masses due to cosmic rays (which would invalidate test mass free-fall due to interactions with the interplanetary magnetic field) is addressed with

ultraviolet radiation, which discharges the test masses via the photoelectric effect [55]. Collectively, drag-free control, ultra-high vacuum and active discharging are sufficient to reduce residual test-mass displacement jitter (with respect to free-fall) below the 10 pm requirements for LISA sensitivity to GWs [53, 56].

Each spacecraft is equipped with two 2 W, 1064 nm laser systems (one for each test mass) which are used to form laser links with the other two spacecraft [53]. Each laser is the input to an optical bench which manipulates the beam (redirection by mirrors and beam-splitters) in order to perform interferometric measurements. The main stages which constitute a laser link between two spacecraft (numbered 1 and 2 in this example) are as follows. First, the laser in spacecraft 1 is reflected by the test mass. Second, the reflected beam is expanded by a 0.3 m telescope and transmitted to spacecraft 2. Third, the transmitted beam (which has decayed in power to $\mathcal{O}(100\,\mathrm{pW})$) is magnified by the telescope of spacecraft 2. Fourth, this beam is reflected by the test mass (of spacecraft 2). Last, the reflected beam from spacecraft 1 and the to-be-transmitted beam of spacecraft 2 are combined and (through interferometric measurements on the optical bench) the phase of the resulting beam is recorded. The two beams are combined in order to measure their phase via heterodyning: while the laser frequency is $\mathcal{O}(10^2\,\mathrm{THz})$, the beatnotes of the combined beam are $\sim$50 MHz (due to Doppler shifting of the transmitted beam by relative spacecraft motion) which is sufficiently low to be sampled alias-free [53]. The signal from GWs enters this measurement by virtue of the fact that the distance travelled by the beam from spacecraft 1 varies with time due to passing GWs, introducing a variation in the phase of this combined beam that encodes the GW strain information. However, extracting this information is made difficult by the laser frequency noises of the two beams, which are independent of one another and introduce phase variations 8 orders of magnitude larger than those from GWs [53]. This is the reason for six laser links: by combining these phase measurements in a specific manner, laser frequency noise can be suppressed and the GW information recovered (see Section 1.2.2).

Scientific data is regularly transmitted to Earth from the constellation by radio link with one of the three spacecraft, which are each equipped with a 0.5 m antenna [53, 57]. Data from the other two spacecraft is transferred to the transmitting spacecraft by encoding it in modulations of the laser links. As the constellation rotates the designated transmission spacecraft changes [53]. Periodic repointing of the antenna to maintain alignment introduces scheduled gaps in the spacecraft data streams, which must be accounted for in subsequent post-processing and data-analysis (see Section 1.6.2 for discussion of the impact of these gaps). The transmitted data requires on-ground processing before GW signals can be extracted. We now outline this procedure and its motivation.

1.2.2 Measurement process

In Section 1.1.2, we described how ground-based Michelson interferometers measure differential arm-motion by recombining a split laser after it has travelled along the detector arms. A crucial element of measuring the recombined beam is that its phase is insensitive to the frequency noise of the laser: it is common between the two beams, as they were produced by splitting one beam from one laser source. This is only true for equal-armlength interferometers; if the arm-lengths differ, the laser frequency noise will not be temporally aligned on recombination and will not cancel out [58]. However, if we were to individually measure the phase of each beam, we can delay these phases and combine them to cancel the laser noise. To demonstrate this, consider the phase of the reflected beam from each arm when combined with the transmitted beam, which has phase $\phi(t)$ [59]. Accounting for the delay due to travel time, the beam from the first arm has phase

$$\phi_1(t) = \phi(t - 2L_1) - \phi(t), \tag{1.2}$$

where L_1 is the length of this arm. The combined beam sourced from the second arm is of similar form. A standard Michelson interferometer would combine these two beams to obtain

$$\phi_{\text{mich}}(t) = \phi_2(t) - \phi_1(t) = \phi(t - 2L_2) - \phi(t - 2L_1), \tag{1.3}$$

which is only zero (i.e. laser phase noise cancellation) for $L_1 = L_2$. However, by further delaying ϕ_1 and ϕ_2 to correct for the path length difference, we restore cancellation of laser noise:

$$\phi_{\text{TDI}}(t) = [\phi_2(t - 2L_1) - \phi_2(t)] - [\phi_1(t - 2L_2) - \phi_1(t)] = 0. \tag{1.4}$$

This is a simple example of TDI, where interferometric measurements are made in post-processing from the recorded phases of the individual beams [58]. A linear combination of time-delayed phase measurements (of which Equation (1.4) is a basic example) is known as a TDI combination. TDI combinations can be physically interpreted as the output of a virtual interferometer; it can be shown that any interferometric configuration can be constructed in this manner, making TDI a flexible and powerful solution for suppressing laser noise [60]. The TDI technique is what enables LISA to detect GWs [61], albeit with far more complex TDI combinations than Equation (1.4), consisting of nested delays of laser phase measurements from six laser links. The additional delays are necessary to correct for the flexing and time-varying arm-lengths of the LISA constellation [59]. The properties of the GW signals in the data may be inferred statistically (see Section 1.4) by applying the same TDI combinations to both the data and the signal model. As TDI incorporates the motion of the LISA constellation, the GW signal will be modulated over time; as this is direction-dependent, information regarding source localisation can be extracted from TDI data products.

Processing LISA data with TDI presents two immediate challenges:

1. The arm-lengths (and their evolution with time) are not known a-priori, and

2. The phase data must be expressed on a common time grid, which can then be interpolated to perform the delays necessary to compute TDI combinations.

The arm-length problem is solved by estimating the light travel time (LTT) of each laser link as a function of time. This is achieved by encoding a pseudo-random code in beatnotes of the laser links (which is performed simultaneously to the data transfer operation described in Section 1.2.1) [57]. On the receiving spacecraft, the received code is compared to a local version (all spacecraft have the same code) and the delay between the two codes is estimated by cross-correlation [62]. This delay is a precise measure of a combination of the LTT and the difference in time-frame between the two spacecraft (known as the measured pseudorange (MPR)), with typical errors $\mathcal{O}(1\,\mathrm{m})$ that are reduced to the sub-millimetre level with the inclusion of ground-tracking measurements of the constellation motion (estimated from telemetry with the transmitting spacecraft) [63]. Despite GWs only imparting a change of $\mathcal{O}(1\,\mathrm{pm})$ on the constellation arm-lengths [53], the requirement for sufficient laser noise suppression is $\sim$30 m [64, 65], which this procedure satisfies with a wide safety margin [66].

The data recorded and encoded by each spacecraft is time-stamped relative to an onboard ultra-stable local oscillator. These clocks provide a reference for the proper time of each spacecraft, which can be converted to the Barycentric Coordinate Time (TCB) time-frame with a relativistic transformation using knowledge of the spacecraft orbits [67]. A complication of this procedure is that the spacecraft clocks are not synchronised and will jitter independently, deviating from one another over time [68]. If clock jitter is left uncorrected, time-stamping errors on phase measurements enter TDI and constitute a limiting source of noise in the LISA band, even for state-of-the art oscillators [69]. Similarly to laser frequency noise, clock jitter noise can be suppressed if timing errors appear in multiple phase measurements at different times: this is enforced by modulating the transmitted laser phase according to the oscillator output, which produces two sidebands in the distant combined beam from which relative clock jitters can be obtained [70]. This information can then be folded into TDI to simultaneously correct for both laser and clock noises within LISA requirements [69, 71].

We now consider LISA noise sources post-TDI, which describe the sensitivity of the instrument to GWs.

1.2.3 Secondary noise sources

By applying TDI, the primary LISA noise sources, being laser frequency and clock stability noise, are greatly suppressed. The resulting instrumental noise floor comprises two secondary noise source groupings (each of which consist of a number of individual noise sources) that directly impinge on the sensitivity of LISA to GWs and therefore constitute the noise floor of

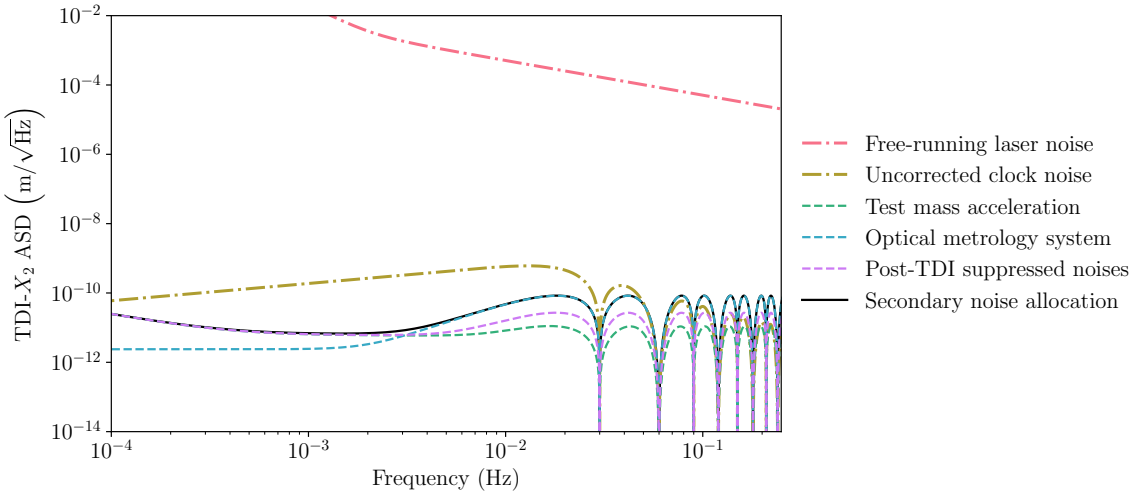


Figure 1.2: A summary of the LISA noise budget as they appear in the second-generation X_2 TDI combination. Dash-dotted lines denote primary noise sources, with laser frequency noise in a single laser link (red) and the clock timing noise (yellow); unless suppressed, these noises will limit LISA sensitivity. Individual secondary noise allocations are indicated as dashed lines, including the test mass acceleration (green) and optical metrology system (blue) noises. Post-TDI, the laser and clock noises must be within a similar tolerance to these noises (i.e., below the purple line). The sum of the two secondary noise sources (black line) is typically quoted as the LISA sensitivity curve for X_2.

the detector. The allocation to the LISA noise budget for each of these sources is shown in Figure 1.2; we describe them below.

The first collective noise source is residual test mass acceleration from imperfect drag-free control by the spacecraft, which directly couples to GW sensitivity as it violates the assumption of free-falling test masses [53, 56]. The capability of LISA to meet the allocated noise requirement was established by the LISA Pathfinder (LPF) mission in 2016 [72, 73]. This satellite consisted of a miniaturised LISA laser link; it was therefore insensitive to GWs, but was a fully-representative analogue for the test-mass acceleration noise that would be observed in the LISA spacecraft [74]. The drag-free control of LPF greatly exceeded expectations, with residual acceleration noise satisfying not only its own requirements but that of LISA as well across the entire LISA band [73, 75]. It is expected that test-mass acceleration noise will limit LISA sensitivity at frequencies below $\sim$2 mHz.

The other set of secondary noise sources is that of the optical metrology system, which is the consequence of noise in the numerous interferometry sub-systems in each spacecraft (mostly optical in nature, for instance scattered light or shot noise) along with thermo-mechanical noise coupling to optical measurements (e.g., adjusting intra-spacecraft path lengths or optics properties) [53]. The noise performance of the LISA interferometry systems has not been demonstrated en-masse, but its many components have been studied in isolation and their collective noise performance estimated via physical modelling. This informs the noise budget allocation of this noise source; it is expected to limit the sensitivity of LISA at higher frequencies [53].

In addition to these two noise sources, Figure 1.2 also displays the noise budget allocated to the post-TDI suppressed primary noise sources. Under nominal performance these noise levels should be roughly an order of magnitude below the overall secondary noise allocation for the entire LISA band (to the extent that they are usually ignored for LISA power spectral density (PSD) models) [76, 77]. However, if the LISA mission does not perform optimally, these noise sources may limit GW sensitivity. It is therefore essential to characterise the performance of LISA, and the post-processing algorithms (such as TDI) that will be used to parse the data it produces, under different operating scenarios. This is the motivation behind the development of increasingly sophisticated simulations of the LISA instrument, which is investigated further in Chapter 2.

Finally, it is expected that bursts of instrumental noise (known as glitches) will be present in LISA data [53]. Glitches associated with high-impulse momentum transfer to the test masses (primarily due to particles from spacecraft outgassing) were observed in LPF, lasting between seconds and hours [56]. It is reasonable to expect that the full LISA dataset will contain a number of glitches of varying morphologies and occurrence rates, as is observed in ground-based GW detectors [78, 79]. Care must be taken to ensure that data analysis algorithms and prospective studies of simulated LISA data sets are robust against the impact of such noise transients.

1.3 Sources of gravitational waves in the mHz band

The mHz band of the GW spectrum, which lies between the frequency bands probed by ground-based interferometers and PTAs, is expected to be occupied by many unique GW sources [53, 80]. The frequency of GWs from compact binary systems generally decreases as either the orbital separation or the total mass of the system increases. Compared to the $\mathcal{O}(1\,\mathrm{M}_\odot$ to $100\,\mathrm{M}_\odot)$ systems observed at higher frequencies, the sources in the mHz band generally include either slowly-inspiralling stellar-mass compact binaries (Section 1.3.1) or systems comprising at least one MBH with mass $M \gtrsim 10^5\,\mathrm{M}_\odot$ (Sections 1.3.2 and 1.3.3). The former evolve slowly in the mHz band, potentially over billions of years, whereas the latter enter this frequency band and chirp over a timescale of months to years. Astrophysical background signals produced by many weak sources overlapping in time are also expected in this frequency band (Section 1.3.4). In this section, we outline the principal source types expected to be observed by LISA; these source classes are summarised in Figure 1.3 (except for stochastic backgrounds, for visual clarity).

1.3.1 Stellar-mass compact binary systems

The majority of solar-mass stars exist in binary systems [81, 82]. As these systems evolve, processes such as common-envelope evolution [83–85] and stable mass-transfer [86–88] can lead to the formation of a hardened binary system containing at least one compact object (such as a white dwarf (WD), NS or BH). Depending on how specific mechanisms contribute to the

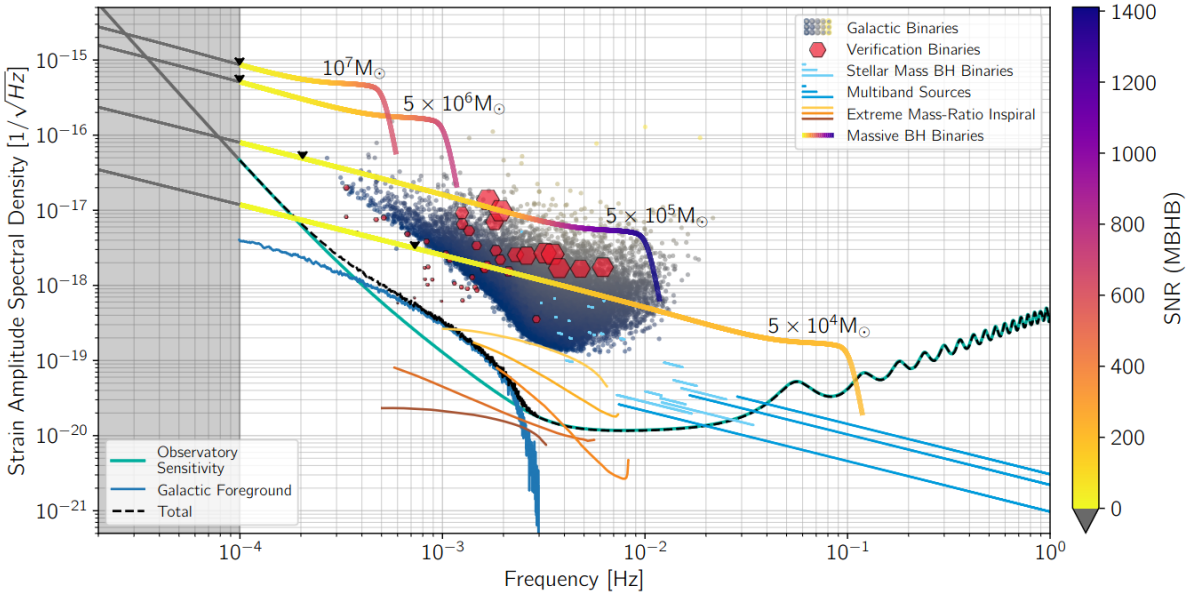


Figure 1.3: The sensitivity curve of LISA (as defined by secondary noise allocations, Section 1.2.3). Shown for reference are typical amplitude spectral densities (ASDs) for GW sources expected to be observed by LISA, which are described throughout Section 1.3. The evolving frequencies of a $z = 3$ MBHBs and the harmonics of a $z = 2$ EMRI are indicated by solid lines. The frequency at which each MBHB becomes detectable is indicated by a black arrow. Stellar-mass BH binaries are indicated by cyan lines; darker lines indicate systems which merge during the LISA mission (and therefore are multi-band sources if observed in ground-based detectors). Quasi-monochromatic GBs are indicated by points; their ASDs are not formally defined but are approximated here as $A\sqrt{f}$ for Fourier amplitude (frequency) A (f). Verification GB systems (known via EM measurements) are indicated by red hexagons (with sizes proportional to their signal-to-noise ratios (SNRs)). GBs with SNR < 7 are assumed to be undetectable, instead forming an unresolvable Galactic foreground of confusion noise (blue line) which limits instrumental sensitivity at mHz frequencies. Adapted from [53].

different stages of the formation of this binary, they can enter the mHz GW frequency band (corresponding to orbital periods $\lesssim 1$ hr) as either detached (interacting only under gravity) or interacting (undergoing mass transfer) systems.

Due to the close orbital separation of the binary at these orbital frequencies, detached systems consist of two compact objects (most commonly, two WDs [89]). As these systems lose angular momentum solely under GR, the chirp mass of the system may be directly inferred if the rate of orbital frequency evolution can be measured [90]. The compact binary components need not be interacting for their orbital evolution to deviate from GR in subtle ways. The presence of additional orbital bodies (e.g., for a triple system) can achieve a similar effect. CO triple systems emitting GWs in the mHz band are likely to be rare [91], but more commonly-occurring smaller bodies such as exoplanets will measurably alter the waveform of the system [92].

Systems undergoing mass transfer exhibit wider diversity than detached systems. Angular momentum transfer within the system leads to changes in orbital frequency that are decoupled from GW-driven evolution [93]. While this limits inference of the chirp mass, it enables binary interaction processes to be probed directly with GW observations. AM Canum Venaticorum (AM CVn) systems consist of a binary in which a WD is periodically transferred He-rich material from either a late-stage main sequence (MS) star or a companion WD [94, 95]. Similar in nature to AM CVn (but more commonly occurring in nature [96]) are cataclysmic variables (CVs): binary systems in which a donor MS star fills its Roche lobe and transfers mass to a WD [97]. A rare class of interacting compact binaries are ultra-compact X-ray binaries, in which the radiation-driven inspiral of a NS and either a WD or helium star after an X-ray binary phase is halted (after a significant decrease in orbital separation) by mass transfer to the NS driven by Roche lobe overflow [98, 99].

Comparing the relative abundances of these various GB subtypes, it is expected that double white-dwarf (DWD) binaries will form a distinct majority [80, 89]. In fact, the number of sources is expected to be sufficiently numerous that their collective emission will constitute a limiting noise source for LISA [100]: see Section 1.3.4 for further discussion of this phenomenon. Beyond the Galaxy, a smaller population of DWDs residing in Milky Way satellites such as the Magellanic clouds are expected to be observed by LISA [101, 102]. The loudest members of the extra-Galactic binary (EGB) populations of other Local Group galaxies such as M31 may also be detectable, with prospects for detection dependent on the duration of the mission [103].

LISA is also expected to observe $\mathcal{O}(100)$ GBs which are WD-NS, WD-BH or double neutron star (DNS) systems, albeit in far fewer numbers than DWDs [104]. As NSs and BHs are formed by supernovae, the resulting natal kick injects orbital eccentricity into the binary [105, 106]. Sources containing a light NS or BH may therefore be distinguishable from DWD binaries (which typically circularise during formation) via orbital eccentricity measurements [107, 108].

Rarer still are NS-BH and BBH systems. As the Galactic number density of these systems

is among the lowest of the stellar-mass compact binary subtypes, only a handful of these GB detections (if any) are expected to be made by LISA [109]. For EGB systems up to ranges of $\mathcal{O}(\mathrm{Gpc})$, the prospective number of detections improves to $\mathcal{O}(100)$ systems due to the larger spacetime volume being observed [110]. Despite the relatively weak SNR of these EGB signals, the duration over which they are observed enables precise constraints to be placed on chirp masses, eccentricity and sky localisation [111]. Of particular interest is the precision with which the coalescence time of the binary can be determined: depending on the remaining time to coalescence, this can range from $\sim$1 d to $\sim$30 s. This precise forecasting of CBCs enables multi-band analyses of these sources to be performed, with observations by both LISA and ground-based detectors such as the LVK network maximising the scientific potential of these sources [112, 113].

Among the population of GBs are systems which have been studied extensively via EM observations. These are known as verification binaries: the amplitude of their GW emission can be directly inferred and therefore their detectability by LISA known in advance of mission launch [114]. The sky locations of verification binaries are known to $\mathcal{O}(\mu\mathrm{as})$ precision [115], orders of magnitude better than can be achieved with GW observations of these systems. EM observations also place moderate constraints on the orbital frequency and inclination. This prior knowledge significantly improves the sensitivity of LISA to these systems, and the accuracy with which their waveforms may be extracted from the data stream. It is expected that GW observations will further refine the constraints on the binary orbital evolution placed by EM measurements; for the highest-frequency GBs, the GW observations may even outperform the EM measurements in this regard [116]. The number of known verification binaries currently sits at 55 [110, 117], which has grown significantly in recent years due to the extensive Galactic surveys produced by the Gaia mission [118]. The loudest of these sources are expected to be detectable in the LISA data stream after less than a month of continuous observation [119]. As the GW strain amplitude of detached verification binaries can be determined in advance, there is also scope for these sources to act as reference points for instrument calibration [120]. However, one must be careful in the assessment of the usefulness of these binaries as verification sources to include ambiguity from the unknown GB population [121].

GBs are a primary scientific target for the LISA mission. They provide a unique view of a stage of stellar evolution that is challenging to probe with EM observations. The role of common envelope and mass transfer processes in the evolution of compact binaries is not fully understood [122]; the large sample of GBs systems observed by LISA will enable high-fidelity parameterised models of these processes to be tested and constrained. The stellar structure of the Galaxy is well-studied local to the Solar System [123], but the combined effect of extinction and high stellar number density in the Galactic plane make studies of the Galaxy as a whole challenging with EM measurements [53, 110]. Unimpeded by the effects of extinction, LISA will probe the structure of the entire Galaxy by mapping the spatial distribution of GBs. Ob-

servations of BH–BH systems in late inspiral presents the opportunity to perform multi-band GW astronomy for the first time, studying the evolution of these systems over a timescale of decades [112, 113]. GBs are guaranteed LISA sources; the sheer number of these systems observable by LISA will ensure that the mission will return impactful scientific conclusions, even in the most pessimistic scenarios for other sources types.

1.3.2 Massive-black-hole binary coalescences

Empirical evidence suggests that virtually all galactic nuclei contain a MBH [124, 125]. The formation and evolution of MBHs is a prominent question in astrophysics [126, 127]. Observations of active galactic nuclei (AGN) indicate that MBHs existed early in the Universe's history, suggesting they formed on similar timescales to the first galaxies [128–130]. EM observations have identified that the MBH mass is strongly correlated with both the luminosity and the velocity dispersion of its host galaxy, which indicates that their evolutionary histories are closely linked [131–134]. As a significant fraction of known galaxies are elliptical [135] (which are thought to be the result of galaxy mergers [135, 136]), it is therefore reasonable to expect that MBHB systems are extremely prevalent across cosmic time.

Astrophysical processes such as dynamical friction with stellar and/or gaseous environments are expected to harden these systems from orbital separations of $\mathcal{O}(\mathrm{kpc})$ to $\mathcal{O}(\mathrm{pc})$ on timescales of $\sim$1 Gyr [137]. Around this point, energy exchange with loss-cone stars in three-body interactions begins to contribute significantly to the evolution of the binary [138] (often ejecting the stars at high velocities in the process [139]), making the inspiral rate of the binary particularly sensitive to its environment [140]. Initial investigations with simpler galaxy models suggested that depletion of the binary loss-cone would stall this stellar hardening process [141] (dubbed the "final parsec problem" [41, 142]), however more recent investigations have suggested this may not occur for more realistic galactic stellar density and velocity dispersion profiles [143–145]. Once the system reaches a separation of $\mathcal{O}(\mathrm{mpc})$, the dissipation of orbital energy due to GW emission is sufficiently rapid to significantly contribute to the evolution of the binary [16, 146]. MBHBs which reach this stage are expected to coalesce within a few Gyr, with the precise time-to-coalescence being strongly dependent on the parameters of the binary and their local environment [146–148]. The sheer number of galaxy mergers occurring throughout the Universe means that the existence of MBHBs at all stages of this evolutionary pathway is essentially guaranteed [149]. It is therefore not a question of if, but how many GW signals from MBHB coalescences are expected to be reaching the Solar System per unit time.

LISA will be particularly sensitive to MBHs in the $10^4\,\mathrm{M}_\odot$ to $10^7\,\mathrm{M}_\odot$ mass range. The MBHB merger rates in this range are expected to be $\mathcal{O}(1\text{–}1000\ \mathrm{yr}^{-1})$, with the vast majority of these being detected due to the high SNR of MBHB coalescence signals [150–152]. Merger rates are uncertain due to the multi-faceted nature of MBHB formation [80, 151], which occurs at the interface of many astrophysical processes at a stage of the Universe's history for which we

have limited observational information. For this reason, observations of MBHB coalescences with LISA are set to revolutionise the fields of MBH and galaxy formation astrophysics. The current sample of MBHs observations are extremely limited for $M \in 10^3\,\mathrm{M}_\odot$ to $10^7\,\mathrm{M}_\odot$ [153], which is the mass range over which LISA is most sensitive [77]. It is hypothesised that MBHs begin as $\sim 10^1\,\mathrm{M}_\odot$ to $10^5\,\mathrm{M}_\odot$ BH seeds which must grow significantly via accretion [154] and mergers [155] in the early Universe to account for the existence of distant quasars [156], but no observational methods currently exist which can probe the BH population at these redshifts. Astrophysical models of BH seeds, which are the foundations upon which MBH population models are built, are therefore poorly constrained at present (hence the large range of potential masses for these seeds quoted above) [126, 127, 157]. LISA aims to fulfil this dire need for observational data by opening an observational window on the seed BH population, enabling constraints to be placed on the nature of this vital stage of MBH formation for the first time.

By observing MBHB coalescences across a large redshift range, LISA will also probe how the MBH population evolves across cosmic time [158]. As this is expected to directly trace the galaxy merger population (and the galaxy population itself), this is an essential ingredient to models of galaxy evolution that is poorly understood at present. The cumulative impact of environmental effects (such as the presence of circumbinary disks) may also be detectable on the population level, further constraining how the environments of MBHs evolve with redshift [159].

The existence of AGN demonstrates that some MBH exist in gas-rich environments. If a circumbinary disk is present, luminosity changes induced by the inspiral of the binary system may be detectable pre-merger [148]. Modulations of accretion-driven X-ray emission due to orbital motion which are coherent with the waveform phase; if these two phenomena can be linked, the host galaxy of the merger can be identified [160]. The energy released by MBH mergers will disrupt accretion disks if they are present, which may lead to changes to observed spectra or jet structure coinciding with the merger of the system [161, 162]. The prospect of detecting EM signatures corresponding to MBH mergers makes them exciting multi-messenger sources. As GWs provide a direct measurement of d_L, an accompanying redshift measurement from EMs observations places constraints on the expansion of the Universe [163] (as was achieved with GW170817, see Section 1.1). The high redshifts ($z \sim 7$) to which both GW and EM signals corresponding to MBH mergers can be detected will enable strong constraints to be placed on parameters for matter density and the dark matter equation of state [164, 165]. For MBHB coalescences observed without EM counterpart, the dark siren [163] and spectral siren [166] techniques will supplement multi-messenger cosmological inference results. However, the limited high-redshift galaxy-catalogue completeness [167] and knowledge of structural features of the MBH population will limit the informativeness of MBH observations in this regard.

The high SNR with which LISA will observe MBH signals also makes them effective probes for modifications to the waveform. Of particular interest is the fundamental nature of BHs, which are predicted by GR to be described solely by their mass, spin and electric charge [7, 168,

169]. Deviations from this description due to alternate theories of gravity, or the existence of ultra-compact objects with no event horizon (such as boson stars [170] or fuzzballs [171]), will modify the quasi-normal mode spectrum of the ringdown phase of the waveform [172, 173]. LISA will measure the ringdown spectra of MBH systems to extreme precision, enabling the predictions of different gravity theories to be tested in the strong-field regime, and for the existence of these exotic objects to be investigated [53]. The environment will also impact the waveform [174]: the presence of a circumbinary disk will modify the orbital evolution of the binary, decoupling the waveform phase with respect to the strain amplitude of the GW emission [147]. By modelling how the presence of gas modifies the waveform, the properties of the circumbinary disks can therefore be studied [147].

MBHB coalescences are an untapped astronomical resource crucial to understanding many astrophysical and cosmological phenomena. Observations of MBHB systems by LISA will enable:

- The formation and evolution of MBHs from the elusive population of BH seeds to be studied for the first time;
- Population studies of unprecedented complexity tracing MBHs across cosmic time, in a poorly-understood region of the BH mass spectrum;
- New constraints to be placed on physics fundamental to cosmic expansion and the nature of gravity; and
- Multi-messenger studies of galactic nuclei environments to be performed.

These sources therefore constitute one of LISA's most informative (and exciting) astrophysical targets [53].

1.3.3 Extreme-mass-ratio inspirals

Nuclear star clusters in the centres of galaxies (where MBHs reside) are dense stellar environments, consisting of both stars and COs [175]. If the orbital trajectory of a CO comes sufficiently close to the MBH, the orbit will decay due to the emission of GWs and become an inspiral that eventually leads to the merger of the CO-MBH system. These EMRIs are long-lived GW sources, evolving in frequency through the mHz band over a timescale of years [176]. EMRIs can be highly eccentric and/or inclined, and therefore typically consist of hundreds of sideband harmonic modes. The phase evolution of these modes is extremely sensitive to the properties of the source, which enables precise constraints to be placed on the parameters of the system even at modest SNRs [177]. For further discussion of EMRI modelling, see Section 3.1.

A number of formation mechanisms for EMRIs have been proposed. Potentially the most significant of these is the loss-cone scattering channel, in which the dynamical evolution of the

nuclear star cluster in which the MBH resides leads to the scattering of COs on highly eccentric orbits (with eccentricities $e > 0.9999$) [178, 179]. For MBHs existing in wide binary systems (which are expected to occur frequently, see Section 1.3.2), the Kozai–Lidov mechanism will further enhance the rate at which these orbits are produced [180]. Supernova kicks may also place newly-formed COs on highly eccentric orbits without the need for dynamical interactions at all [181]. The majority of the scattered objects are expected to directly plunge into the MBH due to the orbital periapsis lying within the region of the MBH in which stable orbits are possible, but some COs by chance skirt this boundary and make complete orbital cycles at these high eccentricities. With each periapsidal passage, bursts of GW emission act to circularise this near-parabolic trajectory, resulting in a more modest eccentricity distribution; simulations with simplified stellar-cusp density models produced inspirals which had eccentricities $e \in [0, 0.9]$ [182] at an orbital period of $\sim 10^4$ s. While the burst emission is unlikely to be detectable by LISA for all but the closest galaxies [183–185], the resulting system is sufficiently compact to constitute a loud GW source in the mHz band. As the system will retain much of its inclination, systems formed by loss-cone scattering are expected to have randomly-distributed inclinations (due to the spherical symmetry of the stellar cusp) and a wide range of eccentricities, necessitating the development of generic EMRI waveform models (Section 3.1).

In the case of AGNs, EMRI formation can occur due to interaction between the accretion disc and orbiting COs or stars. The motion of the orbiting body induces density waves in the disc, leading to a loss of angular momentum and damping of both orbital eccentricity and inclination [186]. The loss of angular momentum leads to the inward migration of the orbiting body, producing EMRIs without the need for loss-cone scattering. As the orbital trajectory will have undergone significant damping before entering the LISA band, these inspirals are expected to be quasi-equatorial, quasi-circular systems (a highly unlikely combination from scattering-based formation). EMRIs produced by other formation channels are still possible in AGN, however due to the small fraction of MBHs residing in AGN these will constitute a small fraction of the EMRI population. Such systems would be particularly interesting due to the interaction of the accretion disc with the eccentric, inclined orbit of the CO in the LISA band, which may be detectable as a deviation in the waveform from one evolving solely under GW emission [187].

There are also formation channels which produce low-eccentricity ($e < 0.01$) but strongly-inclined EMRIs. Stellar-mass BH–BH systems orbiting within a MBHB may be tidally separated, resulting in a near-zero eccentricity inspiral if it is close enough to an MBH. Tidal stripping can remove the hydrogen envelopes of red giant stars, with the remaining helium core inspiralling towards the MBH [188]. As the core undergoes tidal heating, orbital eccentricity decreases and (provided the core is not tidally disrupted) the system will have low orbital eccentricity upon entering the LISA band [189]. The distinct combination of low eccentricity and arbitrary inclination may allow EMRIs formed by these channels to be identified amongst a larger catalogue of EMRI observations.

A key feature of EMRI formation channels is the existence of a relaxed stellar cusp surrounding the MBH, which acts as a source of COs that form EMRIs when they (by one means or another) are directed onto an inspiralling trajectory [190]. However, (as was noted in Section 1.3.2), MBHs frequently merge with other MBHs. This disrupts the stellar cusp, which must then reform and undergo sufficient dynamical evolution to re-assume its relaxed state [190]. During this period of cusp erosion and regrowth, the EMRI rate is highly suppressed due to the lack of COs in the neighbourhood of the MBH. Due to the poorly-constrained nature of MBHB merger rates as a function of both MBH mass and redshift, EMRI rate estimations have wide uncertainties as a result [190, 191]. Predicting the EMRI rate is also made difficult due to other poorly constrained astrophysical inputs. The event rate is particularly sensitive to the rate of direct plunges with respect to inspirals for scattered COs and the form of the MBH mass population as a function of redshift [190, 191]. Estimates of the loss-cone EMRI formation rate span three orders of magnitude, however even the most pessimistic astrophysical models predict that LISA will observe a handful of EMRIs [190]. EMRI formation rates in AGN are expected to be comparable (or slightly higher) than the loss-cone case [192], but these predictions depend on the fraction of MBHs in AGN as a function of redshift which is not well-constrained over the MBH population. In addition to uncertain astrophysics, current EMRI rates have been computed under steady-state assumptions which are unlikely to be valid in practice [190]. Work is ongoing to account for the time-evolution of the MBH with its stellar environment, which introduces an additional layer of complexity in assessing the impact of cusp erosion and regrowth on the overall EMRI rate [191].

The exceptional measurement precision of EMRI observations makes them effective probes of many astrophysical phenomena. As described above, the formation channel of a given EMRI system will be encoded in the orbital parameters of the system; as these can be measured as well as one part in 10^6, the channel which produced a given EMRI signal may be inferrable in some cases. Measuring the branching fractions between different formation channels would enable the relative impact of their underlying astrophysical processes on the EMRI rate to be studied. EMRI observations will predominately be at redshifts $z \in [0,3]$, providing a complementary view of the MBH population to that provided by MBHB coalescence observations. The population of stellar-mass COs probed by EMRIs is also distinct from (but potentially overlapping with) that observed by the LVK network as BBH coalescences [193]. EMRIs observations may therefore enable the breaking of degeneracies which arise when relating these subsets of the BH population to the overall population. The relative precision with which the MBH spin can be inferred, which is $\mathcal{O}(10^{-6})$ [176], is far more precise than what is obtainable with other GW or EM observations of MBHs [194, 195]. EMRIs observations will therefore address the dire need for constraints on the MBH spin population. As the MBH spin is closely linked to the evolutionary histories of galaxies [196, 197], these exquisite spin measurements will have resounding implications for the wider astrophysical community.

The sensitivity with which changes in waveform phase can be measured in EMRIs systems also enables sources of deviation from a two-body GR waveform to be probed with high precision [198]. This can be the result of either environmental effects or an actual deviation in the theory of gravitation from GR. Modifications to GR will impact both the orbital evolution of the CO and the resulting GW emission. EMRIs have been demonstrated to be effective probes of numerous modified gravity theories [199–202]. Deviations of the MBH spacetime from that of the Kerr metric can be determined to one part in $\sim 10^3$ [190, 203, 204]. Weak constraints on the existence of COs with a non-zero dipole moment can also be established with EMRI observations [205]. For exotic massive central compact objects which lack of a horizon, their reflectivity to GWs (zero for BHs) can be determined to one part in 10^6, the most precise constraint identified for GW observations of any compact binary system type [206]. A non-zero graviton mass will lead to GW dispersion, modifying the propagation of GWs in a frequency-dependent manner [207]. As the phasing of each mode of an EMRI waveform will be uniquely effected due to their differing frequencies, they may be particularly sensitive probes of such an effect [208]. GW-damping effects over cosmological distances, a phenomenon common to many modified-gravity theories, can be constrained to within a few percent with LISA observations of EMRIs [209].

Environmental effects measurable with EMRI observations include the presence of gas (from e.g. an accretion disc) or tidal perturbations. Gas imparts additional drag forces on the CO as it inspirals, modifying its trajectory and the resulting GW emission; by measuring this deviation, the properties of the accretion disc can be directly inferred [186]. The presence of nearby stellar-mass objects (stars of COs) can be inferred from resonances in the inspiral trajectory, probing the system's local stellar environment [210, 211]. Dynamical friction due to the presence of a dark-matter halo [212] or dark-matter spike [213, 214] also introduces measurable deviations to the evolution of EMRI systems.

While EMRIs are excellent probes of beyond-GR phenomena, measuring and characterising these effects must be performed with care. Correlations between different effects will lessen the precision with which any individual effect can be measured [215]. It is likely however that perturbations of the waveform due to environmental effects will not be fully degenerate with those predicted by modified gravity theories, enabling them to be separately identified [186].

As long-lived GW sources, EMRIs can be localised on the sky to far better precision than MBHB mergers due to the long duration over which the varying antenna response modulates the signal [190, 216]. This makes EMRI observations effective cosmological probes as part of the dark siren method. A few high-SNR EMRI detections will enable $\mathcal{O}(1\%)$ and $\mathcal{O}(10\%)$ constraints to be placed on the Hubble constant and matter density parameters respectively [217, 218]. Folding in formation channel inference (based on orbital eccentricity measurements), dark siren analyses with AGN catalogues will bolster results obtained with galaxy catalogues [219]. EMRIs occurring in gas-rich environments may produce detectable EM counterparts, potentially

enabling the host galaxy to be directly identified. Although rare, observing a strongly-lensed EMRI signal and identifying its host galaxy through EM follow-up will enable the Hubble constant to be inferred with percent-level precision [220].

Given the plethora of astrophysical phenomena made accessible by EMRI observations, they are a key component of the LISA mission objectives. Precision measurements of MBH mass and spin will enable high-fidelity MBH population models to be constrained, probing the formation and evolution of MBHs and their host galaxies across cosmic time. By observing the strong-field orbital dynamics of the CO over a timescale of years, deviations from the predictions of GR can be probed with exquisite sensitivity. As candidates for multi-messenger astronomy in some formation channels [221], observing the GW signal from EMRIs will complement EM observations of AGNs to break degeneracies and provide a more complete picture of these systems. EMRI are unique sources; fully exploiting their rich scientific potential will entail the development of bespoke techniques for their modelling (Chapter 3) and subsequent analysis with inference frameworks (Chapter 4).

1.3.4 Stochastic signals

As LISA is sensitive to a wide array of sources, numerous quieter members of these source populations which are too weak to be individually resolved will combine to form stochastic signals. The most significant of these signals will come from the GB population [53]. These sources are expected to be sufficiently numerous that bright signals, which would be detectable by LISA in isolation, will be overlapping in frequency to such an extent that they cannot be individually resolved. The GB sub-populations (particularly the DWDs [100, 222] and CVs [96]) will therefore form stochastic foreground signals (with an amplitude larger than the instrumental noise floor) which will ultimately limit the sensitivity of LISA in the 0.1 mHz to 3 mHz band [116]. The optimal extraction of signals from this Galactic foreground will require the use of computationally-costly inference techniques [116, 223]. As the foreground signal of the Galaxy will be anisotropic (with many sources located near the Galactic centre), the modulation of the signal by the LISA response will violate the assumption of stationary noise, further complicating LISA data analysis efforts. However, characterising this foreground signal will probe the GB population, supplementing the results obtained from the bright signals which can be isolated from the foreground. The modulation of the Galactic foreground signal will enable its amplitude spectrum to be probed as a function of sky position, roughly mapping the density of GB sources across the Galactic plane [224]. This will also allow it to be separated to some degree from other anisotropic foreground signals (such as the GB population of the Magellanic clouds [102]) and isotropic background signals.

Other LISA sources are expected to form stochastic background signals. The background signal from BH–BH systems may be detectable in optimistic scenarios [225]. An area of active research is the background signal from EMRIs. Predicting the amplitude spectrum of the EMRI

background is challenging due to rate uncertainties (see Section 1.3.3) [226]. For fiducial estimates, the background is expected to be detectable; in optimistic rate scenarios, it may even constitute a foreground signal in the 1 mHz to 10 mHz band [185, 227].

In addition to stochastic backgrounds associated with individually-detectable LISA sources, there is the possibility of detecting background signals associated with more exotic sources. Some of the more plausible candidates include cosmic strings [228, 229], early-Universe inflation or first-order phase transitions [230] or ergoregion instabilities in horizonless compact objects [231].

The signal-dominated nature of LISA data (see Section 1.6.1 for more discussion) makes it particularly important to account for and model stochastic backgrounds which may be present in the data stream. As the power spectra of these background signals directly affects the parameter estimation of foreground sources, they must be properly considered when assessing prospects for source identification and measurement precision with LISA. Conversely, stochastic backgrounds are challenging to separate from instrumental noise; this is particularly impactful in signal-dominated data where the noise properties cannot be directly quantified in isolation. Recent work has identified that measuring the parameters of stochastic backgrounds in LISA data will depend strongly on the precision with which the instrumental noise can be modelled; in the worst case of weakly-modelled noise, up to two orders of magnitude in sensitivity to stochastic backgrounds is lost [232]. Accurate characterisation of LISA noise will therefore be essential if the full astrophysical potential of stochastic signals is to be exploited.

1.4 Bayesian statistics

Extracting information from GW signals in detector data is inherently probabilistic due to the presence of noise. Our goal is to state a hypothesis (e.g., that a signal is present in noisy data, or that the signal is described by a particular model), incorporate the information encoded by observational data, and examine to what degree the data supports this hypothesis. This requires us to specify a statistical framework within which to express this process.

One such framework is Bayesian statistics [233], in which probability (denoted P) is interpreted as the degree of belief in a particular statement. Probability describes our state of knowledge regarding a statement, which changes when we receive new information regarding the assertion of the statement in question [234]. This process of updating our degree of belief given an observation (when done correctly) is data analysis; it is a foundational principle of the scientific method that underpins humanity's collective understanding of the Universe.

A fundamental component of Bayesian statistics is Bayes' theorem, which describes how the probability of a statement A changes based on observational data B. Bayes' theorem states that [234]

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}. \tag{1.5}$$

The prior probability $P(A)$ describes our degree of belief in A before the observation has been made. The likelihood $P(B \mid A)$ (where $\mid$ denotes conditionality) is the probability that the data B would be observed given A (i.e., if A were true). The evidence $P(B)$ is a normalisation term dependent only on the data and not the statement in question. The posterior $P(A \mid B)$ is related to these three quantities by Equation (1.5) and describes our degree of belief in A given the data B. A key facet of statistical inference is that we should not assert that A is true (or false), given B, with absolute certainty. The posterior probability is a complete representation of our state of knowledge regarding A given B and should be considered as such when presenting conclusions drawn from data analysis [234].

Bayesian probability theory provides us with the mathematical tools necessary to perform logical reasoning in quantitative terms, folding new information into our existing state of knowledge to assess the probability of any statement we wish to assess. We now discuss how this general framework enables us to fit parameterised models to observational data.

1.4.1 Parameter estimation and model selection

In GW astrophysics, we are usually concerned with a specific form of statement: given data $\vec{d}$ that contains a signal of interest, we ask what parameters $\vec{\theta}$ of a model M best describe $\vec{d}$? This is known as parameter estimation (PE) [235].

Let us outline how our degree of belief in statements of this form are described and updated. Typically, $\vec{\theta}$ are not discrete and vary continuously over some pre-defined range. For instance, the mass of a star m is continuous (for all intents and purposes) and non-negative. Our degree of belief in the value of m is represented by a probability density function (PDF) $p(m)$. This non-negative function has the property that the probability of m taking a value within the range $[m_{\min}, m_{\max}]$ is [234]

$$P(m > m_{\min} \,\&\, m < m_{\max}) = \int_{m_{\min}}^{m_{\max}} p(m)\,\mathrm{d}m. \tag{1.6}$$

By construction, PDFs are normalised such that the total probability over the parameter space is 1, viz. [234]

$$\int p(m)\,\mathrm{d}m = 1. \tag{1.7}$$

A multivariate PDF is a function of multiple parameters, and may be written as a product of univariate PDFs if and only if the parameters are statistically independent of one another [236]. Multivariate PDF can be reduced in dimensionality by the process of marginalisation, which produces a PDF agnostic to the values of the marginalised parameters. For instance, a two-dimensional PDF over m and the star's radius R may be re-expressed as [236]

$$p(m) = \int p(m,R)\,\mathrm{d}R. \tag{1.8}$$

For PDFs conditional on additional parameters which are not of interest (e.g., the distance to the star D), these parameters can similarly be marginalised over; it can be shown that [236]

$$p(m) = \int p(D)p(m \mid D)\,\mathrm{d}D. \tag{1.9}$$

Probability densities for parameters are updated via Equation (1.5) in the same manner as for discrete statements. Introducing more commonly-used notation, we may rewrite Equation (1.5) as [236]

$$p(\vec{\theta} \mid \vec{d}, M) = \frac{\pi(\vec{\theta} \mid M)\mathcal{L}(\vec{d} \mid \vec{\theta}, M)}{\mathcal{Z}(\vec{d} \mid M)}, \tag{1.10}$$

where $\pi(\vec{\theta})$, $\mathcal{L}(\vec{d} \mid \vec{\theta})$ and $\mathcal{Z}(\vec{d})$ are the prior, likelihood function and evidence respectively. The evidence is also known as the marginalised likelihood: because it normalises the posterior PDF, it is defined as [236]

$$\mathcal{Z}(\vec{d}) = \int \pi(\vec{\theta})\mathcal{L}(\vec{d} \mid \vec{\theta})\,\mathrm{d}\vec{\theta}. \tag{1.11}$$

The conditional with respect to the model M is usually omitted for brevity if the model is fixed for the problem at hand. However, typically we have a range of methods with which to model the data. Equation (1.10) will provide a posterior PDF for each case, but in some scenarios comparing the models themselves (rather than their parameters) is of greater interest. Therefore, given models M_i, a useful question to answer is which model has the greatest degree of support with respect to the data? It is easy to think of models that fit a given dataset arbitrarily poorly (for instance, fitting a straight line to an exponential curve), so this must be quantifiable in some way by our statistical framework.

Bayesian model selection is easily performed in principle, provided one can compute the evidences corresponding to each model. The models are then compared by computing the posterior odds [234]

$$\mathcal{B}(M_1, M_2) = \frac{p(M_1 \mid \vec{d})}{p(M_2 \mid \vec{d})} = \frac{\pi(M_2)}{\pi(M_1)} \frac{Z(\vec{d} \mid M_1)}{Z(\vec{d} \mid M_2)}, \tag{1.12}$$

where we have again used Bayes' theorem to rewrite the posterior probabilities.

The techniques of Bayesian PE and model selection are essential components of statistical analysis in astrophysics. While we have outlined the mathematical principles underlying these techniques, their application to real problems is not straightforward, as we will now explore.

1.4.2 Stochastic sampling

A challenging aspect of Bayesian probability theory is that, despite their simple form, applying Equations (1.10) and (1.11) is challenging and computationally expensive in practice. The number of parameters $\vec{\theta}$ can be very large: even a simple sinusoid cannot be described by fewer than 3 parameters without loss of generality, and astrophysical models are typically far more complicated. The posterior PDF may be a complete description of the information encoded by the data,

but it is not typically an analytic function and is therefore not particularly useful to us in this representation. One may express the posterior distribution in a more easily manipulated form by producing a discrete set of samples distributed proportionally with respect to the posterior PDF.

Unfortunately, this presents its own difficulties, as efficiently sampling generic PDFs is not straightforward. Multiple general approaches have been developed for this solving this problem. One of the most commonly applied methods is Markov chain Monte Carlo (MCMC), in which a Markov chain is constructed with the PDF to be sampled as its equilibrium distribution [236]. The chain begins at a random point in parameter space and iteratively explores the space. At each step, a random point for the chain to move is proposed by sampling from a proposal PDF (which is usually of a simple form that can be efficiently sampled) and the probability of the chain transitioning to this new point is computed. It can be shown that provided this transition probability satisfies detailed balance (that is, it is identical under reversal), then the posterior PDF can be made the equilibrium distribution of the chain by construction [234]. This process produces posterior samples by recording the position of the chain each iteration. While MCMC is a serial algorithm, evaluating a large number of chains is an embarrassingly parallel problem, so significant efficiency improvements are readily achievable with multi-threading or hardware acceleration techniques (such as the use of graphics processing units (GPUs)) [237, 238].

There are three central issues which can limit the performance of MCMC methods. First, one cannot obtain a set of independent samples by simply taking every sample in the chain, due to the effects of autocorrelation [239]. Instead, one sample is recorded for every τ MCMC iterations, where τ is the autocorrelation time. Reducing τ therefore improves the computational cost of sampling. Second, naive MCMC implementations require significant tuning of the proposal distribution to account for differences in scale-lengths between parameters [240]. Last, the use of overly simple proposal distributions that are agnostic to the specific PDF being sampled significantly hampers efficiency [240].

Modern MCMC algorithms address these challenges in a number of ways. Ensemble sampling methods incorporate the positions of many chains during the proposal stage to improve the mixing of each chain, which reduces τ and lowers convergence times [237]. By constructing these proposals to be affine-invariant, the issue of differing parameter scales can also be avoided [237, 239]. Variations of MCMC such as Hamiltonian Monte-Carlo [241] and Langevin Monte-Carlo [242] achieve lower autocorrelation times by incorporating derivatives of the likelihood function to increase the parameter-space coverage of the chains per iteration. ML techniques such as normalising flows have been applied to approximate the structure of the posterior PDF from the density information provided by the chains, and may be sampled from efficiently be construction [243, 244]. As this produces flexible proposal distributions in a data-driven manner, it is applicable to a wide variety of problems.

The high dimensionality of typical parameter spaces makes the evidence integral, Equation (1.11), difficult to perform accurately and computationally expensive. Commonly-employed

techniques for numerical integration (such as Gaussian quadrature or trapezoidal methods) typically require the posterior PDF to be evaluated an infeasible number of times to achieve good convergence. The evidence can be estimated using the MCMC of thermodynamic integration [245]; however, accurately determining the evidence with this method can significantly increase the number of MCMC chains required, making the analysis prohibitively expensive [246]. An alternative approach designed specifically to compute the evidence is the NS algorithm, a stochastic numerical integration technique well-suited to computing multidimensional integrals [247–249]. In NS, the evidence integral is rewritten as

$$Z(\vec{d}) = \int_0^1 \mathcal{L}(X)\,\mathrm{d}X, \tag{1.13}$$

where we have expressed the likelihood as a function of the prior volume. The prior volume X is a measure of the fraction of the integrated prior probability, such that

$$\mathrm{d}X = \pi(\vec{\theta})\,\mathrm{d}\vec{\theta}. \tag{1.14}$$

This change of variables renders the integral one-dimensional, allowing it to be solved by standard numerical integration techniques (such as those mentioned above), provided that the prior volume contained within a specified likelihood contour can be accurately estimated. The NS algorithm performs this estimation by iteratively updating an ensemble of points in parameter space (known as live points) [247]. To begin, a sample of N_{live} live points are drawn from the prior. At each iteration, new points in parameter space are proposed until one is found with a higher likelihood than the lowest-likelihood live point. That live point is then recorded and removed from the ensemble. As this continues, one considers contours of higher likelihood which by construction enclose increasingly smaller prior volumes. By comparing the reduction in volume enclosed by each successive contour (beginning with the entire prior volume, $X = 1$), the prior volume corresponding to a set of increasing likelihood values is obtained, which provides the necessary information to integrate Equation (1.13) over the prior volume outside of the smallest contour [249]. The evidence integral over the remaining prior volume can be estimated; the algorithm is usually terminated once this quantity crosses a pre-defined threshold [247].

A major benefit of NS is that samples distributed according to the posterior distribution are obtained as a by-product of the algorithm [247]. NS has therefore seen extensive application not only in model selection analyses, but as a general method for PE [248]. As the NS algorithm is inherently sequential, improving performance via parallelisation is more challenging than for MCMC, but can be achieved either via batched likelihood evaluation (which comes with some redundancy) [250] or the combination of independent NS runs [251]. NS typically achieves better performance in multi-modal problems without the need for significant tuning, leading to a greater efficiency per parallelisation-unit. Similarly to MCMC, the performance of NS relies heavily on the choice of proposal distributions. By excluding regions of the parameter space

in which a live point would never be accepted (i.e., it is outside of the likelihood contour for the worst live point), the number of likelihood evaluations can be reduced significantly. Data-driven proposal distributions such as fitted ellipsoids [252, 253] or normalising flows [254, 255] have successfully been applied to this task, improving the convergence time of NS by orders of magnitude [248].

The use of either MCMC or NS to perform a particular PE analysis is to an extent a matter of preference. One particular benefit of NS is the termination criterion, which enables PE runs to be performed with minimal supervision as they have a well-defined stopping point. In this thesis, NS is used to perform $\mathcal{O}(100)$ PE analyses of injected signals in parallel for this reason; this work is described in Sections 2.3 and 2.4. However, there are problems for which MCMC is better-suited due to the flexibility of the algorithm (which requires only that the transition probability satisfies detailed balance). An example of such a problem in GW data analysis is the LISA global fit, which is discussed in more detail in Section 1.6.1.

1.4.3 The GW likelihood

In GW data analysis, our goal is to measure the parameters $\vec{\theta}'$ of a GW signal in the time-domain data stream $d(t)$ from a detector. To perform Bayesian parameter estimation, we must first specify the form of $\mathcal{L}(d \mid \vec{\theta})$. We begin by assuming that $d(t)$ consists of a signal $h(t, \vec{\theta}')$ (described perfectly by our signal model, with parameters $\vec{\theta}'$) embedded in stationary, additive Gaussian noise $n(t)$ with PSD $S_n(f)$. To estimate the parameters of the signal, we generate a candidate template $h(t, \vec{\theta})$. Our hypothesis is that this template matches the signal in the data stream; if this is true, then by our assumption of additive noise

$$d(t) - h(t, \vec{\theta}) = n(t) \iff \vec{\theta} \equiv \vec{\theta}'. \tag{1.15}$$

The likelihood of our hypothesis is therefore the consistency of this residual with our noise assumptions. This is complicated in the time-domain because of correlations in noise between different data points. However, provided the noise can be treated as stationary, each bin in the frequency domain is statistically independent [256]. As the noise is assumed to be Gaussian with zero mean, the likelihood for frequency bin f_i is written as [257]

$$\mathcal{L}(d \mid \vec{\theta}) = \frac{1}{\sqrt{\pi S_n(f_i)}} \exp\left(-\frac{2[\tilde{d}(f_i) - \tilde{h}(f_i)]^* [\tilde{d}(f_i) - \tilde{h}(f_i)]}{2 S_n(f)} \mathrm{d}f \right), \tag{1.16}$$

where $\tilde{x}(f)$ is the Fourier transform of $x(t)$, $\mathrm{d}f$ is the frequency bin spacing, $*$ denotes the complex conjugate and we have dropped the dependence of h on $\vec{\theta}$ for brevity. The factor 2 that is included in the numerator of the exponential (and excluded from the denominator of the normalisation factor) accounts for the single-sidedness of $S_n(f)$. As the joint likelihood of independent observations is simply the product of the likelihoods of each observation, the

likelihood for the data stream is therefore the multivariate normal distribution [258]

$$\mathcal{L}(d \mid \vec{\theta}) = \frac{1}{\sqrt{\det|2\pi S_n(f_i)|}} \exp\left(-\frac{\langle d-h \mid d-h\rangle}{2}\right). \tag{1.17}$$

We have defined the noise-weighted inner product [23]

$$\langle x \mid y\rangle = 4\mathcal{R}\left\{\sum_{i=0}^{N} \frac{\tilde{x}(f_i)^* \tilde{y}(f_i)}{S_n(f_i)}\,\mathrm{d}f\right\}, \tag{1.18}$$

where $\mathcal{R}\{x\}$ denotes the real part. A further factor of 2 appears because $\{x, y\}$ are real, so the sum over negative and positive frequency bins is equal to twice the sum over the positive frequency bins. This template-based fitting technique, known as matched filtering [23], can be shown to optimise the SNR [23]

$$\rho = \sqrt{\langle h \mid h\rangle}, \tag{1.19}$$

which parameterises both the detectability of GW signals and the precision with which their parameters may be estimated (see Chapter 4 for an example of this in practice).

Equation (1.17) has seen widespread application in the analysis of simulated [259] and real [38] GW detector data alike, and has been implemented in numerous GW inference frameworks [246, 260–263]. We make extensive use of this likelihood function throughout this thesis:

- In Chapter 2, Equation (1.17) is used in the validation of LISA PE for multiple GW source classes.
- Chapter 3 focuses on EMRI waveform modelling. The accuracy of these models is assessed with Equation (1.17) via PE and waveform mismatch analyses.
- Both Equations (1.17) and (1.19) are key to the work presented in Chapter 4, in which the EMRI SNR function is approximated with ML methods and an approximate form of Equation (1.17) is employed in the simulation of EMRI catalogues.

The frequent application of matched-filtering techniques in this work underscores its wider importance in the field of GW inference.

1.5 Machine learning

ML is the construction and tuning of (typically large) statistical models for the performance of a specific task [264]. A sub-field of ML is supervised learning, in which the model approximates a relationship between input and output data by fitting it to an existing dataset of input–output pairings (referred to as training) [264]. These models can vary enormously in complexity, from straight-line fits (linear regression) to neural networks with trillions of tunable parameters. Supervised learning methods are a powerful way of modelling relationships

in data [264]. Statistical models trained via supervised learning are capable of generalisation, in which the predictions of the model are accurate even for input data which the model had no access to during training: this makes supervised learning a useful scientific tool for producing empirical models from observation alone. This framework is generic and versatile, applicable to data of arbitrary complexity [264]. Supervised learning can also be used to produce models which emulate a forward process of known form [265, 266]. This is useful in cases where the forward process is computationally expensive and must be evaluated many times. The upfront cost in data generation is large, but it is only incurred once; this cost is then amortised over the many evaluations of the trained model, which will typically be orders of magnitude more efficient than the original forward process. Supervised learning is therefore a useful tool for both the development of new techniques and the acceleration of existing methods.

Applying ML methods to GW data analysis and inference is an area of growing research, especially in the detection of GW sources [267, 268] and rapid estimation of their parameters [269–272]. See [273] for a review which explores the breadth of ML applications to GW science.

Supervised ML techniques are applied in this thesis to solve classification and regression problems in Chapters 3 and 4 respectively. In this section, we establish the base computational framework common to both of these applications. We first define the type of neural network used. We then outline the procedure of fitting these networks to datasets and describe how the trained network's performance can be assessed.

1.5.1 Multi-layer perceptrons

Multi-layer perceptrons (MLPs) are a form of feedforward neural network consisting of neurons arranged in a sequence of L layers [264, 274, 275]. A neuron is a structure described by a real number which represents its activation state. Each neuron in a layer is connected to each neuron one layer ahead of it by a linear mapping (which consists of a weight w and bias b) followed by a non-linear function $f(x)$ known as the activation function. The numeric value of a neuron n in position i of layer $j+1$ is therefore described by

$$n_{j+1,i} = f_{j+1}\left(\sum_{k\in N_j} n_{j,k} w_{j,k,i} + b_{j,i}\right), \tag{1.20}$$

where N_j are the number of neurons in layer j of the MLP, f_j is the activation function of layer j and $w_{j,k,i}$ and $b_{j,i}$ are the layer-j weight and bias parameters of the linear mapping from neuron k of layer j to neuron i of layer $j+1$. This recursive definition of a MLP is completed by defining the values of the neurons in the first layer of the network. These neurons form the input layer of the MLP; the activation state of these neurons determines the behaviour of all subsequent layers by definition of Equation (1.20), with the state of the neurons in the final (output) layer being the

result of recursively applying Equation (1.20) over all layers. In this framework, by setting the input layer values to those of a set of input parameters, the MLP defines a forward mapping from an input space of dimensionality N_0 to one of dimensionality N_L. The nature of this mapping is determined by the free parameters $w_{k,i}$ and b_i. These are the trainable parameters of the MLP. The number of parameters in a given MLP scales quadratically with the number of neurons per layer and grows quickly; for instance, a MLP with 5 layers of 32 neurons has more than 5000 trainable parameters.

Tunability would appear to be sufficient for a MLP to model relationships of arbitrary complexity, provided there are enough parameters to tune, but this is not the case. It can be shown that arbitrarily many sequential layers of linear combinations can be represented as a single layer of linear combinations, which in turn means an MLP of this form would only be capable of performing linear mappings [264]. This issue is avoided by the inclusion of the non-linear activation function $f(x)$, upon which this single-layer representation cannot be performed and the MLP can model non-linear relationships. Many non-linear activation functions have seen use in the ML literature [276, 277]. Not all activation functions may be well-suited to a given problem [276], and some simpler activation functions (such as the rectified linear unit [278]) have undesirable properties such as discontinuous higher derivatives or narrow input ranges. Activation functions are typically determined through experiment with respect to the problem at hand.

A key benefit of MLPs is the efficiency with which they may be evaluated. Equation (1.20) can be represented as a matrix multiplication, an operation that highly optimised algorithms have been constructed to perform [279–282]. Incorporating GPU vectorisation, evaluating MLPs with millions of trainable parameters can be performed in milliseconds [4]. Further performance improvements can be achieved by feeding multiple sets of input parameters into the network at once; this is an embarrassingly parallel operation for which GPU are well-suited, and the resulting efficiency improvement can reduce the per-input computational wall-time to $\mathcal{O}(\mu s)$ [4].

While tunability may in principle allow for extremely flexible non-linear mappings to be constructed as MLPs, the sheer number of possible parameter configurations for all but the most simple neural networks means that identifying a useful mapping by guesswork or brute-force checking is completely infeasible. Identifying a set of values in parameter space which allow an MLP to accurately represent a given mapping is an optimisation problem which must be solved numerically. We now describe how this optimisation is performed.

1.5.2 Fitting machine learning models to data

Our goal is to identify values of the trainable parameters $\vec{\theta} \equiv \{\vec{w}, \vec{b}\}$ for which an MLP mapping $f_{\theta}(x) \equiv f_{\theta}(x \mid \vec{\theta})$ is a close approximation of a true (and potentially unknown) mapping $y(x)$. We must first define quantitatively what a close approximation means. In supervised learning, this is represented by a loss function $J(x \mid \vec{\theta})$ which describes how close in value $f_{\theta}(x)$ and $y(x)$

are [264]. The form of this loss function is problem specific; for example, in a least-squares regression problem it is the L2 norm [264]

$$J(\vec{\theta}) = \overline{(y(x) - f_\theta(x))^2} \approx \frac{1}{N_x} \sum_{i \in N_x} [y(x) - f_\theta(x)]^2 , \tag{1.21}$$

where the overline denotes the mean over all inputs x. In reality, it is not practical to evaluate this function over all x; instead, we may compute a Monte–Carlo approximation of $J(\vec{\theta})$ over N_x discrete inputs (known as the training dataset) [264]. Equation (1.21) takes the minimal value of 0 when the MLP output matches that of the true mapping over all inputs.

Training is the process in which Equation (1.21) is minimised with respect to $\vec{\theta}$. As the dimensionality of this parameter space is typically large, this minimisation problem is not straightforward. An effective technique for performing this optimisation is stochastic gradient descent [283]. Gradient descent is a local minimisation algorithm which iteratively adjusts the parameters $\vec{\theta}$ by a step

$$\delta\vec{\theta} = -\gamma \frac{\mathrm{d}J(\vec{\theta})}{\mathrm{d}\vec{\theta}}, \tag{1.22}$$

where γ is a training hyper-parameter often referred to as the learning rate. The gradient of the loss function with respect to each parameter of the network is computed efficiently via the back-propogation algorithm [284]. Stochastic gradient descent performs the same procedure, but as the loss function is computed by a Monte-Carlo approximation the gradient is approximate as well. The gradient in Equation (1.22) does not need to be determined particularly accurately for an effective $\delta\vec{\theta}$ to be obtained [283]. This is useful because computing the network output over all training data is typically computationally expensive. The training data are therefore typically split into batches of size N_b [264]; this increases the number of iterations per loop over the training data (known as an epoch) by a factor N_x/N_b. With each iteration, the algorithm explores the parameter space and the loss function gradually decreases. This continues until either a convergence criterion is satisfied or maximum number of epochs is reached, at which point the algorithm terminates. Modern optimisation algorithms (such as the Adam optimiser [285]) essentially perform this stochastic gradient descent operation with some modifications to improve the efficiency and robustness of the training process.

For most problems, the amount of training data that is available or can be feasibly generated is finite in size. If the neural network to be trained is sufficiently large, training can lead to an undesirable mode of operation for the resulting network in which the good performance on training data is not maintained when the network is shown new examples. This inability of the network to generalise is known as overfitting [286]. It is challenging to predict when a neural network will overfit to training data. To combat overfitting, validation data is typically included in the training process; once per epoch, the network output for inputs not included in the training data is produced and the loss computed. By comparing how this validation loss evolves with respect

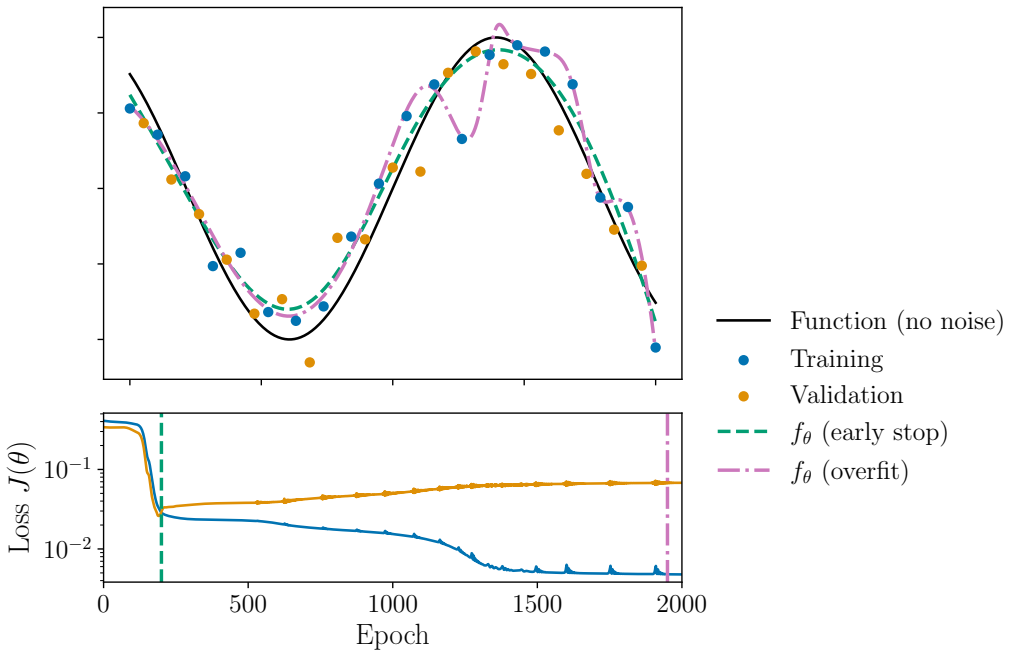


Figure 1.4: An example of overfitting in the training of neural networks. Noisy samples of a one-dimensional function (black) are split into training (blue) and validation (yellow) datasets. The upper panel shows how, during training, the network overfits to the noise in the data and learns a mapping which generalises poorly (purple). This phenomenon can be identified by examination of the loss curves (lower panel). By monitoring the validation loss and halting training when the validation loss begins to increase, a more reasonable solution is obtained (green) that more closely matches the original function.

to the training loss during training, overfitting (which corresponds to an increasing validation loss, i.e. the network is becoming worse at generalising) can be detected and training can be halted [287]. This early-stopping protocol is effective in mitigating the effects of overfitting. An example of the effects of overfitting and its mitigation via early-stopping are shown in Figure 1.4 for a simple one-dimensional regression problem.

