Understanding Gravitational Waves and Their Sources: Robust Inference, Tests of Gravity, and Future Prospects

Thesis by
Ethan Payne

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended June 30, 2025

Ethan Payne
ORCID: 0000-0003-4507-8373

DISCLAIMER

The work presented in this thesis stems from my participation in the LIGO Scientific Collaboration (LSC). This work does not reflect the scientific opinion of the LSC and it was not reviewed by the collaboration.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my adviser, Alan Weinstein, for the continued support over the last four years. His wisdom and knowledge of the field have been a joy to experience first-hand. I appreciate the freedom Alan gave me to explore this broad range of topics presented in this thesis. I would also like to thank Katerina Chatziioannou, who has been an exceptional mentor and inspiration for my research. My journey throughout the last four years would have been a very different experience without both Alan and Katerina's guidance.

Getting to this point has not been straightforward, and any such journey will be influenced by many incredible people. My interest in gravitational-wave astronomy started in 2016, after I was swept up in undergraduate research projects at Monash University under the direction of Paul Lasky and Eric Thrane. I owe an immense amount of gratitude to both of them for their guidance and continued support from the very start of my undergraduate studies until now.

Furthermore, I am incredibly grateful for Eric recommending I pursue an LSC Fellowship at the LIGO Hanford detector in 2019. The first time I felt truly a part of the LIGO-Virgo-KAGRA Collaboration was on my second day at the site. For anyone familiar with the operation of the LIGO detectors during an observing run, it was a maintenance Tuesday and brass plates used for environmental monitoring needed to be removed—a perfect job for the new, jet-lagged fellow on site. Lying on my back under the beam-tube in the corner station LVEA—armed with an Allen key—I could not help but be in disbelief of where I was and what I had become a part of. Over the next seven months at the Hanford detector, I had the pleasure of working with some of the most brilliant and insightful colleagues-turned-friends. In particular, I would like to thank Evan Goetz, Jeff Kissel, Lilli Sun, Niko Lecoeuche, Dripta Bhattacharjee, Sudarshan Karki, and Greg Mendell for their collaboration as I worked on the calibration of the LIGO Hanford detector. Additionally, I owe Rick Savage dearly for his mentorship while I was an LSC fellow.

After returning to Monash University to finish my Honours year—and a short stop-gap Research Assistant role at the Australian National University (thanks Lilli!)—I started my program at Caltech in 2021. During this time, I had the privilege of working with many incredible people within and external to the LIGO-Virgo-KAGRA Collaboration. I would like to thank my incredible collaborators in no particular

order: Max Isi, Will Farr, Luis Lehner, Lee McCuller, Colm Talbot, Kyle Kremer, Michael Zevin, Floor Broekgaarden, Sharan Banagiri, and Yanbei Chen. Being at Caltech also gave me access to many knowledgeable peers, post-docs, and research scientists. I owe many thanks to Jacob Golomb, Isaac Legred, Rhiannon Udall, Simona Miller, Sophie Hourihane, Rico Lo, Alvin Li, Brian Seymour, Andrew Laeuger, Xiang Li, James Gardner, Colin Weller, Su Direkci, Ian MacMillan, Daniel Grass, Ryan Magee, Derek Davis, Arianna Renzini, Lucy Thomas, Virginia D'Emilio, Sophie Bini, Jane Glanzer, Elenna Capote, Sander Vermeulen, Jonah Kanner, and Aidan Brooks for cultivating an inviting, engaging, and (most importantly) entertaining research environment.

On a personal level, it is never particularly easy to uproot and move across the largest ocean on Earth. First, I am indebted to the support of my parents. This body of work is a testament to the values of perseverance and scientific curiosity you both have instilled in me from a young age. I also want to thank my younger brother, who would make the occasional cameo from San Diego and South Dakota over the last four years. Though he cannot read, I owe much to my cat, Alan[1], who has been an integral part of my academic career since the start of my undergraduate studies. My time in Pasadena spent outside of the Institute would not have been nearly as enjoyable as it was without the large network of friends I have made here. From late-night concerts to fourteen hour hikes and thrilling rock climbing trips, I would have burnt out long ago without the many adventures to look forward to each weekend. In such a short time, I have made lifelong friends and I am excited to see what the future holds for all of us. Finally, my late Nana was instrumental in fostering my love of science. From as early as I remember, she would spend her weekends at one of Scienceworks, Melbourne Zoo, and the Melbourne Museum with me. Though it is impossible to attribute my trajectory to any one person, her influence on my academic path cannot be understated.

I thank Alan Weinstein, Katerina Chatziioannou, Lee McCuller, and Ryan Patterson for serving on my doctoral committee. I am grateful for the support from the National Science Foundation Grants PHY-2309200 and PHY-2207758. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. This research has made use of data, software, and/or web tools obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org), a service of LIGO Laboratory, the LIGO

[1]It needs to be clarified that Alan (the cat) was named so prior to my knowledge of Alan (the human)'s existence. The feline Alan was named after the father of modern computing, Alan Turing.

Scientific Collaboration and the Virgo Collaboration. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by Polish and Hungarian institutes. I am grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459.

ABSTRACT

As gravitational-wave detectors have become increasing more sensitive since the first detection in 2015, the now routine observations of gravitational waves have provided a lens through which the field of gravitational-wave astronomy has been able to study the universe. In this thesis, I explore a substantial number of facets regarding the inference challenges associated with observations from binary compact object mergers. I demonstrate the difficulties conducting and interpreting accurate spin measurements from real observations. In addition, I then present a framework for testing general relativity from an ensemble of events without underlying statistical assumptions. This framework is then extended to incorporate theoretically motivated information into these tests. These methods were utilized to analyze observational data from the LIGO-Virgo-KAGRA Collaboration's third observing period. Additionally, I present a novel summary statistic for diagnosing model misspecification in astrophysical compact binary coalescence population studies. Finally, I conclude with a demonstration of the utility of novel detector readout schemes for future gravitational-wave interferometer designs. My thesis presents a sweeping view of a number of current research avenues with current and future gravitational-wave detectors.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] E. Payne, M. Isi, K. Chatziioannou, L. Lehner, Y. Chen, and W. M. Farr. "Curvature Dependence of Gravitational-Wave Tests of General Relativity". In: *Phys. Rev. Lett.* 133.25 (2024), p. 251401. DOI: 10.1103/PhysRevLett.133.251401. arXiv: 2407.07043 [gr-qc].

E.P. helped conceive the idea, undertook all calculations presented, and led the writing of the manuscript.

[2] L. Passenger, E. Thrane, P. D. Lasky, E. Payne, S. Stevenson, and B. Farr. "Are all models wrong? Falsifying binary formation models in gravitational-wave astronomy using exceptional events". In: *Mon. Not. Roy. Astron. Soc.* 535.3 (2024), pp. 2837–2843. DOI: 10.1093/mnras/stae2521. arXiv: 2405.09739 [astro-ph.HE].

E.P. helped conceive the idea presented in the manuscript, and provided feeback on the implementation and manuscript.

[3] A. Gupta et al. (incl. E. Payne). "Possible causes of false general relativity violations in gravitational wave observations". In: *SciPost Phys. Comm. Rep.* (2025), p. 5. DOI: 10.21468/SciPostPhysCommRep.5. URL: https://scipost.org/10.21468/SciPostPhysCommRep.5.

E.P. wrote a section of the review article relating to possible pitfalls of hierarchical analyses in tests of General Relativity.

[4] H. Tong et al. (incl. E. Payne). "Transdimensional Inference for Gravitational-wave Astronomy with Bilby". In: *Astrophys. J. Suppl.* 276.2 (2025), p. 50. DOI: 10.3847/1538-4365/ad9deb. arXiv: 2404.04460 [gr-qc].

E.P. contributed to the initial implementation of methods presented within the manuscript, and provided feedback on the manuscript.

[5] E. Payne, K. Kremer, and M. Zevin. "Spin Doctors: How to Diagnose a Hierarchical Merger Origin". In: *Astrophys. J. Lett.* 966.1 (2024), p. L16. DOI: 10.3847/2041-8213/ad3e82. arXiv: 2402.15066 [gr-qc].

E.P. conceived the project, carried out all the research, and led the writing of the manuscript.

[6] R. Magee, M. Isi, E. Payne, K. Chatziioannou, W. M. Farr, G. Pratten, and S. Vitale. "Impact of selection biases on tests of general relativity with gravitational-wave inspirals". In: *Phys. Rev. D* 109.2 (2024), p. 023014. DOI: 10.1103/PhysRevD.109.023014. arXiv: 2311.03656 [gr-qc].

E.P. carried out the hierarchical analyses presented, and contributed to the writing of the manuscript.

[7] E. Payne, M. Isi, K. Chatziioannou, and W. M. Farr. "Fortifying gravitational-wave tests of general relativity against astrophysical assumptions". In: *Phys. Rev. D* 108.12 (2023), p. 124060. DOI: 10.1103/PhysRevD.108.124060. arXiv: 2309.04528 [gr-qc].

E.P. helped conceive the project, carried out all the analyses presented, and led the writing of the manuscript.

[8] F. S. Broekgaarden, S. Banagiri, and E. Payne. "Visualizing the Number of Existing and Future Gravitational-wave Detections from Merging Double Compact Objects". In: *Astrophys. J.* 969.2 (2024), p. 108. DOI: 10.3847/1538-4357/ad4709. arXiv: 2303.17628 [astro-ph.HE].

E.P. contributed to the discussion regarding the LIGO-Virgo-KAGRA's source classifications and catalog inclusion thresholds in relation to gravitational-wave population analyses and tests of General Relativity.

[9] E. Payne and E. Thrane. "Model exploration in gravitational-wave astronomy with the maximum population likelihood". In: *Phys. Rev. Res.* 5.2 (2023), p. 023013. DOI: 10.1103/PhysRevResearch.5.023013. arXiv: 2210.11641 [astro-ph.IM].

E.P. helped with the conception of the idea, developed and implemented all the of analysis methods, and led the writing of the manuscript.

[10] E. Payne, S. Hourihane, J. Golomb, R. Udall, D. Davis, and K. Chatziioannou. "Curious case of GW200129: Interplay between spin-precession inference and data-quality issues". In: *Phys. Rev. D* 106.10 (2022), p. 104017. DOI: 10.1103/PhysRevD.106.104017. arXiv: 2206.11932 [gr-qc].

E.P. carried out the majority of the parameter estimation analyses using Parallel Bilby for the results presented, and contributed significantly to the writing of the manuscript.

[11] M. Hannam et al. (incl. E. Payne). "General-relativistic precession in a black-hole binary". In: *Nature* 610.7933 (2022), pp. 652–655. DOI: 10.1038/s41586-022-05212-z. arXiv: 2112.11300 [gr-qc].

E.P. contributed to the inclusion of detector calibration uncertainties in LIGO-Virgo-KAGRA analyses used within this research.

[12] X. Li, L. Sun, R. K. L. Lo, E. Payne, and Y. Chen. "Angular emission patterns of remnant black holes". In: *Phys. Rev. D* 105.2 (2022), p. 024016. DOI: 10.1103/PhysRevD.105.024016. arXiv: 2110.03116 [gr-qc].

E.P. advised on the statistical methods used and provided feedback on the manuscript.

[13] R. Abbott et al. (incl. E. Payne). "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run". In: *Phys. Rev. X* 13.4 (2023), p. 041039. DOI: 10.1103/PhysRevX.13.041039. arXiv: 2111.03606 [gr-qc].

E.P. contributed to the inclusion of detector calibration uncertainties in the RIFT LIGO-Virgo-KAGRA analyses.

In addition to the above works, the following highlight contributions from E.P. to the field of gravitational-wave astronomy prior to becoming a graduate student at the California Institute of Technology:

[1] E. Payne, L. Sun, K. Kremer, P. D. Lasky, and E. Thrane. "The Imprint of Superradiance on Hierarchical Black Hole Mergers". In: *Astrophys. J.* 931.2 (2022), p. 79. DOI: 10.3847/1538-4357/ac66df. arXiv: 2107.11730 [gr-qc].

E.P. conceived of the project, developed and all simulations, and wrote the manuscript.

[2] L. Sun et al. (incl. E. Payne). "Characterization of systematic error in Advanced LIGO calibration in the second half of O3". In: (June 2021). arXiv: 2107.00129 [astro-ph.IM].

E.P. developed the photon calibrator data analysis and documentation codebase, contributed to the maintenance of the photon calibrators at LIGO Hanford, and contributed to the writing of the manuscript.

[3] D. Psaltis, C. Talbot, E. Payne, and I. Mandel. "Probing the Black Hole Metric. I. Black Hole Shadows and Binary Black-Hole Inspirals". In: *Phys. Rev. D* 103 (2021), p. 104036. DOI: 10.1103/PhysRevD.103.104036. arXiv: 2012.02117 [gr-qc].

E.P. assisted in the analysis of the gravitational-wave events and the writing of the manuscript.

[4] C. Kimball et al. (incl. E. Payne). "Evidence for Hierarchical Black Hole Mergers in the Second LIGO–Virgo Gravitational Wave Catalog". In: *Astrophys. J. Lett.* 915.2 (2021), p. L35. DOI: 10.3847/2041-8213/ac0aef. arXiv: 2011.05332 [astro-ph.HE].

E.P. contributed the zero-spin analysis of gravitational-wave observations and provided feedback on the manuscript.

[5] E. Payne, C. Talbot, P. D. Lasky, E. Thrane, and J. S. Kissel. "Gravitational-wave astronomy with a physical calibration model". In: *Phys. Rev. D* 102 (2020), p. 122004. DOI: 10.1103/PhysRevD.102.122004. arXiv: 2009.10193 [astro-ph.IM].

E.P. conceived the project idea, developed the analysis code and methodology, and led the writing of the manuscript.

[6] E. Payne, S. Banagiri, P. Lasky, and E. Thrane. "Searching for anisotropy in the distribution of binary black hole mergers". In: *Phys. Rev. D* 102.10 (2020), p. 102004. DOI: 10.1103/PhysRevD.102.102004. arXiv: 2006.11957 [astro-ph.CO].

E.P. led the development of the analysis and the writing of the manuscript.

[7] I. M. Romero-Shaw et al. (incl. E. Payne). "Bayesian inference for compact binary coalescences with bilby: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue". In: *Mon. Not. Roy. Astron. Soc.* 499.3 (2020), pp. 3295–3319. DOI: 10.1093/mnras/staa2850. arXiv: 2006.00714 [astro-ph.IM].

E.P. contributed to the code development of Bilby for this manuscript.

[8] D. Bhattacharjee, Y. Lecoeuche, S. Karki, J. Betzwieser, V. Bossilkov, S. Kandhasamy, E. Payne, and R. L. Savage. "Fiducial displacements with improved accuracy for the global network of gravitational wave detectors". In: *Class. Quant. Grav.* 38.1 (2021), p. 015009. DOI: 10.1088/1361-6382/aba9ed. arXiv: 2006.00130 [astro-ph.IM].

E.P. developed the photon calibrator data analysis and documentation software, as well as the maintenance of the photon calibrators at LIGO Hanford.

[9] L. Sun et al. (incl. E. Payne). "Characterization of systematic error in Advanced LIGO calibration". In: *Class. Quant. Grav.* 37.22 (2020), p. 225008. DOI: 10.1088/1361-6382/abb14e. arXiv: 2005.02531 [astro-ph.IM].

E.P. developed the photon calibrator data analysis and documentation codebase, contributed to the maintenance of the photon calibrators at LIGO Hanford, and contributed to the writing of the manuscript.

[10] E. Payne, C. Talbot, and E. Thrane. "Higher order gravitational-wave modes with likelihood reweighting". In: *Phys. Rev. D* 100.12 (2019), p. 123017. DOI: 10.1103/PhysRevD.100.123017. arXiv: 1905.05477 [astro-ph.IM].

E.P. developed the calculation methodology and led the writing of the manuscript.

[11] G. Ashton et al. (incl. E. Payne). "BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy". In: *Astrophys. J. Suppl.* 241.2 (2019), p. 27. DOI: 10.3847/1538-4365/ab06fc. arXiv: 1811.02042 [astro-ph.IM].

E.P. contributed to the code development of Bilby for this manuscript and contributed diagnostic analyses to the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Number *Page*

Chapter 1

INTRODUCTION

Over the past decade, gravitational-wave astronomy has revolutionized our understanding of the universe. In 2015, the Laser Interferometer Gravitational-Wave Observatory (LIGO) [1] in the United States made the first direct detection of gravitational waves, originating from the coalescence of two black holes [2]. Since this Nobel Prize-winning observation, the field has rapidly expanded. Each successive observing run and corresponding observational catalog [3, 4, 5] has yielded new insights into the formation and population of compact objects [6, 7, 8], the nature of matter at densities beyond nuclear saturation [9, 10, 11], and the fundamental properties of gravity itself [12, 13, 14, 15]. With the cumulative number of detections made by the LIGO-Virgo-KAGRA Collaboration (with addition of the Virgo detector [16]) now approaching three hundred signals as highlighted in Fig. 1.1 [17, 18][1], the primary challenge has shifted. No longer solely an engineering triumph, gravitational-wave astronomy now hinges on our ability to rigorously analyze numerous individual observations and to synthesize these into robust inferences about population-level properties of binary black hole, binary neutron star, and neutron star-black hole mergers [20].

Gravitational waves are a fundamental and emergent prediction of General Relativity [21]—a theory formulated nearly a century before their first direct detection. At its core, General Relativity expresses gravity through the language of differential geometry [22]. Einstein's field equations describe a relationship between spacetime curvature (a second-order derivative of the metric) and the stress-energy tensor. Solving these equations yields the metric, which governs the geodesic trajectories of objects moving under the influence of gravity. This framework predicts a range of phenomena, from the deflection of light [23] to geodetic [24] and Lense-Thirring precession [25], and, crucially for my thesis, gravitational waves [26]. A standard derivation begins with a linear perturbation to a static background spacetime. The perturbation ultimately satisfies a homogeneous wave equation, describing the free propagation of disturbances in the spacetime metric, i.e. gravitational waves [27].

[1] There is some degree of ambiguity to this number given the uncertain nature of observations near the detection threshold. See Ref. [19] for a discussion regarding these statements.

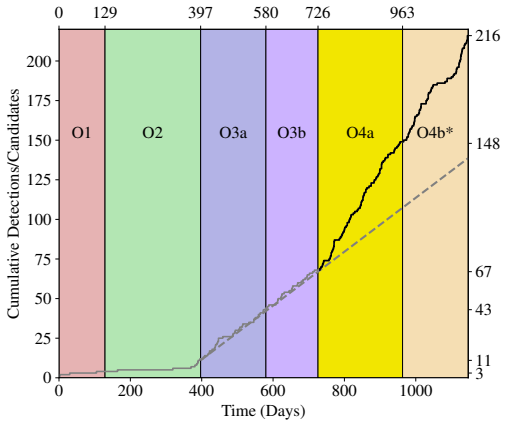

Figure 1.1: Fig. 3 from Ref. [17], summarizing the number of gravitational-wave observations over observing runs O1 through to the first 6 months of O4b. The grey line indicates the average O3 event detection rate projected into the duration of O4. The difference between the cumulative number (black) and the O3 rate projection demonstrates the instrumental improvement between observation periods.

The generation of gravitational waves is considerably more complex. Only non-spherically symmetric, accelerating stress-energy configurations can radiate gravitational waves [27]. In a multipole expansion [28], the leading-order contribution arises from the second time derivative of the mass quadrupole moment, though higher-order terms can, and do, contribute sub-dominantly in astrophysical contexts [29, 5]. The coefficient coupling stress-energy to curvature is necessarily small (G/c^4), and therefore it is not unexpected that detectable gravitational waves are sourced from rapidly moving, compact sources [27]. Such sources include mergers of compact binaries, rapidly rotating aspherical neutron stars, core-collapse supernovae, and other exotic astrophysical phenomena. In Fig. 1.2 from Ref. [2], the inspiral, merger, and ringdown phases of a compact binary coalescence are schematically shown, including the gravitational-wave strain associated with each stage, as well as the velocity and separation of the binary. From this depiction, it is clear that immense systems are required to produce even a quiet gravitational-wave signal.

Ground-based gravitational-wave detectors such as LIGO are designed to observe gravitational radiation from compact binary coalescences [30]. By the time this radiation reaches detectors on Earth, the dimensionless strain is minuscule, of order $O(10^{-21})$. Such an extreme sensitivity requirement poses a formidable engineering challenge. The core design of these detectors is based on a Fabry-Pérot Michelson interferometer [1], which measures the optical path-length differences between two

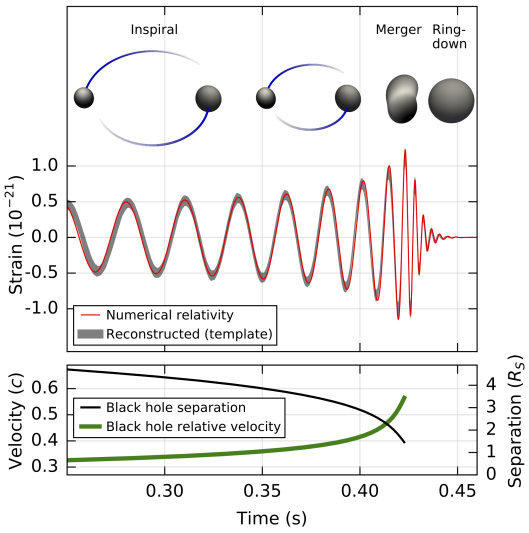

Figure 1.2: Fig. 2 from Ref. [2]. The top panel highlights the different phases of the compact binary inspiral and the associated gravitationa-wave strain. The lower panel shows the separation and velocity of the binary system as a function of time. This schematic figure demonstrates just how energetic systems need to be in order to produce a detectable strain. As the velocities of the balck holes approach $0.6c$, the gravitational-wave strain as measured in ground-based detectors on Earth only reaches $O(10^{-21})$.

perpendicular arms (each 4 km long in the current generation of detectors). These arms house evacuated beam tubes and sustain approximately 400 kW of circulating laser power as of this observing run [17]. Light from each arm is recombined at the beam splitter, producing an interference signal at the output photodiodes. As a gravitational wave passes through the interferometer, it induces differential arm-length changes, modulating the interference pattern. Through precise calibration, the photodiode output is converted back into a strain measurement [31, 32, 33]. While these detectors achieve extraordinary sensitivity—capable of detecting gravitational-wave induced displacements a fraction of a proton's radius—various noise sources inevitably contaminate the measurement. These include stationary Gaussian noise as well as non-Gaussian transient noise artifacts, commonly referred to as "glitches" [34, 17, 35]. Understanding and mitigating these noise sources is critical for conducting statistically robust analyses of astrophysical signals.

Despite the sheer difficulty of building and operating such instruments, the LIGO and Virgo observatories have entered a phase of routine detection [5]. In the ongoing fourth observing run, they are now observing multiple compact binary merger signals each week [18]. Thanks to recent advances in parameter estimation algorithms,

posterior probability distributions for each event can be computed within hours [36, 37]. However, these inferences remain sensitive to both instrumental noise [38, 39, 40] and waveform modeling systematics [41, 42, 43], necessitating careful vetting to avoid biased conclusions.

With the growing catalog of detections, gravitational-wave astronomy has facilitated the development of population studies [6, 7, 8]. Rather than analyzing events in isolation, we can now study ensembles of signals using hierarchical inference techniques. This enables investigations into the underlying astrophysical distributions of binary compact objects [8], the neutron star equation of state [44], and possible deviations from General Relativity [15]. Nevertheless, population-level analyses are not immune to bias. In addition to biases present in individual event analyses, abstract sources of error can arise in hierarchical studies from prior mismatches, incorrect modeling of selection effects, and model misspecification. These subtle effects can propagate through hierarchical pipelines and impact astrophysical conclusions.

Looking toward the future, third-generation detectors such as Cosmic Explorer [45, 46] and the Einstein Telescope [47] promise to extend gravitational-wave astronomy to its cosmic horizon. These instruments will be capable of detecting nearly all compact binary mergers in the observable Universe [46]. Their unprecedented sensitivity will not only improve the precision of parameter estimation but may also enable the detection of new phenomena, such as post-merger gravitational-wave radiation from neutron star collisions. The conceptual design and technical planning of these observatories are active areas of current research.

In this thesis, I will expand upon several aspects of robust gravitational-wave data analysis. I focus on methods for reliable inference of individual binary parameters and for ensemble-level studies relevant to tests of General Relativity and population modeling. I also discuss future prospects for enhancing detector capabilities, with an emphasis on enabling neutron star post-merger science.

1.1 Motivation for testing gravity with gravitational waves

In this thesis, I will spend Chapters 4, 5, and 6 focusing on improvements to tests of gravity with gravitational-wave observations. While these chapters explore my contributions into this topic within the field, here I outline the importance and relevance of such studies using gravitational-wave signals from compact binaries.

To motivate this discussion, we can consider the form of possible extensions to general relativity as higher-order curvature terms in the expansion of the Einstein-

Hilbert action [48],

$$S \sim \int \mathrm{d}^4x\sqrt{-g}\left(R \times \vartheta_0 + \alpha_1\vartheta_1\,\text{curvature}^2 + \alpha_2\,\text{curvature}^3 + ...\right). \tag{1.1}$$

In this expression, R is the Ricci scalar, ϑ_i are additional fields, α_i are coupling coefficients, and "curvature$^\gamma$" is used to represent a number, γ, of Riemann tensor, Ricci scalar and/or tensor terms contracted together. In this expression, as $\alpha_i \to 0$ and $\vartheta_0 \to 1$, the Einstein-Hilbert action, and hence general relativity is recovered.

From the schematic expression in Eq. (1.1), the different classes of modifications to general relativity manifest at each relevant curvature order. First, at the linear curvature order, *cosmological* effects appear corresponding to deviations at arise over large distances. Theories of this class include Brans-Dicke gravity [49] and massive gravity [50, 51]. Solar system tests have greatly constrained these classes of theories [52, 53]. The second class that appears arises from second order curvature terms. Since second order curvature terms—without an additional field—are a topological invariant in four-dimensional spacetimes [54], deviations at this curvature order require the introduction of *additional physics* in the form of additional fields or dimensions. Theories of this nature include many that are often discussed in strong-field tests of gravity such as Einstein-dilaton Gauss-Bonnet gravity [55], dynamical Chern-Simons gravity [56], and Lovelock gravity [57]. Finally, higher-order curvature terms are added to the expansion can be viewed as the most generic term thus far—falling under the broad categorization of general expectations from effective field theory arguments [58, 59, 60]. These terms may not (but can) include additional fields, and so generically these terms can appear at these higher orders. At present, there is still ongoing discourse over whether the cubic order term can exist without introducing a "fifth force" or long-range interactions. See Ref. [58] and references therein for an extended discussion regarding this. The relevance of this discussion of specific theories will become more relevant in Chap. 6.

Now, it is important to understand the relative importance of gravitational-wave observations when attempting to probe these theories. To illustrate their role, Fig. 1.3 from Ref. [61] shows the characteristic gravitational potential and curvature for a number of different tests of gravity. In this figure, the lower left corner corresponds to the curvature and potentials associated with solar system tests such as the perihelion precession of Mercury and the Shapiro delay in signals from the Cassini probe [62]. As we move to higher potentials, past the orbital trajectory observations of the S2 star around Sgr A*, the plot shows the scales associated with massive black holes

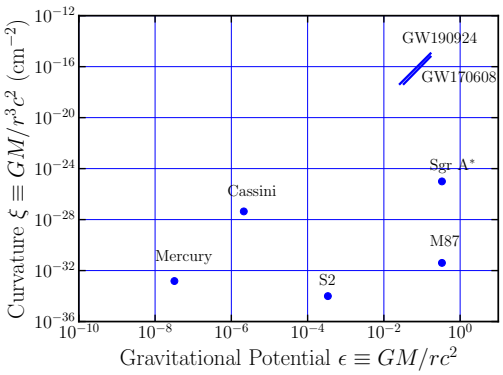

Figure 1.3: Fig. 1 from Ref. [61]. Gravitational potential and curvature for a number of different tests of General Relativity. The plot covers scales and tests from within the Solar System (Mercury's perihelion precession and the Shapiro delay of the Cassini probe's signal) to massive black holes Sgr A* and M87. In the top right corner, the characteristic potential and curvature swept by two binary black-hole inspirals are shown.

such as Sgr A* and the central black hole of M87, both now observed by the Event Horizon Telescope [63, 64]. These observations are probing high potentials but the curvature is still rather small given these black holes have a mass of the order of $\mathcal{O}(10^6 - 10^9)\,M_\odot$. Then, while there is a fundamental limit on the gravitational potential such that it is less than one, higher curvature can still be observed as the mass of the black hole decreases. This leads to observations made of stellar-mass black-hole binaries by ground-based gravitational-wave detectors [12, 13, 14, 15]. Under the expectation that these $\mathcal{O}(10)\,M_\odot$ binary black-hole systems are the lightest such astrophysically-formed systems in the Universe, their gravitational-wave radiation will encode effects present in the highest curvature environments. Relating this to Eq. (1.1), this implies that any higher-order curvature terms will manifest more greatly in these systems, and therefore that gravitational waves from stellar-mass binary black holes will provide the best probes of such effects.