In summary, the training procedure for an MLP is as follows:

1. The available data is split into training and validation datasets, with a ratio based on the number of data points and the accuracy with which network validation must be performed [264]. The training data is further divided into batches.

2. The network output for a batch is generated and fed into the loss function.

3. The optimiser uses the gradient of the loss function to modify the parameters of the network.

4. Steps 2–3 are repeated until the entire training dataset has been processed, completing the epoch.

5. The network output for the validation data is generated and fed into the loss function; if an early-stopping criterion is satisfied (e.g., this loss has increased for 5 epochs) training is terminated.

6. Steps 2–5 continue until either early-stopping or a maximum epoch count is reached.

Throughout this thesis, all MLPs are trained in this manner. This is performed with the POPLAR package [288], software which I designed as part of the work towards this thesis. The underlying ML framework of POPLAR is the PYTORCH package [281], which is highly optimised and has been used extensively to perform ML-related tasks in the literature.

1.6 Challenges in LISA data analysis

Currently, the field of LISA data analysis is undergoing rapid growth; the launch of the LISA mission is roughly a decade away and strategies for maximising its scientific return are still being formulated [53]. While GW data analysis as a field has been developing steadily for half a century (and has been applied to actual interferometer data for more than a decade), LISA presents unique challenges that will require new techniques to be developed. In this section, we outline why LISA data is different and how the work presented in this thesis will contribute to the goal of analysing it.

1.6.1 Analysing signal-dominated datasets

At present, ground-based detector data are only sufficiently sensitive to observe $\mathcal{O}(\text{min})$ (and often $\mathcal{O}(\text{s})$) of a GW signal every few days. The probability that two signals will overlap in time to a non-negligible degree (given this approximate signal duration and mean event rate) is therefore extremely low, so each signal can be treated independently; after a candidate signal is identified, a short segment is excised from the data stream and the signal is analysed. As noise is assumed to be stationary over this timescale, the PSD is obtained from data immediately prior to the event and is treated as known for PE. Therefore, the computational cost of analysing multiple GW signals in the data stream scales linearly with the number of signals. This form of analysis is sufficiently lightweight for a catalogue of ground-based GW detections to be built up one by one [38, 193]

However, this approach is not feasible for LISA. Our assumption that the data consists of a single signal in Gaussian noise no longer holds when the data contains multiple loud signals. When ignoring the presence of additional signals, biases can enter PE because the best-fit signal template must necessarily be adjusted to account for the excess power in the data stream. The degree to which we can expect the presence of a second signal h_2 to affect the inference of a

signal h_1 is parameterised by the overlap [289]

$$\mathcal{O}(h_1,h_2) = \frac{\langle h_1 \mid h_2\rangle}{\sqrt{\langle h_1 \mid h_1\rangle\langle h_2 \mid h_2\rangle}}, \tag{1.23}$$

which is unity when $h_1 = h_2$ and zero when the signals can be treated independently. Under our simplifying assumptions of stationary Gaussian noise over a single data segment, two signals will have an overlap of zero if they do not occupy the same frequency bins over the duration of the data stream. In practice, effects such as spectral leakage mean that this overlap is never zero; however, if $\mathcal{O}(h_1,h_2) \ll 1$ then biases can be expected to be negligible. The overlap at which biases will be comparable to statistical biases is dependent on the signals in question; see [290] for a recent investigation of this problem in the LISA context. In the worst-case scenario, all GW signals in the data stream must be simultaneously fit; this is known as a global fit, and it is a PE problem of immense computational expense and complexity. As we will now explore, this is unfortunately the scenario expected for LISA.

It is expected that the Galaxy contains $\mathcal{O}(10^8)$ GBs (mostly DWDs). The Doppler modulation of GWs by the LISA constellation will smear out the power in each of these signals to occupy ~ 30 frequency bins [291]. Most of the GBs will lie in the 10^{-2} mHz to 10 mHz band, which consists of $\mathcal{O}(10^6)$ frequency bins: by the pigeonhole principle, we therefore expect numerous signals to overlap in frequency simultaneously, so virtually no signals can be analysed independently. However, the narrow frequency band GB signals occupy enables this problem to be segmented into segments in the frequency domain of width $\mathcal{O}(\mu\text{Hz})$ (in which a maximum of $\mathcal{O}(100)$ loud sources can be expected to be present. This reduces the parameter space dimensionality of any one PE analysis to $\lesssim 10^3$; separability of the GB population via frequency-domain segmentation is therefore vital if LISA data analysis is to be computationally feasible [292]. For regions of the spectrum densely packed with GB signals, the information lost due to the correlation between different sources is so great that it effectively acts as a source of noise known as confusion noise (see Section 1.3.4). LISA data is signal-dominated, with confusion noise limiting sensitivity for a significant fraction of the LISA spectrum [222].

For MBHBs, overlap between signals is typically low, so analysing each MBHB individually does not lead to significant parametric biases or lost sensitivity unless their mergers are near-simultaneous [290]. However, we must also consider source confusion between different source classes. The overlap between an MBHB signal and a GB signal taken at random is small; however, the sheer number of GBs in the LISA data stream accumulates to produce significant biases in the recovered MBHB parameters if they are left unaccounted for [194, 293]. The bias incurred from each overlapping GB shifts the recovered parameters randomly, so the accumulated bias from the entire population has mean zero [290]. However, the variance grows with the number of sources, leading to inconsistent PE which significantly degrades the information obtainable from MBH signals [290]. The logical conclusion of this predicament is that MBHBs

and GBs must be fit simultaneously. This is unfortunate as MBHs signals are broadband and therefore the entire LISA band would need to be analysed at once, which as previously mentioned is not feasible due to the number of GBs. Similar conclusions likely also hold for EMRIs (which are also broad-band).

Current approaches to the LISA global fit avoid this problem by performing Gibbs sampling to individually sample the conditional distributions of each source class [294] (from which the joint distribution over source classes can be obtained). After performing an initial fit of the GBs in narrow frequency segments [116], they are subtracted from the data and the MBHBs in the residual are fit [295]. This process continues for other source classes (e.g., EMRIs and stochastic signals), concluding with an estimation of the instrumental noise $S_n(f)$. The initially-subtracted GBs are then re-added to the data (which consisted of only of the residuals from each subtraction stage), re-fit and re-subtracted. The cycle then repeats until the global fit result is no longer changing significantly.

A particularly challenging aspect of performing the global fit is that the total number of sources in the data is unknown and must be inferred alongside the source parameters. As the source count is a property of the assumed signal model, this is a model selection problem. Naive approaches to model selection such as grid-based evidence comparison is computationally expensive even for simple problems (Section 1.4), and is wholly impractical for an analysis of this scale. Global fit pipelines avoid this problem by instead using reversible jump Markov chain Monte Carlo (RJMCMC) methods [116, 296, 297], which explore both the space of possible models and the parameter space of each model in order to reduce the time spent evaluating likelihoods for unfavourable models. Unlike catalogues of ground-based GW observations (which are fixed in size once the criteria for a statistically significant detection are established [38]) this algorithm cannot always return a catalogue of fixed size due to the uncertain number of total sources. This makes summarising the posterior samples returned by RJMCMC for use in later analyses difficult. One approach for producing an approximate catalogue is to compare overlaps between different sources in the posterior distribution, retaining only sources that are of high posterior probability and individually resolvable, and consigning any remaining candidate sources to the astrophysical foreground signal [259, 296]. An alternative possibility (which is arguably the correct choice) is to use the RJMCMC samples in downstream analyses directly, but constructing these analyses around a model-dependent posterior distribution is not straightforward.

The LISA global fit is a problem for which a complete analysis strategy has not yet been developed. Solutions have only been demonstrated for simple LISA data scenarios in which the variety of sources in the data is significantly reduced (to just GBs and MBHBs) and for data with minimal complications regarding instrument operation [116, 259, 296, 298]. Future work towards global LISA data analysis strategies should push these boundaries with the eventual goal of demonstrating consistent PE for realistic LISA data streams. We now discuss the form

of these extensions, and how the work presented in this thesis contributes to their development.

1.6.2 Dealing with real instrument data

At present, full investigations of the LISA global fit problem have only examined relatively simple scenarios for the form of the LISA data stream [259, 295, 296, 298]. Specifically, these analyses have focused on a data stream containing signals embedded in additive stationary Gaussian noise with no data gaps. These two assumptions dramatically simplify the problem, because they are what enable use of highly efficient frequency-domain likelihood functions (Equation (1.17)). A drift in the noise performance of the instrument introduces off-diagonal terms to the noise covariance matrix (i.e., frequency bin noise properties become correlated) which returns the likelihood to a full matrix operation (as opposed to a product along the matrix diagonal). Data gaps introduce similar challenges, completely violating the assumption of stationary noise and introducing significant spectral leakage which must be addressed with careful tapering of each gap. Noise modelling currently consists of spline fits to the LISA PSD, which is taken to be of the smooth form presented as the noise allocation in Figure 1.2; in reality, spectral lines are expected (as were observed in LPF) which introduces rapid changes in noise magnitude that must be modelled carefully. The LISA noise behaviour will be further complicated by glitches (similarly observed in LPF [56]), which must be fit simultaneously to astrophysical sources due to the signal-dominated nature of LISA data (albeit transient sources such as MBHBs may be separable from glitches if they are well-separated in time [299]).

Work is ongoing to address these issues. The LISA data challenges (LDCs), which act as a test-bed for prototype LISA analysis pipelines, have recently been extended to include gaps and glitches in the data stream as part of the *Spritz* challenge [300]. Formally, the requirement for a full matrix operation alters the complexity of an N-point likelihood function from $\mathcal{O}(N)$ to $\mathcal{O}(N^2)$, which would make analysis of LISA data (where $N \sim 10^7$) completely infeasible. However, recent developments in time-frequency [301] and/or wavelet-based methods [302–304] for waveform generation and likelihood evaluation have shown promise in mitigating the impact of non-stationary noise, breaking the problem into day-long segments over which assumptions of noise-stationarity are a better approximation. Alternatively, the likelihood function can be modified from the assumption of Gaussian noise to improve robustness against noise transients in the data stream [305].

While these effects can be directly imposed on simulated TDI observables to approximately produce data with realistic LISA-like features (e.g., directly zero-ing out data to produce a gap), this is not a full representation of the problem. These alterations are emergent properties of changes to the instrument itself, which will couple from their source to the LISA data streams in a complicated manner that is challenging to model accurately. This is the motivation for the development of increasingly sophisticated simulations of LISA that aim to construct observational data from the ground up, modelling the performance of individual subsystems and their interac-

tion in the operation of the instrument [306, 307]. Recent simulation models for LISA have been extended to fully include the effects of generic orbital configurations, as well as the inclusion of a more complete description of instrumental noise source [308]. This model is an area of active research and development [309], an important component of which is the implementation and validation of PE for the variety of sources LISA is expected to observe (Section 1.3). The culmination of this work is presented in Chapter 2, where PE is performed for GBs, MBHBs and EMRIs and validated for a simplified scenario of the GB case. Chapter 2 is foundational to a future LISA simulation and analysis pipeline that will characterise the performance of the instrument under a variety of operating conditions. This will act as a test-bed for the development of data analysis algorithms that are sufficiently robust to be relied upon for the analysis of the real LISA dataset.

1.6.3 Efficient and accurate waveform modelling

Identifying and characterising an individual GW signal typically entails the generation of millions of waveform templates. When fitting multiple signals at once (as in LISA data), the per-signal cost increases substantially as the RJMCMC explores the large parameter-space volume and the number of waveform evaluations required. It is therefore imperative that waveform templates be produced as efficiently as possible. Four particular features of LISA data make this especially challenging:

- The nominal duration of the LISA mission is 4 yr at a sample cadence of 0.25 s, and therefore consists of $\mathcal{O}(10^7)$ data points after resampling to a cadence of 10 s. This makes Equation (1.18) computationally expensive.

- GW sources in the LISA band (Section 1.3) are long-lived, with lengths of order mission duration. This means that GW sources must be modelled over long timescales, with the resulting waveform also being evaluated at $\mathcal{O}(10^7)$ points.

- In addition to their long duration, LISA is expected to measure these GW sources with high SNRs (particularly EMRIs and MBHBs, which may have SNRs in the thousands). Waveform models must therefore be highly accurate to maximise sensitivity and avoid systematic biases in inference.

- The LISA response function, which varies with respect to time, must be applied to each waveform template.

Together, these facets of LISA waveform generation incur significant computational expense.

A number of techniques have been developed for mitigating their impact, many of which rely on generating waveforms and applying the instrument response entirely in the frequency domain [92, 194, 291, 310]. A generic LISA response function has not yet been developed

for use in the frequency domain, so approximations are made in the application of these techniques [291, 310]. GBs evolve sufficiently slowly to merit the use of a fast–slow decomposition in their waveform modelling, in which the frequency-domain response is computed via Fourier transform and applied to a frequency-domain waveform [291]. The process for MBHB waveforms is more complicated, but a frequency-domain response has been developed for these signals as well [216, 310]. Further performance improvement is possible for these waveforms via interpolation in the frequency domain to minimise the number of waveform evaluations [150, 194]. Another area in which significant efficiency can be gained is in the use of accelerator hardware such as GPUs. Batched waveform generation and likelihood evaluation yields greatly improved efficiency for both GB [92] and MBHB [194] systems. While further development of waveforms for GB and MBHB systems is needed to accommodate some of the more challenging aspects of LISA data analysis discussed in Section 1.6.2 (such as data gaps or non-stationary noise, which are challenging for frequency-domain analyses [302]), they are sufficiently rapid for the analysis of current LDCs to be computationally feasible to perform [259, 295, 296].

The development of EMRI waveforms for LISA data analysis is in a less mature state. These signals are among the most challenging to generate [311]; they are long-lived, broadband signals that contain rich harmonic mode structure. Recent work has enabled these waveforms to be generated in $\mathcal{O}(100\,\mathrm{ms})$ via GPUs for systems of limited physical complexity [177, 312]. However, computational bottlenecks inherent to current EMRI waveform generation techniques limit further performance improvements; to make matters worse, introducing necessary physics into EMRI waveform models will further increase computational costs [177]. Such improvements are direly needed: EMRI identification is not yet a solved problem [313, 314], but it is expected that individual signals may require billions of waveform evaluations to fully characterise [177]. In Chapter 3, we identify inefficiencies in current EMRI waveform generation and address them with ML techniques, vastly reducing the computational cost of these bottlenecks. We expect that these modifications will be a foundational component to future EMRI waveform generation frameworks, contributing to the continual effort to make EMRI data analysis a computationally tractable problem.

High waveform generation costs are not only problematic for the direct analysis of LISA data; they are also a hindrance to population studies. Accurately correcting for the presence of selection biases is vital if an unbiased and reliable fit of a GW source population is to be obtained [315]. This correction procedure requires that the selection effects be accurately modelled, which in turn entails a large number of waveform evaluations. If these waveforms are computationally expensive, as in the case of EMRIs, this process becomes impractical. In Chapter 4, we address this problem for EMRI selection bias modelling through the use of ML, laying necessary groundwork for future EMRI population studies to be both unbiased and computationally feasible. We demonstrate that the ML technique we develop is capable of accurately modelling selection biases for a representative EMRI population model, enabling unbiased esti-

mation of the shape of the population and the EMRI event rate.

1.7 Summary

Observing the Universe through GWs provides insights into astrophysical phenomena that are challenging to acquire by other means. Following the successful detections of GWs by ground-based interferometers and PTAs at frequencies $\gtrsim 10\,\mathrm{Hz}$ and $\leq 1\,\mu\mathrm{Hz}$ respectively, the LISA mission will complement these efforts by opening an observational window in the mHz band for the first time. The data LISA collects will contain a rich variety of GW signals, encoding the dynamics of their astrophysical sources. Accessing this trove of information will have deep and far-reaching implications across physics and astronomy, but this is contingent on the development of accurate and efficient waveform models and data analysis techniques.

The goal of this thesis is to contribute to specific areas of these ongoing development efforts. The unbiased extraction of GW signals from LISA data is demonstrated in Chapter 2 for high-fidelity simulations of the instrument, laying the foundations for the extensive validation of LISA data-analysis pipelines under realistic conditions. In Chapter 3, cutting-edge EMRI waveform models are augmented with ML methods, significantly improving their efficiency without sacrificing the physical fidelity of these models that is necessary for EMRI data analysis. Combining ML methods with hierarchical inference techniques, the selection effect modelling framework described in Chapter 4 makes unbiased population studies with LISA observations a computationally feasible prospect.

Chapter 2

Gravitational-wave data analysis for a realistic LISA instrument model

To characterise the performance of LISA and assess its performance as a GW detector, it is necessary to produce and analyse simulated data sets. The LISA instrument is complicated and multi-facted, consisting of many individual subsystems that must be accurately modelled to correctly simulate LISA data. These simulations span many orders of magnitude (with Gm arm-lengths varying by pm) and must retain their accuracy over a timescale of years. Producing high-fidelity simulations of LISA data is therefore a challenging and computationally expensive procedure.

To make progress, simplifying assumptions are usually made regarding the performance of the instrument. Laser and clock noises are assumed to be fully suppressed. The data is assumed to be gap-less and containing stationary Gaussian noise, with PSD described by the secondary noise allocation of the LISA requirements (Figure 1.2). The instrument response is treated with varying levels of rigor depending on the study in question; spacecraft orbits are typically assumed to be known exactly, and often are restricted to the (unphysical) scenario of constant and equal arm-lengths for ease of computation in the frequency domain. If the sky-localisation aspect of PE is not under particular investigation (for instance, in the work presented in Chapter 4), PE is performed directly with the GW strain polarisations and an effective instrument response model [77]. Collectively, these assumptions can dramatically reduce the computational cost of performing inference, making costly prospective studies of LISA measurement precision or detection efficiency feasible.

However, for the real LISA dataset, these challenges must be tackled head-on and overcome in order to perform GW science. We therefore require a testing environment in which the full ramifications of each of these effects can be properly explored, in order for large-scale LISA analysis pipelines to be made robust against any difficulties they will face. This is the motivation behind the development of realistic LISA simulation models. Many have been developed over time with increasing complexity and fidelity, e.g. [306, 307]. In this chapter, we will focus on

the model presented in [308], as this accurately represents the current LISA mission specification by ESA and is under active development.

Combining a simulation of LISA with an algorithm for performing GW data analysis, we obtain an end-to-end pipeline. The LISA pipeline is divided into four stages:

1. **L0-L0.5**. The generation and propagation of laser links between the orbiting spacecraft, including the effects of GW signals, is simulated in this stage. The outputs are the MHz beatnotes measured by each spacecraft, as well as telemetry ground-tracking information required for later processing stages.

2. **L0.5-L1**. In this stage, a noise reduction pipeline is applied to produce TDI combinations from laser beatnote measurements. These combinations are then synchronised from their local timing references to the TCB. Spacecraft orbits are reconstructed from ground-tracking data for use in PE.

3. **L1-L2**. GW signals in the data stream are characterised via PE to produce posterior samples, along with the instrument noise properties. For a full LISA data stream, this will be the result of a trans-dimensional fit that includes an exploration of the number of sources in the data.

4. **L2-L3**. The results of the trans-dimensional fit are processed to produce GW source catalogues of confident detections, for use in future analyses. These are the final data products of the LISA mission.

In this chapter, we make initial steps towards the complete validation of this pipeline, demonstrating that PE is consistent for a LISA simulation of limited complexity and duration. Full verification of pipeline performance (which involves a complete study of the impact of subsystem performance on the measurability of GWs) is a large-scale project that is beyond the scope of this thesis. However, the work presented here addresses many of the major concerns with the operation of LISA (such as the suppression of primary noise sources), and formed the basis of a report submitted to ESA which satisfied the criteria for the adoption of the LISA mission.

We are primarily interested in the validation of stages 1-3 (i.e., L0-L2) via the analysis of individual GW sources. If PE for individual sources is shown be be consistent, this validation can be extended to PE for datasets that contain many sources (i.e., the global fit described in Section 1.6.1) because the interaction between different sources in PE is not dependent on the noise properties of the LISA instrument, only the properties of the signals themselves. The form of the L2-L3 stage is predominantly influenced by the structure of this large-scale analysis, the complexity of which is beyond the scope of this work. Therefore, we do not investigate the process of forming L3 data products from PE outputs.

The layout of this chapter is as follows. In Section 2.1, I outline how the L0 LISA data is simulated and processed (including initial noise reduction operations) to produce L1 data products.

I then describe the general configuration of the L1-L2 analysis in Section 2.2, and demonstrate PE for three different GW source types in Section 2.3. Last, in Section 2.4 I validate PE with this LISA simulation pipeline for GB signals and short-duration simulations with analytic spacecraft orbits. My conclusions and future work are presented in Section 2.5.

2.1 L0-L1: Simulation and initial noise reduction

The process by which initial LISA data is obtained and converted to a form useful for data analysis is complicated. A high-level overview of the LISA measurement process is provided in Section 1.2.1; in this section, we detail which specific quantities must be simulated as part of the L0-L1 segment of the end-to-end pipeline, along with the tools employed at each stage. The work presented in this chapter represents the first steps in the process of validating the performance of a complete LISA simulation, which will be achieved by gradually introducing complexity and additional instrumental effects. Therefore, more challenging aspects of LISA data (discussed in Section 1.6.2), such as data gaps or non-stationary noise properties, are not yet included in the simulation. However, the L0-L1 segment of the end-to-end pipeline has been constructed such that these effects can be easily incorporated in the future, once they are the subjects of these investigations.

2.1.1 Spacecraft orbits

An essential ingredient to a simulation of LISA is a description of the orbital trajectory each spacecraft follows as a function of time. Here, we consider three levels of increasing complexity for the orbit of the constellation:

1. **Equal-armlength orbits**. These are unphysical Keplerian orbital trajectories in which the inter-spacecraft separation is held fixed at a fiducial length of 2.5 Gm. This simple analytic form of LISA orbits has existed in the literature for more than two decades and has seen extensive use in data analysis studies [316].

2. **Keplerian orbits**. Analytic spacecraft orbits are expanded to second order in the orbital eccentricity, yielding unequal spacecraft separations that change minimally with respect to time. While more realistic in terms of complexity, these are still unphysical solutions for the constellation orbits, in the sense that no initial configuration of the spacecraft will yield these orbital trajectories when their equations of motion are integrated numerically. These orbits have been implemented in the most recent LDCs [300].

3. **Generic orbits**. These orbits have no analytic form, and are the results of numerical simulations of orbital motion in the Solar System that account for the positions of all planetary bodies. This specification offers sufficient flexibility for the incorporation of

non-gravitational effects (such as drag) that are expected to affect spacecraft orbits in the real LISA mission. The investigations of this chapter use orbits generated by ESA, which account for gravitational effects and are representative of the trajectories the LISA constellation will be inserted into upon launch [317].

The analytic orbit scenarios enable TDI combinations to be expressed analytically in terms of the beatnote phase measurements, because light travel times (LTTs) (and therefore the required delays) are constant, enabling the LISA response function to be efficiently applied to waveform templates in PE. They are also simpler to implement in terms of simulation accuracy, as they are analytic and no interpolation is required. Numerical orbits are obtained by interpolating sparsely-sampled (10^{-5} Hz) numerical trajectories with a cubic spline. The LISAORBITS package is used for handling all three orbit scenarios [318].

2.1.2 Gravitational-wave response

In order to verify the consistency of GW PE with this end-to-end pipeline configuration, we must simulate how the waveform strain from distant GW sources will alter the propagation of the laser links and enter the beatnote phase measurements. Here, we use the LISAGWRESPONSE package to compute the fractional deviations to the frequency of each laser link, expressed in the time-frame of the receiving spacecraft [316, 319]. This includes the projection of the waveform strain onto the time-dependent antenna response of each laser link (due to the orientation of the link with respect to the wavefront), which incorporates the sky location of the source. This stage is fully generic to the source of the GWs; only an expression of the GW polarisations in the reference frame of the Solar–System barycentre (SSB) is required. Cubic spline interpolation is employed in order to express the uniformly-sampled (0.5 Hz) GW polarisations in the time-frames of each spacecraft.

2.1.3 Measurement of laser beatnote phases

Once spacecraft orbits and any GW sources have been fully specified, the operation of the LISA instrument can be simulated. This stage is complicated and a full exploration of what it entails is beyond the scope of this thesis. A (very high-level) summary is presented here; see [308] for a thorough description of the LISA instrument simulation. The software implementation of this stage of the end-to-end pipeline is provided by the LISAINSTRUMENT package [320], which implements the simulation and modelling techniques for LISA derived in [308].

Simulating the phases of each laser link directly is not feasible due to the 281.6 THz carrier frequency of the beam, as this would lead to floating-point rounding errors in the accumulated phase after $\sim 10^{-1}$ s at double precision, corrupting GW signals and laser noises alike. Instead, lasers are modelled in terms of frequency, described as the sum of a constant frequency (i.e.,

the carrier frequency), large offsets due to Doppler shifts and planned laser frequency variations (which are $\mathcal{O}(\text{MHz})$) and a time-dependent fluctuation that contains phase noises and GW phase information. This decomposition ensures that sufficient numerical precision is maintained throughout the simulation. In line with LISA requirements, we assume laser frequency noise has an ASD of $30\,\text{Hz}\,\text{Hz}^{-0.5}$ [53] (where we have deliberately left the units unfactorised to indicate that this is a spectral density).

All laser beams are modulated with local oscillators by $\sim$2.4 GHz to introduce the sidebands necessary for timing noise correction and spacecraft pseudo-ranging (discussed in Section 2.1.5); imperfect modulation is modelled as an additional noise term in this process, with an ASD of $5.2\times10^{-13}\,\text{s}\,\text{Hz}^{-0.5}$. In LISA, all onboard laser systems will be locked to a single reference laser, which we simulate here via the offset-fluctuation decomposition described above (see [308] for more information). Lasers are not locked to identical frequencies, as Doppler shifting due to relative spacecraft motion would shift the MHz beatnote frequencies out of the measurement band of the phasemeters, which is assumed to be the 5 MHz to 25 MHz band. As the scale of the Doppler shifts can be coarsely predicted ahead of time, the laser frequencies are adjusted as a function of time accordingly (referred to as frequency planning) to maintain near-constant beatnote frequencies throughout the LISA mission [308].

The effects of residual test-mass acceleration are incorporated as a noise term in the frequency of the local laser beam; the PSD of this noise in units of acceleration is [321]

$$S_{\text{TM}}(f) = 2.4^2\left(1+\left[\frac{4\times10^{-3}\,\text{Hz}}{f}\right]^2\right)\left(1+\left[\frac{f}{8\times10^{-3}\,\text{Hz}}\right]^4\right)\frac{(\text{fm}\,\text{s}^{-2})^2}{\text{Hz}}. \tag{2.1}$$

Laser beam propagation between spacecraft is then simulated, including the phase offset due to GWs. At this stage, tilt-to-length coupling (an increase in effective path length due to misalignment of the spacecraft with the beam propagation axis) will occur in the real LISA instrument, but this is not accounted for in current simulations. The combination of laser beams in the various interferometers onboard each spacecraft is then performed, simulating the process of measuring the beatnotes between local and received (Doppler-shifted) beams with on-board interferometers. Variations in beam optical path length (due to a combination of mechanical and thermal noise sources) are folded in as a readout noise term in these measurements, with a combined PSD in units of displacement of [321]

$$S_{\text{OMS}}(f) = 7.3^2\left(1+\left[\frac{2\times10^{-3}\,\text{Hz}}{f}\right]^4\right)\frac{(\text{pm})^2}{\text{Hz}}. \tag{2.2}$$

With carrier-sideband and mis-matched sideband-sideband beatnotes being $\mathcal{O}(\text{GHz})$, and therefore out of the measurement band of the phasemeters, the MHz carrier-carrier and two sideband-sideband (lower and upper) beatnotes are each measured by the phasemeters. This is easily simulated due to the decomposition of the laser frequencies into offsets and fluctuations, enabling

separate treatment of each beatnote measurement. Accurately simulating the measurement process itself is important because this is a major source of timing noise due to clock instability. However, timestamping must be performed at ns resolution over timescales of 10^8 s in order for these timing noise effects to be simulated accurately. To address this issue, a similar approach to that which was employed for laser frequencies is applied in the modelling of clock behaviour, which is split into long-timescale timing drifts and in-band timing fluctuations. The former represents offsets between the timing measurements of each spacecraft, while the latter is clock noise that will be a limiting source of noise if left uncorrected. These corrections are performed in post-processing (Section 2.1.5). The simulation assumes clock jitter noise has an ASD of $6.32 \times 10^{-14}\,\mathrm{Hz}^{-0.5}$, in line with the stability of current state-of-the-art oscillators [71].

One ingredient of this clock-noise correction are the measured pseudoranges (MPRs), which is produced via on-board processing on each spacecraft. A pseudo-random noise code is encoded in each laser link via modulation, which is then measured at the receiving spacecraft and compared to a local version of the code. By performing a cross-correlation between the measured and local codes, the difference in the time of emission (in the emitting spacecraft time-frame) and reception (in the measuring spacecraft time-frame) is computed, which corresponds to the distance between the two spacecraft. This is not a true measure of the distance due to the difference in time-frames, and is therefore referred to as the MPR. Due to the finite length of the pseudo-random codes, MPRs are ambiguous up to a multiple of 400 km; this ambiguity is resolved in L0.5-L1 with the inclusion of ground-tracking information. These ranging measurements are subject to a variety of instrumental noise sources; in the simulation this is modelled as an additive noise term to each ranging measurement with an ASD of $3.9 \times 10^{-9}\,\mathrm{s\,Hz}^{-0.5}$.

Last, the processing of the measured and timestamped beatnotes onboard the spacecraft (i.e., prior to telemetry) is simulated. In the LISA instrument, the MHz beatnote measurements will be sequentially low-pass filtered and downsampled to the final sampling rate of 4 Hz, which is low enough to comply with available data transmission rates. However, simulating this procedure in its entirety is not practical due to the large size of the data stream at MHz sampling rates; instead, beatnote measurements are simulated at 16 Hz and passed through a single filtering and downsampling stage (which is consistent with the final stage of the real LISA filter chain) to obtain data at 4 Hz. It is therefore assumed that the filter chain performs optimally and successfully limits in-band aliasing below requirements. In future, this filtering process will be modelled to more accurately reflect the effectiveness of these filters under different operating conditions [308].

2.1.4 Spacecraft ground-tracking

The L0-L0.5 stage of the end-to-end pipeline is completed by the simulation of spacecraft ground-tracking measurements, which will be performed on Earth. Two principal measurements are performed at this stage. First, spacecraft orbits are reconstructed from the information

provided by the two-way telemetry link, with spacecraft distances, tangential and transverse velocities relative to Earth being provided by time-of-arrival, Doppler-shift and pointing measurements of this link respectively [63]. These reconstructed orbits are required for applying the instrument response function to GW templates in PE. Second, time correlations between the local time-frame of the spacecraft and the TCB are measured (as the receiver time-frame can be easily converted to that of the TCB). These time correlations include both deviations due to timing noise of the spacecraft clock (the correction of which is the motivation for this measurement) and relativistic corrections due to the spacecraft orbit; the latter quantities are simulated using the true orbits of the constellation (and will be corrected for in post-processing using the reconstructed orbits). A combination of these measurements provides the information required to resolve ambiguities in the measurements of pseudo-ranges between spacecraft, which is necessary for TDI to be performed accurately [63].

Orbit reconstruction is currently assumed to contain no errors, an assumption that will be relaxed in future. Noise due to errors in the measured time correlations is simulated as white noise with an ASD of $0.042\,\mathrm{s\,Hz^{-0.5}}$. The received beatnote measurements, reconstructed orbits and time correlation measurements are the final data products of the L0-L0.5 simulation, and are now processed to suppress noise sufficiently for GWs to be extracted.

2.1.5 L0.5-L1: Initial noise reduction

There are multiple approaches by which TDI can be used to suppress laser and clock noises in the phasemeter measurements. Of particular importance is at what stage timing errors due to clock noise are addressed and all laser link measurements are synchronised to a common time frame, which can be performed either before or after the computation of TDI combinations.

The process for performing TDI with pre-synchronised measurements has existed in the literature in some form for two decades [60], and has seen significant refinements over time [59]. Clock noise only affects fluctuations relative to the total frequency, so this fluctuation information is recovered by removing the MHz phase ramps corresponding to the total beatnote frequencies. The detrended beatnote measurements are then synchronised to a common time-frame by combining MPRs and ground-tracking information, which also produces synchronised estimates of the LTTs that are necessary to perform TDI [63]. Clock noise is then corrected for after the construction of TDI variables in a dedicated post-processing stage [69, 322]. Performing TDI and clock noise correction in this manner introduces additional steps into the L0.5-L1 workflow (such as the removal of phase ramps) and requires the original measurements to be synchronised to within 40 ns, which leads to an increase in both the computational complexity and the information required to suppress noises below LISA requirements [71].

An alternative approach is to perform TDI with unsynchronised measurements, instead synchronising the resulting TDI combinations. This method (which is introduced and described in [71]) greatly simplifies the L0.5-L1 procedure:

- The algorithm works with total beatnote frequencies (rather than fluctuations), so no detrending is required;
- The MPRs are used directly in TDI, eliminating the need for any separation or synchronisation of LTTs from clock noises; and
- No dedicated clock noise reduction operation is required at any stage.

This total-frequency TDI method produces unsynchronised TDI combinations. However, the synchronisation timing requirements for the TDI combinations is now determined by the required phase accuracy for GW measurements, which is typically far less stringent than what is necessary for within-requirements suppression of laser noise with fractional-frequency TDI [71]. Given the simplifications enabled by performing this form of TDI, the L0.5-L1 stage of the end-to-end pipeline is structured with this method as its basis. The pipeline implementation, which is based on [71], is as follows:

1. Interpolate the (unsynchronised) LTTs estimated from spacecraft orbit information obtained via ground-tracking.
2. Obtain analytic models of clock drifts by fitting cubic splines to the time-correlation measurements obtained from ground-tracking.
3. Combining estimated LTTs and clock timing models, resolve ambiguity in the MPRs.
4. Delay all data by 4.531 25 s to account for the group delay of the low-pass filter applied at the end of L0-L0.5.
5. Combine intra-spacecraft beatnote measurements and processed MPRs to produce accurate pseudorange estimates.
6. Perform TDI, producing unsynchronised second-generation Michelson TDI combinations.
7. Synchronise TDI combinations using the analytic models for each clock and trim the edges of the resulting data to remove interpolation errors.

The need for step 4 (correction for filter group delay) was only identified by ensemble PE studies (Section 2.4), which identified systematic biases in the estimated phase offsets of injected signals due to an unaccounted-for delay in L0-L1. Some of these stages are currently not fully representative due to simplifications made in L0-L0.5 (such as the assumption of perfect ground-tracking of spacecraft orbits), but are included here for completeness. When these assumptions are relaxed in future work, these stages will be required to produce L1 data products suitable for GW data analysis.

The authors of [71] identify that floating-point rounding errors in total-frequency TDI, arising due to the disparate scales of mHz frequency fluctuations against the THz carrier frequency

of the laser, produce a noise floor that is below the LISA noise allocation. To investigate the impact of this effect on PE, we also implement fractional-frequency TDI, but do not extend this to include a full noise reduction pipeline as in the total-frequency case. By performing PE on simulated data with all instrumental noises disabled (and therefore eliminating the need for other noise-reduction stages) and both forms of TDI, any biases due to numerical precision issues can be investigated. We use the TDI framework of the PYTDI package [323] to implement both fractional- and total-frequency TDI operations.

2.2 L1-L2: Parameter estimation

In this section, we will describe the general structure of the L1-L2 stage of the end-to-end pipeline. This includes some additional pre-processing of the L1 data products, a description of the instrument noise model and how PE is performed. As PE is performed in terms of frequency fluctuations, the L1 TDI combinations $\{X_2, Y_2, Z_2\}$ are normalised by the nominal carrier frequency of the laser (281.6 THz) prior to the following analysis stages.

2.2.1 Detrending

The arm-lengths of the constellation vary on the orbital timescale, corresponding to frequencies $\mathcal{O}(10^{-4}\,\text{mHz})$. Despite being outwith the measurement band of LISA, the large amplitude of the spectral feature due to this effect produces significant spectral leakage in the frequency domain that limits LISA sensitivity and invalidates noise assumptions below $\sim$1 mHz. We combat this effect by detrending each TDI combination to suppress long-timescale variations:

1. The TDI combination is low-pass filtered by a Butterworth low-pass filter (implemented with the SCIPY.SIGNAL module) of order 20 and a critical (turnover) frequency of 10^{-1} Hz.

2. A smoothing cubic spline is fit to the filtered TDI combination, modelling orbital-timescale variations.

3. The spline fit is subtracted from the unfiltered TDI combination, which is passed along for later analysis.

An example of this low-frequency leakage correction with detrending is shown in Figure 2.1 for 3 d and 1 yr simulations (panels (a) and (b) respectively) and the Keplerian orbit configuration. After detrending long simulations with unequal-armlength orbits, we observe some residual structure at lower frequencies (panel (c) of Figure 2.1). The first of these residual peaks (and the uniform spacing between peaks) are of identical frequency to the sampling frequency of the orbital trajectories, which suggests that they may be the result of orbit interpolation errors propagating through the L0-L1 pipeline. These residuals are negligible above $\sim 2 \times 10^{-4}$ Hz and so will be of minimal impact to the PE investigations performed in this chapter. However,

it is not clear how the structure of these residuals will be affected by changes in orbit sampling rate or simulation duration; this investigation is left to future work.

2.2.2 Windowing

In addition to low-frequency power due to orbital variations, the L1 TDI combinations also contain significant power at frequencies $\mathcal{O}(1\,\mathrm{Hz})$ as shown in Figure 2.2. This is due to incomplete suppression of laser noise at these frequencies as a result of interpolation errors in the delayed beatnote measurements, and leads to severe spectral leakage in the mHz band. We combat this leakage by windowing the TDI combinations. The window function used, which is commonly applied in the LDC analyses, is of the form

$$w(t) = \frac{1}{4}\left\{1+\tanh[\kappa(t-t_l)]\right\}\left\{1-\tanh[\kappa(t-t_r)]\right\}, \tag{2.3}$$

where t_l and t_r describe the window start/end positions in seconds, and κ controls the slope of the window taper. Through manual tuning, we find that by setting $\kappa = 10^{-3}\,\mathrm{s}^{-1}$ and $\{x_l, x_r\} = \{5\times10^3\,\mathrm{s}, T-5\times10^3\,\mathrm{s}\}$ leakage effects are suppressed sufficiently for PE. The corresponding suppression of leakage effects with this window function is shown in Figure 2.2.

2.2.3 TDI variable convention and noise properties

Before PE can be performed, a model for the PSD of the instrument noise must be specified. As described in Section 1.4.3, the covariance matrix is diagonal with respect to frequency under the assumption of stationary noise. We therefore transform the TDI channels $\{X_2, Y_2, Z_2\}$ to the frequency domain via fast Fourier transform (FFT). However, the noises in $\{X_2, Y_2, Z_2\}$ are still correlated with respect to one another, i.e. the three-channel covariance matrix of each frequency bin contains off-diagonal terms. If the properties of the instrumental noise processes in each of the three spacecraft are assumed to be identical and stationary, the covariance matrix is defined by two terms (a diagonal and off-diagonal component). For this specific case, the covariance matrix is diagonalised by a change of variables to the alternative TDI combinations [324]

$$A_2 = \frac{Z_2 - X_2}{\sqrt{2}}, \qquad E_2 = \frac{X_2 - 2Y_2 + Z_2}{\sqrt{6}}, \qquad T_2 = \frac{X_2 + Y_2 + Z_2}{\sqrt{3}}. \tag{2.4}$$

The variance of the noise in the A and E TDI channels is identical, provided that the above assumption of identical spacecraft noise behavior holds (otherwise, off-diagonal terms re-enter the covariance matrix; the AET parameterisation may still be used here, but is no more convenient than e.g. XYZ). We model the noise behaviour of the TDI channels by assuming that it is dominated by the secondary noises S_{TM} and S_{OMS} (given by Equations (2.1) and (2.2) respec-

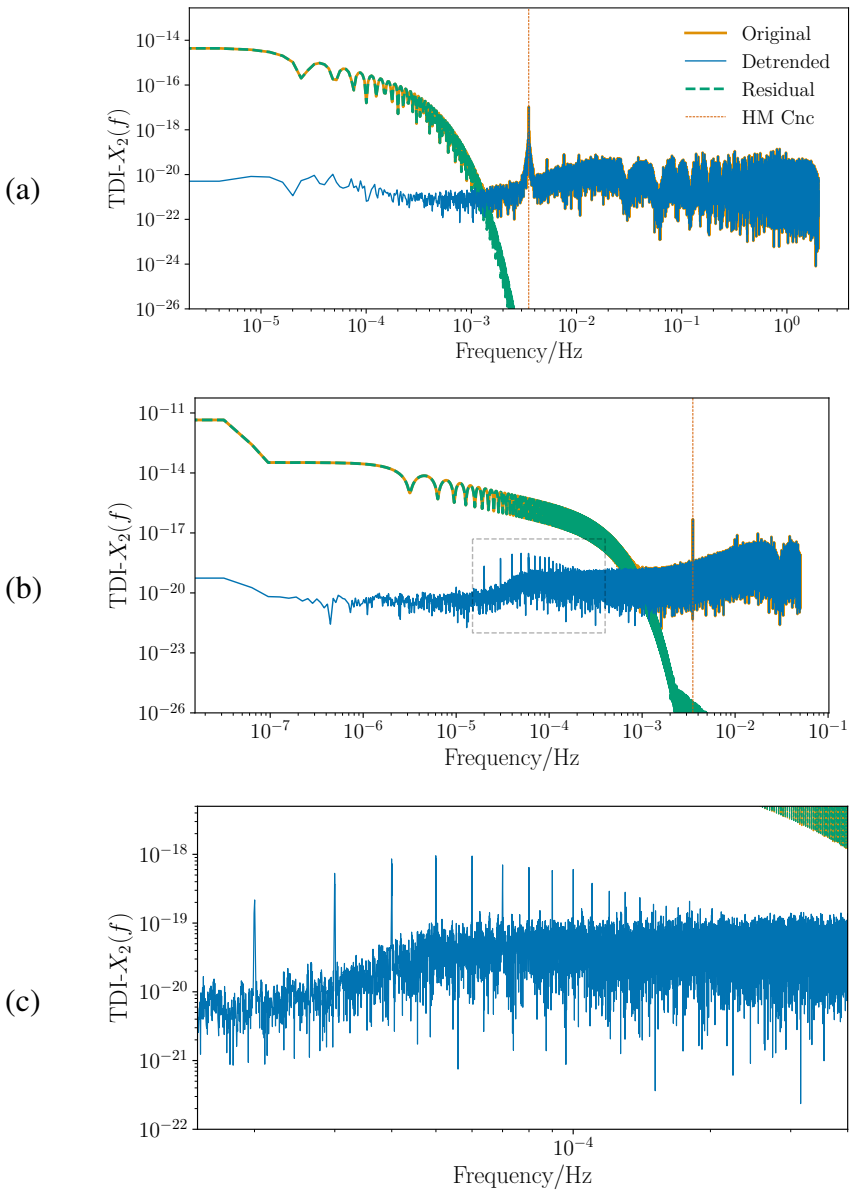


Figure 2.1: Removal of low-frequency spectral leakage with detrending for (a) 3 d- and (b) 1 yr-duration LISA simulations with the Keplerian constellation orbit configuration and all instrumental noises disabled. The L0-L1 output data (orange) is contaminated by spectral leakage that originates from a large spectral feature due to armlength variations on the orbital timescale of the constellation. Fitting and subtracting this low-frequency feature in the time-domain (detrending) significantly suppresses this leakage effect in the resulting data spectra (blue). The residual that was subtracted (green) contains the leakage structure down to a numerical noise floor at $\sim 2 \times 10^{-3}$ Hz, and contains no strong residuals from the loud reference signal HM Cnc (red dashed line, a GB described in Section 2.3.1). In the 1 yr simulation, residual structure associated with the sampling rate of the numerical orbit files (10^{-5} Hz) is revealed by this detrending (indicated by the dashed region). Closer inspection of this region of the spectra (panel (c)) shows that these residuals drop below the numerical precision noise of the simulation at $\sim 2 \times 10^{-4}$ Hz. For all spectra, the window defined in Section 2.2.2 as been applied to better illustrate the leakage and its suppression; in the actual analysis pipeline, the detrending is applied prior to any windowing.

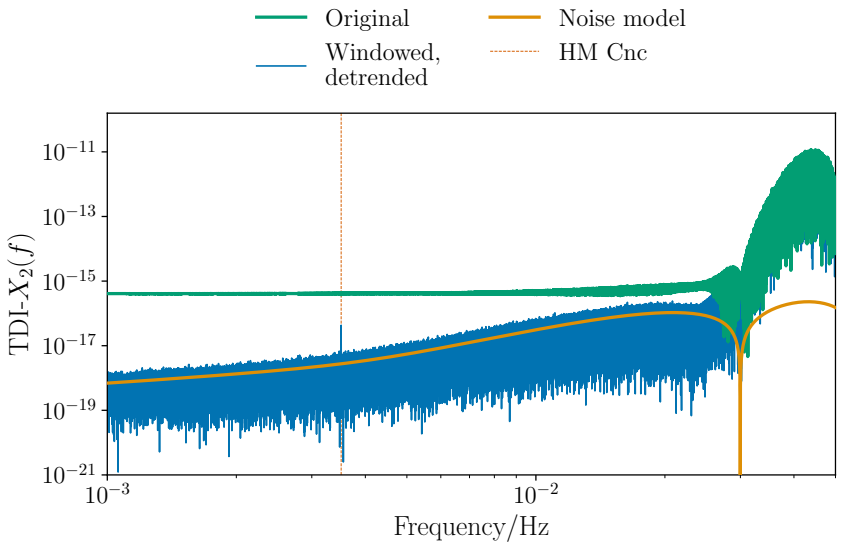


Figure 2.2: Suppression of broadband spectral leakage in the TDI-X_2 data stream with windowing. The raw L0-L1 TDI data (green line) contain large high-frequency spectral features (due to interpolation errors in TDI delay operators) that produce a broadband leakage signal orders of magnitude brighter than GW signals or secondary LISA noise sources. Applying the window function defined in Equation (2.3), the corrected data stream (blue line) contains both the injected GW signal (red dashed line, a GB described in Section 2.3.1) and the expected LISA noise floor below 20 mHz (orange line).

tively) [321]; the corresponding PSD in A and E is

$$S_{\mathrm{AA}}(\omega) = 2C_{\mathrm{XX}}(\omega)\left([2+\cos(\omega_L)]\,S_{\mathrm{OMS}}(\omega) + 2\,[3+2\cos(\omega_L)+\cos(2\omega_L)]\,S_{\mathrm{TM}}(\omega)\right), \quad (2.5)$$

where $\omega = 2\pi f$, $\omega_L = \omega L/c$ and

$$C_{\mathrm{XX}}(\omega) = 16\sin^2(\omega_L)\sin^2(2\omega_L) \quad (2.6)$$

is a common factor of the transfer functions of both secondary noise terms [325]. The PSD of the T channel is of different form; it is written as [325]

$$S_{\mathrm{TT}}(\omega) = 4C_{\mathrm{XX}}(\omega)\left([1-\cos(\omega_L)]\,S_{\mathrm{OMS}}(\omega) + 8\sin^4\left(\frac{\omega_L}{2}\right)S_{\mathrm{TM}}(\omega)\right). \quad (2.7)$$

While $S_{\mathrm{TT}}(\omega) \ll S_{\mathrm{AA}}(\omega)$ in the mHz band, the sensitivity of the T channel to GWs is also much lower than that of the A_2 and E_2 channels in the same frequency range. The change of TDI variables to $\{A_2, E_2, T_2\}$ does not merely alter the noise properties of the TDI channels: it also diagonalises the instrument response below the characteristic frequency $f_* = (2\pi L/c)^{-1} \sim 19.1\,\mathrm{mHz}$, such that $\{A, E, T\} \sim \{h_+, h_\times, h_\odot\}$ [116], where $h_\odot$ is the TT scalar polarisation mode. As $h_\odot = 0$ in GR, the T channel only strongly contributes to the total SNR of a GW signal above f_*, at which point S_{TT} is of similar size to S_{AA}.

With $\{A_2, E_2, T_2\}$ and the noise PSDs of Equations (2.5) and (2.7) computed, we perform PE, sampling the posterior distribution defined by Equations (1.5) and (1.17) with the NESSAI nested sampler [326]. We use 2000 live points, a remaining log-evidence stopping threshold of 10^{-2} and all other settings set to their default values. This process requires both the prior distribution and the signal template model to be specified, both of which depend on the type of GW source being analysed. The LISA simulation is also currently subject to technical limitations (due to available computational resources) which are varied based on the scope of the data analysis task in question. These source-specific assumptions and restrictions are described in the following sections. Section 2.3 describes the PE workflow for three different source types observed for 1 yr, and Section 2.4 focuses on the validation of an ensemble of PE results for GBs observed in 3 d simulations.

2.2.4 A note on reference epochs

Any GW source that evolves with respect to time must necessarily be described by parameters that are defined with respect to a specific point in time, known as a reference epoch. For instance, the quadrupolar emission from an inspiralling compact binary system is characterised by an initial frequency f_0 specified at the reference epoch (along with frequency derivatives, also specified at this epoch up to arbitrary order). When performing PE with simulated data, in which one typically compares the results obtained with the true parameters of the injected signal, it is essential that the reference epoch of the injection and template model are equivalent. This is especially important for quantities which vary quickly with time, such as GW phase: even a slight change to the reference epoch may be sufficient for a measured phase offset to be completely inconsistent with that of the injected signal.

A logical choice of reference epoch is the first time-stamp of the instrument data, which we refer to as t_0. However, if t_0 changes after the GW signal is simulated and injected into the data stream, there will be an epoch mismatch. We found this issue to be particularly pertinent for PE with the end-to-end pipeline due to the presence of both delays and edge-trimming operations in the L0-L1 simulation and noise reduction segments described in Section 2.1. To avoid this problem, we define an epoch t_{init} with respect to which all GW-related parameters are defined. We choose t_{init} to be equal to t_0 at the beginning of the pipeline; because these quantities are adjusted independently, no epoch mismatches will occur even if the simulation time-stamps are manipulated or samples are removed. In other words, we define the GW epoch separately to the grid of measurement time-stamps, ensuring that both injection and recovery waveform models share a consistent reference point during PE.

2.3 LISA data analysis for different source classes

The aim of the work presented in this chapter is to examine the consistency of PE in our end-to-end LISA simulation and analysis pipeline. A first-cut approach towards this goal is simply to perform PE and examine whether the injected parameter values are consistent with the recovered posterior distribution. This process is useful for identifying significant errors in either simulation (e.g., incorrectly injected signals) or data analaysis (e.g., mis-specified template or noise models). However, PE consistency can only be established to a coarse degree with a single posterior distribution, given that provided the injection lies somewhere in the posterior bulk the results will appear reasonable. The confidence of this statement can be improved by performing a statistical analysis of an ensemble of parameter estimation run results, but this is computationally expensive. We also perform such an analysis for GB signals, albeit for simulations over a shorter duration due to memory constraints, in Section 2.4.

For the simulations produced in this section, we assume a LISA mission duration of 1 yr and a sampling rate of 10^{-1} Hz. These parameters are enforced by the random-access memory (RAM) of the computing resources available; a simulation of this duration and cadence requires $\mathcal{O}(50\,\mathrm{GB})$ in memory. Efforts are ongoing to reduce the memory footprint of LISAINSTRUMENT by an order of magnitude, but they are not yet complete as of the writing of this thesis. We do not expect the reduced sampling rate or mission duration to greatly affect the conclusions drawn from the results presented in this section, because the sources considered are sufficiently low in frequency to be unaffected by this limitation. However, other GW sources targeted by LISA (particularly stellar-mass BH binaries) do occupy this higher frequency band; it is therefore important that memory limitations are overcome in future work to properly investigate PE for such sources in the end-to-end pipeline framework. It is also a defining characteristic of this effort that the LISA simulation be as close to reality as possible in order to identify and quantify unknown instrumental effects empirically, so it is important that any deviations from the instrument specification are corrected in order to facilitate this.

For each source type considered, we characterise the recovery of source parameters from the simulated data stream by considering two detector scenarios. In the first, we disable all instrumental noise sources in order to examine whether the efficiently-generated templates we use in PE are a good match to the signal injected as part of the L0-L0.5 simulation. The posterior distribution recovered for PE in the absence of noise is equivalent to the expectation of the posterior distribution recovered in the presence of noise: the absence of noise is equivalent to the noise taking its expected value of zero for every data point. For noise-free analyses, we assume uniform prior distributions in all parameters such that the recovered posterior is directly proportional to the likelihood. By removing the majority of the statistical bias in the posterior distribution (some still remains due to residual numerical noise in the simulation output), we can therefore identify the significance of systematic biases attributable to the template model used. This is an important point to examine, because if the template does not match to such

a degree that measurable biases (i.e., larger than the width of the posterior bulk) are present, the sensitivity with which we can test the validity of our LISA noise assumptions is greatly diminished. We repeat this process for each of our orbital scenarios described in Section 2.1.1 to establish the validity of the template instrument response function in each case.

Once any issues identified in this first scenario have been rectified (if possible), we then enable the full suite of noise sources modelled by the L0-L0.5 simulation. We perform PE to obtain samples of the posterior distribution which (if our assumptions regarding the suppression of LISA noise are correct) will overlap with the parameter-space location of the injected parameters. If the injected parameter values are found to be consistent, then we can confidently rule out the presence of any large systematic biases in the end-to-end pipeline and state that our LISA PE is optimal. This conclusion holds only if the properties of the simulation do not significantly change for different noise realisations, which we found to be the case in our development of the pipeline. For simplicity, we do not estimate the noise level in any of these analyses, as this is not formally required if our objective is to validate PE for signal parameters. In future work, properties of the LISA instrument will be varied to understand the impact of these changes on parameter estimation, which will entail noise modelling to account for changes to the PSD; inclusion of noise modelling in these analyses is therefore left to this future work.

We examine three classes of GW sources with this analytic framework: GBs, MBHBs and EMRIs. The following subsections describe these analyses in turn.

2.3.1 Galactic binaries

We first consider PE for individual GB sources, which are the simplest GW sources LISA is expected to observe. The waveforms of GB systems are quasi-monochromatic, predominantly being composed of the quadrupole mode evolving slowly in frequency. For a system of chirp mass $\mathcal{M}_c$ emitting at frequency f_0 and evolving purely under GW emission, the source-frame frequency derivative to first post-Newtonian (PN) order is [327]

$$\dot{f} \approx 5.8 \times 10^{-18} \left(\frac{\mathcal{M}_c}{\mathrm{M}_\odot}\right)^{5/3} \left(\frac{f_0}{\mathrm{mHz}}\right)^{11/3} \mathrm{Hz\,s^{-1}}. \tag{2.8}$$

which accumulates to less than the width of a frequency bin ($\sim 10\,\mathrm{nHz}$) for typical GB systems over a duration of 4 yr. The bandwidth of a GB source in the detector frame is wider than this due to Doppler modulation due to the motion of the constellation, which varies the observed frequency of the GB signals by

$$\Delta f \approx 2 f_0 \left(\frac{v_{\mathrm{orb}}}{c}\right) \cos\beta, \tag{2.9}$$

where v_{orb} is the orbital velocity of the constellation, β is the ecliptic latitude of the source and we have neglected the evolution of the GW frequency with respect to time. For $v_{\mathrm{orb}} \sim 10^{-4} c$

(treating it to be approximately equal to the orbital velocity of the Earth), the observed frequency of GB signals varies by less than $\mathcal{O}(\mu\text{Hz})$ over the course of a year. These signals are therefore highly compact in the frequency domain ($\mathcal{O}(100)$ frequency bins) compared to the time domain ($\mathcal{O}(10^7)$ samples), making frequency-domain generation of GB waveforms orders of magnitude more efficient.

This feature of GB waveforms motivated the development of the FastGB waveform model, which generates both the waveform strain and the LISA response function directly in the frequency domain [291]. The waveform is not analytic in the frequency domain, but is obtained by performing a fast-slow decomposition (or equivalently, heterodyning the signal with respect to f_0). The signal is described as the product of a sinusoid oscillating on the GW timescale and a slowly-varying component that changes on the orbital timescale of the constellation. The slowly-varying part is represented alias-free in the time domain with low sampling rates $\mathcal{O}(10^{-5}\,\text{Hz})$ and is therefore computationally inexpensive to manipulate. The GB waveform is obtained by Fourier transforming the slowly-varying part and reversing the heterodyne in the frequency domain. In FastGB, this is implemented in terms of a sum over Fourier coefficients (in the form of a FFT) that is truncated after N_{GB} terms. See [291, 328] for explicit derivations of these operations and discussion of their validity. The tuning parameter N_{GB} is set as low as possible to maximise computational efficiency without loss of accuracy; values of $N_{\text{GB}} \in [128, 512]$ are commonly used for 1 yr data segments [92, 328]. The FastGB model has seen widespread application in LISA data analysis for both the analysis of individual GB signals [92, 119, 329] and global analyses of GB populations [259, 295, 298, 328]. We therefore use FastGB as our GB template model throughout this chapter.

We improve upon existing implementations of FastGB by interfacing the orbit computation with LISAOrbits, extending the model to support generic numerical orbit configurations, and optimise the model code by rewriting it to use vectorisation operations where possible for increased efficiency when evaluating multiple waveform templates at once. As the data has been windowed, we convolve each template with the frequency-domain representation of Equation (2.3) to ensure our waveform model is consistent with the data. Currently, FastGB is only capable of incorporating the effects of unequal and time-varying armlengths (as are present in realistic orbits) in the slow part of the decomposition; the TDI response is still applied to the fast part assuming equal arm-lengths. Dedicated PE studies investigating the validity of this approximation identify that it leads to significant biases for SNRs above $\mathcal{O}(500)$ [330], which will only be problematic for the loudest GB sources [331].

To assess the impact of the accuracy of FastGB on parameter estimation, we generate the injected GB signals in the time domain rather than inverse Fourier transforming the FastGB output. We set the parameters of the injected GB source to those of HM Cnc, a loud (SNR ~ 100 over 4 yr) AM CVn source in the verification GB catalogue [117, 332]. These parameters are given in Table 2.1: inclination, polarisation and initial phase are not well constrained with

Parameter	Symbol	Reference value
Strain amplitude	A	6.3×10^{-23}*
Frequency	f_0	$3.5 \times 10^{-3}\,\mathrm{Hz}$
Frequency derivative	$\dot{f}$	$5.6 \times 10^{-17}\,\mathrm{Hz\,s^{-1}}$
Initial phase	ϕ_0	$\pi/3$
Inclination	ι	$\pi/3$
Polarisation angle	ψ	$\pi/4$
Ecliptic latitude	β	0.82
Ecliptic longitude	λ	5.15

Table 2.1: Source parameters for our chosen reference GB system, HM Cnc. Intrinsic parameters and sky locations are taken from [117]. Inclination, phase and polarisation are poorly constrained with existing observations, so we choose representative values for these quantities. All angular quantities are expressed in radians. *The source has SNR ~ 100 when observed over a 4 yr duration and nominal LISA sensitivity; for our 1 yr simulations, we double A such that the SNR is unchanged.

EM observations and are assigned values representative of typical GB systems. The SNR of the signal is maintained for our shorter 1 yr simulations by doubling the strain amplitude of the source (as SNR grows approximately with the square root of the observing time for slowly-evolving sources).

The prior distributions assumed for the GB analyses presented in this section are shown in Table 2.1. To improve PE convergence times, the prior on f_0 is narrowly centred around that of the source. The range chosen is still a factor of $\mathcal{O}(10^3)$ wider than the expected marginal posterior in f_0 for HM Cnc, and is conservative when compared to the size of the frequency bands over which GBs are analysed as part of global fit studies, which are $\mathcal{O}(5\,\mu\mathrm{Hz})$ [328]. We therefore do not expect this choice to affect any results obtained from PE analyses.

The lower bound of the prior on A is higher than what would correspond to a typical GB SNR-based detection threshold. However, as our goal is to sample the posterior distribution (which will have almost zero support at these low amplitudes), this lower bound may be set arbitrarily to reduce PE complexity provided this does not truncate the posterior bulk.

The LISA PSD does not change significantly over a 20 µHz frequency band at mHz frequencies. We therefore treat the noise variance in each channel as constant (equal to $S_{\mathrm{CC}}(2\pi f_0)$ for TDI channel $C \in \{A, E, T\}$) over all frequency bins for GB analyses, an approximation found to be reasonable in previous analyses that further reduces computational costs [295, 328].

We first prospect for systematic biases due to any mismatch between FastGB and the injected GB signal for each spacecraft orbit configuration described in Section 2.1.1. For each GB parameter, we compute the difference between the median of the marginal posterior samples and the injected value and normalise this deviation by the standard deviation of the marginal posterior samples. We additionally compute the total normalised deviation by summing the deviations in quadrature to compute the Euclidean distance between the median of the normalised

Parameter	Prior distribution	Lower bound	Upper bound
A	Uniform	10^{-23}	10^{-21}
f_0	Uniform	$f_0 - 10^{-5}\,\mathrm{Hz}$	$f_0 + 10^{-5}\,\mathrm{Hz}$
$\dot{f}$	Uniform	$-10^{-15}\,\mathrm{Hz\,s^{-1}}$	$10^{-15}\,\mathrm{Hz\,s^{-1}}$
β	$\cos\beta$	$-\pi/2$	$\pi/2$
λ	Uniform	0	2π
ι	$\sin\iota$	0	π
ψ	Uniform	0	π
ϕ_0	Uniform	0	2π

Table 2.2: Prior distributions and bounds used for 1 yr GB PE analyses. All angular quantities are expressed in radians; see Table 2.1 for symbol definitions.

posterior and the injected parameters. Previous results in the literature suggest that we only expect systematic biases to be observable from isolated analyses at SNRs ~ 500 [330]. To determine whether we will observe systematic biases at large (yet astrophyically plausible) SNRs, we repeat these three analyses for an injected GB signal with amplitude rescaled by a factor of 10 such that the overall SNR is 1000. The six resulting sets normalised deviations are shown in Figure 2.3 for all orbit configurations. The SNR-100 analyses do not show conclusive signs of systematic bias. However, there is significant scatter of similar order to the posterior width, even for the equal-armlength case (where we expect systematic biases to be extremely small); this is due to the presence of numerical precision noise in the simulation, which persists despite all instrumental noise being disabled. This residual noise is a feature of total-frequency TDI at double precision, and was identified when this method was first derived [71]. At higher SNRs, bias is observed for both the Keplerian and generic orbit configurations. The bias is more severe in the generic orbit scenario, totalling more than three posterior widths, whereas the Keplerian orbit configuration only has a bias of $\sim$ one posterior width. This agrees with our expectations: as the equal-armlength TDI response (which is applied in FastGB) is a better approximation for constant arm-length orbits [330], we expect any resulting biases to be less significant in this case. These results are fully consistent with those obtained in [330] and suggest that our pipeline faithfully represents the simulation and analysis of LISA data containing GB signals.

As any systematic biases present are minimal for the actual SNR of HM Cnc (100), we enable all instrumental noise sources and subsequent correction steps in L0-L1 and repeat PE for all three orbit configurations. The recovered posteriors are shown in Figure 2.4. All three posterior distributions are consistent with the injected parameters, allowing us to rule out the presence of any large systematic biases that would produce deviations larger than the scale of the posterior bulk. While the equal-armlength marginal posterior PDF on f_0 is only marginally consistent with the injected parameter, this is expected; for three PE analyses with a seven-parameter model, we should expect at least one marginal posterior PDF to deviate to such a degree. The TDI-X_2 ASD for the generic orbit scenario (shown inset in Figure 2.4) agrees closely with both the injected

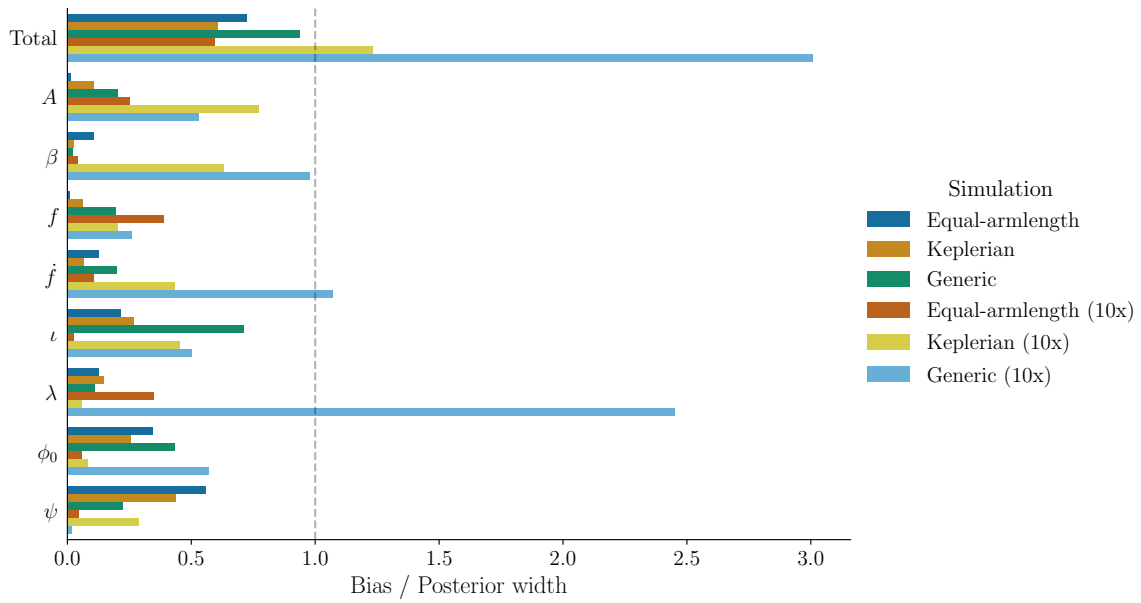


Figure 2.3: Biases in inferred GB parameters for PE on six LISA simulations of duration 1 yr and with no instrumental noises enabled. We investigate systematic biases for both our reference signal HM Cnc (parameters in Table 2.1) and another signal of identical parameters, except for a ten-fold increase in strain amplitude which raises the SNR to 10^3. Biases are obtained by computing the difference between the median marginal posterior sample and the injected value of each parameter and normalising it by the standard deviation of the marginal posterior samples. These quantities are summed in quadrature to calculate the total bias. We confirm that FastGB is not sufficiently accurate (greater than one posterior width of bias) for unequal and/or time-varing armlength orbits at SNR 10^3, in line with the results of previous work [330]. Biases of up to ~ 0.5 posterior widths are observed for some parameters, even for equal-armlengths at SNR 10^2. These may be the result of numerical precision noise in the simulated data or statistical errors from stochastic sampling.

and recovered signal (both generated with FASTGB for this visualisation). These results indicate that our expectations regarding the LISA noise floor post-TDI are consistent with those realised by an accurate simulation of the LISA instrument.