Finally, given that gravitational-wave observations provide the potential to most deeply probe the nature of gravity, it is important to consider the impact such extensions to general relativity might have on the gravitational radiation. An example of the impact of extensions to general relativity, Fig. 1.4 from Ref. [65] shows numerical simulations of gravitational-wave radiation generated from binary black-hole coalescences under dynamical Chern-Simons gravity [56] with varying coupling lengthscale, ℓ. The crucial detail from these results—which, in turn, is typically the case for many realistic extensions to general relativity—is that the extension leads

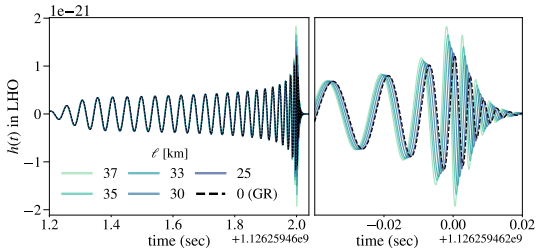

Figure 1.4: Fig. 2 from Ref. [65]. The numerical relativity simulations of a binary black-hole merger assuming dynamical Chern-Simons gravity with different coupling lengthscales, ℓ. The observed gravitational-wave strain experiences an increased dephasing of the signal as the lengthscale increases. There are also amplitude modifications that occur closer to merger, though these are subdominant in any statistical analyses of gravitational-wave observations.

to an accumulated dephasing of the signal. This anticipated modification allows for the construction of statistical tests that primarily search for these dephasings. The current, most-common approach in gravitational-wave astronomy is to use a post-Newtonian parameterization [66, 67, 68, 69, 70, 71]. This parameterization is discussed in Sec. 4.5 and used extensively throughout Chapters 4, 5, and 6. There are additional effects that can also be considered such as additional polarizations [72] and birefringence [73], but for these studies, I have focused my time on direct modifications to the tensorial gravitational-wave strain.

1.2 Overview of Thesis' Contents

The advent and rapid growth of the field of gravitational-wave astronomy has lent itself well to allowing for many impactful studies to be undertaken across a range of topics. Since many of the analysis methods rely on Bayesian inference techniques at their core [74], many approaches are transferrable from exploring individual-event spin parameter constraints to formulating statistical methods for potentially radical improvements to future gravitational-wave detector designs. In my thesis, the primary focus has been on a series of improvements made to heirarchical methods for testing General Relativity. The overview of my thesis is as follows.

Chapter 2 contains published work from Ref. [75], in which we revisit the potential evidence for spin-orbit coupling and therefore spin-precession in GW200129_065458 (henceforth GW200129) [76]. This gravitational-wave observation was known to possess a glitch in the 20-50 Hz frequency range in LIGO Livingston (in addition to noise artifacts in Virgo) [77]. We show that the difference between a spin-precessing and a non-precessing interpretation for GW200129 is smaller than the statistical and

systematic uncertainty of the subtraction, finding that the support for spin-precession depends sensitively on the glitch modeling. We conclude that while our analysis does not disprove the possibility of spin-precession in GW200129, we argue any inference is contingent upon the statistical and systematic uncertainty of the glitch mitigation.

Chapter 3 contains published work from Ref. [78]. In this work, we aim to determine the utility of spin alignment, quantified by $\chi_{\rm eff}$, and spin-precession, quantified by $\chi_{\rm p}$, at classifying the formation channel through which a binary black-hole system formed. In dense stellar environments, black-hole binaries can form through dynamical captures, leading to isotropic orientations of their spins [79, 80, 81]. Whereas in isolated binary formation channels, black hole binaries are expected to form with small spin magnitudes preferentially aligned with the orbital angular momentum. Given the expected isotropic distribution of component spins of binary black hole in gas-free dynamical environments [81], the presence of anti-aligned or in-plane spins with respect to the orbital angular momentum is considered a tell-tale sign of a merger's dynamical origin. Using Monte Carlo cluster simulations to generate a realistic distribution of hierarchical merger parameters from globular clusters [82, 83], we simulate the detection and inference of the binary black-hole mergers' parameters. Using a "likelihood-ratio"-based statistic, we find that $\sim 2\%$ of the recovered population by the current gravitational-wave detector network has a statistically significant $\chi_{\rm p}$ measurement, whereas no $\chi_{\rm eff}$ measurement was capable of confidently determining a system to be anti-aligned with the orbital angular momentum at current detector sensitivities. These results indicate that measuring spin-precession is a more detectable signature of hierarchical mergers and dynamical formation than anti-aligned spins.

Chapters 4 and 5 contain published works from Ref. [84] and Ref. [85], respectively. These works focus on important improvements to tests of gravity with ensembles of observations. In Chapter 4, we incorporate, for the first time in the literature, information about the underlying astrophysical population to avoid biases in the inference of deviations from general relativity. Current tests assume that the astrophysical population follows an unrealistic fiducial prior chosen to ease sampling of the posterior which is is inconsistent with both astrophysical expectations and the distribution inferred from observations [15]. We propose a framework for fortifying tests of general relativity by simultaneously inferring the astrophysical population using a catalog of detections. Using observations from LIGO-Virgo-KAGRA's third

observing run, we show that concurrent inference of the astrophysical distribution strengthens constraints and improves the overall ability to detect deviations from general relativity if present.

In Chapter 5 we estimate the impact of selection effects for tests of gravity using the inspiral phase evolution of compact binary signals with a simplified version of the GstLAL search pipeline. Leveraging the work presented in Chapter 4, we find that selection biases affect the search for very large values of the deviation parameters, much larger than the hierarchical constraints implied by the detected gravitational-wave signals. Therefore, combined population constraints for deviations from general relativity from confidently detected events are mostly unaffected by selection biases. These findings suggest that current population constraints on the inspiral phase are robust without factoring in selection biases.

Chapter 6 is based on work in Ref. [86]. In this study, we consider the role theoretical information high-energy extensions to General Relativity play in more theoretically motivated tests of gravity. High-energy extensions to General Relativity modify the Einstein-Hilbert action with higher-order curvature corrections and theory-specific coupling constants [87, 88]. The order of these corrections imprints a universal curvature dependence on observations while the coupling constant controls the deviation strength. We leverage the theory-independent expectation that modifications to the action of a given order in spacetime curvature (Riemann tensor and contractions) lead to observational deviations that scale with the system length-scale to a corresponding power. We incorporate this universal scaling into the theory-agnostic tests of General Relativity presented in Chapter 4 with current gravitational-wave observations, thus enabling constraints on the curvature scaling without compromising the agnostic nature of these tests.

Chapter 8 is based on work currently in preparation for publication. In this work, we turn our attention to improvements to future third-generation gravitational-wave detectors in the coming decades for improving their observational capabilities. Binary neutron star post-merger signals are a high-priority target for third-generation gravitational-wave detectors, as they encode valuable information about the dense-matter equation of state in their high-frequency gravitational-wave signatures [89, 45, 90]. Future detectors like Cosmic Explorer and the Einstein Telescope are expected to observe millions of compact binary coalescences; however, their sensitivity is dominated by quantum noise above $\sim 1\,\mathrm{kHz}$ [91], hindering the detection of these signals [17]. We present the statistical background and methodology for utilizing a

newly suggested readout scheme for future detectors, known as photon counting [92]. In such a readout, signals and noise become quantized into discrete distributions corresponding to the detection of single photons measured in a chosen basis of modes for each event. Through simulations of realistic binary neutron star signals, we demonstrate that photon counting can extract meaningful information even from signals with low signal-to-noise ratios. Furthermore, we show that this capability improves the prospects for equation-of-state parameter inference from post-merger signals. We also demonstrate that the hierarchical constraints scale more favorably with possible detector classical noise sensitivity improvements. These results indicate that photon counting offers a promising alternative to traditional homodyne readout techniques for extracting information from low signal-to-noise ratio post-merger signals. These methods can potentially be extended to other observational science targets in the third-generation detector era.

Finally, in Chapter 9, I summarize the work presented and I outline possible future avenues for the ideas put forward in this thesis.

Chapter 2

THE CURIOUS CASE OF GW200129: INTERPLAY BETWEEN SPIN-PRECESSION INFERENCE AND DATA-QUALITY ISSUES

E. Payne, S. Hourihane, J. Golomb, R. Udall, D. Davis, and K. Chatziioannou. "Curious case of GW200129: Interplay between spin-precession inference and data-quality issues". In: *Phys. Rev. D* 106.10 (2022), p. 104017. DOI: 10.1103/PhysRevD.106.104017. arXiv: 2206.11932 [gr-qc].

E.P. carried out the majority of the parameter estimation analyses using Parallel Bilby for the results presented, and contributed significantly to the writing of the manuscript.

2.1 Introduction

GW200129_065458 (henceforth GW200129) is a gravitational wave (GW) candidate reported in GWTC-3 [77]. The signal was observed by all three LIGO-Virgo detectors [1, 16] operational during the third observing run (O3) and it is consistent with the coalescence of two black holes (BHs) with source-frame masses $34.5^{+9.9}_{-3.2}\,M_\odot$ and $28.9^{+3.4}_{-9.3}\,M_\odot$ at the 90% credible level. Though the masses are typical within the population of observed events [8], the event's signal-to-noise-ratio (SNR) of $26.8^{+0.2}_{-0.2}$ makes it the loudest binary BH (BBH) observed to date. Additionally, it is one of the loudest triggers in the Virgo detector with a detected SNR of 6–7 depending on the detection pipeline [77]. The signal temporally overlapped with a glitch in the LIGO Livingston detector, which was subtracted using information from auxiliary channels [38]. The detection and glitch mitigation procedures for this event are recapped in App. 2.6.

The interpretation of some events in GWTC-3 was impacted by waveform systematics, with GW200129 being one of the most extreme examples. As part of the catalog, results were obtained with the `IMRPhenomXPHM` [93] and `SEOBNRv4PHM` [94] waveform models using the parameter inference algorithms `Bilby` [36, 95] and `RIFT` [96] respectively. Both waveforms correspond to quasicircular binary inspirals and include high-order radiation modes and the effect of relativistic spin-precession arising from interactions between the component spins and the orbital angular momentum.

All analyses used the glitch-subtracted LIGO Livingston data. The `IMRPhenomXPHM` result was characterized by large spins and a bimodal structure with peaks at ~ 0.45 and ~ 0.9 for the binary mass ratio. The `SEOBNRv4PHM` results, on the other hand, pointed to more moderate spins and near equal binary masses. Both waveforms, however, reported a mass-weighted spin aligned with the Newtonian orbital angular momentum of $\chi_{\rm eff} \sim 0.1$, and thus the inferred large spins with `IMRPhenomXPHM` corresponded to spin components in the binary orbital plane and spin-precession. Such differences between the waveform models are not unexpected for high SNR signals [41]. Waveform systematics are also likely more prominent when it comes to spin-precession, as modeling prescriptions vary and are not calibrated to numerical relativity simulations featuring spin-precession [93, 97, 94]. Data quality issues could further lead to evidence for spin-precession [98]. Due to differences in the inference algorithms and waveform systematics, GWTC-3 argued that definitive conclusions could not be drawn regarding the possibility of spin-precession in this event [77].

Stronger conclusions in favor of spin-precession [76] and a merger remnant that experienced a large recoil velocity [99] were put forward by means of a third waveform model. `NRSur7dq4` [100] is a surrogate to numerical relativity simulations of merging BHs that is also restricted to quasicircular orbits and models the effect of high-order modes and spin-precession. This model exhibits the smallest mismatch against numerical relativity waveforms, sometimes comparable to the numerical error in the simulations. It is thus expected to generally yield the smallest errors due to waveform systematics [100]. This fact was exploited in Hannam *et al.* [76] to break the waveform systematics tie and argue that the source of GW200129 exhibited relativistic spin-precession with a primary component spin magnitude of $\chi_1 = 0.9^{+0.1}_{-0.5}$ at the 90% credible level.

During a binary inspiral, spin-precession is described through post-Newtonian theory [101, 102]. Spin components that are not aligned with the orbital angular momentum give rise to spin-orbit and spin-spin interactions that cause the orbit to change direction in space as the binary inspirals, e.g., [103, 104, 105, 106, 107, 108, 71, 109, 110, 111]. The emitted GW signal is modulated in amplitude and phase, and morphologically resembles the beating between two spin-aligned waveforms [112] or a spin-aligned waveform that has been "twisted-up" [105, 106]. As the binary reaches merger, numerical simulations suggest that the direction of peak emission continues precessing [113]. Parameter estimation analyses using `NRSur7dq4` find

that spins and spin-precession can be measured from merger-dominated signals for certain spin configurations [114], however the lack of analytic understanding of the phenomenon means that it is not clear how such a measurement is achieved.

The main motivation for this study is to revisit GW200129 and attempt to understand how spins and spin-precession can be measured from a heavy BBH with a merger-dominated observed signal. In Sec. 2.2 we use `NRSur7dq4` to conclude that the evidence for spin-precession originates exclusively from the LIGO Livingston data in the 20–50 Hz frequency range, where the inferred signal amplitude is lower than what a spin-aligned binary would imply given the rest of the data. This range coincides with the known data quality issues described in App. 2.6 and first identified in GWTC-3 [77]. LIGO Hanford is consistent with a spin-aligned signal, causing an inconsistency between the inferred mass ratio q and precession parameter χ_p inferred from each LIGO detector separately. By means of simulated signals, we argue that such $q - \chi_p$ inconsistency is unlikely to be caused solely by the different Gaussian noise realizations in each detector at the time of the signal, rather pointing to remaining data quality issues beyond the original glitch-subtraction [77]. We also re-analyze the LIGO Livingston data above 50 Hz, (while keeping the original frequency range of the LIGO Hanford data) and confirm that all evidence for spin-precession disappears.

In the process, we find that the Virgo trigger, though consistent with a spin-aligned BBH, is *inconsistent* with the signal seen in the LIGO Hanford and LIGO Livingston detectors. Specifically, the Virgo data are pointing to a much heavier BBH that merges ∼20 ms earlier than the one observed by the LIGO detectors. We discuss Virgo data quality considerations in Sec. 2.3 within the context of a potential glitch that affects the inferred binary parameters if unmitigated. As a consequence, we do not include Virgo data in the sections examining spin-precession unless otherwise stated. The Virgo-LIGO inconsistency can be resolved if we use `BayesWave` [115, 116, 117] to simultaneously model a CBC signal and glitches with CBC waveform models and sine-Gaussian wavelets respectively [118, 39]. The Virgo data are now consistent with the presence of both a signal that is consistent with the one in the LIGO detectors and an overlaping glitch with SNR ∼ 4.6.

In Sec. 2.4 we revisit the LIGO Livingston data quality issues and compare the original glitch-subtraction based on `gwsubtract` [119, 38] that uses information from auxiliary channels and the glitch estimate from `BayesWave` that uses only strain data. Though the CBC model used in `BayesWave` does not include the effect

of spin-precession, we show that differences between the reconstructed waveforms from a non-precessing and spin-precessing analysis for GW200129 are *smaller* than the statistical uncertainty in the glitch inference. Such differences can therefore not be reliably resolved in the presence of the glitch and its subtraction procedure. The two glitch estimation methods give similar results within their statistical errors, however `gwsubtract` yields typically a lower glitch amplitude. We conclude that any evidence for spin-precession from GW200129 is contingent upon the systematic and statistical uncertainties of the LIGO Livingston glitch subtraction. Given the low SNR of the LIGO Livingston glitch and the glitch modeling uncertainties, we can at present not conclude whether the source of GW200129 exhibited spin-precession or not.

In Sec. 2.5 we summarize our arguments that remaining data quality issues in LIGO Livingston cast doubt on the evidence for spin-precession. Besides data quality studies (i.e., spectrograms, glitch modeling, auxiliary channels), our investigations are based on comparisons between different detectors as well as different frequency bands of the same detector. We propose that similar investigations in further events of interest with exceptional inferred properties could help alleviate potential contamination due to data quality issues.

2.2 The origin of the evidence for spin-precession

Our main goal is to pinpoint the parts of the GW200129 data that are inconsistent with a non-precessing binary and understand the relevant signal morphology. Due to different orientations, sensitivities, and noise realizations, different detectors in the network do not observe an identical signal. The detector orientation, especially, affects the signal polarization content and thus the degree to which spin-precession might be measurable in each detector. Motivated by this, we begin by examining data using different detector combinations.

We perform parameter estimation using the `NRSur7dq4` waveform and examine data from each detector separately (left panel) as well as the relation between the LIGO and the Virgo data (right panel) and show posteriors for select intrinsic parameters in Fig. 2.1. Analysis settings and details are provided in App. 2.6 and in all cases we use the same LIGO Livingston data as GWTC-3 [77] where the glitch has been subtracted. Though we do not expect the posterior distributions for the various signal parameters inferred with different detector combinations to be identical, they should have broadly overlapping regions of support. If the triggers recorded by the

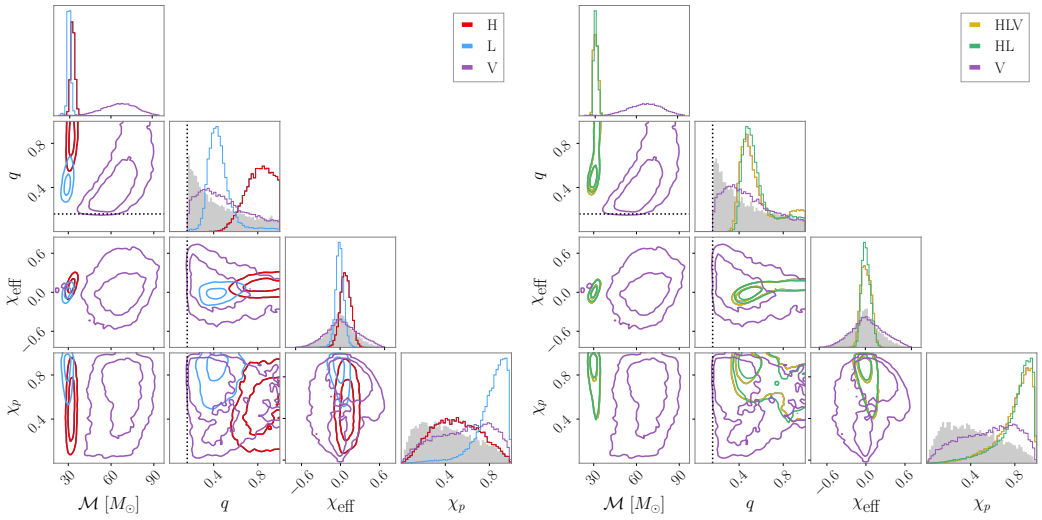

Figure 2.1: One- and two-dimensional marginalized posteriors for select intrinsic binary parameters: detector frame chirp-mass $\mathcal{M}$, mass ratio q, effective spin $\chi_{\rm eff}$, and precessing spin χ_p. See Table 2.1 for analysis settings and App. 2.6 for detailed parameter definitions. Two-dimensional panels show 50% and 90% contours. The black dashed line marks the minimum bound of q=1/6 in `NRSur7dq4`'s region of validity. Shaded regions shows the prior for q, $\chi_{\rm eff}$, χ_p. The $\mathcal{M}$ prior increases monotonically to the maximum allowed value (see App. 2.6 for details on choices of priors). Left panel: comparison between analyses that use solely LIGO Hanford (red; H), LIGO Livingston (blue; L), and Virgo (purple; V) data. Right panel: comparison between analyses of all three detectors (yellow; HLV), only LIGO data (green; HL) and only Virgo data (purple; V). The evidence for spin-precession originates solely from the LIGO Livingston data as the other detectors give uninformative χ_p posteriors. Additionally, the binary masses inferred based on Virgo only are inconsistent with those from the LIGO data.

different detectors are indeed consistent, any shift between the posteriors should be at the level of Gaussian noise fluctuations.

The left panel shows that the evidence for spin-precession arises primarily from the LIGO Livingston data, whereas the precession parameter χ_p posterior is much closer to its prior when only LIGO Hanford or Virgo data are considered. A similar conclusion was reached in Hannam *et al.* [76]. There is reasonable overlap between the two-dimensional distributions that involve the chirp mass $\mathcal{M}$, the mass ratio q, and the effective spin $\chi_{\rm eff}$ inferred by the two LIGO detectors, as expected from detectors that observe the same signal under different Gaussian noise realizations. The discrepancy between the spin-precession inference in the two LIGO detectors, however, is evident in the $q-\chi_p$ panel. The two detectors lead to non overlapping distributions that point to either unequal masses and spin-precession

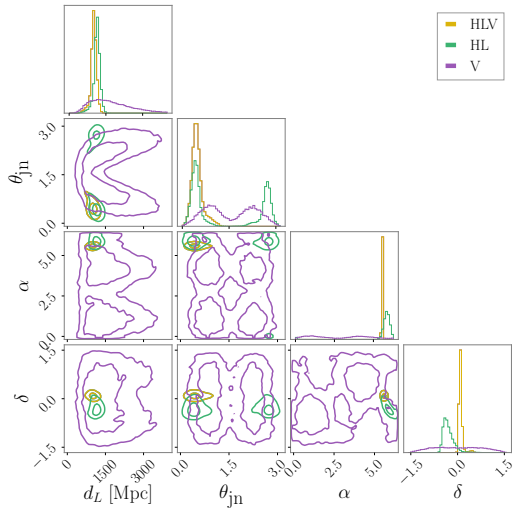

Figure 2.2: Similar to the right panel of Fig. 2.1 but for select extrinsic parameters: luminosity distance d_L, angle between total angular momentum and line of sight $\theta_{\rm jn}$, right ascension α, and declination δ. For reference, the median optimal SNR for each run is HLV: 27.6, HL: 26.9, V: 6.7.

(LIGO Livingston), or equal masses and no information for spin-precession (LIGO Hanford).

Besides an uninformative posterior on χ_p, the left panel points to a bigger issue with the Virgo data: inconsistent inferred masses. The right panel examines the role of Virgo in more detail in comparison to the LIGO data. Due to the lower SNR in Virgo, the intrinsic parameter posteriors are essentially identical between the HL and the HLV analyses. The lower total SNR means that the Virgo-only posteriors will be wider, but they are still expected to overlap with the ones inferred from the two LIGO detectors. However, this is not the case for the mass parameters as is most evident from the two dimensional panels involving the chirp mass. While the LIGO data are consistent with a typical binary with (detector-frame) chirp mass $30.3^{+2.5}_{-1.6}\,M_\odot$ at the 90% credible level, the Virgo data point to a much heavier binary with $66.7^{+19.7}_{-22.6}\,M_\odot$ at the same credible level.

The role of Virgo data on the inferred binary extrinsic parameters is explored in Fig. 2.2. In general, Virgo data have a larger influence on the extrinsic than the intrinsic parameters as the measured time and amplitude helps break existing degeneracies. The extrinsic parameter posteriors show a large degree of overlap. The Virgo distance posterior does not rail against the upper prior cut off, suggesting that this detector does observe some excess power. The HL sky localization also overlaps with the Virgo-only one, though the latter is merely the antenna pattern

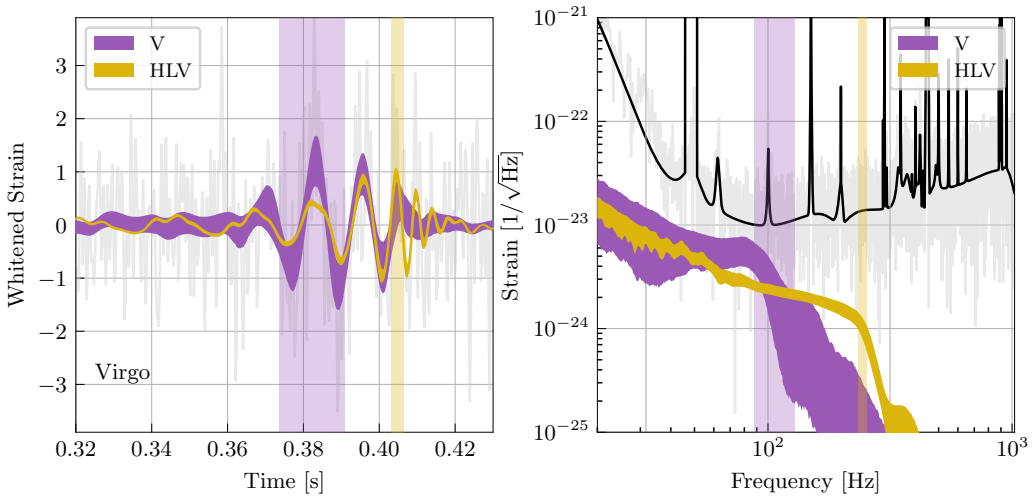

Figure 2.3: 90% credible intervals for the whitened time-domain reconstruction (left) and spectrum (right) of the signal in Virgo from a Virgo-only (purple; V) and a full 3-detector (yellow; HLV) analysis; see Table 2.1 for analysis settings. The data are shown in gray and the noise PSD in black. The time on the left plot is relative to GPS 1264316116. The high value of the PSD at ~ 50 Hz was imposed due to miscalibration of the relevant data [77]. Vertical shaded regions at each panel correspond to the 90% credible intervals of the merger time (left; defined as the time of peak strain amplitude) and merger frequency (right; approximated via the dominant ringdown mode frequency as computed with `qnm` [120], merger remnant properties were computed with `surfinBH` [121]). The Virgo data point to a heavier binary that merges $\sim$ 20ms earlier than the full 3-detector results that are dominated by the LIGO detectors.

of the detector that excludes the four Virgo "blind spots." We use the HL results to calculate the projected waveform in Virgo and calculate the 90% lower limit on the signal SNR to be 4.2. This suggests that given the LIGO data, Virgo should be observing a signal with at least that SNR at the 90% level.

In order to track down the cause of the discrepancy in the inferred mass parameters, we examine the Virgo strain data directly. Figure 2.3 shows the whitened time-domain reconstruction (left panel) and the spectrum (right panel) of the signal in Virgo from a Virgo-only and a full 3-detector analysis. Compared to Figs. 2.1 and 2.2, here we only consider a 3-detector analysis as the reconstructed signal in Virgo inferred from solely LIGO data would not be phase-coherent with the data, and thus would be uninformative. Given the higher signal SNR in the two LIGO detectors, the signal reconstruction morphology in Virgo is driven by them, as evident from the intrinsic parameter posteriors from the right panel of Fig. 2.1.

The two reconstructions in Fig. 2.3 are morphologically distinct. The 3-detector inferred signal is dominated by the LIGO data and resembles a typical "chirp" with increasing amplitude and frequency. This signal is, however, inconsistent with the Virgo data as it underpredicts the strain at $t \sim 0.382$ s in the left panel. The Virgo-only inferred signal matches the data better by instead placing the merger at earlier times to capture the increased strain at $t \sim 0.382$ s as shown by the shaded vertical region denoting the merger time. Rather than a "chirp", the signal is dominated by the subsequent ringdown phase with an amplitude that decreases slowly over ~2 cycles. As also concluded from the inferred masses in Fig. 2.1, the Virgo data point to a heavier binary with lower ringdown frequency (vertical regions in the right panel).