2.3.2 Massive-black-hole binaries

While GBs are a useful source for verifying some aspects of LISA PE due to the low computational cost incurred in their analysis, their waveforms only occupy a narrow frequency band. This means that their analysis does not require knowledge of broad-band noise properties of the detector (other than the assumption of stationary noise). It is therefore prudent to also consider PE for broad-band GW sources; one such source class is MBHBs. As a MBHB system inspirals, the resulting waveform sweeps through the LISA band on timescales of days to months (depending on the mass of the system). The majority of the SNR of a MBHB accumulates in the final stages of inspiral, merger and ringdown (as indicated by the sharp increase in SNR for MBHB systems in Figure 1.3), which occurs over a shorter timescale of hours to days (again, depending on the properties of the system). MBHB analyses therefore serve as a useful intermediate test of LISA signal recovery in our end-to-end pipeline, being of wider frequency band than GBs but only over a relatively short timescale compared to the mission duration.

In this section, we perform PE on a representative MBHB source with parameters given in Table 2.3. Unlike the GB case, this is not a real GW source as we do not have accurate EM observations of near-coalescence MBHB systems. Instead, we choose astrophysically-motivated values for the source parameters: masses are of order $10^6\,\mathrm{M}_\odot$, spin magnitudes are of moderate size and the luminosity distance to the source is representative of MBH binary sources LISA is expected to observe [150, 216]. The SNR of the source is $\mathcal{O}(1000)$, which is not exceptionally high for MBH signals [216]. This high SNR makes MBH analyses particularly sensitive to sources of systematic bias, including waveform mismatch biases and inaccurate modelling of the instrument noise.

To perform PE on MBHB signals we must specify an efficient waveform template model for use in inference. It is well understood that significant efficiency improvements are attained with frequency-domain BBH waveform generation [333, 334], especially when combined with interpolation techniques [150]. However, including the LISA response function in these frequency-domain waveform models is challenging: unlike GBs, one cannot consider MBHBs as narrow-band systems for which a heterodyne approximation will hold with any reasonable accuracy. Frequency-domain LISA response functions for MBHB waveforms have been derived in recent work. However, software implementations of the techniques developed for this purpose (such as the LISABETA [216, 310] and BBHX [150, 194] packages) currently only support analytic spacecraft orbits with constant arm-lengths, and have yet to be expressed as part of a framework that is compatible with LISAORBITS, a vital component to the LISA response function.

We must therefore generate our waveform templates in the time domain. While accurate,

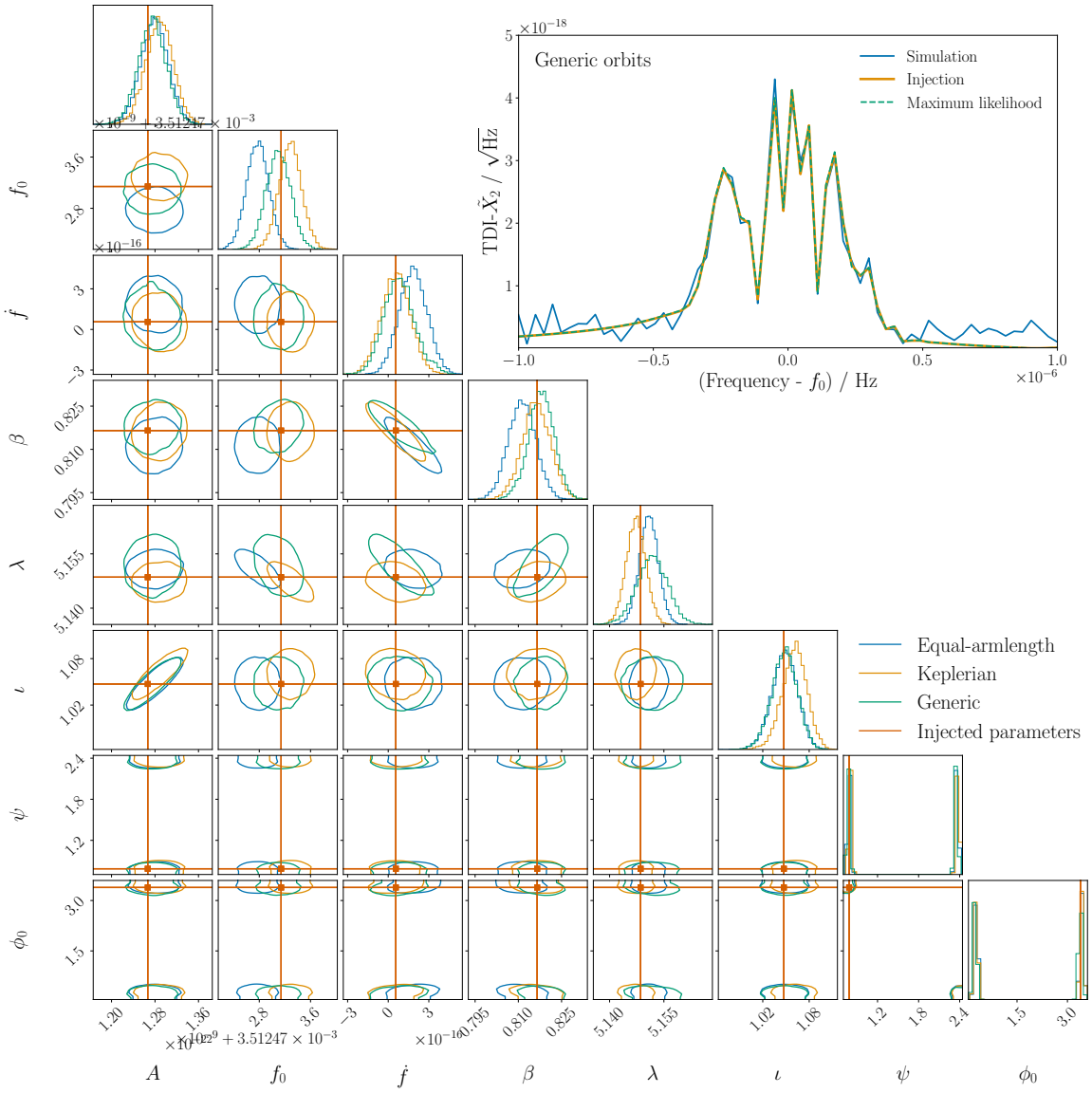


Figure 2.4: Posterior distributions obtained via PE for GB signals injected in 1 yr simulations of LISA and different orbital configurations. The injected signal is HM Cnc (parameters indicated by red line, see Table 2.1 for definitions) with amplitude rescaled such that it is recovered with an SNR of 100. For each of the equal-armlength (blue line), Keplerian (orange line) and generic (green line) orbital configurations, we obtain posteriors that are consistent with the injected values for all parameters, indicating that no significant systematic biases (i.e., much larger than the statistical biases) are present. Contours contain 90% of the total posterior probability. Inset: the TDI-X_2 ASD for the generic orbit configuration (blue line), centred on the initial frequency of HM Cnc. The injected signal (orange line) and maximum likelihood template (green dashed line, obtained from the posterior PDF) agree closely with the simulated data.

Parameter	Symbol	Reference value
Total mass	M	$1.5 \times 10^6\,\mathrm{M}_\odot$
Mass ratio	q	0.5
Primary spin	a_1	0.5
Secondary spin	a_2	0.3
Inclination	ι	$\pi/3$
Polarisation angle	ψ	$\pi/4$
Ecliptic latitude	β	$\pi/6$
Ecliptic longitude	λ	$3\pi/2$
Luminosity distance	d_L	30 Gpc
Time of coalescence	t_c	40 d
Phase at coalescence	ϕ_c	$\pi/5$

Table 2.3: Source parameters for our chosen reference MBHB system. As no such systems near coalescence are known, we construct a system with astrophysically representative parameters with SNR ~ 1000. The spin of each MBH is expressed in dimensionless form (i.e., normalised with respect to its maximal value). Phase at coalescence is defined with respect to the $(\ell, m) = (2,2)$ mode of the waveform, whereas time of coalescence is defined with respect to the epoch t_init. All angular quantities are expressed in radians.

the response implemented by LISAGWRESPONSE (designed for use with central processing units (CPUs)) costs $\gg 1$ s per waveform evaluation, which is prohibitively expensive for use in PE. Instead, we use the recently-developed FASTLISARESPONSE package, a GPU-accelerated time-domain LISA response code that implements fractional-frequency TDI for arbitrary specifications of waveforms and spacecraft orbits [330]. Pairing this response code with the IMRPHENOMXPHM BBH waveform model [335], implemented via PYCBC [263, 336], we obtain an accurate MBHB waveform model that costs $\mathcal{O}(100\,\mathrm{ms})$ per evaluation (using an Nvidia V100 GPU), which is sufficiently inexpensive for use in data analysis. This computational cost is achieved for a lower frequency cut-off of 10^{-3} Hz, which is not fully representative as these systems will nominally be examined from 10^{-4} Hz in the real LISA data stream [53]. However, the length of the waveform in the time domain increases significantly for a frequency cut-off of 10^{-4} Hz, which makes the inverse-FFT operation within PYCBC that represents the IMRPHENOMXPHM in the time domain prohibitively expensive (taking $\mathcal{O}(1\,\mathrm{s})$). This limitation will be addressed in the future when the aforementioned fast MBHB waveform models have been extended to support generic orbits.

As we are now considering a broad-band GW source, we cannot reasonably approximate the noise level as constant as we did in the GB case. We instead express the PSD as a vector over the frequency bins in the 10^{-3} Hz to 10^{-2} Hz frequency band, with values defined according to Equations (2.5) and (2.7). To complete our definition of the MBHB model, we define priors on the MBHB parameters: these are given in Table 2.4. The prior bounds are centered on the parameters of the injected source for efficiency. This is necessary for rapid convergence of our

Parameter	Prior distribution	Lower bound	Upper bound
M	Uniform	$-10^4\,\mathrm{M}_\odot$	$10^4\,\mathrm{M}_\odot$
q	Uniform	-0.05	0.05
a_1	Uniform	-0.05	0.05
a_2	Uniform	-0.1	0.1
ι	$\sin\iota$	-0.1	0.1
ψ	Uniform	-0.2	0.2
β	$\cos\beta$	-0.2	0.2
λ	Uniform	-0.2	0.2
d_L	d_L^{-3}	$-5\,\mathrm{Gpc}$	$5\,\mathrm{Gpc}$
t_c	Uniform	$-20\,\mathrm{s}$	$20\,\mathrm{s}$
ϕ_c	Uniform	-0.1	0.1

Table 2.4: Prior distributions and bounds used for 1 yr MBHB PE analyses. All angular quantities are expressed in radians; see Table 2.3 for symbol definitions. For notational convenience, lower and upper prior bounds are defined with respect to the parameters of the injected source (also defined in Table 2.3); these bounds were identified through manual tuning to minimise PE computational costs without significant truncation of the posterior distribution.

MBHB analyses, as the high SNR of the injected source enables one to constrain parameters very precisely. We tuned the posterior bounds manually such that the posterior bulk would not be truncated in any of our PE runs. As our objective is not to perform a search, but instead to identify any biases in parameter recovery, we believe this tuning of the prior bounds to be acceptable and to have minimal impact on the results obtained.

Even with this GPU-accelerated waveform model setup, PE for MBHBs is still fairly expensive, typically taking $\mathcal{O}(1\,\mathrm{d})$ per run. Given this high wall-time, we limit the scope of our analysis to two constellation orbit scenarios: equal-armlength and generic orbits. We perform PE for both noise-free and noisy simulations of LISA and each set of constellation orbits. The recovered posterior distributions are shown in Figure 2.5. For the noise-free PE runs, we recover posterior distributions that are well-centred on the injected parameters. The posteriors differ somewhat in both their shape and overall width for the two orbit scenarios considered; this is because the antenna response of the LISA instrument (and thus the recovery of the signal SNR and parameters) differs between the two cases. As neither of the posteriors shown any significant deviation from the injected parameters, we can assume that (at this point in the MBHB parameter space) our recovery of MBHB signals is free from any significant systematic biases.

For the noisy simulations, PE results are generally consistent with the injected parameter values. Some scatter is expected, and for an 11 parameter model it is not unlikely for one or two parameters to deviate at the 90% level. This is more or less what is observed for the equal-armlength orbit results, but the generic orbit posteriors appear to deviate a little more frequently than this (albeit, not to the degree that any biases larger than $\sim 90\%$ are observed). There may therefore be some minor source of systematic bias associated with the instrument

noise model that has not been fully accounted for. Examining the whitened strain shown in Figure 2.5, the maximum-likelihood template closely fits both the injection and the noisy simulation with generic orbits, reinforcing the idea that any biases present are likely to be extremely small (and only detectable due to the high SNR of MBHB signals). Further investigation of MBHB recovery with our end-to-end pipeline is an area we plan to investigate in future work, once current limitations regarding waveform model and simulation efficiency are addressed and more extensive analyses may be performed.

2.3.3 Extreme-mass-ratio inspirals

Last, we consider the injection and recovery of EMRIs. These GW signals are particularly challenging to analyse due to the extreme precision with which their parameters may be recovered, which exceeds one part in a million for reasonable astrophysical sources (see Section 3.4 for further discussion). Unlike MBHBs, the SNR of EMRIs accumulates gradually over a timescale of years as their broad-band spectra of harmonic modes sweeps across the LISA band. This precise recovery implies that the phase of the waveform (which occupies a large portion of the measurement band) must be accurately modelled throughout this long duration. Such stringent requirements make EMRIs excellent candidate GW sources for testing the PE consistency of our end-to-end pipeline: if small systematic biases are present, we can expect them to appear most strongly in EMRI analyses.

As with MBHBs, no observations of EMRI sources have been made; we therefore construct a representative EMRI source with parameters given in Table 2.5. To reduce the computational cost of PE to acceptable levels, we neglect the spin of the MBH and inject a non-spinning EMRI system. The SNR of the source is $\mathcal{O}(50)$, which lies well within the expected SNR distribution of EMRI sources [190].

The current state-of-the-art in EMRI waveform modelling (in terms of the construction of the source-frame GW strain) is the FEW software package [177], which is described in Section 3.1. We use the FASTSCHWARZSCHILDECCENTRICFLUX waveform model [177], and therefore neglect the spin of the MBH in the analyses performed in this section, as this significantly increases the cost of each waveform evaluation. As even non-spinning EMRI systems are challenging to analyse (the requirement of long-duration broad-band phase coherence has not changed), we do not expect this choice to affect the conclusions we draw from these analyses. As with other GW sources, in order to apply EMRI waveform models to the task of LISA data analysis we require a model of the LISA response function. This is an area of ongoing research for EMRI waveforms; as of the writing of this thesis, no specific method for applying the LISA response that is tailored to these sources has yet been developed. Therefore, as with our MBHB analyses, we turn to the FASTLISARESPONSE code to perform this stage of the likelihood computation. As this code is generic to the form of the waveform strain, applying it to EMRI waveforms is no more complex than the MBHB case.

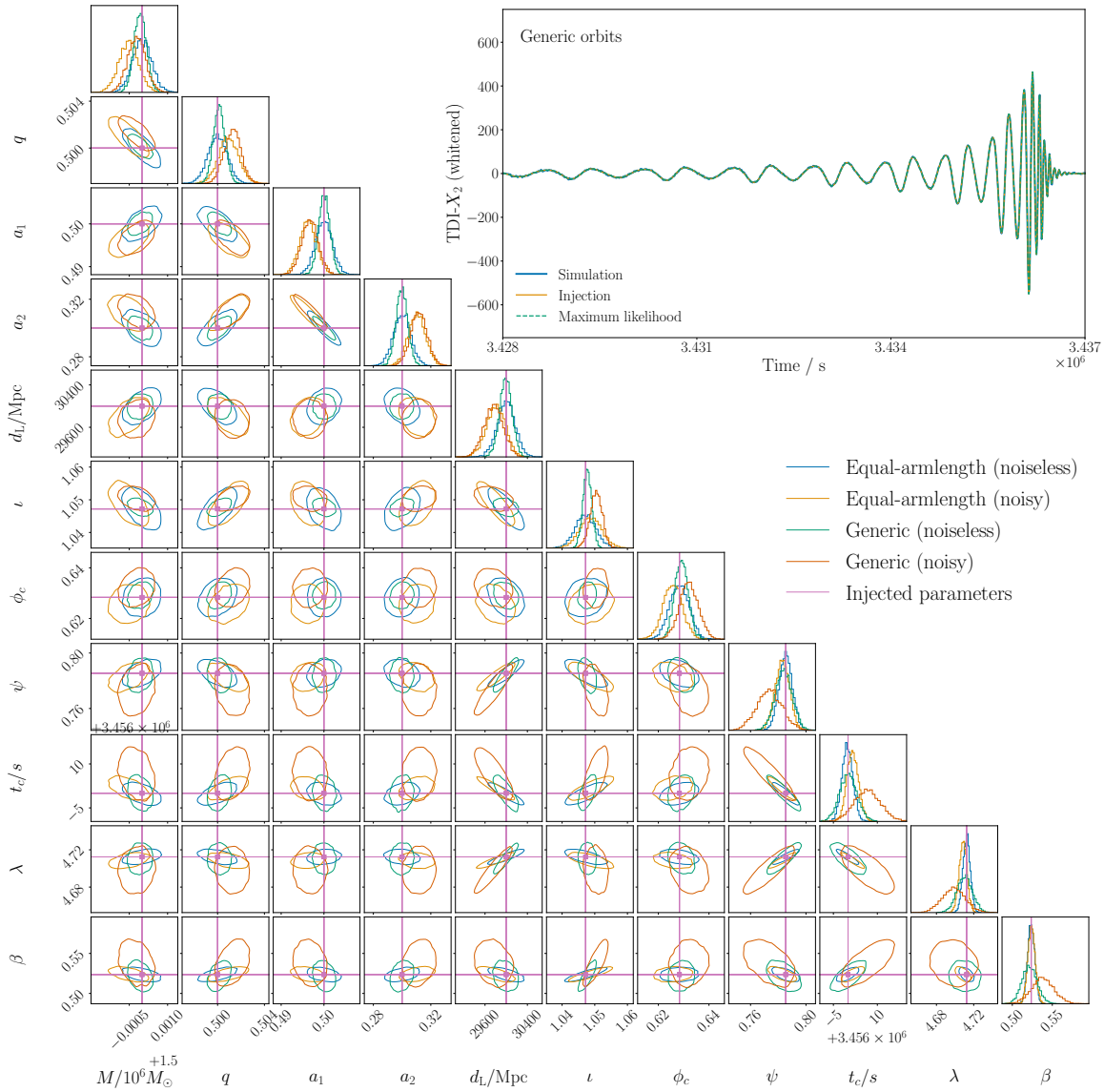


Figure 2.5: Posterior distributions obtained from PE analyses of simulated data containing a single injected MBHB signal (of parameters described in Table 2.3 and indicated by the pink crosshairs). We perform four analyses, examining both noiseless and noisy simulations with equal-armlength and generic spacecraft orbits. No systematic biases due to waveform template mismatch are observed in the noiseless simulations for either orbit scenario considered. Introducting instrumental noises, the equal-armlength orbit posteriors appear to be more consistent with the injected values than those of the generic orbit analysis, which are marginally consistent with the injected values (with many parameters only agreeing with the results obtained at the 90% level). The inset panel shows the whitened time-domain strain in the TDI-X_2 channel, for a noisy simulation with generic orbits (blue line). The injected signal (yellow line) and maximum-likelihood template (green dashed line) are nearly indistinguishable from the data, indicating that any systematic biases present are small and only apparent due to the high SNR of the injected signal (~ 1000).

Parameter	Symbol	Reference value
Primary mass	M	$10^6\,\mathrm{M}_\odot$
Secondary mass	μ	$10^1\,\mathrm{M}_\odot$
Initial semi-latus rectum of inspiral	p_0	10
Initial eccentricity of inspiral	e_0	0.3
Luminosity distance	d_L	1 Gpc
Ecliptic latitude	β	$\pi/4$
Ecliptic longitude	λ	$\pi/3$
Polar spin angle	θ_K	$\pi/5$
Azimuthal spin angle	ϕ_K	$\pi/6$
Initial azimuthal phase of inspiral	$\Phi_{\phi,0}$	$\pi/5$
Initial radial phase of inspiral	$\Phi_{r,0}$	$\pi/2$

Table 2.5: Source parameters for our chosen reference EMRI system. As no such systems near coalescence are known, we construct a system with astrophysically representative parameters with SNR ~ 100. The parameters describing the initial conditions of the inspiral (denoted with subscript 0) are defined with respect to the epoch t_init. All angular quantities are expressed in radians.

The FEW package is GPU-accelerated and is sufficiently rapid that we achieve computational wall-times of $\mathcal{O}(200\,\mathrm{ms})$ per likelihood evaluation over the $10^{-4}\,\mathrm{Hz}$ to $10^{-2}\,\mathrm{Hz}$ frequency band. While methods for fast EMRI waveform generation are investigated in this thesis (in Chapter 3), these techniques are not useful here because we must include the response function via FASTLISARESPONSE. Once a frequency-domain LISA response code for EMRIs has been developed, waveform costs will be reduced significantly with the inclusion of such techniques. The computational cost obtained per likelihood call translates to an inference wall-time of $\sim 1.5\,\mathrm{d}$, which is of similar speed to the MBHB case.

We begin by running a noise-free simulation with equal-armlength orbits, which is our most simple scenario. Unlike with other GW sources, we observe systematic biases in the recovered EMRI posteriors that are dependent on the size of the MHz beatnote frequency measurements. As these simulations are noise-free, any biases present cannot be associated with our model of the instrument noise. One possibility is that the numerical precision noise floor of the simulations, which will scale with the size of the beatnote frequencies, interferes more strongly with the EMRI analysis than those of other GW sources due to the weak nature of the signal. This noise floor was identified in [71], but was assumed to be safely ignorable due to the ASD of this noise lying two orders of magnitude below the LISA noise floor. However, if this noise is the result of rounding errors due to insufficient numerical precision, it cannot simply be treated as additive Gaussian noise in the same manner as instrumental noise sources. If a GW signal is present, the noise will be correlated with the signal in some way, which can be considered as a direct modification of the injected signal; as our template model is no longer consistent with the injected signal, we will observe a systematic bias due to waveform mismatch.

Parameter	Prior distribution	Lower bound	Upper bound
M	Uniform	$-10^2\,\mathrm{M}_\odot$	$10^2\,\mathrm{M}_\odot$
μ	Uniform	$-10^{-3}\,\mathrm{M}_\odot$	$10^{-3}\,\mathrm{M}_\odot$
p_0	Uniform	-10^{-4}	10^{-4}
e_0	Uniform	-10^{-4}	10^{-4}
d_L	d_L^{-3}	$-0.1\,\mathrm{Gpc}$	$0.1\,\mathrm{Gpc}$
β	$\cos\beta$	-0.1	0.1
λ	Uniform	-0.1	0.1
θ_K	$\sin\theta_K$	-0.1	0.1
ϕ_K	Uniform	-0.1	0.1
$\Phi_{\phi,0}$	Uniform	-0.1	0.1
$\Phi_{r,0}$	Uniform	-0.1	0.1

Table 2.6: Prior distributions and bounds used for 1 yr EMRI PE analyses. All angular quantities are expressed in radians; see Table 2.5 for symbol definitions. For notational convenience, lower and upper prior bounds are defined with respect to the parameters of the injected source (also defined in Table 2.5); these bounds were identified through manual tuning to minimise PE computational costs without significant truncation of the posterior distribution.

To examine this hypothesis, we perform more noise-free simulations with an idealised locking configuration (known as *six*) in which all lasers are independently locked to separate reference frequencies. These simulations are otherwise identical to the first simulation we performed above, which used a realistic LISA locking scheme referred to as *N1-12* (see [308] for further details). By tuning the independently-locked laser frequencies, we can finely adjust the scale of the beatnote frequencies. In one simulation, we set the reference frequencies of each laser such that two of the inter-spacecraft beatnote frequencies are $\sim \pm 25\,\mathrm{MHz}$, close to the upper end of the measurement band of the onboard phasemeters. We then perform a second simulation with laser reference frequencies set such that these beatnote frequencies are $\sim \pm 10\,\mathrm{MHz}$, similarly to those of the "N1-12" simulation; we denote this second simulation as *six-alt*. As there are no instrumental noises (which also means that onboard clocks are assumed to be perfect), these locking schemes should perform identically. The posteriors obtained for the PE analyses of the *N1-12*, *six* and *six-alt* simulations are shown in Figure 2.6, with the inter-spacecraft beatnote frequencies as a function of observation duration indicated in the inset panel of the figure. Strong systematic biases exceeding the 90% level are observed in the *six* posteriors, which are almost completely corrected for with the adjustment of the beatnote frequencies in *six-alt*. These results strongly indicate that numerical precision errors in the simulation and analysis pipeline are a source of significant systematic biases. Some systematic bias still remains in the *six-alt* results despite the beatnote frequencies being of similar magnitude to those in the *N1-12* simulation. This may be due to more complex interactions between the reference frequencies and the beatnote measurements in other on-board interferometry systems (such as the reference or test-mass interferometers). A more extensive examination of the impact of laser locking on the numerical

precision noise in the simulation output is an area of investigation for future work.

It is not clear precisely where the loss of numerical precision enters the data products of the L0-L1 stage of the pipeline. If this effect enters at the measurement of the beatnote frequencies by the on-board phasemeters, it would be particularly concerning as there would be no way to correct for the loss of resolution in on-ground processing and modifications to the precision of beatnote measurement and transmission would be required, which would have further implications for instrument hardware and data transmission rates. However, it is more likely that the loss of numerical precision enters in this on-ground processing stage; the delay and linear combination operations of total-frequency TDI entail interpolations which place further requirements on the numerical precision of the data. These operations are already performed a double-precision, but could be upgraded to extended precision if this is required; while this would make TDI significantly less performant, this is preferable over a loss of precision at the beatnote measurement stage as GW information can still be recovered. The stage that numerical precision errors degrade the GW measurements in the simulation of the instrument will be investigated further in future work.

As we have identified that systematic biases are present even for noise-free simulations with our simplest orbit scenario, it would be hard to assess the results of simulations with instrumental noises included. We therefore leave these investigations to future work, which should be carried out once the precise mechanism by which numerical precision errors couple to the L1 data products has been identified and corrected for.

2.4 Validation of LISA parameter estimation performance

While the simulation and analysis of a single GW signal facilitates a coarse appraisal of PE consistency, it is not an effective approach for identifying subtle systematic biases in these analyses associated with incorrect modelling of the noise. Inclusion of the noise adds statistical fluctuations to PE results by construction. To detect systematic biases of a scale smaller than these fluctuations requires that PE be repeated many times on new instances of simulated data in order to separate a deterministic shift in parameter estimates from stochastic variations due to changing detector noise. In this section, we repeatedly simulate and analyse LISA data containing injected GW sources, and examine whether there is significant evidence for a systematic bias in the ensemble of PE results obtained. In Section 2.4.1 we introduce and demonstrate an example of the probability–probability (P–P) test, which we use to probe for the existence of significant systematic biases given a set of PE results. We then describe how these results are obtained: changes in the configuration of these simulations (with respect to those of Section 2.3) are described in Section 2.4.2, and the procedure for performing each PE analys is described in Section 2.4.3. Last, the results of the P–P tests are presented in Section 2.4.4.

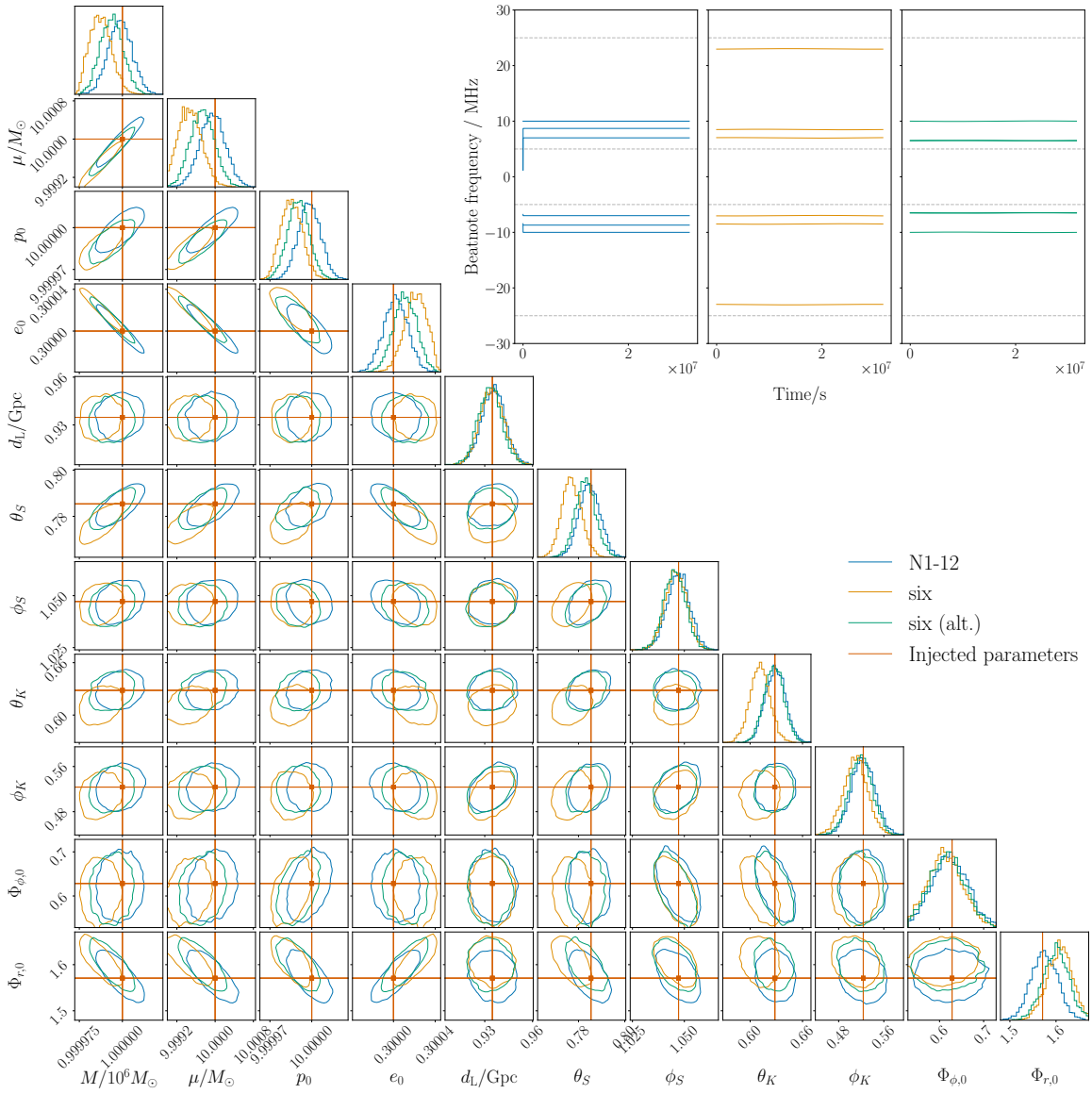


Figure 2.6: Posteriors obtained from the analysis of noise-free equal-armlength orbit simulations containing a single EMRI source (with parameters described in Table 2.5 and indicated by the red crosshairs). Each simulation uses a different locking scheme for the on-board laser systems: *N1-12* (blue line) uses a realistic locking scheme where all lasers are locked to a single reference; *six* (yellow line) uses an idealised scheme where all lasers are locked to individual reference frequencies; and *six-alt* (green line) uses this same scheme but with these reference frequencies tuned to reduce the scale of the inter-spacecraft beatnote frequencies. We observe systematic biases in the recovered posteriors that scale with the size of the inter-spacecraft beatnote frequencies, suggesting that the effects of limited numerical precision on the injected EMRI signal are a source of bias in the end-to-end pipeline. The inter-spacecraft beatnote frequencies are shown in the inset panel; the sharp changes to the *N1-12* beatnote frequencies at the start of the simulation are edge-effects due to the interpolation of simulated orbits and are not physical. This initial segment of the data is trimmed at the L0.5-L1 stage and does not enter our PE analysis.

2.4.1 Probability-probability plots

The P–P test is a statistical test of PE consistency [337]. It examines the results of an ensemble of PE analyses of simulated datasets and compares statistically how the recovered posteriors are distributed with respect to the parameters of the signals injected in the simulations. This statistical comparison makes use of a defining property of Bayesian posterior distributions: if the posterior has been specified correctly (i.e., the assumed model and noise properties are accurate with respect to the data), then for a given simulation the parameters of the injected signal will lie within the q% credible interval (CI) of the posterior with a probability of q. This is simply the definition of the posterior distribution, with an added conditional statement that is only relevant for simulated data where the concept of a true signal or noise model is well-defined. The use of simulated data in P–P tests also makes them sensitive to the accuracy of these simulations, as if the produced data does not match expectations the P–P test may fail. The P–P test serves as a useful tool for validating the performance of the sampling method employed in PE, and has been used to confirm the performance of sampler implementations in the literature [246, 254, 338]. The wide range of systematic effects that may be probed with P–P tests makes them a versatile statistical technique, but can make it challenging to determine the precise cause for a failed P–P test. It is therefore important to develop strategies for ruling out each potential contributor to a failed test, for instance by using a well-tested sampling method or performing noiseless simulations to identify systematic biases that are independent of the noise model.

The comparison of the injected parameters with the posterior CI is probabilistic in nature; fundamentally, it is a measure of the similarity of two PDFs. The hypothesis that these two distributions are equivalent can be tested with a Kolmogorov–Smirnov (K-S) test, which examines the similarity of two probability distributions given samples from one distribution and the PDF of the other. In this case, these samples are injected parameter CIs (from the ensemble of PE analyses), which are compared against a uniform PDF (the expected distribution of these CIs if PE is optimal). The p-value returned by the K-S test indicates whether the distribution of CIs obtained is sufficiently inconsistent with a uniform distribution to suggest that the PE process is sub-optimal; this is subjective to an extent, requiring that one specifies a criterion for a statistically significant p-value. This only presents an issue for marginal P–P test failures or successes (as significant biases will result in extremely small p-values), and may be rectified by repeating the test with a larger sample size.

A P–P test is typically accompanied by a graphical representation of the cumulative density functions (CDFs) of the two distributions, known as a P–P plot. This plot provides additional information useful for identifying why a set of PE analyses do not pass the P–P test. There are four primary failure modes of a P–P test. Two are associated with a systematic shift in the posterior with respect to the injected parameters (i.e., preferentially underestimating or overestimating a parameter). The other two are the result of the recovered posterior having a width that is not consistent with the injected value. For instance, if the posterior is too wide, the P–P test will fail

despite the injected parameters consistently lying within the posterior bulk; this is the expected outcome, as the statistical fluctuations in the PE results are of a magnitude that is inconsistent with those of an optimal PE process.

As an illustrative example of these four modes of failure, consider a scenario in which we are attempting to measure a parameter θ_0 given noisy data. We simulate the posterior PDF from each PE run as a normal distribution such that if PE is being performed correctly, each posterior will have unit variance and a mean drawn from a unit normal distribution with respect to θ_0. We can simulate each failure mode of the P–P test by adding the corresponding systematic effect to our simulated posteriors, with parameter shifts and incorrect posterior widths being represented as a constant offset to the posterior mean and a scaling of the posterior variance respectively. The CI of the injection with respect to each posterior is obtained from the posterior CDF, which is simply a normal CDF in this example. The P–P plots for both an optimal PE process and each of these failure modes for this toy problem are shown in Figure 2.7, for sets of $N = 128$ simulated PE results.

In principle a P–P test is only satisfied for a P–P plot which follows the diagonal, which corresponds to uniform distributions for both CIs and their corresponding fractions of the injected parameters. However, there will be statistical uncertainty because we only have N PE results; upper bounds of the confidence intervals for deviations from the diagonal due to this uncertainty are obtained by treating the P–P plot process as a Brownian bridge [339]. In Figure 2.7, the 68%, 90% and 99% confidence intervals are indicated by the shaded regions. The optimal PE analysis lies within the confidence intervals, while the intentionally-biased PE analyses characteristically deviate from this region. These significant differences are reflected in the p-values produced by K-S tests of each distribution with respect a uniform distribution; the optimal PE process has $p = 0.3$ in this case (which is acceptable), whereas the biased PE processes have $p \ll 10^{-2}$ (suggesting that the hypothesis that they are consistent with a uniform distribution be rejected). For multivariate posterior PDFs, P–P tests are performed independently on the one-dimensional marginal PDFs, producing a set of p-values that must be considered collectively to judge whether the P–P test has passed overall. This process can be simplified by combining these p-values with e.g. Fisher's method [340], but this must be done with care as it can obfuscate the failure of the P–P test for one parameter in a large model. One should therefore consider both the p-values and the P–P plot when judging the results of the P–P test.

Performing a P–P test in practice entails the specification of a prior range of parameter values over which to validate PE performance. However, simply drawing a uniform sample of injection parameters from this prior range is incorrect: if the prior specified in the definition of the posterior is non-uniform, the P–P test will fail for sufficiently large N, depending on how informative the prior is. For instance, a particularly uninformative analysis will return a prior-dominated posterior PDF that is largely insensitive to the injected parameter values, leading to systematic biases unless the injected parameter distribution matches that of the prior. This issue

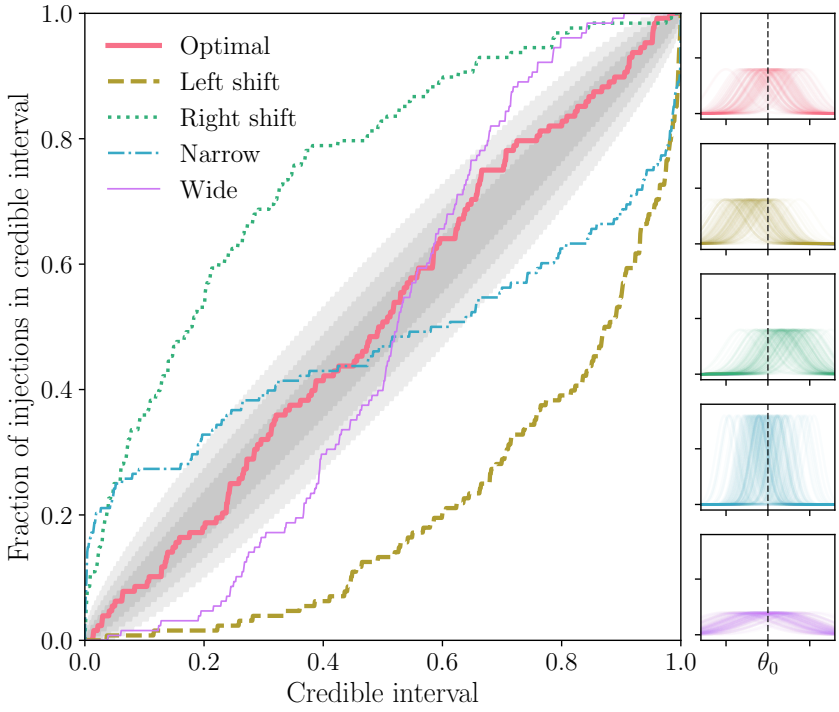


Figure 2.7: P–P plots (left column) from five sets of $N = 128$ simulated posterior distributions (right column) for a toy problem of estimating a single parameter θ_0 (black dashed line). If PE is performing optimally (pink line), the P–P plot should be consistent with a diagonal line (with some deviation due to finite N, as indicated by the 68%, 90% and 99% confidence interval shaded regions). A systematic shift to the left (olive) or right (green) of the posterior with respect to θ_0 leads to a complete deviation of the P–P plot from the diagonal, with the deviation direction indicating the nature of the shift. If the recovered posteriors are systematically too narrow (blue) or too wide (purple) a deviation which crosses through the diagonal will be observed. These two effects can mix together for PE analyses which recover shifted posteriors of improper width to produce a combination of these deviations.

can therefore be avoided by sampling the injection parameter values from the same prior that is assumed in PE.

2.4.2 Simulations for an ensemble of parameter estimation analyses

Performing many LISA simulations and PE analyses is a computationally expensive task. While these analyses are independent (and therefore an ideal target for parallel processing), the high RAM usage of each instance limits the number of simulations that may be run simultaneously. Each simulation of 1 yr duration at a sampling cadence of 10 s requires more than half of the available memory of each cluster node, further restricting the number of processes that may be run in tandem; if the resources are sufficiently busy, simulation computing requests can remain stalled for long periods of time. We find that constructing a set of PE results from 1 yr is therefore highly impractical as a result of these technical limitations.

Efforts are ongoing to reduce the memory footprint of LISAINSTRUMENT; ensemble PE studies for full-duration LISA simulations will be performed once this work is complete. In the meantime, to make progress we reduce the duration of our simulations for the ensemble PE analysis to 3 d. Memory requirements scale linearly with the observation time, so this choice enables many simulations to be run in parallel. This also reduces memory usage sufficiently that we are able to set a realistic sampling cadence of 0.25 s and enable the final filtering stage described in Section 2.1.3. However, these shorter simulations do not retain all aspects of the LISA mission complexity, as long-timescale effects such as clock drift and orbit reconstruction errors have less impact on this timescale. We nevertheless expect these short-duration simulations to capture the essence of the LISA data analysis problem. The fundamental challenges inherent to LISA data (such as the suppression of primary laser noise via TDI, reduction of clock jitter and unbiased recovery of GW signals) will still be limiting on this timescale if post-processing or source modelling is performed incorrectly.

The P–P test requires that all GW signal injections are produced with the same model (and are therefore all of the same source class). It is also essential for PE to be sufficiently computationally inexpensive that it does not bottleneck the simulation and analysis pipeline. We choose to inject GB signals for the following reasons:

- FASTGB is highly computationally efficient, being a factor $\mathcal{O}(10^3)$ faster than our MBHB or EMRI waveform models.
- On short timescales, these signals can be described by 5 well-constrained parameters, compared to ~ 10 parameters for EMRIs and MBHBs. The computational cost of PE increases as the parameter-space volume grows.
- The posterior structure of GBs in these parameters is extremely simple (when compared to MBHBs and EMRIs) and therefore easy to sample efficiently with NESSAI.

- As narrowband sources, the inner product in the GB likelihood is computationally inexpensive, facilitating a likelihood CPU wall-time of $\mathcal{O}(100\,\mu\text{s})$.

As systematic biases with the FASTGB model were observed for generic constellation orbits at high SNRs in Section 2.3.1, we generate sets of PE results for both the Keplerian and generic orbit models to examine whether the impact of these biases is detectable for each case. Collectively, adjusting the observation duration and injecting only GB signals reduces the cost of simulation and analysis sufficiently for an ensemble of 128 PE results to be obtained over a timescale of $\mathcal{O}(2\,\text{d})$, with each result taking $\mathcal{O}(2\,\text{hr})$ of CPU time to produce. We find this wall-time to be acceptable for the development and application of the P–P test framework.

The parameters of the injected GB signal are sampled from the prior distributions specified in Table 2.7. Due to the short duration of the simulation, some parameters of the GB cannot be strongly constrained with PE:

- The frequency derivative $\dot{f}$ has negligible impact on the waveform over 3 d and cannot be measured (except for the placing of a high upper limit).
- The sky position angles of the source (β, λ) are poorly constrained, with the two-dimensional marginal likelihood containing multiple modes of high probability that are connected in a complicated manner (see Figure 2.8 for an example of this). This increases the computational cost of PE by orders of magnitude, as this complicated shape is very difficult to sample efficiently.

These features of the PE analyses are not representative of the full GB PE problem; they are artefacts of our shortening of the simulation timespan. For simplicity, we therefore elect to fix these three parameters at their true values during PE and sample over the remaining GB parameters in Table 2.7. For all injected sources, we set astrophysically-motivated values of $\dot{f} = 10^{-17}\,\text{Hz}\,\text{s}^{-1}$ and $(\beta, \lambda) = (-0.10, 4.66)$; the latter are set to the position of Galactic Center, where we expect that a significant fraction of the GB population are located [331]. We do not expect this choice to significantly affect the conclusions of our P–P test analyses. The capability of LISA to accurately recover these parameters will be investigated in future work, once full-duration simulations with LISAINSTRUMENT are computationally feasible to perform.

2.4.3 Ensemble parameter estimation procedure

The procedure for estimating the parameters of GBs in 3 d simulations is not significantly different to that of the 1 yr simulations described in Section 2.3.1. As the in-band ASD of the high-frequency spectral leakage has decreased as a result of the increased sampling rate of the data stream, we adjust the taper width of the window function (Equation (2.3)) accordingly. We find that $x_{l,r} = \pm 10^3$ s improves the frequency resolution attainable during PE without compromising instrumental noise behaviour. The detrended and windowed TDI observables are then

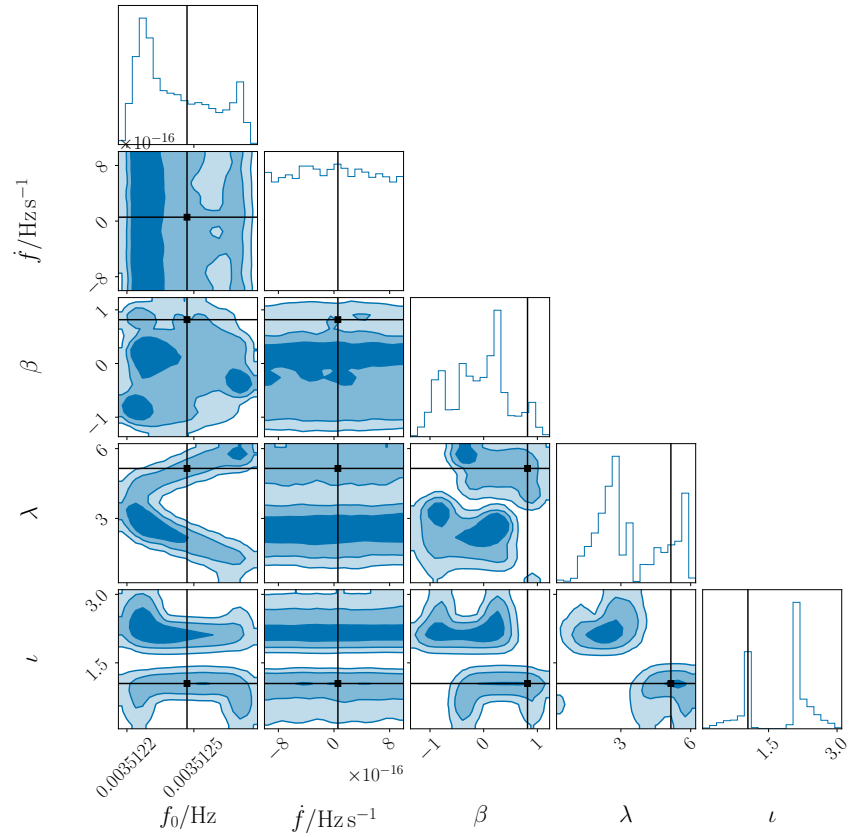


Figure 2.8: A subset of the PE results for our reference source HM Cnc (with amplitude rescaled such that the SNR is ~ 100) and no parameters fixed. The parameters of HM Cnc are indicated by the black crosshairs. Despite the representative SNR of the signal, the frequency resolution of the simulation (which scales inversely with the observation length) is too poor for $\dot{f}$ to be constrained to within the wide prior bounds. The LISA response does not significantly change over 3 d, so the sky location of the source (β, λ) is poorly constrained and highly correlated with other parameters of the GB system. These correlations complicate PE, making it significantly more computationally expensive. However, they are not representative of an SNR 100 signal observed over e.g. 1 yr; we therefore choose to fix these parameters for our P–P tests.

Parameter	Prior distribution	Lower bound	Upper bound
A	Uniform	10^{-22} (5×10^{-22})	5×10^{-21} (10^{-20})
f_0	Uniform	10^{-4} Hz	10^{-2} Hz
ι	$\sin\iota$	0	π
ψ	Uniform	0	π
ϕ_0	Uniform	0	2π

Table 2.7: Prior distributions and bounds used for the ensemble of PE analyses performed on 3 d simulations of a single GB source. Brackets indicate the original prior amplitude range assumed, which led to a failing P–P test due to the high SNR of some of the injected signals. See Table 2.1 for symbol definitions.

transformed to the frequency domain and a 20 μHz segment containing the signal is extracted for use in PE. While this segment is narrow, the bandwidth of the GB waveform is narrower still (< 1 μHz) because the simulation is too short for significant Doppler modulation of GW signals.

We slightly adjust FASTGB to account for the change in the duration of the data stream. We reduce the model parameter N (which controls the number of frequency bins in the slow part of the decomposed waveform) from 256 to 64. The reasons for this choice are as follows. First, the frequency bin width of the data has increased by two orders of magnitude (such that 256 bins now spans a 1 mHz frequency band), so we should reduce N accordingly to more accurately reflect the bandwidth of the GB waveform. Second, the GW signal has not been modulated to the same extent as in longer simulations (as was noted above) and therefore fewer frequency bins are required in the slow part of FASTGB. The impact of both the adjustment of N and this truncation at the edge of the band was examined via PE of noiseless simulations and was not found to be significant for the SNRs considered in the analyses of this section.

The prior distributions assumed in PE are identical to those given in Table 2.7 for all parameters except f_0, which we restrict to a narrow range of width 10 μHz. We chose the size of this range such that it fits within the 20 μHz segment width, leaving some room near the edges of the segment to account for the bandwidth of the waveform templates. As in the 1 yr simulations, setting a narrow range on f_0 significantly improves PE convergence times, which is vital for our P–P test analyses in which PE is repeated many times. However, this must be performed with care; one cannot simply place a symmetric prior range with respect to f_0 in each instance, as the injected values would not be distributed according to the prior and a bias would enter the P–P test. We retain this property of the injected signals by dividing the LISA frequency spectrum into 10 μHz segments. Before the analysis of each simulation, we identify which segment contains f_0 and use the bounds of this segment as our prior bounds in PE. As the distribution of injected f_0 values is uniform throughout the wide frequency band specified in Table 2.7, it will also be uniform within each segment, restoring the property of prior-distributed injections in the PE analyses. Accounting for this effect is particularly important in the analysis of low-SNR GBs in the injection set, for which the signal is unresolvable and PE recovers prior-dominated posterior distributions.

With our modified prior and likelihood functions defined, we perform PE on each simulated dataset with NESSAI in the same manner as was described in Section 2.2.3. We then compare the posterior distributions obtained with their corresponding injected values to produce P–P plots, as we will now describe.

2.4.4 Results

Bringing together Sections 2.4.1 to 2.4.3, the process for performing a P–P test in the framework of the end-to-end pipeline is as follows:

1. We draw N sets of parameters from their priors (Table 2.7), and inject the corresponding signals at the start of the pipeline, performing short-duration simulations configured according to Section 2.4.2.

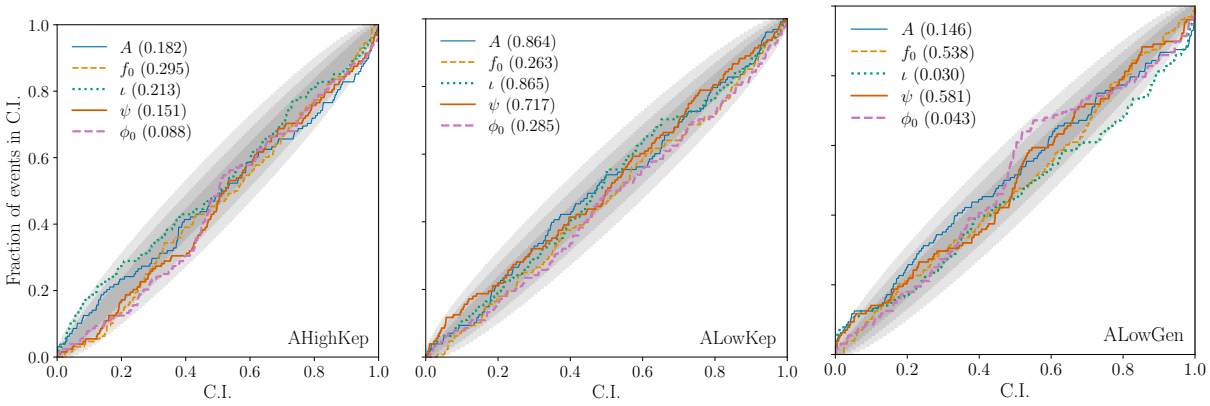
2. Posterior distributions are obtained for each injection via PE, as described in Section 2.4.3.

3. For each source parameter θ, we obtain a series of N confidence intervals q_θ of the injection parameter's value. The confidence intervals are estimated by marginalising the posterior over all parameters except θ and computing the quantile of the injection parameter's value with respect to this marginalized posterior.

4. We produce a P–P plot and compute a p-value (from K-S tests) for each q_θ according to the procedure described in Section 2.4.1.

First, we perform a P–P test for the Keplerian orbit scenario and inject sources with amplitudes drawn from the bracketed prior range in Table 2.7; we denote this orbit-prior combination as AHighKep. The corresponding P–P plot is shown in panel (a) of Figure 2.9. The AHighKep P–P test marginally passes, with a combined p-value of 6×10^{-2}; ϕ_0 is recovered especially poorly, but none of the parameters have particularly high p-values. While statistical error due to the number of PE analyses is certainly a contributing factor, it is worth considering sources of systematic bias in these analyses. One such source is waveform modelling error: to investigate this possibility, we compute the SNR distribution of the AHighKep injections, which is shown in Figure 2.10 by the blue line. The tail of this distribution reaches an SNR of $\sim 5 \times 10^2$, which approaches the level where significant PE biases were observed due to the assumptions underlying the FastGB waveform model in Section 2.3.1. While the systematic biases in the PE for signals in the upper-tail of the SNR distribution would be small compared to the posterior width, they will cumulatively impact the ensemble of PE results and cause a P–P test to fail. To examine this hypothesis, we perform a second P–P test on an injection set with the lower (unbracketed) amplitude prior in Table 2.7, which we denote as ALowKep. The highest SNRs of the ALowKep injection set (indicated by the orange dotted line in Figure 2.10) are roughly a factor of two lower than those of the AHighKep analysis; we therefore expect the effects of waveform mismatch bias to be significantly diminished in the ALowKep P–P test. The P–P plot obtained (shown in panel (b) of Figure 2.9) is more consistent with the diagonal than in the AHighKep case, indicating that the PE results for the high-SNR sources in the injection set of that analysis were not satisfying the requirements of the P–P-test. None of the p-values obtained in the ALowKep P–P test are particularly low, which supports the conclusion that the PE results in this sample of 128 analyses are consistent with those that would be obtained from an optimal PE process. This confirms that the LISA simulation and analysis pipeline produces reliable parameter estimates that are consistent with our statistical expectations.

We finish by considering a third scenario, in which we use the same injection set as the ALowKep analysis but switch to the generic orbit configuration; we therefore refer to this as the



(a) (b) (c)

Figure 2.9: P–P plots for sets of 128 PE analyses of 3-day LISA simulations containing individual GB signals. The p-values returned from K-S tests performed on each curve are indicated in the legend in brackets. The parameters of the injected signals are distributed according to the prior distributions specified in Table 2.7. We consider three scenarios: Keplerian spacecraft orbits and higher amplitude prior bounds (AHighKep, a); Keplerian spacecraft orbits and lower amplitude prior bounds (ALowKep, b); and generic spacecraft orbits and lower amplitude prior bounds (ALowGen, c). The P–P test marginally passes for AHighKep, passes well for ALowKep and marginally fails for ALowGen; this is reflected in the p-values obtained in each case. These results can be attributed to the impact of waveform mismatch biases in high-SNR members of the injection sets (c.f. Figure 2.10), which are more severe at higher SNRs and for the generic orbit scenario.

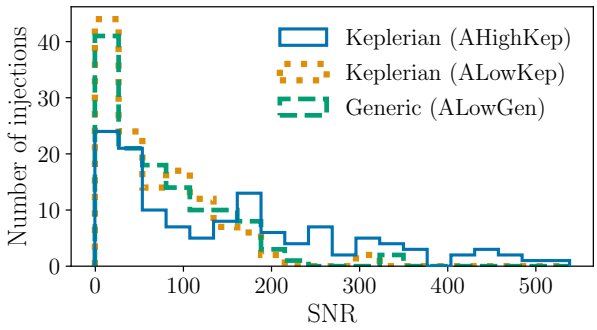


Figure 2.10: The SNR distributions of GB injection sets corresponding to the P–P tests performed in this section. The AHighKep injection set (blue solid line) has an upper amplitude prior bound a factor of two larger than the ALowKep (orange dotted line) or ALowGen (green dashed line) sets, leading to a proportional increase in the SNRs of the loudest signals. The ALowKep and ALowGen populations contain identical signals, but differ slightly in SNR due to the different spacecraft orbit configurations of each case.

ALowGen analysis. The SNR distribution of the injection set (green dashed line, Figure 2.10) is largely unchanged despite the change in the LISA response, as the spacecraft orbits mainly impact the GW phase. The P–P plot obtained from this analysis is shown in panel (c) of Figure 2.9. We once again obtain a marginal result from the P–P test that is only slightly worse than what was obtained for the ALowKep analysis. It is likely that a similar explanation is warranted here, as the accuracy of the FastGB waveform model significantly decreases for generic orbits (as was demonstrated in Section 2.3.1). The systematic biases from waveform mismatch will therefore have greater impact on the ALowGen analysis, despite all other aspects of the simulations and injected signals remaining unchanged with respect to the ALowKep analysis. We therefore expect that a P–P test would pass if the amplitude prior bounds were further reduced, but leave this investigation to future work.

The SNRs of the ALowGen injection set are not particularly extreme, and we expect the loudest GB sources to have SNRs of similar order to the upper tail of this distribution [331]. While the SNRs of the injected signals are below the level at which biases were identified for mis-modelled GBs in previous work, this level was inferred from the PE analysis of individual GB signals [330]. Despite the small size of PE biases at lower SNRs, our P–P tests demonstrate that they can still accumulate to constitute a significant source of error in PE. This is particularly relevant for LISA, where joint analysis of multiple signals is necessary and small systematic errors in the fitting of loud sources may significantly degrade PE for quieter sources at the limits of instrumental sensitivity. This highlights an urgent need for the improvement of the FastGB waveform model to properly model the effects of generic orbital configurations on the frequency-domain TDI response for GB signals. Such an extension of FastGB is an area of active research, and an upgraded version of the waveform model is under development.

In this section, we have demonstrated that the assumptions commonly made in LISA data analysis regarding the cancellation of primary noise sources and the PSD of the remaining instrumental noise are sufficiently accurate for optimal PE to be performed. These initial investigations into the performance of the instrument on short timescales confirm that essential components of initial noise reduction processing, such as TDI and the construction and application of MPR estimates, successfully produce observables from which GW signals may be extracted without biases. This essential work was presented to ESA as a component of the LISA adoption review, a landmark stage in the development of the LISA mission that was completed successfully in early 2024.

2.5 Conclusions

In this chapter, we have verified the performance of an end-to-end LISA simulation and analysis pipeline via the PE of injected GW signals. The pipeline framework was introduced in Section 2.1, consisting of: the implementation of the L0-0.5 simulation in Sections 2.1.1 and 2.1.4;

initial noise reduction and pre-processing of telemetred spacecraft data in Section 2.1.5; and the identification and performance of additional pre-processing steps prior to PE (such as detrending and windowing) in Section 2.2.3.

In Section 2.3, we performed PE for 1 yr simulations of LISA at the reduced sampling rate of 0.1 Hz and different GW source classes. Due to resource limitations, these investigations consisted of individual simulations of different instrument and orbit configurations, and were therefore limited to the identification of strong systematic biases. We found no evidence for these biases in our analysis of GB signals in Section 2.3.1 and verified that existing rapid GB waveform models are not sufficiently accurate at high SNRs, in line with the results of previous work. Similar investigations for MBHBs, described in Section 2.3.2, returned posterior distributions consistent with the parameters of the injected signals, with marginal evidence for small systematic biases that warrant follow-up in future work. Our analyses of injected EMRI signals in Section 2.3.3 yielded biased posteriors that were independent of instrumental noise; through subsequent investigation we determined that the scale of these biases was correlated with the size of the on-board beatnote measurements. We interpret this behaviour as the result of rounding errors due to insufficient numerical precision in either the L0-L0.5 or L0.5-L1 segments of the pipeline (or some combination of the two). However, we were unable to fully account for the biases we observed in our results. We plan to investigate this behaviour further in future work.

To perform a more extensive investigation of LISA PE performance, in Section 2.4 we mitigated the computational cost of our simulations by reducing the duration of our simulations to 3 d. Simulating and recovering 128 GB signals for three spacecraft orbit scenarios, we performed P–P tests to probe for small systematic biases in the PE results obtained from these analyses. We demonstrated that the waveform mismatch biases observed in Section 2.3.1 were also identifiable in our P–P test analyses, and constituted a statistically significant deviation in the PE results. This occurred despite our GB injections being of lower SNRs than this was previously observed for in isolated cases, providing further evidence that current GB waveform models are not sufficiently accurate for unbiased inference of GB signals with LISA. Once we corrected for these biases associated with waveform mismatch, we obtained a passing P–P test result, suggesting that the LISA data obtained via realistic simulations possessed both the signal and noise properties that were assumed during PE. This important result confirms that the suppression of laser and clock noises may be performed via TDI to a sufficient degree that the resulting LISA data streams are dominated by secondary noise sources.

The PE tests we performed in this chapter were far from exhaustive due to the restrictions imposed on their extent by the resources available. In future work, once the resource requirements of LISAINSTRUMENT have been reduced, P–P tests will be repeated for GBs injected in full-length LISA simulations. These investigations will be extended to include MBHBs and EMRIs once fast waveform modelling techniques for these source classes are available, examining LISAs noise performance in a wide frequency band over longer periods of time. A similar ap-

proach may be employed to examine the systematics of assumptions regarding the extraction of multiple signals such as inter-source confusion noise or the convergence of analysis algorithms that attempt to break the problem down by source type and iterate through them sequentially. The analysis framework developed in this chapter provides a useful foundation upon which these more complex analyses of the systematics of LISA data analysis can be tested. It is expected that these investigations will continue up to (and beyond) the launch of the LISA mission.

Chapter 3

Accelerating EMRI waveform generation with neural networks

I provided a high-level overview of the scientific importance of EMRI systems in Section 1.3 of Chapter 1. Modelling EMRI waveforms accurately is computationally expensive. In this chapter, I present a novel technique for mitigating this computational cost with ML techniques. First, in Section 3.1 I explain how EMRI waveforms are modelled in the gravitational self-force (GSF) formalism, and describe how cutting-edge techniques have accelerated EMRI waveform generation. Identifying that harmonic mode selection bottlenecks further performance improvements, in Section 3.2 I propose a novel technique for mitigating this computational cost which leverages ML to model this mode selection process, accelerating EMRI waveform generation. I then validate the performance of the trained model and apply it in an exploratory study of EMRI mode retention across parameter space. In Section 3.3, I verify the accuracy of EMRI waveforms generated using my approach through statistical comparisons with the original waveform models, and evaluate their computational cost via benchmarking. Through PE studies in Section 3.4, I demonstrate that the results of these benchmarks and comparisons hold for representative data analysis scenarios. Last, I present the conclusions of the work described in this chapter in Section 3.5.

3.1 Modelling extreme-mass-ratio systems

The exquisite complexity of EMRI systems owes primarily to the small mass ratio $q = \mu/M$ of the system. The CO of mass μ completes tens of thousands of orbital cycles in the strong field of the MBH of mass M before the two objects merge under radiation reaction [2, 341]. The eccentricity of this inspiral trajectory is typically $\sim$0.7 as the inspiral enters the LISA band [182, 342], giving rise to significant apsidal precession on the orbital timescale, which is further accentuated by the spin of the MBH due to frame-dragging effects [343]. For MBHs with non-zero dimensionless spin magnitude a, inclination of the inspiral trajectory with respect to the plane

of the MBH spin-angular momentum leads to Lense–Thirring precession of the orbital plane by spin-orbit coupling [343, 344]. The energy and angular momentum of the system is gradually lost as gravitational radiation, a dissipative process which causes the inspiral that eventually leads to the coalescence of the two objects [16, 345]. Together, the action of these phenomena over many orbital cycles gives rise to complicated behaviour: a short segment for an eccentric and inclined EMRI into a rotating MBH (sometimes referred to as a generic Kerr inspiral) is shown in Figure 3.1.

The properties and dynamics of an EMRI system can be studied by observing the GWs it produces [176]. Accurately modelling the dynamical evolution of the system, including the emitted gravitational waveform, is vital to their analysis [177, 313]. Waveform modelling errors lead to a decrease in detection sensitivity and biases in the estimation of source parameters from detector data, hampering the scientific potential of EMRI observations [347–349].

The two-body problem in general relativity is complicated; the effectiveness of a given method for solving it is strongly dependent on the properties of the system in question [327, 350–352]. Some features of EMRI systems make them particularly challenging to model accurately. The tens of thousands of orbital cycles completed by the CO during the inspiral must be modelled to great precision, as numerical errors will build up over the course of the inspiral [353, 354]. The extreme-mass-ratio (EMR) of the system introduces a significant difference in length scale between the two objects, which further challenges direct numerical methods [355]. Numerical relativity (NR) techniques, which have been used extensively in the modelling of comparable-mass binaries, are computationally intractable for the accurate simulation of such long coalescences at extreme mass ratios [356]. This immediately presents challenges for phenomenological waveform modelling techniques employed in the analysis of CBC signals for ground-based detectors, which rely heavily on calibration to NR simulations in the rapid production of accurate waveforms for use in data analysis [335, 357]. The effective-one-body formalism aims to capture the dynamics of generic CBCs in analytic form; in principle, it is capable of modelling extreme-mass-ratio systems [351]. This analytic description is made complete by free parameters which may be adjusted to match behaviour obtained with other methods, such as NR simulations [358]. Therefore, they suffer from the lack of numerical results in the EMR regime in the same way as phenomenological methods. However, they are more easily extended to EMRI modelling due to the ease with which this numerical data may be incorporated once it is available [359]. If the high computational cost of effective-one-body waveform models can be brought under control, they may see application to EMRI waveform generation and data analysis in the future. To develop data analysis strategies and assess EMRI science prospects in advance of the LISA mission, however, EMRI waveforms are required now.