Despite these large inconsistencies, the issues with the Virgo data do not affect our main goal, which is identifying the origin of the evidence for spin-precession. In order to avoid further ambiguities for the remainder of this section we restrict to data from the two LIGO detectors unless otherwise noted. In Fig. 2.1 we concluded that LIGO Livingston alone drives this measurement and here we attempt to further zero in on the data that support spin-precession by comparing results from a spin-precessing and a spin-aligned analysis with `NRSur7dq4`, see App. 2.6 for details. Figure 2.4 shows the whitened time-domain reconstruction (left panel) and the spectrum (right panel) in LIGO Hanford (top) and LIGO Livingston (bottom). The two reconstructions remain phase-coherent, however there are some differences in the inferred amplitudes, with the spin-aligned amplitude being slightly larger at ~30–50 Hz and slightly smaller for $\gtrsim$ 100 Hz. Comparison to the estimate for the glitch that was subtracted from the data based on information from auxiliary channels with `gwsubtract` shows that the glitch overlaps with the part of the signal where the spin-precessing amplitude is smaller than the spin-aligned one. The glitch subtraction and data quality issues are therefore related to the evidence for spin-precession.

We confirm that the low-frequency data in LIGO Livingston (in relation to the rest of the data) are the sole source of the evidence for spin-precession, by carrying out analyses with a progressively increasing low frequency cutoff in LIGO Livingston only, while leaving the LIGO Hanford data intact. Figure 2.5 shows the effect on the posterior for χ_p, q, and $\chi_{\rm eff}$. When we use the full data bandwidth, $f_{\rm low}(L) = 20$ Hz, we find that q and χ_p are correlated and their two-dimensional posterior appears similar to the combination of the individual-detector posteriors from Fig. 2.1.

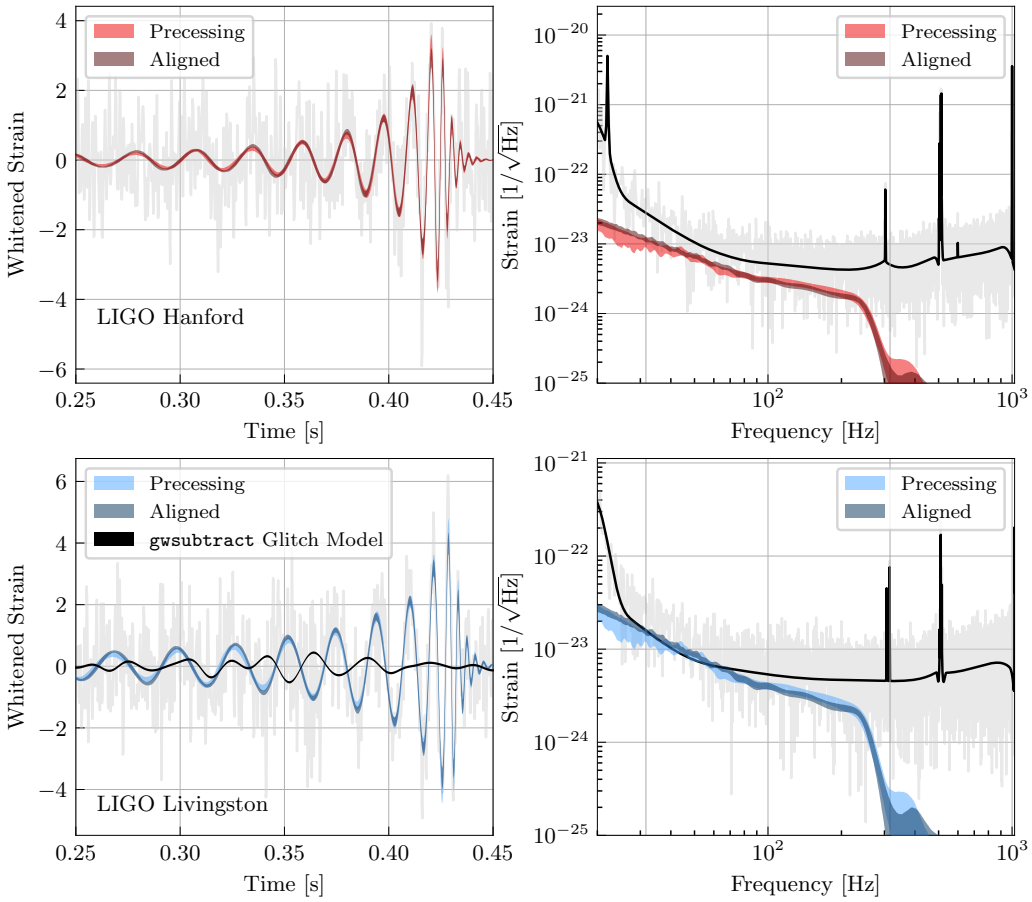

Figure 2.4: Whitened time-domain reconstruction (left) and spectrum (right) of GW200129 in LIGO Hanford (top) and LIGO Livingston (bottom). Shaded regions show the 90% credible intervals for the signal using a spin-precessing (light blue and red) and a spin-aligned (dark blue and red) analysis based on `NRSur7dq4`, see Table 2.1 for run settings. In gray we show the analyzed data where the `gwsubtract` estimate for the glitch (black line) has already been subtracted. The black line in the right panels is the noise PSD. The glitch overlaps with the part of the inferred signal where the spin-aligned amplitude is on average larger than the spin-precessing one.

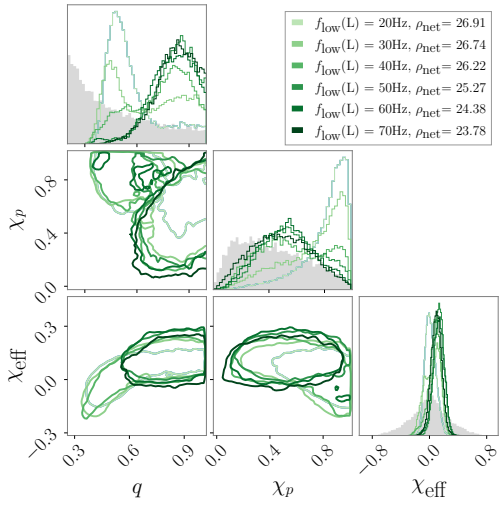

Figure 2.5: One- and two-dimensional marginalized posterior for the mass ratio q, the precession parameter χ_p, and the effective spin parameter $\chi_{\rm eff}$ for analyses using a progressively increasing low frequency cutoff in LIGO Livingston but all the LIGO Hanford data, see Table 2.1 for details. The median network SNR for each value of the frequency cutoff is given in the legend. Contours represent 90% credible regions and the prior is shaded in gray. As the glitch-affected data are removed from the analysis, the posterior approaches that of an equal-mass binary and becomes uninformative about χ_p. This behavior does not immediately indicate data quality issues and we only use this increasing-$f_{\rm low}(L)$ analysis to isolate the data which contribute the evidence of spin-precession when compared to the rest of the data to within 20–50 Hz.

However, as the low frequency cutoff in LIGO Livingston is increased and the data affected by the glitch are removed, the posterior progressively becomes more consistent with an equal-mass binary and χ_p approaches its prior. By $f_{\rm low}(L) = 50$ Hz, χ_p is similar to its prior and further increasing $f_{\rm low}(L)$ has a marginal effect. This confirms that *given all the other data*, the LIGO Livingston data in 20–50 Hz drive the inference for spin-precession.

The signal network SNR (i.e., the SNR in both detectors added in quadrature) is given in the legend for each value of the low frequency cutoff. By $f_{\rm low}(L) = 50$ Hz where all evidence for spin-precession has been eliminated, the SNR reduction is only 1.5 units, suggesting that the large majority of the signal is consistent with a non-precessing origin. This might also suggest that χ_p inference is not degraded solely due to loss of SNR, as the latter is very small. The $\chi_{\rm eff}$ posterior is generally only minimally affected, with a small shift to higher values driven by the $q - \chi_{\rm eff}$ correlation [122]. We have verified that these conclusions are robust against re-

including the Virgo data (using their full bandwidth).

The above analysis is *not* on its own an indication of data quality issues in LIGO Livingston, but we now turn to an observation that might be more problematic: the $q - \chi_p$ inconsistency between LIGO Hanford and LIGO Livingston identified in Fig. 2.1. In order to examine whether such an effect could arise from the different Gaussian noise realizations in each detector, we consider simulated signals. We use 100 posterior samples obtained from analyzing solely the LIGO Livingston data, make simulated data that include a noise realization with the same noise PSDs as GW200129, and analyze data from the two LIGO detectors separately. To quantify the degree to which the LIGO Hanford and LIGO Livingston posteriors overlap, we compute the Bayes factor for overlapping posterior distributions relative to if the two distributions do not overlap [123, 124],

$$\mathcal{B}^{\text{overlapping}}_{\text{not overlapping}} = \iint \mathrm{d}\chi_p \mathrm{d}q \, \frac{p_L(\chi_p, q|d) p_H(\chi_p, q|d)}{\pi(\chi_p, q)}, \tag{2.1}$$

where we compute the overlap within the $q-\chi_p$ plane, $p_L(\chi_p, q|d)$ and $p_H(\chi_p, q|d)$ are the LIGO Livingston and LIGO Hanford posteriors, and $\pi(\chi_p, q)$ is the prior. While evaluating this quantity is subject to sizeable sampling uncertainty for events where the two distributions are more distinct (i.e., the case of GW200129), we find $\mathcal{O}(5/100)$ injections have a similar overlap as GW200129 (Fig. 2.1). Figure 2.6 shows a selection of $q-\chi_p$ posteriors for 10 injections as inferred from each detector separately. The posteriors typically overlap, though they are shifted with respect to each other as expected from the different noise realizations.

We conclude that the evidence for spin-precession originates exclusively from the LIGO Livingston data that overlapped with a glitch. This causes an inconsistency between the LIGO Hanford and LIGO Livingston that we typically do not encounter in simulated signals in pure Gaussian noise. This inconsistency suggests that there might be residual data quality issues in LIGO Livingston that were not fully resolved by the original glitch subtraction. Though inconsequential for the spin-precession investigation, we also identify severe data quality issues in Virgo. Before returning to the investigation of spin-precession, we first examine the Virgo data in detail in Sec. 2.3 and argue that they should be removed from subsequent analyses. We reprise the spin-precession investigations in Sec. 2.4.

2.3 Data quality issues: Virgo

Having established that the Virgo trigger is coincident but not fully coherent with the triggers in the two LIGO detectors, we explore potential reasons for this discrepancy.

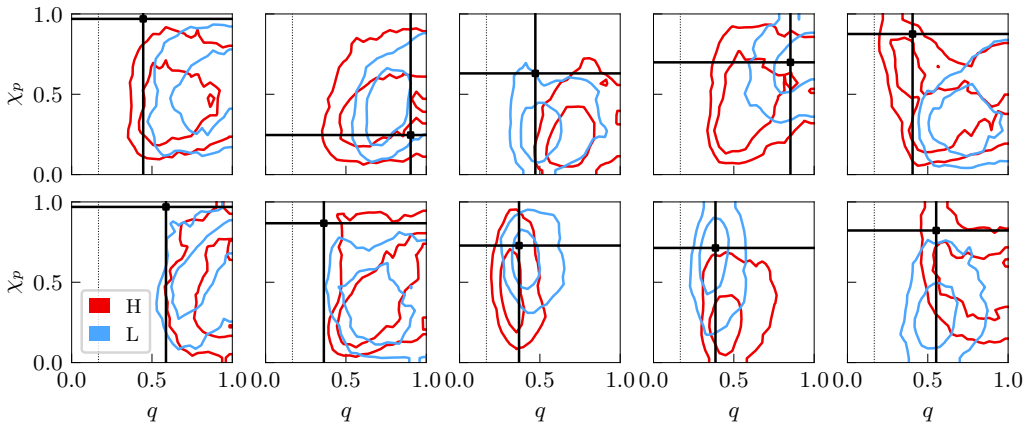

Figure 2.6: 90% contours for the two-dimensional marginalized posteriors for the mass ratio q and the precessing parameter χ_p obtained from analyzing data from each LIGO detector separately for 10 simulated signals. The signal parameters are drawn from the posterior for GW200129 when using LIGO Livingston data only and true values are indicated by black lines. Due to the spin priors disfavoring large χ_p, the injected value is outside the two-dimensional 90% contour in some cases. We only encounter an inconsistency between LIGO Hanford (red; H) and LIGO Livingston (blue; L) as observed for GW200129 in Fig. 2.1 in $O(5/100)$ injections.

Figure 2.7 shows a spectrogram of the data in each detector centered around the time of the event. A clear chirp morphology is visible in the LIGO detectors but not in Virgo, though this might also be due to the low SNR of the Virgo trigger. Within a few seconds of the trigger, however, a number of other glitches are also present in Virgo, mostly assigned to scattered light. We estimate the SNR of the Virgo trigger without assuming it is a CBC signal (i.e., without using a CBC model) through Omicron [125] and `BayesWave` using its glitch model that fits the data with sine-Gaussian wavelets, see Table 2.2 for run settings[1]. The former finds a matched-filter Omicron SNR[2] of 7.0, while the latter finds an optimal SNR of 7.3 for the median glitch reconstruction.

Given the prevalence of glitches, the first option is that the Virgo trigger is actually a detector glitch that happened to coincide with a signal in the LIGO detectors. To estimate the probability that such a coincidence could happen by chance, we consider the glitch rate in Virgo. In O3, the median rate of glitches in Virgo was 1.11/min,

[1]The `BayesWave` analyses described here do not concurrently marginalize over the PSD uncertainty.

[2]The SNR reported by Omicron is normalized so that the expectation value of the SNR is 0, rather than $\sqrt{2}$ [125]. To highlight this difference, we use the phrase "Omicron SNR" whenever a reported result uses this normalization.

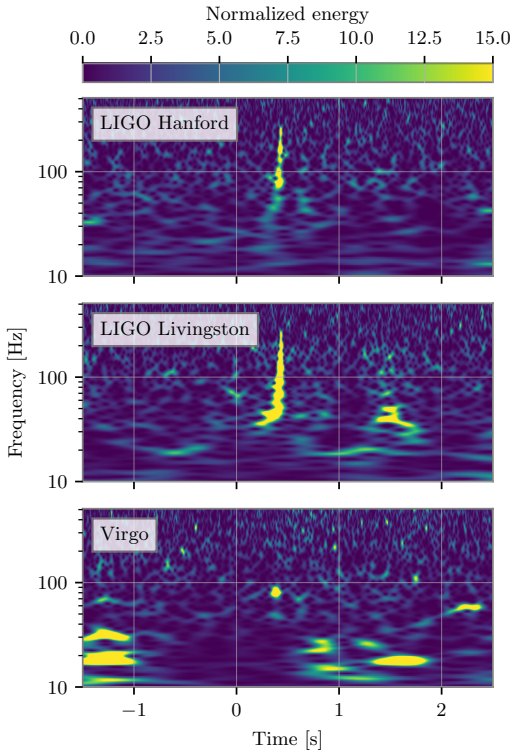

Figure 2.7: Spectrogram of the data in each detector, plotted using plotted using the Q-transform [126, 127]. Listed times are with respect to GPS 1264316116. Besides the clear chirp morphology in LIGO, there is visible excess power $\sim$ 1 s after the signal in LIGO Livingston. Virgo demonstrates a high rate of excess power, though most is due to scattered light and concentrated at frequencies < 30 Hz. The excess power in Virgo that is coincident with GW200129 does not have a chirp morphology.

with significant variation versus time [77]. When we consider the hour of data around the event, the rate of glitches with Omicron SNR > 6.5 is 10.2/min. Most of the glitches in Virgo at this time are due to scattered light [128, 129, 130, 131, 132]. While Fig. 2.7 shows that there are scattered light glitches in the Virgo data near the time of GW200129, the excess power from these glitches are concentrated at frequencies < 30 Hz. To account for the excess power corresponding to GW200129 in Virgo, there must be a different type of glitch present in the data. The rate of glitches at frequencies similar to the signal is much lower; using data from 4 days around the event, the rate of glitches with frequency 60-120 Hz is only 0.06/hr. Given this rate, we calculate the probability that a glitch occurred in Virgo within a 0.06 s window (roughly corresponding to twice the light-travel time between the LIGO detectors and Virgo) around a trigger in the LIGO detectors. We find that if glitches at any frequency are considered, the probability of coincidence per event

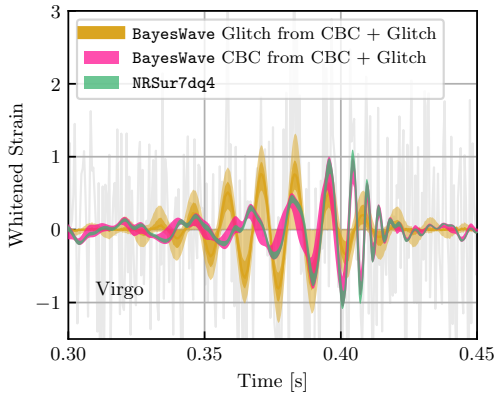

Figure 2.8: Whitened time-domain reconstruction of the signal in Virgo obtained after analysis of data from all three detectors relative to GPS 1264316116. Shaded regions correspond to 90% and 50% (where applicable) credible intervals. Green corresponds to the same 3-detector result obtained with `NRSur7dq4` as Fig. 2.3, while pink and gold correspond to the CBC and glitch part of the "CBC+glitch" analysis with `BayesWave`. See Tables 2.1 and 2.2 for run settings. The two CBC reconstructions largely overlap, suggesting that the lack of spin-precession in `BayesWave`'s analysis does not affect the reconstruction considerably. A glitch overlapping with the signal is, however, recovered.

is $O(0.01)$, and if only glitches with similar frequencies are considered, the same probability is $O(10^{-5})$.

Another option is that the Virgo trigger is a combination of a genuine signal and a detector glitch. We explore this possibility using `BayesWave` [115, 116, 117] to simultaneously model a potential CBC signal that is coherent across the detector network and overlapping glitches that are incoherent [118, 39]. In this "CBC+glitch" analysis, `BayesWave` models the CBC signal with the `IMRPhenomD` waveform [133, 134] and glitches with sine-Gaussian wavelets. Details about the models and run settings are provided in App. 2.6. An important caveat here is that `IMRPhenomD` does not include the effects of higher-order modes and spin-precession. A concern is, therefore, that the CBC model could fail to model precession-induced modulations in the signal amplitude and instead assign them to the glitch model. This precise scenario is tested in Hourihane *et al.* [39] where the analysis was shown to be robust against such systematics. Below we argue that the same is true here for the Virgo data, especially since they are consistent with a spin-aligned binary as shown in Fig. 2.1.

Figure 2.8 compares `BayesWave`'s reconstruction in Virgo with the one obtained with the `NRSur7dq4` analysis from Fig. 2.3 that ignores a potential glitch but models

spin-precession and higher order modes. All results are obtained using data from all three detectors. The CBC reconstruction from `BayesWave` with `IMRPhenomD` is consistent with the one from `NRSur7dq4` to within the 90% credible level at all times. This is unsurprising given Fig. 2.1 that shows that Virgo data are consistent with a spin-aligned BBH. Crucially, there is no noticeable difference between the two CBC reconstructions for times when the inferred glitch is the loudest. This suggests that the lack of higher-order modes and spin-precession in `IMRPhenomD` does not lead to a noticeable difference in the signal reconstruction and could thus not account for the inferred glitch. The differences between the inferred signals using `IMRPhenomD` and `NRSur7dq4` are much smaller than the amount of incoherent power present in Virgo. In fact, the glitch reconstruction is larger than the signal at the 50% credible level, though not at the 90% level. This result suggests that a potential explanation for the trigger in Virgo is a combination of a signal consistent with the one in the LIGO detectors and a glitch.

Figure 2.9 summarizes the various SNR estimates for the excess power in Virgo. We plot an estimate of the SNR in Virgo suggested by LIGO data; in other words it is the SNR that is consistent with GW200129 as observed by LIGO. In comparison, we also show the SNR from a Virgo-only analysis and the SNR from `BayesWave`'s "glitchOnly" analysis that models the excess power with sine-Gaussian wavelets without the requirement that it is consistent with a CBC. The fact that the SNR inferred from HL data is smaller than the other two again suggests that the Virgo trigger is not consistent with the one seen by LIGO and contains additional power. `BayesWave`'s "CBC+glitch" analysis is able to separate the part of the trigger that is consistent with a CBC and recovers a CBC SNR that is consistent to the one inferred from LIGO only. The "remaining" power is assigned to a glitch with SNR ~ 4.6 (computed through the median `BayesWave` glitch reconstruction).

Based on the glitch SNR calculated by the `BayesWave` "CBC+glitch" model, we revisit the probability of overlap with a signal based on the SNR distribution of Omicron triggers. Since the lowest SNR recorded in Omicron analyses is 5.0, we fit the SNR distribution of glitches with Omicron SNR > 5.0 with a power-law and extrapolate to SNR 4.6. We find that the rate of glitches with frequencies similar to the one in Fig. 2.8 with SNR > 4.6 is $0.31/\mathrm{min}$ and the probability of overlap with a signal in Virgo is $O(10^{-3})$. Given the 60 events from GWTC-3 that were identified in Virgo during O3, the overall chance of at least one glitch of this SNR overlapping a signal is $O(0.1)$.

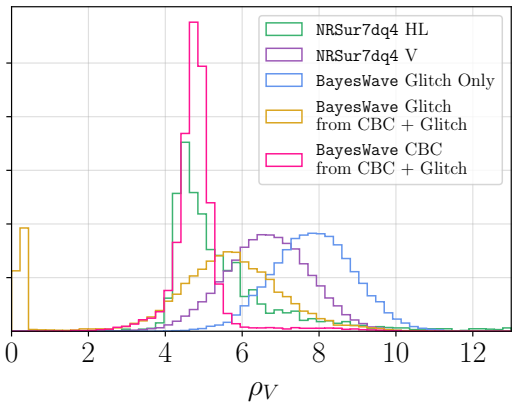

Figure 2.9: Comparison of optimal SNR estimates for Virgo from different analyses. In green is the posterior for the expected SNR in Virgo from just the LIGO data using the `NRSur7dq4` waveform (HL analysis of Fig. 2.1), while purple corresponds to the SNR from an analysis of the Virgo data only (V analysis of Fig. 2.1). The CBC and glitch SNR posterior from `BayesWave`'s full "CBC+glitch" model (Fig. 2.8) are shown in pink and orange respectively. Part of the latter is consistent with zero, which corresponds to no glitch (as also seen from the 90% credible interval in Fig. 2.8). The SNR posterior from a "glitchOnly" `BayesWave` is shown in blue.

The above studies suggest that the most likely scenario is that the Virgo trigger consists of a signal and a glitch. However, due to the low SNR of both, this interpretation is subject to sizeable statistical uncertainties and we therefore do not attempt to make glitch-subtracted Virgo data. Such data would be extremely dependent on which glitch reconstruction we chose to subtract, for example the median or a fair draw from the `BayesWave` glitch posterior. For these reasons and due to its low sensitivity, we do not include Virgo data in what follows.

2.4 Data quality issues: LIGO Livingston

The data quality issues in LIGO Livingston were identified and mitigated in GWTC-3 [77] through use of information from auxiliary channels [119, 38] and the `gwsubtract` pipeline as also described in App. 2.6. The comparison of Figs. 2.1 and 2.6, however, suggest that residual data quality issues might remain, as the two LIGO detectors result in inconsistent inferred $q - \chi_p$ parameters beyond what is expected from typical Gaussian noise fluctuations. Here we revisit the LIGO Livingston glitch with `BayesWave` and again model both the CBC and potential glitches. This analysis offers a point of comparison to `gwsubtract` as it uses solely strain data to infer the glitch instead of auxiliary channels. Additionally, `BayesWave` computes a posterior for the glitch, rather than a single point estimate, and thus al-

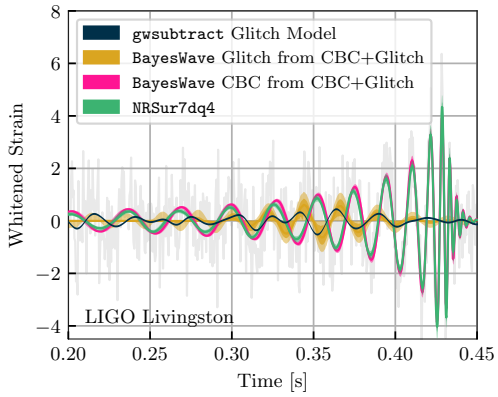

Figure 2.10: Whitened time-domain reconstruction of the data in LIGO Livingston obtained after analysis of data from the two LIGO detectors. Shaded regions correspond to 90% and 50% (where applicable) credible intervals and gray gives the original data without any glitch mitigation. Green corresponds to the same 2-detector result obtained with `NRSur7dq4` as Fig. 2.4, while pink and gold correspond to the CBC and glitch part of the joint "CBC+glitch" analysis with `BayesWave`. The black line shows an estimate for the glitch obtained through auxiliary channels. All analyses use only LIGO data.

lows us to explore the statistical uncertainty of the glitch mitigation. In all analyses involving `BayesWave` we use the original LIGO Livingston data without any of the data mitigation described in App. 2.6.

Figure 2.10 shows `BayesWave`'s CBC and glitch reconstructions in LIGO Livingston compared to the one based on the `NRSur7dq4` (from glitch-mitigated data) and the glitch model computed with `gwsubstract`. All analyses use data from the two LIGO detectors only. Unsurprisingly, now, the CBC reconstructions based on `IMRPhenomD` and `NRSur7dq4` do not fully overlap around t=0.3 s, though they are consistent during the signal merger phase. This is expected from the fact that LIGO Livingston supports spin-precession as well as Fig. 2.4. However, this difference is *smaller* than the statistical uncertainty in the inferred glitch from `BayesWave` (yellow) and well as differences between the `BayesWave` and the `gwsubtract` glitch estimates. This suggests that even though the `BayesWave` glitch estimate might be affected by the lack of spin-precession in its CBC model, this effect is smaller than the glitch uncertainty.

We also model the signal as a superposition of coherent wavelets in addition to the incoherent glitch wavelets using `BayesWave` [115, 116, 117]. This approach has been previously utilized for glitch subtraction [77]. However, we do not recover

strong evidence for a glitch overlapping the signal in LIGO Livingston when running with this "signal+glitch" analysis. The "signal+glitch" analysis attempts to describe both the signal and the glitch with wavelets and hence it is significantly less sensitive than the "CBC+glitch" model. In the data of interest, both the signal and the glitch whitened amplitudes are $\sim 1\sigma$ and as such they are difficult to separate using coherent and incoherent wavelets. Given that we know (based on the auxiliary channel data) that there is some non-Gaussian noise in LIGO Livingston, we find that the "signal+glitch" analysis is not sensitive enough for our data.

The large statistical uncertainty in the glitch reconstruction (yellow bands in Fig. 2.10) implies that the difference between the spin-precessing and non-precessing interpretation of GW200129 cannot be reliably resolved. To confirm this, we select three random samples from the glitch posterior of Fig. 2.10, subtract them from the unmitigated LIGO Livingston data, and repeat the parameter estimation analysis with `NRSur7dq4`. The `BayesWave` glitch-subtracted frames and associated `NRSur7dq4` parameter estimation results are available in [135]. For reference, we also analyze the original unmitigated data (no glitch subtraction whatsoever). Figure 2.11 confirms that the spin-precession evidence depends sensitively on the glitch subtraction. The original unmitigated data and the `gwsubtract` subtraction yield the largest evidence for spin-precession, but this is reduced -or completely eliminated- with different realizations of the `BayesWave` glitch model. In general, larger glitch amplitudes lead to less support for spin-precession, suggesting that the evidence for spin-precession is increased when the glitch is *undersubtracted*.

Figure 2.12 compares the corresponding $q - \chi_p$ posterior inferred from LIGO Hanford and LIGO Livingston separately under each different estimate for the glitch. Each of the 3 `BayesWave` glitch draws results in single-detector posteriors that fully overlap, thus resolving the inconsistency seen in $q - \chi_p$ when using the `gwsubtract` glitch estimate. Due to the lack of spin-precession modeling in the "CBC+glitch" analysis of Fig. 2.10, however, we cannot definitively conclude that any one of the new glitch-subtracted results is preferable. The 3 `BayeWave` glitch draws results in different levels of support for spin-precession. It is therefore possible that GW200129 is still consistent with a spin-precessing system. We do conclude, though, that the evidence for spin-precession is contingent upon the large statistical uncertainty of the glitch subtraction.

As a further check of whether the lack of spin-precession in `BayesWave`'s CBC model could severely bias a potential glitch recovery, we revisit the 10 simulated

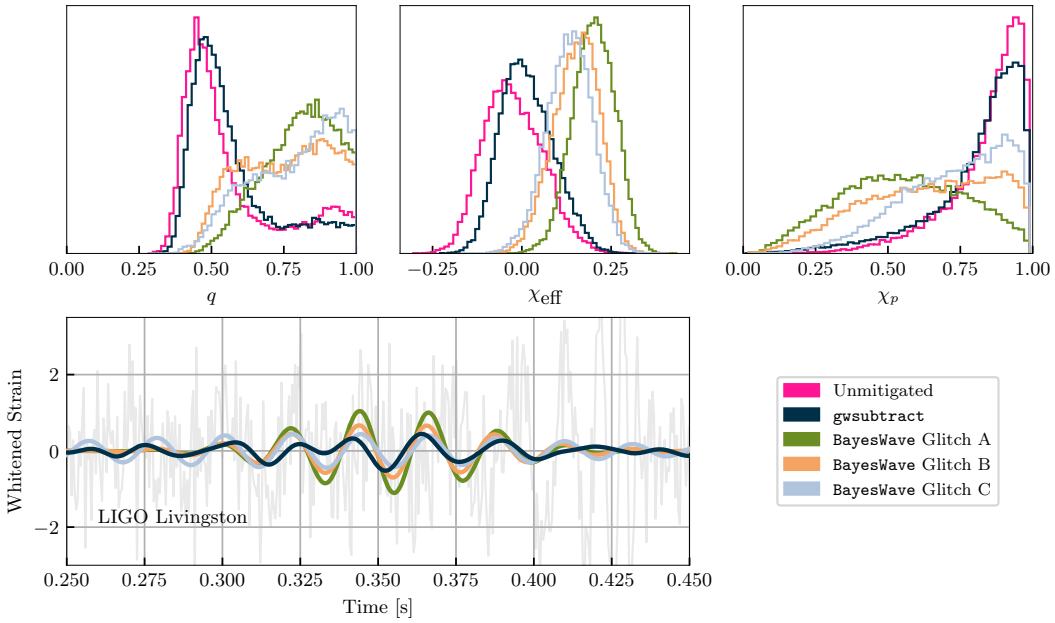

Figure 2.11: Bottom: Whitened, time domain reconstructions of various glitch reconstructions subtracted from LIGO Livingston data. The green line corresponds to the glitch reconstruction obtained from auxiliary data using `gwsubtract`. The rest are glitch posterior draws from the `BayesWave` "CBC+Glitch" analysis on HL unmitigated data. Top: Marginalized posterior distributions corresponding to parameter estimation performed with the `NRSur7dq4` waveform model on HL data where each respective glitch realization was subtracted from LIGO Livingston (same colors). Pink corresponds to the original data without any glitch subtraction. Larger glitch reconstruction amplitudes roughly lead to less informative χ_p posteriors and eliminate the $q - \chi_p$ inconsistency between LIGO Hanford and LIGO Livingston.

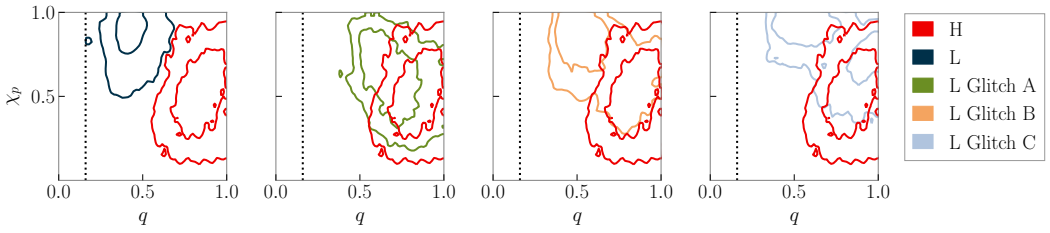

Figure 2.12: Two-dimensional posterior distributions for χ_p and q (50% and 90% contours) from single-detector parameter estimation runs. The far left panel shows the same tension as the LIGO Hanford and LIGO Livingston data plotted in Fig. 2.1 when using the `gwsubtract` estimate for the glitch. Subsequent figures show inferred posterior distributions using data where the same three different `BayesWave` glitch models as Fig. 2.11 have been subtracted. These results show less tension between the two posterior distributions.

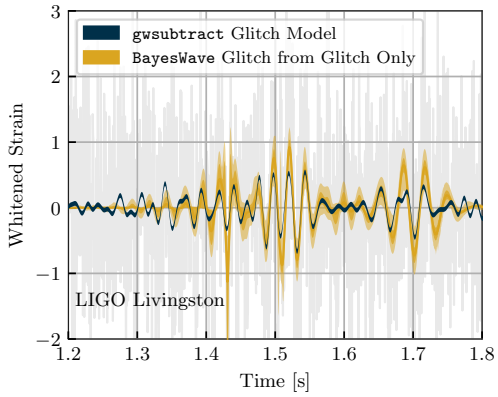

Figure 2.13: Comparison between the two glitch reconstruction and subtraction methods for a glitch in LIGO Livingston $\sim$ 1 s after GW200129, see the middle panel of Fig. 2.7. We plot the original data with no glitch mitigation (grey), the glitch reconstruction obtained from auxiliary channels with 90% confidence intervals (black), and the 50% and 90% credible intervals for the glitch obtained with `BayesWave` that uses only the strain data (gold).

signals from Fig. 2.6 and analyze them with the "CBC+glitch" model. These signals are consistent with GW200129 as inferred from LIGO Livingston data only, and thus exhibit the largest amount of spin-precession consistent with the signal. In all cases we find that the glitch part of the "CBC+glitch" model has median and 50% credible intervals that are consistent with zero at all times. This again confirms that the differences between the spin-precessing and the spin-aligned inferred signals in Fig. 2.10 is smaller than the uncertainty in the glitch. This test suggests that the glitch model is not strongly biased by the lack of spin-precession, however it does not preclude small biases (within the glitch statistical uncertainty); it is therefore necessary but not sufficient.

As a final point of comparison between `BayesWave`'s glitch reconstruction that is based on strain data and the `gwsubtract` glitch reconstruction based on auxiliary channels, we consider a *different* glitch in LIGO Livingston approximately 1s after the signal (see Fig. 2.7). Studying this glitch offers the advantage of direct comparison of the two glitch reconstruction methods without contamination from the CBC signal and uncertainties about its modeling. We analyze the original data with no previous glitch mitigation around that glitch using `BayesWave`'s glitch model and plot the results in Fig. 2.13. For the `gwsubtract` reconstruction we also include 90% confidence intervals, as described in App. 2.6.

The two estimates of the glitch are broadly similar but they do not always overlap

within their uncertainties. The main disagreement comes from the sharp data "spike" at $t = 1.43$ s that is missed by `gwsubtract`, but recovered by `BayesWave`. The reason is that the the maximum frequency considered by `gwsubtract` was 128 Hz and thus cannot capture such a sharp noise feature [38]. Away from the "spike," the two glitch estimates are approximately phase-coherent. On average `BayesWave` recovers a larger glitch amplitude as the `gwsubtract` result typically falls on `BayesWave`'s lower 90% credible level.

Figures 2.10 and 2.13 broadly suggest that `BayesWave` recovers a higher-amplitude glitch. Figure 2.11 shows that the evidence for spin-precession is indeed reduced, the LIGO Hanford-LIGO Livingston inconsistency is alleviated (Fig. 2.12), and the LIGO Livingston data become more consistent across low and high frequencies (Fig. 2.5) if the glitch was originally undersubtracted. However, due to the low SNR of the glitch and other systematic uncertainties it is not straightforward to select a "preferred" set of glitch-subtracted data. All studies, however, indicate that the statistical uncertainty of the glitch amplitude is larger than the difference between the inferred spin-precessing and spin-aligned signals.

2.5 Conclusions

Though it might be possible to infer the presence of spin-precession and large spins in heavy BBHs, our investigations suggest that in the case of GW200129 any such evidence is contaminated by data quality issues in the LIGO Livingston detector. In agreement with [76] we find that the evidence for spin-precession originates exclusively from data from that detector. However, we go beyond this and also demonstrate the following.

1. The evidence for spin-precession in LIGO Livingston is localized in the 20–50 Hz band in comparison to the rest of the data, precisely where the glitch overlapped the signal. Excluding this frequency range from the analysis, we find that GW200129 is consistent with an equal-mass BBH with an uninformative χ_p posterior; it is thus similar to the majority of BBH detections [6, 7, 8]. However, the fact that there is no evidence for spin-precession if $f_{\rm low}(L) > 50$ Hz is not on its own cause for concern as it might be due to Gaussian noise fluctuations or the precise precessional dynamics of the system.

2. LIGO Hanford is not only uninformative about spin-precession (which again could be due to Gaussian noise fluctuations or the lower signal SNR in that

detector), but it also yields an *inconsistent* $q-\chi_p$ posterior compared to LIGO Livingston. Using simulated signals, we find that the latter, i.e., the $q-\chi_p$ inconsistency, is larger than $O(95\%)$ of results expected from Gaussian noise fluctuations.

3. Given the LIGO Livingston glitch's low SNR, the statistical uncertainty in modeling it is *larger* than the difference between a spin-precessing and a non-precessing analysis for GW200129. Inferring the presence of spin-precession requires reliably resolving this difference, something challenging as we found by using different realizations of the glitch model from the `BayesWave` glitch posterior. Crucially, any evidence for spin-precession in GW200129 depends sensitively on the glitch model and priors employed.

4. Given the large statistical uncertainty in modeling the glitch, evidence for systematic differences between `BayesWave` and `gwsubtract` that use strain and auxiliary data respectively is tentative. However, the `BayesWave` estimate typically predicts a larger glitch amplitude, which would reduce the evidence for spin-precession and alleviate the tension between LIGO Hanford and LIGO Livingston. Additionally, we do not recover any support for a glitch when injecting spin-precessing signals from the LIGO Livingston-only posterior distribution into Gaussian noise. This indicates that `BayesWave` is unlikely to be strongly biasing the glitch recovery due to its lack of spin-precession.

Overall, given the uncertainty surrounding the LIGO Livingston glitch mitigation, we cannot conclude that the source of GW200129 was spin-precessing. We do not conclude the opposite either, however. Though we obtain tentative evidence that the glitch was undersubtracted, we can at present not estimate how much it was undersubtracted by due to large statistical and potential systematic uncertainties. It is possible that some evidence for spin-precession remains, albeit reduced given the glitch statistical uncertainty.

In addition, we verify that this uncertainty in the glitch modeling is larger than uncertainty induced by detector calibration. We repeat select analyses in Appendix 2.6 and confirm that the inclusion of uncertainty in the calibration of the gravitational-wave detectors negligibly impacts the spin-precession inference, as expected. Indeed, the glitch impacts the data at a level comparable to the signal strain, c.f., Fig. 2.10, whereas the calibration uncertainty within 20 to 70 Hz is only $\sim 5\%$ in amplitude

and 5° in phase [33]. Therefore, the glitch in LIGO Livingston's data dominates over uncertainties about the data calibration.

Though not critical to the discussion and evidence for spin-precession, we also identified data quality issues in Virgo. The inconsistency between Virgo and the LIGO detectors is in fact more severe than the one between the two LIGO detectors, however the Virgo data do not influence the overall signal interpretation due to the low signal SNR in Virgo. Nonetheless, we argue that the most likely explanation is that the Virgo data contain both the GW200129 signal and a glitch.

These conclusions are obtained with `NRSur7dq4`, which is expected to be the more reliable waveform model including spin-precession and higher-order modes in this region of the parameter space [100, 76]. We repeated select analyses with `IMRPhenomXPHM` which also favored a spin-precessing interpretation for GW200129 [77]. We found largely consistent but not identical results between `NRSur7dq4` and `IMRPhenomXPHM`, suggesting that there are additional systematic differences between the two waveform models. Appendix 2.7 shows some example results. Nonetheless, our results are directly comparable to the ones of [76, 99] as they were obtained with the same waveform model.

Our analysis suggests that extra caution is needed when attempting to infer the role of subdominant physical effects in the detected GW signals, for example spin-precession or eccentricity. Low-mass signals are dominated by a long inspiral phase that in principle allows for the detection of multiple spin-precession cycles or eccentricity-induced modulations. However, the majority of detected events, such as GW200129, have high masses and are dominated by the merger phase. The subtlety of the effect of interest and the lack of analytical understanding might make inference susceptible not only to waveform systematics, but also (as argued in this study) potential small data quality issues.

Indeed, Fig. 2.11 shows that a difference in the glitch amplitude of $< 0.5\sigma$ can make the difference between an uninformative χ_p posterior and one that strongly favors spin-precession. This also demonstrates that low-SNR glitches are capable of biasing inference of these subtle physical effects. Low-SNR departures from Gaussian noise have been commonly observed by statistical tests of the residual power present in the strain data after subtracting the best-fit waveform of events [13, 14, 15]. If indeed such low-SNR glitches are prevalent, they might be individually indistinguishable from Gaussian noise fluctuations. Potential ways to safeguard our analyses and conclusions against them are (i) the detector and frequency band

consistency checks performed here, (ii) extending the `BayesWave` "CBC+glitch" analysis to account for spin-precession and eccentricity while carefully accounting for the impact of glitch modeling and priors especially for low SNR glitches, (iii) and modeling insight on the morphology of subtle physical effects of interest such as spin-precession and eccentricity in relation to common detector glitch types.

2.6 Appendix: analysis details

In this appendix we provide details and settings for the analyses presented in the main text. All data are obtained via the GW Open Science Center [136]. Throughout we use geometric units, $G = c = 1$.

Detection and Glitch-subtracted data

GW200129 was identified in low latency [137] by GstLAL [138, 139], cWB [140], PyCBC Live [141, 142], MBTAOnline [143], and SPIIR [144]. The quoted false alarm rate of the signal in low latency was approximately 1 in 10^{23} years, making this an unambiguous detection. Below we recap the detection and glitch mitigation process from [77].

Multiple data quality issues were identified in the data surrounding GW200129. As a part of the rapid response procedures, scattered light noise [128, 131] was identified in the Virgo data, as seen in Fig. 2.7 in the frequency range 10–60 Hz. These glitches did not overlap the signal, and no mitigation steps were taken with the Virgo data. During offline investigations of the LIGO Livingston data quality, a malfunction of the 45 MHz electro-optic modulator system [145] was found to have caused numerous glitches in the days surrounding GW200129. To help search pipelines differentiate these types from glitches, a data quality flag was generated for this noise source [146]. These data quality vetoes are used by some pipelines to veto any candidates identified during the data quality flag time segments [147]. The glitches from the electro-optic modulator system directly overlapped GW200129, meaning that the time of the signal overlapped the time of the data quality flag.

Although clearly an astrophysical signal, the data quality issues present in LIGO Livingston introduced additional complexities into the estimation of the significance of this signal [77]. Due to the data quality veto, the signal was not identified in LIGO Livingston by the PyCBC [148, 149] MBTA [150], and cWB [140] pipelines. PyCBC was still able to identify GW200129 as a LIGO Hanford – Virgo detection, but the signal was not identified by MBTA due to the high SNR in LIGO Hanford and cWB due to post-production cuts. The GstLAL [151, 152] analysis did not

incorporate data quality vetoes in its O3 analyses and was therefore able to identify the signal in all three detectors.

The excess power from the glitch directly overlapping GW200129 in LIGO Livingston was subtracted before estimation of the signal's source properties [77, 38] using the `gwsubtract` algorithm [119]. This method relies on an auxiliary sensor at LIGO Livingston that also witnesses glitches present in the strain data. The transfer function between the sensor and the strain data channel is measured using a long stretch of data by calculating the inner product of the two time series with a high frequency resolution and then averaging the measured value at nearby frequencies to produce a transfer function with lower frequency resolution [153]. This transfer function is convolved with the auxiliary channel time series to estimate the contribution of this particular noise source to the strain data. Therefore, the effectiveness of this subtraction method is limited by the accuracy of the auxiliary sensor and the transfer function estimate. This tool was previously used for broadband noise subtraction with the O2 LIGO dataset [119], but this was the first time it was used for targeted glitch subtraction. Additional details about the use of `gwsubtract` for the GW200129 glitch subtraction can be found in Davis *et al.* [38].

The `gwsubtract` glitch model does not include a corresponding interval that accounts for all sources of statistical errors as is done by `BayesWave`. However, a confidence interval based on only uncertainties due to random correlations between the auxiliary channel and the strain data can be computed. For the GW200129 glitch model, this interval is ± 0.022 in the whitened strain data [38]. Additional systematic uncertainties due to time variation in the measured transfer function and effectiveness of the chosen auxiliary channel are expected to be present but are not quantified. The relative size of these uncertainties is dependent on the specific noise source that is being modeled and chosen auxiliary channel.

`Bilby` parameter estimation analyses

Quasicircular BBHs are characterized by 15 parameters, divided into 8 intrinsic and 7 extrinsic parameters. Each component BH has source frame mass m_i^s, $i \in \{1, 2\}$. In the main text we mainly use the corresponding detector frame (redshifted) masses $m_i = (1 + z)m_i^s$, where z is the redshift, as we are interested in investigating data quality issues and detector frame quantities better relate to the signal as observed. Each component BH also has dimensionless spin vector $\vec{\chi}_i$, and χ_i is the magnitude of this vector. We also use parameter combinations that are useful in various

Figure(s)	Waveform Model	Detector Network	Glitch mitigation	$f_{\rm low}$ (Hz)
2.1, 2.12	`NRSur7dq4`	H	`gwsubtract`	20
2.1, 2.12	`NRSur7dq4`	L	`gwsubtract`	20
2.1, 2.2, 2.3	`NRSur7dq4`	V	`gwsubtract`	20
2.1, 2.2, 2.3, 2.8	`NRSur7dq4`	HLV	`gwsubtract`	20
2.1, 2.2, 2.4, 2.10, 2.11, 2.14	`NRSur7dq4`	HL	`gwsubtract`	20
2.4	`NRSur7dq4` spin-aligned	HL	`gwsubtract`	20
2.5	`NRSur7dq4`	HL	`gwsubtract`	{20,30,40,50,60,70} in L, 20 in H
2.11	`NRSur7dq4`	HL	No mitigation	20
2.11	`NRSur7dq4`	HL	`BayesWave` fair draws	20
2.12	`NRSur7dq4`	L	`BayesWave` fair draws	20
2.14	`IMRPhenomXPHM`	HL	`gwsubtract`	20

Table 2.1: Table of `Bilby` runs and settings. All analyses use 4 s of data, and a sampling rate of 4096 Hz. Columns correspond to the main text figures each analysis appears in, the waveform model, the detector network used (H: LIGO Hanford, L: LIGO Livingston, V: Virgo), the type of glitch mitigation in LIGO Livingston, and the low frequency cutoff of the analysis. Figure 2.6 also presents results for a set of 10 injections drawn from the LIGO Livingston only posterior distribution with $f_{\rm low}(L) = 20\,$Hz. These analyses use the same settings as above with $f_{\rm low}(L) = 20\,$Hz.

contexts: total mass $M = m_1 + m_2$, mass ratio $q = m_2/m_1 < 1$, chirp mass $\mathcal{M} = (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5}$ [154, 155, 156], effective orbit-aligned spin parameter [157, 158, 159]

$$\chi_{\rm eff} = \frac{\vec{\chi}_1 \cdot \vec{L} + q\vec{\chi}_2 \cdot \vec{L}}{1+q}\,, \tag{2.2}$$

where $\vec{L}$ is the Newtonian orbital angular momentum, and effective precession spin parameter [107, 160]

$$\chi_p = \max\left(\chi_{1\perp}, q\chi_{2\perp}\frac{3q+4}{4q+3}\right)\,, \tag{2.3}$$

where $\chi_{1\perp}$ is the $\vec{\chi}_i$ component that is perpendicular to $\vec{L}$. The remaining parameters are observer dependent, and hence referred to as extrinsic. The right ascension α and declination δ designate the location of the source in the sky, while the luminosity distance to the source is d_L. The angle between total angular momentum and the observer's line of sight is θ_{jn}; for systems without perpendicular spins it reduces to the inclination ι, the angle between the orbital angular momentum and observer's line of sight. The time of coalescence t_c is the geocenter coalescence time of the binary. The phase of the signal ϕ is defined at a given reference frequency, and the polarization angle ψ completes the geometric description of the sources position and orientation relative to us; neither of these are used directly in this work.

Parameter estimation results are obtained with `parallel Bilby` [36, 95, 161] using the nested sampler, `Dynesty` [162]. The numerical relativity surrogate, `NRSur7dq4` [100], is used for all main results due to its accuracy over the regime of highly precessing signals. Its space of validity is limited by the availability of numerical simulations [163] to $q > 1/4$ and component spin magnitudes $\chi < 0.8$, though it maintains reasonable accuracy when extrapolated to $q > 1/6$ and $\chi < 1$ [100].

The majority of our analyses use the publicly released strain data, including the aforementioned glitch subtraction in LIGO Livingston [38], and noise power spectral densities (PSDs) [77]. The exception to the publicly released data was the construction of glitch-subtracted strain data using `BayesWave` for LIGO Livingston, as discussed in Sec. 2.4. We do not incorporate the impact of uncertainty about the detector calibration as the SNR of the signal is far below the anticipated regime where calibration uncertainty is non-negligible [164, 165, 166, 167]. Furthermore, we confirm that including marginalization of calibration uncertainty does not qualitatively change the recovered posterior distributions or our main conclusions by also directly repeating select runs.

Figure(s)	Models	Detector Network
2.8, 2.9	CBC+glitch	HLV
2.10, 2.11	CBC+glitch	HL
2.9	glitch	V
2.13	glitch	L

Table 2.2: Table of `BayesWave` runs and settings. All analyses use 4 s of data, a low frequency cut-off of $f_{\rm low} = 20$ Hz, a sampling rate of 2048 Hz, and the `IMRPhenomD` waveform when the CBC model is used. Furthermore, all analyses use the original strain data without the glitch mitigation described in Sec. 2.6. Columns correspond to the main text figures each analysis appears in, the `BayesWave` models that are used, and the detector network (H: LIGO Hanford, L: LIGO Livingston, V: Virgo). While not plotted in any figure, we also performed "CBC+Glitch" analyses on injections into the HL detector network as a glitch background study on GW200129-like sources; see Sec. 2.4.

As is done in GWTC-3 [77], we choose a prior that is uniform in detector frame component masses, while sampling in chirp mass and mass ratio. The mass ratio prior bounds are 1/6 and 1, where we utilize the extrapolation region of `NRSur7dq4`. Since `NRSur7dq4` is trained against numerical relativity simulations which typically have a short duration, only a limited number of cycles are captured before coalescence. With a reduced signal model duration, our analysis is restricted to heavier systems so that the model has content spanning the frequencies analyzed (20 Hz and above). We therefore enforce an additional constraint on the total detector-frame mass to be greater than $60\,M_\odot$. We verify that our posteriors reside comfortably above this lower bound. The luminosity distance prior is chosen to be uniform in comoving volume. The prior distribution on the sky location is isotropic with a uniform distribution on the polarization angle. Finally, for most analyses, the prior on the spin distributions is isotropic in orientation and uniform in spin magnitude up to $\chi = 0.99$. For the spin-aligned analyses, a prior is chosen on the aligned spin to mimic an isotropic and uniform spin magnitude prior. These settings and data are utilized in conjunction with differing GW detector network configurations and minimum frequencies in LIGO Livingston. The differences between runs and their corresponding figures are presented in Tab. 2.1.

`BayesWave` CBC and glitch analyses

`BayesWave` [115, 116, 117] is a flexible data analysis algorithm that models combinations of coherent generic signals, glitches, Gaussian noise, and most recently, CBC signals that appear in the data [39, 118, 168]. To sample from the multi-

dimensional posterior for all the different models, BayesWave uses a "Gibbs sampler" which cycles between sampling different models while holding the parameters of the non-sampling model(s) fixed.

For this analysis, we mainly use the CBC and glitch models (a setting we refer to as "CBC+Glitch"). The CBC model parameters (see App. 2.6) are sampled via a fixed-dimension Markov Chain Monte Carlo sampler (MCMC) using the priors described in Wijngaarden *et al.* [168]. The glitch model is based on sine-Gaussian wavelets and samples over both the parameters of each wavelet (central time, central frequency, quality factor, amplitude, phase [115]) and the number of wavelets via a trans-dimensional or Reverse-jump MCMC. In some cases, we also make use of solely the glitch model (termed "glitchOnly" analyses) that assumes no CBC signal and the excess power is described only with wavelets. The differences between runs and the figures in which they appear are presented in Tab. 2.2.

Though BayesWave typically marginalizes over uncertainty in the noise PSD [116], in this work we use the same fixed PSD as the Bilby runs for more direct comparisons. Additionally, we use identical data as App. 2.6 for the LIGO Hanford and Virgo detectors. However, when it comes to LIGO Livingston we use the original (i.e., "unmitigated," without any glitch subtraction) data in order to independently infer the glitch. We do not marginalize over uncertainty in the detector calibration.

2.7 Appendix: Select results with IMRPhenomXPHM

In this Appendix, we present select results obtained with the IMRPhenomXPHM [93] waveform model that also resulted in evidence for spin-precession in GWTC-3 [77]. Even though IMRPhenomXPHM and NRSur7dq4 both support spin-precesion, in contrast to SEOBNRv4PHM, there are still noticeable systematic differences between them. Figure 2.14 shows that while NRSur7dq4 and IMRPhenomXPHM generally have overlapping regions of posterior support, IMRPhenomXPHM shows slightly more preference for higher q and less support for extreme precession when compared to NRSur7dq4. Waveform systematics are expected to play a significant role in GW200129's inference (e.g. Refs. [77, 76, 169]), which motivates utilizing NRSur7dq4 for all of our main text results.

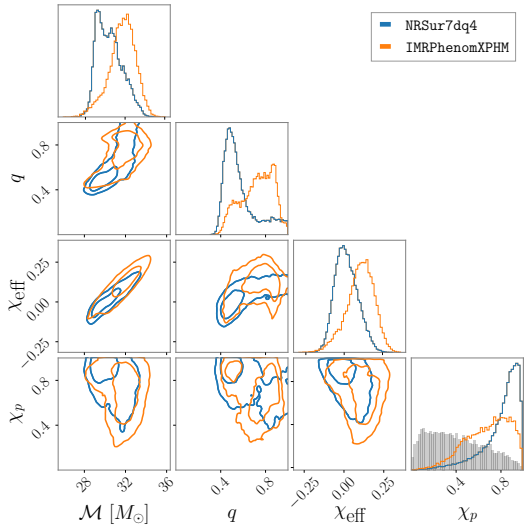

Figure 2.14: Similar to Fig. 2.1, using data from LIGO Livingston and LIGO Hanford. The comparison shows slight tension between results when using `NRSur7dq4` and `IMRPhenomXPHM`, though qualitatively `IMRPhenomXPHM` also seems to support the evidence for spin-precession.

Chapter 3

HOW TO DIAGNOSE A HIERARCHICAL MERGER ORIGIN VIA SPIN PARAMETERS

E. Payne, K. Kremer, and M. Zevin. "Spin Doctors: How to Diagnose a Hierarchical Merger Origin". In: *Astrophys. J. Lett.* 966.1 (2024), p. L16. DOI: 10.3847/2041-8213/ad3e82. arXiv: 2402.15066 [gr-qc].

E.P. conceived the project, carried out all the research, and led the writing of the manuscript.

3.1 Introduction

Following the first handful of observations of BBH mergers through their gravitational wave (GW) emission [2, 3, 4], many studies predicted that the dominant formation channel of BBHs would be determined after $O(10-100)$ observations [170, 171, 172, 173, 174, 175, 176, 177, 178]. However, despite the LVK detector network accumulating nearly 100 confident BBH observations [77], prominent formation pathways for BBH mergers remains an open question in GW astrophysics. The incongruity between prior expectation and reality can be attributed to a number of factors:

1. The diversity in the gravitational-wave events detected thus far does not show a strong preference for any one formation channel, with observations spanning a broad range of masses and mass ratios [e.g. 3, 4, 77, 179, 179].