Perturbative methods offer an alternative route towards the solution of the relativistic two-body problem, in which the system's dynamics are described analytically by that of a simplified system modified by corrective terms of increasing order. The most well known of these methods

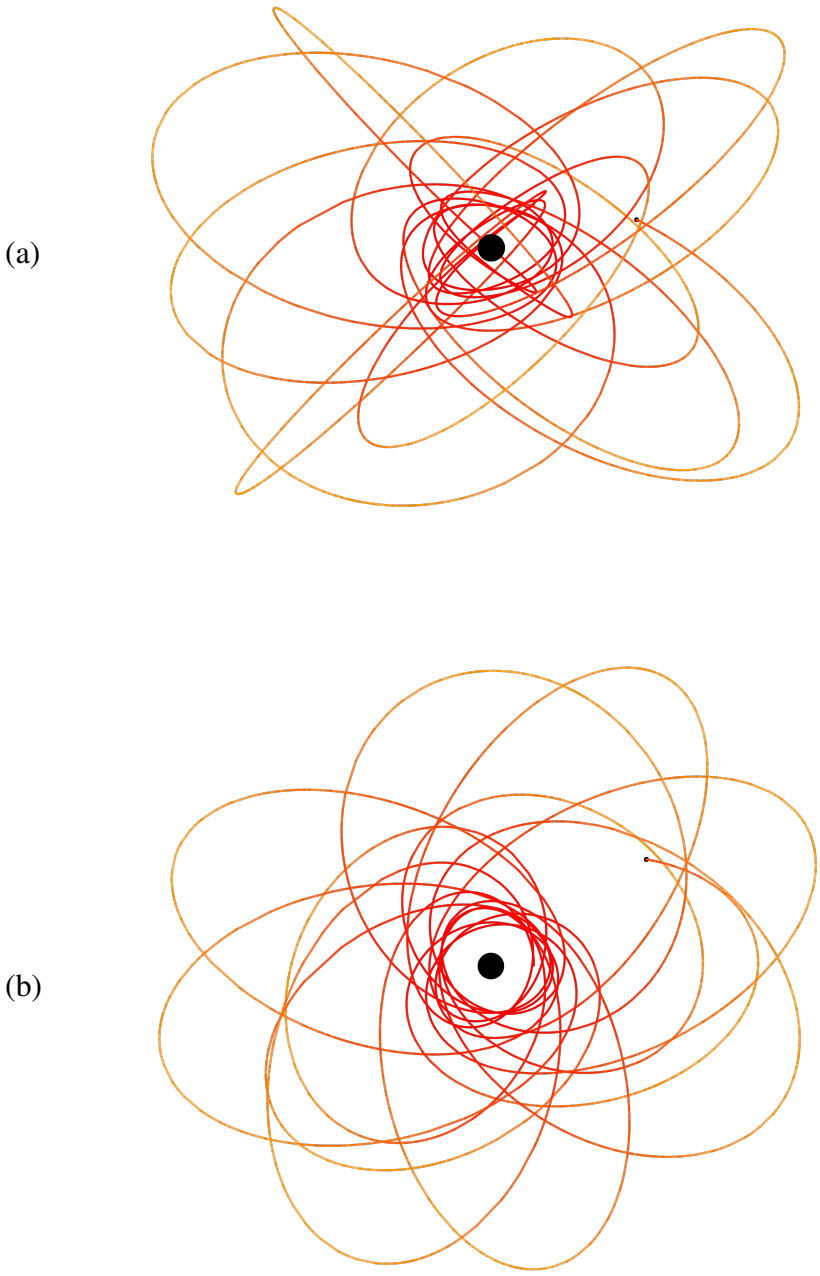


Figure 3.1: A short segment of an inclined, eccentric inspiral of a CO into a rotating MBH, sometimes referred to as a generic Kerr inspiral, viewed (a) edge-on and (b) face-on with respect to the plane of the MBH spin-angular momentum. The MBH is shown to scale with respect to the orbit, whereas the CO is increased in size to render it visible. The line tracing the orbit changes from yellow to red as separation decreases. The strongly-curved spacetime near the MBH manifests as apsidal and Lense–Thirring precession on the orbital timescale, leading to complex dynamical evolution of the system. On longer timescales, the dissipative action of the GSF carries energy and orbital angular momentum from the system, eventually leading to the plunge of the CO into the MBH. Produced using [346].

for modelling gravitational effects is the PN approximation, in which (for EMRI systems) the motion of the system is described by a series expansion in the eccentricity of the inspiral [360]. These corrective terms correspond to deviations from Newton's law of universal gravitation in the weak-field limit where the ratio between the velocities of the two objects are small with respect to the speed of light [361], which has a larger maximum value per orbital cycle as eccentricity increases. This weak-field approximation is wholly inadequate for EMRI systems, in which the motion of the CO typically reaches relativistic velocities (i.e. the maximum orbit velocity is a significant fraction of the speed of light) due to the eccentricity of the inspiral in the strong field of the MBH. However, the expansion produces sensible results for systems of low eccentricity with larger orbital separations, making it a useful tool for understanding the dynamics of these systems in the weak-field limit on longer timescales [3].

Fortunately, there is another approach rooted in black-hole perturbation theory (BHPT) that is well-suited to the modelling of EMRI systems by construction: the GSF formalism, which we examine in more detail below. The following description is a brief examination of the larger picture of GSF waveform modelling: for a more thorough exploration of this field, see [352] for an excellent review.

3.1.1 Inspirals under gravitational self-forces

Our goal is to accurately describe the evolution of an EMRI system, including the emitted gravitational radiation, from an initial point until merger occurs between the MBH and the CO. This task requires us to determine the equations of motion for the CO and integrate them forwards in time.

Before we can consider inspirals, we must outline the basic principles of orbital mechanics in GR. Free from external forces, particles move along geodesics — the equivalent of straight lines for curved spacetimes. Unlike Keplerian mechanics (which describes geodesic orbits in flat spacetime), GR predicts that bound orbits in curved spacetime are never closed due to the effects of precession [362]. For orbits around a Schwarzschild black hole, the system's gravitational potential retains the same underlying symmetries as the Keplerian case, ensuring that the system's total energy E and total angular momentum $\vec{L}$ are conserved. For Kerr black holes, this is not the case; the potential's spherical symmetry is demoted to an axial symmetry, so only the momentum in the direction of the MBH spin (L_z) is conserved. This is the root cause of the Lense–Thirring precession of the orbital plane. In Keplerian mechanics, a third constant of motion (the Laplace–Runge–Lenz vector) exists due to an additional rotational symmetry which is specific to the construction of the two-body problem in Newtonian gravity [363] and therefore does not extend to orbital mechanics in black hole spacetimes. However, the geometry of the Kerr black hole gravitational potential still admits a third constant of motion known as the Carter constant Q. Details regarding its derivation can be found in [364].

The existence of four constants of motion (including the CO mass μ), combined with the

knowledge that the geodesic equations of motion are integrable for Kerr geometry under the Mino time parameterisation [365] (which separates the radial and polar motions of the particle [366]), permits the use of the action–angle variables [367, 368]. Defining the action variables $J_\alpha = \{m, E, L_z, Q\}$ and their corresponding angle variables $\gamma_\alpha = \{\gamma_t, \gamma_r, \gamma_\theta, \gamma_\phi\}$, we write the equations of motion in Boyer–Lindquist coordinates [369] as

$$\dot{J}_\alpha = 0 \tag{3.1}$$

and

$$\dot{\gamma}_\alpha = \omega_\alpha, \tag{3.2}$$

where the overdot refers to a derivative with respect to Mino time, and ω_α is the rate of change of the angle γ_α (and is therefore a frequency) which is dependent on the actions J_α. The explicit form of $\dot{\gamma}_\alpha$ is not reproduced here for brevity; a detailed mathematical prescription can be found in [370] and references therein. These frequencies correspond directly to the fundamental frequencies of geodesic orbits in Kerr spacetime, denoted $\{\Omega_r, \Omega_\theta, \Omega_\phi\}$, which uniquely define an orbit in Boyer–Lindquist coordinates [371, 372]. This completes our framework; for a given initial condition, we can evolve the motion of an orbiting particle forwards in time.

For an orbiting CO, the situation is more complicated. The orbit no longer takes place in the pristine spacetime geometry of a Kerr black hole: the CO itself deforms the spacetime metric to a small degree, which in turn imparts a small acceleration on the CO and modifies its trajectory. As the source of this acceleration (and resulting deviation from a bound geodesic) is the CO itself, this is known as a gravitational self-force (GSF). Dissipative components of the GSF dampen the orbit (which we now refer to as an inspiral) by doing work on the CO and extracting energy, which (in the absence of any other external forces) places the terminus of the inspiral below the innermost stable orbit of the MBH leading to a CBC [373]. This extracted energy is released in the form of GWs. Throughout this inspiral, conservative pieces of the GSF act to modify the rate of apsidal advance, manifesting on the orbital timescale as small adjustments to the shape of the orbit [374]. Dissipative and conservative parts of the GSF must both be considered to model the inspiral phase accurately [375].

The GSF is computed by treating the deformation imparted on the spacetime metric by the CO in a perturbative fashion. This is a viable approach due to the length scale of this deformation (which is defined by μ) with respect to the metric itself (which is defined by M). The resulting series expansion is therefore in terms of the mass ratio $q = \mu/M$, which is folded into our prescription for Kerr geodesic motion order-by-order, to obtain

$$\dot{J}_\alpha = qG^{(1)}_\alpha(\gamma_r, \gamma_\theta) + q^2 G^{(2)}_\alpha(\gamma_r, \gamma_\theta) + \mathcal{O}(q^3) \tag{3.3}$$

and

$$\dot{\gamma}_\alpha = \omega_\alpha + qg^{(1)}_\alpha(\gamma_r, \gamma_\theta) + q^2 g^{(2)}_\alpha(\gamma_r, \gamma_\theta) + \mathcal{O}(q^3). \tag{3.4}$$

The terms $G_\alpha^{(i)}$ and $g_\alpha^{(i)}$ describe the contributions of order q to the actions and angle variables respectively. Radiation-reaction (the results of which are GWs) enter at first order, with successively higher-order terms capturing more of the inherently non-linear behaviour of the GSF and the resulting GW emission [376].

A useful technique for tackling the inherent complexity of Equations (3.3) and (3.4) is to separate the timescales of each forcing term into a time-averaged component and an oscillatory component. The time-averaged component of the first-order term in the action derivatives $\langle G_\alpha^{(1)} \rangle = \{\langle \dot{E} \rangle, \langle \dot{L}_z \rangle, \langle \dot{Q} \rangle\}$, which we refer to as fluxes, captures the inspiral-length behaviour of the system to within $\mathcal{O}(q)$ because the oscillating component (which varies on the orbital timescale) almost completely averages away on this timescale [377]. This gives rise to a simple but accurate prescription for self-forced inspirals known as the adiabatic approximation [376], in which we solve modified versions of Equations (3.3) and (3.4) of the form

$$\dot{J}_\alpha = q \langle G_\alpha^{(1)}(\gamma_r, \gamma_\theta) \rangle \tag{3.5}$$

and

$$\dot{\gamma}_\alpha = \omega_\alpha. \tag{3.6}$$

The fluxes $\langle G_\alpha^{(1)} \rangle\}$ and frequencies $\{\omega_\alpha\}$ are dependent on the values of the actions: as the actions are changing with time, the fluxes and frequencies are time-dependent as well. The explicit form of these fluxes and their derivation is omitted for brevity, but may be found in [345, 378]. They are functions of $\{Z^\infty_{\ell mkn}, Z^H_{\ell mkn}, \omega_{mkn}\}$: the former two are coefficients obtained by solving the Teukolsky equation for the GW amplitude far from the source (∞) and at the MBH event horizon (H) [345], and

$$\omega_{mkn} = m\Omega_\phi + k\Omega_\theta + n\Omega_r. \tag{3.7}$$

The integers (ℓ, m, k, n) are harmonic mode indices, with ℓ corresponding to the orbital angular momentum mode and (m, k, n) the azimuthal, polar and radial modes of the orbital motion respectively. This mode-decomposed description is the result of solving the Teukolsky equation by separating it in a spheroidal harmonic basis, and is useful for expressing EMRI systems as being comprised of harmonics.

The system of ordinary differential equations (ODEs) Equations (3.5) and (3.6) can be integrated numerically to obtain solutions which are functions of time. Physically, this corresponds to modelling the evolving orbit of the CO as a continuous sequence of bound geodesics: at each instance, we compute the GSF for the geodesic tangent to the inspiral and make the corresponding adjustment to the orbital elements. This framework, known as the method of osculating geodesics, is the relativistic analogue of osculating orbits in Keplerian orbital mechanics [379]. The orbital phases $\{\Phi_r, \Phi_\theta, \Phi_\phi\}$, required for the reconstruction of the emitted gravitational radiation, are obtained from the integration of the correspondingly evolving frequencies in Equation (3.6). Errors in this assumption are of $\mathcal{O}(q)$ due to our neglect of long-time variations of the

oscillatory component in $G_\alpha^{(1)}$ [377], typically resulting in the accumulation of several radians of error in the orbital phases [2]. Excluding this oscillatory behaviour is vital to efficient trajectory computations because orbital-timescale variations require ODE integrators to take small steps in time to capture this structure accurately, the cost of which accumulates over tens of thousands of orbital cycles resulting in a framework which is prohibitively expensive for use in data-analysis. Addressing these complications with near-identity-transformations, which exploit the underlying periodicity of the oscillatory behaviour to restore the capacity of ODE integrators to take larger steps in time, is the subject of ongoing work [380, 381].

A useful property of osculating geodesics is that we can perform a change of variables to parameters that are more physically intuitive in their description of orbital geometry. One such reparameterisation is a conversion to the (osculating) orbital elements via $\{\dot{E},\dot{L}_z,\dot{Q}\} \rightarrow \{\dot{p},\dot{e},\dot{x}_I\}$, where p is the semi-latus rectum of the orbit, e is the orbital eccentricity and x_I is the cosine of the orbital inclination with respect to the plane of the MBH spin-angular momentum [370, 371]. The sign of x_I denotes whether the orbit is aligned (prograde, $+$) or anti-aligned (retrograde, $-$) with the spin of the MBH. Throughout this thesis, we will describe orbital trajectories in terms of these geometrical parameters; the details of this transformation are outlined in [2] and references therein.

This EMRI modelling framework cannot be extended all the way to the merger between the two objects. Eventually, the inspiral reaches a point at which the tangent geodesic is no longer a bound orbit, instead intersecting with the event horizon of the MBH [382]. This location in parameter space is where the inspiral trajectory intersects with the separatrix of this ODE system [383]. The separatrix traces a parametric curve in the inspiral parameter space which must be computed numerically for generic orbits of a spinning MBH [384]. After passing through the separatrix, the CO rapidly plunges into the MBH on the orbital timescale, the modelling of which requires a different approach to what is presented here [196]. For a non-spinning MBH, it takes the limiting value of $p_{\rm sep} = 6+2e$; as the MBH spin increases, the separatrix decreases (increases) for prograde (retrograde) inspirals [373]. This limitation of osculating geodesics is acceptable in the EMR regime because the post-separatrix stage takes place over a far shorter timescale than the duration of the inspiral in which ω_{mkn} is in the frequency band of space-based detectors [177].

The adiabatic approximation describes the strong-field inspiral dynamics and the resulting GW emission sufficiently well for use in prospective data analysis studies, but is limited in its application to the analysis of real signals in detector data. For many years it was expected that the orbital phases Φ_α must be accurately modelled to within one radian over the course of the inspiral [368]. However, recent studies examining the relationship between phase error and biases in PE have suggested this may be overly conservative in practice, although this is limited to circular inspirals of non-spinning MBHs where the impact of this phase error is minimal [353]. Such a strict error tolerance demands second-order contributions to the GSF to be modelled,

as their exclusion introduces $\mathcal{O}(1)$ radians of deviation in orbital phase [377]. Furthermore, Equations (3.5) and (3.6) are poor approximations for a special class of geodesics known as orbital resonances, in which two of the fundamental frequencies Ω_α are integer in ratio, because the averaging-away assumption regarding the oscillating component of $G_\alpha^{(i)}$ is no longer valid here [348]. As of the writing of this thesis, fast inspiral modelling codes implemented in efficient waveform models have only been developed under the adiabatic approximation [177]; the extension of these models to incorporate these higher-order effects as corrective terms to the adiabatic treatment is the subject of ongoing work [354, 381]. The adaptation of this scheme to include the plunge of the CO and subsequent merger-ringdown of the binary system is also an area of research interest [385, 386]. This stage of the CBC is of minor importance for EMRI data analysis, because the vast majority of the SNR is accumulated by the end of the inspiral, but it is of greater relevance for the modelling of systems of less extreme mass-ratios [372]. Other higher order effects are also present, including spin on the CO and the growth of the MBH due to the absorption of GWs during the inspiral; these effects are typically expected to be sub-dominant for EMRIs, but work is ongoing to include them in GSF models to extend their domain of applicability to more comparable mass-ratio systems [377, 387, 388].

Equipped with this mathematical machinery, we are now able to evolve an EMRI system forwards in time from initial conditions to plunge. To complete our computational framework for EMRI modelling in the GSF formalism, we now turn to the corresponding prescription for the gravitational radiation at null infinity.

3.1.2 Efficient gravitational waveforms

The emitted gravitational radiation encodes the evolving curvature of the spacetime in the neighbourhood of the inspiral. The GW strain, decomposed into the transverse-traceless plus and cross polarisations as $h = h_+ - ih_\times$, is related in its second derivative to the curvature at a large distance from the source by [389]

$$\psi_4(t) = \frac{1}{2}\ddot{h}(t). \tag{3.8}$$

The Weyl scalar at null infinity $\psi_4(t)$ varies over the course of an inspiral. For the geodesic tangent to the inspiral trajectory at time t', it is the solution of the Teukolsky equation corresponding to that geodesic and may be written as

$$\psi_4(t') = \frac{\mu}{d_\mathrm{L}} \sum_{\ell mkn} Z^\infty_{\ell mkn}(t') S_{\ell m}(\theta, \omega_{mkn}(t')) e^{im\phi} e^{-i\Phi_{mkn}(t')}, \tag{3.9}$$

where $\Phi_{mkn} = m\Phi_\phi + k\Phi_\theta + n\Phi_r$ [345], and non-subscripted θ and ϕ refer to the source-frame polar and azimuthal viewing angles respectively, and d_L is the luminosity distance to the source. The spin-weighted spheroidal harmonics $S_{\ell m}$ of spin weight -2 are the eigenfunctions of the separated Teukolsky equation, and are also obtained from this solution [345]. They vary with

respect to time as the frequencies ω_{mkn} evolve during an inspiral. Combining Equations (3.8) and (3.9), we obtain the gravitational waveform

$$h(t') = \frac{\mu}{d_{\mathrm{L}}} \sum_{\ell mkn} A_{l\ell mkn}(t') S_{\ell m}(\theta, \omega_{mkn}(t')) e^{im\phi} e^{-i\Phi_{mkn}(t')}, \tag{3.10}$$

where we have defined that

$$A_{\ell mkn}(t') = -\frac{2Z^{\infty}_{\ell mkn}}{\omega^2_{mkn}}. \tag{3.11}$$

The summation is defined over $(-\infty, \infty)$ for ℓ, k, n and from $[-\ell, \ell]$ for m. This summation can be interpreted as a sum of harmonic modes, each of amplitude $A_{\ell mkn}(t) S_{\ell mkn}(t, \theta)$ and phase $\Phi_{mkn}(t)$ (where we have re-expressed the frequency-dependence of $S_{\ell m}(\theta, \omega_{mkn}(t))$ as a time-dependence for ease of notation). Each mode can be evaluated separately before summing the results to construct the gravitational waveform [2].

A naive application of Equation (3.10) to EMRI waveform modelling would be prohibitively expensive in practice. Even under the adiabatic approximation (which allows for Equations (3.5) and (3.6) to be efficiently integrated with adaptive time-stepping), derivatives must each still be evaluated thousands of times for a given inspiral. This computational cost is acceptable for the evaluation of the fundamental frequencies Ω_α, which are analytic in form [366], but the same is not true of the fluxes. Simulating waveforms on the timescale of years at reasonable sampling rates also requires Equation (3.10) to be evaluated at $\sim 10^7$ points in time. Solving the Teukolsky equation numerically at each required point in both the trajectory integration and subsequent waveform generation would be extremely computationally costly with existing techniques [343, 390, 391]. However, the fluxes and waveform mode amplitudes are both smoothly-varying over the inspiral parameter space and are functions of only four parameters for standard EMRI systems, making them ideal candidates for interpolation on data grids. This decision requires the range of mode indices (ℓ, m, k, n) to be defined (the grids must be finite, after all), which is typically performed empirically by examining how the result converges with the inclusion of modes of sequentially higher orders and truncating the series when the result is no longer changing by more than a pre-defined threshold [2]. The up-front computational cost for producing these grids is quite high, requiring these quantities to be computed for tens of thousands of parameter sets, but the resulting interpolants are simple (typically consisting of tensor product cubic splines or neural networks) and therefore highly efficient [177, 392]. Further savings are possible due to a symmetry in the mode amplitude space in which

$$A_{\ell -m -k -n} = (-1)^{(\ell+k)} A^*_{\ell mkn}. \tag{3.12}$$

Equation (3.12) halves the number of mode amplitudes which must be directly evaluated by substituting them for an inexpensive complex conjugate operation [393, 394].

The spin-weighted spheroidal harmonics are expensive to compute as well, and introduce

undesirable complications due to their effective time dependence in the form of the fundamental frequencies. The computational expense and time dependence can be converted into further up-front costs by exploiting the property that the spin-weighted spheroidal harmonics can be written as a sum over spin-weighted spherical harmonics of spin weight -2, viz.

$$S_{\ell mkn}(t,\theta)e^{im\phi} = \sum_{j=l_{\min}}^{\infty} b^{j}_{\ell mkn}(t)\,{}_{-2}Y_{\ell m}(\theta,\phi). \tag{3.13}$$

The coefficients $b^{j}_{\ell mkn}$ map between the spheroidal and spherical harmonic bases, and $\ell_{\min} = \max(2,|m|)$. As the $b^{j}_{\ell mkn}$ can be absorbed by the mode amplitudes $A_{\ell mkn}$, we are effectively mapping the mode amplitudes from one basis to another. This is desirable because the spin-weighted spherical harmonics are analytic in form and are constant throughout the inspiral, whereas the amplitude terms are already computed numerically and time-varying. The details of this procedure are derived and explored in [395].

Further computational gains can be made by capitalising on the sparse nature of the solutions to Equations (3.5) and (3.6) returned by adaptive ODE integrators. The trajectory and fundamental phases will be evaluated at ~ 100 points in time, spaced non-uniformly to better capture their rapid variation towards the end of the inspiral [177]. For waveform generation, these orbital parameters must first be interpolated to the sampling rate of the data stream. However, we can exploit this property by instead evaluating $A_{\ell mkn}$ at these few points in time and interpolating the result, saving $\mathcal{O}(10^5)$ interpolations per mode amplitude for year-length waveforms. The number of interpolations required can be reduced by a factor of two with the symmetry described in Equation (3.12), which is best applied to the interpolated mode amplitudes to maximise computational savings. This representation of the waveform as a set of sparsely-sampled mode amplitude and phase trajectories can be interpolated very efficiently because this is an embarrassingly parallel problem [396]. Recent work has taken advantage of this by using GPU vectorisation techniques to evaluate thousands of cubic splines in parallel, as demonstrated in the open-source software package FEW [1, 177].

We complete our description of waveform generation by outlining a simple yet effective technique for directly reducing the computational cost of Equation (3.10) at the cost of accuracy, hinted at by the truncation of the mode summation employed in the construction of the data grids. That truncation is essentially a "first cut" in the mode space, discarding modes that are not expected to contribute significantly to the final waveform. This idea can be taken further to identify on a waveform-specific basis if the mode count can be further reduced by examining the mode amplitude spectrum $A_{\ell mkn}(t)$, accelerating the mode interpolation and summation stage (which is typically a significant fraction of the waveform's computational cost). In other words, we are able to reduce the cost of the summation in Equation (3.10) by carefully picking and removing small terms in the series. This was first applied in [177] for FEW, in which the authors perform a mode-sorting operation and truncate the resulting series when the difference

between a cumulative and total sum over the power contribution of each mode falls below a pre-defined threshold ε. This final step typically reduces mode counts by a factor of a few for eccentric inspirals to more than an order of magnitude for quasi-circular systems for $\varepsilon = 10^{-5}$, with further reductions possible by raising ε. Evidence suggests that aggressive mode thresholds as high as $\varepsilon = 10^{-2}$ may be sufficient for data analysis of EMRI systems in recovery of the intrinsic parameters [177], but (as we shall show later in Section 3.4) the same is not necessarily true for extrinsic parameters.

In practice, a computational framework for computing Equation (3.10) has not yet been completed for generic Kerr systems consisting of eccentric, inclined inspirals into spinning MBHs. This is primarily due to the complexity inherent to the efficient and accurate construction of interpolation schemes for four-dimensional data, but no hard limitations are present, and this work is expected to be completed in the near future [177]. As preliminary steps, work has focused on scenarios of reduced complexity, which we briefly outline below.

3.1.3 Schwarzschild eccentric inspirals

The simplest case we will describe here consists of eccentric inspirals into non-spinning MBHs, as implemented in the FEW software. The resulting spherical symmetry of the MBH spacetime metric means that all orbits are equatorial with $x_I = 1$. As this implies that there is no precession of the orbital plane, the fundamental phase Φ_θ and mode index k which describe this motion may be ignored as all $k \neq 0$ terms in Equation (3.10) are zero. The basis mapping described by Equation (3.13) is trivial in this limit, with the summation being equal to ${}_{-2}Y_{\ell m}(\theta,\phi)$ everywhere. Bringing together these changes, we can rewrite Equation (3.10) as

$$h(t') = \frac{\mu}{d_{\rm L}} \sum_{\ell mn} A_{\ell mn}(t') Y_{\ell m}(\theta,\phi) e^{-i\Phi_{mn}(t')}, \tag{3.14}$$

where here (and in what follows) we have omitted the (-2) subscript of ${}_{-2}Y_{\ell m}(\theta,\phi)$ for brevity. The elimination of the k modes reduces the number of modes that must be considered in this summation by more than an order of magnitude. The mode amplitude spectrum is more sharply peaked as a decreases, further reducing mode counts at $a = 0$ with respect to the general case. Given this behaviour, the FEW implementation of Equation (3.14) truncates the mode summation as follows: $\ell \in [2,10]$, $m \in [-\ell,\ell]$ and $n \in [-30,30]$. The total number of modes considered (from counting all mode combinations) is therefore 7137.

For equatorial orbits, the Carter constant Q and its derivative are both zero, reducing Equations (3.5) and (3.6) to a system of four ODEs which can be easily solved. The remaining two fluxes can be easily evaluated with two-dimensional interpolation on a grid over (p,e), and the fundamental frequencies are particularly simple in form in the Schwarzschild limit [397]. Further computational savings are made by the simple analytic form of the separatrix of $p_{\rm sep} = 6+2e$ for Schwarzschild inspirals. Waveforms in the Schwarzschild eccentric limit can therefore

be computed extremely rapidly: FEW is capable of sub-second waveform generation timings for year-duration waveforms [177]. Throughout this chapter, this waveform model will be referred to as the SchwarzEcc model.

3.1.4 Kerr equatorial eccentric inspirals

Unfortunately, despite their low computational cost it is unlikely that Schwarzschild eccentric EMRI waveforms will be particularly applicable beyond their use as a playground in which EMRI data analysis strategies can be formulated and tested. This is because in reality we expect MBHs to be spinning rapidly [197, 398], for which the Schwarzschild waveforms are a poor approximation. For prograde inspirals, the CO completes more orbital cycles deep within the strong-field region of the MBH spacetime. The waveform SNR increases significantly as a result, so to properly model interesting statistical properties such as detection rates and measurement precision MBH spin cannot be neglected [190]. Fortunately, an extension to Kerr equatorial inspirals (where $x_I = 1$ is still maintained, but $a \neq 0$) from the Schwarzschild eccentric case is fairly straightforward in principle. The mapping described by Equation (3.13) is no longer trivial and must be performed numerically, which increases up-front costs but does not substantially alter the computational cost of the resulting waveform model. The waveform equation is therefore unchanged in form from Equation (3.14), provided we map $A_{\ell mkn}$ from the spheroidal to the spherical harmonic basis in our data grids.

The inclusion of spin increases the dimensionality of the parameter space over which the fluxes, fundamental frequencies and waveform mode amplitudes must be evaluated to three. As the fluxes must be interpolated to a relative accuracy of $\sim 1/q$ for sub-radian biases in the orbital phases [374], a precise interpolant must be used; a tensor product spline is a reasonable choice, but Chebyshev spectral interpolation has also seen recent success [399]. The interpolation error requirements for the waveform mode amplitudes are less strict than those of the fluxes, as errors in the orbital phase lead to more significant biases in PE than mode amplitude errors of corresponding size. Additionally, a shift in the evolution of the geometric parameters of the orbit will couple to the mode amplitudes (as they will have been estimated at the wrong point in parameter space), further accentuating the resulting systematic biases. Further work is ongoing to examine if a densely spaced configuration of bicubic grids in (p, e) with linear interpolation over a will provide sufficiently precise interpolation without the need for excessively large numbers of grid points.

As the separatrix for Kerr inspirals must be obtained numerically, trajectory integration is somewhat more expensive than the Schwarzschild case, although costs are mitigated due to the simpler form of the separatrix polynomial for equatorial inspirals [373].

For prograde orbits, we expect that the mode selection procedure will retain more modes as spin increases. To account for this effect at high spin, the range of n modes included in the data grids is increased to $n \in [-50, 50]$, raising the total number of modes in the data grid

to 11817 (an increase of roughly 65% over the SchwarzEcc model). The larger typical mode counts combined with the greater computational cost of the trajectory integration means that prograde Kerr equatorial eccentric waveforms are inherently more computationally costly than their Schwarzschild counterparts of zero spin, but overall costs are still sufficiently low for them to be applied in data analysis studies.

3.1.5 Approximate waveform models for generic Kerr inspirals

Prior to the development of fast GSF waveform models, a number of approximants were constructed to provide a means to explore the detection, analysis and astrophysical potential of EMRIs. The earliest, so-called "kludge" waveform models, offer vastly simplified descriptions of EMRI waveforms, achieving low computational costs at the cost of accuracy. This is illustrated well by the first of the kludge models, known as the analytic kludge [341]: in this approach, the orbital dynamics are taken to be Keplerian, with relativistic precession effects approximated by first-order PN formulae applied semi-physically to the orbital motion. The gravitational radiation is described solely by the quadrupole approximation for a Keplerian orbit [16]. The analytic kludge is a good start to the problem of EMRI modelling, but is now only useful if one wishes to estimate the SNR of a given system to within an order of magnitude.

Over the last two decades, kludge models have undergone significant development [313, 400]. The latest incarnation (which is implemented in FEW) is the augmented analytic kludge (AAK), which has improved consistency with GSF waveforms with respect to the original analytic kludge [401]. This implementation uses fifth-order PN approximations of the GSF fluxes $\{\dot{p}, \dot{e}, \dot{Y}\}$, but uses exact formulae for the fundamental frequencies Ω_α [360]. The parameter Y differs slightly from the standard inclination parameter x_I described earlier, but the two may be converted between numerically through the integrals of motion [177]. This waveform model is not accurate enough for the analysis of real EMRI signals due to the poor accuracy of the trajectory (as a result of extending PN fluxes beyond their domain of validity [3]) and severely limited treatment of the amplitudes (which still relies on the quadrupole approximation [313]). However, it has still seen extensive use because, unlike current GSF waveform models, it is both applicable to generic Kerr inspirals and rapid enough for use in data analysis studies. In this thesis, the AAK is applied to population modelling (see Chapter 4); it has seen recent application in the forecasting of the observability of environmental effects [186, 213], lensing parameter recovery [220] and astrophysical background spectra [226, 227]. Its widespread use in EMRI data analysis studies will continue for a number of years until a fully-relativistic generic Kerr waveform model is constructed.

An extension of the PN trajectory to include PN approximants of equal order for the waveform mode amplitudes and spin-weighted spheroidal harmonics has recently been initiated [3] and is in the process of being implemented into the FEW framework. This waveform model, which we denote as PN5, is currently in a developmental state. In qualitative terms, it is far

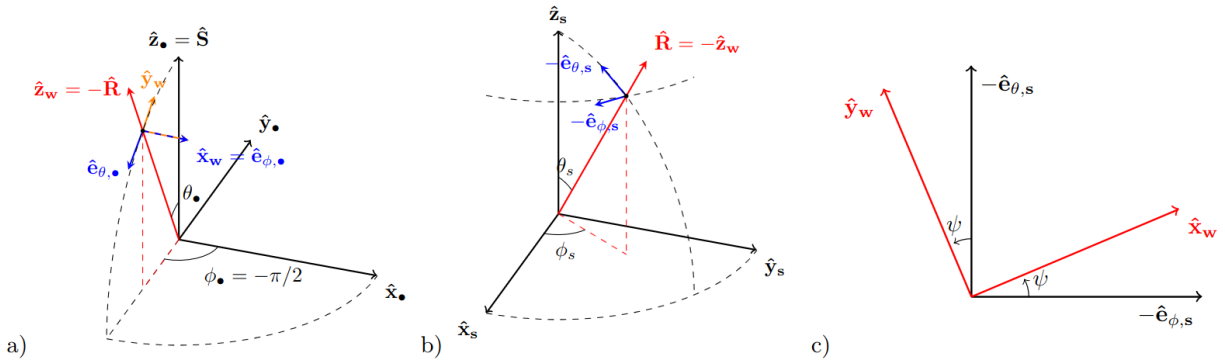


Figure 3.2: Frame transformation procedure for the EMRI waveforms defined in this chapter. The orientation of the source frame (a) is defined such that the spin-angular momentum vector $\hat{S}$ points in the $\hat{z}$ direction, and $\phi_\bullet$ is always offset from the sky-position vector $\hat{R}$ by $\pi/2$ (see Equations (3.15) to (3.17)). The position of the source in the reference frame of the SSB (b) is then defined by the sky position angles (θ_S, ϕ_S). To complete the frame transformation, the waveform is rotated according to the polarisation angle ψ (Equations (3.18) and (3.19)) as shown in panel (c). Adapted from [177].

more representative of the challenges inherent to EMRI waveform modelling in generic Kerr than the AAK, specifically with respect to mode amplitude generation and mode selection. It is therefore a useful analogue with which techniques for addressing these challenges can be validated, as we will examine later in this chapter. The mode index ranges for this model are more complex because each mode amplitude is described analytically rather than on a pre-computed grid: $\ell \in [2, 10]$, $m \in [-\ell, \ell]$, but the ranges for k and n vary as a function of ℓ. In general, they increase from ($k \in [-8, 8]$, $n \in [-5, 11]$) to ($k \in [-20, 20]$, $n \in [-9, 11]$) as ℓ increases from 2 to 10. Counting each set of mode indices in this irregular grid, the total number of modes in the PN5 waveform model is 38042.

3.1.6 Detector frame transformation

So far, we have only considered the parameters of an EMRI system in the reference frame of the source. To complete this prescription, we must project the transverse-traceless polarisations of the waveform onto the reference frame of the detector. This procedure can be separated into two stages. The first stage is a reference frame transformation from the source frame to the SSB. The position of the source on the sky is described by the ecliptic longitude and the complement of the ecliptic latitude, denoted ϕ_S and θ_S respectively. The orientation of the source with respect to the SSB frame is described with respect to the spin-angular momentum vector of the MBH, decomposed into an azimuthal component ϕ_K and a polar component θ_K. The luminosity distance d_L is unchanged under this transformation. The projection of the waveform as a function of these four angles is demonstrated in Figure 3.2 and is performed as follows.

Following [177], we first define the sky-position vector

$$\hat{R} = (\sin\theta_S \cos\phi_S, \sin\theta_S \sin\phi_S, \cos\theta_S), \tag{3.15}$$

and the spin-angular momentum vector

$$\hat{S} = (\sin\theta_K \cos\phi_K, \sin\theta_K \sin\phi_K, \cos\theta_K). \tag{3.16}$$

Second, we define the Cartesian basis vectors $(\hat{x}_\bullet, \hat{y}_\bullet, \hat{z}_\bullet)$ of the source frame as

$$(\hat{x}_\bullet, \hat{y}_\bullet, \hat{z}_\bullet) = \left(\frac{\hat{R}\times\hat{S}}{\sqrt{1-(\hat{S}\cdot\hat{R})^2}}, \frac{\hat{R}-(\hat{R}\cdot\hat{S})\hat{S}}{\sqrt{1-(\hat{S}\cdot\hat{R})^2}}, \hat{S} \right). \tag{3.17}$$

Third, we write the polarisation angle of the incoming GWs as

$$\psi = \frac{\cos\theta_S \sin\theta_K \cos(\phi_S-\phi_K) - \sin\theta_S \cos\theta_K}{\sin\theta_K \sin(\phi_S-\phi_K)}. \tag{3.18}$$

Last, the rotation of the transverse-traceless waveform polarisations to the SSB frame is written as

$$h_{+,\mathrm{SSB}} = \cos(2\psi)h_+ - \sin(2\psi)h_\times \tag{3.19}$$

and

$$h_{\times,\mathrm{SSB}} = \sin(2\psi)h_+ + \cos(2\psi)h_\times. \tag{3.20}$$

The source-frame viewing angles are related to this description by $\theta_\bullet \equiv \theta = \arccos\left(-\hat{R}\cdot\hat{S}\right)$ and $\phi_\bullet \equiv \phi = -\pi/2$ [177].

The second stage is to project the SSB-frame waveform onto the detector via the response function. The LISA response function is derived by projecting the GWs onto each laser link in the constellation, before constructing TDI combinations (see Chapter 2) which encode the GW signal in the form it will appear in processed LISA data [65, 306, 316]. It is a computationally expensive operation and has only been derived for generic signals in the time domain [330] (however, see [310] for a frequency-domain response function valid for short-duration signals such as MBH binaries). In this chapter, we focus on the generation of EMRI waveforms; the application of the response function is a secondary step which is not unique to these signals, so we do not discuss it further here.

3.1.7 Frequency-domain waveform generation

The mode summation described by Equation (3.10) is defined in the time domain. However, the form of this equation implies that a frequency-domain description may be more appropriate: each harmonic mode must be evaluated at the sampling rate of the data stream in the time

domain, but many are more compactly represented in the frequency domain by a small number of frequency bins. It is possible to derive a frequency-domain representation of Equation (3.10) by systematically applying the stationary phase approximation to each harmonic mode [402] and summing over the results. The specifics of this derivation and the resulting frequency-domain representation of Equation (3.10) are not reproduced here: they can be found in [312].

Performing the time-domain summation in the frequency domain can be more performant depending on the parameters of the EMRI in question [312]. A particularly interesting application of frequency-domain waveform generation is the evaluation of the waveform strain on a sparse set of frequency bins. As EMRI signals are rich in information, it is possible that the number of frequency bins required to represent the signal for analysis purposes may be significantly smaller than what is obtained from the FFT of the time-domain data. The computational cost of the mode summation for these downsampled waveforms is significantly lower than their time-domain counterpart, shifting the computational burden to the prior waveform generation steps. As this mode summation step is highly computationally costly when performed on a single CPU core, eliminating this bottleneck enables CPU-based waveform generation, which is desirable as it is more easily distributed over wide pools of computational resources. Carefully constructed GPU summation kernels capable of handling many sets of waveform modes simultaneously will also benefit from this reduced computational cost.

Frequency-domain EMRI waveforms are currently limited in their application to the analysis of detector data due to the lack of accurate frequency-domain detector response functions; existing implementations are only suitable for use in small data segments [216] or for signals that vary slowly with time [291]. This may be addressed in future work by combining existing techniques in order to generate EMRI waveforms in the time-frequency or wavelet domains. For now, some systematics of EMRI PE and data analysis may still be investigated with frequency-domain waveforms, as we will explore in Section 3.4.

3.2 Mode selection for rapid waveform generation

The various stages of EMRI waveform generation described in Section 3.1, written in the order in which they are performed in FEW, are

1. Given initial fundamental phases $\Phi_{\alpha,0}$ and initial orbital conditions $(p_0, e_0, x_{I,0})$, integrate Equations (3.5) and (3.6) until either the desired waveform duration T or the separatrix is reached to obtain a sparse trajectory with respect to time.

2. Evaluate the mode amplitudes $A_{\ell mkn}$ for each sparse point in the trajectory.

3. Perform a sorting and cumulative summation of these mode amplitudes (multiplied by the spin-weighted spherical harmonics $Y_{\ell m}$ to account for the viewing angle), truncating the

summation when the residual power is less than the threshold ε. Any modes that are not retained for at least one point in the sparse trajectory (to ensure continuity) are discarded.

4. The remaining modes are interpolated to the sampling rate of the data stream and summed to produce the gravitational waveform.

There is a clear inefficiency in the above process which leads to wasted computational effort: the discarding of mode amplitudes. For time-domain EMRI waveform modelling, mode amplitude generation and selection are of similar cost to the summation step, but frequency-domain waveforms (particularly when evaluated on sparser grids of frequency bins) are heavily bottlenecked by this stage of the computation (see Section 3.3). This issue could be addressed by re-ordering the second and third stages: if the required mode content (i.e. the modes retained after the mode selection stage) of a given EMRI waveform could somehow be obtained without the need for a set of mode amplitudes, only these mode amplitudes would need to be computed. We refer to this form of mode selection, performed in advance of mode amplitude generation, as mode prediction.

Accurately predicting which modes should be kept is challenging due to the inherent complexity of the mode selection process. Whether a given mode is retained or discarded is a complicated function of the initial parameters of the system, dependent on the evolving mode amplitude spectrum (which itself depends on the evolving orbital parameters of the system), the viewing angle of the observer (via the spin-weighted spheroidal harmonics) and the threshold ε with respect to which the mode summation series is truncated. Nonetheless, this complicated function is deterministic, and we can already compute its inverse: given a set of initial parameters, we can define the set of modes which pass the selection threshold (the output of step 3. in the sequence above). We therefore turn to ML methods, which provide the machinery required to model the mode selection process and are well-suited to problems in which computational cost is of particular importance.

3.2.1 Mode prediction as a classification problem

To understand how we can model the mode selection process, we must first explore the nature of the problem. For a given waveform of parameters $\vec{\theta}$ and duration T, mode selection with respect to a threshold ε returns a vector of labels η_i of length N_{modes}, where the subscript i enumerates over all combinations of the mode indices (ℓ, m, k, n) included in the data grid. Collectively, $\{\vec{\theta}, T, \varepsilon\}$ are referred to as the features which correspond to the labels η_i. These labels indicate which modes were retained after the selection process, with retained modes being indicated by a 1 and discarded modes by a 0. Mode prediction is therefore a classification problem: given a set of waveform parameters and a mode power threshold, our goal is to construct a statistical model that predicts the probability of each mode being retained given a set of input parameters. It is a specific type of classification problem known as a multi-label classification problem,

because a given set of features maps to multiple labels (i.e., η_i is a vector quantity, not a scalar one). Multi-label classification problems are difficult to solve optimally, fully accounting for the correlations between the probabilities of retaining each mode. This can be addressed by expanding the problem into the form of a multi-class classification problem (which may be tractably solved) by listing every permutation of η_i and assigning it a new label. However, the total number of permutations is $2^{N_{\text{modes}}}$: as N_{modes} grows, defining a classification problem by this enumeration quickly becomes computationally intractable if many label combinations are possible.

To make progress, we can take the simplifying assumption of binary relevance, in which the classification of each mode label is treated as being independent of the others [403]. This discards the information encoded in the correlation between labels, but yields a computationally tractable process by reducing the multi-label classification problem into N_{modes} binary classification problems. Binary classification is the simplest form of classification problem and can be efficiently solved with a variety of techniques depending on the context of the problem. In the case of mode prediction for EMRI waveform generation, the accuracy of the method we choose is important, but of foremost importance is computational cost. A suitable approach must be capable of processing all N_{modes} binary classification problems in parallel on the timescale of a millisecond. Additionally, for use in waveform generation this performance must be achieved for a single mode prediction example: the computational cost cannot be reduced by evaluating thousands of sets of features at once. Many approaches commonly employed for classification problems are too expensive to meet this requirement. Decision trees, which are an algorithm consisting of a sequence of tuneable if-else statements, are too expensive for the complexity of this problem on a CPU and are not readily vectorisable on a GPU due to the irregular shapes of each tree structure. The dimensionality of the feature space (which in this case is on the order of 10) also renders more flexible approaches such as Gaussian processes too expensive to be practical for our purposes.

Deep neural networks are well-suited to solving many classification or regression problems in parallel [264]. By constructing a MLP with N_{modes} in its output layer, we can model and solve the binary classification problems of every mode in an inherently parallelised format. They are highly computationally efficient and can be readily deployed on accelerator hardware if faster performance is required. Their primary drawback is the need for a large training dataset, which is not a significant limitation for this problem because we can generate data at a relatively low computational cost by performing the mode selection process manually. We now quantitatively describe how this training data is distributed in parameter space.

3.2.2 Training datasets

Training data is constructed by performing the mode selection process in the same manner as it is used to construct EMRI waveforms. This takes significantly less time than what is required to

Parameter	Range (SchwarzEcc)	Range (KerrEqEcc)	Range (PN5)
$\log(M/M_\odot)$	5 to 7		
$\mu/M_\odot$	1 to 100		
a	-	0 to 0.99	
$y(p_0)$	Varies (see text)		
e_0	0.1 to 0.7		0.1 to 0.4
Y_0	-	-	$\pm$(0.1 to 0.99)*
θ	0 to π		
ϕ	0 to 2π		
T	0.01 to 4		
$\log\varepsilon$	-6 to -1		

Table 3.1: Training data bounds which define the neural network's domain of validity for each model. Training data is distributed uniformly within these bounds. The parameter bounds for p_0 are complicated because they are parameterised in terms of the coordinates of flux data grids: they are explained further in Section 3.2.2. *To avoid numerical issues in the PN5 adiabatic fluxes for near-polar orbits, the region (-0.1 to 0.1) is excluded from the domain of the model.

fully generate waveforms because the costly mode summation procedure is not required. Inputs are a set of waveform parameters $\vec{\theta}$, data length T and a mode selection threshold ε. The number of parameters in $\vec{\theta}$ depends on the waveform model. For the PN5 model, our most general case, $\vec{\theta} = \{M, \mu, a, p_0, e_0, Y_0, \theta, \phi\}$. Stepping to the KerrEqEcc model, Y_0 is no longer included as a parameter. Typically, Y_0 takes on negative or positive values to represent retrograde or prograde orbits; to ensure the training data is smoothly distributed and to eliminate Y_0 entirely, we assign this sign to a instead for this model. A further reduction to $\vec{\theta}$ is applied for the SchwarzEcc model in which a is removed, as inspirals all have $a = 0$ in this regime.

The parameter space ranges over which the training data is distributed for each model are shown in Table 3.1. To account for the increase in physical complexity as we step from the SchwarzEcc to PN5 models, we slightly increase the number of training data points in tandem. In total, we generate 10^6 , 2×10^6 and 4×10^6 data points for the SchwarzEcc, KerrEqEcc and PN5 models respectively, using 95% of the dataset during training and 5% for validation to monitor overfitting and halt training when appropriate. We employ a number of reparameterisations to increase the density of training data points in regions where waveform mode content is larger and more complicated. By taking the base-10 logarithm of M and ε, good coverage is maintained at lower masses and lower mode thresholds respectively, which is important as the mode content of the waveform changes more rapidly in this region of parameter space. We choose to parameterise p_0 identically to how it is expressed in the data grid structure, which is written as a function of the separatrix $p_{\rm sep}$ and varies between different models. We denote this reparameterisation in the general form $y(p_0)$. For the SchwarzEcc model, it takes the form

$$y_{\rm SE}(p_0, e_0) = \ln(p_0 - p_{\rm sep}(e_0) + 3.9), \tag{3.21}$$

with $y_{\mathrm{SE}} \in [1.37, 3.75]$. In the KerrEqEcc model, it is written as

$$y_{\mathrm{KEE}}(p_0, e_0, x_I) = \ln\left(\frac{p_0 - p_{\mathrm{sep}}(a, e_0, x_I) + 3.95}{4}\right), \tag{3.22}$$

with $y_{\mathrm{KEE}} \in [0, 1.365]$. The PN5 model is not based on data grids, but the KerrEqEcc reparameterisation for p_0 is still a convenient way to distribute points near the separatrix in this regime, so we employ the same reparameterisation (and limits) for p_0 in the PN5 model as in the KerrEqEcc model, with Y_0 substituted for x_I. The PN5 model can be extended to arbitrarily high p_0, but modelling the strong-field behaviour of the mode spectrum is by far the most complicated part of the parameter space and so we focus on this region as a proof-of-principle.

The parameter space bounds in Table 3.1 push the PN5 model beyond its domain of validity. In particular, the model has been shown to be inaccurate for $e_0 \gtrsim 0.3$ and $p_0 \lesssim 6$ [3]. For the purposes of mode selection and prediction however, this inaccurate waveform model is sufficient for demonstrating that our approach will be practical even for accurate generic Kerr waveform models because the inherent complexity of the problem is similar in both cases. In other words, we are only interested in whether the predicted mode content is sufficiently accurate for the resulting waveforms to be useful in data analysis; at no point do we require the waveforms themselves to be accurate enough for use in the analysis of real signals.

In principle, ε could be fixed to a specific value to improve the predictive accuracy of the MLP at the cost of flexibility. We find that the neural networks are capable of learning how the mode selection process depends on this threshold sufficiently well for use in waveform modelling, and therefore choose to keep it as a free parameter for generality. It is likely that adjustment of the mode selection threshold will form a component of EMRI data analysis as the PE results are iteratively refined from coarse to fine levels of accuracy; it is therefore important that our approach is able to incorporate such an adjustment.

Generating the training dataset for the SchwarzEcc (PN5) model takes roughly 2 hr (12 hr) on an Nvidia A100 GPU. After the raw training data has been constructed for each model, some pre-processing steps are applied. Some trajectories were numerically unstable due to starting extremely close to the separatrix and crossing into unphysical regions of parameter space (e.g., negative eccentricity), which resulted in not-a-number trajectory outputs. These erroneous trajectories (typically less than 10 points) are removed from the dataset. The mode retention of the training data points is then examined to identify which modes were not retained for any data point. These modes are of sufficiently low amplitude that, even for $\varepsilon = 10^{-6}$, they never pass the mode selection threshold. As these modes do not meaningfully contribute to waveform generation (and more importantly, cannot be learned by the neural networks due to a lack of examples in which they were retained), they are removed from the training dataset. This is effectively a pre-computed mask over mode indices, similar to what is discussed in the conclusions of [177]; however, this pre-computed mask is conditioned on the bounds of the

training data. Constructing this mode-index mask eliminates roughly two thirds of the modes for all three waveform models, which is vital to the performance of the neural networks in terms of both speed and accuracy. The number of modes remaining in the training data is indicated for each waveform model in the form of the output layer size of the neural networks, shown in Table 3.2.

With these pre-processed training datasets in hand, it is now possible to fit three neural networks that can accurately and rapidly predict the mode content of EMRI waveforms. This procedure for their training and implementation are described in the following subsection.

3.2.3 Neural network structure and training

To construct a MLP in the form of a statistical model, we must first define how the MLP output (which is a vector of real numbers) is converted to a vector of probabilities defined on the range $[0,1]$. Any unique mapping between these domains is fit for this purpose: a common choice is the logistic function (commonly referred to as the sigmoid function),

$$\sigma(x) = \frac{1}{1+e^{-x}}, \tag{3.23}$$

due to its desirable mathematical properties in statistical optimisation and ease of computation [277, 284]. Adopting the logistic function for our output mapping, the probability predicted by the MLP of retaining a given mode i is therefore written as

$$q_i = \sigma(f_{\theta,i}), \tag{3.24}$$

where $f_{\theta,i} \equiv f_\theta(\vec{\theta}, T, \varepsilon)$ represents the functional mapping of the MLP from the input parameters to a vector of real numbers. These predictions may be compared to the actual results of the mode selection process, which is a vector of probabilities p_i. The elements of p_i are either 0 or 1 because mode selection is deterministic. The consistency between the predicted and true probabilities (i.e., a measure of the difference between the probability distributions described by p_i and q_i) is represented by the binary cross-entropy

$$H(p_i, q_i) = -p_i \ln q_i - (1-p_i)\ln(1-q_i), \tag{3.25}$$

which takes the minimal value of 0 for $p_i = q_i$ [404]. Equation (3.25) is a suitable choice of loss function to be optimised during the training of the MLP: it is differentiable, inexpensive and has a global minimum which matches our objective.

We employ the same general network structure, with different numbers of layers and neurons based on the waveform model which the mode predictor will be used in, because N_{modes} varies by an order of magnitude between the SchwarzEcc and PN5 models (with the KerrEqEcc model lying in between the two). All three networks employ the sigmoid linear unit (SiLU) activation

Waveform model	Neuron and layer structure	Number of parameters	GPU wall-time [ms]	CPU wall-time [ms]
SchwarzEcc	[16,64,256,1024,1024, 1024,1024]	4.7×10^6	0.26	0.43
KerrEqEcc	[16,64,256,1024,2048, 2048,2048]	1.6×10^7	0.26	1.96
PN5	[16,64,256,1024,2048, 2048,2048,2048]	3.7×10^7	0.31	4.52

Table 3.2: Structure, number of parameters and execution wall time for the MLP neural networks developed to perform mode prediction for each of the three waveform models considered. The neuron structure describes the number of neurons in each hidden layer of the MLP.

function [276]. These network parameters are shown for each of the three MLPs in Table 3.2, along with total parameter count and timing statistics for each network on both CPU and GPU hardware.

Each network is trained for 500 epochs, at which point the validation loss plateaus and further training is observed to result in overfitting. This training operation is relatively quick, taking 0.5 hr (4 hr) for the SchwarzEcc (PN5) model. This speed is obtained due to two choices made during network training. The first is to exploit the compact nature of the training data. Although $\sim 10^6$ data points each consisting of $\sim 10^4$ mode labels is extremely large when expressed in the floating-point format required for backpropagation algorithms, by storing it as unsigned 8-bit integers (which are each a byte in size) we reduce the memory footprint of the data by a factor of 4. This enables transfers between CPU and GPU RAM within the training loop to be completely eliminated, greatly reducing the wall-time per training iteration. The second choice is to use automatic mixed-precision (AMP) methods to optimise the performance of the network (in both the evaluation and subsequent backpropagation steps), which reduces the numerical precision for network evaluations to half-precision (16-bit floating point) [281]. This approach is highly effective in the training of MLPs because 32-bit floating-point precision is not necessary for the optimisation of these networks [405], and GPUs with hardware optimised for half-precision operations have been designed to exploit this fact (such as the Nvidia A100, which was used here).

Rapid data generation and subsequent training is vital to the incorporation of mode prediction with neural networks into EMRI waveform models. An important component of GW data analysis is to introduce additional parameters which modify the waveform model (for instance, to represent deviations from GR or environmental effects) and then attempt to constrain these parameters given some detector data. As these parametric modifications will alter the mode selection behaviour of the waveform, a new MLP must be trained for each model. The practicality of our approach is therefore strongly dependent on the computational cost of this process. This is not necessary until the developmental stage of the modified waveform model is com-

plete, as standard mode selection is sufficient for testing purposes on the scale of thousands of waveforms.

Given a set of trained mode predictor networks, we now detail how they are applied in practice and their performance is quantified.

3.2.4 Network validation

Once the network has been trained, the mode content of any EMRI waveform with parameters inside the convex hull described by the ranges in Table 3.1 can be rapidly predicted with the neural network. Feedforward networks such as MLPs typically perform poorly at extrapolation [406], so mode predictions made outside these bounds are unlikely to be accurate. The probabilities q_i must be converted to a binary label to form the mode retention vector η_i and identify which modes should be retained. This conversion is performed with respect to a numerical threshold ζ, viz.

$$\eta_i = \mathcal{H}(q_i - \zeta), \tag{3.26}$$

where $\mathcal{H}(x)$ is the Heaviside step function. In practice, the $\sigma(x)$ evaluation of the output layer is not performed for computational efficiency: instead, we may replace ζ such that

$$\eta_i = \mathcal{H}(f_{\theta,i} - \sigma^{-1}(\zeta)), \tag{3.27}$$

where $\sigma^{-1}(x)$ is the inverse of the sigmoid function (also known as the logit function), written as

$$\sigma^{-1}(x) = \ln\left(\frac{x}{1-x}\right). \tag{3.28}$$

After this step function has been applied, the prediction vector of mode labels η_i is passed forwards to the summation stage of waveform generation and the waveform is constructed.

Due to the form of Equation (3.25), the network has been implicitly optimised with respect to the value $\zeta = 0.5$ during training. Adjusting ζ from this point allows for the mode prediction algorithm to be tuned, with a lower ζ corresponding to a more conservative prediction which retains more modes and vice versa. This can be seen by examining the true positive rate (TPR) of the network predictions varies as a function of ζ. The TPR for a binary estimator is defined in terms of the number of true positives (TP) and false negatives (FN) as [407]

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} - \mathrm{FN}}. \tag{3.29}$$

In other words, the TPR expresses how often the mode predictor correctly chooses to retain a given mode. CDFs of the TPR over mode index, evaluated over all validation data, are shown for each waveform model in Figure 3.3. At $\zeta = 0.5$, at least 80% of modes are retained correctly more than 80% of the time for all three models. This accuracy further improves as ζ decreases

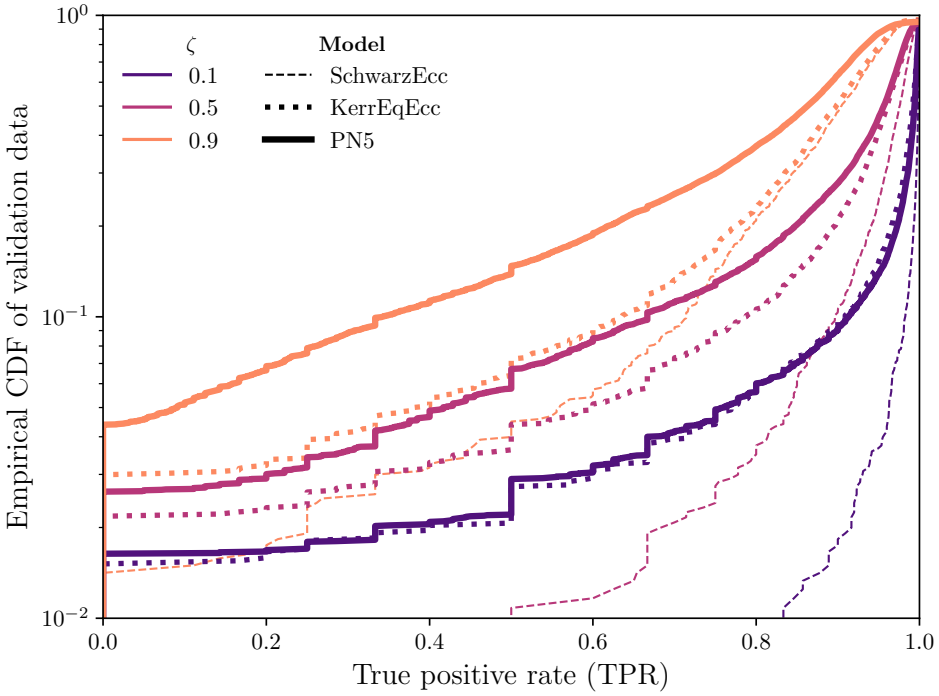


Figure 3.3: CDFs of the TPRs over mode index achieved by the neural mode predictor for the SchwarzEcc (dashed lines), KerrEqEcc (dotted lines) and PN5 (solid line) waveform models when evaluated over their validation datasets. As the mode retention threshold ζ is lowered, the networks preferentially retain more modes, increasing the TPR. For $\zeta = 0.5$, at least 80% of modes have a TPR of at least 0.8 in all cases.

because this causes the mode predictors to preferentially retain modes. However, this does not come without a cost, as more incorrectly retained modes slip into each result for the same reason. A useful metric by which we can explore this effect in more quantitative terms is to examine how the number of false positives or false negatives, summed over all mode indices, is distributed across the validation data. The false negatives are missed modes which the prediction algorithm has wrongfully rejected. Likewise, false positives are extra modes which have been retained when an actual mode selection would have discarded them. As we adjust ζ, the changing CDFs of these two quantities for each waveform model are shown in row (a) of Figure 3.4. The computational cost of waveform generation is roughly proportional to the number of modes included in the waveform. A more insightful expression of these two quantities is therefore to normalise them with respect to the number of modes retained by mode selection, which provides a strong indication of the relative increase (or decrease) in waveform computational cost. The CDFs of the normalised quantities evaluated over the validation dataset are shown in row (b) of Figure 3.4. For the PN5 model, setting $\zeta = 0.1$ ensures that 50% of mode predictions in the validation dataset do not exclude any modes which should have been retained, compared to a value of $\sim 10\%$ for $\zeta = 0.5$). However, this also leads to $\sim 50\%$ of mode predictions containing at least $\sim 10\%$ unnecessary modes, compared to an excess of just 2% for $\zeta = 0.5$. When necessary, waveform performance can be made more robust against outliers for less than 20% additional computational cost in most cases. The cost incurred by setting $\zeta = 0.1$ is even lower for simpler waveform models. For the rest of this chapter, we set $\zeta = 0.5$ to examine the baseline performance of the neural mode predictors; tuning ζ does not greatly affect the conclusions we draw as a result of this choice.

3.2.5 Exploring the mode retention space

When evaluating thousands of sets of input parameters on a GPU, the computational cost per mode prediction operation is roughly four orders of magnitude lower than mode selection. This paves the way for high-resolution exploration of EMRI mode retention as a function of parameters, which essentially expresses how the computational cost of EMRI waveforms varies over the parameter space. Correlations between different parameters can be examined by taking two-dimensional slices through the parameter space. We expect certain features to be immediately apparent in these slices of the mode retention function. For instance, there is a general trend that inspirals that reach the separatrix (and therefore complete many cycles deep within the strong field of the MBH) will have noticeably different mode retention behaviour to those that do not. This can be seen clearly in a heatmap of the mode retention in the (p_0, T) plane (with other parameters held fixed), shown for the PN5 model in Figure 3.5. At a given initial separation, the inspiral will take a certain amount of time to reach the separatrix: by varying the maximum length of the inspiral T we can identify this parametric boundary in parameter space. Aside from this broad feature, it is clear that there is significant structure on both small and large

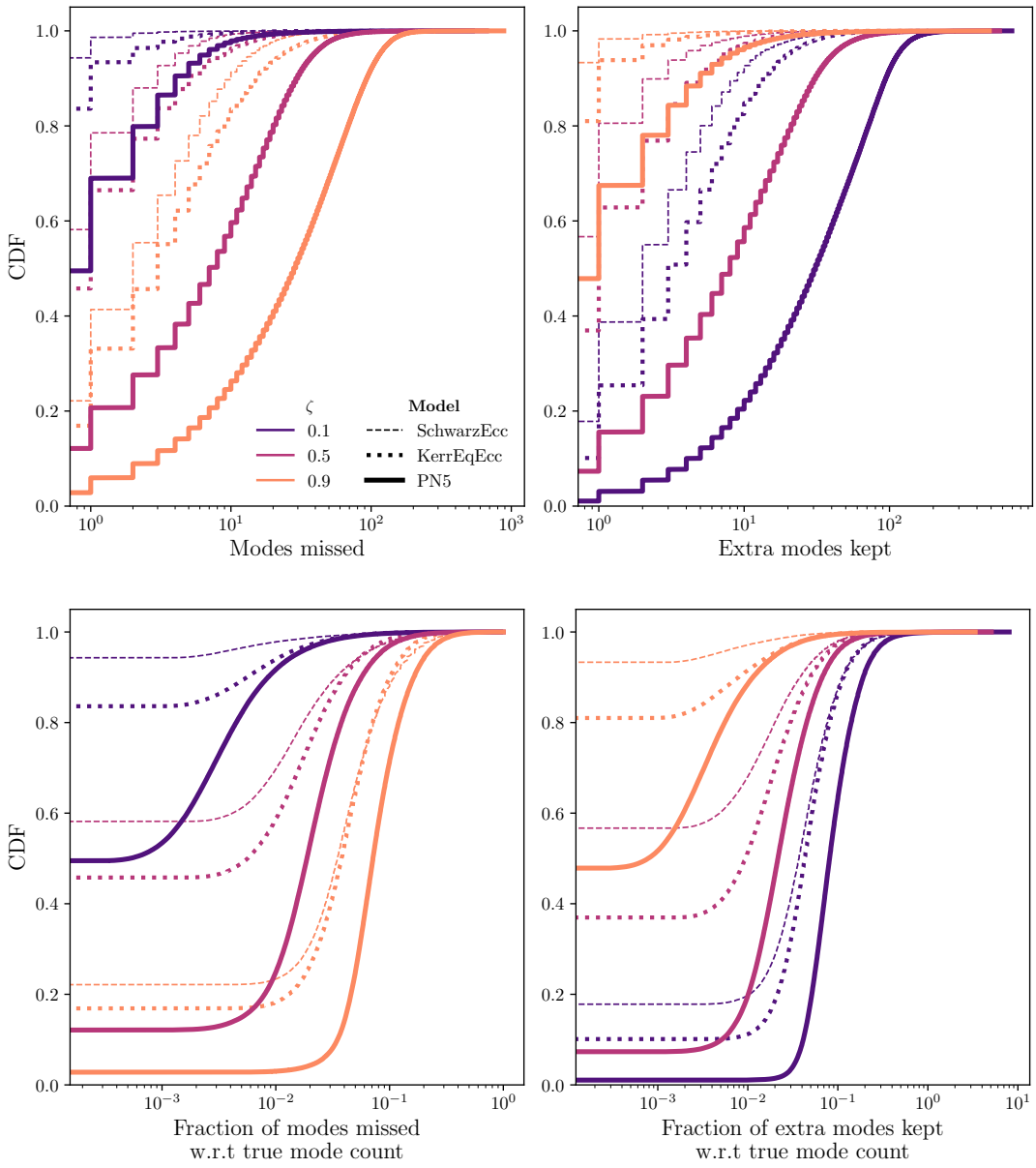


Figure 3.4: CDFs showing the mode retention performance of the mode prediction networks over the validation dataset for the SchwarzEcc (dashed lines), KerrEqEcc (dotted lines) and PN5 (solid lines) models. Left-hand panels of both rows indicate the number of modes missed per prediction, whereas right-hand panels describe the number of additional modes retained per prediction. The CDF input data for row (b) are normalised with respect to the true number of modes retained for each validation data point. From row (a), we can see that for $\zeta = 0.5$ at least one extra mode is missed or retained for 90% of cases in the PN5 model, with better performance for the simpler waveform models. Row (b) shows how changing ζ can improve mode prediction accuracy at the expense of waveform computational cost (or vice versa), because waveforms with more modes are more expensive to generate.

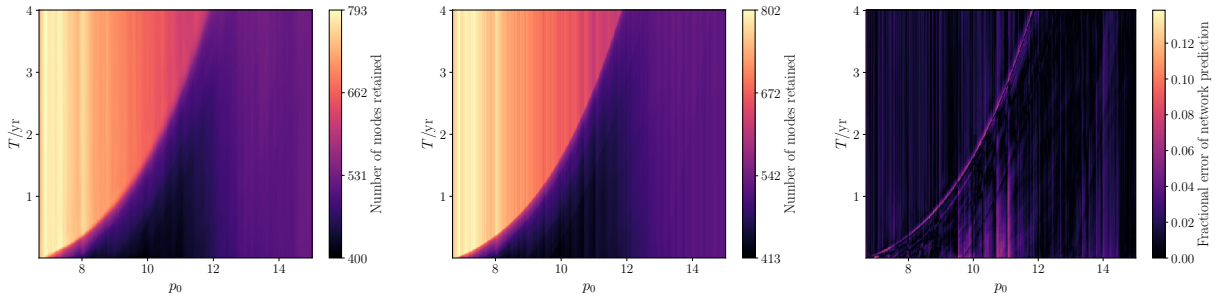
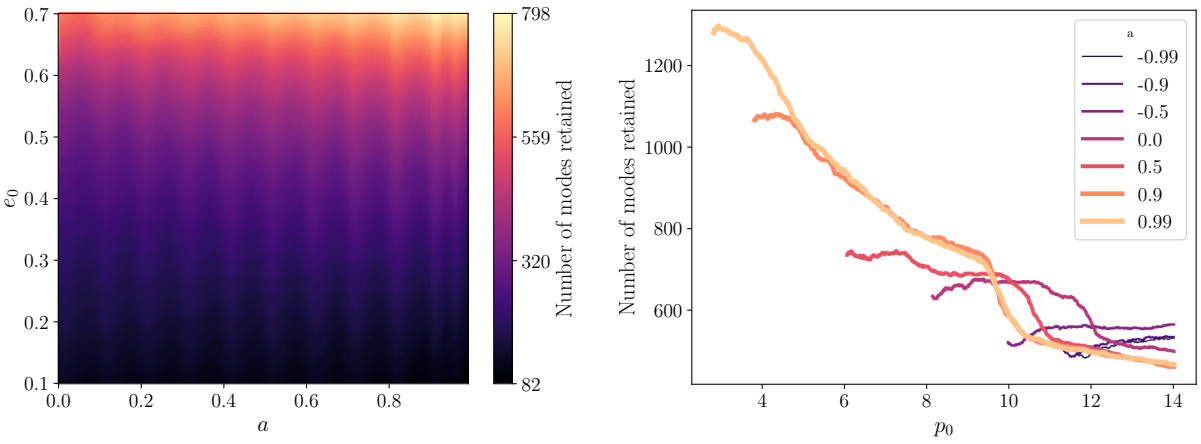


Figure 3.5: Total number of modes retained by (a) neural mode prediction, (b) mode selection, and (c) the fractional error (in terms of the number of modes retained) between these values as a function of initial semi-latus rectum p_0 and maximum inspiral duration T in years for the PN5 waveform model The boundary between plunging and non-plunging inspirals appears clearly as a sharp change in the mode content of the waveform. Additional structure on both large and small scales is common between (a) and (b), indicating the neural mode predictor is capable of accurately modelling the complex mode selection behaviour across the EMRI parameter space. The fractional error heatmap (c) is < 0.05 almost everywhere; along a sharp boundary in the mode retention space this error increases to ~ 0.13. The generation times of the first two plots were (a) 1 s and (b) 5 hr on a GPU.

scales within this two-dimensional slice. It is worth verifying that this structure is indeed present and not an artefact of the neural network's reconstruction of the mode selection behaviour. We therefore perform mode selection for each parameter set in the grid to obtain the heatmap shown in panel (b) of Figure 3.5. The wall time for these operations is ~ 5 hr on a GPU, which is computationally expensive compared to the neural network which completes the entire grid in ~ 1 s. Comparing the mode prediction output with the truth, panel (c) shows that the fractional error is less than 0.05 for most of this slice; along the parametric boundary linked to plunging inspirals, this fractional error reaches ~ 0.13 due to the rapid change in the mode content of the waveform in this region. The neural network encodes a smoothed form of the mode selection process, which is expected because a trained MLP is fundamentally an interpolation algorithm which operates over the training dataset. Despite this loss of fine detail, it is clear that the neural network successfully reproduces the structure in the mode selection function even on small scales.