2. Additional potential formation channels have been proposed in addition to the canonical "dynamical-versus-isolated" distinction [see e.g. 180, for a review], as well as subchannels to these canonical birth environments, which muddles the ability to pin down specific birth environments [181].

3. Uncertainties in massive-star evolution, binary physics, and formation environments are more vast than previously appreciated, translating to larger uncertainties in expected parameter distributions and generally making inference difficult [see e.g. 182, 183, for reviews].

4. Unlike black holes (BHs) in high-mass X-ray binaries in the Milky Way, which have been argued to have spin estimates that are near extremal [184, 185, 186],

the population of spins for GW-detected BHs are relatively small [7], making it difficult to distinguish between small, aligned spins expected from isolated evolution and moderate, in-plane spins expected from dynamical assembly.

In addition to spins, trends in the mass spectrum [e.g. 171, 170, 187, 8, 188, 189, 190], redshift evolution [e.g. 191, 192, 193], orbital eccentricity [e.g. 194], and correlations between BBH parameters [e.g. 195, 196, 197, 198, 199, 200, 201, 202, 203] have been investigated to elucidate the contribution of the various proposed BBH formation channels, although a robust conclusion is still far from being reached.

Although the holistic approach of examining features of the *full* BBH population holds promise for constraining formation scenarios [204], a complementary approach is the identification of *individual* merger events with distinguishing features uniquely associated with one or a subset of formation pathways. One example of this is eccentricity: BBH mergers with measurable eccentricity in the LVK sensitive frequency range ($\gtrsim 0.05$ at 10 Hz, [205, 206]) strongly point to a recent dynamical interaction, as orbital eccentricity quickly dissipates if a BBH system inspirals over a long timescale. Although no eccentric BBH mergers have been confidently detected to date (though see [207]), the detection of a small number of eccentric mergers (or non-detection of eccentric mergers) would place stringent constraints on the contribution of dynamical formation pathways [194].

Another possible smoking-gun signal of dynamical formation is the presence of hierarchical mergers—BBH mergers where one or both of the component BHs have gone through a previous merger event. Hierarchical mergers have masses that are typically larger than their "first-generation" progenitors that were born from massive stars as well as distinctive signatures in their spin magnitudes ($a \approx 0.7$, with a dispersion based on the mass ratio and component spins of the prior merger) and spin orientations (an isotropic distribution assuming a gas-free dynamical formation environment). Although hierarchical mergers are predicted to contain black holes with masses in the (pulsational) pair instability mass gap and studies have attempted to quantify the likelihood of particular GW systems being hierarchical merger [208, 209, 210, 211], uncertainties in the size and location of the gap [212, 8, 213], measurement uncertainties for high-mass black holes [214], and prior considerations [215, 216, 217] make mass alone difficult to pin down whether a particular system contains a black hole that was the result of a prior merger.

To identify the tell-tale signatures of hierarchical mergers, it is useful to consider

the leading-order (i.e., typically best-measured) spin terms from the post-Newtonian expansion of the GW waveform: the *effective spin* [218, 219]

$$\chi_{\rm eff} = \frac{a_1 \cos\theta_1 + q a_2 \cos\theta_2}{1+q}, \tag{3.1}$$

and *precessing spin* [220]

$$\chi_p = \max\left(a_1 \sin\theta_1, q\frac{3+4q}{4+3q} a_2 \sin\theta_2\right), \tag{3.2}$$

parameters where q is the mass ratio between the secondary and primary black holes, and a_1 and a_2 are the primary and secondary black holes' spins, respectively. The effective spin encodes a mass-weighted projection of the spin vectors on the orbital angular momentum axis, whereas χ_p depends on the projection of the spin vector on the plane of the orbit and is related to the strength of precession of the orbital angular momentum about the total angular momentum.

Hierarchical mergers are expected to have distinctive signatures in both of these spin parameters; due to generally large spin magnitudes (acquired during their first generation merger, [79, 80]) and isotropic spin orientations (a natural feature of dynamical formation in gas-poor environments, e.g. [81]), some hierarchical mergers should show evidence for negative $\chi_{\rm eff}$, and others for large χ_p. While a positive $\chi_{\rm eff}$ is possible, such systems may not be distinguishable from other formation channels whereas spin anti-alignment is difficult to form in the field [221]. Being a typically better-measured parameter [222, 114], studies have focused on negative $\chi_{\rm eff}$ as a potential sign for a hierarchical merger event [e.g., 223, 224, 225]. However, due to the inherent isotropic spin orientation distribution that is expected for hierarchical mergers in most astrophysical environments, many more systems will have large in-plane spins as opposed to large spins anti-aligned with the orbital angular momentum. For example, from cluster population simulations (see Sec. 3.2), $\sim 0.5\%$ ($\sim 20\%$) of hierarchical systems will have $\chi_{\rm eff} < -0.5$ ($\chi_{\rm eff} < -0.2$) whereas $\sim 67\%$ ($\sim 96\%$) of systems will have $\chi_p > 0.5$ ($\chi_p > 0.2$). So while $\chi_{\rm eff}$ is expected to be better measured, a significantly higher fraction of the hierarchical population will have the distinct signature of precession.

In this letter, we investigate the ability to measure each of these parameters for the purpose of identifying specific BBH mergers that are likely of a hierarchical origin. We take synthetic BBH mergers from realistic models of globular clusters, performing full parameter estimation on 6×10^3 events. Using these realistic measurement uncertainties, we quantify the fraction of hierarchical mergers that

confidently exhibit negative $\chi_{\rm eff}$ and large χ_p. Despite larger typical measurement uncertainties, we show that χ_p is a better indicator of hierarchical mergers than $\chi_{\rm eff}$—a consequence of the generic properties of hierarchically-formed BBHs.

The remainder of this letter is as follows. We outline the cluster population models used to construct the simulated set of first-generation (1G1G) and hierarchical BBH mergers in Sec. 3.2 before discussing how we quantify the measurements of the spin parameters in Sec. 3.3. The results of this calculation using the simulated population of BBH mergers as well as a selection of observed gravitational-wave signals are presented in Sec. 3.3. Finally, concluding remarks and implications of this study are presented in Sec. 3.4.

3.2 Cluster population models

We assemble our synthetic sample of dynamically-formed binary black hole mergers using the `CMC Cluster Catalog`, a suite of N-body cluster simulations spanning the parameter space of globular clusters observed in the local universe [82]. This catalog of models is computed using `CMC` [83], a Hénon-type Monte Carlo code that includes various physical processes specifically relevant to the dynamical formation of black hole binaries in dense star clusters including two-body relaxation, stellar and binary evolution [computed using `COSMIC`; 226], and direct integration of small-N resonant encounters including post-Newtonian effects [227]. In total, this catalog contains 148 independent simulations with variation in total cluster mass, initial virial radius, metallicity, and cluster truncation due to galactic tidal forces. The chosen values for these parameters reflect the observed properties of the Milky Way globular clusters [e.g., 228], but also serve as reasonable proxies for extragalactic clusters [e.g., 229] enabling a robust exploration of the formation of gravitational-wave sources in dense star clusters throughout the local universe.

To obtain a realistic astrophysically-weighted sample of binary black hole mergers, we follow Rodriguez and Loeb [191] and Zevin et al. [194]: each of the 148 simulations are placed into equally-spaced bins in cluster mass and logarithmically-spaced bins in metallicity. Each cluster model is then assigned a relative astrophysical weight corresponding to the number of clusters expected to form in its associated 2D mass-metallicity bin across cosmic time, assuming that initial cluster masses follow a $\propto M^{-2}$ distribution [e.g., 230] and that metallicities (as well as corresponding cluster formation times) follow the hierarchical assembly distributions of El-Badry et al. [231]. For all binary black hole mergers in a given model, the drawn cluster

formation time is then added to the black hole binary's merger time, yielding a realistic distribution of binary black hole merger events as a function of redshift. This scheme yields a predicted binary black hole merger rate of roughly $20\,Gpc^{-3}\,yr^{-1}$ in the local universe from dense star clusters.

We account for detectability of the simulated binary systems by generating colored Gaussian noise corresponding to a three-detector LIGO-Virgo gravitational-wave detector network at both design sensitivity [1, 16] and at the sensitivity the network achieved during the first half of LVK's third observing period [O3; 4]. We then add the simulated signals, randomly generating the binary's orientation and sky position, to the detector network noise and calculate the matched-filter signal-to-noise ratio [122]. Signals which pass the threshold signal-to-noise ratio (SNR) of ten are kept within the set of simulated detections.

In the `CMC` simulations, all black holes formed via stellar evolution are assumed to have negligible birth spin, a reasonable assumption if angular momentum transport in their massive-star progenitors is highly efficient [e.g., 232, 233]. However, spin can be imparted to cluster black holes through previous black hole merger events [79]. We assume all spin tilts are isotropically distributed. In addition to the non-spinning first-generation mergers, we consider two additional populations—the population of hierarchical BBHs formed consistently from these non-spinning first-generation systems, and first-generation mergers with black hole spins artificially included between [0, 0.2]. The latter population is included as a "worst-case" scenario for first-generation mergers that are not formed with small spins. While we do not self-consistently generate a fourth population corresponding to hierarchical mergers from this spinning first-generation population, modifications to the spin properties of first-generation BHs only marginally change the distribution of hierarchical merger parameters (cf. Figs. 4, 6, and 7 from [81]). The dominant impact of a spinning first-generation population is a significant reduction in the rate of hierarchical mergers, which does not affect our conclusions significantly regarding distinguishing the mergers within the hierarchical population but would affect their rates via the number of systems that are retained [81, 211, 234].

In Fig. 3.1 we show the spin parameters, $\chi_{\rm eff}$ and χ_p, of the O3-detected set of simulations from the low-spinning first-generation (purple), and hierarchical BBHs. The black lines indicate reasonable thresholds beyond which no 1G1G systems reside in the $\chi_{\rm eff}$-χ_p parameter space. While χ_p is typically less well-measured in gravitational-wave observations [222, 114], hierarchical systems overwhelming

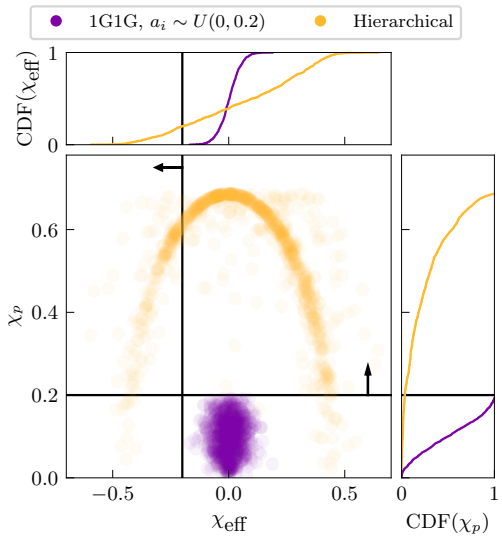

Figure 3.1: Two-dimensional distribution of spin parameters, $\chi_{\rm eff}$ and χ_p, for detectable low-spinning first-generation BBHs (1G1G; purple), and hierarchically formed BBHs (yellow). The one-dimensional marginal cumulative distribution functions (CDFs) are shown in the top and right panels. The spins of the low-spinning population are drawn uniformly and isotropically with spin-magnitudes from 0 to 0.2 in post-processing. All black hole masses are determined from the cluster simulations. We have selected for signals that are detectable by enforcing a signal-to-noise ratio threshold of 10 across the three detector LIGO-Virgo network at the LVK's sensitivity during their third observing period. The threshold of $\chi_{\rm thres} = 0.2$ used throughout the manuscript is indicated by the black lines for $\chi_{\rm eff}$ and χ_p. A significantly greater fraction of the hierarchical systems possess $\chi_p > 0.2$ than $\chi_{\rm eff} < -0.2$.

produce more moderate-to-high χ_p BBHs and occupy a unique region of the $\chi_{\rm eff}-\chi_p$ plane [235]. Therefore, in the following section, we explore the use of both $\chi_{\rm eff}$ and χ_p as potential "smoking-gun" signatures of a BBH's hierarchical origin.

3.3 Distinguishing hierarchical mergers

In this section, we turn our attention to how we might observationally identify the hierarchical mergers predicted from cluster populations using only the effective and precession spin parameters inferred from the observed GW signals. We first outline how we quantify the significance of the measurement before applying the calculation to the simulated cluster populations following Sec. 3.2 in addition to a number of gravitational-wave events from the LVK's third observing period which may present evidence of hierarchical origin based on their leading-order spin measurements.

Quantifying spin measurement significance

To understand the detectability of χ_p and $\chi_{\rm eff}$ in the simulated populations produced in Sec. 3.2, we infer the 15 binary parameters (assuming quasi-circular orbits) for each merger injected into the two gravitational-wave networks considered. We then calculate the posterior distributions on $\chi_{\rm eff}$ and χ_p directly from the inferred spin parameters, using Eqs. (7.29) and (3.2).

To quantify how significantly $\chi_{\rm eff}$ and χ_p are measured beyond the chosen thresholds, we utilize a "likelihood-ratio"-based statistic, denoted LR. This threshold boundary is somewhat arbitrary but can be motivated from the cluster simulations in Sec. 3.2. We compute LR by integrating over the marginal single-event likelihood and a uniform prior bounded between the threshold and the parameter boundaries (here denoted χ_L and χ_U for the lower and upper edges respectively). For example,

$$\mathrm{LR}^{\chi>\chi_{\rm thres}}_{\chi\leq\chi_{\rm thres}} = \frac{\int_{\chi_{\rm thres}}^{\chi_U} \mathcal{L}(d|\chi) U(\chi_{\rm thres}, \chi_U)\,\mathrm{d}\chi}{\int_{\chi_L}^{\chi_{\rm thres}} \mathcal{L}(d|\chi) U(\chi_L, \chi_{\rm thres})\,\mathrm{d}\chi} \tag{3.3}$$

computes the likelihood-ratio for support above the threshold, $\chi_{\rm thres}$, compared to below the threshold. Here, $\mathcal{L}(d|\chi)$ is the marginal likelihood for the observed event data, d given the spin parameter χ (either χ_p or $\chi_{\rm eff}$). We use the analytical expressions from [195] to construct the marginal likelihood (i.e. all prior dependence, $\pi(\chi|q)$, is removed). It is important to note, however, that in marginalizing over all other degrees of freedom we have made implicit choices for the prior distributions on other parameters, such as the individual black hole masses and redshift. We use uniform-in-detector-frame component mass priors when sampling in chirp mass and mass ratio [95, 195], and a Euclidean luminosity distance prior ($\propto d_L^2$). While these choices will inevitably have an impact on the inferred LR values, we are aiming to identify *unequivocally spinning* systems. Equation (3.3) can also be inverted to compute the likelihood-ratio for support below the threshold.

Upon close examination of Eq. (3.3), astute readers would note that it closely resembles a Bayes factor between two possible hypotheses (a spin parameter either above or below $\chi_{\rm thres}$)[1]. Therefore, we can interpret the inferred value in a similar way—the likelihood-ratio quantifies the amount of support above (below) the threshold against the support below (above) it. A common metric in the field of Bayesian statistics is that a $\ln \mathrm{LR}^{\chi>\chi_{\rm thres}}_{\chi\leq\chi_{\rm thres}} > 8$ quantifies significant evidence, corresponding to

[1]We have opted for the terminology "likelihood-ratio" here as we are removing the explicit and

a $\sim$ 3000:1 preference for $\chi > \chi_{\rm thres}$ [236]. Due to the nature of this calculation, there is statistical uncertainty due to a finite number of posterior distribution samples above $\chi_{\rm thres}$. The uncertainty in ln LR scales approximately, ignoring the impact of the removal of the prior, as $\sim \sqrt{\rm LR/\mathit{N}}$, where N in the posterior samples. Since we have $\sim 4 \times 10^4$ samples per event, this corresponds to an uncertainty of ~ 0.3 at ln LR = 8. This may slightly modify the exact percentage of systems passing the chosen ln LR = 8 threshold, though the broader conclusions of the Letter are unaffected.

There are, of course, many other parameters and methods to quantify this significance [237, 112, 238, 239]. Here we utilize this straightforward approach for two reasons. The first is that it is intuitive to interpret from the one-dimensional marginal distribution—*how much support is above or below a threshold?* And the second is that this statistic is more directly understood by the leading order terms in the gravitational-wave radiation due to both $\chi_{\rm eff}$ and χ_p, rather than being related first to the noise properties as in [237, 112]. Therefore, with a choice of spin threshold for the LR ($\chi_{\rm thres}$; motivated by Sec. 3.2) and under the assumption that all systems which pass $\chi_{\rm thres}$ are hierarchical mergers, we can use measurements of LR as a proxy for a definitive detection of a hierarchical merger. A $\chi_{\rm thres}$ value of 0.2 is motivated by confidently bounding observations above the expected small spins from [232] and [233]. We further choose more conservative thresholds ($\chi_{\rm thres} = 0.3, 0.4$) in the case where first-generation black holes might have some mechanism of being spun up (e.g. Ref. [240]). However, these systems still typically possess spins below 0.4 and are rare (e.g. see App. A.1.3 of Ref. [204]). Additionally, it is expected that the presence of hierarchical mergers formed from first generation BBH mergers with birth spins above 0.2 is heavily suppressed due to ejection of the merger remnant from the cluster environment [81]. Finally, the more conservative bound of $\chi_{\rm thres} = 0.4$ is consistent with the population observed thus far by the LVK [8] being consistent with only first generation black holes. This measure relies heavily on only the spins of the system, and so the statements in following sections are conservative. Information about the masses could be incorporated to boost the significance, though a threshold on masses will then need to be chosen as well, may

complex behavior of the posterior distribution with respect to the prior. If interested,

$$\mathrm{LR}^{\chi>\chi_{\rm thres}}_{\chi\leq\chi_{\rm thres}} = \frac{\int_{\chi_{\rm thres}}^{\chi_U} p(\chi|d)\,\mathrm{d}\chi}{\int_{\chi_L}^{\chi_{\rm thres}} p(\chi|d)\,\mathrm{d}\chi} \tag{3.4}$$

could be computed instead, where $p(\chi|d)$ is the marginal posterior distribution.

be less motivated given large uncertainties in the underlying first-generation mass distributions [182, 183], and will inadvertently remove lighter hierarchical systems from consideration.

Application to cluster population models

To explore how effectively hierarchical mergers can be selected out from a given population using spin parameters, we infer the properties of 1000 mergers from each of the three simulated populations (1G1G, 1G1G with uniform spin magnitudes in the range [0, 0.2], and hierarchical systems; as described in Sec. 3.2) in the current gravitational-wave detector network (from the third observing period; [4]) and at design sensitivity [1, 16]. We simulate these signals using the gravitational waveform model IMRPhenomXPHM [241], which we add into Gaussian noise colored by the respective noise power spectral densities. We arrive at 6×10^3 posterior distributions[2], using the nested sampling algorithm `dynesty` [162] embedded within the Bayesian inference library `Bilby` [36, 95], from which we calculate the LR following Eq. (3.3). From these results, we can then construct the complementary cumulative distribution function indicating the recovered fraction of observations that have a LR above a given value. The result of this calculation is shown in Fig. 3.2 for both χ_p (top) and $\chi_{\rm eff}$ (bottom). We find little difference in the inferred distribution of values of LR for 1G1G systems, independent of detector sensitivity and only slightly dependent on the choice of threshold and spin distribution. We therefore group all such possible distributions into the hatched purple region in Fig. 3.2. The fraction of hierarchical binaries for different thresholds are shown in black and grey for LVK's gravitational-wave detector network at O3 sensitivity and at design sensitivity, respectively. The complementary cumulative distribution function as a function of the LR represents the fraction of simulated observations above a LR value. Finally, we also include the relevant values from three gravitational-wave observations with ticks above the curves: GW190521 (purple; [214, 242, 243]), GW191109_010717 (pink), and GW200129_065458 (yellow; [77]).

From Fig. 3.2, we can identify the fraction of hierarchical binaries which pass a particular threshold of likelihood-ratio for both χ_p and $\chi_{\rm eff}$. Focusing on observations in the third LVK observing period (O3), $\sim 2\%$ of hierarchical mergers possess $\ln \mathrm{LR}^{\chi_p>0.2}_{\chi_p\leq 0.2} > 8$, indicating a confident detection. Signal-to-noise ratio has a mild

[2]Publicly available posterior samples are available at `https://doi.org/10.5281/zenodo.10558308`.

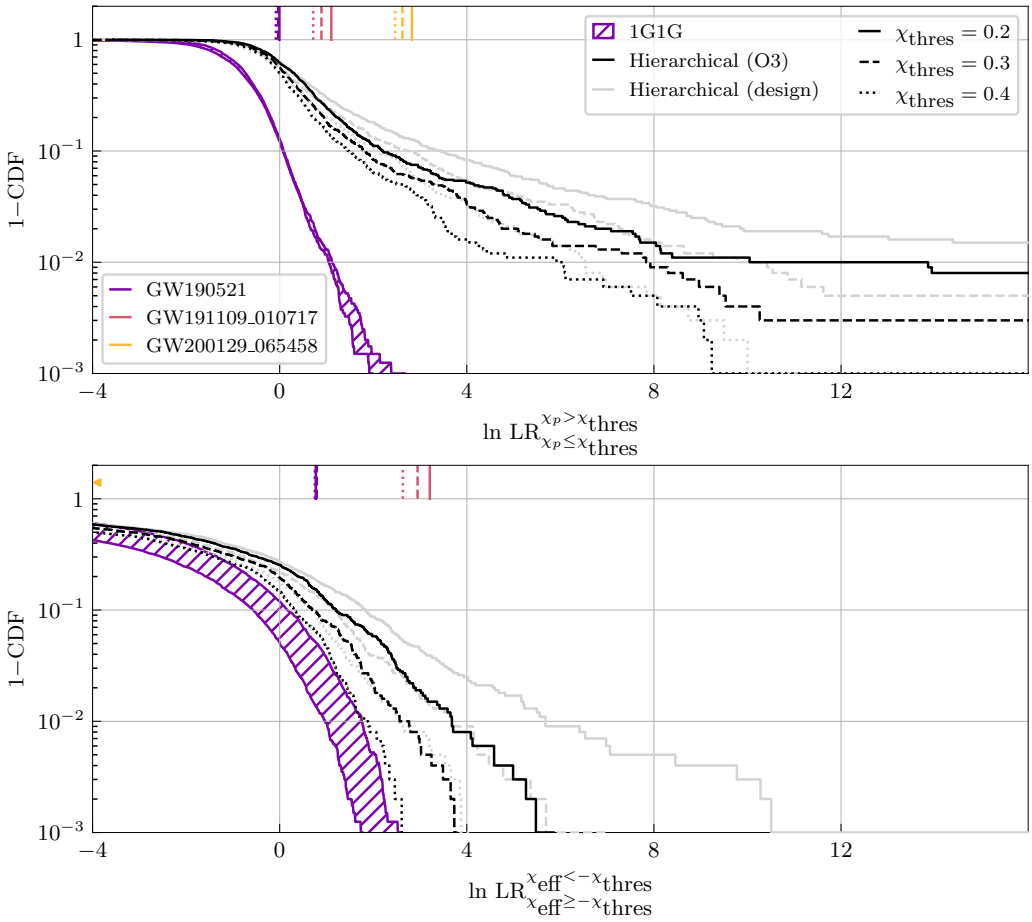

Figure 3.2: The complementary cumulative distribution function (1 – CDF) of detectable 1G1G (shaded; purple) and hierarchical BBH mergers (lines) as a function of the logarithmic likelihood ratio, ln LR, defined in Eq. (3.3). The three different linestyles correspond to different threshold choices ($\chi_{\rm thres} = 0.2, 0.3, 0.4$), and shadings correspond to simulated signals detected in the first half of the LVK's third observing period (O3) sensitivity (dark), or a three-detector LIGO-Virgo network at design sensitivity (light). The top and bottom panels correspond to the complementary cumulative distribution functions for χ_p and $\chi_{\rm eff}$, respectively. Finally, the observed values of ln LR at the different thresholds for three gravitational-wave observations made during O3—GW190521 (purple), GW191109_010717 (pink), and GW200129_065458 (yellow)—are marked. A significantly larger fraction of the hierarchical population possess a confidently measurable value of χ_p, whereas only the most relaxed threshold at design sensitivity can lead to a confident negative $\chi_{\rm eff}$ measurement in a single event.

impact on the systems with high LRs, with higher SNR systems somewhat more likely to have a higher LRs. For example, 15% of hierarchical systems have SNR > 20, whereas 63% of all hierarchical systems with $\ln \mathrm{LR}^{\chi_p > 0.2}_{\chi_p \leq 0.2} > 4$ possess an SNR > 20. We anticipate much of the support for higher values of χ_p in these systems is also a product of clear imprints of spin precession in the waveform from specific spin configurations. However, no choice of $\chi_{\rm thres}$ can provide a confident measurement for negative $\chi_{\rm eff}$ except with the most liberal threshold ($\chi_{\rm thres} = 0.2$) at design sensitivity of the three-detector LIGO-Virgo detector network. Therefore, from the simulated population of binary black hole mergers from globular clusters, $\chi_{\rm eff}$ is a wholly ineffectual parameter for distinguishing individual[3] hierarchical mergers[4]. Furthermore, if we instead treat the 1G1G population as a "null" background distribution from which to define a threshold (which is a very liberal threshold—requiring complete confidence in the population model), we still arrive at similar conclusions. With a detection threshold informed from the 1G1G LR distribution ($\ln \mathrm{LR}^{\chi > 0.2}_{\chi \leq 0.2} > 3$), we find $\sim 8\%$ of hierarchical mergers would be distinguishable via precession effects, while only $\sim 3\%$ would be distinguishable from $\chi_{\rm eff}$ measurements. While we believe it to be difficult to claim any one observation is of a hierarchical origin with $\ln \mathrm{LR} \sim 3$, an ensemble of such observations would indicate some number of these observations were hierarchical. This may lead to hints at the level of a population of hierarchical BBH mergers in the LVK's current fourth observing period—even if we are not confident in the origin of any one event.

Finally, we briefly turn our attention to a select few events from the LVK's third observing period (O3) that have been discussed in the literature as potential systems with anti- or mis-aligned spins: GW190521, GW191109_010717, and GW200129_065458 [214, 242, 4, 243, 77]. For simplicity and direct comparison to the simulated mergers, we use only posteriors constructed using IMRPhenomXPHM[5]. Using the LR calculation, no events surpass $\ln \mathrm{LR} > 8$ for either χ_p or $\chi_{\rm eff}$, although with a reduced threshold of $\ln \mathrm{LR} > 3$, GW200129_065458 and GW191109_010717 pass the thresholds for χ_p and $\chi_{\rm eff}$, respectively. However, since the impact of data quality

[3]This does not invalidate hierarchical studies where a population of potentially anti-aligned systems may be identified, as more information is extracted from a population of sources [e.g. 7, 224, 244].

[4]We also computed LR with the primary black-hole spin magnitude (a_1). We find that $\sim 4\%$ of hierarchical mergers possess $\ln \mathrm{LR}^{a_1 > 0.2}_{a_1 \leq 0.2} > 8$. While insightful, this does not factor in spin alignment and therefore such a measure may be contaminated by other channels.

[5]While GW190521 [4] and GW200129_065458 [77, 245] have results with waveform models more closely resembling numerical relativity (NRSur7dq4; [100]), using these samples for these two results only marginally affects these conclusions.

issues impacting the interpretation of these events is still an open question, caution should be taken when interpreting these results [see 38, 75, 246, 247].

3.4 Conclusions

Unequivocal detections of a hierarchical BBH merger via gravitational-wave observations will help understand the formation channels and histories of such systems. While studies often focus on identifying a hierarchical merger from anti-aligned spins [see e.g., 225, 224], we have focused on both the measurement of spin-precession in addition to anti-alignment in a simulated BBH merger population from realistic cluster models [83]. From this study, the key insights are as follows:

1. We have demonstrated that, in a realistic cluster population, *determining a system to be hierarchical will likely first come from the measurement of spin-precession* (cf. Fig. 3.2).