The mode content strongly depends on the MBH spin, eccentricity and inclination of the system. The dependence on a and e_0 is best shown by the KerrEqEcc model, valid up to an eccentricity of 0.7. The number of retained modes for this model with $p_0 = 8$ is shown in panel (a) of Figure 3.6. Mode retention appears to be more strongly correlated with e_0 here, but the separatrix decreases as a increases so its impact is not as visibly apparent for fixed p_0. Taking the upper eccentricity value of $e_0 = 0.7$, the number of retained modes for $p_0 \in (p_{\rm sep} + 0.75, 14)$ is shown in panel (b) of Figure 3.6. In the range $p_0 \in [10, 12]$, prograde four-year inspirals reach the separatrix and the mode count increases, most sharply for highly spinning systems. The opposite

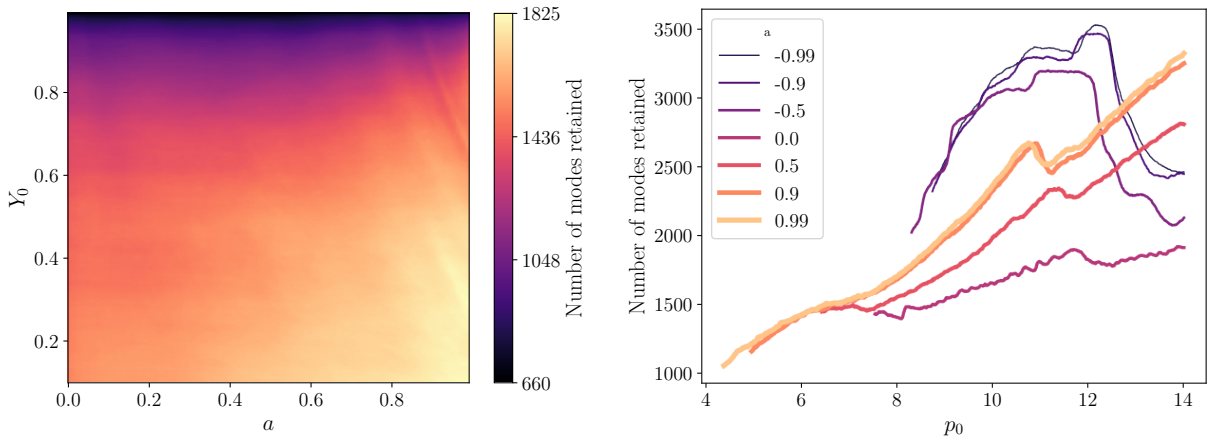


(a) (b)

Figure 3.6: Total number of modes retained by neural mode prediction over (a) a segment over MBH spin a and initial eccentricity e_0, and (b) segments over initial semi-latus rectum p_0 (minimum value $p_{\rm sep}+0.75$) for the KerrEqEcc waveform model. The mode content of the waveform depends strongly on both e_0 and a, with the latter correlation more apparent as p_0 approaches the separatrix $p_{\rm sep}$. Mode count is enhanced for prograde inspirals and suppressed for retrograde inspirals in proportion to a. Some periodic structure in mode count with respect to a is apparent, likely due to the spacing between data grids in the a-direction being too large.

is true for retrograde inspirals, which exhibit a suppression in mode count as a increases due to the inspiral plunging more quickly for higher MBH spin. As the initial p_0 approaches $p_{\rm sep}$, the inspiral effectively becomes more eccentric at a closer orbital separation, which manifests here as increasing (decreasing) mode counts for prograde (retrograde) inspirals. These highly eccentric inspirals are quite short ($\ll 4$ yr in duration) but could be extended backwards in time to a larger (p_0, e_0) which is beyond the edges of the numerical flux grids. As some EMRIs formation mechanisms are expected to have large initial eccentricities, this highlights the need for flux grids to be extended in the near future.

A similar comparison can be made to understand the impact of inclination on mode retention via the PN5 model. Setting $e_0 = 0.4$ (the edge of our domain of validity for this model), a slice in (a, Y_0) is shown in panel (a) of Figure 3.7. The contribution of modes with high k as orbital inclination increases leads to a steadily increasing total mode count, which also appears slightly correlated with spin. Akin to the KerrEqEcc case, to better illustrate the correlation between mode count and *spin* we evaluate the number of retained modes in $p_0 \in (p_{\rm sep}+0.75, 14)$, with $(e_0, Y_0) = (0.4, 0.5)$. The resulting mode retention behaviour, shown in panel (b) of Figure 3.7, is very different to that of the KerrEqEcc model. Total mode count largely decreases as a function of p_0 for prograde inspirals, with a small peak which trends inwards as a function of MBH spin. For retrograde inspirals, a similar behaviour is observed but with a larger peak which trends outwards for higher MBH spin. This behaviour is the result of a mode frequency cut-off within the PN5 model, which sets $Z_{\ell mkn}$ to zero if $|\Omega_{mkn}| > 0.205\,{\rm rad\,s^{-1}}$. Therefore, as inspirals enter the strong field, higher-order modes (which would normally begin to contribute more strongly

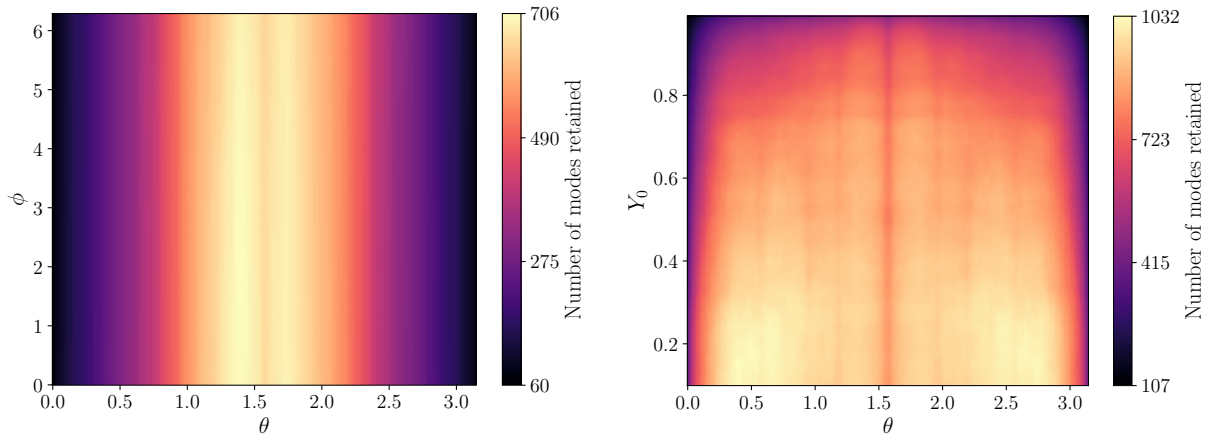


(a) (b)

Figure 3.7: Same as for Figure 3.6, but varying the cosine of the initial orbital inclination Y_0 rather than e_0 and for the PN5 waveform model. Mode content rapidly grows as orbital inclination deviates from that of an equatorial inspiral. Unphysical frequency cut-offs in the PN5 amplitude model suppress the number of modes at lower p_0, preferentially for prograde inspirals as they have larger ω_{mkn} and therefore have more modes above the cut-off.

and pass mode selection, as for the KerrEqEcc model in Figure 3.6) are instead discarded, causing the mode spectrum to rail up against the frequency cut-off and suppressing the total number of modes retained. For retrograde inspirals, ω_ϕ is negative, so Ω_{mkn} is typically smaller which permits more higher-order modes. In future work, we plan to remove the frequency cut-off to properly forecast the performance of our method for a generic Kerr waveform model. A strong cut-off in mode retention itself presents a challenge for neural networks to model, as this discontinuity forms a boundary in parameter space which differs for every set of mode indices. We therefore expect that the neural network will have some capacity to accommodate the additional modes which will be present after the removal of the frequency cut-off, as this will also remove the discontinuity.

We can also examine how waveform mode count varies with respect to the source-frame viewing angle of the observer. Examining a slice on the (θ, ϕ) plane for the KerrEqEcc model (panel (a) of Figure 3.8), it is clear that the azimuthal viewing angle has no bearing on the mode content of the waveform. The mode content with respect to θ for this equatorial waveform model peaks around $\pi/2 \pm 0.2$ independent of the other parameters of the inspiral. Introducing inclination into the system via the PN5 model (panel (b) of Figure 3.8), we see that the mode count flattens with respect to θ as the orbit deviates further from an equatorial configuration. In general, the mode count is highly suppressed for near-polar viewing angles (due to the symmetry of the system) as well as edge-on viewing angles, albeit to a lesser extent. Panel (b) of Figure 3.8 also highlights the challenges inherent to the extension of existing models to generic Kerr inspirals, as the number of modes increases as Y_0 deviates from 1, changing by an order of magnitude with respect to equatorial inspirals.



(a) (b)

Figure 3.8: Total number of modes retained by neural mode prediction over (a) viewing angles θ and ϕ for the KerrEqEcc model, and (b) θ and cosine of initial orbital inclination Y_0 for the PN5 waveform model. The mode content of the waveform is independent of ϕ. For equatorial inspirals (panel (a)), the mode content as a function of θ peaks around $\pi/2 \pm 0.2$, but introducing orbital inclination into the system (panel (b)) broadens this peak into a flat distribution (with the exception of edge-on and face-on viewing angles, for which the mode content is suppressed).

Last, a coarse view of mode selection behaviour across the parameter space is shown by scatter-corner plots for the KerrEqEcc and PN5 models in Figure 3.9 and Figure 3.10 respectively. The pairwise correlations described above can be seen clearly here, but no other large-scale correlations are apparent. This is not because the mode selection behaviour with respect to these other parameter combinations is simple; rather, it is due to the complexity being on a scale too fine to be resolved by this plot. For instance, no correlation is apparent between total mode retention and the masses (M, μ), but a two-dimensional slice (Figure 3.11) over these parameters reveals that a strong feature is present (which again corresponds to a boundary between plunging and non-plunging inspirals). Capturing this fine-scale complexity is precisely why the flexibility of neural networks is necessary to emulate the mode selection process.

3.3 Waveform mismatch and timing analysis

To properly characterise the performance of the neural network mode predictors, we must examine their effectiveness at the level of waveform generation. There are two key metrics by which we can examine their performance. For the first, a starting point is to measure the wall-time for the neural mode predictor waveform model. This can then be compared to the wall-time for the corresponding waveform model when performing mode selection to directly quantify the performance improvement obtainable with our method. A figure of merit which directly quantifies the improvement realised with the use of neural networks is to take the ratio of these two wall-times, obtaining a speed-up fraction $\mathcal{T}$; an equivalent cost for the two waveforms corresponds

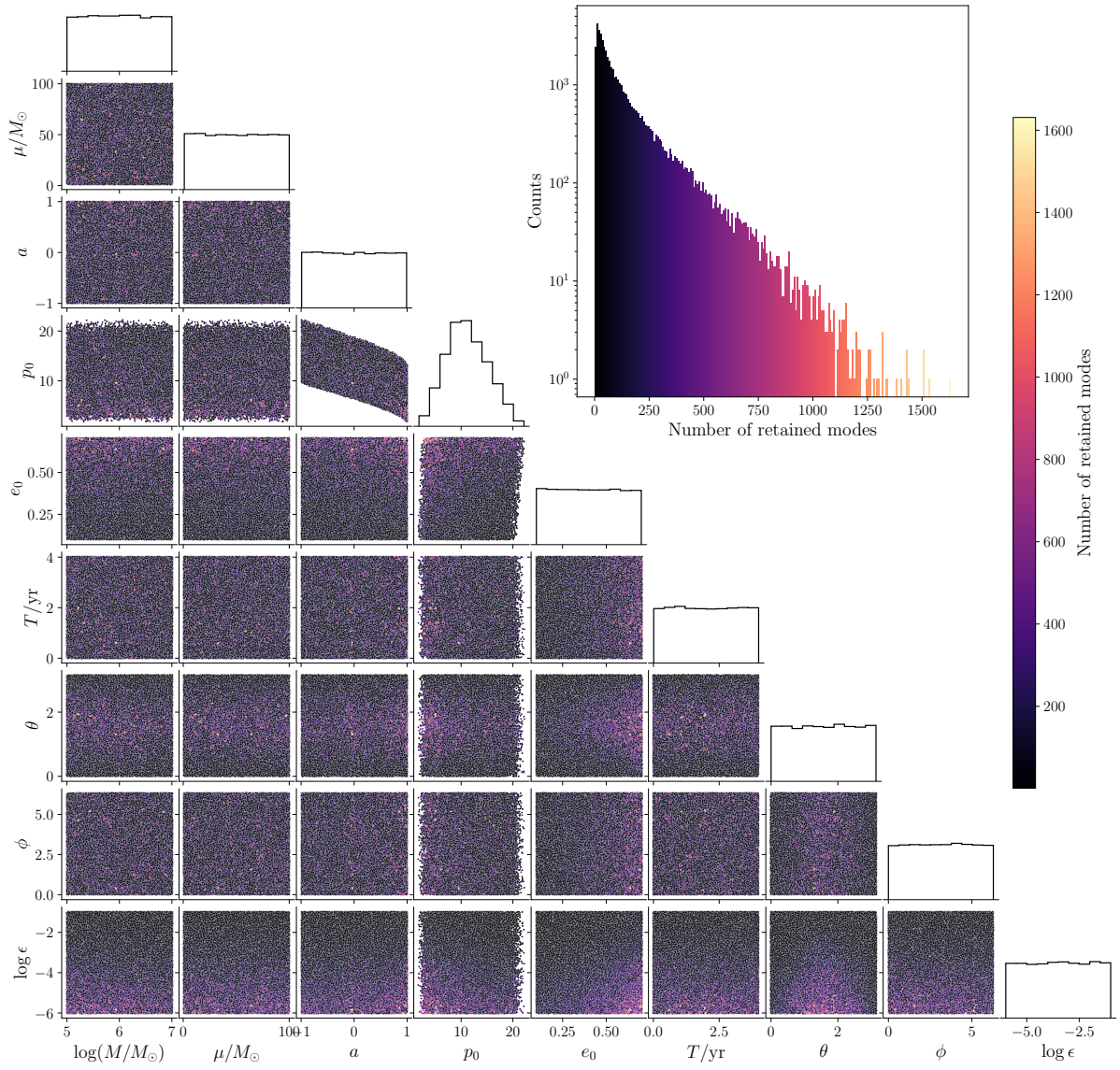


Figure 3.9: Total number of modes retained by mode prediction for 50,000 random draws from the EMRI parameter space. The marginal distribution of each sampled parameter is indicated by the histograms along the main diagonal. For high ε and/or low e_0 only ~ 100 modes are retained, but as $e_0 \to 0.7$ at low ε waveforms contain up to ~ 1000 modes. The complexity of the mode selection process is only visible on large scales in this plot.

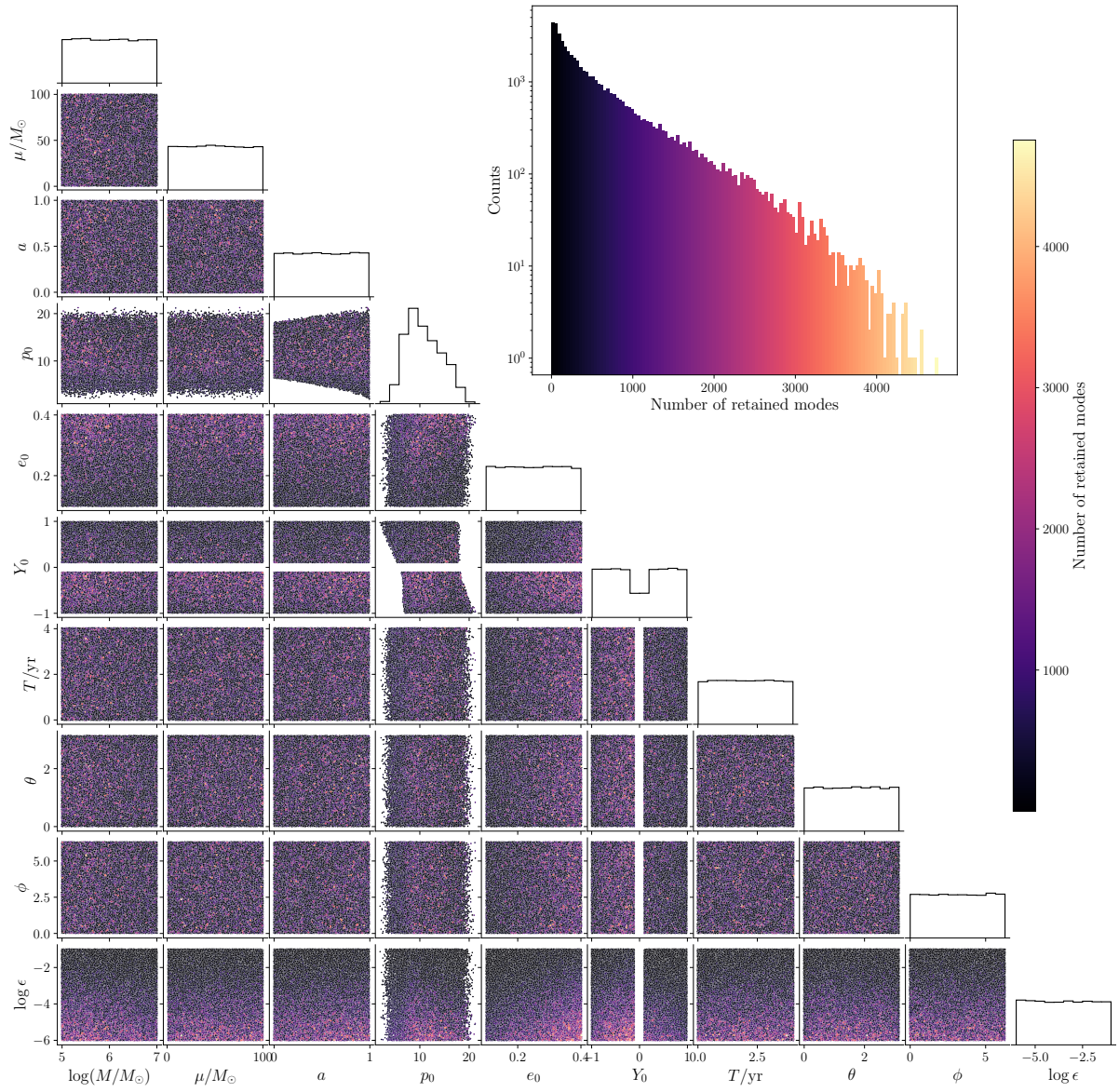


Figure 3.10: Same as Figure 3.9, but for the PN5 waveform model. At high ε and/or low e_0 relatively few modes are retained, but this grows quickly as a function of orbital eccentricity and inclination. A general disparity in mode retention between prograde and retrograde inspirals is apparent, which is explored in Figure 3.7.

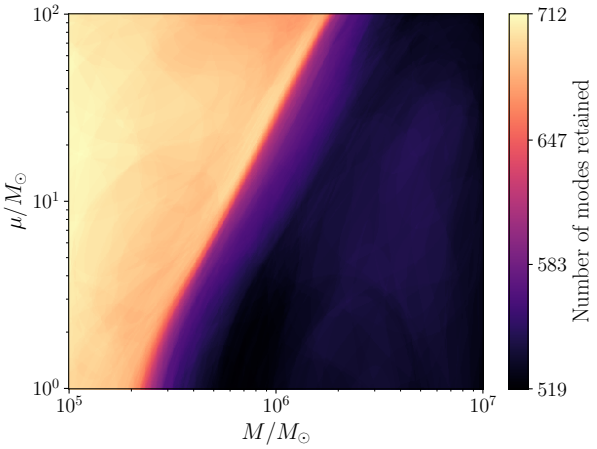


Figure 3.11: Total number of modes retained by neural mode prediction for a segment over MBH mass M and CO mass μ for the PN5 waveform model. Similarly to Figure 3.5, the boundary between plunging and non-plunging inspirals appears here as a rapid change in the mode content of the waveform, despite no strong correlation between M and μ appearing in a two-dimensional projection of the full parameter space in Figure 3.9. This emphasises the need for highly expressive and flexible algorithms if mode selection is to be modelled accurately.

to $\mathcal{T} = 1$, with $\mathcal{T} > 1$ denoting a lower wall-time for the neural mode prediction waveform. The improvement realised will depend both on the hardware used to compute the waveform (i.e., CPU or GPU) and the domain in which the waveform is computed (either time or frequency domain). The analysis of EMRI signals is expensive, and reducing the computational cost of waveform models is crucial in the construction and application of search and PE pipelines.

A useful performance metric is the waveform mismatch, which is defined between two waveforms h_1 and h_2 as [289]

$$\mathcal{M}(h_1, h_2) = 1 - \frac{\langle h_1 | h_2 \rangle}{\sqrt{\langle h_1 | h_1 \rangle \langle h_2 | h_2 \rangle}}. \tag{3.30}$$

The use of inner products naturally weights this comparison metric according to the sensitivity curve of the detector. However, as we are not weighting the selection of modes with respect to a given noise model for generality, these inner products will likewise be computed with a uniform PSD. This quantity is zero for $h_1 = h_2$, which in this case will only occur if the two waveforms consist entirely of the same set of modes. Any discrepancies between the modes retained in each case will lead to a non-zero mismatch. We expect there to be disagreement between the results of mode prediction and mode selection (the neural network is not a perfect emulator, after all) but the power of the erroneously discarded/retained modes will be of order ε. The second metric by which we will quantify the performance of our method is therefore the normalised mismatch $\tilde{\mathcal{M}} = \mathcal{M}/\varepsilon$; if this is less than one, the two waveforms will be of similar accuracy when used in PE despite the small differences between them. In other words, provided that the mode prediction output is sufficiently close to what would be obtained from mode selection, any biases incurred when the waveform is used in the analysis of a signal will primarily be the

result of discarding low-power waveform modes, irrespective of the method which was used to perform this task.

We now examine each of the three waveform models (SchwarzEcc, KerrEqEcc and PN5) one at a time with respect to both wall-time and normalised mismatch. By starting with a simple waveform model, we establish a baseline cost which provides context for the increased computational cost of more complex waveform models.

3.3.1 Schwarzschild Eccentric

The SchwarzEcc model is of low computational cost; the inspiral trajectory is inexpensive to compute for Schwarzschild geometry, typically taking under 1 ms, and the number of modes included in waveforms can be relatively small at lower eccentricities. The cost breakdown for a typical Schwarzschild eccentric waveform is shown in Figure 3.12 for various domain, hardware and parameter configurations. In each case, $(M,\mu) = (10^6, 10^1)$ and $(\theta,\phi) = (\pi/3, \pi/4)$. The waveform duration is 4 yr, with p_0 adjusted such that the CO plunges at the end of the observation window. Time-domain computation of the waveform on a CPU is prohibitively expensive (taking more than 1 s for all parameter sets) and is not shown. Frequency-domain waveforms are computed on a uniform grid of frequency bins which contains 0.01% of the number of bins in the FFT of the corresponding time-domain waveform, spaced uniformly between the minimum frequency $f_{\min} = 1/T$ and the maximum frequency of any harmonic mode in the waveform $f_{\max}$. For convenience, $f_{\max}$ is determined empirically by generating a waveform over all frequency bins and identifying the largest frequency bin with a non-zero value. As $f_{\max}$ can be small for extreme mass ratios with large p_0, we enforce a minimum frequency bin count of 100. Additionally, we use the same frequency grid for both mode selection and mode prediction waveforms so that mismatch calculations can be made. The frequency bin grid generation step is not included in waveform timing.

The bicubic spline code used in this waveform model is CPU-based, so only the mode selection and summation stages differ in cost between different hardware choices. The amplitude generation and subsequent mode selection is a clear bottleneck for this waveform, especially at lower eccentricities and larger mode power thresholds. This is particularly pronounced for the frequency-domain CPU-based waveforms, where the mode selection operation alone takes more than 100 ms. We therefore expect our method to be extremely impactful on the computational cost of the SchwarzEcc waveform, as it directly addresses this issue.

The reduction in wall-time per waveform evaluation for the amplitude and mode selection stages are indicated by the hatched regions in Figure 3.12. The computational cost of both mode selection and amplitude generation are reduced by more than an order of magnitude for all configurations shown, with an additional order of magnitude speed-up achieved in some cases. The improvement is especially noticeable for waveforms with higher ε, which is expected because more modes are discarded in this case (meaning more compute time is wasted by waveforms

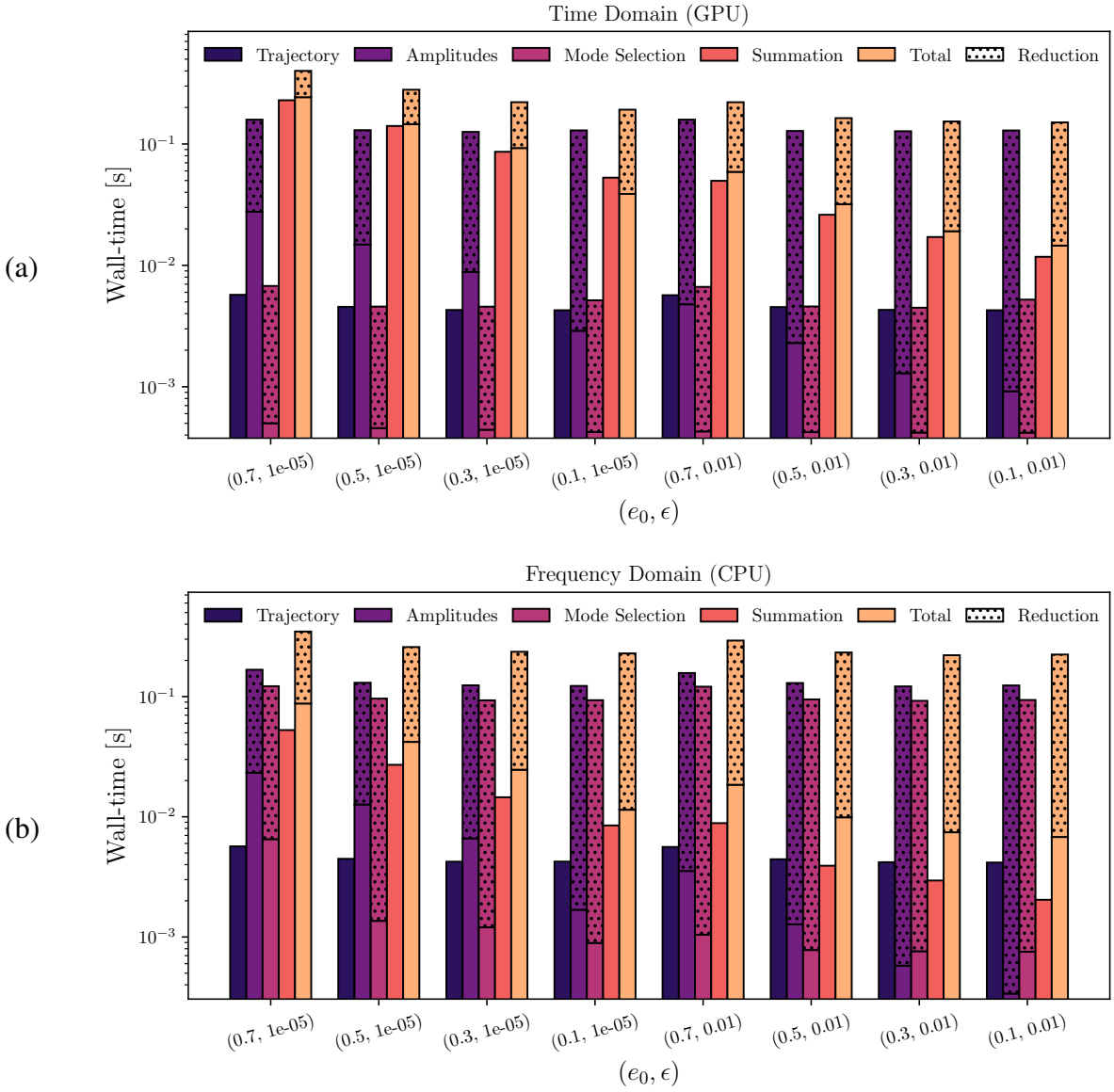


Figure 3.12: A breakdown of the wall-time for each stage of SchwarzEcc waveform generation, in the (a) time domain with a GPU, or (b) frequency domain with a single CPU core. As these waveforms always use CPU-based amplitude generation, time-domain GPU-based waveforms are improved significantly, particularly at high mode selection thresholds. In frequency-domain CPU-based waveforms, the cost of both amplitude generation and mode selection stages (which originally bottleneck waveform generation) are significantly reduced. At high mode selection thresholds, the trajectory stage forms a significant percentage of the computational cost for waveforms which use neural mode predictors.

performing mode selection after amplitude generation). In fact, the integration of the trajectory, which initially formed a small fraction of the overall cost, now constitutes a significant portion of the wall-time. This suggests that this stage may need to be more efficient in the future if sub-millisecond waveform costs are to be achievable. The most significant overall reduction in waveform cost is achieved for frequency-domain CPU-based waveforms with higher ε, where waveform wall-time is reduced from $\sim 250\,\mathrm{ms}$ to $\sim 5\,\mathrm{ms}$, an improvement of nearly two orders of magnitude.

The values of our figures of merit ($\mathcal{T}$ and $\tilde{\mathcal{M}}$) over the domain of validity of the waveform model are shown in panels (a) and (b) of Figure 3.13 for time-domain GPU- and frequency-domain CPU-based waveform generation respectively. With the exception of < 10 outliers, all waveforms generated have $\tilde{\mathcal{M}} \lesssim 10^{-1}$, indicating that the mode predictions of the neural network are sufficient for the primary contribution to the mismatch between the waveform and one containing all modes to be the act of discarding modes in the first place (which leads to a mismatch of $\mathcal{O}(\varepsilon)$). A slight correlation between $\tilde{\mathcal{M}}$ and M is present for waveforms generated in either domain, which is likely due to waveforms of larger M being of more extremal mass-ratios on average and therefore less likely to plunge during the maximum duration T of the waveform. As a consequence of this, these waveforms consist of relatively few dominant modes and the incorrect mode predictions of the network penalise mismatch more heavily. The small number of outliers with $\tilde{\mathcal{M}} \sim 0$ occur at the most extreme mass ratios, and are systems with large p_0 that evolve slowly and contain few modes. In this scenario, excluding or including even a few modes can harm mismatches considerably. Signals with these parameters will have wide posterior distributions because such slow evolution far from plunge does not depend on the parameters of the system as strongly as for plunging signals. The less accurate mode prediction performance in this regime is therefore acceptable as it is unlikely to induce noticeable biases on any recovered parameters.

In line with what is seen in Figure 3.12, $\mathcal{T}$ is highest for waveforms with low overall mode retention (high ε, low e_0). Speedup ratios of more than an order of magnitude are commonplace across the parameter space. As the cost of search and PE analyses is largely tied to waveform generation, the computational cost of these tasks will decrease accordingly. The use of neural mode predictors improves the efficiency of amplitude generation sufficiently that this is no longer the leading bottleneck of the waveform model, which enhances further computational gains made in the future as other stages of waveform generation become more efficient.

3.3.2 Kerr Equatorial Eccentric

In the KerrEqEcc model, additional computational cost enters every stage of waveform generation. The data grids over which the adiabatic fluxes are interpolated increase in dimension by one, the fundamental frequency equations become more complex and numerical separatrix calculations are necessary to accurately identify the termination point of the inspiral. Additionally,

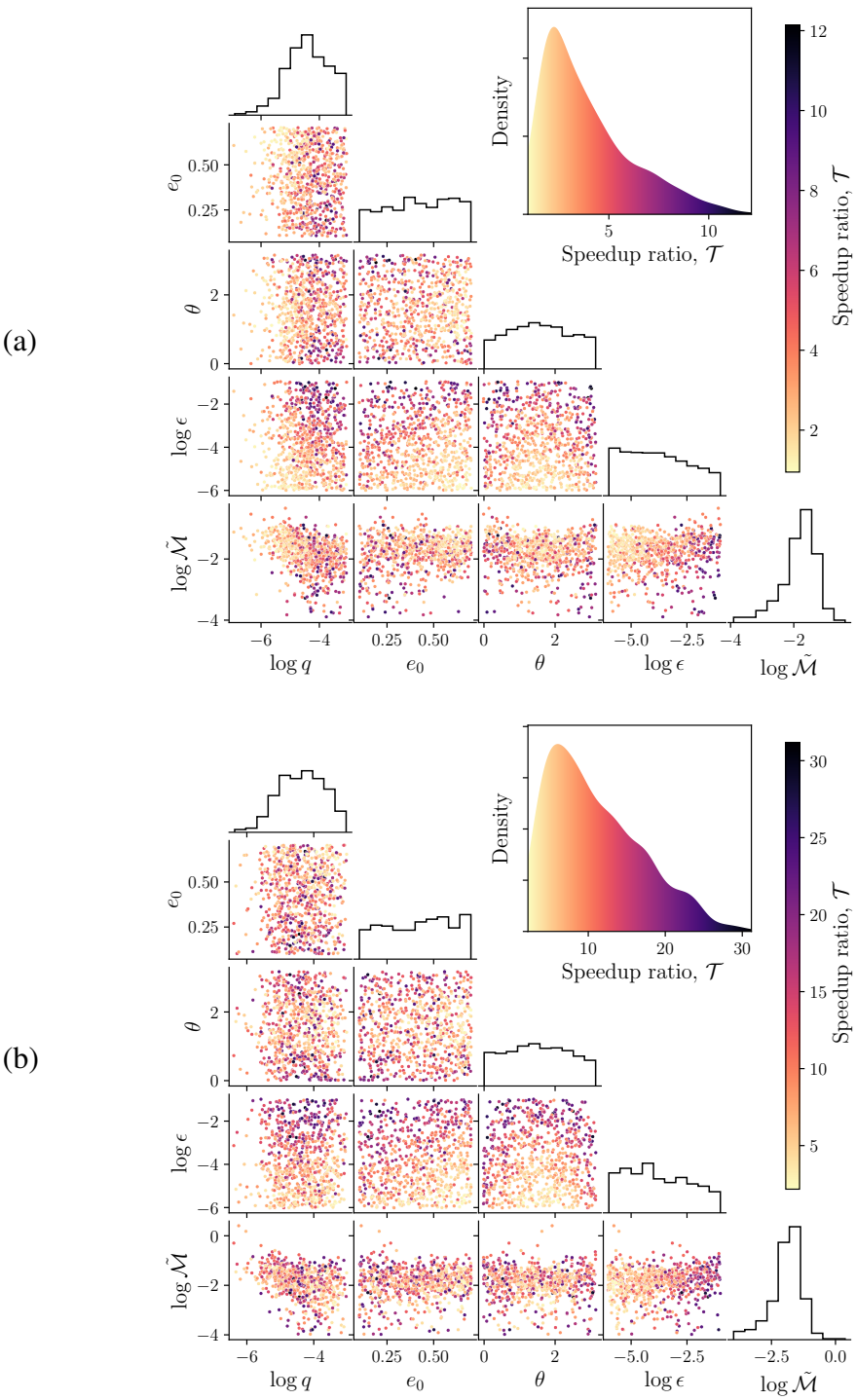


Figure 3.13: The performance of neural mode prediction in the SchwarzEcc model, compared to a mode selection waveform in terms of normalised mismatch $\tilde{\mathcal{M}}$ and speedup ratio $\mathcal{T}$, as a function of source parameters. Panels (a) and (b) correspond to time-domain GPU- and downsampled frequency-domain CPU-based waveform generation respectively. Greater improvements are observed in the frequency domain case due to the lower cost of the summation step for these waveforms. As $\tilde{\mathcal{M}} \lesssim -1$ for most of the parameter space, we expect waveforms augmented with neural mode prediction to perform well in PE analyses (see Section 3.4).

as $a \to 1$ the ODEs for $p(t)$ and $e(t)$ become stiff as $p(t) \to p_{\text{sep}}$ and $e(t) \to 0$. At present, FEW solves the trajectory ODE system with the explicit embedded Runge–Kutta Prince–Dormand (8, 9) method [408], which is not well-suited to solving stiff equations of motion and must take smaller adaptive steps [409]. The larger number of trajectory points incurs additional cost in both the integration of the trajectory and subsequent interpolated summation of waveform modes (as there are now more mode amplitude values and spline coefficients to compute).

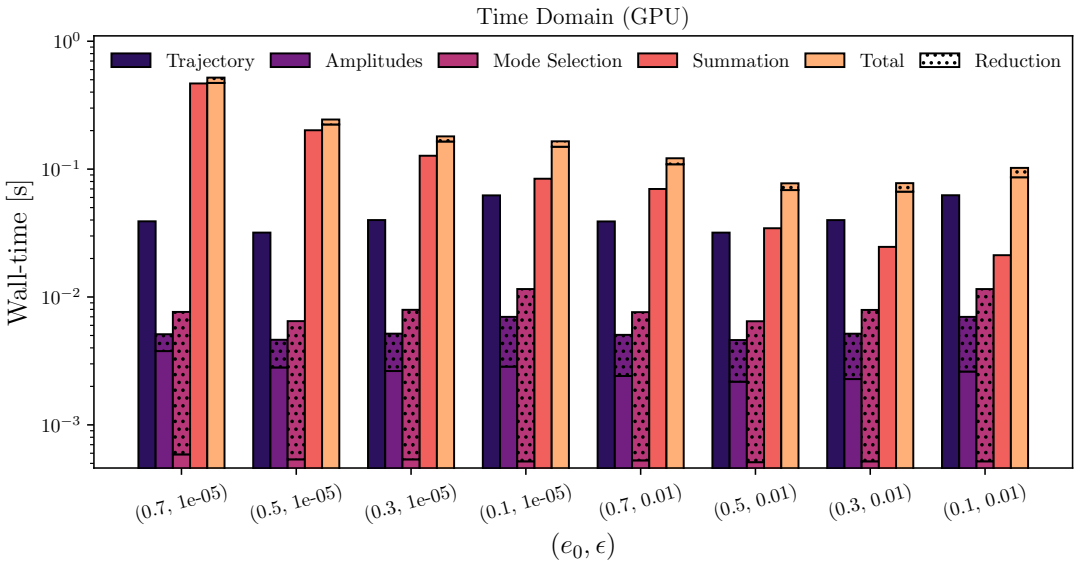
Performing linear interpolation of mode amplitudes over a reduces the computational cost incurred per mode amplitude to a factor of two higher than in the Schwarzschild eccentric regime. However, the larger total number of mode indices which must be considered leads to a further cost increase in the mode amplitude generation, mode selection and summation stages.

The wall-time of each stage of the KerrEqEcc model are demonstrated in Figure 3.14, with the above increases in computational cost being clearly apparent when compared to the SchwarzEcc timings of Figure 3.12. For all waveforms, $a = 0.91$ is chosen such that the adiabatic fluxes and mode amplitudes must be evaluated between grid points in the a direction (as will be typical of waveforms in any realistic data analysis scenario). The trajectory computation is now a significant bottleneck to GPU-based waveform generation, especially for higher ε, which is in stark contrast to the benchmarks of the SchwarzEcc model. As trajectory wall-time increases with fixed T are largely due to an increase in the number of locations in time at which the adaptive solver must place a point, the computational cost of the mode amplitude step increases accordingly. This behaviour can be seen in the case of $e_0 = 0.1$, where ODE stiffness leads to more points in the sparse trajectory. In agreement with the SchwarzEcc benchmarks, our neural mode selection scheme is most effective for frequency-domain CPU-based waveform generation.

The variation of $\mathcal{T}$ and $\tilde{\mathcal{M}}$ for the KerrEqEcc waveform over its domain of validity are shown in Figure 3.15. The global behaviour of $\mathcal{T}$ matches the snapshot presented by Figure 3.14: for time-domain GPU waveforms, the trajectory and summation stages are the most expensive part of waveform generation and accelerating amplitude generation and mode selection only has a minor impact (10–20%) on waveform generation wall-time, with the greatest improvement in for higher ε because the summation is least expensive in that region. For frequency-domain CPU waveforms, $\mathcal{T} > 5$ almost everywhere and $\mathcal{T} > 20$ is obtained for $\varepsilon > 10^{-2}$. This result is slightly better overall than what was obtained for the SchwarzEcc model, despite the increased trajectory cost; this is due to the number of modes being higher in the KerrEqEcc model combined with the increased cost of amplitude generation (due to the additional linear interpolation step for each mode). Assuming that the cost distribution per waveform generation at higher ε is similar everywhere to what is observed in Figure 3.14, optimisation of the trajectory is of high priority, potentially yielding between 100% and 400% reduced computational costs depending on the value of e_0.

Normalised mismatch behaviour is also similar to the SchwarzEcc case, with $\mathcal{M} < -1$ al-

(a)



(b)

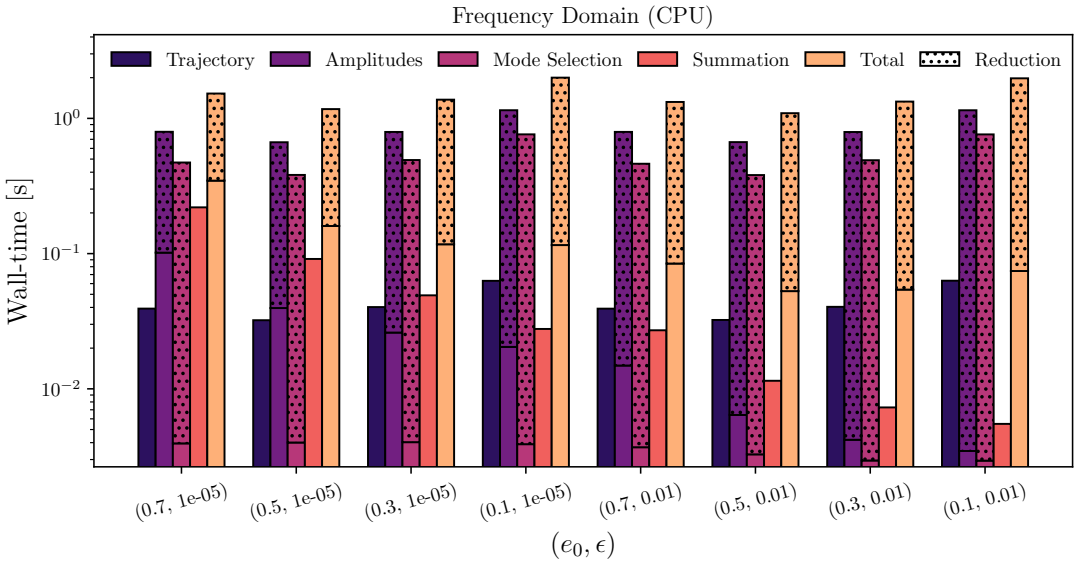


Figure 3.14: A breakdown of the wall-time for each stage of KerrEqEcc waveform generation, in the (a) time domain with a GPU, or (b) frequency domain with a single CPU core. The efficiency of the GPU in amplitude generation means that only a moderate reduction in computational cost is achieved for time-domain GPU waveforms. On the CPU, significant improvements of up to an order of magnitude in total cost are achieved because amplitude generation and mode selection are the primary bottleneck.

most everywhere and a few outliers present at extreme mass ratios. We therefore expect that a KerrEqEcc waveform model augmented with neural mode prediction will be suitable for search and PE analyses.

3.3.3 Generic Kerr

The computational cost of the PN5 waveform, when generated in the time domain, is dominated by the cost of the summation stage due to the large number of additional harmonic modes. The PN5 trajectory is of similar computational cost to the KerrEqEcc trajectory, and also suffers from stiffness issues as $e \to 0$. Computational wall-times for each stage of the PN5 waveform model are shown in Figure 3.16 for time-domain GPU-based waveform generation.

Amplitude generation is fast due to the analytic form of the PN5 amplitudes. For a grid-based generic Kerr model, this step is expected to be of higher cost; scaling the computational cost of this stage of the KerrEqEcc model by the increase in mode count yields costs of $\sim$ 100 ms on a GPU, or a few seconds on a CPU. Mode selection cost is reduced significantly from $\sim$ 10 ms to under 1 ms for all sets of parameters. This leads to modest improvements in waveform computational cost at higher ε.

A generic Kerr frequency-domain summation kernel is currently under development. We therefore restrict our timing and mismatch comparisons to time-domain GPU-based waveform generation. This will allow us to assess whether the mode prediction results will be sufficiently accurate for use in data analysis. By applying the fractional reduction in computational cost for the KerrEqEcc waveform when switching a to CPU-based frequency-domain summation to the summation wall-times in Figure 3.16, we predict that the waveform summation will not be the primary bottleneck to CPU waveform generation. Even without this CPU-based waveform generation framework, we can still examine the computational savings achieved with neural mode prediction by timing this stage in isolation. We find that mode selection takes between 0.5–1 s depending on the value of e_0, whereas mode prediction takes 4 ms in all cases because it is not sensitive to the number of points in the sparse trajectory output. This reduction in computational cost by more than two orders of magnitude is therefore essential if CPU-based waveform generation is to be practical for use in data analysis.

Given the reduction in cost achieved for mode selection when performed on the CPU, the trajectory and amplitude stages are likely to form the majority of the computational cost of CPU waveform generation for grid-based generic Kerr waveform models.

Examining our figures of merit for the PN5 waveform model over its domain of validity (Figure 3.15), neural mode selectors yield roughly a 10% reduction in computational cost for time-domain GPU-based waveform generation. Waveforms with $Y_0 < 0.3$ are not included because ODE stiffness causes the sparse trajectory to have too many points for the waveform summation kernel, which is limited by the number of GPU threads available. In a small number of cases the waveform cost increases because the mode prediction returns sufficient additional

(a)

(b)

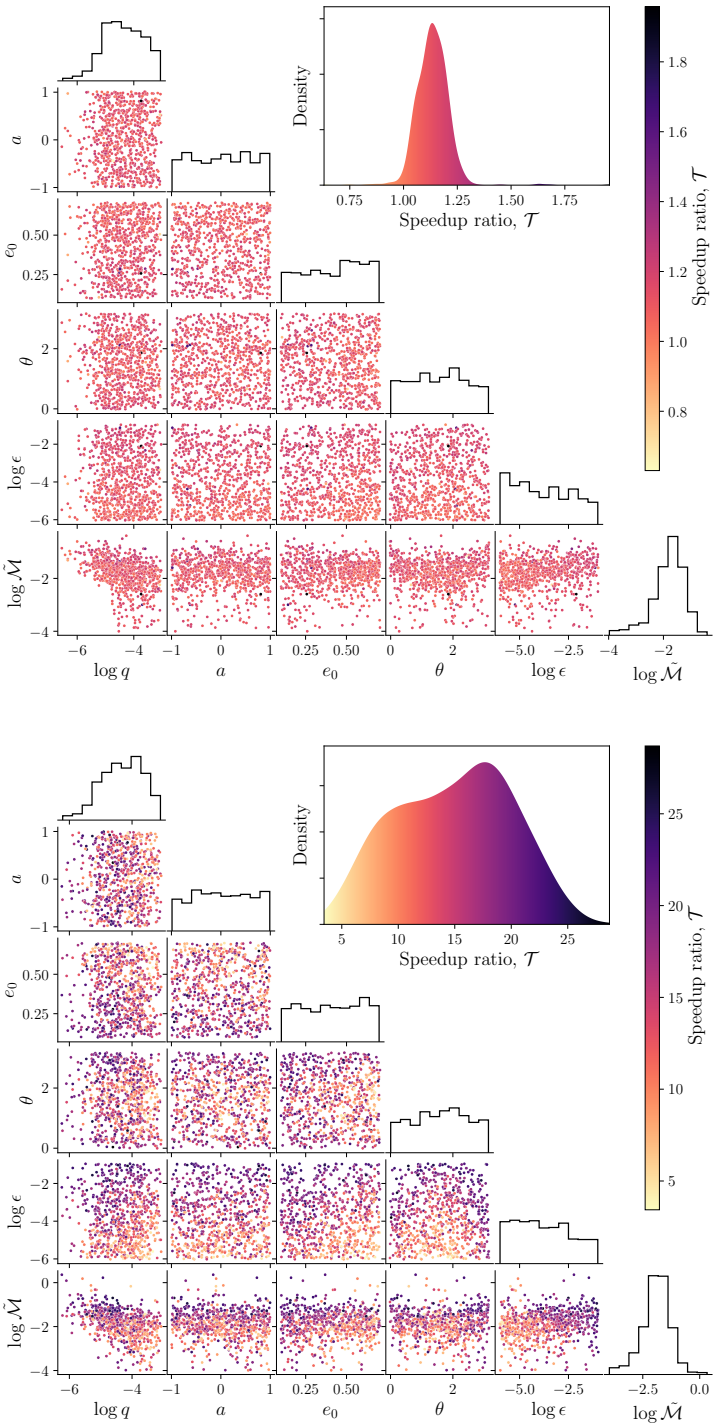


Figure 3.15: The same as Figure 3.13, but for the KerrEqEcc waveform model. In line with the results shown in Figure 3.14, time-domain waveforms on GPU generally only see an improvement of $\sim 10\%$ in computational cost across the parameter space. Frequency-domain CPU-based waveform cost sees further relative improvement when using neural mode prediction compared to the SchwarzEcc waveform model, with $\mathcal{T} \sim 20$ for higher ε. As $\tilde{\mathcal{M}} \lesssim -1$ for most of the parameter space, we expect waveforms augmented with neural mode prediction to perform well in PE analyses (see Section 3.4).

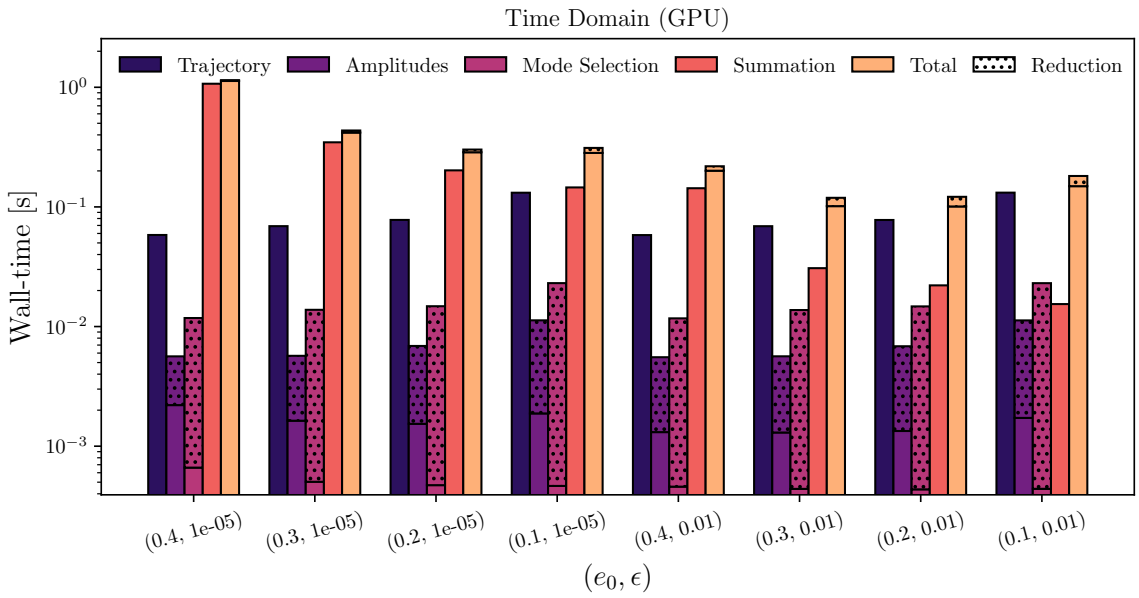


Figure 3.16: A breakdown of the wall-time for each stage of PN5 waveform generation, in the time domain with a GPU. Frequency-domain waveform generation is not yet supported for the PN5 model and is therefore not shown. For high mode count, the time-domain summation is the primary bottleneck; for a higher mode selection threshold the trajectory and summation stages are of roughly equivalent cost.

modes with respect to mode selection that the resulting cost increase of the summation stage eliminates any savings achieved during amplitude generation and mode selection. The cause of these outliers may be due to the discontinuity caused by the frequency cut-off discussed in Section 3.2.5. Near this discontinuity, the neural network will be likely to retain additional modes as it attempts to model the sharp change in mode retention across the parameter space. This instability is likely the cause of the upper tail in the $\tilde{\mathcal{M}}$ histogram. The majority of waveforms have acceptable mismatch performance; we therefore expect that once the frequency cut-off is removed, our method will yield a waveform model with global performance sufficient for use in data analysis.

3.4 Rapid parameter estimation

The results of the previous section suggest that EMRI waveforms augmented with neural mode prediction are significantly more performant than their mode-selection counterparts without compromising accuracy. We now examine the impact of mode prediction errors on EMRI parameter recovery by performing PE runs and comparing the results between mode-prediction and mode-selection waveforms.

PE for EMRI systems in LISA data is extremely challenging if no prior information about the signal parameters is known. The intrinsic parameters of the system can typically be con-

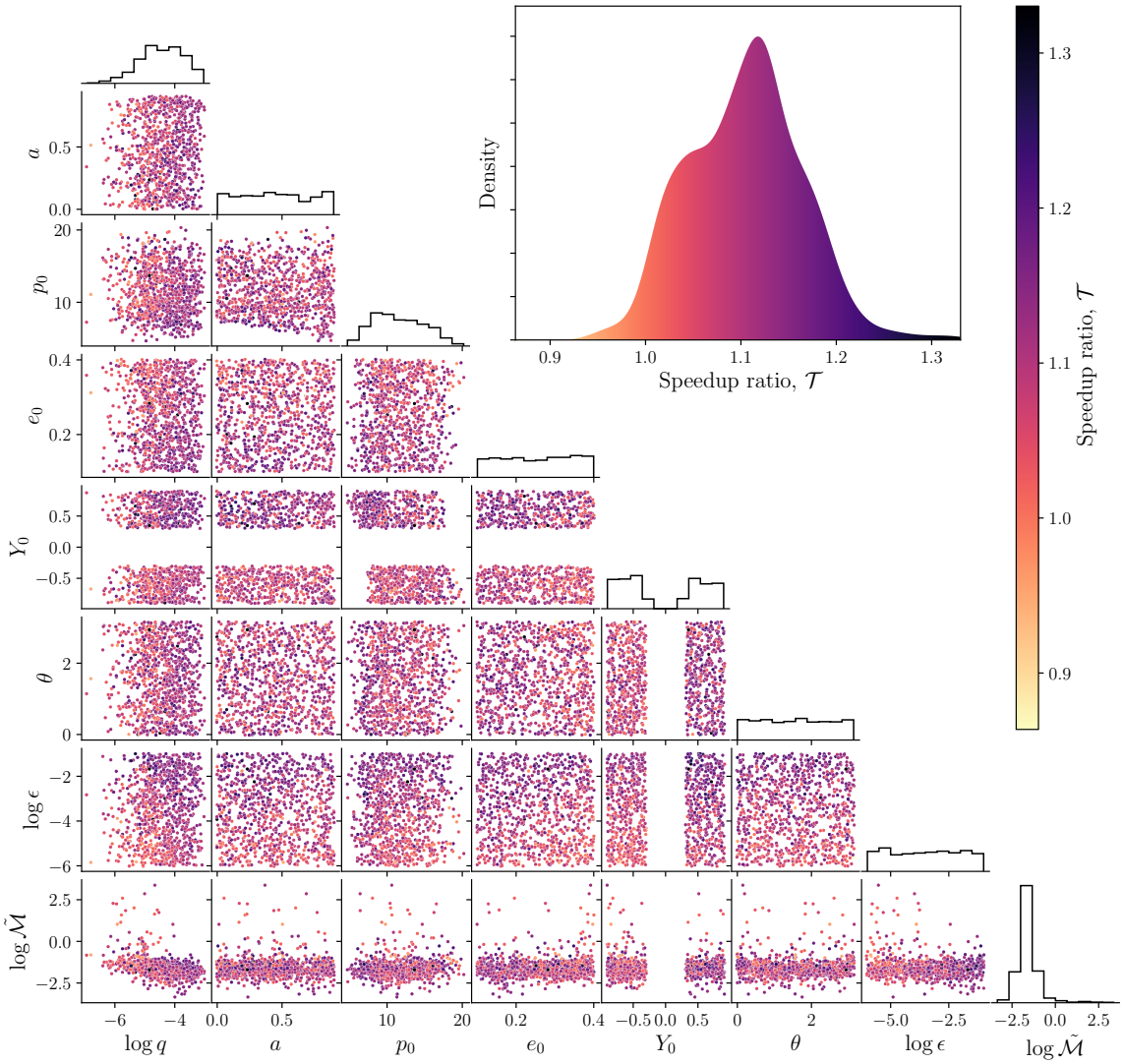


Figure 3.17: The same as Figure 3.13, but for the PN5 waveform model. As a frequency-domain summation kernel for generic Kerr waveforms has not yet been developed, only time-domain benchmarks are available. Waveforms with $|Y_0| < 0.3$ were problematic due to ODE stiffness leading to trajectories with too many sparse points for the GPU summation kernel, so this region is excised from the analysis. Similar to the results shown for KerrEqEcc, time-domain waveform generation wall-time sees an improvement of $\sim 10\%$ when using neural mode prediction. While $\tilde{\mathcal{M}} \lesssim -1$ for most of the parameter space, some outliers are present towards lower ε, which may be due to the presence of a sharp unphysical cut-off in mode selection in the PN5 model (see Section 3.2.5). Once addressed, we expect waveforms augmented with neural mode prediction to perform well in PE analyses. We expect that frequency-domain waveforms will show similar improvements in computational cost as was observed in the case of the KerrEqEcc waveform model, once generation in this domain becomes available.

strained to one part in a million (of order $1M_\odot$ for the MBH mass), with extrinsic parameters and fundamental phases being determined to 1–10% precision [176]. Naively searching for a region of parameter space of this size with respect to the prior is of similar scale to searching for a grain of sand hidden somewhere in the inner Solar System. To make matters worse, the EMRI likelihood is populated by many secondary maxima due to parameter degeneracies intrinsic to the waveform model [410]. The EMRI search problem is formidable; it has been solved for simplified scenarios and general strategies are currently under development. To make progress, we can consider the refinement stage of an EMRI parameter inference, following a global search of the parameter space which has localised the posterior bulk to within an order of magnitude. This final step might consist of the switching from an approximate likelihood or waveform model to one which is accurate, in order to place updated constraints on the parameters of the source. It is a useful testing ground for waveform models, as systematic biases due to modelling error will appear as an offset in this localised posterior distribution when PE is performed for a signal in noiseless data [177]. See Section 1.4 for an introduction to Bayesian PE for GW signals.

To characterise the performance of our neural mode predictors in PE, we inject a signal containing all modes in the data grids into a noiseless data stream and sample the posterior distribution for each waveform model. We perform four runs where ε varies from 10^{-5} to 10^{-2} in ten-fold steps for both the mode-prediction and mode-selection waveform models. For all runs, we examine frequency-domain waveforms due to their computational efficiency, neglecting the instrumental response function (and therefore the sky position of the source). As the mismatch performance of mode-prediction frequency-domain waveforms is in line with their time-domain counterparts we expect conclusions drawn from the results of these PE runs to be equally applicable to the time-domain case. All runs are performed with the `nessai` nested sampler, using 1000 live points and default settings [254, 326]. For the sensitivity curve $S_n(f)$ of the LISA detector, we use the Michelson-style model derived in [77] appropriate for use with the TT-gauge polarisations.

3.4.1 Schwarzschild eccentric inspirals

We first consider a typical EMRI source modelled by the SchwarzEcc waveform model. The intrinsic parameters of the source are $\{M, \mu, p_0, e_0\} = \{3\times10^6, 50, 9.5, 0.521\}$, with initial fundamental phases $\{\Phi_{\phi,0}, \Phi_{r,0}\} = \{1.047, 4.238\}$ We set the spin-angular momentum vector orientation parameters $\{\theta_K, \phi_K\} = \{1.047, 1.775\}$, and (as we are not modelling the detector response) arbitrarily set $\{\theta_S, \phi_S\} = \{1.047, 5.498\}$.

The results of the PE runs for this source are shown in Figure 3.18. PE runs using mode selection took roughly an order of magnitude longer than their corresponding analyses using mode prediction waveforms. Mode-prediction waveform results indicated by solid lines and mode-selection waveforms indicated by dashed lines. For $\varepsilon = 10^{-2}$, some discrepancy is apparent between mode-prediction and mode-selection waveforms, particularly in the extrinsic

parameters. However, comparing to the posteriors with lower ε, this discrepancy is of similar order to the overall bias inherent to excluding so many of the waveform modes. The extrinsic parameters are particularly sensitive to the inclusion of weaker modes in the waveform. The only parameter that shows a clear difference in posterior distribution between $\varepsilon = 10^{-3}$ and lower ε is d_{L}. Neglecting higher-order modes has been demonstrated to produce clear biases in the analysis of comparable-mass binaries [411, 412]. As removing these modes can be compensated to a degree by rescaling the amplitude of the signal, recovery of d_{L} is particularly sensitive to their inclusion. The ability to perform this rescaling is particularly apparent for cases where a given mode (ℓ, m, n) is included, but a somewhat weaker mode $(\ell - 1, m, n)$ is not: as the latter mode has identical frequency evolution to the former, its contribution to the waveform may be entirely replaced by an amplitude rescaling of the former mode. This is no longer true for a waveform with multiple modes which differ in frequency, but if the overall amplitudes of these modes are generally lower than they should be this will result in a bias in the recovery of d_{L}.

Examining the marginal posterior distribution for the orientation parameters $\{\theta_K, \phi_K\}$, it also appears to exhibit biases with $\varepsilon = 10^{-2}$ for both waveform models, which is somewhat expected by the same argument as for d_{L} because the amplitude of each mode also depends on the viewing angle with respect to the source. There is a degeneracy between these two parameters corresponding to a rotation of the source which leaves the GW signal unchanged. Focusing on one mode of the marginal posterior (inset panel, top right of Figure 3.18), we see that the biases incurred at $\varepsilon = 10^{-2}$ are severe, with the truth lying outside of the 99% region of the posterior bulk. The mode-prediction posterior for $\varepsilon = 10^{-2}$ also has an additional mode, which would not be expected from a simple amplitude-rescaling bias.

To further examine the origin of this strong bias and multimodal behaviour, we examine the posterior structure of the θ_K parameter more closely. We first set $\{\theta_S, \phi_S\} = \{0, 0\}$: due to the conventions of the detector-frame transformation described in Section 3.1.6, this choice eliminates any correlation between θ_K and ϕ_K and therefore eliminates any multimodality in θ_K. A one-dimensional slice through the likelihood with respect to θ_K (where all other parameters are set to their true values) is shown in Figure 3.19 for $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-5}$ as solid and dashed lines respectively. The waveform mismatch with respect to the injected waveform and the number of modes included in the waveform at each point in parameter space are also shown. When ε is high and the waveform contains few modes, adding or removing a single mode produces significant residuals in the likelihood due to the mismatch changing by order ε. At lower ε, adding or removing modes at the mode selection threshold does not significantly change the waveform mismatch and the likelihood is stable. This behaviour was not observed in previous examinations of the impact of mode selection on PE because these investigations focused on the intrinsic parameters of the system [177]. For these parameters, the scale of the posterior is small enough that in many cases the mode content of the waveform does not change, and no residuals are introduced. Given this fact, it may be possible to fix the mode indices included in waveform

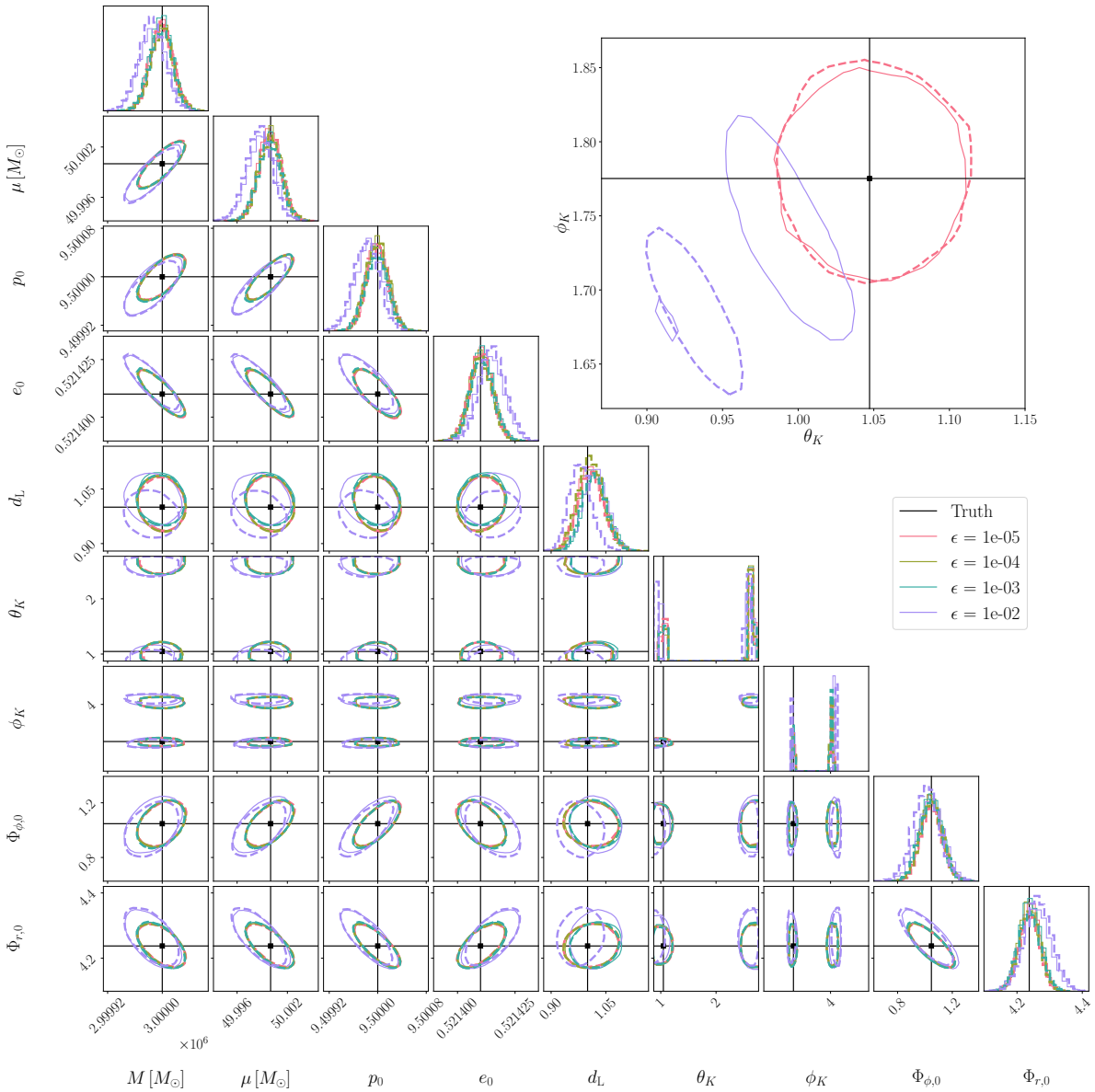
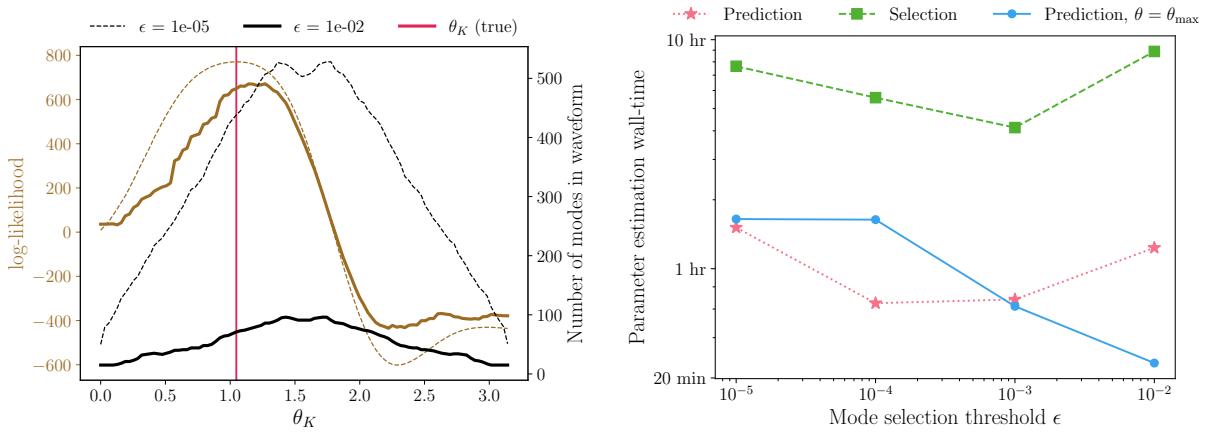


Figure 3.18: Posterior distributions obtained from PE analyses performed on an EMRI signal in a noiseless data-stream, when using either a mode-prediction (solid lines) or mode-selection (dashed lines) SchwarzEcc waveform template model and varying the mode selection threshold ε. The injected signal contains all modes included in the waveform data grids. Lines in off-diagonal panels indicate the 90% contours of the marginal posterior distribution for each parameter pair. Diagonal panels show binned one-dimensional marginal distributions for each parameter. The mode-prediction and mode-selection waveform models both show little bias for $\varepsilon \leq 10^{-3}$, with stronger biases present which differ between models for $\varepsilon = 10^{-2}$. Biases are most apparent in the extrinsic parameters; the true orientation of the source lies well outside the posterior for either model at $\varepsilon = 10^{-2}$ (inset) due to artefacts in the likelihood from mode selection. No detector response function was included in this analysis, and that the sky location of the source is taken to be fixed.



(a) (b)

Figure 3.19: (a) A one-dimensional slice of the EMRI likelihood with respect to θ_K. At higher ε, addition and removal of modes due to mode selection modifies the structure of the posterior distribution, leading to biases and sampling efficiency during PE. (b) Wall-time of SchwarzEcc PE runs for the second source considered in this section. Mode prediction (dotted line) reduces the overall cost of PE by a factor of ~ 4 with respect to mode selection (dashed line), but both waveform models show an increase in PE wall-time at higher ε due to the artefacts in the likelihood shown in panel (a). Correcting for these artefacts by setting $\theta = \theta_{\max} \sim 1.39$ (solid line) restores the reduction in computational cost at higher ε, reducing the wall-time to 23 min.

generation at this scale, eliminating the mode selection process entirely after an initial evaluation of the waveform somewhere in the prior space to further accelerate PE. However, on a larger scale (e.g., a prior an order of magnitude larger than the posterior for the intrinsic parameters) the mode content will change, which may lead to significant biases if the mode content is fixed at a particular point in parameter space which happens to be far from the posterior bulk.

An alternative approach, robust against the risks inherent to fixing the mode content of the waveform entirely, is to instead fix the mode content with respect to the viewing angle. To ensure that the waveform contains at least the number of modes required by mode selection over the entire parameter space, the mode count should be maximised over viewing angle. As the mode count does not vary with respect to ϕ, this amounts to maximising it over θ. One could instead attempt to strike a balance by sampling waveform modes in proportion to the probability that they should be retained (which is computed by the mode prediction network, as in Equation (3.24)), but this comes at the cost of having a noisy likelihood function (i.e., it can take multiple values for the same set of input parameters). As the PE analyses performed in this chapter use nested sampling, which is ill-posed for a likelihood function that is not deterministic, we opt to maximise mode content over θ and leave the investigation of alternative approaches to future work.

For equatorial waveforms, we demonstrated empirically in Figure 3.8 that the mode count peaked strongly around $\theta_{\max} \approx 1.39$. The neural mode prediction framework allows us to fold this into the waveform model with ease; when we evaluate the mode content of a prospective

waveform, we simply replace θ with $\theta_{\rm max}$. This incurs an elevated computational cost for more inclined viewing angles, as they would normally contain fewer modes, but restores stability to the posterior distribution.

We validate the effectiveness of this remedial adjustment by performing PE on a second source with identical parameters to the first, except for the change to the sky position parameters. The results of PE for both this modified mode-prediction waveform model and the standard mode-selection waveform are shown in Figure 3.20. The mode-selection waveform exhibits severe biases at $\varepsilon = 10^{-2}$, with multimodal structure present even at $\varepsilon = 10^{-3}$. The structure of the posterior is modified to such an extent that biases of roughly one standard deviation are observed in the intrinsic parameters, significantly degrading the overall recovery of the source parameters: the impact of this phenomenon on PE can therefore not be ignored even if extrinsic parameters are of no interest. The multimodality introduced by the mode selection process significantly degrades sampling efficiency, leading to a PE wall-time of almost 9 hr on 20 CPU cores. Modifying the neural mode prediction operation restores Gaussian posteriors, even for $\varepsilon = 10^{-2}$. The sampling efficiency is much higher for the mode-prediction posterior, leading to a computational wall-time of only 23 min for a PE run.

3.4.2 Kerr equatorial eccentric and generic Kerr inspirals

As the KerrEqEcc model is similarly an equatorial waveform model, we can make the same modification to the mode prediction procedure to guard against the same problem. We set the parameters of the source to be the same as the second source considered for the SchwarzEcc model, then set $a = 0.94$ and $p_0 = 5.5$ (such that the CO still plunges during the observing window). The decrease in orbital separation raises the SNR from 40 to 300; as we do not adjust the prior for these runs compared to the one used for the SchwarzEcc analyses sampling is made slightly more expensive as the prior volume must be reduced further to resolve the posterior bulk. As the luminosity distance remains set at 1 Gpc, observing a source of this SNR is a realistic prospect for the LISA mission. The results of PE with both mode-selection and neural mode prediction (maximsed over θ) waveforms are shown in Figure 3.21. The increase in the SNR of source combined with the richer mode content of the KerrEqEcc waveform model exacerbates the likelihood artefacts present due to mode selection. For $\varepsilon = 10^{-2}$, multi-modal marginal posterior distributions in $d_{\rm L}$ and θ_K significantly degrade sampling efficiency, which results in a PE wall-time of almost 10 d. As was observed for the SchwarzEcc model, maximising the neural mode prediction output over θ eliminates this problem; we instead recover somewhat biased (albeit uni-modal) marginal posteriors in these parameters for $\varepsilon = 10^{-2}$ with a PE wall-time of 3 h (a factor of ~ 75 less than what was required with a mode-selection waveform). Both mode-selection and neural mode prediction waveforms return near-identical posteriors for $\varepsilon \leq 10^{-4}$, where we observe a factor ~ 7 reduction in PE wall-time due to the improved efficiency of neural mode prediction waveforms. Given these results, we expect that our neural mode prediction

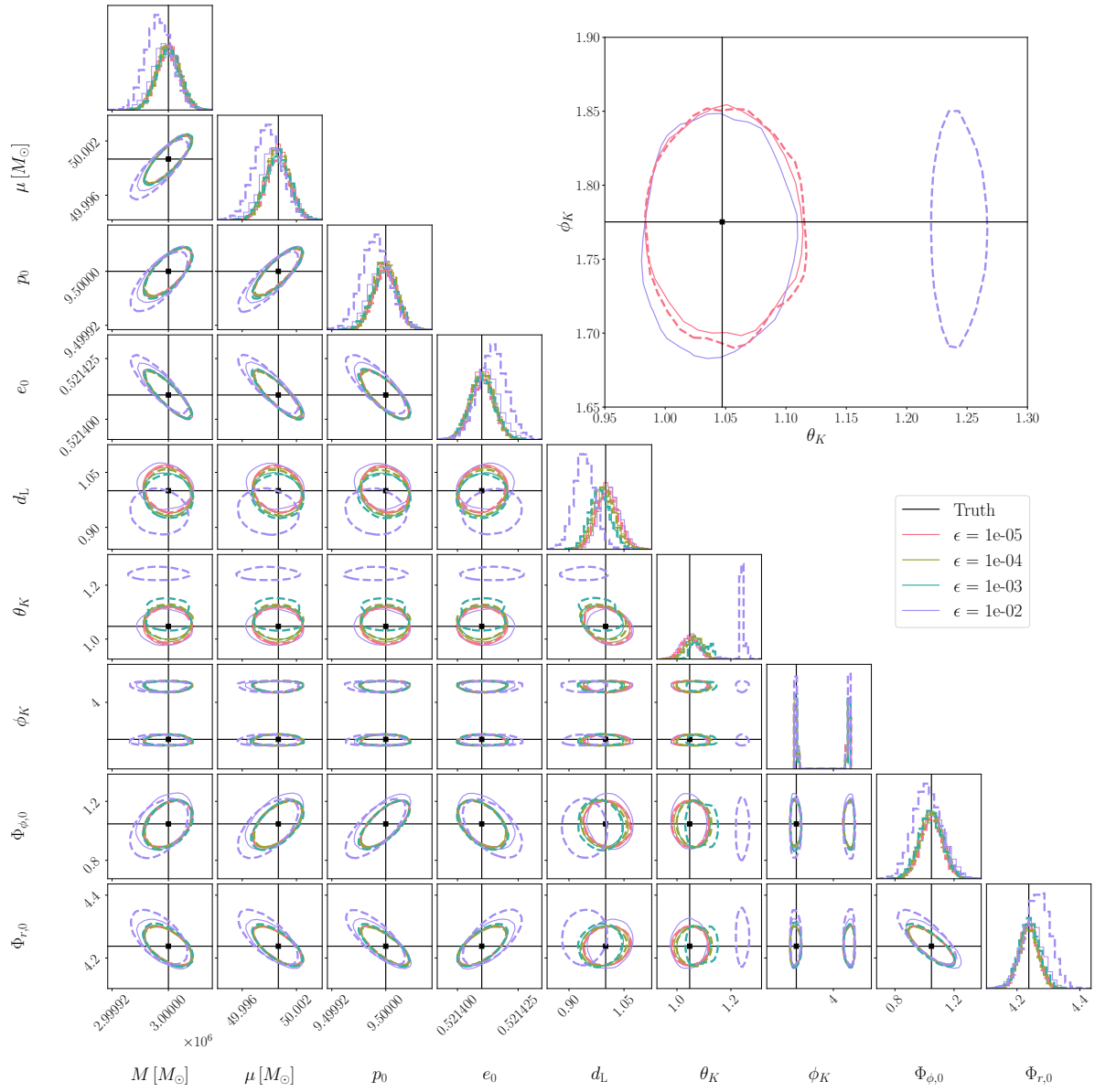


Figure 3.20: The same as Figure 3.18, but for a sky position chosen such that θ_K only has a single posterior mode. In this case, the neural mode-prediction waveform (solid lines) evaluates the mode content of the waveform at a fixed source-frame viewing angle. This eliminates the impact of mode selection on the likelihood surface, enabling accurate recovery of the source parameters with $\varepsilon = 10^{-2}$. The posterior obtained with a mode-selection waveform exhibits noticeable biases, even for $\varepsilon = 10^{-3}$.

method will significantly reduce the computational cost of both search and PE analyses for EMRI signals without significantly comprimising sensitivity or accuracy.

For generic Kerr waveform models the maximisation over θ is not appropriate, as the mode count flattens with respect to θ as the orbital inclination deviates from that of an equatorial inspiral. However, examining a one-dimensional slice through the likelihood for the PN5 model (Figure 3.22) with $Y_0 = 0.7$ and identical source parameters to the system examined above, we do not observe the same instability in the posterior distribution at $\varepsilon = 10^{-2}$. This is because generic inspirals will contain a larger number of modes for a given value of ε (due to the presence of modes with non-zero k index), so the artefacts of adding or removing individual modes are much smaller in scale than for an equatorial waveform with the same ε. We therefore expect that the maximisation over viewing angle will only be necessary to perform for quasi-equatorial inspirals (which we expect to approximately exist in nature, for instance in AGN [186]). The PN5 model is prohibitively expensive to perform multiple PE runs and validate the neural mode prediction in this regime. Once a frequency-domain prescription for this waveform model is available, it will be straightforward to examine the performance of mode-prediction waveforms in PE in the same manner as was performed for the SchwarzEcc and KerrEqEcc waveform models.

3.5 Conclusions

In this chapter, we described how EMRI waveforms are constructed in the GSF formalism in theory and their computation in practice. A significant bottleneck during this computation, which prohibits millisecond waveform-generation wall-times, is the identification and selection of dominant modes in an EMRI waveform as a function of parameter space, which currently involves an inefficient sort-and-discard operation over numerous mode amplitudes. We outlined a novel technique which addresses this inefficiency by using a MLP machine learning model to predict the mode content of EMRI waveforms based on their parameters, significantly reducing the computational cost of this stage of waveform generation. The accuracy of these MLP mode-prediction networks was examined at three levels:

1. Their prediction accuracy and fractional error with respect to validation data was computed,

2. The topology of the mode-retention surface as a function of parameter space was explored and compared both to the results of mode selection and what is expected physically from EMRI systems, and

3. The impact of mode-prediction error on parameter recovery for simulated injections of fiducial systems, as well as the resulting improvement in the computational cost of these analyses.

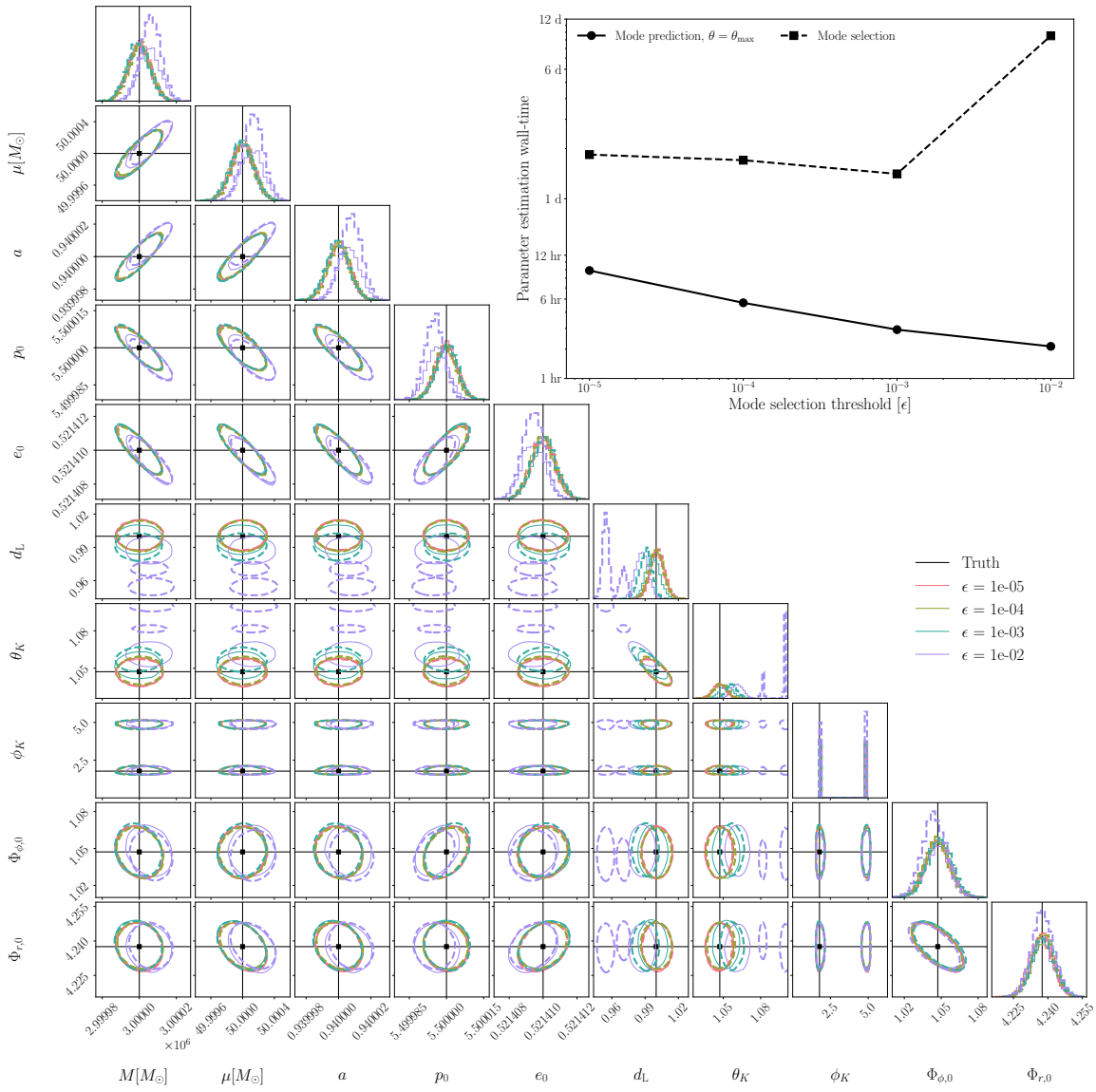


Figure 3.21: Posterior distributions obtained from PE analyses performed on an EMRI signal in a noiseless data-stream, using a mode-prediction KerrEqEcc waveform model as a template and varying the mode selection threshold ϵ. The injected signal contains all modes included in the waveform data grids. Lines in off-diagonal panels indicate the 90% contours of the marginal posterior distribution for each parameter pair. Diagonal panels show binned one-dimensional marginal distributions for each parameter. The larger SNR of the injection (300) compared to the SchwarzEcc injection (40, see Figure 3.20) leads to significant biases for $\epsilon = 10^{-2}$ which are somewhat mitigated by setting $\epsilon = 10^{-3}$. The PE wall-time (inset) improves by a factor of ~ 4 as the ϵ is increased.

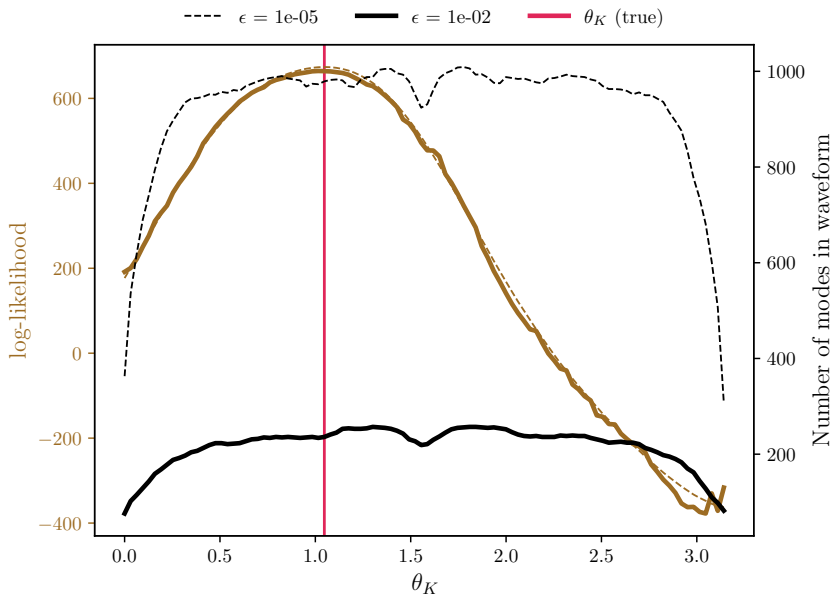


Figure 3.22: The same as panel (a) of Figure 3.19, but for the PN5 waveform model. The inclination of the inspiral ($Y_0 = 0.7$) increases the number of modes in the waveform at $\varepsilon = 10^{-2}$ sufficiently that the likelihood surface is relatively stable as a function of viewing angle. We therefore predict that maximisation of the mode content over viewing angle will not be required when analysing most generic Kerr systems.

These steps were performed for three waveform models representing the state-of-the-art in rapid EMRI waveform generation: the SchwarzEcc model for eccentric inspirals into spin-zero MBHs, as presented in the FEW package [1, 177]; the KerrEqEcc model which extends the SchwarzEcc model to spinning MBH, currently in development for integration with the FEW package; and the PN5 model, which uses PN expansions to build an approximate waveform model for generic Kerr inspirals. We examined these three waveform models to establish a baseline for the effectiveness of our method for a simple scenario, before introducing additional complexity in the form of MBH-spin and orbital inclination, with the goal of demonstrating that our approach will be applicable in the generation of generic Kerr inspirals.

The mode-prediction networks were demonstrated to be sufficiently accurate in their predictions that the complicated structure of the mode-selection function was accurately constructed on both large and small scales. The efficiency of our method reduced the computational cost of waveforms by up to an order of magnitude, with more pronounced improvements in the generation of frequency-domain waveforms on a CPU. Mismatches (normalised by the mode-selection threshold ε) between mode-prediction and mode-selection waveform models were below what we expected to be required for reliability in parameter estimation for most of the parameter space, with a few outliers present in the generic Kerr model which may be explained by unphysical cut-offs specific to the PN5 model. We demonstrated that the good normalised-mismatch performance of the mode prediction neural networks meant that waveforms built with mode-prediction results perform consistently and effectively in PE studies as expected. We also iden-

tified a previously unknown issue in the recovery of extrinsic parameters when using aggressive mode-selection thresholds, and provided a method for avoiding this problem when using neural mode prediction in waveform generation. Using mode-prediction waveforms in PE analyses reduces the computational cost by more than an order of magnitude for the SchwarzEcc model, with further improvements demonstrated for the KerrEqEcc model which are accentuated by the higher SNR of prograde EMRIs signals with spinning MBHs. Given the prediction accuracy and computational cost of the neural mode prediction network for the PN5 model, it is reasonable to expect that this result will extend to generic Kerr waveform models built in the frequency domain once they have been constructed.

Chapter 4

Rapid and accurate modelling of EMRI selection biases for population inference studies

A large fraction of Chapter 3 is devoted to (successfully) addressing the significant computational cost of EMRI waveforms. In this chapter, we turn to an application of EMRI observations for which the computational cost of their modelling is particularly relevant: population studies.

As we demonstrated in Chapter 3, the amplitude and phase evolution of EMRI is strongly dependent on the parameters of the MBH–CO system, enabling precise measurements of these parameters [190]. In particular, the (redshifted) MBH and CO masses, MBH spin and orbital eccentricity may be determined to accuracies of $\sim 10^{-3}\%$, and the sources localised in space to better than 10% relative precision [176, 342]. This level of measurement precision for MBHs surpasses both existing electromagnetic techniques [413] and LISA observations of MBH binaries [158].

The number of EMRIs that will be detected is uncertain, largely due to poorly constrained astrophysical parameters in current formation channel theories, but the detection rate is likely to be of order 10^1–10^3 yr^{-1} [80, 190, 192, 414]. During LISA's proposed 4 yr mission, we therefore expect to attain a sufficiently large catalogue of EMRIs (each of which providing excellent measurements of their parameters) to probe and resolve features of the MBH mass and spin populations. For example, we expect to match the precision of current observational estimates of an MBH mass power-law spectral index, even for pessimistic EMRI detection rate predictions [415]. Such a catalogue also enables the testing of the wide array of EMRI formation channel theories. Several EMRI formation channels have been suggested, including loss-cone scattering of COs into inspiral orbits [179], radial migration of COs by dynamical interaction with an accretion disc [192], capture via the Kozai–Lidov mechanism due to the presence of a binary MBH system [180], and the tidal disruption of main-sequence or helium stars on highly eccentric orbital trajectories around MBHs [181]. For a given formation channel, the rate of

EMRI production depends on astrophysical parameters [e.g., 191, for loss-cone or TDE channels] that are poorly constrained by existing observations. Determining the relative contributions of the formation channels to the overall EMRI rate therefore places constraints on these parameters; this could be estimated from a catalogue of EMRI observations in a similar manner to how ground-based GW observations are used to constrain their source astrophysics [e.g., 416].

Extracting information about the astrophysical population requires a hierarchical inference where the parameters of each EMRI in the catalogue (and their associated uncertainties) are collectively used to constrain the parameters of a chosen population model. However, the catalogue only contains sources loud enough to cross a detection threshold, and these are generally not representative of the underlying population: if these selection effects are ignored, the inferred population parameters will be biased. EMRIs that are intrinsically fainter (depending on their masses, spins or orbital parameters) and more distant are less likely to be detected than their nearer and louder counterparts. In practice, one may correct for this by determining the fraction of the proposed population that is detectable, and re-weighting the population likelihood accordingly [315, 417].

In the absence of a standard EMRI detection pipeline, the detectability of a given EMRI is typically assessed with respect to the SNR of the EMRI waveform [190, 415, 418]. The detectable fraction of a proposed population (the selection function) may therefore be estimated by randomly drawing EMRI events from the population and computing their SNRs to determine the fraction of these samples that are detectable. These selection function estimates may then be used to re-weight the population likelihood and obtain unbiased inferences of the population parameters.

The function that maps EMRI parameters to the waveform SNR is complicated, so bias correction is computationally expensive. This high cost comes from both the generation of complex long-duration waveforms and the manipulation of these large data sets. Even exploiting graphics processing unit GPU acceleration and vectorisation, each SNR takes of order 1 s to compute for a 4 yr data-stream duration. Using conservative estimates, if one draws 10^5 EMRIs for each candidate population, and a population inference sampling run consists of 10^5 candidate populations, full selection bias correction would require 10^{10} GPU seconds. This is too costly for analyses including selection bias correction in this manner to be conducted in a reasonable amount of time.

Previous studies have addressed the issue of computational cost by indirectly approximating the behaviour of the selection function via: proxy distance thresholds [217]; a reduction in EMRI parameter space complexity, for instance by neglecting the dependence of eccentricity or inclination on detectability [415]; a reduction of waveform complexity by only computing a small number of sideband modes with faster, less accurate waveform models [410]. These approaches permit the rapid computation of the selection function, but do not account for more complex correlations between EMRI parameters and may introduce systematic biases due to the

approximations made. For instance, the evolution of orbital eccentricity and inclination during an inspiral is correlated with mass ratio, and the mode amplitudes (and therefore, the overall SNR of the waveform) are correlated with both eccentricity and inclination evolution. Therefore, even for parameters not directly of interest to a given population study, the correlations between these parameters and event detectability must still be taken into account to avoid biases in the results obtained.

In this chapter, we describe an alternative approach that leverages the speed of FEW [177] (which we highlighted in Chapter 3) and machine-learning techniques to interpolate the EMRI SNR function, directly correcting for an SNR-based selection bias without the need for major systematic approximations or simplifications. In Section 4.1 we outline the Bayesian population inference framework employed, including the correction for selection biases. Models for the EMRI population and selection effects are defined in Section 4.2. We then describe our methodology for simulating catalogues of EMRI PE results in Section 4.3. In Section 4.4 we introduce our approach for estimating the selection bias with machine learning, which we achieve by replacing the bottleneck in the selection bias calculation (the EMRI SNR function) with a neural network interpolator. The effectiveness of our framework is demonstrated in Section 4.5 for a typical EMRI population, presenting clearly the manifestation of the selection bias in the obtained results and how this is corrected for in practice; the corrected results provide an unbiased estimate for the precision to which LISA observations could constrain the astrophysical EMRI population. Finally, in Section 4.6, we perform a global posterior consistency check to verify the analysis. Our conclusions are presented in Section 4.7.

The majority of this chapter has been published as a peer-reviewed paper [4], with the method implemented in the open-source code package POPLAR [288]. Section 4.A contains unpublished material appearing for the first time in this thesis.

4.1 Hierarchical Bayesian inference framework

Our goal is to infer the properties of an EMRI population model using a catalogue of many EMRI observations.

For each EMRI in the catalogue, information about its parameters $\vec{\theta}_i$ is encoded in the data, $\vec{d}$, where the subscript $i \in [1, N_{\rm obs}]$ identifies the particular EMRI in the catalogue of $N_{\rm obs}$ detections. The posterior distribution for $\vec{\theta}_i$ given the data is

$$p(\vec{\theta}_i|\vec{d}) = \frac{\pi(\vec{\theta}_i)\mathcal{L}(\vec{d}|\vec{\theta}_i)}{\mathcal{Z}(\vec{d})}, \tag{4.1}$$

where $\pi(\vec{\theta}_i)$ is the prior distribution on $\vec{\theta}_i$, $\mathcal{L}(\vec{d}|\vec{\theta}_i)$ is the likelihood of observing the data given a set of source parameters, and $\mathcal{Z}(\vec{d})$ is the evidence (marginalised likelihood). We estimate the parameters of the EMRI by stochastically sampling the posterior distribution, obtaining a set of

posterior samples $\{^k\theta_i\}$ [419], where the superscript $k \in [1, S_i]$ denotes each posterior sample for a given event. The posterior $p(\vec{\theta}_i|\vec{d})$ provides information about a single EMRI source; by combining the properties of the catalogue of sources, we can constrain a population model.

The population model $p_{\rm pop}(\vec{\theta}|\vec{\lambda})$ describes the astrophysical distribution of EMRI source parameters. It is described by a set of hyperparameters $\vec{\lambda}$ that determine the shape of the population, and a Poissonian mean event rate $\mathcal{R}$ that parameterises how often EMRIs occur. We use the *hyper* prefix to differentiate these population-level (hyper)parameters from the event-level EMRI parameters. By estimating the hyperparameters, we constrain the relative probabilities of different population shapes and event rates in accordance with the contents of the catalogue. We perform this hyperparameter estimation in a hierarchical Bayesian inference framework [315].

To obtain an estimate of $\vec{\lambda}$ and $\mathcal{R}$, we sample the hyperparameter posterior distribution

$$p(\vec{\lambda}, \mathcal{R}|\{\vec{\theta}\}) = \frac{\pi(\vec{\lambda})\pi(\mathcal{R})\mathcal{L}(\{\vec{\theta}\}|\vec{\lambda})\mathcal{L}(\{\vec{\theta}\}|\mathcal{R})}{\mathcal{Z}(\{\vec{\theta}\})}, \tag{4.2}$$

where $\pi(\vec{\lambda})$ and $\pi(\mathcal{R})$ are hyperprior distributions, and $\mathcal{Z}(\{\vec{\theta}\})$ is the hyperevidence. The hyperparameter likelihood $\mathcal{L}(\{\vec{\theta}\}|\vec{\lambda})$ is defined as

$$\mathcal{L}(\{\vec{\theta}\}|\vec{\lambda}) = \prod_{i=1}^{N_{\rm obs}} \frac{1}{S_i \alpha(\vec{\lambda})} \sum_{j=1}^{S_i} \frac{p_{\rm pop}({}^j\vec{\theta}_i|\vec{\lambda})}{\pi(\vec{\theta}_i)}, \tag{4.3}$$

where. in general, the population probability of each posterior sample must be re-weighted by the prior used in the EMRI parameter estimation step [315]. In our case, we adopt uniform priors on all EMRI parameters and this re-weighting simplifies to a proportionality constant. The rate likelihood $\mathcal{L}(\{\vec{\theta}\}|\mathcal{R})$ is

$$\mathcal{L}(\{\vec{\theta}\}|\mathcal{R}) = \exp\left[-\mathcal{R}\alpha(\vec{\lambda})\right] \left[\mathcal{R}\alpha(\vec{\lambda})\right]^{N_{\rm obs}}. \tag{4.4}$$

Here, the selection function $\alpha(\vec{\lambda})$ is a corrective factor applied to account for the presence of selection bias on the observations: of the (unknown) number of events that occurred, only a subset $N_{\rm obs}$ were detected. It may be written as

$$\alpha(\vec{\lambda}) = \int p_{\rm det}(\vec{\theta}) p_{\rm pop}(\vec{\theta}|\vec{\lambda})\, {\rm d}\theta, \tag{4.5}$$

for some detection probability $p_{\rm det}(\vec{\theta})$, and represents the fraction of a population (described by a particular set of hyperparameters $\vec{\lambda}$) that is detectable. Performing an inference including the selection effects should produce results unbiased by detectability [315].

To make use of this inference framework, one must specify:

1. The form of $p_{\rm pop}(\vec{\theta}|\vec{\lambda})$;

2. The selection effects for the observations;

3. A method for obtaining $\{\vec{\theta}\}$.

Once these three ingredients have been formally defined, we can construct (and ultimately, sample) the hyperposterior, Equation (4.2). We detail the form of our population model in Section 4.2, outlining our treatment of detectability and the modelling of selection effects in Section 4.2.1. We then introduce the approach we use to simulate EMRI catalogues in Section 4.3, describing the method by which we obtain posterior samples for each detected EMRI in Section 4.3.2.

4.2 Population model

For simplicity, we choose a population model that is a product of independent univariate subpopulations, such that

$$p_{\rm pop}(\vec{\theta}|\vec{\lambda}) = \prod_{x\in\vec{\theta}} p_x(x|\vec{\lambda}_x), \tag{4.6}$$

where x denote EMRI parameters and $\vec{\lambda}_x$ the corresponding hyperparameters that describe the shape of the subpopulation. The mathematical form of these subpopulations is summarised in Table 4.1, and in detail:

- Mass functions for both MBHs and stellar-mass black holes are well approximated by power laws, albeit with additional substructure present when examined in detail [193, 420]. We therefore model the MBH and CO mass distributions $p_M(M|\lambda_M, M_{\rm min}, M_{\rm max})$ and $p_\mu(\mu|\lambda_\mu, \mu_{\rm min}, \mu_{\rm max})$ as power laws, with index λ_x and limits $[x_{\rm min}, x_{\rm max}]$, which have the form

$$p_x(x|\lambda_x, x_{\rm min}, x_{\rm max}) = \frac{1+\lambda_x}{x_{\rm max}^{1+\lambda_x} - x_{\rm min}^{1+\lambda_x}} x^{\lambda_x}. \tag{4.7}$$

- The form of the MBH spin magnitude distribution $p_a(a|\mu_a, \sigma_a)$ is dependent on a number of astrophysical processes during the formation and evolution of MBHs and their host galaxies [197, 421]. Incorporating these into our population model and characterising their impact on inference results is beyond the scope of this study. For simplicity, we instead choose a truncated normal distribution with mean μ_a and variance σ_a^2 as has been done in previous analyses of the stellar-mass BBH mergers [377, 422]. This is written as

$$p_x(x|\mu_x, \sigma_x) = \frac{1}{\sigma} \frac{\psi\left[(x-\mu_x)/\sigma_x\right]}{\Psi\left[(B-\mu_x)/\sigma_x\right] - \Psi\left[(A-\mu_x)/\sigma_x\right]}, \tag{4.8}$$

  where $\psi(x)$ and $\Psi(x)$ are the PDFs and CDFs of the standard normal distribution, respectively. The limits $[A, B]$ are chosen to be $[0.001, 0.999]$ as waveform generation is unstable at extremal spins beyond these limits.

- High initial orbital eccentricities (> 0.99) are expected for EMRIs formed by relaxation mechanisms, but significant orbital eccentricity will be lost before the GW emission of the system enters the LISA band [16], broadening the distribution and shifting it to lower eccentricities [423]. To reflect this behaviour, we choose a uniform eccentricity distribution $e_0 \in [0.1, 0.5]$, with upper limit chosen to reflect that the waveform model is a series expansion in eccentricity and should therefore not be trusted for high eccentricities [3, 360]. The waveform model also consists of a system of ODEs that must be solved [177]. The lower limit of the eccentricity distribution is chosen due to increasing stiffness in this ODE system at lower eccentricities leading to high computational cost [424]. We do not anticipate the validity of our approach to be affected by this lower eccentricity cut-off.

- Orbital inclination ι_0 is similarly truncated due to ODE stiffness issues, but is otherwise distributed uniformly on the unit sphere along with other angular parameters.

- We choose a redshift distribution that is uniform in comoving volume and in comoving time [425]; this has the form

$$p_z(z) \propto \frac{1}{(1+z)E(z)} \left(\int_0^z \frac{\mathrm{d}z'}{E(z')} \right)^2, \tag{4.9}$$

  where

$$E(z) = \sqrt{\Omega_M (1+z)^3 + \Omega_\Lambda}, \tag{4.10}$$

  and we assume a standard cosmology with $\Omega_M = 0.3$ and $\Omega_\Lambda = 1 - \Omega_M = 0.7$. The upper redshift limit for this distribution is chosen to be $z = 6$ such that the detectable region of parameter space is not significantly truncated (otherwise, selection effects will be artificially suppressed). Increasing the redshift limit leads to high computational costs as the event rate $\mathcal{R}$ must also be increased accordingly, as the event rate density has remained constant but the comoving volume over which we are distributing events has grown.

The chosen form of these subpopulations is motivated primarily by computational simplicity. However, our approach is flexible and can be applied to the hierarchical inference of any population model.

4.2.1 Selection effects

In the absence of a specific EMRI search pipeline, we model the detection probability as a binary SNR threshold, as is typical for EMRI studies [190, 226, 415]. This may be written as

$$p_{\mathrm{det}}(\vec{\theta}) = \mathcal{H}(\rho_{\mathrm{n}} - \rho_{\mathrm{t}}), \tag{4.11}$$

Table 4.1: The functions, free parameters and limits of the sub-population distributions $p_x(x|\vec{\lambda}_x)$, the product of which is the EMRI population chosen. The hyperparameters $\vec{\lambda}_x$ are estimated via population inference. The upper limit for $t_{\rm plunge}$ is reduced to 2 yr for our validation analysis (Section 4.6).

x	$p_x(x\|\vec{\lambda}_x)$	$\vec{\lambda}_x$	$[x_{\rm min}, x_{\rm max}]$
M	Power-law	λ_M, $M_{\rm min}$, $M_{\rm max}$	$[M_{\rm min}, M_{\rm max}]$
μ	Power-law	λ_μ, $\mu_{\rm min}$, $\mu_{\rm max}$	$[\mu_{\rm min}, \mu_{\rm max}]$
a	Trunc. Normal	μ_a,σ_a^2	$[0.001, 0.999]$
e_0	Uniform	—	[0.1, 0.5]
$\cos\iota_0$	Uniform	—	$[0, \pi/3]$
$\sin\theta_S$	Uniform	—	$[0, \pi]$
$\sin\theta_K$	Uniform	—	$[0, \pi]$
$\Delta\phi$	Uniform	—	$[0, 2\pi]$
$t_{\rm plunge}$	Uniform	—	$[0, (2, 10)]$ yr
z	$p_z(z)$	—	$[0, 6]$

where $\rho_{\rm n}$ is a (noise-realized) SNR, $\mathcal{H}(x)$ is the Heaviside step function and $\rho_{\rm t}$ is a chosen threshold SNR. We obtain $\rho_{\rm n}^2$ by drawing a sample from a non-central χ^2 distribution with two degrees of freedom and non-centrality parameter $\rho_{\rm opt}^2$ [23],

$$p\left(\rho_{\rm n}^2|\rho_{\rm opt}^2\right) = \frac{1}{2}\exp\left(-\frac{\rho_{\rm n}^2+\rho_{\rm opt}^2}{2}\right) I_0\left(\rho_{\rm n}\rho_{\rm opt}\right), \tag{4.12}$$

where $I_0(x)$ is a modified Bessel function of the first kind [426], and $\rho_{\rm opt}^2$ is the square of the optimal matched-filter SNR described by Equation (1.19). We assume that $\rho_{\rm n}$ is the positive square root of $\rho_{\rm n}^2$; while noise fluctuations can lead to negative values, this is not expected for large values around our detection threshold. One may analytically compute the mean detection probability in Equation (4.11) over all noise realizations by directly computing the non-central χ^2 CDF $p(\rho_{\rm n}^2 > \rho_{\rm t}^2|\rho_{\rm opt}^2)$, such that

$$\overline{p_{\rm det}}(\vec{\theta}) = 1 - p\left(\rho_{\rm n}^2 > \rho_{\rm t}^2|\rho_{\rm opt}^2\right), \tag{4.13}$$

where an overline denotes the mean.

We approximate the selection function Equation (4.5) by evaluating the Monte Carlo sum

$$\alpha(\vec{\lambda}) \approx \frac{1}{N_{\rm t}}\sum_{k=0}^{N_{\rm t}} \overline{p_{\rm det}}(\vec{\theta}_k), \tag{4.14}$$

where $\{\vec{\theta}_k\}$ are sampled from $p_{\rm pop}(\vec{\theta}|\vec{\lambda})$. As the variance on this approximation scales inversely with $N_{\rm t}$, one must compute $\overline{p_{\rm det}}(\vec{\theta})$ (and therefore $\rho_{\rm opt}$) on the order of 10^5 times for each computation of $\alpha(\vec{\lambda})$ to achieve percent-level accuracy; even with parallelisation, this would

be prohibitively expensive with typical computing resources (taking on the order of minutes) for use in a typical sampling run, in which $\alpha(\vec{\lambda})$ must be computed once per hyperlikelihood call. We address this problem by replacing the SNR function with an accurate and rapid interpolator, allowing for Equation (4.14) to be evaluated in parallel at a sufficiently low computational cost to be practical for use in inference problems.

4.3 Simulating catalogues of EMRI observations

The generation of a catalogue of EMRI observations is multi-faceted. We first describe the generation of EMRI waveforms, including our treatment of initial conditions and our choices regarding waveform model and detector response, in Section 4.3.1. In Section 4.3.2, we outline our procedure for obtaining posterior samples for EMRIs that pass the detection threshold, along with consistency checks we perform to ensure the approximations made in this process are justifiable. These two steps are performed for each set of EMRI parameters drawn from the population to construct a catalogue of EMRI detections and their corresponding posterior samples.

4.3.1 Parametric conventions and waveform generation

The EMRI parameter space is complicated, consisting of 17 source parameters and an additional parameter describing the epoch they are defined with respect to. While these were previously discussed in Section 3.1, we re-define them here for clarity:

- The intrinsic parameters of the EMRI describe the properties of the two objects and their initial orbital configuration. The primary MBH is described by its mass M and spin vector $\vec{a}$, and the secondary CO similarly by mass μ and spin vector $\vec{a}_{\rm CO}$. As $\vec{a}_{\rm CO}$ is not predicted to have a significant effect on EMRI detectability [387, 427], it is not currently included in state-of-the-art waveform models, including the model used in this study [177]. The inspiral orbit is described by the initial eccentricity e_0, semi-latus rectum p_0 and orbital inclination ι_0, along with three orbital phases Φ_r, Φ_θ and Φ_ϕ [360].

- The extrinsic parameters describe the orientation of the system and its location with respect to the detector. The position of the system is described by the luminosity distance vector $\vec{d}_{\rm L}$, the magnitude of which may be described by a redshift z via the relation

$$d_{\rm L} = (1+z)\frac{c}{H_0}\int_0^z \frac{{\rm d}z'}{E(z')}, \tag{4.15}$$

  where H_0 is the Hubble constant [425]. We separate out vectors into their magnitudes and angular components, such that $\vec{d}_{\rm L} = \{d_{\rm L}, \theta_{\rm S}, \phi_{\rm S}\}$. Similarly, we decompose the primary spin magnitude vector $\vec{a} = \{a, \theta_{\rm K}, \phi_{\rm K}\}$. The sets of angles describe the orientation of the

MBH spin-angular momentum vector and the sky position vector respectively, with θ and ϕ referring to polar and azimuthal angles respectively.

- The plunge time of the system with respect to the start of the observation is described by the parameter t_{plunge}. However, the concept of initial conditions on a population level is not well-defined unless a common reference point in the waveform is set. Therefore, we adjust p_0 such that the EMRI waveform will plunge after 10 yr [373]. For our fiducial example outlined in Section 4.5, we assume a LISA observing window of 4 yr in line with the current mission proposal [428]. Our choice to allow for EMRI plunges to occur up to 10 yr after the beginning of LISA observation is made to accommodate the presence of EMRIs in the data that plunge after the end of the observation window. Neglecting these events as being undetectable is not typically a reasonable approximation. In many cases, $\rho_{\text{opt}} > \rho_{\text{t}}$ even for $t_{\text{plunge}} = 10$ yr, so some detectable events in the data are ignored despite this extension of EMRI plunge times to the post-window regime. By excluding these events that plunge after 10 yr, the results of our population inference will be conservative. We include this effect to demonstrate that our approach is capable of accommodating a post-window cut on t_{plunge}, but acknowledge that the tuning of such a cut-off point with respect to the detectability of the excluded signals is an issue that warrants further investigation in future work.

Waveform generation also includes some additional considerations to transform from the source frame to the detector frame. Prior to waveform generation, we convert source-frame masses to detector-frame masses with the mapping $M_{\text{det}} = (1+z)M$ [429]. When we performed the work described in this chapter, the FASTLISARESPONSE code applied extensively in Section 2.3.3 was not publicly available. We therefore do not include a detector response in our EMRI waveform modelling and instead work in terms of the waveform strain (as opposed to the TDI combinations that the LISA detector outputs will be used to construct [59]). This choice leads to a degeneracy between ϕ_{K} and ϕ_{S}, which we navigate by defining a new parameter $\Delta\phi = \phi_{\text{S}} - \phi_{\text{K}}$. Our approach can incorporate a chosen LISA response by reverting back to the separate angles and including the response function in waveform generation. As the addition of a response function does not significantly alter EMRI SNRs, we do not expect its exclusion to affect the validity of our approach.

For our waveform model, we choose the fifth-order PN AAK implemented in the FEW package [177, 313]. The validity of our population inference framework should not depend strongly on this choice, as the EMRI SNR function should remain well-behaved and smooth for any reasonable choice of waveform model, although the specific numerical results may vary for different waveforms.

To accommodate our choice of initial conditions, we generate 10 yr EMRI waveforms in the time domain with a sampling rate of 0.1 Hz, and crop them according to their (randomly sampled) t_{plunge} values. We calculate the waveform's ρ_{opt} value via Equation (1.19) and produce

a noise-realized SNR estimate ρ_n by drawing a sample from Equation (4.12). Detection is evaluated via Equation (4.11).

4.3.2 Parameter estimation

For waveforms that pass the detection threshold, we proceed to draw samples from the posterior distribution on $\vec{\theta}$. Despite recent reductions in EMRI waveform computation time to the sub-second level [177], standard Bayesian parameter estimation techniques are too costly for the event posteriors to be sampled directly en-masse as is required in population studies. We instead opt to approximate the EMRI likelihood (and by extension, posterior) with a Fisher matrix approach, operating under the linear signal approximation (LSA) [397], in which the likelihood is approximated by a multivariate normal distribution,

$$p(d|\vec{\theta}) \approx \mathcal{N}\left(\vec{\theta}, \Gamma^{-1}\right), \tag{4.16}$$

with mean equal to the vector of source parameters $\vec{\theta}$ and covariance matrix equal to the inverse of the Fisher information matrix (FIM) Γ of the EMRI waveform. The LSA is only valid in the high-SNR limit (which may lie far above the detection threshold), which should be verified before it is used to approximate likelihoods [430]. The FIM is given by

$$\Gamma_{\ell m} = \langle \partial_\ell h | \partial_m h \rangle, \tag{4.17}$$

where $\partial_\ell h$ refers to the derivative of the waveform strain $h(t)$ with respect to the ℓ-th parameter of $\vec{\theta}$, evaluated at $\vec{\theta}$. The noise-weighted inner product between time-domain data streams $\langle x|y\rangle$ is defined by Equation (1.18). We adopt the analytic fit to the LISA PSD derived in [77]. Using the FIM, we can rapidly produce posterior distributions for a catalogue of EMRIs.

Operating under the LSA, one can determine the FIM Γ from waveform derivatives computed using finite-differencing methods. Due to the large condition number of EMRI FIMs, it is essential that the waveform derivatives are computed as accurately as possible. We therefore choose to use a nine-point stencil central difference, accurate to $\mathcal{O}(\delta^8)$, where δ is the step size between points in the stencil (and typically takes different values for each parameter) [431, 432]. The accuracy of this finite difference relies strongly on the choice of step-size used for each parameter: if δ is too large, the derivative will be inaccurate, but too small a step size will be unstable due to floating-point rounding errors. We address this problem by computing numerical derivatives on a grid of δ values and identifying step-size regions where the derivative is stable (and assumed to be close to the true value of the derivative). Once identified, we assume that these δ values will hold for all systems after linearly rescaling them according to the parameters (such that their proportionality to these parameters is fixed). We found this assumption to be sufficiently accurate across the parameter space. During those investigations, we identified

two issues that significantly impacted FIM accuracy and made the necessary adjustments to our FIM generation framework to correct them. Further discussion of these challenges encountered during FIM computation can be found in Section 4.A.

Given tuned step-sizes for all parameters, we calculate waveform derivatives and obtain Γ via Equation (4.17). With Γ computed, we then invert it to obtain the covariance matrix of the LSA likelihood. FIM for EMRI waveforms typically have large condition numbers, which can cause issues when performing matrix inversion; we mitigate this by employing singular value decomposition (SVD) to compute the pseudoinverse of Γ [433]. The numerical stability of this inversion can also be problematic, even for double precision; to alleviate this, we perform the SVD with 500-point decimal precision using the MPMATH package [434]. We also perform some additional verification of the validity of the LSA likelihood as a substitute for the full likelihood as recommended in [430]. This check is performed by examining the ratio between the LSA and full likelihood, which we denote r. By sampling the 1 standard deviation isoprobability contour in the LSA likelihood and computing $|\log_{10} r|$ for each of these samples, we can examine whether the high-probability regions of the LSA and the full likelihood are consistent. Following [430], we accept the LSA likelihood as a suitable approximation if and only if 90% of the samples satisfy $|\log_{10} r| < 0.1$. We found that for waveforms with $\rho_{\rm opt} > 10$ that this condition was satisfied in all cases, which is well below our chosen SNR threshold of 20.

Our goal is to obtain posterior samples via the LSA that are representative of those that would be obtained from sampling the true posterior. In the absence of noise, the likelihood will peak on the true values provided there are no degeneracies in the parameter space. For EMRIs, this is not generally true, but the degeneracies are non-local and therefore do not affect the morphology of the posterior near the true values [410]. However, when noise is included, the likelihood shifts in a random direction in parameter space such that the true values are no longer at the maximum likelihood point. To simulate this measurement effect, we draw one sample from the posterior (which, assuming broad uniform priors, is equivalent to the event likelihood) and recompute the FIM at this new point [435]. Finally, with our second FIM computed, we invert it and draw $S = 10^4$ samples from the corresponding multivariate normal distribution. The posterior samples in $d_{\rm L}$ are converted to samples in z assuming a standard cosmology, and detector-frame masses then converted back to the source frame by dividing through by $(1+z)$ on a per-sample basis. This modified set of posterior samples is the final product of the individual EMRI event simulation.

4.4 Interpolating over signal-to-noise ratio

The principal requirements for our SNR interpolator are that it must be accurate and unbiased across the EMRI parameter space: inaccuracies may bias the results of our population inference. It must also be sufficiently fast as to not bottleneck the sampling process, capable of estimating

SNRs for 10^5 sets of EMRI parameters in < 1 s. These constraints are particularly challenging to meet due to the high dimensionality of the EMRI parameter space (13 dimensions, as defined in Section 4.3).

Fortunately we can reduce the number of parameters that we need to interpolate over by considering how the SNR of an EMRI waveform depends on each parameter. We can ignore the orbital phase parameters $(\Phi_r, \Phi_\theta, \Phi_\phi)$ due to their negligible correlation with SNR, as the initial phase becomes relatively unimportant for an inspiral with $\sim 10^4$ orbital cycles. Additionally, as SNR scales inversely with luminosity distance $d_{\rm L}$, we may further reduce the dimensionality of the parameter space by fixing $d_{\rm L}$ in training data and applying this scaling post-interpolation: we use $d_{\rm L} = 1$ Gpc for convenience.

Despite eliminating four dimensions of the EMRI parameter space, we are still in a regime where standard interpolation schemes are ineffective. As a representative example, we consider spline interpolation schemes with piecewise polynomials of zeroth, first and third order: these are more commonly known as nearest neighbour, linear and cubic spline interpolation respectively [436]. Our requirement for a fast interpolator prevents us from interpolating over points randomly distributed in the parameter space, as the complexity of the algorithms used for this scales quadratically with the number of basis points [437] and the computational cost of these methods quickly becomes impractical. Instead, we may use grid-based versions of these techniques. However, these methods suffer from the curse of dimensionality: the Euclidean distances between neighbouring grid verticies grows as the dimensionality of the space increases, which leads to poor interpolation accuracy.

To demonstrate the unsuitability of linear interpolation in practice, we generate $\rho_{\rm opt}$ on a regular grid with 10^6 total grid points and construct the aforementioned spline interpolators with this grid as a basis. We then compute a testing set of 10^6 SNRs from randomly sampled sets of EMRI parameters and compare the interpolator output at these points by calculating the discrepancies between the true and predicted SNRs, denoted $\rho_{\rm true}$ and $\rho_{\rm pred}$, respectively. The cumulative distribution of the (i) absolute and (ii) relative differences between prediction and truth for the three spline interpolation schemes we consider are shown in Figure 4.1. As expected, the grid-based interpolator performance is poor regardless of the order of the piecewise polynomial used. While linear or spline interpolation offers marginal improvement over nearest neighbour interpolation, the low spatial resolution of the grid limits the improvements. The majority of the interpolated SNRs are inaccurate by at least 50%, with absolute errors typically exceeding 10 (or even as high as 100 in extreme cases). As we will demonstrate in Section 4.6, this performance is inadequate for unbiased population inference.

4.4.1 Interpolation with neural networks

Neural networks, are highly flexible mathematical tools that are capable of learning complex relationships in high-dimensional spaces [264]. For our purposes, we need a neural network that

Table 4.2: Architecture and training settings chosen for our two MLP neural networks introduced in Section 4.4.

Setting	SNR MLP	$\alpha(\lambda)$ MLP
Number of (hidden) layers	10	8
Neurons per layer	128	128
Activation function	SiLU	SiLU
Rescaling	Unit normal	Unit normal
Optimiser	Adam	Adam
Learning rate	5×10^{-4}	5×10^{-4}
Batch size	10^4	10^5
Max epochs	10^5	10^3
Loss function	L1	L1

takes a vector as an input (the EMRI parameters) and produces a scalar output (the SNR estimate). We opt for the MLP algorithm [438], which we introduced and described in Section 1.5, as it fits this specification. MLPs are fast and capable of high accuracy, satisfying our requirements well. The number of required neurons and layers (which describe the dimensions of each matrix multiplication) in an MLP depends on the complexity of the function to be interpolated and the number of interpolation dimensions. Due to the stochastic nature of training neural networks, tuning of the learning rate, batch size and number of iterations employed during training is required to maximise performance. For complicated problems, optimization techniques may be employed to explore the space of network settings and identify a sensible configuration [439]. In our case, the problem is sufficiently low-dimensional that we were able to obtain effective MLPs through the manual tuning of network settings. Network complexity was gradually increased through the addition of neurons or layers until overfitting was observed [264]. This is characterised by the performance of the network on testing data degrading despite continued improvement in performance on training data. At this point, training settings were adjusted to minimise this overfitting. The choice to rescale training data to that of a unit normal distribution, and to use the Adam optimisation algorithm [285], were made following current best practices [264]: use of other optimisation or rescaling functions was not found to significantly affect network performance. The resulting network settings chosen for the SNR MLP are given in the second column of Table 4.2.

The trained network achieves two orders of magnitude of improvement in accuracy compared to other interpolation approaches, as shown in Figure 4.1; the majority of the test data is predicted to percent-level accuracy. This network is capable of producing 10^5 SNR estimates in < 0.1 s, which is six orders of magnitude faster than calculating the SNR directly.

As the MLP was trained with an L1 loss function, which minimises the absolute difference between the prediction and truth [264], it does not perform as well in terms of fractional error for $\rho_{\text{opt}} \ll 1$. This manifests as a larger upper tail in the relative CDF. However, this does not translate to a reduction in performance, as these signals are too weak to be detectable across the

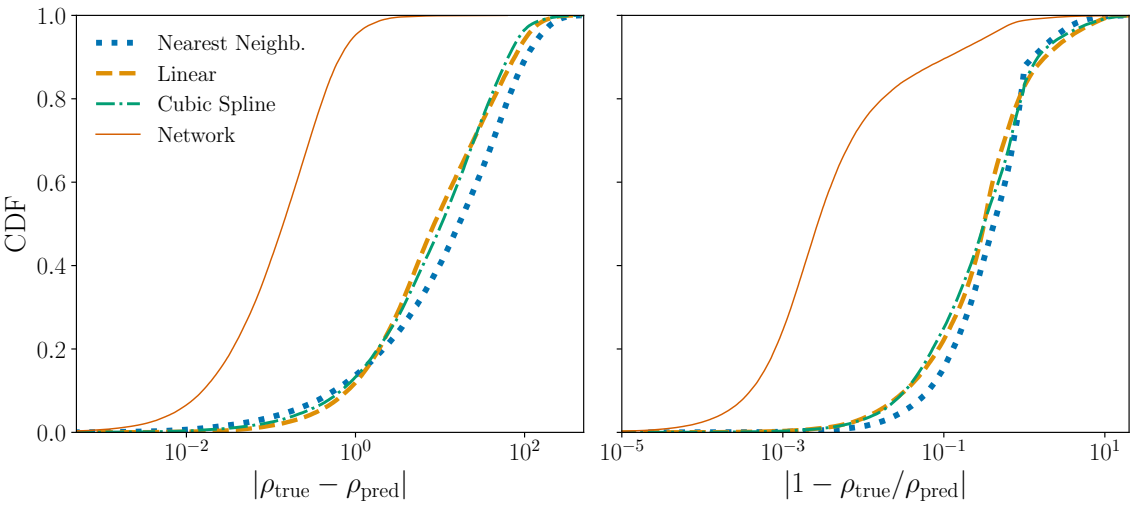


Figure 4.1: Cumulative distribution functions for the (*left*) absolute and (*right*) fractional differences between the SNR predictions of (linear, nearest neighbour, cubic spline and neural network) interpolation methods and the true values. The former three grid-based interpolation methods perform poorly, with at least 47% of SNRs inaccurate by at least 10, and at least 75% predicted to less than 10% accuracy. Conversely, the latter method precisely estimates SNRs: 95.3% of SNRs are determined to within 1, and 89.6% within 10% of the true values.

majority of the luminosity distance distribution. If adequate performance across all SNR scales is required, this may be achieved with the appropriate choice of loss function, for instance by training on the log of the SNR. By choosing not to train on log SNRs, we prioritise the regions of parameter space corresponding to larger SNRs in the data set (at the fiducial luminosity distance of 1 Gpc). As these SNRs will be pushed towards the detection threshold at larger distances, and the majority of our luminosity distance distribution is above 1 Gpc, estimating these larger SNRs well has the greatest impact on accurate detectability estimates.

The other main contributor to variations in the SNR function is the mass ratio, as the waveform amplitude is proportional to the mass of the secondary CO [177]. As the mass ratio spans ~ 3 orders of magnitude in the EMRI parameter space we consider, we distribute the training data logarithmically with respect to the mass ratio to ensure that this variation is well captured at all scales.

4.4.2 Interpolating over the selection function

Using our trained interpolator network to produce SNR estimates, we are now able to approximate $\alpha(\vec{\lambda})$ via Equation (4.14) with the interpolated SNR as an input. However, two issues still remain that must be addressed before this may be applied in practice. First, this setup still requires that the parameter set $\{\vec{\theta}\}$ is drawn from $p_{\rm pop}(\vec{\theta}|\vec{\lambda})$ for each hyperlikelihood call; this is a slow operation, even for the relatively simple population models in Table 4.1. It is also inefficient, because $\alpha(\vec{\lambda})$ will typically not vary significantly across the high-probability region

of the posterior, where the majority of samples are drawn. Second, the stochastic nature of the Monte Carlo selection function estimates itself presents challenges in sampling: the hyperlikelihood surface becomes noisy, which can be problematic for the reliable convergence of sampling algorithms. These issues prohibit the use of our stochastic selection function estimates in hyperposterior sampling and must be solved.

To address these problems and further accelerate our bias-corrected likelihood, we extend the idea of interpolating over high-dimensional spaces with MLPs further. A second MLP trained prior to sampling can be used to interpolate directly over $\alpha(\vec{\lambda})$. The hierarchical likelihood after substituting $\alpha(\vec{\lambda})$ for this second MLP is

$$\mathcal{L}(\{\vec{\theta}\}|\vec{\lambda}) = \prod_{i=1}^{N_{\rm obs}} \frac{1}{S_i \alpha_{\rm MLP}(\vec{\lambda})} \sum_{j=1}^{S_i} \frac{p_{\rm pop}({}^j\vec{\theta}_i|\vec{\lambda})}{\pi(\vec{\theta}_i)}, \tag{4.18}$$

where $\alpha_{\rm MLP}(\vec{\lambda})$ is the MLP that is trained to interpolate the selection function. The architecture and training settings for this MLP, given in the third column of Table 4.2, are nearly identical to those of the SNR MLP introduced in Section 4.4.1. We use two fewer hidden layers for the selection function MLP to account for the higher dimensionality of the EMRI parameter space with respect to the hyperparameter space. Using this second MLP step greatly reduces the time per likelihood call, achieving a further order of magnitude of speedup with respect to stochastic estimation of the selection function. For our chosen population model, the resulting computational cost of the numerator and denominator in Equation (4.2) become roughly equivalent. Further speedup is achieved with vectorised evaluation of the hyperlikelihood (i.e., evaluating Equation (4.18) for multiple sets of hyperparameters $\vec{\lambda}$ at once), as the selection function MLP is capable of handling many sets of hyperparameters at no additional cost provided that sufficient GPU memory is available. This vectorisation would not be practical if one were to estimate the selection function values with the stochastic approach. With the main limitations of our method addressed, we are now able to produce selection function estimates usable in sampling at low computational cost.

4.5 Unbiased population inference results

With our strategy for selection bias correction formulated, we are now properly equipped to tackle an EMRI population inference problem. We begin by simulating a catalogue of EMRI observations. To estimate how well LISA will resolve the shape of the EMRI population in a realistic scenario, we choose hyperparameter values supported by recent black hole population studies:

- While constraints have been placed on the slope of the MBH mass function by current observations [420], the mass function for MBHs hosting EMRIs is subject to additional se-

lection effects that are poorly understood at present [190]. Recent work has estimated that the spectral index of this power law after the inclusion of selection effects is $\lambda_M \approx -1.43$ in the mass range $[M_{\rm min}, M_{\rm max}] = [10^5, 10^7]\ M_\odot$ [190]. We assume that the slope of the CO mass function is equal to the median value observed in stellar-mass BBH mergers of $\lambda_\mu \approx -3.50$ in the mass range $[3, 90]\ M_\odot$ [193], assuming that the progenitors of these mergers are representative of the universal stellar-mass BH population. The selection effects that translate this into the EMRI CO mass function are poorly understood and would require a dedicated set of N-body simulations of stellar cusps to properly quantify [190, 191], so we do not consider them in this study. Despite this caveat, the slope observed via BBH mergers is the strongest constraint placed on the mass function for black holes in this mass range available and is a reasonable starting point for estimating LISA's ability to resolve the CO mass function with EMRI observations.

- The MBH spin magnitude distribution is also poorly constrained by observational data. Current measurements are limited to MBHs in AGN [413], which may not be representative of the full MBH spin magnitude population as different formation channels will yield different MBH spin magnitude distributions [80]. Self-consistent simulations of MBH growth with cosmic evolution predict that most MBHs have spins greater than 0.9 in the MBH mass range quoted above, with a fairly narrow spread below $10^7\ M_\odot$ [197]. We include this characteristic shape of a narrow spin distribution above $a = 0.9$ in our population by choosing $[\mu_a, \sigma_a] = [0.93, 0.06]$.

- The time that each EMRI plunges with respect to the start of the observational data $t_{\rm plunge}$ is randomly distributed in the range $[0, 10]$ yr. In line with the planned LISA mission duration, we assume a 4-year observational window [428]: some EMRIs will not plunge until after the end of our observational data, but may still be detectable if they are bright enough. We assume that EMRIs occur at a rate of $\mathcal{R} = 700\ {\rm yr}^{-1}$, which is conservative (considering our redshift cut-off of $z = 6$) when compared with EMRI rate estimates from astrophysical modelling [190, 191, 414].

After discarding the signals too faint to be detected, we obtain a catalogue of 116 EMRIs.

To demonstrate the selection biases present, we perform two sampling runs: one in which selection biases are corrected for with our interpolation scheme, and another in which selection effects are not accounted for, i.e., replacing $\alpha(\vec{\lambda})$ with 1 in Equation (4.3) and Equation (4.4). We sample the hyperposterior Equation (4.2) with the NESSAI nested sampler [254, 326], using default settings. The convergence of all sampling runs with these settings was verified by examining the results of internal consistency checks built into NESSAI. The hyperposteriors obtained from these sampling runs for a subset of hyperparameters are shown in Figure 4.2. The full hyperposterior is shown in Section 4.B, which demonstrates minor discrepancies between the two hyperposteriors for the other hyperparameters, with marginal posteriors that are too narrow

(over-constrained) but otherwise fairly consistent with the set hyperparameter values.

The bias that results from ignoring selection effects on the observations is visible here as a discrepancy between the credible interval contours of the two hyperposteriors at the 99% level. For λ_M, the uncorrected posterior is inconsistent with the true value at the 99% credible level; the marginal posteriors for λ_μ that include or exclude selection bias correction disagree to an even greater extent. There is also a clear difference between the marginal posteriors on $\mathcal{R}$: this is symptomatic of the presence of selection effects, as it indicates that predicting the overall event rate solely from the size of the detection catalogue will result in a significant underestimate of the actual event rate.

After accounting for selection effects, our results serve to probe how well LISA can resolve the form of this EMRI population. We estimate (quoting the median and the 90% credible interval) that $\lambda_M = -1.39^{+0.12}_{-0.12}$ and $\lambda_\mu = -3.58^{+0.16}_{-0.17}$, corresponding to precisions of 8.8% and 4.6% respectively. The MBH spin distribution is well recovered, with $\mu_a = 0.924^{+0.008}_{-0.007}$ and $\sigma_a = 0.054^{+0.006}_{-0.005}$; these hyperparameters are recovered to within 0.87% and 10% respectively. The EMRI rate is estimated with 12% precision to be $\mathcal{R} = 678^{+88}_{-75}$ yr^{-1}. The precision achievable by LISA will roughly scale with the square root of the true event rate, which we have assumed a conservative value for in this study: for the most optimistic scenarios, this could improve by as much as an order of magnitude [190].

The number of detected events depends on the underlying population. For example, if the number of events is skewed to high redshift with respect to our assumed distribution of uniform in comoving volume and comoving time, then the number of detected events will decrease accordingly. Our choice of redshift distribution is equivalent to assuming that the probability of an EMRI occurring for a given MBH is constant across cosmic time. In reality, we expect that the physics of EMRI formation, such as cusp erosion [190], will lead to deviations away from this. Similarly, we expect that the distributions of MBH and CO masses will differ in reality from our assumptions. Hence, the results presented here should only be considered illustrative. A comprehensive study of how population inference results vary with the underlying population (which would require a computationally efficient method, such as ours) is necessary to fully map out how well LISA could measure the EMRI source population.

4.6 Verifying the accuracy of results

In the previous section, we demonstrated the capability of our approach for a single example. However, this is not sufficient to ensure that the selection function estimates output by our MLP are sufficiently accurate and unbiased that population inference will return hyperposteriors that are consistent with the truth across the hyperparameter space.