2. Additionally, from these simulated BBH mergers from 1G1G and hierarchical systems, we can approximately discern the number of gravitational-wave observations needed to uncover a hierarchical system in such a manner. We generally find that we should not yet have expected to confidently identify a hierarchical merger. Since $\sim$ 25% of the detectable BBHs from the cluster population are hierarchical, and $\sim$ 2% are confidently detectable at current sensitivity of the gravitational-wave network (from Fig. 3.2), there is only a 25% chance one or more hierarchical mergers would have been *detectable* in the LVK's third observing run [4, 77, 243]. This probability should be considered a generous upper limit, as it assumes dynamical formation in globular clusters as the only channel and environment.

3. Future observations appear much more fruitful. At design sensitivity $\sim$ 4% of hierarchical mergers become distinguishable. With an increased number of detections (ranging from ~200–1000; [248]), one can reasonably expect $\sim$ 2–10 identifiably hierarchical systems. Crucially, this analysis cannot be undertaken using anti-alignment of spins (i.e. $\chi_{\rm eff}$), as such effects will not be detectable, even in the most optimistic of circumstances.

As the ground-based gravitational-wave detector network evolves and approaches its design sensitivity, the tangible possibility of observing an unequivocally spinning,

hierarchical merger will become a reality. As we enter this era, the conclusions drawn here will be important in future discussions about the hierarchical origins of yet-to-be-detected BBH mergers. When discussing such a system, in this Letter we find it will be significantly more advantageous to investigate the spin-precession than spin misalignment. This motivates current and future research into both population modelling for hierarchical systems (and their first-generation progenitors) and waveform modeling to accurately capture this effect.

Chapter 4

FORTIFYING GRAVITATIONAL-WAVE TESTS OF GENERAL RELATIVITY AGAINST ASTROPHYSICAL ASSUMPTIONS

E. Payne, M. Isi, K. Chatziioannou, and W. M. Farr. "Fortifying gravitational-wave tests of general relativity against astrophysical assumptions". In: *Phys. Rev. D* 108.12 (2023), p. 124060. DOI: 10.1103/PhysRevD.108.124060. arXiv: 2309.04528 [gr-qc].

E.P. helped conceive the project, carried out all the analyses presented, and led the writing of the manuscript.

4.1 Motivation

Gravitational-wave observations from compact binary mergers have provided a unique laboratory to test Einstein's theory of gravity in the strong-field regime [249, 250, 251, 12, 13, 14, 15]. These individual detections by the Advanced LIGO [1] and Advanced Virgo [16] detectors allow for various tests—such as inspiral-merger-ringdown consistency [252, 253], parameterized inspiral deviations [69, 70, 254], gravitational-wave dispersion [255, 256], birefringence [257, 73] and nontensorial polarizations [258, 258, 259, 260, 261], among many more; see Ref. [15] for recent results—to both target specific properties of general relativity (GR) as well as broadly explore its consistency with observations. Beyond analyzing events individually, the ensemble of detections can be analyzed collectively to study the possibility of deviations from GR at the population level [262, 263, 14, 15]. Hierarchical population tests rely on inferring the distribution of deviation parameters across all events and confirming that it is consistent with a globally vanishing deviation [262, 264, 265].

In this study we explore the systematic impact of astrophysical population assumptions on these studies, show that they already come into play for current catalogs due to the increasing number of detections, and offer a solution under the framework of hierarchical population modeling.

In inferences about deviations from GR, there are strong likelihood-level correlations between the deviation parameters and the astrophysical parameters of the source,

such as the masses and spins of compact binaries [249, 266, 267]. Therefore, any inference of deviations from GR signals from black hole coalescences will be affected by assumptions about the distribution of binary black-hole masses and spins in the universe—otherwise known as the astrophysical population distribution [8]. This is true at both the individual-event and catalog levels, regardless of the specific assumptions made in combining deviation parameters across events, whether the analysis is hierarchical or not. Even when astrophysical parameters do not explicitly appear in the catalog-level test of GR, assumptions about these parameters are implicitly encoded in the individual-event deviation posteriors through the prior. As the catalog of gravitational-wave observations grows and the precision of the measurements improves, these systematic effects become more important.

In presence of correlations between deviation and astrophysical parameters, we must simultaneously model the astrophysical population distribution in conjunction with testing GR. By not explicitly doing so, as has been the case in previous tests of GR [249, 250, 251, 12, 13, 14, 262, 15], the astrophysical population is typically implicitly assumed to be uniform in detector-frame masses and uniform in spin magnitude. This fiducial sampling prior is adopted to ensure broad coverage of the sampled parameter-space, and not to represent a realistic astrophysical population. In reality, the primary-black hole mass population more closely follows a decreasing power-law with an excess of sources at $\sim 35\,M_{\odot}$, and preferentially supports low spins [7, 8]. This mismatch can lead to biased inference regarding deviations from GR. Simultaneously modeling the astrophysical and deviation distributions will not eliminate the influence of the former on the latter, but it will ensure that this interplay is informed by the data and not arbitrarily prescribed by analysis settings.

While this insight applies to all tests of GR, for concreteness we devote our attention to constraints on the mass of the graviton [255, 256] and deviations in parameterized post-Newtonian (PN) coefficients [66, 67, 68, 69, 70, 71]. A massive graviton would affect the propagation of a gravitational wave over cosmological distances; this leads to a frequency-dependent dephasing of the gravitational wave which is related to the mass of the graviton, m_g, and the propagated distance. The PN formalism describes the Fourier-domain phase of an inspiral signal under the stationary phase approximation through an expansion in the orbital velocity of the binary system; each $k/2$ PN expansion order can then be modified by a deviation parameter, $\delta\varphi_k$, which vanishes in GR. See App. 4.5 for further details about both calculations. We focus on these tests as they target the signal inspiral phase, which also primarily

informs astrophysical parameters such as masses and spins; we leave other tests [252, 253, 255, 256, 258, 258, 259, 260, 261, 13, 14, 15] to future work.

As motivation, Fig. 4.1 shows how inference on the 0PN coefficient of a real event (GW191216_213338) depends on astrophysical assumptions. This figure compares measurements with (blue) and without (red) a simultaneous measurement of the population of black hole masses and spins (see Sec. 4.2). The observed binary black-hole population shows a preference for systems with comparable masses; as a consequence of the strong correlation between the 0PN deviation coefficient and the mass ratio of GW191216_213338, this preference then "pulls" the system towards more equal masses and a more negative deviation coefficient. This is a direct manifestation of the fact that tests of GR are contingent on our astrophysical assumptions. Higher PN orders are expected to display similar correlations as in Fig. 4.1 with these and other parameters. For example, spins are known to be correlated with the coupling constant of dynamical Chern-Simons gravity which modifies the phase at the 2PN order [268, 269, 270, 65]. While we have constructed the posterior informed results here, it is more robust to simultaneously infer the astrophysical population while also testing GR. Fixing the prior to one astrophysical population realization or marginalizing over possible distributions from other analyses will not capture any correlated structure between the inferred deviation parameters and the astrophysical distributions. The above example serves only to illustrate the impact of the arbitrary choices previously made.

The remainder of the manuscript focuses on combining information from many observations to simultaneously infer the astrophysical population while testing GR; it is structured as follows. We first introduce our hierarchical analysis framework, as well as astrophysical and GR deviation models, in Sec. 4.2. We then demonstrate the impact of incorporating astrophysical information by constraining the graviton mass and inferring the PN deviation properties with an ensemble of gravitational-wave observations in Sec. 4.3. We analyze events from LIGO-Virgo-KAGRA (LVK)'s third observing run with individual-event results from Ref. [15] (the posterior samples are available in Ref. [275]) —a subset of the events in GWTC-3 [77]. The simultaneous modeling of the astrophysical population while testing GR tightens the graviton mass upper limit by 25%, and improves consistency with GR on the PN coefficients by $\sim 0.4\sigma$, when using a modified SEOBNRv4 waveform [271, 272, 273, 274, 254]. Finally, we conclude in Sec. 4.4, where we summarize the case for jointly modeling the astrophysical population when testing GR in order to

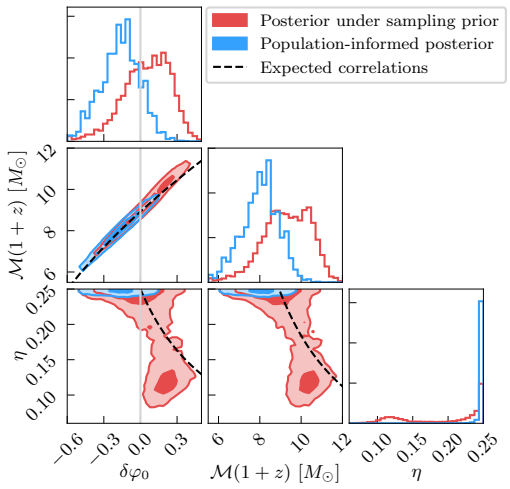

Figure 4.1: Posterior distributions for the 0PN deviation coefficient $\delta\varphi_0$, detector-frame chirp mass $\mathcal{M}(1+z)$, and symmetric mass ratio η for the gravitational-wave event GW191216_213338 [14, 4], as inferred by a modified SEOBNRv4 waveform [271, 272, 273, 274, 254]. Posteriors are conditioned on two different astrophysical assumptions: the broad prior used during parameter estimation (red), and the astrophysical population inferred by the data using the model in Sec. 4.2 (blue). The black dashed curves show the expected correlation (App. 4.6). Due to the correlations between astrophysical and deviation parameters, different astrophysical populations lead to different posteriors for $\delta\varphi_0$.

avoid biases and hidden assumptions, and comment on how the same is true for gravitational-wave studies of cosmology or nuclear matter.

4.2 Population Analyses

In this section, we introduce the fundamentals of inferring a population distribution from individual observations and discuss the population models we employ. We also outline the implementation and importance of observational selection effects in accounting for the events used within the analysis.

Preliminaries

We infer the astrophysical population distribution and deviations from GR (see Refs. [276, 74, 277] for a discussion of hierarchical inference in the context gravitational-wave astronomy). This framework has already been extensively applied to tests of GR and astrophysical population inference separately [262, 263,

14, 15, 6, 7, 8, 278, 279, 280, 281, 195, 224, 202, 174, 172, 282, 283, 284, 285, 286, 287, 288]. Here we focus on combining both methods to jointly infer the astrophysical population while testing GR.

Our approach is based on a *population likelihood*, $p(\{d\}|\Lambda)$, for the ensemble of observations, $\{d\}$, given population hyperparameters, $\Lambda = \{\Lambda_{\rm astro}, \Lambda_{\rm nGR}\}$. We separate the hyperparameters into the parameters describing the astrophysical population distribution, $\Lambda_{\rm astro}$, and parameters describing the deviation to GR, $\Lambda_{\rm nGR}$. The hyperparameters encode the shape of the population distribution, $\pi(\theta|\Lambda)$, where θ are parameters of a single event; we describe our population models in the following subsections. This hierarchical approach allows us to test GR while concurrently inferring the astrophysical population from the data. Given the likelihoods of individual events, $p(d_i|\theta_i)$, the population likelihood is

$$p(\{d\}|\Lambda) = \frac{1}{\xi(\Lambda)^N} \prod_{i=1}^{N} \int {\rm d}\theta_i \, p(d_i|\theta_i) \, \pi(\theta_i|\Lambda) \,, \tag{4.1}$$

where d_i and θ_i are respectively the data and parameters for the ith event, and $\{d\}$ is the collection of data for the ensemble of N observations[1]. We address the technical aspects of the likelihood calculation in App. 4.7.

In Eq. (4.1), $\xi(\Lambda)$ is the detectable fraction of observations given a set of population hyperparameters and accounts for selection biases [276]. It is defined as

$$\xi(\Lambda) = \int {\rm d}\theta \, p_{\rm det}(\theta) \, \pi(\theta|\Lambda) \,. \tag{4.2}$$

Here $p_{\rm det}(\theta)$ is the probability of detecting a binary black-hole system with parameters θ. The selection factor in Eq. (4.2) accounts for both the intrinsic selection bias of a gravitational-wave detector (e.g., heavier binaries are more detectable), as well as selection thresholds used when deciding which gravitational-wave events to analyze. The detected fraction can also be framed as a "normalizing factor", which relaxes the need for normalizable population distributions (so long as the integrals in Eqs. 4.1 and 4.2 are finite) [289]. This correction will become important in Sec. 4.2 when discussing the selection criteria for events to be included in the analysis.

In theory, the selection factor should account for the effect of both astrophysical and deviation parameters. However, we ignore the latter here, the effect of which

[1] Equation (4.1) assumes a prior on the rate of observations as $\pi(R) \propto 1/R$, which was analytically marginalized [284].

is subject of ongoing research [85]. For the former, we compute the detectable fraction, $\xi(\Lambda)$, from a set of recovered injections,

$$\xi(\Lambda) = \frac{1}{N_{\rm inj}} \sum_{i=1}^{N_{\rm rec}} \frac{\pi(\theta_i|\Lambda)}{\pi_{\rm draw}(\theta_i)}\,, \tag{4.3}$$

where $N_{\rm inj}$ is the number of injected signals, $N_{\rm rec}$ is the number of recovered signals, and $\pi_{\rm draw}(\theta_i)$ is the distribution from which the injected signals were drawn (for more details see Refs. [74, 276, 277, 6, 7, 8]). The subset of injected signals that are recovered is determined by the particular thresholds used to determine which gravitational-wave observations to use within the hierarchical analysis. To avoid biases, the criteria on the threshold for the detectable fraction calculation must match that of the observed signals. We address the specifics of the relevant criteria for our analysis in Sec. 4.2.

Finally, Eq. (4.1) explicitly shows the need for jointly modeling the astrophysical population when testing GR. While the astrophysical population may be separable from the deviation distribution so that $\pi(\theta|\Lambda) = \pi(\theta_{\rm astro}|\Lambda_{\rm astro})\,\pi(\theta_{\rm nGR}|\Lambda_{\rm nGR})$, this factorization cannot be undertaken for individual event likelihoods, as the deviations are often correlated with astrophysics (see Fig. 4.1), i.e. $p(\{d_i\}|\theta) \neq p(\{d\}|\theta_{\rm nGR})\,p(\{d\}|\theta_{\rm astro})$. Therefore, the integrals of Eq. (4.1) do not separate and tests of GR cannot be undertaken in isolation from the astrophysics.

From the hyperposterior distribution on the population parameters, we can construct the individual event population-informed posteriors following Refs. [290, 291, 292] (and references therein). Such distributions represent our best inference about the properties of a given event in the context of the entire catalog of observed signals. These calculations are subtle as they avoid "double-counting" the gravitational-wave events which also used to infer the population distribution.

Population models

In this subsection, we outline the population models for both the GR deviations and the astrophysical population. While many astrophysical population models have been proposed [6, 7, 8, 278, 279, 280, 281, 195, 224, 202, 174, 172, 282, 283, 284, 285, 286, 287, 288] as a product of the increasing number of observations [77, 8], in this work we restrict ourselves to standard parameterized models motivated by previous analyses.

GR deviation population models

There are two typical approaches to combining posteriors on GR deviation parameters obtained from different gravitational-wave observations, each stemming from different assumptions behind the deviations (see, e.g., discussions in [14, 15]). The first, more general approach is to assume that the population describing deviations from GR is, to the lowest order, a Gaussian distribution with a mean, μ, and standard deviation, σ [264, 262]. In the limit that all observations are consistent with GR, $(\mu, \sigma) \to (0, 0)$ and the inferred distribution approaches a Dirac delta function at the origin. Since a Gaussian distribution encapsulates the lowest order moments of more complicated distributions, given enough events any deviation from a delta function at the origin will be identified as a violation of GR, even if the exact shape of the deviation distribution is not captured by a Gaussian [262, 265]. This approach is now routinely applied to post-Newtonian deviations tests, inspiral-merger-ringdown consistency tests and ringdown analyses [262, 14, 15], but it can be naturally extended to any analysis that recovers GR in the limit of some vanishing parameter. This method provides a null test in cases where the exact nature of the deviation is unknown.

The second approach assumes all observations share the same value of the deviation parameter [13, 69, 70, 293, 294, 252, 253, 295, 274, 254]. This is the limiting case of the aforementioned Gaussian model when $\sigma \to 0$. This model (in the absence of astrophysical information) is equivalent to simply multiplying the marginal likelihoods of the deviation parameter obtained from the individual events. The assumption of a shared parameter is only suitable in the context of specific theories or models, in which case the expected degree of deviation for each event can be predicted exactly as a function system specific parameters (e.g., BH masses and spins) and universal, theory-specific parameters (e.g., coupling constants), the second of which can be measured jointly from a catalog of detections by multiplying likelihoods. In practice, the lack of complete waveform models beyond GR means that this approach has so far only been well-suited for measurements such as the mass of the graviton, and features of the propagation of gravitational waves whose observational signatures are independent of specific source properties by construction [13, 14, 15].

Astrophysical population models

Following Refs. [6, 7, 8], we model the primary black-hole mass (m_1) distribution as a power-law whose slope is given by an index α, with a sharp cut-off governed by the minimum mass, $m_{\rm min}$, and a higher-mass Gaussian peak,

$$\pi(m_1|\Lambda) = (1 - f_{\rm peak})\,\mathcal{P}[\alpha, m_{\rm min}](m_1) + f_{\rm peak}\,\mathcal{N}[\mu_{\rm peak}, \sigma^2_{\rm peak}](m_1)\,. \tag{4.4}$$

Here, $f_{\rm peak}$ is the fraction of binaries in the Gaussian peak, the powerlaw is given by

$$\mathcal{P}[\alpha, m_{\rm min}](m_1) \propto \begin{cases} m_1^{-\alpha}, & m_1 \geq m_{\rm min} \\ 0, & m_1 < m_{\rm min}\,, \end{cases} \tag{4.5}$$

and $\mathcal{N}[\mu, \sigma^2](x)$ is the probability density function for a Gaussian with mean μ and variance σ^2. We fix $m_{\rm min} = 5\,M_\odot$ for simplicity. Unlike other studies [280, 7, 8], we do not infer much structure in the Gaussian peak as higher mass features become unresolvable when looking at the light binary systems that provide constraints of PN coefficients (see Sec. 4.2).

We parameterize the distribution of mass ratios, $q \equiv m_2/m_1$, as a conditional power-law, with index β, and a sharp cut-off imposed by $m_{\rm min}$, such that

$$\pi(q|m_1;\Lambda) \propto \begin{cases} q^\beta, & 1 \geq q \geq m_{\rm min}/m_1 \\ 0, & q \leq m_{\rm min}/m_1\,. \end{cases} \tag{4.6}$$

Here β can take any value without leading to a singularity due to the lower bound on the mass ratio.

We adopt a truncated Gaussian population model for the component spins with a mean, μ_χ, and standard deviation, σ_χ, bounded between zero and one, assuming both spins are drawn independently from the same population distribution. This differs from standard Beta distribution utilized in many recent analyses [296, 281, 6, 7, 8]. as it allows for non-zero support at the edges of the spin-magnitude domain [297]. Furthermore, adopting a Gaussian model allows for efficient computation of the population likelihood via analytic integration (see App. 4.7). For individual-event analyses where the spins are assumed to be aligned with the orbital angular momentum (as is the case for posteriors using a modified SEOBNRv4 waveform [271, 272, 273, 274, 254]), this model treats the measured spin along the orbital angular momentum as the total spin magnitude.

For analyses where the individual event inferences also possess information about the spin-precession degrees of freedom, we adopt a model for the spin tilts, $\cos\theta_{1/2}$, whereby the population is parameterized as a mixture of isotropically distributed and preferentially aligned spins [281],

$$\pi(\cos\theta_1, \cos\theta_2|\Lambda) = \frac{f_{\rm iso}}{4} + (1 - f_{\rm iso})\times \\ \mathcal{N}[1, \sigma_\theta^2](\cos\theta_1)\,\mathcal{N}[1, \sigma_\theta^2](\cos\theta_2)\,, \tag{4.7}$$

where $f_{\rm iso}$ is the mixing fraction, and σ_θ is the standard deviation of the preferentially aligned Gaussian component. This model is only relevant for analyses with precessing spins. In this manuscript, this includes the massive graviton constraints (Sec. 4.3), and PN deviation tests with the IMRPhenomPv2 [68, 134, 71] waveform (App. 4.8).

Finally, we also adopt a power-law model for the merger-rate density as a function of redshift [284],

$$\pi(z|\Lambda) \propto \frac{1}{1+z}\frac{{\rm d}V_c}{{\rm d}z}(1+z)^\lambda\,, \tag{4.8}$$

where ${\rm d}V_c/{\rm d}z$ denotes the evolution of the comoving volume with redshift, and λ is the power-law index. When $\lambda = 0$, the binary black-hole population is uniformly distributed within the source-frame comoving volume.

Selection criteria and observations

We limit ourselves to binary black-hole observations made during LIGO-Virgo-KAGRA's third observing run [77] with false-alarm-rates of less than 10^{-3} per year[2]. This mirrors the selection criteria chosen for the tests of GR within Refs. [13, 14, 15], and therefore we do need not reanalyze any individual gravitational-wave observations [298, 275]. The events that pass these criteria are listed in Table IV of Ref. [14] and Table V of Ref. [15]. In future studies, the false-alarm-rate threshold could be raised to increase the number of included gravitational-wave events. This would likely improve inference of the astrophysical population and GR deviation constraints due to the larger catalog of observations. In our analyses, we exclude gw190814 [299] as it is an outlier from the binary black-hole population [7] and GW200115_042309 since it is a black hole-neutron star merger [300]. It

[2]For comparison, the population analyses presented Ref. [8] used a false alarm rate threshold of 1 per year. A more stringent false-alarm-rate threshold is often adopted when testing GR to avoid contaminating from false detections.

is straightforward to extend this analysis to additionally incorporate binary neutron star and neutron star-black hole mergers by adopting a mixture model of the different source classifications (see Ref. [8] for one example). We then use all events except GW200316_215756[3] when inferring the mass of the graviton, mirroring the analysis in Ref. [15]. When constraining the PN deviation coefficients, we include the additional criterion that signal-to-noise ratios (SNRs) during the binaries' inspiral must be greater than 6, again mirroring previous analyses [14, 15].

We use posteriors for the graviton's mass inferred using a modified IMRPhenomPv2 [68, 134, 71] waveform, whereas we use both modified SEOBNRv4 [271, 272, 273, 274, 254] (for results in Sec. 4.3) and modified IMRPhenomPv2 [68, 134, 71, 70, 69, 294, 295] (for results in App. 4.8)[4] waveform models when inferring the PN deviations. We summarize these events and their relevant properties in Tab. 4.1. We do not include gravitational-wave events from the first and second LIGO-Virgo observing runs, as a semi-analytic approximation was used to estimate the sensitivity of the detector network during that time [13]. This approximation does not compute a false-alarm rate and therefore cannot be unambiguously incorporated into this methodology.

As described in Sec. 4.2, selection effects are estimated through an injection campaign. While we know the total network SNR of the individual injections, part of our selection criteria is based on the inspiral network SNR. We approximate the inspiral SNR from the total SNR by constructing a linear fit to their ratio as a function of detector-frame total mass (Fig. 4.2). This fit is constructed by inferring the slope and offset of the line, as well as the uncertainty on the data points. We assume identical uncertainties on all SNR ratios, and marginalize over this parameter to fit the line. We validate this approximation by computing the detection probability $p_{\rm det}(\theta)$ with different draws of the linear fit. We find that different realizations of the approximation do not change the detection probability, and so we consider this approximation to be sufficiently accurate for our purposes. Future injection campaigns may also opt to compute the inspiral SNR directly.

4.3 Results

[3] GW200316_215756 was excluded from propagation tests within Ref. [15] due to poor sampling convergence.

[4] Single-event results with IMRPhenomPv2 were only produced during the first half of the third observing run [14, 15].

Table 4.1: Observations from the LIGO-Virgo-KAGRA's third observing run that pass our selection criteria [4, 77, 14, 15]. The different columns outline the gravitational-wave event, the detector-frame chirp mass, the total and inspiral *maximum a posteriori* SNRs ($\rho_{\rm tot}$ and $\rho_{\rm insp}$ respectively), and whether it was included in the graviton constraint calculation (m_g) or the post-Newtonian deviation tests (PN). Horizontal lines split events from the two halves of the third observing period. While we use all events marked under "PN" in Sec. 4.3, we are limited to the first half of observing run when using IMRPhenomPv2 posterior samples in App. 4.8.

Event	$(1+z)\mathcal{M}$ $[M_\odot]$	$\rho_{\rm tot}$	$\rho_{\rm insp}$	m_g	PN
GW190408_181802	$23.7^{+1.4}_{-1.7}$	15.0	8.3	✓	✓
LIGOScientific:2020stg	$30.1^{+4.7}_{-5.1}$	19.1	15.1	✓	✓
GW190421_213856	$46.6^{+6.6}_{-6.0}$	10.4	2.9	✓	-
GW190503_185404	$38.6^{+5.3}_{-6.0}$	13.7	4.3	✓	-
GW190512_180714	$18.6^{+0.9}_{-0.8}$	12.8	10.5	✓	✓
GW190513_205428	$29.5^{+5.6}_{-2.5}$	13.3	5.1	✓	-
GW190517_055101	$35.9^{+4.0}_{-3.4}$	11.1	3.4	✓	-
GW190519_153544	$65.1^{+7.7}_{-10.3}$	15.0	0.0	✓	-
GW190521_074359	$39.8^{+2.2}_{-3.0}$	25.4	9.7	✓	✓
GW190602_175927	$72.9^{+10.8}_{-13.7}$	13.1	0.0	✓	-
GW190630_185205	$29.4^{+1.6}_{-1.5}$	16.3	8.1	✓	✓
GW170706_222641	$75.1^{+11.0}_{-17.5}$	12.7	0.0	✓	-
GW190707_093326	$9.89^{+0.1}_{-0.09}$	13.4	12.2	✓	✓
GW190708_232457	$15.5^{+0.3}_{-0.2}$	13.7	11.1	✓	✓
GW190720_000836	$10.4^{+0.2}_{-0.1}$	10.5	9.2	✓	✓
GW170727_060333	$44.7^{+5.3}_{-5.7}$	12.3	2.0	✓	-
GW190728_064510	$10.1^{+0.09}_{-0.08}$	12.6	11.4	✓	✓

Table 3.1 continued on the following page...

Event	$(1+z)\mathcal{M}\ [M_\odot]$	$\rho_{\rm tot}$	$\rho_{\rm insp}$	m_g	PN
GW190828_063405	$34.5^{+2.9}_{-2.8}$	16.2	6.0	✓	✓
GW190828_065509	$17.4^{+0.6}_{-0.7}$	9.9	6.3	✓	✓
GW190910_112807	$43.9^{+4.6}_{-3.6}$	14.4	3.3	✓	-
GW190915_235702	$33.1^{+3.3}_{-3.9}$	13.1	3.7	✓	-
GW190924_021846	$6.44^{+0.04}_{-0.03}$	12.2	11.8	✓	✓
GW191129_134029	$8.49^{+0.06}_{-0.05}$	14.1	12.8	✓	✓
GW191204_171526	$9.70^{+0.05}_{-0.05}$	18.0	16.3	✓	✓
GW191215_223052	$24.9^{+1.5}_{-1.4}$	10.6	5.5	✓	-
GW191216_213338	$8.94^{+0.05}_{-0.05}$	17.9	15.6	✓	✓
GW191222_033537	$51.0^{+7.2}_{-6.5}$	13.1	3.1	✓	-
GW200129_065458	$32.1^{+1.8}_{-2.6}$	25.7	10.4	✓	✓
GW200202_154313	$8.15^{+0.05}_{-0.05}$	11.1	10.5	✓	✓
GW200208_130117	$38.8^{+5.2}_{-4.8}$	9.9	3.0	✓	-
GW200219_094415	$43.7^{+6.3}_{-6.2}$	11.2	2.8	✓	-
GW200224_222234	$40.9^{+3.5}_{-3.8}$	19.4	4.7	✓	-
GW200225_060421	$17.7^{+1.0}_{-2.0}$	12.9	6.8	✓	✓
GW200311_115853	$32.7^{+2.7}_{-2.8}$	17.5	6.5	✓	✓
GW200316_215756	$10.7^{+0.1}_{-0.1}$	11.5	10.7	-	✓

In this section we simultaneously infer the astrophysical population while testing GR and quantify the impact of fixing the population distribution to the sampling prior. Throughout, we use the nomenclature "fixed" and "inferred" to refer to whether the analysis uses the fixed sampling prior or infers the distribution from data, respectively. We implement the analyses using NumPyro [301, 302] and JAX [303], leveraging AstroPy [304, 305, 306] and SciPy [307] for additional calculations, and matplotlib [308], arViz [309] and corner [310] for plotting purposes. The code for the hierarchical tests is available in Ref. [311].

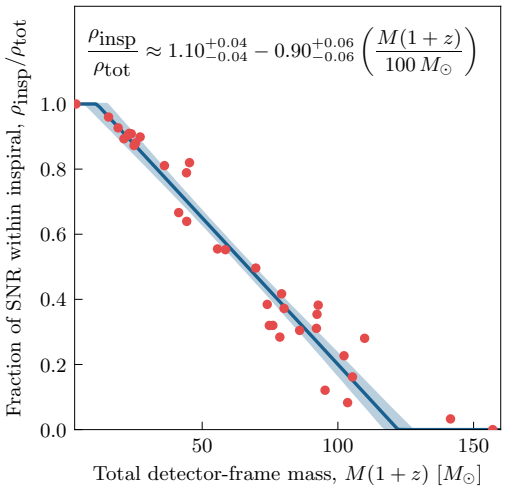

Figure 4.2: Ratio between the network *maximum a posteriori* gravitational-wave inspiral and the total SNRs as a function of detector-frame total mass, $M(1+z) \equiv (m_1+m_2)(1+z)$, for all gravitational-wave observations in the LIGO-Virgo-KAGRA third observing run [4, 77, 14, 15] with a false-alarm rate less than 10^{-3}/yr. The solid blue line is the median best-fit line to the observations, with the band representing the 90%-credible uncertainty. While computing this fit, we also estimate the uncertainty in the individual data points. We use this fit to compute the inspiral SNR for the injections used to estimate the detection probability, $p_{\rm det}(\theta)$, as described in Sec. 4.2.

Massive graviton constraints

We begin by demonstrating that astrophysical assumptions are crucial even in the simplest scenarios, where a global deviation parameter is shared across events. This is the case for the mass of the graviton, m_g [255, 256] (see App. 4.5), for which we produce an updated upper limit by simultaneously inferring the astrophysical distribution.

We combine results from individual-event likelihoods under the assumption of a shared deviation parameter as described in Sec. 4.2. In practice, we compute this as the limit of a vanishing standard deviation of the hierarchical analysis described in Sec. 4.2. For technical reasons, we assume a uniform prior distribution on $\log_{10}(m_g)$ when combining observations, which differs from Refs. [13, 14, 15] which applied a uniform prior prior on m_g itself; this is to avoid poor convergence when reweighting between individual-event posterior distributions. In the end, we reweight the shared graviton mass inference to a uniform prior to report upper limits on m_g. We compare this to results obtained assuming the sampling prior for the astrophysical parameters.

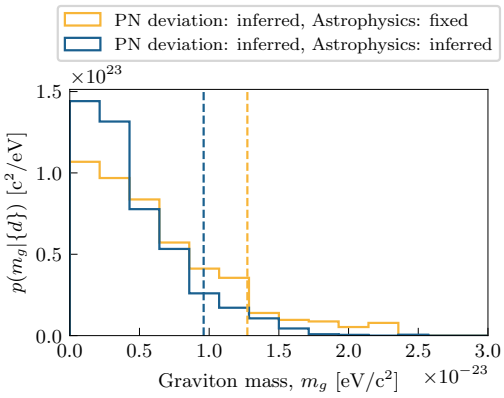

Figure 4.3: Marginal one-dimensional posterior distributions for the mass of a massive graviton. In practice, we compute the shared value of graviton mass by assuming a shared deviation parameter $\log_{10}(m_g c^2/\mathrm{eV})$ then reweighting to a uniform graviton mass prior. The dashed lines correspond to the 90% upper limits from the two analyses. We compare the result when astrophysical information is not included, equivalent to multiplying individual event likelihood functions (yellow), to also modeling the astrophysical population (dark blue). The result shifts towards smaller values of m_g if simultaneously modelling the astrophysical population and the graviton's mass.

The one-dimensional marginal distributions of the shared mass of the graviton are shown in Fig. 4.3. The inclusion of astrophysical information changes the inferred distributions of the graviton's mass increasing support for $m_g = 0$. When using the sampling prior for the astrophysical population (and thereby assuming the incorrect distribution), the graviton's mass is constrained to be $m_g \leq 1.3 \times 10^{-23}\,\mathrm{eV}/c^2$ at the 90% level[5]; however, upon inferring the astrophysical population the graviton's mass becomes more constrained, with $m_g \leq 9.6 \times 10^{-24}\,\mathrm{eV}/c^2$ at the 90% credible level. Under the expectation that GR is correct and $m_g = 0$, a reduced constraint is generically expected as we have included the correct information regarding the astrophysical population. This highlights the effect of unreasonable astrophysical assumptions, which are inconsistent with the observed population, on tests of GR.

Hierarchical post-Newtonian deviation constraints from SEOBNRv4

We repeat the population analysis, this time measuring the hierarchical PN deviation

[5]This constraint differs from the 90% upper limit of $1.27 \times 10^{-23}\,\mathrm{eV}/c^2$ calculated in Ref. [15], which is determined by additionally incorporating observations from the first and second LIGO-Virgo-KAGRA observing periods [3, 13]. We do not include these observations due to the ambiguity in the detector network sensitivity during these periods.

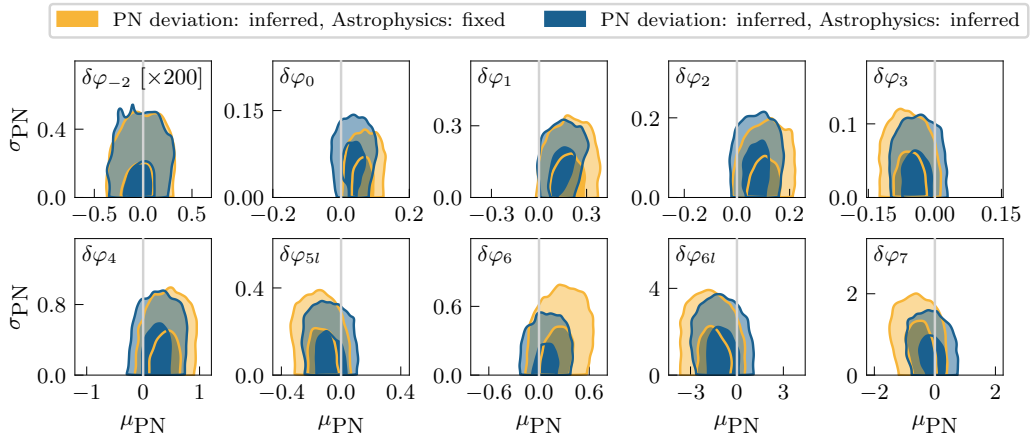

Figure 4.4: Two-dimensional marginal posterior distributions for the hyperparameters of the Gaussian PN deviation distribution informed by the 20 events in the third LIGO-Virgo-KAGRA observing run passing the selection criteria, analysed with a modified SEOBNRv4 [271, 272, 273, 274, 254] waveform. The contours indicate the 50% and 90% credible regions. Each panel corresponds to a separate analysis where the coefficient varied was at a different PN order. The analysis was undertaken with an implicitly assumed, astrophysically-unrealistic population (yellow), and a model which simultaneously infers the astrophysical population model (dark blue). Modelling both the astrophysical population and the PN deviation population systematically shifts the inferred mean, $\mu_{\rm PN}$, closer to zero.

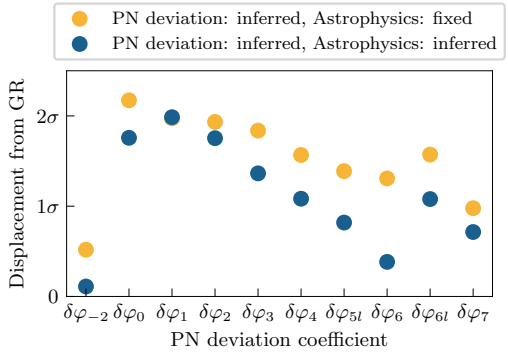

Figure 4.5: Displacement of the deviation parameter distribution from GR for each PN deviation coefficient. The displacement corresponds to the credible levels at which the hyperparameter values corresponding to GR, $(\mu_{\rm PN}, \sigma_{\rm PN}) = (0, 0)$, reside for two different models as shown in Fig. 4.4. This quantity is indicative of the relative position of the posterior to the GR value. Incorporating the astrophysical population as well as the hierarchical model for the PN deviation leads to an inferred result more consistent with GR for most cases.

distribution with a mean, $\mu_{\rm PN}$, and standard deviation, $\sigma_{\rm PN}$, for all PN orders. This is corresponds to ten separate analyses where only one PN deviation coefficient is allowed to vary. To compare with the default approach (which implicitly assumes a flat-in-detector-frame mass, uniform mass ratio, uniform spin-magnitude aligned spin, and comoving volume redshift distributions), we also fit the GR deviation in isolation under the assumption of the (astrophysically unrealistic) sampling prior [14, 15].

Figure 4.4 shows the two-dimensional posterior distribution of the deviation hyperparameters for -1 through to 3.5 PN orders. The standard results implicitly using the sampling prior are shown in yellow, while the results from the simultaneous modeling of the astrophysical and deviation populations are shown in dark blue. When concurrently modeling the astrophysical distribution, in all PN deviation parameters the inferred mean resides closer to zero, i.e., the expected value from GR, while there is no clear trend in $\sigma_{\rm PN}$. Overall, $(\mu_{\rm PN}, \sigma_{\rm PN}) = (0, 0)$ is retained with greater significance for almost all PN orders.

We quantify this improvement by comparing the two-dimensional credible level[6] at which the expected GR value, $(\mu_{\rm PN}, \sigma_{\rm PN}) = (0, 0)$, resides in Fig. 4.5. A lower value for the credible region implies that the value of hyperparameters expected from GR resides closer to the bulk of the distribution. In all but one PN order, jointly inferring the astrophysical and PN deviation distributions moves the inferred distribution to be more consistent with GR. For the 0.5PN deviation coefficient, $\delta\varphi_1$, there is little change in the credible level at which GR is recovered. Generally, inference of the astrophysical population allows our inferences of GR deviations to be more consistent with GR, with an average improvement of 0.4σ.

To shed further light on the interaction between the GR and astrophysics parameters, we focus on two specific deviation parameters. In particular, we draw attention to the 3PN coefficient (which shows the largest tightening of the supported hyperparameter space in Fig. 4.4) and the 0PN coefficient (where the PN deviation is most inconsistent with GR in Fig. 4.5).

[6]This "displacement" is the quantile, $Q_{\rm GR}$, reported in Refs. [14, 15] as (displacement)^2 = $-2\ln(1 - Q_{\rm GR})\,\sigma^2$. The quantile is computed by integrating over all regions of the hyperposterior distribution which are at a higher probability than $(\mu_{\rm PN}, \sigma_{\rm PN}) = (0, 0)$. We report values in terms of the standard deviation in two dimensions, 1σ and 2σ correspond to ~39.3% and ~86.5% credibility, respectively.

Example: 3PN deviation coefficient, $\delta\varphi_6$

To understand the origin of the improved measurement for $\delta\varphi_6$ when modeling astrophysics in Fig. 4.4, we show an expanded corner plot in Fig. 4.6 with an additional subset of the hyperparameter posterior distributions. The top left corner reproduces the corresponding panel in Fig. 4.4, wherein the yellow posterior distribution is obtained under the assumption of the astrophysical population given by the sampling priors, while the dark blue is obtained by simultaneously inferring the astrophysical-population and the GR deviation parameters.

Additionally, we use the same set of individual-event posterior samples to *separately* infer the astrophysical population independently of the PN deviation parameters, which amounts to assuming a uniform distribution of deviations across events (solid green). This differs from standard astrophysical population inference, which assumes that GR is correct *a priori* and thus starts from individual-event posteriors conditioned on $\delta\varphi = 0$ [6, 7, 8]. Finally, we also compute the astrophysical population under the assumption that GR is correct, $(\mu_{\rm PN}, \sigma_{\rm PN}) = (0, 0)$ (dashed green). The result assuming GR is correct is computed by fixing $(\mu_{\rm PN}, \sigma_{\rm PN}) \to (0, 0)$ to ensure equivalent samples are used between analyses, and is consistent with the usual population inference modulo model choices at the individual-event and population levels [6, 7, 8].

From the two-dimensional marginal distributions, the most apparent feature is that inferring the astrophysical population under the assumption of a broad uniform GR deviation population (shown in solid green) leads to inferences consistent with broad spin populations (large σ_{χ_z}) and populations favoring uneven mass ratios ($\beta < 0$). This can be straightforwardly explained by the presence of correlated structure between $\delta\varphi_6$, mass ratio, and the component spins at the individual-event level.

To demonstrate this, Fig. 4.7 shows four different posteriors for GW191216_213338 under different priors. The four distributions shown are the posterior obtained with the sampling priors (red), the one informed by the GR deviation population only analysis (yellow), the one informed by the astrophysical population only analysis (green), and the one informed by the jointly-inferred GR deviation and astrophysical populations (blue). The posteriors which involve information from inferred populations are computed following Ref. [290], and do not double-count the data from GW191216_213338, as discussed in Sec. 4.2.

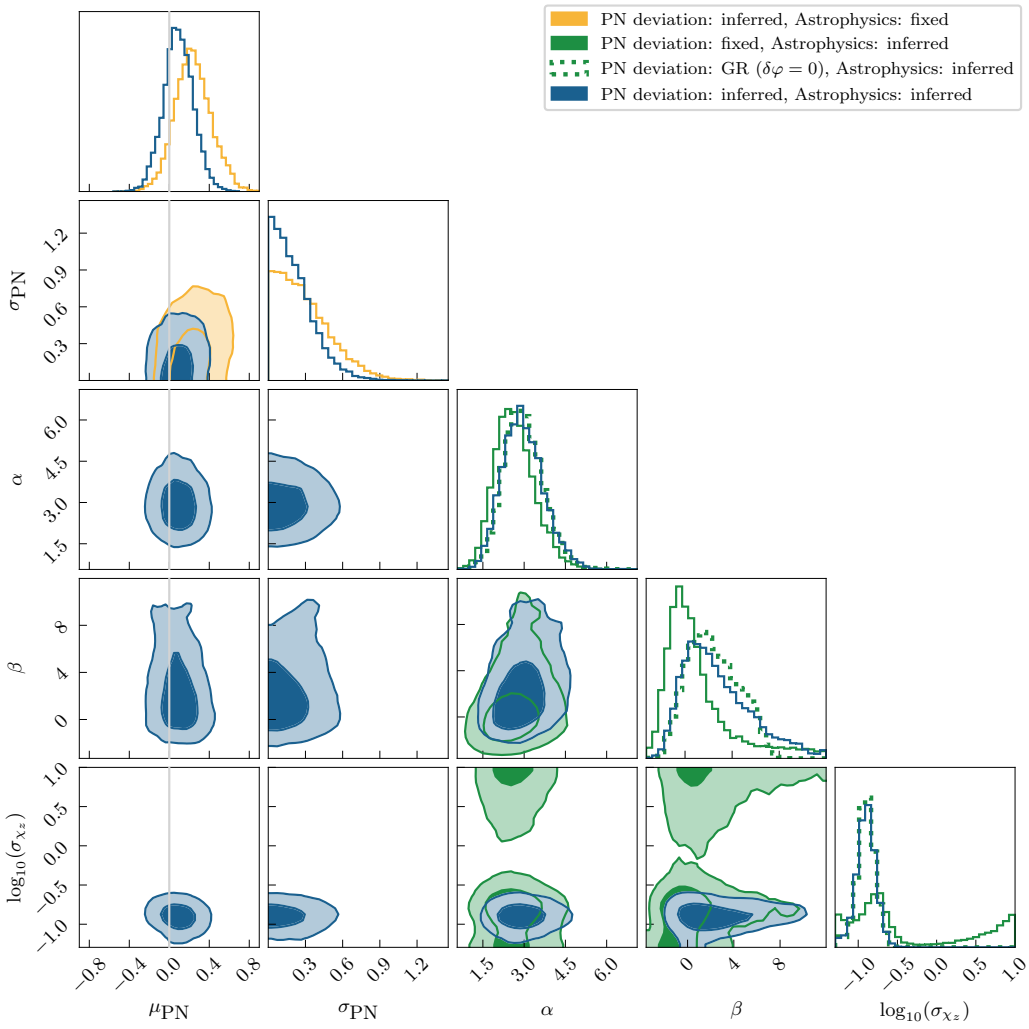

Figure 4.6: Marginal one- and two-dimensional posterior distributions for the $\delta\varphi_6$ PN deviation and a subset of astrophysical population hyperparameters. Contours correspond to the 50% and 90% credible regions. Results from four analyses are shown—population inference using the PN deviation population only with the "default" sampling prior astrophysical population (yellow), astrophysical population only (green), astrophysical population under the assumption that GR is correct (dashed green), and the joint analysis inferring the post-Newtonian deviation and astrophysical populations simultaneously (dark blue). No strong correlations exist between either the mean or standard deviation of the deviation Gaussian and astrophysical population parameters. The starkest difference is that inferring the population when the PN deviation population is ignored leads to broad spin magnitude populations.

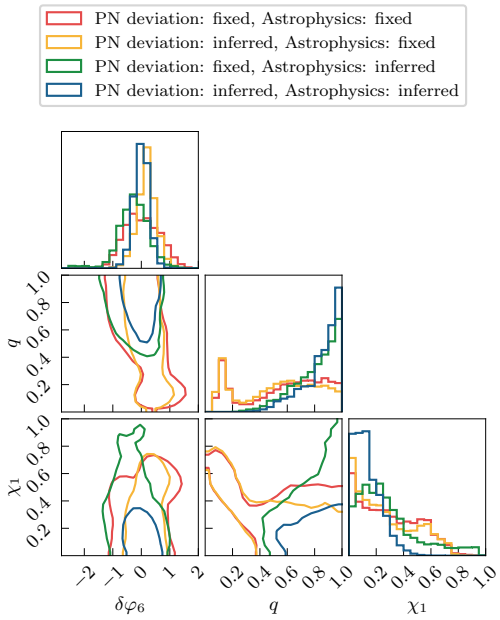

Figure 4.7: One- and two-dimensional posterior distributions for the 3PN deviation parameter, the mass ratio, and the primary black-hole spin for GW191216_213338 under four different assumptions: broad sampling priors (red), informed by the GR deviation population analysis (yellow), informed by the astrophysical population (green), informed by the joint inference of PN deviation and astrophysical populations (dark blue). Contours indicate the 90% credible region. Evidence for both a low mass ratio and larger primary spins is strongly contingent upon the astrophysical assumptions. Broad priors such as those used while sampling the posterior distribution have significant support for lower mass ratios. Inclusion of information from both the deviation population and the astrophysics leads to an inferred result with both low primary spin and high mass ratio.

Under the sampling astrophysical prior, posteriors exhibit a low-q, high-χ_1 mode. Since the inferred astrophysical population is inconsistent with low mass ratios and high spin magnitudes, the astrophysical-population-informed posteriors have reduced support for unequal masses (compare the red contour to the green one). Additionally incorporating the GR deviation information (blue), the population-informed posterior further reduces support for high-spinning systems. The similarity of the results under the sampling prior (red) with those in which only the GR deviation population is inferred (yellow) suggests that inferring small GR deviations is on its own not enough to significantly affect the inference of the astrophysical parameters in this case.

The tightening of the σ_{χ_z} hyperposterior distribution (i.e., inferring a more narrow spin population) when jointly inferring the GR deviation and astrophysical popu-

lations is precisely what we observe at the population level in Fig. 4.6 comparing the dark blue and green contours. Additionally, when enforcing that $\delta\varphi_6 = 0$ for all events (dashed green), we no longer recover support for broad spin populations. Interestingly, the astrophysical population inferred jointly with the GR deviation population is very similar to the result obtained when fixing $\delta\varphi_6 = 0$. This illustrates that, if we allow the model to infer that the scale of GR deviations is small, we will recover similar inferences overall as if we had fixed $\delta\varphi = 0$ *a priori*: we are learning *both* that spins are small *and* that any GR deviation must be small at this PN order. Conversely, an assumption of a broad GR deviation population leads to unrealistic astrophysical populations to account for the far-fetched astrophysical systems such analyses allow. We can also use this example to understand why inferring the deviation population in the absence of astrophysical modelling leads to a different deviation population with a larger inferred mean. Figure 4.7 shows that q and $\delta\varphi_6$ are correlated at the individual-event level, and therefore a broader q distribution will lend more support to the higher values of $\delta\varphi_6$. This correlation then systematically pulls the mean of the PN deviation distribution to higher values.

Example: 0PN deviation coefficient, $\delta\varphi_0$

We now turn to $\delta\varphi_0$, for which the standard analysis with a fixed astrophysical prior finds the least consistency with GR, at the 2.2σ credible level (yellow circle for $\delta\varphi_0$ in Fig. 4.5), driven by a displacement away from $\mu_{\rm PN} = 0$ (Fig. 4.4). Since this parameter is strongly correlated with the chirp mass and mass ratio (Fig. 4.1), we expect improvements when jointly modeling the astrophysical and deviation distributions; indeed that is the case, with GR recovered at the 1.6σ level (blue circle in Fig. 4.5). This analysis infers a $\sigma_{\rm PN}$ distribution that peaks slightly away from zero.

We can understand this behavior with Fig. 4.8, where we plot a subset of the two-dimensional marginal population posterior distributions in the same color scheme as Fig. 4.6. The structure of the PN deviation distribution is directly correlated with the mass ratio power-law index, β: steeper power-laws correspond to more variance in the GR deviation (larger β, larger $\sigma_{\rm PN}$). This is also manifested in the fact that when the PN deviation is assumed to be uniformly distributed (solid green), the astrophysical inference prefers steeper mass ratio power-laws (larger β), and that the analysis with deviations fixed to zero (dashed green) leads to a shallower

slope ($\beta \lesssim 6$). There is also a correlation between $\sigma_{\rm PN}$ and the width of the spin distribution, σ_{χ_z}, by which a narrower spin distribution demands for a greater spread in deviation parameters within the population.

Such correlations highlight precisely why we need to account for the astrophysical population when testing GR. By assuming a particular, fixed model for the astrophysical population, the hyperparameter correlations will not be captured in the marginal posterior for the GR deviation population. The analysis assuming the sampling prior for the astrophysical population (yellow), infers a value of $\sigma_{\rm PN}$ which peaks at zero. Among other hyperparameters, the sampling prior corresponds to a uniform ($\beta = 0$) mass-ratio distribution. Fixing the astrophysical population in such a way will lead to the hyperparameter posterior peaking at $\sigma_{\rm PN} = 0$, as seen in Fig. 4.8.

4.4 Conclusions

In this study, we have shown the importance of modeling the astrophysical population when testing GR with gravitational waves. Current tests do not explicitly model the astrophysical population, and therefore implicitly treat the prior used for sampling the posterior distribution as the assumed astrophysical population. Due to the presence of correlations between many GR deviations and astrophysical parameters, inappropriate astrophysical population choices will bias the test of GR. Like other sources of systematics, including waveform modeling [312, 313, 314, 315], the severity of this bias increases with the number of detections. We have shown that the effect of this bias is already being felt in the present catalog. This issue can only be fully addressed by simultaneously modelling both the astrophysical population in addition to the GR deviations.

We demonstrate the effect of inappropriate astrophysical models using constraints of the graviton's mass and tests of PN deviations as concrete examples. We show that jointly modeling the astrophysical population distribution while testing GR leads to results more consistent with GR. Furthermore, for some deviations at various PN orders there are correlations between hyperparameters governing the astrophysical and deviation populations. The impact of the astrophysical distribution is not just important for these parameters and these hierarchical models: any test of GR should accurately account for the astrophysical population. In fact, this problem is not unique to tests of GR; attempts to infer cosmological properties [316] or the

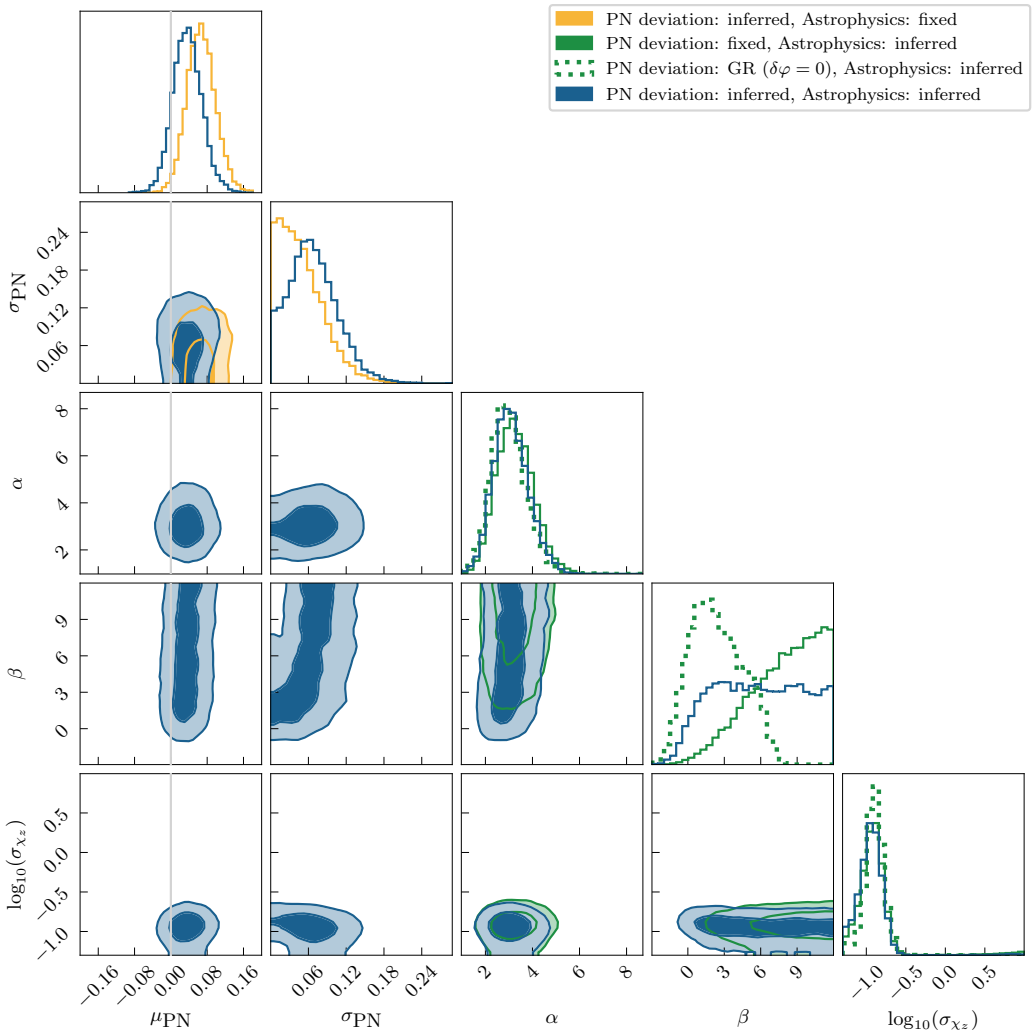

Figure 4.8: Similar to Fig. 4.6, one- and two-dimensional posterior distributions for the $\delta\varphi_0$ deviation and a subset of astrophysical population hyperparameters. A strong correlation is found between the width of the inferred post-Newtonian deviation population and the index of the mass ratio power-law when jointly inferring the deviation and astrophysical population models. There is also a less pronounced correlation between the deviation and spin population standard deviations. In the absence of modelling the astrophysical population, the inferred PN population is pulled to a higher mean with a reduced width.