To assess whether this is the case, we opt for the P–P plot test [337]. An overview of P–P tests and plots is given in Section 2.4.1 First, we draw N sets of hyperparameters from the

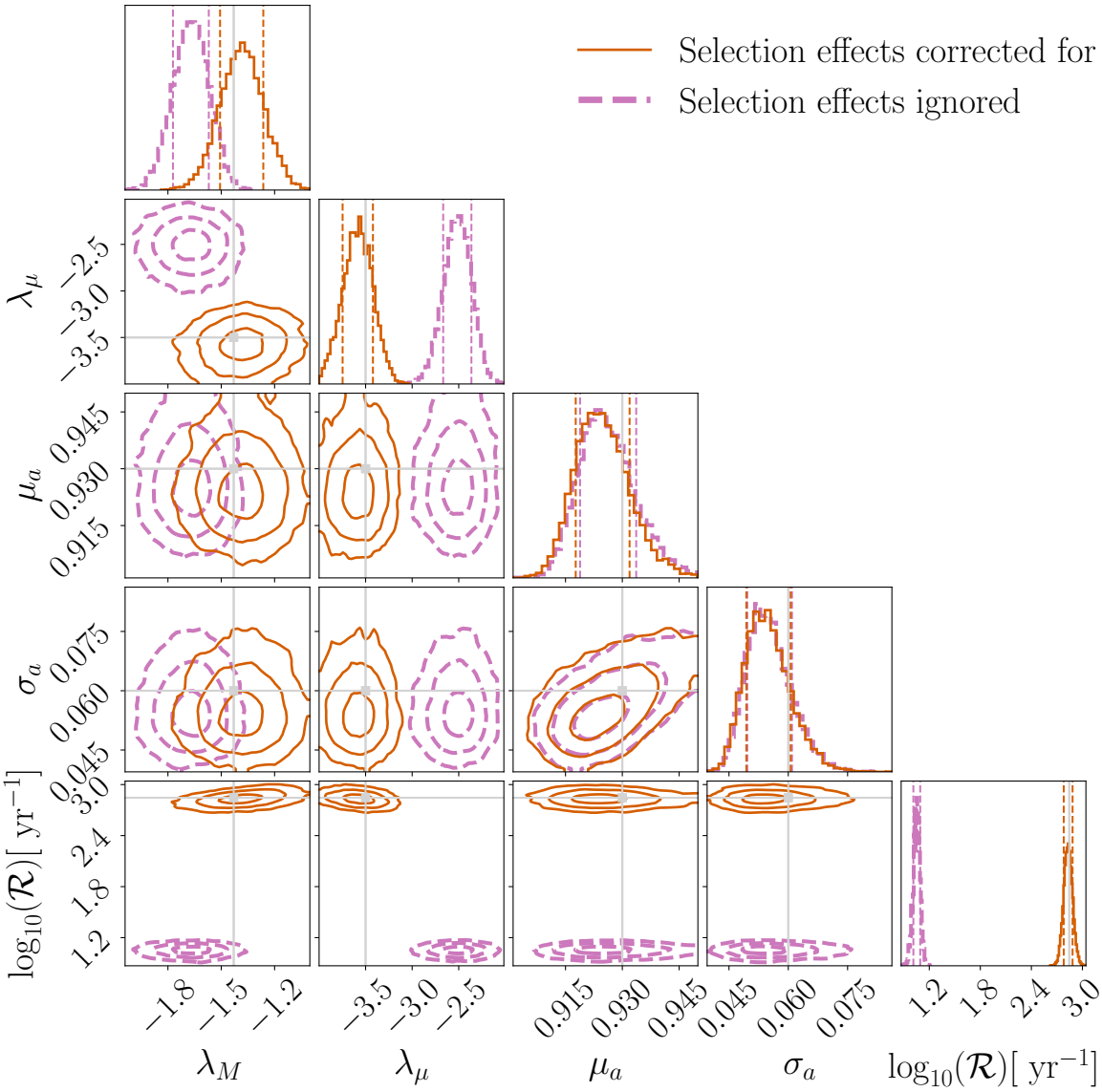


Figure 4.2: Recovered hyperposteriors for our example scenario with selection effects corrected for (solid) or ignored (dashed). The true values of each hyperparameter are indicated by the cross-hairs. Significant bias is present in the recovery of λ_μ, with minor bias also visible in the recovery of λ_M. The rate $\mathcal{R}$ is accurately recovered with the inclusion of selection effect correction.

Table 4.3: Hyperprior distributions chosen for all sampling runs. As the range of plunge times is reduced by a factor of 5 for our P–P plot analysis (Section 4.6), our prior bounds on the EMRI rate are adjusted accordingly: this is indicated by $(^*)$.

Parameter	Distribution	Limits
λ_M	Uniform	$[-4,-1]$
$M_{\rm min}$	Uniform	$[5,50]\times10^4\,M_\odot$
$M_{\rm max}$	Uniform	$[5,50]\times10^6\,M_\odot$
λ_μ	Uniform	$[-4,1]$
$\mu_{\rm min}$	Uniform	$[1,5]\,M_\odot$
$\mu_{\rm max}$	Uniform	$[80,100]\,M_\odot$
μ_a	Uniform	$[0.05,0.95]$
σ_a	Uniform	$[10^{-3},2]$
$\mathcal{R}$	Log-uniform	$[350,1050]$ $([75,150]^*)$ yr^{-1}

hyperpriors described in Table 4.3, and generate the corresponding population catalogues. We then perform hyperposterior sampling runs to produce estimates of the hyperparameters in each case, and determine the confidence interval q_λ of the true hyperparameters with respect to the posterior obtained. Last, we plot the CDF of q_λ. When the trial sets of hyperparameters are drawn from the hyperprior, we expect that the true value of a hyperparameter will fall within the $x\%$ credible interval in $x\%$ of realizations (i.e., a plot of q_λ against its CDF will be diagonal) if our hyperposteriors are consistent with the true values in all cases. We test the accuracy of our inference framework by comparing the calculated CDF with the expected diagonal trend.

Some variation of each CDF from the diagonal due to small-number statistics is expected. For the P–P plot to be meaningful N needs to be large, so we modify our population to reduce the computational cost of waveform generation by reducing the length of the observational window from 4 yr to 2 yr and the range of EMRI plunge times from $[0,10]$ yr to $[0,2]$ yr. To further reduce the cost of generating each population, we lower the overall event rate by limiting our population to a maximum redshift of $z=1$. Adjusting the event rate to account for these reductions in both duration and sensitive volume, the number of expected EMRIs for each population decreases by a factor of 35. The simulation and analysis configuration remains otherwise unchanged from the analysis described in Section 4.5. We perform three analyses to compare their results: first, we exclude selection effect correction; second, we include selection effect correction by means of a linear interpolation scheme, and last, we include selection effect correction with our neural network interpolation scheme. The P–P plots obtained from these analyses are shown in Figure 4.3, broken down by hyperparameter and compared to the expected $68\%, 90\%$ and 99% deviations for the $N=208$ sets of drawn hyperparameters [440].

We first examine the P–P plot for the hyperposteriors obtained when selection effects are excluded, setting $\alpha(\vec{\lambda})=1$, as shown in the top panel of Figure 4.3. The resulting posteriors exhibit strong biases with respect to the true values, and the P–P test fails; the majority of the hyperparameter CDFs deviate beyond the 99% confidence interval.

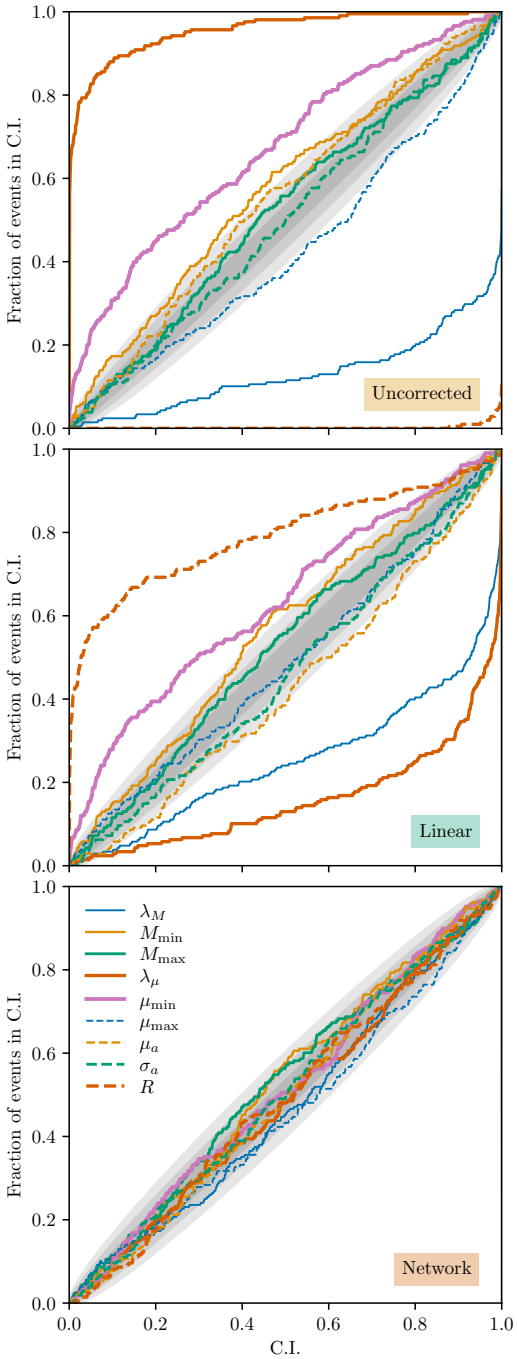


Figure 4.3: P–P plots showing the fraction of hyperparameters within a given CI for $N = 208$ hyperprior draws. Leaving selection biases uncorrected, setting $\alpha(\vec{\lambda}) = 1$, demonstrates the presence of significant selection biases in the population inference (*top*). These biases are successfully rectified with our approach (*bottom*), whereas standard interpolation techniques fail to produce unbiased results (*middle*). The expected 68%, 90% and 99% confidence intervals are shown in grey.

The P–P plot is a useful diagnostic for how the posterior deviates from what is expected on a dimension-by-dimension basis. We can identify that $\lambda_\mu, m_{\rm min}, \mu_{\rm min}$ and μ_a are typically overestimated when selection effects are ignored, and conversely that $\mu_{\rm max}$ and λ_M are underestimated. This can be understood by considering how the SNR of an EMRI waveform is correlated with the parameters associated with these hyperparameters. The SNR is positively correlated with μ, so the power-law slope is flattened off by the suppression of low CO mass events; similarly, SNR is higher for larger MBH spins, so the mean of the truncated Gaussian shifts to the right due to the suppression of the lower spin events. For λ_M, the opposite is true: the number of detectable higher MBH mass EMRIs (which are at lower frequencies) is suppressed due to the shape of the LISA sensitivity curve and so the observed power-law steepens. The parameters that pass the P–P test, $M_{\rm max}$ and σ_a are not strongly influenced by selection bias. The mean of the MBH spin truncated Gaussian may shift, but the change in the width of the distribution will be proportionally smaller, and is therefore less sensitive to this selection effect. Likewise, as the high MBH mass EMRIs are typically unobserved and occupy a small fraction of the overall EMRI population, small changes to the maximum MBH mass do not strongly affect the detectable fraction of the population. This is not the case for the CO mass distribution: the high mass events in the upper end of the power law are also the brightest events in the population, so adjusting the upper limit of the power law leads to larger changes in the fraction of events expected to be detectable. As the observed deviations from consistent hyperposteriors align with our expectations, we are confident they are the result of selection effects.

To demonstrate the impact of the low interpolation accuracy seen in Figure 4.1 on the resulting selection function estimates, we repeat this analysis with an MLP selection function trained on SNR estimates produced by the linear interpolator described in Section 4.4. The resulting P–P plot, shown in the middle panel of Figure 4.3, demonstrates that although modest correction is achieved in this case, it is still far too biased and inaccurate to result in consistent hyperposteriors. In the case of λ_μ, this even results in an over-correction of the selection bias when compared to the uncorrected case. The danger of over-correction implies that it is not sufficient to include a selection function term in the population likelihood: the selection function must also be accurately calculated to obtain good results.

Finally, we present the P–P plot obtained for this analysis with our MLP selection function estimator in the bottom panel of Figure 4.3. In stark contrast to the previous two plots, the hyperparameter CDFs are fully consistent with the expected confidence intervals. This conclusion is supported by the results of K-S [441] tests for each hyperparameter. Combining the p-values from each test with Fisher's method [340] returned a combined p-value of 0.3. This indicates that all hyperparameters (including the EMRI rate $\mathcal{R}$) are consistently recovered: the MLP is capable of producing selection function estimates that are sufficiently accurate for consistent posteriors to be obtained. This result verifies the application of our method in the treatment of selection biases in population inference.

4.7 Conclusions

Population inference with EMRIs has the potential to probe the evolution of both MBHs and their galactic neighbourhoods to unprecedented precision. However, the computational cost of components of this analysis are prohibitively high. Estimating selection biases in EMRI populations is computationally expensive due to a combination of the need for costly waveform models and the resources required to perform SNR calculations for long-duration data. As the SNR calculation is the computational bottleneck, we substitute it for an interpolation over pre-computed SNRs. We find that commonly employed interpolation schemes are not sufficiently fast or accurate for this problem, so we instead use machine-learning techniques. By using a neural network trained on a data set of SNRs distributed uniformly in the EMRI parameter space, we achieve a speedup of six orders of magnitude over direct SNR evaluation. We then replace the stochastic estimation of the selection function obtained via Monte Carlo integration with a second neural network that is trained on these stochastic estimates distributed uniformly in the hyperparameter space. This further improves the speed of hyperlikelihood evaluation by an order of magnitude, and enables vectorised estimation of the selection function for further reductions in computational cost. To verify the robustness of our approach against systematic biases, we globally evaluate hyperposterior consistency by simulating 208 EMRI populations and checking the results of selection bias-corrected hyperparameter estimation with a P–P test. This test confirmed that (i) the presence of selection effects significantly biased inferences that did not correct for them appropriately, and (ii) our approach successfully corrected for selection effects to produce unbiased results.

We apply our method to the inference of an astrophysically-motivated EMRI population (assuming sources distributed uniformly in comoving volume and time) to study LISA's ability to probe the structure of such populations. We estimate that $\lambda_M = -1.39^{+0.12}_{-0.12}$ (a precision of 8.8%) and that $\lambda_\mu = -3.58^{+0.16}_{-0.17}$ (a precision of 4.6%). For the MBH spin magnitude distribution, we find that $\mu_a = 0.924^{+0.008}_{-0.007}$ and that $\sigma_a = 0.054^{+0.006}_{-0.005}$, resolving the width of the MBH spin magnitude distribution to within 10%. The event rate is estimated to be $\mathcal{R} = 678^{+88}_{-75}$ yr^{-1} (a precision of 12.0%).

The capability of our approach for treating selection effects in the case of a simple population model, which excludes substructure or correlations due to astrophysical effects, paves the way for future work to investigate more complex EMRI population models. The EMRI population we expect to be present in reality is multi-faceted [190], and by introducing these features systematically to the population inference problem we can begin to characterise their measurability with space-based detectors such as LISA. As population inference is intimately tied to other hierarchical inference problems, including cosmological inference [217, 218] and tests of general relativity [442], proper treatment of selection effects for EMRIs has direct implications for these analyses as well. Ultimately, a joint hierarchical inference over this problem space may be required, of which our approach can be an integral part.

As our method is not predicated on a particular population model, it is applicable to a wide variety of population inference problems. Similarly, as any waveform model may be used, this approach is capable of accommodating future changes to EMRI waveform models with little tuning required. While our approach specifically targets the EMRI population inference problem, it may be generalised to any problem with an SNR-threshold selection bias due to this model-agnostic nature. The reduction in computational cost achieved by employing our method will be most pronounced in cases where the SNR function is expensive to compute and of high dimensionality, but the ability to form a vectorised SNR approximant will still offer a notable speedup for waveform models that are not easily parallelisable or vectorisable (e.g., due to memory constraints).

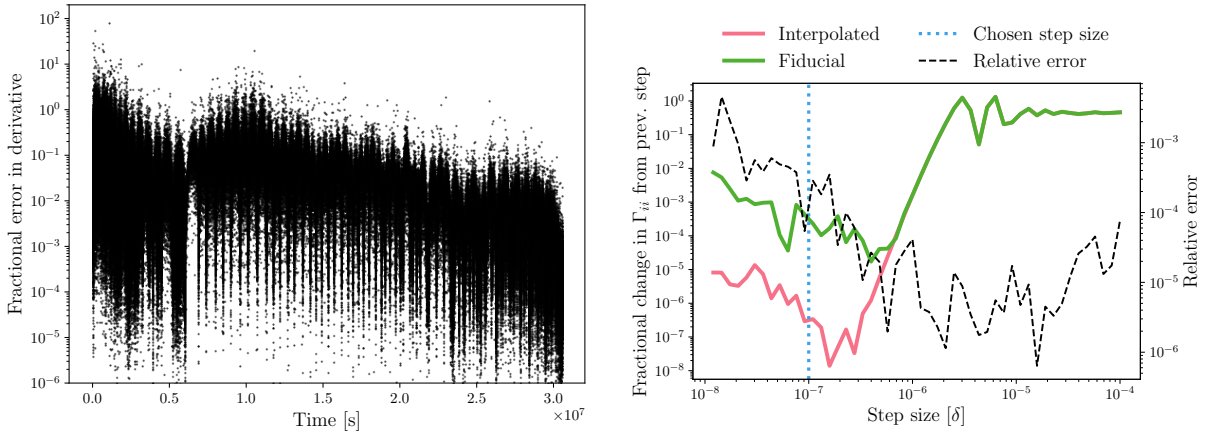
Our code package POPLAR, containing the tools used in this chapter, has been released as an open-source package [288] for use in future population studies.

4.A Challenges encountered in FIM generation

During the step-size tuning process described in Section 4.3.2, we identified two issues specific to the generation of FIMs with FEW. If left unaccounted for, these issues led to significant underestimation of the width of the EMRI likelihood, so we developed techniques for mitigating their impact on FIM accuracy. As these investigations required an in-depth investigation of the systematics of EMRI waveform generation with FEW in the context of finite-differencing operations, it constitutes a significant departure from the main body of work presented in this chapter and has therefore been presented as an Appendix. We now describe the two issues encountered during FIM computation in turn.

4.A.1 Interpolation error in FEW

The first issue we identify derives from a simplifying assumption made during the summation stage of waveform generation in this waveform generation framework. See Section 3.1 for a description of EMRI waveform modelling with FEW; we focus specifically on the summation step here. In the summation kernel, the sparse time points output by the trajectory integrator are interpolated with cubic splines to obtain a waveform at the sampling rate of the data stream. This is an computational cost-reducing approximation: because an eighth-order Runge-Kutta integrator was used to evolve the trajectory, the appropriate choice of interpolating function is a seventh-order spline as opposed to a third-order (cubic) spline [443]. This choice has negligible impact in most data analysis scenarios, such as PE or detectability estimation, because the waveform mismatch incurred is typically extremely small [177]. However, waveform derivatives are particularly sensitive to this interpolation error; the fractional error in the waveform derivative with respect to M due to the mis-matched interpolation order is shown in panel (a) of Figure 4.4. Due to the large spacing between samples early in the trajectory, the relative error in



(a) (b)

Figure 4.4: Errors in EMRI waveform derivatives and FIMs due to the use of a third-order of spline interpolant in the mode summation stage of waveform generation. The EMRI waveform model used approximates the seventh-order spline over a sparsely sampled trajectory with a third-order (cubic spline); this does not cause biases in inference [177], but leads to significant errors in numerical waveform derivatives. Panel (a) shows the fractional error in the derivative of an EMRI waveform with respect to M as a function of time. Early in the waveform, where the sparse trajectory points are widely separated, fractional errors of more than an order of magnitude are observed. At late times, the trajectory is more densely sampled and derivative errors reduce to $\sim 10\%$. The impact of these derivative errors on the determination of FIM elements is shown in panel (b), which shows how the FIM diagonal element with respect to M (green line) does not stabilise to better than one part in 10^4 regardless of the finite-differencing step size used. This is shown in Figure 4.5 to lead to highly inaccurate FIMs. By implementing an additional pre-processing step in which we upsample the sparse trajectory with a fifth-order spline (curbing any significant interpolation errors), we improve FIM stability by more than three orders of magnitude (pink line). The relative error between the FIM element values for each approach are indicated by the black dashed line.

the derivative early in the waveform regularly exceeds 10. Closer to plunge, use of the incorrect interpolant leads to derivative errors of $\sim 10\%$; although smaller in scale, the rapid evolution of the waveform near-plunge greatly improves the precision with which EMRI parameters may be estimated [190], so accurate waveform derivatives are more important here. Collectively, these deviations are sufficiently large to introduce significant errors into the calculation of EMRI FIM calculations and must be corrected for.

We alleviate these errors with an intermediate processing step which we apply to the trajectory before it is fed to the summation kernel. By interpolating the trajectory with a fifth-order B-spline and increasing the density of points by a factor of 10, the interpolation error incurred in the summation kernel is significantly reduced, improving both the accuracy and the stability of the waveform derivatives with respect to the intrinsic parameters. The effectiveness of this intermediate step are apparent when examining the stability of the FIM diagonal elements with respect to the finite-difference step-size, as shown in panel (b) of Figure 4.4 for the diagonal element corresponding to M. At each step-size value, we examine the change in the value of the diagonal element with respect to what was obtained with the previous step-size. We expect this quantity to be minimised within a stable region of step-sizes (if one exists at this point in parameter space). In the step-size region $[10^{-7}, 4 \times 10^{-7}]$, the FIM diagonal element for the interpolated waveform is stable to less than one part in a million. The fiducial waveform exhibits changes two orders of magnitude larger than this in the same region. We also examine the relative error between the fiducial and interpolated waveform derivative, which is $\sim 10^{-4}$ in the FIM diagonal element within the region of stability ($\delta \in [10^{-8}, 10^{-6}]$). Errors of this scale may appear insignificant, but their impact on the accuracy of the LSA likelihood are clearly apparent when compared with the results of PE. One such comparison is shown in Figure 4.5, in which PE is performed on a noise-free injection with the fiducial waveform.

Wide, flat prior distributions are assumed for all parameters, enabling direct comparisons to be drawn between the sampled posterior and the LSA likelihoods. The LSA likelihood corresponding to the interpolated waveform is reasonably accurate, whereas the LSA likelihood of the fiducial waveform significantly underestimates the uncertainties on the intrinsic parameters. We quantify the disparity between each LSA likelihood and the sampled posterior by computing the KL divergence between the two distributions in each case. The KL divergence (or relative entropy) is a statistical measure of how different two probability distributions are [444]. We express all KL divergences in units of nats. In this case, the KL divergences between each of these approximate likelihoods and the posterior obtained from PE are 23.8 and 0.02 respectively, supporting the conclusion that the LSA likelihood is a significantly poorer approximation of the true likelihood when interpolation error in the waveform model is not addressed. Based upon this result, we compute FIMs for all EMRIs systems with this additional interpolation step included.

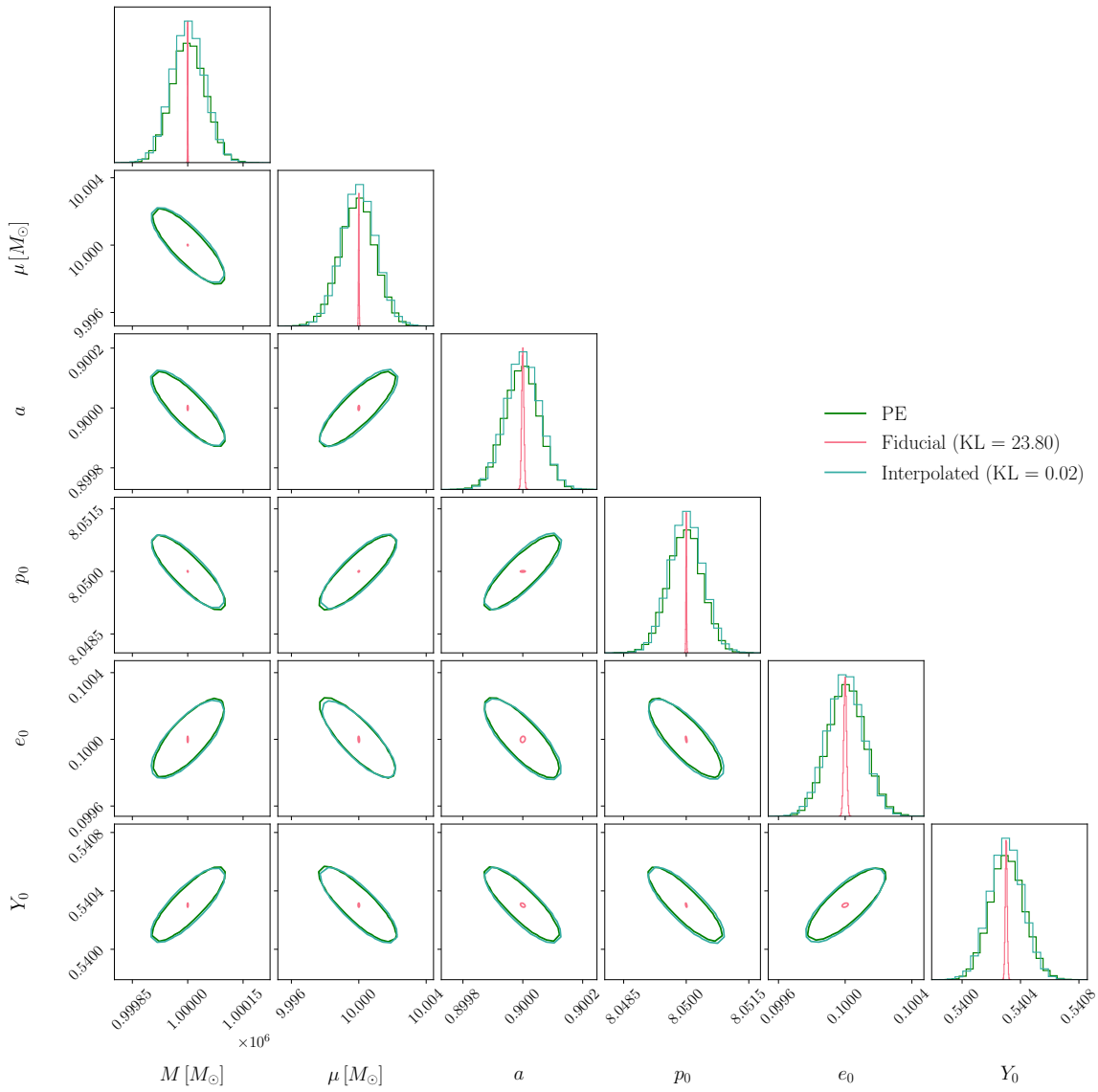
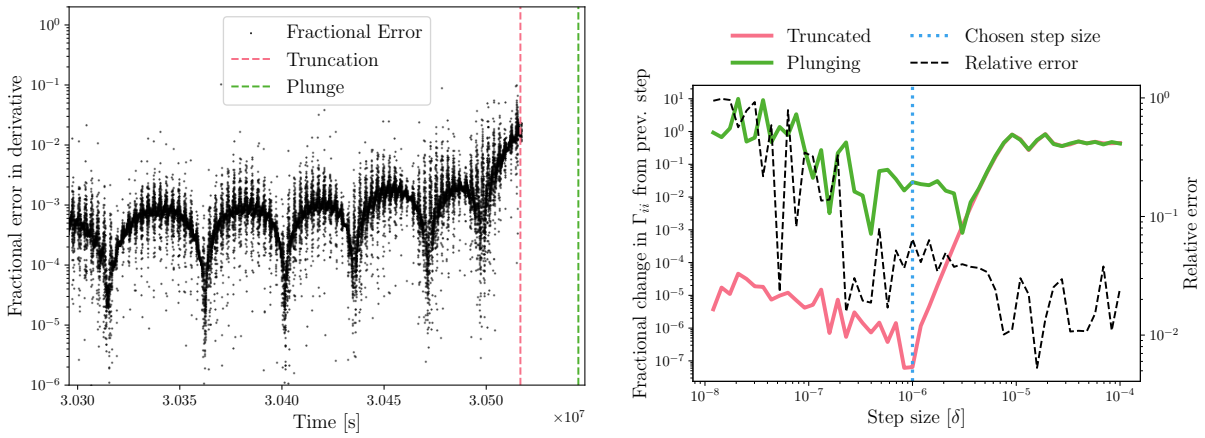


Figure 4.5: Addressing errors in EMRI FIMs due to interpolation errors. The approximation of a seventh-order spline over sparse inspiral trajectories with a cubic spline leads to erroneous FIMs (pink line) that significantly underestimate the width of the likelihood (obtained via PE, green line). Following the stability improvements presented in Figure 4.4, we verify that the FIMs we obtain (cyan line) closely approximate the true likelihood. This is reflected in the Kullback–Leibler (KL) divergences between each FIM and the true likelihood (given in the figure legend), which reduces by three orders of magnitude with our addition of an additional fifth-order interpolation stage in the waveform model.

4.A.2 Biases for plunging EMRIs

The second problem we observed in FIM generation occurs when attempting to construct the FIM for an EMRI waveform which plunges in the observation window. It has been previously identified in the literature that computing accurate FIMs for signals of this morphology is an ill-posed problem [445]. The root cause of the bias in the FIM is the impact of a detectable cut-off (i.e. one which is above the noise floor) on parameter estimation [446]. This issue will be addressed once the merger-ringdown stages of coalescence are added to EMRI waveform models, which will smoothly transition the large waveform amplitude at plunge into a low-amplitude exponential decay and therefore eliminate the measurability of an abrupt termination point entirely. Regularisation of the FIM is possible by tapering this cut-off [446]. However, for EMRIs, the majority of the waveform SNR has accumulated in the inspiral phase, so we can simply truncate the observation window a short duration before plunge and suffer minimal biases to the FIM. We demonstrate the impact of the abrupt termination of an EMRI waveform on the derivative with respect to p_0 in Figure 4.6; in panel (a), we show how this waveform derivative error reaches $\sim 10^{-2}$ late in the inspiral. The impact of this error on the stability of the FIM is demonstrated in panel (b) of Figure 4.6 for the FIM diagonal element corresponding to p_0. When the observation window is truncated 6 hr prior to the plunge of the CO (denoted *Truncated*), we find a region of stable step sizes in $[10^7, 5\times10^7]$ in which the FIM element varies by less than one part in a million. However, if the waveform is allowed to end abruptly within the observation window (denoted *Plunging*) no such stable region can be found, and the relative error in the FIM element rapidly increases, exceeding 10^{-2} at the truncation point.

The observation window must be truncated sufficiently early that plunge does not occur for any waveforms within the nine-point stencil. In Figure 4.7, we examine the change in the LSA likelihood as this truncation buffer is increased in length from 0 hr to 12 hr and compare it with a likelihood obtained in PE. We observe that in the 3 hr to 4 hr range of truncation durations, the accuracy of the LSA likelihood improves significantly, indicating that the shortest-duration waveforms in the nine-point stencil plunge within the observation window if less than ~ 4 hr of the waveform is truncated. While truncating the end of the waveform reduces the total information that may be extracted from the signal (especially as the EMRI SNR increases rapidly towards the end of the inspiral [190]), even 12 hr is a relatively short period of time with respect to the $\mathcal{O}(4\,\mathrm{yr})$ duration of the waveform. The SNR of the waveform only decreases by a fraction of a percent, which has minimal impact on the width of the posterior distribution. This is reflected in the KL divergences obtained between the LSA posteriors and the likelihood obtained via PE, which sharply drops from > 10 to $\mathcal{O}(10^{-2})$ as the duration of waveform truncated approaches this value. As with Section 4.A.1, we find that the intrinsic parameters of the EMRI system are significantly underestimated (by more than one order of magnitude for some parameters) for systems that plunge within the observation window, which will lead to overconstrained population posteriors if this behaviour is unaccounted for. We therefore safeguard our



(a) (b)

Figure 4.6: Errors in EMRI waveform derivatives and FIMs due to the inclusion of near-plunge waveform data. In panel (a) we identify rapid deviation between the derivative of an EMRI waveform with respect to p_0 depending on whether the waveform evolution is truncated 6 hr prior to plunge (at the pink dashed line) or allowed to continue until plunge (green dashed line). The impact of this deviation is observed in the stability of the FIM diagonal element corresponding to p_0 in panel (b). When the waveform evolves until plunge (green line), the fractional variation in the FIM element value does not stabilise below 10^{-3} for any value of δ, leading to a significantly biased approximation of the likelihood (Figure 4.7). Once the final 6 hr of data are removed (pink line), maximal stability improves to one part in 10^{-7} for $\delta = 10^{-7}$. The relative error between the FIM element values for each approach are indicated by the black dashed line.

FIM computations against this issue by truncating 6 hr of data from the end of every waveform that plunges within the observation window.

4.B Full hyperposterior obtained from 4-year scenario

The full hyperposteriors from the population inference in Section 4.5 are shown in Figure 4.8. By including the mass range parameters, we can observe a more subtle consequence of the presence of selection biases: over-constrained hyperposteriors. Neglecting selection effects leads to an underestimation of the error on parameters: this is reflected in the P–P plot analysis of Section 4.6, where it is demonstrated that these effects lead to globally inconsistent hyperposterior effects at a statistically significant level.

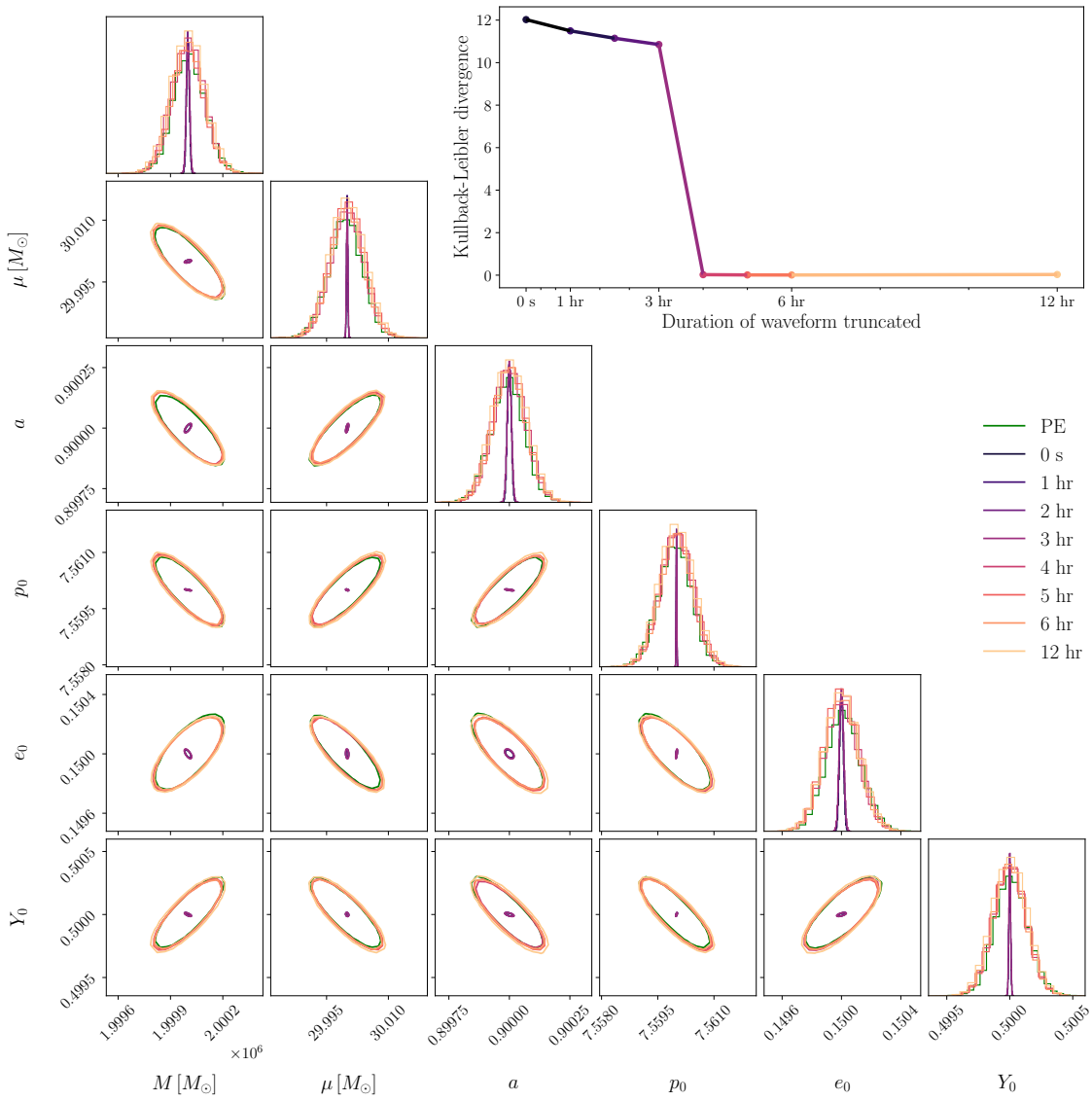


Figure 4.7: LSA likelihood approximations for an EMRI waveform with varying durations of pre-plunge data truncated are shown in the main panel. We observe that including the last few hours of the waveform prior to plunge results in an erroneous FIM, leading to significant discrepancies in the LSA likelihood when compared to the results of PE on the noise-free waveform. For this signal, we identify that $\sim 4\,\mathrm{hr}$ of data must be removed for computed FIMs to be accurate. This is reflected by a sharp decrease from > 10 to $\mathcal{O}(10^{-2})$ in the KL divergence (inset panel) between the approximate and sampled likelihoods between 3 hr and 4 hr of truncation. The PE results were obtained from an analysis of the entire waveform (i.e., no truncation).

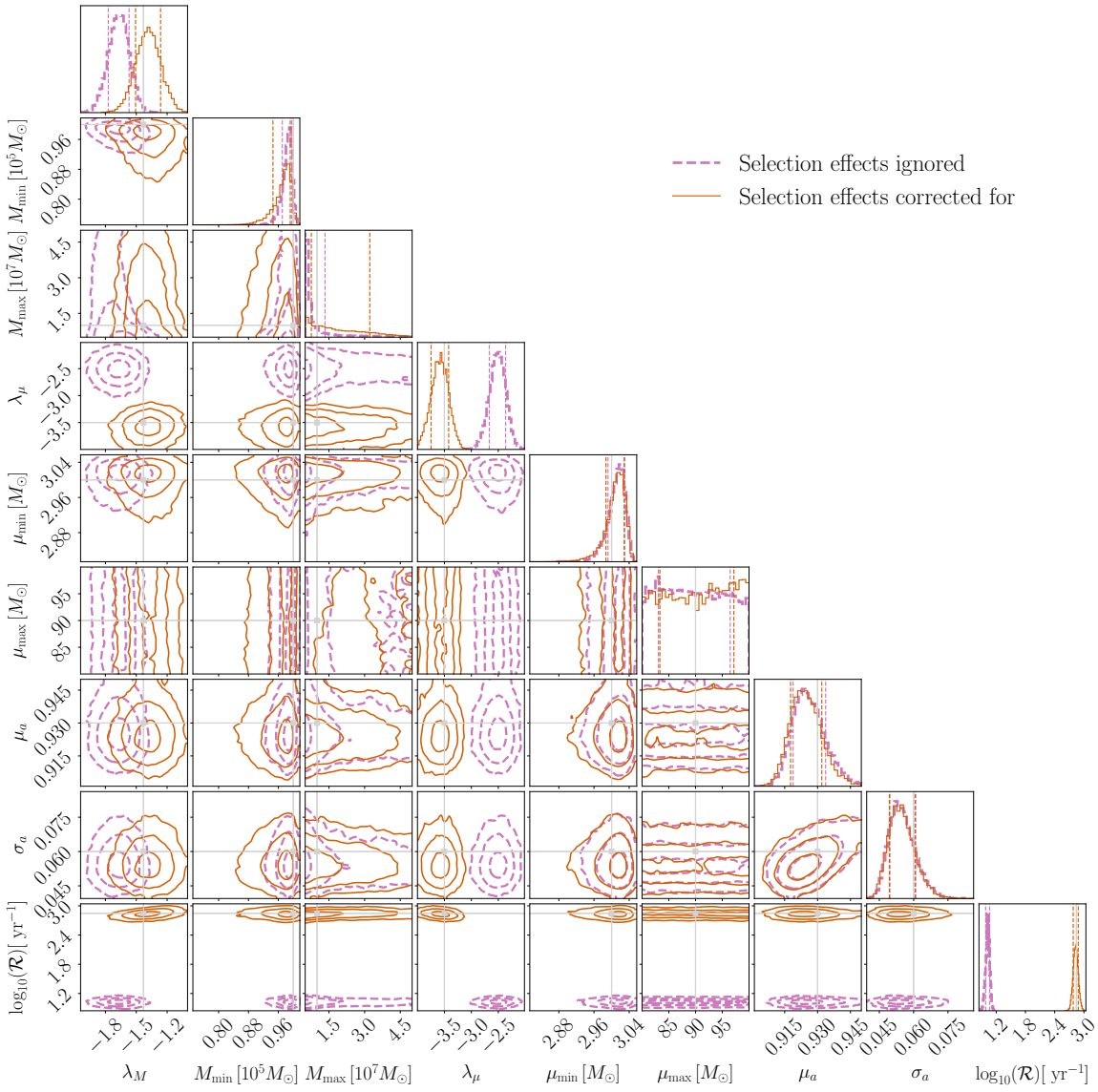


Figure 4.8: Full recovered hyperposteriors for our example scenario with selection effects corrected for (solid) or ignored (dashed). The true values of each hyperparameter are indicated by the cross-hairs. Significant bias is present for λ_μ, with more minor bias for λ_M. The rate $\mathcal{R}$ is accurately recovered with the inclusion of selection effect correction. For other hyperparameters, the dashed hyperposteriors are over-constrained when compared to the corrected hyperposteriors.

Chapter 5

Conclusions

In order to fully exploit the rich astrophysical information encoded in LISA observations, we identify three inter-connected goals towards which LISA data-analysis research efforts should be directed:

- Methods for the analysis of LISA data, which will take the form of a global fit, must be robust with respect to the properties of the LISA instrument. Data gaps, non-stationary noise (in the form of long-term drifts or transient glitches) and sources of uncertainty in the LISA data processing workflow significantly complicate global fit analyses. In order to inform the development of (and assess the performance of) analysis pipelines in advance of the LISA mission, we must develop tools for accurate high-fidelity simulations of the LISA instrument.

- As LISA will observe GW signals with large SNRs, waveform models for these signals must be sufficiently accurate that astrophysical results are not significantly affected by systematic bias. Inaccurate waveform models will also lead to post-subtraction residuals in LISA data, limiting the sensitivity of global fit analyses.

- Waveform models must also be computationally efficient. Any improvements in computational cost must be achieved without compromising accuracy, due to the nature of LISA data analysis. As a global fit analysis comprises many waveform evaluations, reducing the cost of waveform models is essential if these analyses are to be computationally feasible.

The work presented in this thesis is strongly aligned with these objectives, and advances the state of the art in the simulation and analysis of LISA data. We now summarise the content of each chapter in turn.

The need for accurate LISA simulations is addressed by the work presented in Chapter 2, in which we validate LISA data analysis results for a realistic LISA instrument simulation and analysis pipeline. We first outline the modular components of the simulation stage in Section 2.1, describing how they work in isolation and how they fit together to form a complete simulation of

the instrument. Following this simulation, in Section 2.1.5 we describe how the phasemeter data from each spacecraft are processed via TDI in order to suppress laser and clock noises. We then validate the PE analysis of the output TDI variables, probing for the presence of systematic bias in the results obtained from these analyses. Balancing the sensitivity with which we can identify these biases with the computational cost of the simulation and analysis pipeline, we consider two scenarios:

1. In the first scenario (Section 2.3), we study GB, MBHB and EMRI signals in 1 yr of LISA data. Due to the significant resource requirements of LISAINSTRUMENT, we restrict our study of PE for each source to individual analyses with different simulation configurations, limiting the scope of our bias investigations to that of large (i.e., posterior width-scale) biases. We observe significant biases for GB signals with SNRs $\mathcal{O}(10^3)$, in agreement with the literature. We identify no strong biases in our analyses of GBs of lower (and more representative) SNRs. For MBHBs, we obtain PE results that are marginally consistent with an absence of systematic bias for generic spacecraft orbits. Unlike for MBHBs and GBs, we observe significant systematic biases in our analysis of EMRIs, finding that the scale of these biases is correlated with the beatnote frequencies measured at each spacecraft. We conclude that these biases are attributable to numerical precision errors in either the simulation or pre-processing of LISA data; further investigation of the source of these biases is an avenue for future research.

2. For the second scenario (Section 2.4), we reduce the duration of our simulations to 3 d and consider only GB signals. These modifications reduce the computational cost of each simulation sufficiently that we can perform P–P tests on ensembles of PE results. We demonstrate that approximations in the FASTGB waveform model measurably bias GB PE at lower SNRs than previously identified in the literature, indicating that more sophisticated waveform models will be required for the analysis of these systems. Otherwise, we find no evidence for systematic bias in our PE results. This significant result was reported to ESA as part of the LISA adoption review in early 2024; it demonstrates that within-requirements suppression of laser and clock noises is achievable (without degrading GW signals) for a realistic LISA simulation.

We plan to combine aspects of these two scenarios in future work, extending our P–P test investigations of LISA performance to longer simulations (and to incorporate gaps and non-stationary noise features) once the resource requirements of the simulation have been addressed.

Chapter 3 focuses on the reduction of computational cost for the FEW EMRI waveform generation framework using ML techniques. We first outline EMRI waveform generation via FEW, within the context of GSF theory, in Section 3.1. In Section 3.2, we identify an inefficiency in FEW regarding the selection of dominant harmonic modes in an EMRI waveform, which is performed to reduce the computational cost of the mode summation stage. The FEW package

performs this operation on-the-fly, requiring all prospective mode amplitudes to be computed for the selection stage, which is computationally expensive. We address this limitation of FEW with ML: treating mode selection as an ensemble of binary classification problems, we train an MLP to predict the outcome of mode selection given the parameters of the EMRI system and rearrange FEW to generate mode amplitudes and build waveforms based on these predictions. To examine how this approach scales with the complexity of the EMRI system, we repeat this process for a range of waveform models. We train and validate MLPs for each waveform model in Sections 3.2.2 to 3.2.4, demonstrating that they accurately predict EMRI mode content in $\mathcal{O}(1\,\mathrm{ms})$. In Section 3.2.5, we use each MLP to explore the behaviour of the EMRI mode selection process, demonstrating that it successfully reproduces physically-motivated features of this process in parameter space. As the computational cost of an EMRI waveform scales with the number of modes needed to represent it, the trained MLP is a highly efficient approximant of EMRI computational cost, which is in itself a useful tool in EMRI data analysis studies.

Implementing the trained MLPs in our modified version of FEW, we then validate the performance of the resulting waveform model with respect to the standard FEW package. We begin in Section 3.3 by examining (for each waveform model in turn) waveform accuracy and computational cost reduction (via wall-time measurements) over the EMRI parameter space. These investigations confirm that waveform computational cost is reduced without compromising accuracy over the majority of the parameter space, with some outliers for the most complicated waveform model considered that are likely attributable to non-physical features of this waveform model identified in Section 3.2.4. We find that waveform generation is sped up by roughly one order of magnitude (excluding the cost of the LISA response function), significantly reducing the computational cost of EMRI data analysis and enabling rapid and accurate inference of EMRI parameters. To conclude the validation of our modified EMRI waveform generation scheme, we inject and recover EMRI signals via PE in Section 3.4. We find that the computational cost reduction achieved in Section 3.3 leads to a similar reduction in the wall-time per PE run. These investigations also identify a problem related to the recovery of EMRI signals with templates containing few harmonic modes that has not previously been described in the literature. We observe that the likelihood with respect to extrinsic parameters contains significant unphysical structure in these cases, leading to significant biases in PE and hampered sampling efficiency. While this problem is challenging to overcome for standard FEW, we show that it can be addressed at negligible computational cost in our modified waveform generation framework, enabling PE to be performed for eccentric equatorial inspirals into spinning MBHs in $\mathcal{O}(1\,\mathrm{hr})$ with minimal systematic biases.

In Chapter 4, we address the impact of computationally expensive waveform models on population inference studies. In order for population inference results to be free of systematic bias, waveform-dependent selection effects must be modelled and corrected for. However, computing selection effects for a single candidate population entails a large number of waveform

evaluations, making hierarchical inference methods computationally infeasible. We address these significant computational costs by modelling selection biases with ML methods, examining EMRI population inference as a representative example. After introducing the hierarchical inference framework we employ in Section 4.1, we outline the population model we consider in Section 4.2 and describe how selection effects (in the form of an SNR threshold) modify the observed form of this population in Section 4.2.1. We then describe our approach for producing simulated catalogues of EMRI observations, including both waveform generation (Section 4.3.1) and approximate parameter estimation under the LSA using FIMs (Section 4.3.2), and identify and correct for two effects that lead to biases in the LSA likelihood (Section 4.A). These parts of our analysis framework are standard and representative of typical EMRI population inference studies.

In Section 4.4, we focus on rapid and accurate modelling of selection biases. We identify that the bottleneck in the computation of the selection function (which accounts for selection effects) is the EMRI SNR function, which costs $\mathcal{O}(100\,\mathrm{ms})$ per evaluation. By training a MLP on a dataset of EMRI parameter-SNR pairings, we construct a rapid and accurate approximant of the EMRI SNR function capable of evaluating 10^6 SNRs in under 1 s. We demonstrate that the MLP is able to estimate $\sim 90\%$ of EMRI SNRs to within 10%, and is two orders of magnitude more accurate than classical grid-based interpolation techniques without compromising efficiency. To further reduce the computational cost of selection effect evaluation during PE, we generate a dataset of Monte Carlo selection function estimates computed using our MLP SNR approximant and train a second MLP on this dataset of selection function values, producing a smooth and low-cost approximant of the selection function. The selection function approximant is sufficiently inexpensive that the evaluation of selection biases contributes negligibly to the cost of population inference. Applying our selection effect modelling framework in Section 4.5, we perform PE on an astrophysically-representative EMRI population. From this analysis, we predict that LISA will measure the MBH and mass function slopes to precisions of 8.8% and 4.6% respectively, the MBH spin magnitude distribution parameters to within 10%, and the event rate to within 12%. We validate the accuracy of our selection bias approximant in Section 4.6 by conducting a P–P test over the hyperparameter space, an investigation made computationally feasible by the low cost of population inference due to our selection bias modelling. The results of this P–P test indicate that the population inference results are free of significant systematic bias, confirming that we have corrected for selection biases without introducing biases due to the mis-modelling of selection effects. The accuracy, efficiency and flexibility of our selection bias modelling framework will ensure that it is applicable in future population inference studies; as the constituent parts of this framework are neural networks (and therefore, non-linear) it is reasonable to expect that the performance of this approach will be maintained for more complex problems of higher dimensionality. A crucial feature of our approach is that the method by which the aforementioned neural networks are constructed is completely independent of the waveform

and population models used, making it applicable to any GW population study. Population models for LISA GW sources will need to account for a wide range of astrophysical effects (Section 1.3), so it is vital that any LISA population inference framework is sufficiently robust to remain accurate and efficient when paired with complicated astrophysical population models.

The work presented in Chapters 2 to 4 of this thesis addresses present limitations in LISA simulation and data analysis and provides a foundation upon which further ideas can be developed. An accurate and robust framework for LISA simulations is vital for the development of realistic LISA data analysis techniques, and validation of such a framework through extensive PE consistency testing is essential in achieving this robustness. While the work towards this goal in Chapter 2 does not cover the full scope of the problem, it lays the groundwork required for increasingly extensive studies to be performed. The end-to-end simulation and analysis pipeline will continue to be developed up to the launch of the LISA mission, providing the necessary framework with which to assess the performance of large-scale LISA data analysis methods. For the identification and extraction of EMRIs from LISA data to be a realistic prospect, waveform models must be made as efficient as possible. This is especially pertinent in the construction of EMRI search techniques, which is an area of ongoing research that is computationally expensive even in the development phase. The framework developed in Chapter 3 provides a practical solution for efficient waveform generation that is both readily applicable and scalable in order to accommodate the higher complexity of future waveform models. Similarly, LISA will present population inference challenges of large scale and complexity; efficient techniques for bringing the cost of hierarchical inference studies to reasonable levels are therefore vital if the astrophysical potential of LISA observations is to be fully exploited. Addressing the high computational cost of selection bias modelling, the ML techniques developed in Chapter 4 are highly scalable and applicable to a wide range of population inference problems. The capability of this framework to make EMRI population inference sufficiently inexpensive that P–P tests are possible underscores its applicability to a wide range of GW sources, even those for which efficient waveform models have not yet been constructed.

Bibliography

[1]M. L. Katz, L. Speri, A. J. K. Chua, C. E. A. Chapman-Bird, N. Warburton, and S. A. Hughes, *BlackHolePerturbationToolkit/FastEMRIWaveforms: Frequency Domain Waveform Added!*, version v1.5.1, July 2023.

[2]S. A. Hughes, N. Warburton, G. Khanna, A. J. K. Chua, and M. L. Katz, "Adiabatic waveforms for extreme mass-ratio inspirals via multivoice decomposition in time and frequency", Phys. Rev. D **103**, 104014 (2021).

[3]S. Isoyama, R. Fujita, A. J. K. Chua, H. Nakano, A. Pound, and N. Sago, "Adiabatic Waveforms from Extreme-Mass-Ratio Inspirals: An Analytical Approach", Phys. Rev. Lett. **128**, 231101 (2022).

[4]C. E. A. Chapman-Bird, C. P. L. Berry, and G. Woan, "Rapid determination of LISA sensitivity to extreme mass ratio inspirals with machine learning", Mon. Not. Roy. Astron. Soc. **522**, 6043–6054 (2023).

[5]A. Einstein, "Approximative Integration of the Field Equations of Gravitation", Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.) **1916**, 688–696 (1916).

[6]L. Blanchet, "Quadrupole-quadrupole gravitational waves", Class. Quant. Grav. **15**, 89–111 (1998).

[7]C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (W. H. Freeman, San Francisco, 1973), pp. 941–986.

[8]D. Christodoulou, "Nonlinear nature of gravitation and gravitational wave experiments", Phys. Rev. Lett. **67**, 1486–1489 (1991).

[9]M. Bailes et al., "Gravitational-wave physics and astronomy in the 2020s and 2030s", Nature Rev. Phys. **3**, 344–366 (2021).

[10]N. V. Krishnendu and F. Ohme, "Testing General Relativity with Gravitational Waves: An Overview", Universe **7**, 497, 497 (2021).

[11]P. Murdin and L. Murdin, *Supernovae* (Cambridge University Press, 1985).

[12]D. R. Lorimer and M. Kramer, *Handbook of Pulsar Astronomy*, Vol. 4 (2004).

[13]M. H. van Kerkwijk and D. L. Kaplan, "Isolated neutron stars: magnetic fields, distances,and spectra", Astrophys. and Space Science **308**, 191–201 (2007).

[14]R. A. Hulse and J. H. Taylor, "Discovery of a pulsar in a binary system", Astrophys. J. Lett. **195**, L51–L53 (1975).

[15]J. M. Weisberg and Y. Huang, "Relativistic Measurements from Timing the Binary Pulsar PSR B1913+16", Astrophys. J. **829**, 55 (2016).

[16]P. C. Peters and J. Mathews, "Gravitational radiation from point masses in a Keplerian orbit", Phys. Rev. **131**, 435–439 (1963).

[17]J. H. Taylor and J. M. Weisberg, "A new test of general relativity: Gravitational radiation and the binary pulsar PS R 1913+16", Astrophys. J. **253**, 908–920 (1982).

[18]J. M. Weisberg and J. H. Taylor, "Relativistic binary pulsar B1913+16: Thirty years of observations and analysis", ASP Conf. Ser. **328**, 25 (2005).

[19]M. Kramer et al., "Strong-Field Gravity Tests with the Double Pulsar", Phys. Rev. X **11**, 041050 (2021).

[20]I. H. Stairs, S. E. Thorsett, J. H. Taylor, and A. Wolszczan, "Studies of the relativistic binary pulsar psr b1534+12: I. timing analysis", Astrophys. J. **581**, 501–508 (2002).

[21]B. P. Abbott et al. (LIGO Scientific), "LIGO: The Laser interferometer gravitational-wave observatory", Rept. Prog. Phys. **72**, 076901 (2009).

[22]J. Aasi et al. (LIGO Scientific), "Advanced LIGO", Class. Quant. Grav. **32**, 074001 (2015).

[23]M. Maggiore, "Gravitational waves volume 1: theory and experiments", in (Oxford, 2008), pp. 352–358.

[24]B. P. Abbott et al. (LIGO Scientific, Virgo), "Observation of Gravitational Waves from a Binary Black Hole Merger", Phys. Rev. Lett. **116**, 061102 (2016).

[25]R. Abbott et al. (LIGO Scientific, VIRGO), "GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run", Phys. Rev. D **109**, 022001 (2024).

[26]B. P. Abbott et al. (LIGO Scientific, Virgo), "The Rate of Binary Black Hole Mergers Inferred from Advanced LIGO Observations Surrounding GW150914", Astrophys. J. Lett. **833**, L1 (2016).

[27]B. P. Abbott et al. (LIGO Scientific, Virgo), "Tests of general relativity with GW150914", Phys. Rev. Lett. **116**, 221101 (2016).

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

[29]F. Acernese et al. (VIRGO), "Advanced Virgo: a second-generation interferometric gravitational wave detector", Class. Quant. Grav. **32**, 024001 (2015).

[30]B. P. Abbott et al. (LIGO Scientific, Virgo), "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral", Phys. Rev. Lett. **119**, 161101 (2017).

[31]D. J. Thompson and C. A. Wilson-Hodge, "Fermi Gamma-ray Space Telescope", 10.1007/978-981-16-4544-0_58-1 (2022).

[32]B. P. Abbott et al. (LIGO Scientific, Virgo, Fermi-GBM, INTEGRAL), "Gravitational Waves and Gamma-rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A", Astrophys. J. Lett. **848**, L13 (2017).

[33]B. P. Abbott et al. (LIGO Scientific, Virgo, ASKAP, BOOTES, DES, Fermi GBM, Fermi-LAT, GRAWITA, INTEGRAL, iPTF, InterPlanetary Network, J-GEM, La Silla-QUEST Survey, Liverpool Telescope, LOFAR, MASTER, MAXI, MWA, Pan-STARRS, PESSTO, Pi of the Sky, SkyMapper, Swift, C2PU, TOROS, VISTA), "Localization and broadband follow-up of the gravitational-wave transient GW150914", Astrophys. J. Lett. **826**, L13 (2016).

[34]B. P. Abbott et al. (LIGO Scientific, Virgo, Fermi GBM, INTEGRAL, IceCube, AstroSat Cadmium Zinc Telluride Imager Team, IPN, Insight-Hxmt, ANTARES, Swift, AGILE Team, 1M2H Team, Dark Energy Camera GW-EM, DES, DLT40, GRAWITA, Fermi-LAT, ATCA, ASKAP, Las Cumbres Observatory Group, OzGrav, DWF (Deeper Wider Faster Program), AST3, CAASTRO, VINROUGE, MASTER, J-GEM, GROWTH, JAGWAR, CaltechNRAO, TTU-NRAO, NuSTAR, Pan-STARRS, MAXI Team, TZAC Consortium, KU, Nordic Optical Telescope, ePESSTO, GROND, Texas Tech University, SALT Group, TOROS, BOOTES, MWA, CALET, IKI-GW Follow-up, H.E.S.S., LOFAR, LWA, HAWC, Pierre Auger, ALMA, Euro VLBI Team, Pi of Sky, Chandra Team at McGill University, DFN, ATLAS Telescopes, High Time Resolution Universe Survey, RIMAS, RATIR, SKA South Africa/MeerKAT), "Multi-messenger Observations of a Binary Neutron Star Merger", Astrophys. J. Lett. **848**, L12 (2017).

[35]B. P. Abbott et al. (LIGO Scientific, Virgo, 1M2H, Dark Energy Camera GW-E, DES, DLT40, Las Cumbres Observatory, VINROUGE, MASTER), "A gravitational-wave standard siren measurement of the Hubble constant", Nature **551**, 85–88 (2017).

[36]C. Nicolaou, O. Lahav, P. Lemos, W. Hartley, and J. Braden, "The Impact of Peculiar Velocities on the Estimation of the Hubble Constant from Gravitational Wave Standard Sirens", Mon. Not. Roy. Astron. Soc. **495**, 90–97 (2020).

[37]"Observation of Gravitational Waves from the Coalescence of a $2.5-4.5\,M_{\odot}$ Compact Object and a Neutron Star", (2024).

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

[39]M. Maggiore et al., "Science Case for the Einstein Telescope", JCAP **03**, 050 (2020).

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

[41]M. C. Begelman, R. D. Blandford, and M. J. Rees, "Massive black hole binaries in active galactic nuclei", Nature **287**, 307–309 (1980).

[42]S. Burke-Spolaor et al., "The Astrophysics of Nanohertz Gravitational Waves", Astron. Astrophys. Rev. **27**, 5 (2019).

[43]J. G. Hartnett and A. Luiten, "Colloquium: Comparison of Astrophysical and Terrestrial Frequency Standards", Rev. Mod. Phys. **83**, 1–9 (2011).

[44]R. S. Foster and D. C. Backer, "Constructing a Pulsar Timing Array", Astrophys. J. **361**, 300 (1990).

[45]R. w. Hellings and G. s. Downs, "UPPER LIMITS ON THE ISOTROPIC GRAVITATIONAL RADIATION BACKGROUND FROM PULSAR TIMING ANALYSIS", Astrophys. J. Lett. **265**, L39–L42 (1983).

[46]N. Steinle, H. Middleton, C. J. Moore, S. Chen, A. Klein, G. Pratten, R. Buscicchio, E. Finch, and A. Vecchio, "Implications of pulsar timing array observations for LISA detections of massive black hole binaries", Mon. Not. Roy. Astron. Soc. **525**, 2851–2863 (2023).

[47]J. Antoniadis et al. (EPTA, InPTA:), "The second data release from the European Pulsar Timing Array - III. Search for gravitational wave signals", Astron. Astrophys. **678**, A50 (2023).

[48]G. Agazie et al. (NANOGrav), "The NANOGrav 15 yr Data Set: Evidence for a Gravitational-wave Background", Astrophys. J. Lett. **951**, L8 (2023).

[49]D. J. Reardon et al., "Search for an Isotropic Gravitational-wave Background with the Parkes Pulsar Timing Array", Astrophys. J. Lett. **951**, L6 (2023).

[50]H. Xu et al., "Searching for the Nano-Hertz Stochastic Gravitational Wave Background with the Chinese Pulsar Timing Array Data Release I", Res. Astron. Astrophys. **23**, 075024 (2023).

[51]J. P. W. Verbiest, S. J. Vigeland, N. K. Porayko, S. Chen, and D. J. Reardon, "Status Report on Global Pulsar-Timing-Array Efforts to Detect Gravitational Waves", (2024).

[52]G. Agazie et al. (NANOGrav), "The NANOGrav 15 yr Data Set: Bayesian Limits on Gravitational Waves from Individual Supermassive Black Hole Binaries", Astrophys. J. Lett. **951**, L50 (2023).

[53]M. Colpi et al., "LISA Definition Study Report", (2024).

[54]G. Congedo, R. Dolesi, M. Hueller, S. Vitale, and W. J. Weber, "Space-borne gravitational wave detectors as time-delayed differential dynamometers", Phys. Rev. D **88**, 082003 (2013).

[55]K.-X. Sun, B. Allard, S. Buchman, S. Williams, and R. L. Byer, "LED deep UV source for charge management of gravitational reference sensors", Class. Quant. Grav. **23**, edited by N. Mio, S141–S150 (2006).

[56]M. Armano et al. (LISA Pathfinder), "Transient acceleration events in LISA Pathfinder data: Properties and possible physical origin", Phys. Rev. D **106**, 062001 (2022).

[57]G. Heinzel, J. José Esteban, S. Barke, M. Otto, Y. Wang, A. F. Garcia, and K. Danzmann, "Auxiliary functions of the LISA laser link: ranging, clock noise transfer and data communication", Classical and Quantum Gravity **28**, 094008, 094008 (2011).

[58]M. Tinto and J. W. Armstrong, "Cancellation of laser noise in an unequal-arm interferometer detector of gravitational radiation", Phys. Rev. D **59**, 102003 (1999).

[59]M. Tinto and S. V. Dhurandhar, "Time-delay interferometry", Living Rev. Rel. **24**, 1 (2021).

[60]M. Vallisneri, "Geometric time delay interferometry", Phys. Rev. D **72**, 042003 (2005).

[61]M. Tinto, F. B. Estabrook, and J. W. Armstrong, "Time delay interferometry for LISA", Phys. Rev. D **65**, 082003 (2002).

[62]J. J. Esteban, I. Bykov, A. F. Garcia Marin, G. Heinzel, and K. Danzmann, "Optical ranging and data transfer development for LISA", J. Phys. Conf. Ser. **154**, edited by A. Lobo and C. F. Sopuerta, 012025 (2009).

[63]J. N. Reinhardt, M. Staab, K. Yamamoto, J.-B. Bayle, A. Hees, O. Hartwig, K. Wiesner, S. Shah, and G. Heinzel, "Ranging sensor fusion in LISA data processing: Treatment of ambiguities, noise, and onboard delays in LISA ranging observables", Phys. Rev. D **109**, 022004 (2024).

[64]J. J. Esteban, A. F. García, S. Barke, A. M. Peinado, F. G. Cervantes, I. Bykov, G. Heinzel, and K. Danzmann, "Experimental demonstration of weak-light laser ranging and data communication for lisa", Opt. Express **19**, 15937–15946 (2011).

[65]M. Tinto, F. B. Estabrook, and J. W. Armstrong, "Time delay interferometry with moving spacecraft arrays", Phys. Rev. D **69**, 082001 (2004).

[66]Y. Wang, G. Heinzel, and K. Danzmann, "First stage of LISA data processing: Clock synchronization and arm-length determination via a hybrid-extended Kalman filter", Phys. Rev. D **90**, 064016 (2014).

[67]S. Prieaux, "Time scales in LISA", Class. Quant. Grav. **24**, 2271–2281 (2007).

[68]R. Hellings, G. Giampieri, L. Maleki, M. Tinto, K. Danzmann, J. Hough, and D. Robertson, "Heterodyne laser tracking at high doppler rates", Optics Communications **124**, 313–320 (1996).

[69]O. Hartwig and J.-B. Bayle, "Clock-jitter reduction in LISA time-delay interferometry combinations", Phys. Rev. D **103**, 123027 (2021).

[70]R. W. Hellings, "Elimination of clock jitter noise in spaceborne laser interferometers", Phys. Rev. D **64**, 022002 (2001).

[71]O. Hartwig, J.-B. Bayle, M. Staab, A. Hees, M. Lilley, and P. Wolf, "Time-delay interferometry without clock synchronization", Phys. Rev. D **105**, 122008 (2022).

[72]P. McNamara, S. Vitale, and K. Danzmann (LISA), "LISA Pathfinder", Class. Quant. Grav. **25**, edited by S. M. Scott and D. E. McClelland, 114034 (2008).

[73]M. Armano et al., "Sub-Femto- g Free Fall for Space-Based Gravitational Wave Observatories: LISA Pathfinder Results", Phys. Rev. Lett. **116**, 231101 (2016).

[74]S. Anza et al., "The LTP experiment on the LISA Pathfinder mission", Class. Quant. Grav. **22**, edited by O. Jennrich, S125–S138 (2005).

[75]M. Armano et al., "Beyond the Required LISA Free-Fall Performance: New LISA Pathfinder Results down to 20 μHz", Phys. Rev. Lett. **120**, 061101 (2018).

[76]L. Consortium, *Lisa performance model and error budget, lisa-lcst-inst-tn-003*, tech. rep. (ESA, 2020).

[77]T. Robson, N. J. Cornish, and C. Liu, "The construction and use of LISA sensitivity curves", Class. Quant. Grav. **36**, 105011 (2019).

[78]M. Zevin et al., "Gravity Spy: lessons learned and a path forward", Eur. Phys. J. Plus **139**, 100 (2024).

[79]J. Glanzer et al., "Data quality up to the third observing run of advanced LIGO: Gravity Spy glitch classifications", Class. Quant. Grav. **40**, 065004 (2023).

[80]P. A. Seoane et al. (LISA), "Astrophysics with the Laser Interferometer Space Antenna", Living Rev. Rel. **26**, 2 (2023).

[81]G. Duchêne and A. Kraus, "Stellar Multiplicity", Ann. Rev. Astron. Astrophys. **51**, 269 (2013).

[82]H. Sana, S. E. de Mink, A. de Koter, N. Langer, C. J. Evans, M. Gieles, E. Gosset, R. G. Izzard, J. .-. Le Bouquin, and F. R. N. Schneider, "Binary Interaction Dominates the Evolution of Massive Stars", Science **337**, 444 (2012).

[83]B. Paczynski, "Common Envelope Binaries", in Structure and evolution of close binary systems, Vol. 73, edited by P. Eggleton, S. Mitton, and J. Whelan (Jan. 1976), p. 75.

[84]I. Iben Jr. and M. Livio, "Common envelopes in binary star evolution", Publ. Astron. Soc. Pac. **105**, 1373–1406 (1993).

[85]K. Belczynski, D. E. Holz, T. Bulik, and R. O'Shaughnessy, "The first gravitational-wave source from the isolated evolution of two 40-100 Msun stars", Nature **534**, 512 (2016).

[86]E. P. J. van den Heuvel, S. F. Portegies Zwart, and S. E. de Mink, "Forming short-period Wolf–Rayet X-ray binaries and double black holes through stable mass transfer", Mon. Not. Roy. Astron. Soc. **471**, 4256–4264 (2017).

[87]S. S. Bavera et al., "The impact of mass-transfer physics on the observable properties of field binary black hole populations", Astron. Astrophys. **647**, A153 (2021).

[88]M. Gallegos-Garcia, C. P. L. Berry, P. Marchant, and V. Kalogera, "Binary Black Hole Formation with Detailed Modeling: Stable Mass Transfer Leads to Lower Merger Rates", Astrophys. J. **922**, 110 (2021).

[89]S. Nissanke, M. Vallisneri, G. Nelemans, and T. A. Prince, "Gravitational-wave emission from compact Galactic binaries", Astrophys. J. **758**, 131 (2012).

[90]V. Korol and M. Safarzadeh, "How can LISA probe a population of GW190425-like binary neutron stars in the Milky Way?", Mon. Not. Roy. Astron. Soc. **502**, 5576–5583 (2021).

[91]M. Perpinyà-Vallès, A. Rebassa-Mansergas, B. T. Gänsicke, S. Toonen, J. J. Hermes, N. P. Gentile Fusillo, and P. .-. Tremblay, "Discovery of the first resolved triple white dwarf", Mon. Not. Roy. Astron. Soc. **483**, 901–907 (2019).

[92]M. L. Katz, C. Danielski, N. Karnesis, V. Korol, N. Tamanini, N. J. Cornish, and T. B. Littenberg, "Bayesian characterization of circumbinary sub-stellar objects with LISA", Mon. Not. Roy. Astron. Soc. **517**, 697–711 (2022).

[93]K. Breivik, K. Kremer, M. Bueno, S. L. Larson, S. Coughlin, and V. Kalogera, "Characterizing Accreting Double White Dwarf Binaries with the Laser Interferometer Space Antenna and Gaia", Astrophys. J. Lett. **854**, L1 (2018).

[94]J.-E. Solheim, "AM CVn Stars: Status and Challenges", Publications of the Astronomical Society of the Pacific **122**, 1133 (2010).

[95]G. Nelemans, L. R. Yungelson, and S. F. Portegies Zwart, "Short- period AM CVn systems as optical, x-ray and gravitational wave sources", Mon. Not. Roy. Astron. Soc. **349**, 181 (2004).

[96]S. Scaringi, K. Breivik, T. B. Littenberg, C. Knigge, P. J. Groot, and M. Veresvarska, "Cataclysmic variables are a key population of gravitational wave sources for LISA", Mon. Not. Roy. Astron. Soc. **525**, L50–L55 (2023).

[97]C. Knigge, I. Baraffe, and J. Patterson, "The Evolution of Cataclysmic Variables as Revealed by their Donor Stars", Astrophys. J. Suppl. **194**, 28 (2011).

[98]R. F. Webbink, "The evolution of low-mass close binary systems. VI. Population II W Ursae Majoris systems.", Astrophys. Journal **227**, 178–184 (1979).

[99]S. Rappaport, P. C. Joss, and R. F. Webbink, "The evolution of highly compact binary stellar systems.", Astrophys. Journal **254**, 616–640 (1982).

[100]A. J. Ruiter, K. Belczynski, M. Benacquista, S. L. Larson, and G. Williams, "The LISA Gravitational Wave Foreground: A Study of Double White Dwarfs", Astrophys. J. **717**, 1006–1021 (2010).

[101] V. Korol et al., "Populations of double white dwarfs in Milky Way satellites and their detectability with LISA", Astron. Astrophys. **638**, A153 (2020).

[102] S. Rieck, A. W. Criswell, V. Korol, M. A. Keim, M. Bloom, and V. Mandic, "A Stochastic Gravitational Wave Background in LISA from Unresolved White Dwarf Binaries in the Large Magellanic Cloud", (2023).

[103] V. Korol, O. Koop, and E. M. Rossi, "Detectability of double white dwarfs in the Local Group with LISA", Astrophys. J. Lett. **866**, L20 (2018).

[104] K. Belczynski, M. Benacquista, and T. Bulik, "Double Compact Objects as Low-frequency Gravitational Wave Sources", Astrophys. J. **725**, 816–823 (2010).

[105] V. Kalogera, U. Kolb, and A. R. King, "Supernova kicks, magnetic braking, and neutron star - binaries", Astrophys. J. **504**, 967 (1998).

[106] H. .-. Janka, "Natal Kicks of Stellar-Mass Black Holes by Asymmetric Mass Ejection in Fallback Supernovae", Mon. Not. Roy. Astron. Soc. **434**, 1355 (2013).

[107] M. Y. M. Lau, I. Mandel, A. Vigna-Gómez, C. J. Neijssel, S. Stevenson, and A. Sesana, "Detecting Double Neutron Stars with LISA", Mon. Not. Roy. Astron. Soc. **492**, 3061–3072 (2020).

[108] C. J. Moore, E. Finch, A. Klein, V. Korol, N. Pham, and D. Robins, "Discovering neutron stars with LISA via measurements of orbital eccentricity in Galactic binaries", (2023).

[109] A. Sesana, A. Lamberts, and A. Petiteau, "Finding binary black holes in the Milky Way with LISA", Mon. Not. Roy. Astron. Soc. **494**, L75–L80 (2020).

[110] T. Kupfer, V. Korol, S. Shah, G. Nelemans, T. R. Marsh, G. Ramsay, P. J. Groot, D. T. H. Steeghs, and E. M. Rossi, "LISA verification binaries with updated distances from Gaia Data Release 2", Mon. Not. Roy. Astron. Soc. **480**, 302–309 (2018).

[111] R. Buscicchio, A. Klein, E. Roebber, C. J. Moore, D. Gerosa, E. Finch, and A. Vecchio, "Bayesian parameter estimation of stellar-mass black-hole binaries with LISA", Phys. Rev. D **104**, 044065 (2021).

[112] N. Muttoni, A. Mangiagli, A. Sesana, D. Laghi, W. Del Pozzo, D. Izquierdo-Villalba, and M. Rosati, "Multiband gravitational wave cosmology with stellar origin black hole binaries", Phys. Rev. D **105**, 043509 (2022).

[113] A. Sesana, "Prospects for Multiband Gravitational-Wave Astronomy after GW150914", Phys. Rev. Lett. **116**, 231102 (2016).

[114] A. Stroeer and A. Vecchio, "The LISA verification binaries", Class. Quant. Grav. **23**, edited by M. Diaz and S. Mohanty, S809–S818 (2006).

[115] A. G. A. Brown, "Microarcsecond Astrometry: Science Highlights from Gaia", Annual Review of Astronomy and Astrophysics **59**, 59–115 (2021).

[116]T. B. Littenberg and N. J. Cornish, "Prospects for Gravitational Wave Measurement of ZTF J1539+5027", Astrophys. J. Lett. **881**, L43 (2019).

[117]T. Kupfer et al., "LISA Galactic Binaries with Astrometry from Gaia DR3", Astrophys. J. **963**, 100 (2024).

[118]T. Prusti et al. (Gaia), "The Gaia Mission", Astron. Astrophys. **595**, A1 (2016).

[119]E. Finch, G. Bartolucci, D. Chucherko, B. G. Patterson, V. Korol, A. Klein, D. Bandopadhyay, H. Middleton, C. J. Moore, and A. Vecchio, "Identifying LISA verification binaries among the Galactic population of double white dwarfs", Mon. Not. Roy. Astron. Soc. **522**, 5358–5373 (2023).

[120]E. Savalle, J. Gair, L. Speri, and S. Babak, "Assessing the impact of instrumental calibration uncertainty on LISA science", Phys. Rev. D **106**, 022003 (2022).

[121]T. B. Littenberg and A. K. Lali, "Have any LISA verification binaries been found?", (2024).

[122]N. Ivanova, S. Justham, X. Chen, O. De Marco, C. L. Fryer, E. Gaburov, H. Ge, E. Glebbeek, Z. Han, X. .-. Li, G. Lu, T. Marsh, P. Podsiadlowski, A. Potter, N. Soker, R. Taam, T. M. Tauris, E. P. J. van den Heuvel, and R. F. Webbink, "Common envelope evolution: where we stand and how we can move forward", Astron. and Astrophys. Reviews **21**, 59, 59 (2013).

[123]G. et al., "Gaia Data Release 3. Summary of the content and survey properties", Astron. and Astrophys. **674**, A1, A1 (2023).

[124]A. King, "Black holes, galaxy formation, and the $M_{\rm BH}$ - σ relation", Astrophys. J. Lett. **596**, L27–L30 (2003).

[125]L. Ferrarese and H. Ford, "Supermassive Black Holes in Galactic Nuclei: Past, Present and Future Research", Space Science Reviews **116**, 523–624 (2005).

[126]K. Inayoshi, E. Visbal, and Z. Haiman, "The Assembly of the First Massive Black Holes", Ann. Rev. Astron. Astrophys. **58**, 27–97 (2020).

[127]M. Volonteri, M. Habouzit, and M. Colpi, "The origins of massive black holes", Nature Rev. Phys. **3**, 732–743 (2021).

[128]D. J. Mortlock et al., "A luminous quasar at a redshift of z = 7.085", Nature **474**, 616 (2011).

[129]F. Wang, J. Yang, X. Fan, J. F. Hennawi, A. J. Barth, E. Banados, F. Bian, K. Boutsia, T. Connor, F. B. Davies, R. Decarli, A.-C. Eilers, E. P. Farina, R. Green, L. Jiang, J.-T. Li, C. Mazzucchelli, R. Nanni, J.-T. Schindler, B. Venemans, F. Walter, X.-B. Wu, and M. Yue, "A Luminous Quasar at Redshift 7.642", Astrophys. Journal Letters **907**, L1, L1 (2021).

[130]O. E. Kovacs et al., "A candidate supermassive black hole in a gravitationally-lensed galaxy at $z \approx 10$", (2024).

[131]J. Magorrian et al., "The Demography of massive dark objects in galaxy centers", Astron. J. **115**, 2285 (1998).

[132]N. J. McConnell, C.-P. Ma, K. Gebhardt, S. A. Wright, J. D. Murphy, T. R. Lauer, J. R. Graham, and D. O. Richstone, "Two ten-billion-solar-mass black holes at the centres of giant elliptical galaxies", Nature **480**, 215 (2011).

[133]J. Kormendy and L. C. Ho, "Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies", Ann. Rev. Astron. Astrophys. **51**, 511–653 (2013).

[134]M. Habouzit, M. Onoue, E. Banados, M. Neeleman, D. Angles-Alcazar, F. Walter, A. Pillepich, R. Dave, K. Jahnke, and Y. Dubois, "Co-evolution of massive black holes and their host galaxies at high redshift: discrepancies from six cosmological simulations and the key role of JWST", Mon. Not. Roy. Astron. Soc. **511**, 3751–3767 (2022).

[135]F. Bournaud, C. J. Jog, and F. Combes, "Multiple minor mergers: Formation of elliptical galaxies and constraints for the growth of spiral disks", Astron. Astrophys. **476**, 1179 (2007).

[136]G. De Lucia, V. Springel, S. D. M. White, D. Croton, and G. Kauffmann, "The formation history of elliptical galaxies", Mon. Not. Roy. Astron. Soc. **366**, 499–509 (2006).

[137]J. Binney and S. Tremaine, *Galactic Dynamics: Second Edition* (2008).

[138]E. Bortolas, A. Franchini, M. Bonetti, and A. Sesana, "The Competing Effect of Gas and Stars in the Evolution of Massive Black Hole Binaries", Astrophys. J. Lett. **918**, L15 (2021).

[139]A. Sesana, F. Haardt, and P. Madau, "Interaction of massive black hole binaries with their stellar environment. 1. Ejection of hypervelocity stars", Astrophys. J. **651**, 392–400 (2006).

[140]G. D. Quinlan, "The dynamical evolution of massive black hole binaries - I. hardening in a fixed stellar background", New Astron. **1**, 35–56 (1996).

[141]M. Milosavljevic and D. Merritt, "Long term evolution of massive black hole binaries", Astrophys. J. **596**, 860 (2003).

[142]M. Milosavljevic and D. Merritt, "The Final parsec problem", AIP Conf. Proc. **686**, edited by J. M. Centrella, 201–210 (2003).

[143]A. Gualandris, J. I. Read, W. Dehnen, and E. Bortolas, "Collisionless loss-cone refilling: there is no final parsec problem", Mon. Not. Roy. Astron. Soc. **464**, 2301–2310 (2017).

[144]E. Bortolas, A. Gualandris, M. Dotti, M. Spera, and M. Mapelli, "Brownian motion of massive black hole binaries and the final parsec problem", Mon. Not. Roy. Astron. Soc. **461**, 1023–1031 (2016).

[145]P. Berczik, D. Merritt, R. Spurzem, and H.-P. Bischof, "Efficient merger of binary supermassive black holes in non-axisymmetric galaxies", Astrophys. J. Lett. **642**, L21–L24 (2006).

[146]L. Zwick, P. R. Capelo, E. Bortolas, V. Vazquez-Aceves, L. Mayer, and P. Amaro-Seoane, "Improved gravitational radiation time-scales II: Spin–orbit contributions and environmental perturbations", Mon. Not. Roy. Astron. Soc. **506**, 1007–1018 (2021).

[147]M. Garg, A. Derdzinski, S. Tiwari, J. Gair, and L. Mayer, "Measuring eccentricity and gas-induced perturbation from gravitational waves of LISA massive black hole binaries", (2024).

[148]Y. Tang, Z. Haiman, and A. Macfadyen, "The late inspiral of supermassive black hole binaries with circumbinary gas discs in the LISA band", Mon. Not. Roy. Astron. Soc. **476**, 2249–2257 (2018).

[149]C. Lacey and S. Cole, "Merger rates in hierarchical models of galaxy formation", Mon. Not. Roy. Astron. Soc. **262**, 627–649 (1993).

[150]M. L. Katz, L. Z. Kelley, F. Dosopoulou, S. Berry, L. Blecha, and S. L. Larson, "Probing Massive Black Hole Binary Populations with LISA", Mon. Not. Roy. Astron. Soc. **491**, 2301–2317 (2020).

[151]E. Barausse, I. Dvorkin, M. Tremmel, M. Volonteri, and M. Bonetti, "Massive Black Hole Merger Rates: The Effect of Kiloparsec Separation Wandering and Supernova Feedback", Astrophys. J. **904**, 16 (2020).

[152]E. Berti, "Lisa observations of massive black hole mergers: event rates and issues in waveform modelling", Class. Quant. Grav. **23**, edited by M. Diaz and S. Mohanty, S785–S798 (2006).

[153]Q. Wu and Y. Shen, "A Catalog of Quasar Properties from Sloan Digital Sky Survey Data Release 16", Astrophys. J. Suppl. **263**, 42 (2022).

[154]T. Alexander and P. Natarajan, "Rapid growth of seed black holes in the early universe by supra-exponential accretion", Science **345**, 1330 (2014).

[155]O. Piana, P. Dayal, M. Volonteri, and T. R. Choudhury, "The mass assembly of high-redshift black holes", Mon. Not. Roy. Astron. Soc. **500**, 2146–2158 (2020).

[156]Z. Haiman and A. Loeb, "What is the highest plausible redshift of luminous quasars?", Astrophys. J. **552**, 459 (2001).

[157]B. Carr and J. Silk, "Primordial Black Holes as Generators of Cosmic Structures", Mon. Not. Roy. Astron. Soc. **478**, 3756–3775 (2018).

[158]A. Klein et al., "Science with the space-based interferometer eLISA: Supermassive black hole binaries", Phys. Rev. D **93**, 024003 (2016).

[159]M. Siwek, L. Z. Kelley, and L. Hernquist, "Signatures of Circumbinary Disk Dynamics in Multi-Messenger Population Studies of Massive Black Hole Binaries", (2024).

[160]T. Dal Canton, A. Mangiagli, S. C. Noble, J. Schnittman, A. Ptak, A. Klein, A. Sesana, and J. Camp, "Detectability of modulated X-rays from LISA's supermassive black hole mergers", Astrophys. J. **886**, 146 (2019).

[161]A. Khan, V. Paschalidis, M. Ruiz, and S. L. Shapiro, "Disks Around Merging Binary Black Holes: From GW150914 to Supermassive Black Holes", Phys. Rev. D **97**, 044036 (2018).

[162]E. M. Rossi, G. Lodato, P. J. Armitage, J. E. Pringle, and A. R. King, "Black hole mergers: the first light", Mon. Not. Roy. Astron. Soc. **401**, 2021 (2010).

[163]B. F. Schutz, "Determining the Hubble Constant from Gravitational Wave Observations", Nature **323**, 310–311 (1986).

[164]A. Mangiagli, C. Caprini, S. Marsat, L. Speri, R. R. Caldwell, and N. Tamanini, "Massive black hole binaries in LISA: constraining cosmological parameters at high redshifts", (2023).

[165]N. Tamanini, "Late time cosmology with LISA: probing the cosmic expansion with massive black hole binary mergers as standard sirens", J. Phys. Conf. Ser. **840**, edited by D. Giardini and P. Jetzer, 012029 (2017).

[166]J. M. Ezquiaga and D. E. Holz, "Spectral Sirens: Cosmology from the Full Mass Distribution of Compact Binaries", Phys. Rev. Lett. **129**, 061102 (2022).

[167]G. Dálya et al., "GLADE+ : an extended galaxy catalogue for multimessenger searches with advanced gravitational-wave detectors", Mon. Not. Roy. Astron. Soc. **514**, 1403–1411 (2022).

[168]W. Israel, "Event horizons in static electrovac space-times", Commun. Math. Phys. **8**, 245–260 (1968).

[169]B. Carter, "Axisymmetric Black Hole Has Only Two Degrees of Freedom", Phys. Rev. Lett. **26**, 331–333 (1971).

[170]S. L. Liebling and C. Palenzuela, "Dynamical boson stars", Living Rev. Rel. **26**, 1 (2023).

[171]D. R. Mayerson, "Fuzzballs and Observations", Gen. Rel. Grav. **52**, 115 (2020).

[172]E. Berti, K. Yagi, H. Yang, and N. Yunes, "Extreme Gravity Tests with Gravitational Waves from Compact Binary Coalescences: (II) Ringdown", Gen. Rel. Grav. **50**, 49 (2018).

[173]F. Echeverria, "Gravitational Wave Measurements of the Mass and Angular Momentum of a Black Hole", Phys. Rev. D **40**, 3194–3203 (1989).

[174]L. Z. Kelley, L. Blecha, and L. Hernquist, "Massive Black Hole Binary Mergers in Dynamical Galactic Environments", Mon. Not. Roy. Astron. Soc. **464**, 3131–3157 (2017).

[175]N. Neumayer, A. Seth, and T. Böker, "Nuclear star clusters", Astron. and Astrophys. Reviews **28**, 4, 4 (2020).

[176]C. P. L. Berry, S. A. Hughes, C. F. Sopuerta, A. J. K. Chua, A. Heffernan, K. Holley-Bockelmann, D. P. Mihaylov, M. C. Miller, and A. Sesana, "The unique potential of extreme mass-ratio inspirals for gravitational-wave astronomy", (2019).

[177]M. L. Katz, A. J. K. Chua, L. Speri, N. Warburton, and S. A. Hughes, "Fast extreme-mass-ratio-inspiral waveforms: New tools for millihertz gravitational-wave data analysis", Phys. Rev. D **104**, 064047 (2021).

[178]P. Amaro-Seoane, "Relativistic dynamics and extreme mass ratio inspirals", Living Rev. Rel. **21**, 4 (2018).

[179]T. Alexander, "EMRIs and the relativistic loss-cone: The curious case of the fortunate coincidence", J. Phys. Conf. Ser. **840**, edited by D. Giardini and P. Jetzer, 012019 (2017).

[180]S. Naoz, S. C. Rose, E. Michaely, D. Melchor, E. Ramirez-Ruiz, B. Mockler, and J. D. Schnittman, "The Combined Effects of Two-body Relaxation Processes and the Eccentric Kozai–Lidov Mechanism on the Extreme-mass-ratio Inspirals Rate", Astrophys. J. Lett. **927**, L18 (2022).

[181]E. Bortolas and M. Mapelli, "Can supernova kicks trigger EMRIs in the Galactic Centre?", Mon. Not. Roy. Astron. Soc. **485**, 2125–2138 (2019).

[182]C. Hopman and T. Alexander, "The Orbital statistics of stellar inspiral and relaxation near a massive black hole: Characterizing gravitational wave sources", Astrophys. J. **629**, 362–372 (2005).

[183]C. P. L. Berry and J. R. Gair, "Expectations for extreme-mass-ratio bursts from the Galactic Centre", Mon. Not. Roy. Astron. Soc. **435**, 3521–3540 (2013).

[184]C. P. L. Berry and J. R. Gair, "Extreme-mass-ratio-bursts from extragalactic sources", Mon. Not. Roy. Astron. Soc. **433**, 3572–3583 (2013).

[185]D. J. Oliver, A. D. Johnson, J. Berrier, K. Glampedakis, and D. Kennefick, "Gravitational Wave Peeps from EMRIs and their Implication for LISA Signal Confusion Noise", (2023).

[186]L. Speri, A. Antonelli, L. Sberna, S. Babak, E. Barausse, J. R. Gair, and M. L. Katz, "Probing Accretion Physics with Gravitational Waves", Phys. Rev. X **13**, 021035 (2023).

[187]A. Franchini, M. Bonetti, A. Lupi, G. Miniutti, E. Bortolas, M. Giustini, M. Dotti, A. Sesana, R. Arcodia, and T. Ryu, "Quasi-periodic eruptions from impacts between the secondary and a rigidly precessing accretion disc in an extreme mass-ratio inspiral system", Astron. Astrophys. **675**, A100 (2023).

[188]Y. Y. Wang, F. Y. Wang, Y. C. Zou, and Z. G. Dai, "A bright electromagnetic counterpart to extreme mass ratio inspirals", Astrophys. J. Lett. **886**, L22 (2019).

[189]T. Bogdanovic, R. M. Cheng, and P. Amaro-Seoane, "Disruption of a Red Giant Star by a Supermassive Black Hole and the Case of PS1-10jh", Astrophys. J. **788**, 99 (2014).

[190]S. Babak, J. Gair, A. Sesana, E. Barausse, C. F. Sopuerta, C. P. L. Berry, E. Berti, P. Amaro-Seoane, A. Petiteau, and A. Klein, "Science with the space-based interferometer LISA. V: Extreme mass-ratio inspirals", Phys. Rev. D **95**, 103012 (2017).

[191]L. Broggi, E. Bortolas, M. Bonetti, A. Sesana, and M. Dotti, "Extreme mass ratio inspirals and tidal disruption events in nuclear clusters – I. Time-dependent rates", Mon. Not. Roy. Astron. Soc. **514**, 3270–3284 (2022).

[192]Z. Pan, Z. Lyu, and H. Yang, "Wet extreme mass ratio inspirals may be more common for spaceborne gravitational wave detection", Phys. Rev. D **104**, 063007 (2021).

[193]R. Abbott et al. (KAGRA, VIRGO, LIGO Scientific), "Population of Merging Compact Binaries Inferred Using Gravitational Waves through GWTC-3", Phys. Rev. X **13**, 011048 (2023).

[194]M. L. Katz, "Fully automated end-to-end pipeline for massive black hole binary signal extraction from LISA data", Phys. Rev. D **105**, 044055 (2022).

[195]L. Brenneman, "Measuring Supermassive Black Hole Spins in Active Galactic Nuclei", 10.1007/978-1-4614-7771-6 (2013).

[196]P. Amaro-Seoane, C. F. Sopuerta, and M. D. Freitag, "The role of the supermassive black hole spin in the estimation of the EMRI event rate", Mon. Not. Roy. Astron. Soc. **429**, 3155–3165 (2013).

[197]A. Sesana, E. Barausse, M. Dotti, and E. M. Rossi, "Linking the spin evolution of massive black holes to galaxy kinematics", Astrophys. J. **794**, 104 (2014).

[198]J. R. Gair, M. Vallisneri, S. L. Larson, and J. G. Baker, "Testing General Relativity with Low-Frequency, Space-Based Gravitational-Wave Detectors", Living Rev. Rel. **16**, 7 (2013).

[199]T. Jiang, N. Dai, Y. Gong, D. Liang, and C. Zhang, "Constraint on Brans-Dicke theory from intermediate/extreme mass ratio inspirals", JCAP **12**, 023 (2022).

[200]J. Tan, J.-d. Zhang, H.-M. Fan, and J. Mei, "Constraining the EdGB Theory with Extreme Mass-Ratio Inspirals", (2024).

[201]P. Canizares, J. R. Gair, and C. F. Sopuerta, "Testing Chern-Simons Modified Gravity with Gravitational-Wave Detections of Extreme-Mass-Ratio Binaries", Phys. Rev. D **86**, 044010 (2012).

[202]C. Zhang, Y. Gong, D. Liang, and B. Wang, "Gravitational waves from eccentric extreme mass-ratio inspirals as probes of scalar fields", JCAP **06**, 054 (2023).

[203]T. Zi and P.-C. Li, "Gravitational waves from extreme-mass-ratio inspirals around a hairy Kerr black hole", Phys. Rev. D **108**, 084001 (2023).

[204]E. Babichev, C. Charmousis, D. D. Doneva, G. N. Gyulchev, and S. S. Yazadjiev, "Testing disformal non-circular deformation of Kerr black holes with LISA", (2024).

[205]J. Lestingi, E. Cannizzaro, and P. Pani, "Extreme mass-ratio inspirals as probes of fundamental dipoles", Phys. Rev. D **109**, 044052 (2024).

[206]E. Maggio, M. van de Meent, and P. Pani, "Extreme mass-ratio inspirals around a spinning horizonless compact object", Phys. Rev. D **104**, 104026 (2021).

[207]C. M. Will, "Bounding the mass of the graviton using gravitational wave observations of inspiralling compact binaries", Phys. Rev. D **57**, 2061–2068 (1998).

[208]V. Cardoso, G. Castro, and A. Maselli, "Gravitational waves in massive gravity theories: waveforms, fluxes and constraints from extreme-mass-ratio mergers", Phys. Rev. Lett. **121**, 251103 (2018).

[209]C. Liu, D. Laghi, and N. Tamanini, "Probing modified gravitational-wave propagation with extreme mass-ratio inspirals", Phys. Rev. D **109**, 063521 (2024).

[210]B. Bonga, H. Yang, and S. A. Hughes, "Tidal resonance in extreme mass-ratio inspirals", Phys. Rev. Lett. **123**, 101103 (2019).

[211]P. Gupta, L. Speri, B. Bonga, A. J. K. Chua, and T. Tanaka, "Modeling transient resonances in extreme-mass-ratio inspirals", Phys. Rev. D **106**, 104001 (2022).

[212]C. Zhang, G. Fu, and N. Dai, "Detecting dark matter with extreme mass-ratio inspirals", (2024).

[213]G.-L. Li, Y. Tang, and Y.-L. Wu, "Probing dark matter spikes via gravitational waves of extreme-mass-ratio inspirals", Sci. China Phys. Mech. Astron. **65**, 100412 (2022).

[214]Z.-C. Zhang and Y. Tang, "Velocity Distribution of Dark Matter Spike around Schwarzschild Black Holes and Effects on Gravitational Waves from EMRIs", (2024).

[215]S. Kejriwal, L. Speri, and A. J. K. Chua, "Impact of Correlations on the Modeling and Inference of Beyond Vacuum-GR Effects in Extreme-Mass-Ratio Inspirals", (2023).

[216]S. Marsat, J. G. Baker, and T. Dal Canton, "Exploring the Bayesian parameter estimation of binary black holes with LISA", Phys. Rev. D **103**, 083011 (2021).

[217]D. Laghi, N. Tamanini, W. Del Pozzo, A. Sesana, J. Gair, S. Babak, and D. Izquierdo-Villalba, "Gravitational-wave cosmology with extreme mass-ratio inspirals", Mon. Not. Roy. Astron. Soc. **508**, 4512–4531 (2021).

[218]C. L. MacLeod and C. J. Hogan, "Precision of Hubble constant derived using black hole binary absolute distances and statistical redshift information", Phys. Rev. D **77**, 043512 (2008).

[219]L.-G. Zhu, H.-M. Fan, X. Chen, Y.-M. Hu, and J.-d. Zhang, "Improving the Cosmological Constraints by Inferring the Formation Channel of Extreme-mass-ratio Inspirals", (2024).

[220]M. Toscani, O. Burke, C. Liu, N. B. Zamel, N. Tamanini, and F. Pozzoli, "Strongly-Lensed Extreme Mass-ratio Inspirals", (2023).

[221]A. Derdzinski and L. Zwick, "Multimessenger Astronomy with Black Holes: Extreme mass ratio inspirals", in (Oct. 2023).

[222]S. Staelens and G. Nelemans, "Likelihood of white dwarf binaries to dominate the astrophysical gravitational wave background in the mHz band", Astron. Astrophys. **683**, A139 (2024).

[223]R. Umstatter, N. Christensen, M. Hendry, R. Meyer, V. Simha, J. Veitch, S. Vigeland, and G. Woan, "LISA source confusion: Identification and characterization of signals", Class. Quant. Grav. **22**, edited by B. Mours and F. Marion, S901–S912 (2005).

[224]S. Banagiri, A. Criswell, T. Kuan, V. Mandic, J. D. Romano, and S. R. Taylor, "Mapping the gravitational-wave sky with LISA: a Bayesian spherical harmonic approach", Mon. Not. Roy. Astron. Soc. **507**, 5451–5462 (2021).

[225]S. Babak, C. Caprini, D. G. Figueroa, N. Karnesis, P. Marcoccia, G. Nardini, M. Pieroni, A. Ricciardone, A. Sesana, and J. Torrado, "Stochastic gravitational wave background from stellar origin binary black holes in LISA", JCAP **08**, 034 (2023).

[226]M. Bonetti and A. Sesana, "Gravitational wave background from extreme mass ratio inspirals", Phys. Rev. D **102**, 103023 (2020).

[227]F. Pozzoli, S. Babak, A. Sesana, M. Bonetti, and N. Karnesis, "Computation of stochastic background from extreme-mass-ratio inspiral populations for LISA", Phys. Rev. D **108**, 103039 (2023).

[228]P. Auclair, K. Leyde, and D. A. Steer, "A window for cosmic strings", JCAP **04**, 005 (2023).

[229]C.-F. Chang and Y. Cui, "Stochastic Gravitational Wave Background from Global Cosmic Strings", Phys. Dark Univ. **29**, 100604 (2020).

[230]N. Christensen, "Stochastic Gravitational Wave Backgrounds", Rept. Prog. Phys. **82**, 016903 (2019).

[231]E. Barausse, R. Brito, V. Cardoso, I. Dvorkin, and P. Pani, "The stochastic gravitational-wave background in the absence of horizons", Class. Quant. Grav. **35**, 20LT01 (2018).

[232]M. Muratore, J. Gair, and L. Speri, "Impact of the noise knowledge uncertainty for the science exploitation of cosmological and astrophysical stochastic gravitational wave background with LISA", Phys. Rev. D **109**, 042001 (2024).

[233]T. Bayes, "An essay towards solving a problem in the doctrine of chances", Phil. Trans. of the Royal Soc. of London **53**, 370–418 (1763).

[234]D. S. Sivia and J. Skilling, *Data analysis: a bayesian tutorial*, 2nd. (Oxford University Press, New York, 2006), pp. 14–102.

[235]E. Thrane and C. Talbot, "An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models", Pub. of the Astron. Soc. of Australia **36**, e010, e010 (2019).

[236]C. A. L. Bailer-Jones, *Practical bayesian inference: a primer for physical scientists* (Cambridge University Press, Cambridge, 2017), pp. 4–35.

[237]D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, "emcee: The MCMC Hammer", Publ. Astron. Soc. Pac. **125**, 306–312 (2013).

[238]N. Karnesis, M. L. Katz, N. Korsakova, J. R. Gair, and N. Stergioulas, "Eryn: a multipurpose sampler for Bayesian inference", Mon. Not. Roy. Astron. Soc. **526**, 4814–4830 (2023).

[239]J. Goodman and J. Weare, "Ensemble samplers with affine invariance", Communications in Applied Mathematics and Computational Science **5**, 65–80 (2010).

[240]P. Gregory, *Bayesian logical data analysis for the physical sciences*, 1st. (Cambridge University Press, Cambridge, 2005), pp. 313–346.

[241]M. J. Betancourt and M. Girolami, "Hamiltonian Monte Carlo for Hierarchical Models", arXiv e-prints, arXiv:1312.0906, 10.48550/arXiv.1312.0906 (2013).

[242]U. Grenander and M. I. Miller, "Representations of knowledge in complex systems", Journal of the Royal Statistical Society. Series B (Methodological) **56**, 549–603 (1994).

[243]K. W. K. Wong, M. Gabrié, and D. Foreman-Mackey, "flowMC: Normalizing-flow enhanced sampling package for probabilistic inference in Jax", arXiv e-prints, arXiv:2211.06397, 10.48550/arXiv.2211.06397 (2022).

[244]N. Korsakova, S. Babak, M. L. Katz, N. Karnesis, S. Khukhlaev, and J. R. Gair, "Neural density estimation for Galactic Binaries in LISA data analysis", (2024).

[245]N. Lartillot and H. Philippe, "Computing Bayes Factors Using Thermodynamic Integration", Systematic Biology **55**, 195–207 (2006).

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

[247]J. Skilling, "Nested Sampling", in Bayesian inference and maximum entropy methods in science and engineering: 24th international workshop on bayesian inference and maximum entropy methods in science and engineering, Vol. 735, edited by R. Fischer, R. Preuss, and U. V. Toussaint, American Institute of Physics Conference Series (Nov. 2004), pp. 395–405.

[248]G. Ashton et al., "Nested sampling for physical scientists", Nature **2**, 10.1038/s43586-022-00121-x (2022).

[249]J. Buchner, "Nested Sampling Methods", Statistics Surveys **17**, 169–215 (2023).

[250]W. Del Pozzo and J. Veitch, *CPNest: Parallel nested sampling*, Astrophysics Source Code Library, record ascl:2205.021, May 2022.

[251]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).

[252]J. S. Speagle, "DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences", Mon. Not. Roy. Astron. Soc. **493**, 3132–3158 (2020).

[253]F. Feroz, M. P. Hobson, and M. Bridges, "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics", Mon. Not. Roy. Astron. Soc. **398**, 1601–1614 (2009).

[254]M. J. Williams, J. Veitch, and C. Messenger, "Nested sampling with normalizing flows for gravitational-wave inference", Phys. Rev. D **103**, 103006 (2021).

[255]M. J. Williams, J. Veitch, and C. Messenger, "Importance nested sampling with normalising flows", Mach. Learn. Sci. Tech. **4**, 035011 (2023).

[256]P. Whittle, "The analysis of multiple stationary time series", Journal of the Royal Statistical Society. Series B (Methodological) **15**, 125–139 (1953).

[257]L. S. Finn, "Detection, measurement and gravitational radiation", Phys. Rev. D **46**, 5236–5249 (1992).

[258]P. Whittle, *Hypothesis testing in time series analysis*, 1st. (Almquist & Wiksell, Uppsala, 1951).

[259]K. Lackeos, T. B. Littenberg, N. J. Cornish, and J. I. Thorpe, "The LISA Data Challenge Radler analysis and time-dependent ultra-compact binary catalogues", Astron. Astrophys. **678**, A123 (2023).

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

[261]G. Ashton et al., "BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy", Astrophys. J. Suppl. **241**, 27 (2019).

[262]K. Cannon et al., "GstLAL: A software framework for gravitational wave discovery", (2020).

[263]S. A. Usman et al., "The PyCBC search for gravitational waves from compact binary coalescence", Class. Quant. Grav. **33**, 215004 (2016).

[264]I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, `http://www.deeplearningbook.org` (MIT Press, Cambridge, MA, USA, 2016).

[265]K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators", Neural Networks **2**, 359–366 (1989).

[266]S. Liang and R. Srikant, "Why Deep Neural Networks for Function Approximation?", arXiv e-prints, arXiv:1610.04161, `10.48550/arXiv.1610.04161` (2016).

[267]H. Gabbard, M. Williams, F. Hayes, and C. Messenger, "Matching matched filtering with deep networks for gravitational-wave astronomy", Phys. Rev. Lett. **120**, 141103 (2018).

[268]M. B. Schäfer et al., "First machine learning gravitational-wave search mock data challenge", Phys. Rev. D **107**, 023021 (2023).

[269]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 Phys. **18**, 112–117 (2022).

[270]A. J. K. Chua and M. Vallisneri, "Learning Bayesian posteriors with neural networks for gravitational-wave inference", Phys. Rev. Lett. **124**, 041102 (2020).

[271]M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, "Real-Time Gravitational Wave Science with Neural Posterior Estimation", Phys. Rev. Lett. **127**, 241103 (2021).

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

[273]E. Cuoco et al., "Enhancing Gravitational-Wave Science with Machine Learning", Mach. Learn. Sci. Tech. **2**, 011002 (2021).

[274]F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain", Psychological Review **65**, 386–408 (1958).

[275]S. Amari, "A theory of adaptive pattern classifiers", IEEE Transactions on Electronic Computers **EC-16**, 299–307 (1967).

[276]P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for Activation Functions", (2017).

[277]S. Narayan, "The generalized sigmoid activation function: competitive supervised learning", Information Sciences **99**, 69–82 (1997).

[278]A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)", arXiv e-prints, arXiv:1803.08375 10.48550/arXiv.1803.08375 (2018).

[279]V. Strassen, "Gaussian elimination is not optimal", Numer. Math, **13**, 354–356 (1969).

[280]K. Fatahalian, J. Sugerman, and P. Hanrahan, "Understanding the efficiency of gpu algorithms for matrix-matrix multiplication", in Proceedings of the acm siggraph/eurographics conference on graphics hardware, HWWS '04 (2004), pp. 133–137.

[281]A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", in Advances in neural information processing systems 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (2019), pp. 8024–8035.

[282]O. Zachariadis, N. Satpute, J. Gómez-Luna, and J. Olivares, "Accelerating Sparse Matrix-Matrix Multiplication with GPU Tensor Cores", arXiv e-prints, arXiv:2009.14600, 10.48550/arXiv.2009.14600 (2020).

[283]S. Ruder, "An overview of gradient descent optimization algorithms", (2016).

[284]D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", nature **323**, 533–536 (1986).

[285]D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", arXiv e-prints, arXiv:1412.6980 (2014).

[286]X. Ying, "An Overview of Overfitting and its Solutions", in Journal of physics conference series, Vol. 1168, Journal of Physics Conference Series (Feb. 2019), p. 022022.

[287]Y. Bai, E. Yang, B. Han, Y. Yang, J. Li, Y. Mao, G. Niu, and T. Liu, "Understanding and Improving Early Stopping for Learning with Noisy Labels", arXiv e-prints, arXiv:2106.15853, 10.48550/arXiv.2106.15853 (2021).

[288]C. Chapman-Bird, *Poplar*, version v0.1.0, Jan. 2023.

[289]B. J. Owen, "Search templates for gravitational waves from inspiraling binaries: Choice of template spacing", Phys. Rev. D **53**, 6749–6761 (1996).

[290]A. Antonelli, O. Burke, and J. R. Gair, "Noisy neighbours: inference biases from overlapping gravitational-wave signals", Mon. Not. Roy. Astron. Soc. **507**, 5069–5086 (2021).

[291]N. J. Cornish and T. B. Littenberg, "Tests of Bayesian Model Selection Techniques for Gravitational Wave Astronomy", Phys. Rev. D **76**, 083006 (2007).

[292]T. B. Littenberg, "A detection pipeline for galactic binaries in LISA data", Phys. Rev. D **84**, 063009 (2011).

[293]C. R. Weaving, L. K. Nuttall, I. W. Harry, S. Wu, and A. Nitz, "Adapting the PyCBC pipeline to find and infer the properties of gravitational waves from massive black hole binaries in LISA", Class. Quant. Grav. **41**, 025006 (2024).

[294]A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities", Journal of the American Statistical Association **85**, 398–409 (1990).

[295]T. B. Littenberg and N. J. Cornish, "Prototype global analysis of LISA data with multiple source types", Phys. Rev. D **107**, 063004 (2023).

[296]M. L. Katz, N. Karnesis, N. Korsakova, J. R. Gair, and N. Stergioulas, "An efficient GPU-accelerated multi-source global fit pipeline for LISA data analysis", (2024).

[297]P. J. GREEN, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", Biometrika **82**, 711–732 (1995).

[298]S. H. Strub, L. Ferraioli, C. Schmelzbach, S. C. Stähler, and D. Giardini, "Global Analysis of LISA Data with Galactic Binaries and Massive Black Hole Binaries", (2024).

[299]A. Spadaro, R. Buscicchio, D. Vetrugno, A. Klein, D. Gerosa, S. Vitale, R. Dolesi, W. J. Weber, and M. Colpi, "Glitch systematics on the observation of massive black-hole binaries with LISA", Phys. Rev. D **108**, 123029 (2023).

[300]Q. Baghi (LDC Working Group), "The LISA Data Challenges", in 56th Rencontres de Moriond on Gravitation (Apr. 2022).

[301]M. C. Edwards, P. Maturana-Russel, R. Meyer, J. Gair, N. Korsakova, and N. Christensen, "Identifying and Addressing Nonstationary LISA Noise", Phys. Rev. D **102**, 084062 (2020).

[302]N. J. Cornish, "Time-Frequency Analysis of Gravitational Wave Data", 10.1103/PhysRevD.102.124038 (2020).

[303]M. C. Digman and N. J. Cornish, "Parameter estimation for stellar-origin black hole mergers in LISA", Phys. Rev. D **108**, 023022 (2023).

[304]M. C. Digman and N. J. Cornish, "LISA Gravitational Wave Sources in a Time-varying Galactic Stochastic Background", Astrophys. J. **940**, 10 (2022).

[305]A. Sasli, N. Karnesis, and N. Stergioulas, "Heavy-tailed likelihoods for robustness against data outliers: Applications to the analysis of gravitational wave data", Phys. Rev. D **108**, 103005 (2023).

[306]M. Vallisneri, "Synthetic LISA: Simulating time delay interferometry in a model LISA", Phys. Rev. D **71**, 022001 (2005).

[307]A. Petiteau, G. Auger, H. Halloin, O. Jeannin, E. Plagnol, S. Pireaux, T. Regimbau, and J.-Y. Vinet, "LISACode: A Scientific simulator of LISA", Phys. Rev. D **77**, 023002 (2008).

[308]J.-B. Bayle and O. Hartwig, "Unified model for the LISA measurements and instrument simulations", Phys. Rev. D **107**, 083019 (2023).

[309]J.-B. Bayle, O. Hartwig, M. Lilley, A. Hees, C. Chapman-Bird, G. Woan, and P. Wolf, "End-to-end simulation and analysis pipeline for LISA", in 57th Rencontres de Moriond on Gravitation (May 2023).

[310]S. Marsat and J. G. Baker, "Fourier-domain modulations and delays of gravitational-wave signals", (2018).

[311]N. Afshordi et al. (LISA Consortium Waveform Working Group), "Waveform Modelling for the Laser Interferometer Space Antenna", (2023).

[312]L. Speri, M. L. Katz, A. J. K. Chua, S. A. Hughes, N. Warburton, J. E. Thompson, C. E. A. Chapman-Bird, and J. R. Gair, "Fast and Fourier: Extreme Mass Ratio Inspiral Waveforms in the Frequency Domain", (2023).

[313]A. J. K. Chua, C. J. Moore, and J. R. Gair, "Augmented kludge waveforms for detecting extreme-mass-ratio inspirals", Phys. Rev. D **96**, 044005 (2017).

[314]J. R. Gair, E. K. Porter, S. Babak, and L. Barack, "A Constrained Metropolis-Hastings Search for EMRIs in the Mock LISA Data Challenge 1B", Class. Quant. Grav. **25**, edited by S. Hughes and E. Katsavounidis, 184030 (2008).

[315]I. Mandel, W. M. Farr, and J. R. Gair, "Extracting distribution parameters from multiple uncertain observations with selection biases", Mon. Not. Roy. Astron. Soc. **486**, 1086–1093 (2019).

[316]N. J. Cornish and L. J. Rubbo, "The LISA response function", Phys. Rev. D **67**, 022001 (2003).

[317]W. Martens and E. Joffre, "Trajectory Design for the ESA LISA Mission", 10.1007/s40295-021-00263-2 (2021).

[318]J.-B. Bayle, A. Hees, M. Lilley, C. Le Poncin-Lafitte, W. Martens, and E. Joffre, *Lisa orbits*, version 2.3, Mar. 2023.

[319]J.-B. Bayle, Q. Baghi, A. Renzini, and M. Le Jeune, *Lisa gw response*, version 2.3, Sept. 2023.

[320]J.-B. Bayle, O. Hartwig, and M. Staab, *Lisa instrument*, version 1.5, Sept. 2023.

[321]S. Babak, A. Petiteau, and M. Hewitson, "LISA Sensitivity and SNR Calculations", (2021).

[322]M. Tinto and O. Hartwig, "Time-Delay Interferometry and Clock-Noise Calibration", Phys. Rev. D **98**, 042003 (2018).

[323]M. Staab, J.-B. Bayle, and O. Hartwig, *Pytdi*, version 1.3.1, Oct. 2023.

[324]T. A. Prince, M. Tinto, S. L. Larson, and J. W. Armstrong, "The LISA optimal sensitivity", Phys. Rev. D **66**, 122002 (2002).

[325]D. Q. Nam, Y. Lemiere, A. Petiteau, J.-B. Bayle, O. Hartwig, J. Martino, and M. Staab, "TDI noises transfer functions for LISA", (2022).

[326]M. J. Williams, *Nessai: nested sampling with artificial intelligence*, version latest, Feb. 2021.

[327]L. Blanchet, "Gravitational radiation from post-Newtonian sources and inspiralling compact binaries", Living Rev. Rel. **9**, 4 (2006).

[328]T. Littenberg, N. Cornish, K. Lackeos, and T. Robson, "Global Analysis of the Gravitational Wave Signal from Galactic Binaries", Phys. Rev. D **101**, 123021 (2020).

[329]L. Sberna, A. Toubiana, and M. C. Miller, "Golden galactic binaries for LISA: mass-transferring white dwarf black hole binaries", Astrophys. J. **908**, 1 (2021).

[330]M. L. Katz, J.-B. Bayle, A. J. K. Chua, and M. Vallisneri, "Assessing the data-analysis impact of LISA orbit approximations using a GPU-accelerated response model", Phys. Rev. D **106**, 103001 (2022).

[331]V. Korol, E. M. Rossi, P. J. Groot, G. Nelemans, S. Toonen, and A. G. A. Brown, "Prospects for detection of detached double white dwarf binaries with Gaia, LSST and LISA", Mon. Not. Roy. Astron. Soc. **470**, 1894–1910 (2017).

[332]S. C. C. Barros, T. R. Marsh, V. S. Dhillon, P. J. Groot, S. Littlefair, G. Nelemans, G. Roelofs, D. Steeghs, and P. J. Wheatley, "ULTRACAM Photometry of the ultracompact binaries V407 Vul and HM Cnc", Mon. Not. Roy. Astron. Soc. **374**, 1334–1346 (2007).

[333]S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, "Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era", Phys. Rev. D **93**, 044007 (2016).

[334]S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, "Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal", Phys. Rev. D **93**, 044006 (2016).

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

[336]T. Dal Canton et al., "Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors", Phys. Rev. D **90**, 082004 (2014).

[337]S. Cook, A. Gelman, and D. Rubin, "Validation of software for bayesian models using posterior quantiles", Journal of Computational and Graphical Statistics **15**, 675 (2006).

[338]G. Ashton and C. Talbot, "B ilby-MCMC: an MCMC sampler for gravitational-wave inference", Mon. Not. Roy. Astron. Soc. **507**, 2037–2051 (2021).

[339]J. Beirlant and P. Deheuvels, "On the approximation of p—p and q—q plot processes by brownian bridges", Statistics and Probability Letters **9**, 241–251 (1990).

[340]F. Mosteller and R. A. Fisher, "Questions and answers", The American Statistician **2**, 30–31 (1948).

[341]L. Barack and C. Cutler, "LISA capture sources: Approximate waveforms, signal-to-noise ratios, and parameter estimation accuracy", Phys. Rev. D **69**, 082005 (2004).

[342]P. Amaro-Seoane, J. R. Gair, M. Freitag, M. Coleman Miller, I. Mandel, C. J. Cutler, and S. Babak, "Astrophysics, detection and science applications of intermediate- and extreme mass-ratio inspirals", Class. Quant. Grav. **24**, R113–R169 (2007).

[343]S. Drasco and S. A. Hughes, "Gravitational wave snapshots of generic extreme mass ratio inspirals", Phys. Rev. D **73**, 024027 (2006).

[344]L. E. Kidder, "Coalescing binary systems of compact objects to postNewtonian 5/2 order. 5. Spin effects", Phys. Rev. D **52**, 821–847 (1995).

[345]S. A. Teukolsky and W. H. Press, "Perturbations of a rotating black hole. III - Interaction of the hole with gravitational and electromagnetic radiation", Astrophys. J. **193**, 443–461 (1974).

[346]*Black Hole Perturbation Toolkit*, (bhptoolkit.org).

[347]C. P. L. Berry, R. H. Cole, P. Cañizares, and J. R. Gair, "Importance of transient resonances in extreme-mass-ratio inspirals", Phys. Rev. D **94**, 124042 (2016).

[348]L. Speri and J. R. Gair, "Assessing the impact of transient orbital resonances", Phys. Rev. D **103**, 124032 (2021).

[349]C. Cutler and M. Vallisneri, "LISA detections of massive black hole inspirals: Parameter extraction errors due to inaccurate template waveforms", Phys. Rev. D **76**, 104018 (2007).

[350]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", Phys. Rev. Lett. **96**, 111102 (2006).

[351]A. Buonanno and T. Damour, "Effective one-body approach to general relativistic two-body dynamics", Phys. Rev. D **59**, 084006 (1999).

[352]L. Barack and A. Pound, "Self-force and radiation reaction in general relativity", Rept. Prog. Phys. **82**, 016904 (2019).

[353]O. Burke, G. A. Piovano, N. Warburton, P. Lynch, L. Speri, C. Kavanagh, B. Wardell, A. Pound, L. Durkan, and J. Miller, "Accuracy Requirements: Assessing the Importance of First Post-Adiabatic Terms for Small-Mass-Ratio Binaries", (2023).

[354]A. Albertini, A. Nagar, A. Pound, N. Warburton, B. Wardell, L. Durkan, and J. Miller, "Comparing second-order gravitational self-force, numerical relativity, and effective one body waveforms from inspiralling, quasicircular, and nonspinning black hole binaries", Phys. Rev. D **106**, 084061 (2022).

[355]E. Schnetter, S. H. Hawley, and I. Hawke, "Evolutions in 3-D numerical relativity using fixed mesh refinement", Class. Quant. Grav. **21**, 1465–1488 (2004).

[356]C. O. Lousto and Y. Zlochower, "Orbital Evolution of Extreme-Mass-Ratio Black-Hole Binaries with Numerical Relativity", Phys. Rev. Lett. **106**, 041101 (2011).

[357]L. Pompili et al., "Laying the foundation of the effective-one-body waveform models SEOBNRv5: Improved accuracy and efficiency for spinning nonprecessing binary black holes", Phys. Rev. D **108**, 124035 (2023).

[358]A. Buonanno, Y. Pan, H. P. Pfeiffer, M. A. Scheel, L. T. Buchman, and L. E. Kidder, "Effective-one-body waveforms calibrated to numerical relativity simulations: Coalescence of nonspinning, equal-mass black holes", Phys. Rev. D **79**, 124028 (2009).

[359]M. van de Meent, A. Buonanno, D. P. Mihaylov, S. Ossokine, L. Pompili, N. Warburton, A. Pound, B. Wardell, L. Durkan, and J. Miller, "Enhancing the SEOBNRv5 effective-one-body waveform model with second-order gravitational self-force fluxes", Phys. Rev. D **108**, 124038 (2023).

[360]R. Fujita and M. Shibata, "Extreme mass ratio inspirals on the equatorial plane in the adiabatic order", Phys. Rev. D **102**, 064005 (2020).

[361]C. M. Will, "On the unreasonable effectiveness of the post-Newtonian approximation in gravitational physics", Proc. Nat. Acad. Sci. **108**, 5938 (2011).

[362]J. Brink, M. Geyer, and T. Hinderer, "Astrophysics of resonant orbits in the Kerr metric", Phys. Rev. D **91**, 083001 (2015).

[363]H. Goldstein and C. Poole, *Classical mechanics /*, 3rd. (Pearson Education, New Delhi, 2002), pp. 102–106.

[364]B. Carter, "Global structure of the Kerr family of gravitational fields", Phys. Rev. **174**, 1559–1571 (1968).

[365]Y. Mino, "Perturbative approach to an orbital evolution around a supermassive black hole", Phys. Rev. D **67**, 084027 (2003).

[366]R. Fujita and W. Hikida, "Analytical solutions of bound timelike geodesic orbits in Kerr spacetime", Class. Quant. Grav. **26**, 135002 (2009).

[367]M. Kerachian, L. Polcar, V. Skoupý, C. Efthymiopoulos, and G. Lukes-Gerakopoulos, "Action-angle formalism for extreme mass ratio inspirals in Kerr spacetime", Phys. Rev. D **108**, 044004 (2023).

[368]T. Hinderer and E. E. Flanagan, "Two timescale analysis of extreme mass ratio inspirals in Kerr. I. Orbital Motion", Phys. Rev. D **78**, 064028 (2008).

[369]R. H. Boyer and R. W. Lindquist, "Maximal analytic extension of the Kerr metric", J. Math. Phys. **8**, 265 (1967).

[370]M. van de Meent, "Analytic solutions for parallel transport along generic bound geodesics in Kerr spacetime", Class. Quant. Grav. **37**, 145007 (2020).

[371]W. Schmidt, "Celestial mechanics in Kerr space-time", Class. Quant. Grav. **19**, 2743 (2002).

[372]N. Warburton, A. Pound, B. Wardell, J. Miller, and L. Durkan, "Gravitational-Wave Energy Flux for Compact Binaries through Second Order in the Mass Ratio", Phys. Rev. Lett. **127**, 151102 (2021).

[373]L. C. Stein and N. Warburton, "Location of the last stable orbit in Kerr spacetime", Phys. Rev. D **101**, 064007 (2020).

[374]T. Osburn, N. Warburton, and C. R. Evans, "Highly eccentric inspirals into a black hole", Phys. Rev. D **93**, 064024 (2016).

[375]N. Warburton, S. Akcay, L. Barack, J. R. Gair, and N. Sago, "Evolution of inspiral orbits around a Schwarzschild black hole", Phys. Rev. D **85**, 061501 (2012).

[376]S. A. Hughes, S. Drasco, E. E. Flanagan, and J. Franklin, "Gravitational radiation reaction and inspiral waveforms in the adiabatic limit", Phys. Rev. Lett. **94**, 221101 (2005).

[377]J. Miller and A. Pound, "Two-timescale evolution of extreme-mass-ratio inspirals: waveform generation scheme for quasicircular orbits in Schwarzschild spacetime", Phys. Rev. D **103**, 064048 (2021).

[378]N. Sago, T. Tanaka, W. Hikida, K. Ganz, and H. Nakano, "The Adiabatic evolution of orbital parameters in the Kerr spacetime", Prog. Theor. Phys. **115**, 873–907 (2006).

[379] A. Pound and E. Poisson, "Osculating orbits in Schwarzschild spacetime, with an application to extreme mass-ratio inspirals", Phys. Rev. D **77**, 044013 (2008).

[380] M. Van De Meent and N. Warburton, "Fast Self-forced Inspirals", Class. Quant. Grav. **35**, 144003 (2018).

[381] L. V. Drummond, P. Lynch, A. G. Hanselman, D. R. Becker, and S. A. Hughes, "Extreme mass-ratio inspiral and waveforms for a spinning body into a Kerr black hole via osculating geodesics and near-identity transformations", (2023).

[382] R. W. O'Shaughnessy, "Transition from inspiral to plunge for eccentric equatorial Kerr orbits", Phys. Rev. D **67**, 044004 (2003).

[383] P. Blanchard, "Differential equations", in (Brooks/Cole, 2012), p. 469.

[384] P. A. Sundararajan, "The Transition from adiabatic inspiral to geodesic plunge for a compact object around a massive Kerr black hole: Generic orbits", Phys. Rev. D **77**, 124050 (2008).

[385] A. Pound, G. Compere, L. Durkan, L. Kuchler, N. Warburton, and B. Wardell, in prep.

[386] L. Küchler, G. Compère, L. Durkan, and A. Pound, "Self-force framework for transition-to-plunge waveforms", (2024).

[387] G. A. Piovano, A. Maselli, and P. Pani, "Extreme mass ratio inspirals with spinning secondary: a detailed study of equatorial circular motion", Phys. Rev. D **102**, 024041 (2020).

[388] J. Mathews, A. Pound, and B. Wardell, "Self-force calculations with a spinning secondary", Phys. Rev. D **105**, 084031 (2022).

[389] M. Maggiore, "Gravitational waves volume 2: astrophysics and cosmology", in (Oxford, 2018), pp. 174–176.

[390] R. Fujita and H. Tagoshi, "New Numerical Methods to Evaluate Homogeneous Solutions of the Teukolsky Equation", eConf **C041213**, edited by P. Chen, E. D. Bloom, G. Madejski, and V. Petrosian, 1211 (2004).

[391] R. Fujita and H. Tagoshi, "New Numerical Methods to Evaluate Homogeneous Solutions of the Teukolsky Equation II. Solutions of the Continued Fraction Equation", Prog. Theor. Phys. **113**, 1165–1182 (2005).

[392] Z. Nasipak, "BHPWAVE: An adiabatic gravitational waveform model for compact objects undergoing quasi-circular inspirals into rotating massive black holes", (2023).

[393] S. A. Hughes, "The Evolution of circular, nonequatorial orbits of Kerr black holes due to gravitational wave emission", Phys. Rev. D **61**, 084004 (2000).

[394] K. Glampedakis and D. Kennefick, "Zoom and whirl: Eccentric equatorial orbits around spinning black holes and their evolution under gravitational radiation reaction", Phys. Rev. D **66**, 044002 (2002).

[395]N. Yunes, A. Buonanno, S. A. Hughes, Y. Pan, E. Barausse, M. C. Miller, and W. Throwe, "Extreme Mass-Ratio Inspirals in the Effective-One-Body Approach: Quasi-Circular, Equatorial Orbits around a Spinning Black Hole", Phys. Rev. D **83**, 044044 (2011).

[396]C. Moler, "Matrix computation on distributed memory multiprocessors", Hypercube Multiprocessors **86**, 31 (1986).

[397]C. Cutler, D. Kennefick, and E. Poisson, "Gravitational radiation reaction for bound motion around a Schwarzschild black hole", Phys. Rev. D **50**, 3816–3835 (1994).

[398]E. Barausse, "The evolution of massive black holes and their spins in their galactic hosts", Mon. Not. Roy. Astron. Soc. **423**, 2533–2557 (2012).

[399]P. Lynch, M. van de Meent, and N. Warburton, "Self-forced inspirals with spin-orbit precession", (2023).

[400]J. R. Gair and K. Glampedakis, "Improved approximate inspirals of test-bodies into Kerr black holes", Phys. Rev. D **73**, 064037 (2006).

[401]S. Babak, H. Fang, J. R. Gair, K. Glampedakis, and S. A. Hughes, "'Kludge' gravitational waveforms for a test-body orbiting a Kerr black hole", Phys. Rev. D **75**, 024005 (2007).

[402]S. Droz, D. J. Knapp, E. Poisson, and B. J. Owen, "Gravitational waves from inspiraling compact binaries: Validity of the stationary phase approximation to the Fourier transform", Phys. Rev. D **59**, 124016 (1999).

[403]M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification", Pattern Recognition **37**, 1757–1771 (2004).

[404]I. J. Good, "Rational decisions", Journal of the Royal Statistical Society. Series B (Methodological) **14**, 107–114 (1952).

[405]P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed Precision Training", arXiv e-prints, arXiv:1710.03740, 10.48550/arXiv.1710.03740 (2017).

[406]K. Xu, M. Zhang, J. Li, S. S. Du, K.-i. Kawarabayashi, and S. Jegelka, "How neural networks extrapolate: from feedforward to graph neural networks", (2021).

[407]T. Fawcett, "Introduction to roc analysis", Pattern Recognition Letters **27**, 861–874 (2006).

[408]P. Prince and J. Dormand, "High order embedded runge-kutta formulae", Journal of Computational and Applied Mathematics **7**, 67–75 (1981).

[409]G. G. Dahlquist, "A special stability problem for linear multistep methods", BIT Numerical Mathematics **3**, 27–43 (1963).

[410]A. J. K. Chua and C. J. Cutler, "Nonlocal parameter degeneracy in the intrinsic space of gravitational-wave signals from extreme-mass-ratio inspirals", Phys. Rev. D **106**, 124046 (2022).

[411]E. Payne, C. Talbot, and E. Thrane, "Higher order gravitational-wave modes with likelihood reweighting", Phys. Rev. D **100**, 123017 (2019).

[412]S. Khan, F. Ohme, K. Chatziioannou, and M. Hannam, "Including higher order multipoles in gravitational-wave models for precessing binary black holes", Phys. Rev. D **101**, 024056 (2020).

[413]R. A. Daly, "Estimates of Black Hole Spin Properties of 55 Sources", Mon. Not. Roy. Astron. Soc. **414**, 1253 (2011).

[414]V. Vázquez-Aceves, L. Zwick, E. Bortolas, P. R. Capelo, P. Amaro-Seoane, L. Mayer, and X. Chen, "Revised event rates for extreme and extremely large mass-ratio inspirals", Mon. Not. Roy. Astron. Soc. **510**, 2379–2390 (2022).

[415]J. R. Gair, C. Tang, and M. Volonteri, "LISA extreme-mass-ratio inspiral events as probes of the black hole mass function", Phys. Rev. D **81**, 104014 (2010).

[416]M. Zevin, S. S. Bavera, C. P. L. Berry, V. Kalogera, T. Fragos, P. Marchant, C. L. Rodriguez, F. Antonini, D. E. Holz, and C. Pankow, "One Channel to Rule Them All? Constraining the Origins of Binary Black Holes Using Multiple Formation Pathways", Astrophys. J. **910**, 152 (2021).

[417]J. Alsing, H. Peiris, D. Mortlock, J. Leja, and B. Leistedt, "Forward Modeling of Galaxy Populations for Cosmological Redshift Distribution Inference", Astrophys. J. Suppl. **264**, 29 (2023).

[418]J. R. Gair, L. Barack, T. Creighton, C. Cutler, S. L. Larson, E. S. Phinney, and M. Vallisneri, "Event rate estimates for LISA extreme mass ratio capture sources", Class. Quant. Grav. **21**, S1595–S1606 (2004).

[419]N. Christensen and R. Meyer, "Parameter estimation with gravitational waves", Rev. Mod. Phys. **94**, 025001 (2022).

[420]F. Shankar, "Black Hole Demography: From scaling relations to models", Class. Quant. Grav. **30**, 244001 (2013).

[421]M. Volonteri, "Formation of Supermassive Black Holes", Astron. Astrophys. Rev. **18**, 279–315 (2010).

[422]J. Roulet and M. Zaldarriaga, "Constraints on binary black hole populations from LIGO–Virgo detections", Mon. Not. Roy. Astron. Soc. **484**, 4216–4229 (2019).

[423]P. Amaro-Seoane, "The gravitational capture of compact objects by massive black holes", 10.1007/978-981-15-4702-7_17-1 (2020).

[424]R. L. Burden and J. D. Faires, *Numerical analysis* (Brooks Cole, Boston, 1993).

[425]D. W. Hogg, "Distance measures in cosmology", (1999).

[426]M. Abramowitz and I. A. Stegun, "Handbook of mathematical functions with formulas, graphs, and mathematical tables", in (Dover, New York, 1964), p. 375.

[427]E. A. Huerta and J. R. Gair, "Importance of including small body spin effects in the modelling of extreme and intermediate mass-ratio inspirals", Phys. Rev. D **84**, 064023 (2011).

[428]P. Amaro-Seoane et al. (LISA), "Laser Interferometer Space Antenna", (2017).

[429]A. Krolak and B. F. Schutz, "Coalescing binaries — Probe of the universe", Gen. Rel. Grav. **19**, 1163–1171 (1987).

[430]M. Vallisneri, "Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects", Phys. Rev. D **77**, 042001 (2008).

[431]T. Sauer, *Numerical analysis* (Pearson, Michigan, 2012).

[432]B. Fornberg, "Generation of finite difference formulas on arbitrarily spaced grids", Mathematics of Computation **51**, 699–706 (1988).

[433]A. Ben-Israel and T. N. E. Greville, *Generalized inverses: theory and applications* (Springer, 2003).

[434]F. Johansson et al., *Mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 0.18)*, `http://mpmath.org/` (Dec. 2013).

[435]S. Stevenson, C. P. L. Berry, and I. Mandel, "Hierarchical analysis of gravitational-wave measurements of binary black hole spin–orbit misalignments", Mon. Not. Roy. Astron. Soc. **471**, 2801–2811 (2017).

[436]L. Piegl and W. Tiller, "Curve and surface constructions using rational B-splines", Computer-Aided Design **19**, 485–498 (1987).

[437]C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, *Qhull: Quickhull algorithm for computing the convex hull*, Astrophysics Source Code Library, record ascl:1304.016, Apr. 2013.

[438]T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer Series in Statistics (Springer New York Inc., New York, NY, USA, 2001).

[439]M. Feurer and F. Hutter, "Hyperparameter Optimization", in *Automated machine learning: methods, systems, challenges* (Springer International Publishing, Cham, 2019), pp. 3–33.

[440]O. Ibe, *Markov processes for stochastic modeling* (Elsevier, 2013).

[441]Y. Dodge, *The concise encyclopedia of statistics* (Springer New York, New York, NY, 2008), pp. 283–287.

[442]A. J. K. Chua, S. Hee, W. J. Handley, E. Higson, C. J. Moore, J. R. Gair, M. P. Hobson, and A. N. Lasenby, "Towards a framework for testing general relativity with extreme-mass-ratio-inspiral observations", Mon. Not. Roy. Astron. Soc. **478**, 28–40 (2018).

[443]E. Hairer, G. Wanner, and S. P. Norsett, "Solving ordinary differential equations i", in (Springer, Berlin, 1993), pp. 188–194.

[444]S. Kullback and R. A. Leibler, "On information and sufficiency", The Annals of Mathematical Statistics **22**, 79–86 (1951).

[445]C. L. Rodriguez, B. Farr, W. M. Farr, and I. Mandel, "Inadequacies of the Fisher Information Matrix in gravitational-wave parameter estimation", Phys. Rev. D **88**, 084013 (2013).

[446]I. Mandel, C. P. L. Berry, F. Ohme, S. Fairhurst, and W. M. Farr, "Parameter estimation on compact binary coalescences with abruptly terminating gravitational waveforms", Class. Quant. Grav. **31**, 155005 (2014).