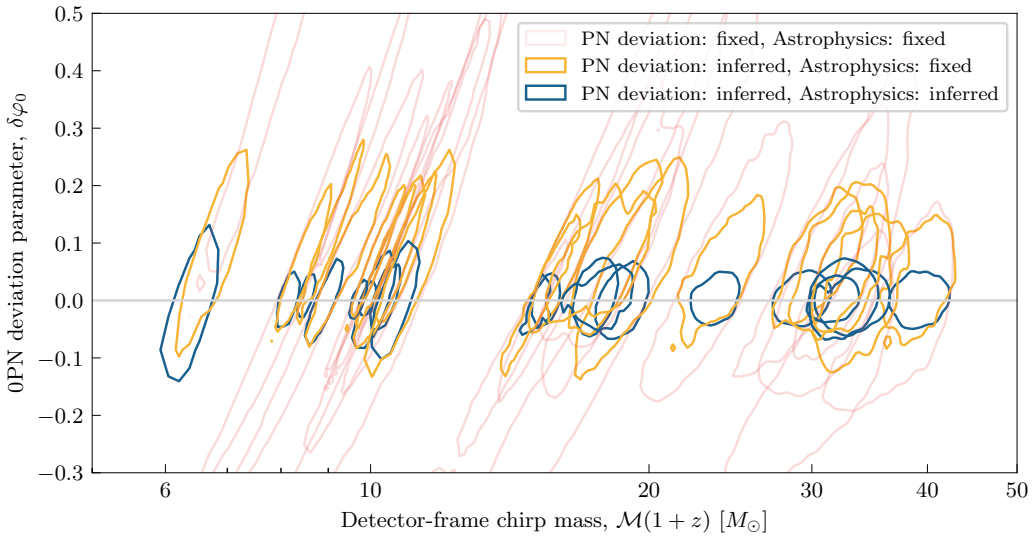

Figure 4.9: Marginal two-dimensional posterior distributions for the 0PN deviation coefficient and the detector-frame chirp mass for the events analyzed under the broad prior assumptions (light red), informed PN deviation population only (yellow), and informed by the jointly inferred deviation and astrophysical populations (dark blue). Contours indicate the 90% credible regions. This result demonstrates that as additional information is incorporated into the population distribution, more stringent constraints on the deviation parameters are placed on an individual event level. In the case demonstrated here, this pulls the inferred value towards $\delta\varphi = 0$ for all events.

equation of state of dense nuclear matter [317] are also impacted by these same considerations.

We can generically understand the impact of folding in the astrophysical population as follows. The standard sampling prior is chosen to broadly cover the parameter range of interest, and not to accurately represent the true astrophysical population. The actual population distribution will then typically provide support on a more narrow region of parameter space than the sampling prior. As a result, population-informed posteriors will not only avoid systematic biases but will also provide more stringent constraints on GR due to the additional information from the associated narrower population.

This posterior shrinkage is illustrated in Fig. 4.9, which shows the 0PN deviation parameter and detector frame chirp mass for the 20 events considered in our study (Table 4.1). The three sets of distributions correspond to the posteriors under different priors: fixed sampling priors (light red), fixed astrophysical prior and an inferred PN deviation population (yellow), and the case where both PN-deviation

and astrophysics distributions are inferred (blue). As more information about the GR deviation distribution is included, the inferred posterior of 0PN deviation parameter and the detector-frame chirp mass is more constrained. The posteriors are then constrained further still as additional information regarding the astrophysical population is included.

There are a number of directions in which to extend our work. The first would be to account for selection effects on the hyperparameters of the GR deviation distribution; this is to be addressed in upcoming work [85]. Additionally, here we have assumed a strongly parameterized model for the astrophysical population, with a power law and a Gaussian peak. This model is currently flexible enough given the number of events, with the primary mass Gaussian peak not impacting the inferred PN deviations with the selection of events considered. As the number of events used with these tests increases, and subtle features in the astrophysical population reveal themselves, we will likely need more flexible models [285, 286, 287, 288] to further avoid biases from misspecified population models [318, 319, 320]. Furthermore, in the case of PN coefficients, one would ideally constrain all orders simultaneously, in addition to the astrophysical parameters [249, 321, 322, 323, 324, 325].

Concurrently modeling the astrophysical population when testing GR is inevitable. Models that do not include a parameterized astrophysical population are implicitly assuming the sampling prior as the fixed population model. Such an assumption may induce systematic biases, cause false detections of GR violations, or incorrectly claim a stronger confirmation of GR than is warranted by the data. Moreover, even when accounting for the astrophysical population, correlations between GR deviation and astrophysical hyperparameters suggest that a true deviation could be absorbed into an unphysical inferred astrophysical population, a case that can only be noticed in studying the hyperposterior relating astrophysical to deviation parameters. Hierarchically modeling the astrophysical population while testing GR provides the solution to the implicit bias of assuming a fixed astrophysical population, and allows us to explore correlations between astrophysical parameters and deviations from GR, with fewer hidden assumptions.

4.5 Appendix: Formulation of parameterized tests of general relativity

In this appendix we outline the calculations required to constrain the graviton's mass and infer the PN deviation parameters.

Massive graviton measurements

The impact of a massive graviton on the propagation of gravitational waves has been studied in Refs. [255, 256] and references therein. A graviton with mass m_g modifies the dispersion relation of the gravitational wave. In a cosmological background, $g_{\mu\nu}$,

$$g_{\mu\nu}p^\mu p^\nu = -m_g^2, \tag{4.9}$$

where p^μ is the 4-momentum of the graviton. This leads to a dephasing of the gravitational wave, $\delta\Phi(f)$, that scales with the distance over which the signal propagates,

$$\delta\Phi(f) = -\frac{\pi(1+z)D_L^2 m_g^2 c^3}{D_0 h^2} f^{-1}\,, \tag{4.10}$$

where D_L is the luminosity distance, h is Planck's constant, and

$$D_0 = \frac{c(1+z)}{H_0}\int_0^z \mathrm{d}z' \,\frac{(1+z')^{-2}}{\sqrt{\Omega_m(1+z')^3+\Omega_\Lambda}}\,. \tag{4.11}$$

Here, $H_0 = 67.9\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ is the Hubble constant, and $\Omega_m = 0.3065$ and $\Omega_\Lambda = 0.6935$ are the matter and dark energy density parameters, respectively, adopting the values used in previous analyses [77, 15, 326].

Post-Newtonian deviation tests

Current parameterized PN tests are constructed by single-parameter modifications to the post-Newtonian description of the inspiral gravitational-wave phase in the frequency domain. This is given by [327, 68]

$$\Phi(f) = 2\pi f t_c - \phi_c - \frac{\pi}{4} + \frac{3}{128}\times \sum_{k=0}^{7}\frac{1}{\eta^{k/5}}\left(\varphi_k + \varphi_{k,l}\ln\tilde{f}\right)\tilde{f}^{(k-5)/3}\,. \tag{4.12}$$

Here, $\Phi(f)$ is the frequency-domain gravitational-wave phase under the stationary-phase approximation, $\tilde{f} = \pi G\mathcal{M}(1+z)f/c^3$, where $\mathcal{M}(1+z)$ is the redshifted chirp mass, $\mathcal{M} = (m_1m_2)^{3/5}/(m_1+m_2)^{1/5}$ is the source-frame chirp mass, $\eta = m_1m_2/M^2$ is the symmetric mass ratio, t_c and ϕ_c are the coalescence time and phase of the binary; finally, k indexes the $k/2$ PN order, and φ_k and $\varphi_{k,l}$ are the PN coefficients. The logarithmic coefficients, $\varphi_{k,l}$ only enter at 2.5 and 3.5 PN orders and otherwise

vanish [328, 329]. In GR, the coefficients are functions of the intrinsic parameters of the binary, their masses and spins. From this prescription, modifications to GR are incorporated by modifying [66, 69, 70]

$$\varphi_k \to (1 + \delta\varphi_k)\,\varphi_k\,, \tag{4.13}$$

except for k's for which $\varphi_k = 0$ in GR ($k = -2, 1$); in these cases, the modification is $\varphi_k \to \delta\varphi_k$, and $\delta\varphi_k$ is an absolute deviation [330].

In practice, modifications to IMRPhenomPv2 [68, 134, 71, 70, 69, 294, 295] and SEOBNRv4 [271, 272, 273, 274, 254] waveforms are computed differently, then the latter is transformed to the former. For the modified SEOBNRv4 waveform, the deviation is applied as above [254]. While, IMRPhenomPv2 is modified to only apply the deviation is onto the nonspinning portion of the PN coefficient [70, 69]. We translate all inferred deviation parameters to the IMRPhenomPv2 deviation parameter $\delta\varphi_k^{\rm IMR}$ for consistency,

$$\delta\varphi_k^{\rm IMR} = \delta\varphi_k \frac{\varphi_k}{\varphi_k^{\rm NS}}\,, \tag{4.14}$$

where $\varphi_k^{\rm NS}$ is the nonspinning value of the PN coefficient—calculated by setting the spins to zero for a particular set of compact binary masses. Additionally, care needs to be taken when translating to a uniform prior on $\delta\varphi_k^{\rm IMR}$, as the appropriate Jacobian,

$$\frac{{\rm d}\delta\varphi_k^{\rm IMR}}{{\rm d}\delta\varphi_k} = \frac{\varphi_k}{\varphi_k^{\rm NS}}\,, \tag{4.15}$$

is necessary. If the original prior is uniform on $\delta\varphi_k$, then the $\delta\varphi_k^{\rm IMR}$ must be weighted by the Jacobian to be effectively translated to another uniform prior.

4.6 Appendix: Computing expected parameter correlations

Correlations between GR deviation and astrophysical parameters can be analytically approximated by identifying regions of the parameter space that lead to a similar frequency evolution [266] and signal duration. The dominant correlation is the one between the detector-frame chirp mass, $\mathcal{M}(1+z)$, and the symmetric mass ratio, η. The duration of a gravitational-wave signal is related to the detector-frame chirp mass and some fiducial cut-off frequency [122],

$$T \propto \mathcal{M}^{5/3}(1+z)^{5/3} f_{\rm cut}^{-8/3}\,. \tag{4.16}$$

If we relate the final frequency to the innermost stable orbit or any cut-off which scales inversely with the binary's total mass, then $T \propto \eta^{-8/5}\mathcal{M}^{13/3}(1+z)^{13/3}$. A constant duration then implies

$$\mathcal{M}(1+z) \propto \eta^{-24/65} \,. \tag{4.17}$$

Here we have ignored both the contributions of a spin-induced "hang-up" effect [331] and GR deviations.

Correlations between astrophysical parameters and GR deviations can then be computed at lowest order [266] by enforcing that the second-order derivative of the phase evolution as a function of frequency be constant. As an example, for the correlation in Fig. 4.1, we compare the phase evolution when $\delta\varphi_0 = 0$ and when varying $\delta\varphi_0$ at the leading PN order, resulting in

$$\mathcal{M}_0^{-5/3}(1+z_0)^{-5/3} \sim (1+\delta\varphi)\mathcal{M}^{-5/3}(1+z)^{-5/3} \,. \tag{4.18}$$

Here $\mathcal{M}_0$ and z_0 are the values of the chirp mass and redshift when there is no deviation. We find the 0PN deviation coefficient to only be directly correlated with the detector frame chirp mass,

$$\delta\varphi_0 \sim \left(\frac{\mathcal{M}(1+z)}{\mathcal{M}_0(1+z_0)}\right)^{5/3} - 1 \,. \tag{4.19}$$

This calculation can be repeated for higher PN orders as well, however care needs to be taken as lower PN orders need to be retained when computing higher PN deviation coefficient correlations.

4.7 Appendix: Population likelihood approximation

In practice, we carry out single-event parameter estimation with a fiducial sampling prior, $\pi(\theta)$, before the hierarchical population analysis. We therefore do not possess representations of the individual event likelihoods, $p(d|\theta)$, but rather samples drawn from the fiducial posterior distribution $p(\theta|d) \propto p(d|\theta)\,\pi(\theta)$. Therefore, it is common to instead reformulate the integral within Eq. (4.1) as an average over samples drawn from each event's posterior distribution [276, 74, 277],

$$p(\{d\}|\Lambda) \propto \frac{1}{\xi(\Lambda)^N} \prod_{i=1}^{N} \frac{1}{M_i} \sum_{k=1}^{M_i} \frac{\pi(\theta_{i,k}|\Lambda)}{\pi(\theta_{i,k})} \,, \tag{4.20}$$

where M_i is the number of posterior samples for the ith event. It is possible for this Monte Carlo integration to not converge—particularly if the population distribution $\pi(\theta|\Lambda)$ is narrower than posterior distributions for individual events [289, 74, 283, 297, 332, 333]. This is particularly important in our scenario, since the inferred population of deviations from GR is typically narrower than marginal measurements from many individual events. This leads to a dearth of samples within the inferred GR deviation population, which subsequently leads to unreliable Monte Carlo integration in Eq. (4.20).

To address this issue, we use Gaussian kernel density estimates to represent the individual-event posteriors in a number of parameters, and simplify the calculation analytically by leveraging Gaussian population models. Dividing the parameters into the subset described by the Gaussian population distributions, $\theta^{\rm G}$, and the non-Gaussian distributions, $\theta^{\rm NG}$, we can analytically integrate over the former without resorting to Eq. (4.20). The Gaussian population parameters are the GR deviation parameter and the binary-hole spin magnitudes, whereas the black-hole primary mass and mass ratio, redshift, and spin tilts (for the analysis in App. 4.8) are included in the non-Gaussian set of parameters. For the kernel density estimation, we determine the corresponding covariance matrix for each individual event's distribution using Scott's rule [334],

$$\Sigma_{BW,i} \approx \frac{\Sigma_i}{n_{{\rm eff},i}^{2/(d+4)}}\,, \tag{4.21}$$

where Σ_i is the weighted covariance matrix of the parameters being estimated, d is the number of Gaussian dimensions, and $n_{\rm eff}$ is the effective number of samples [335, 336],

$$n_{{\rm eff},i} = \frac{\left(\sum_{k=1}^{M_i} w(\theta^G_{i,k})\right)^2}{\sum_{k=1}^{M_i} w(\theta^G_{i,k})^2}\,, \tag{4.22}$$

with the weights, $w(\theta^G_{i,k}) = 1/\pi(\theta^G_{i,k})$.

Since the integrand in the $\theta^{\rm G}$-space is a product of Gaussian distributions, the resulting integral is also a Gaussian [337]. This leads to the straightforward expression for the likelihood function

$$\begin{aligned} p(\{d\}|\Lambda) \propto & \frac{1}{\xi(\Lambda)^N} \prod_{i=1}^{N} \frac{1}{M_i} \sum_{k=1}^{M_i} \frac{\pi(\theta^{\rm NG}_{i,k}|\Lambda)}{\pi(\theta_{i,k})} \times \\ & \mathcal{N}[\mu(\Lambda), \Sigma_{BW} + \Sigma(\Lambda)](\theta^{\rm G}_{i,k})\,, \end{aligned} \tag{4.23}$$

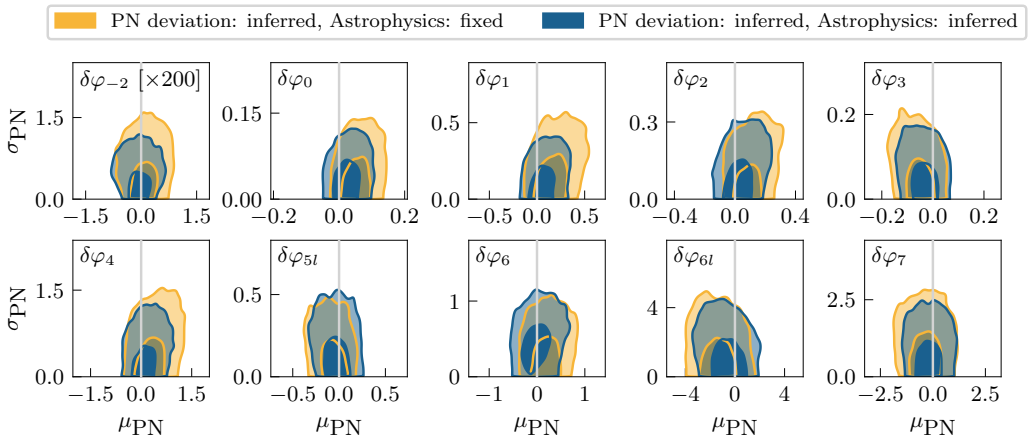

Figure 4.10: Same figure as Fig. 4.4 but using 12 events from the first half of the third LIGO-Virgo-KAGRA observing run, with individual event posterior distributions constructed with IMRPhenomPv2. We generally observe similar structure to the results with SEOBNRv4, although parameters are less constrained—likely due to fewer observations incorporated.

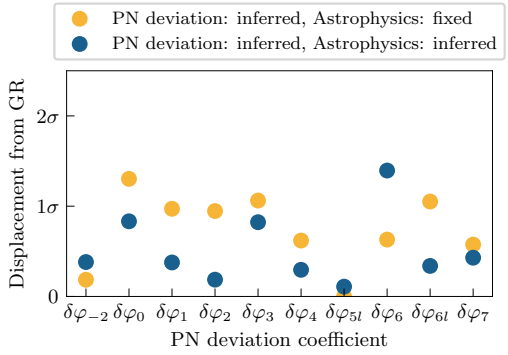

Figure 4.11: Same as Fig. 4.5, for the results from the IMRPhenomPv2 analysis. As seen throughout the manuscript, inclusion of the astrophysical population model in general leads to improved consistency with GR. Furthermore, the posterior distributions sit closer to GR for IMRPhenomPv2 than SEOBNRv4, likely as a result of analyzing fewer events.

where $\mu(\Lambda) = (\mu, \mu_\chi, \mu_\chi)$ and $\Sigma(\Lambda) = \mathrm{diag}(\sigma^2, \sigma_\chi^2, \sigma_\chi^2)$, though more complicated structure can be imposed on the population model. Since this integral is computed analytically, we empirically find improved convergence.

4.8 Appendix: Constraints from IMRPhenomPv2

While we have focused on results from SEOBNRv4 [271, 272, 273, 274, 254], these analyses do not include precessing degrees of freedom. However, evidence

for precession has been found at the population level within gravitational-wave observations [7, 8]. Therefore, to explore if there are any major changes when incorporating precession effects, we use the 12 events from the first half of the third observing run analysed with IMRPhenomPv2 [68, 134, 71, 70, 69, 294, 295] which meet our selection criteria [14]. There are no equivalent results from the second half of the third observing run [15]. We show the summary of the marginal two-dimensional posterior distribution for the Gaussian population hyperparameters with and without the inclusion of astrophysical information in Fig. 4.10. Generally, these results are less constrained due to the smaller number of events, though we still witness a similar shift in the means of the Gaussian populations as in Fig. 4.4. We also summarize the quantiles at which the expectation from GR presides in Fig. 4.11. Generally, the IMRPhenomPv2 results are more consistent with GR than the equivalent SEOBNRv4 results presented in Sec. 4.3. This could be a product of this waveform model incorporating precession, or simply that fewer events were analyzed, leading to a decrease in precision.

Chapter 5

THE IMPACT OF SELECTION BIASES ON TESTS OF GENERAL RELATIVITY WITH GRAVITATIONAL-WAVE INSPIRALS

R. Magee, M. Isi, E. Payne, K. Chatziioannou, W. M. Farr, G. Pratten, and S. Vitale. "Impact of selection biases on tests of general relativity with gravitational-wave inspirals". In: *Phys. Rev. D* 109.2 (2024), p. 023014. DOI: 10.1103/PhysRevD.109.023014. arXiv: 2311.03656 [gr-qc].

E.P. carried out the hierarchical analyses presented, and contributed to the writing of the manuscript.

5.1 Introduction

Gravitational wave (GW) signals detected by LIGO [1] and Virgo [16] have provided otherwise-inaccessible constraints on deviations from general relativity (GR) in the dynamical and strong-field regimes [249, 338, 13, 14, 15]. When considered in aggregate, the set of detected binary black hole (BBH) signals is fully consistent with the null hypothesis of quasicircular mergers in vacuum GR. However, existing constraints apply only to signals that have been confidently detected and identified as compact binaries by pipelines based on GR. Even though generic searches exist [140, 117, 339, 340, 341], all current BBH signals have been detected with search pipelines that are based on templates produced within Einstein's theory. It remains possible that there exist binaries whose signals depart from GR but have been selected against by searches [342, 343, 344]. This raises two interrelated questions: (i) what is the largest deviation from GR that current searches can detect, and (ii) are current constraints on deviations from GR artificially narrow because they are based on signals that were detected in the first place?

Answering these questions amounts to quantifying the *selection biases* that modulate the probability of signal detection as a function of its parameters. The impact of regular binary parameters within GR—such as black hole (BH) masses or spins—can be approximated through their influence on the expected signal-to-noise ratio (SNR) of a given signal [6, 7], or more robustly by assessing the performance of the search pipeline on simulated signals [8]. The resulting selection function is an indispensable ingredient in inferring the astrophysical distributions of the detected

events [6, 7, 8]. While this effect is well understood for GR parameters, the selection on beyond-GR parameters is currently largely unknown and generally unquantified. Nevertheless, studies under specific models suggest searches have nonnegligible selection for sufficiently large deviations [342, 343, 344].

In the absence of a quantified selection function for tests of GR, current constraints are restricted to assessing agreement of the population properties of detected events with GR. Such an analysis can be performed without reference to any specific alternative theory of gravity by inferring the general shape of the population of deviations using hierarchical inference [345, 346, 347, 348]. This procedure can detect anomalies in a collection of signals even if the deviation manifests differently for each individual event [264, 262, 265]. However, without selection effects, this procedure does not infer the *intrinsic* population of deviations, which could contain undetectable signals [262, 14, 15]. Furthermore, if selection biases are strong, these population constraints do not formally correspond to the *detected population* either on account of detector noise [349]. This concern also extends to cases in which events can be combined by simply multiplying likelihoods for a shared deviation parameter.

In this paper, we study the selection function within template-based search pipelines for parameterized tests of the inspiral phasing parameters [66, 69, 350, 351]. Among the wide array of possible GR tests, we focus on post-Newtonian (PN) modifications to the waveform phasing, $\varphi(f)$, due to anomalous dynamics [352, 69, 350, 351, 353, 354, 355, 356, 357, 358, 359, 360, 361], which could arise from corrections to the theory or due to exotic sources following other nonstandard physics, such as BH mimickers. We use the deviation parameters $\delta\varphi_i$, where $i/2$ denotes the associated PN order. We focus on PN modifications as they are one of the flagship tests of GR with LIGO, Virgo, and KAGRA [362], and their effect is to modify the full inspiral, which dominates the detectability of all but the most massive systems. The latter can more easily be detected by theory-agnostic burst pipelines, potentially reducing the expected impact of selection biases induced by deviations from GR.

We generate simulated signals (also called *injections*) and recover them with a simplified version of the GstLAL pipeline [138, 151, 152, 363]in Sec. 5.2. Rather than evaluating the computationally expensive likelihood ratio that would normally be computed by GstLAL as a detection statistic, we approximate detection efficiency with a proxy ranking statistic based on the recovered SNR and an autocorrelation-based consistency check. In Sec. 5.3 we find that, under these circumstances,

selection biases affect the detectability of signals only for very large values of the deviation parameters. These values are significantly higher than the precision achieved by current tests; we therefore expect that incorporating selection effects in population inference will have a minimal impact on the resulting constraints.

Armed with the results from our injection campaign, we confirm this expectation by enhancing existing hierarchical tests of GR [15] with a selection factor, and compute the resulting astrophysical distribution of deviation parameters in Sec. 5.4. We parametrize the deviation population with a Gaussian and infer its mean and standard deviation while taking into account selection effects. Following [84], we simultaneously model the astrophysical distribution of the binary component masses. For most phase deviation terms we consider, the inferred astrophysical distributions for beyond-GR parameters are identical to those obtained by ignoring the GR selection effects. We recover the strongest impact for the -1PN term, where incorporating selection effects widens the inferred population distribution by 10%. We therefore conclude that the quantitative impact of ignoring selection effects in tests of GR with GW inspirals is small.

This conclusion may be surprising given the crucial role of selection effects in estimating, for example, the mass distribution of BBHs. The crucial difference between deviation parameters and BBH masses is that the former population is inferred to be intrinsically very narrow as all events are consistent with a vanishing deviation. Indeed, after a dozen high-significance BBHs, the population for all deviation parameters inferred from LIGO-Virgo data is already narrower than the impact of selection effects. As more events are detected (and assuming they remain consistent with GR), the inferred deviation population will continue to narrow, making selection effects even less relevant. In other words, selection effects do exist in the population, but their impact is only appreciable for deviation values that are already ruled out. Other population distributions, such as those for the mass and spin, are not inherently narrow and selection effects remain important no matter how many events are detected. These considerations suggest that our conclusions only apply under the assumption that all events come from a narrow, unimodal population of deviation parameters. They do not rule out a disjoint population with deviations large enough to remain hidden to searches; such extreme non-GR signals can only be ruled out with a dedicated search [342, 343, 344]. We further this argument in our concluding remarks (Sec. 5.5).

5.2 Estimating the matched-filter selection function for signals with GR deviations

In this section, we describe the procedure for quantifying the effect of GR deviations on the GW selection function. In summary, we follow the standard practice of estimating detection efficiency by simulating a large set of signals, analyzing them with a detection pipeline, and determining which signals are detectable.

Injection Set

We start with the publicly available set of 156 878 BBH injections associated with GWTC-3, which target only GR parameters [364]; we leave detailed explorations of binary neutron stars and neutron-star–black-hole binaries to future work. In this injection set, the primary and secondary binary masses are distributed as $p(m_1) \propto m_1^{-2.35}$ and $p(m_2|m_1) \propto m_2$ and bounded, in the source frame, such that $2\,M_\odot < m_2 \leq m_1 < 100\,M_\odot$; the BH spins are isotropically distributed with uniformly distributed magnitudes $|\chi_{1,2}| \leq 0.998$. Further specifics of the within-GR population are described in Table XII of [77]. The simulations are generated using a baseline IMRPhenomPv2 waveform approximant [133, 134, 107], which includes the effects of spins misaligned with the orbital angular momentum. We implement deviations from GR using the TIGER framework [69, 350, 351], as in [15].

To reduce the computational burden on the original GWTC-3 analysis [77], these injections have already been selected against a minimum optimal network signal-to-noise ratio (SNR) threshold of 6. The network SNR was calculated by adding the LIGO-Livingston and LIGO-Hanford SNRs in quadrature. Systems with a lower optimal network SNR are considered "hopeless" for detection. To further enhance computational efficiency, we only consider BBHs that have optimal LIGO Livingston SNRs ≥ 6 and redshifted total masses below $300\,M_\odot$. For our purposes, restricting the total mass injected has negligible effect due to the additional inspiral SNR selection criterion typically applied in PN tests of GR [15, 84]; we return to this in Sec. 5.4. These initial cuts result in 84 119 injections.

To measure the selection bias against beyond-GR populations, we perturb the inspiral phasing of the injections and recover them with an approximation of the GstLAL-based inspiral pipeline described in Sec. 5.2. Following the standard parametrized post-Einsteinian test [66], we perturb each PN order and repeat the analysis separately. Each simulation is assigned a random fractional[1] deviation drawn from a

[1] In GR, the coefficients corresponding to the -1PN and 0.5 PN terms are exactly zero. $\delta\varphi_{-2}$ and

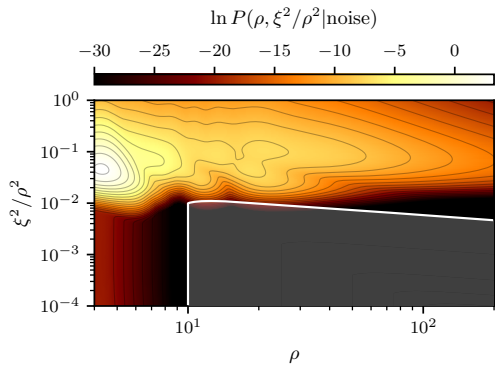

Figure 5.1: A representative background distribution for BBHs collected for the LIGO Livingston detector. The background is parameterized in ξ^2/ρ^2 vs ρ space. Regions with high $\ln P$ indicate where noise is most likely (brighter color). The shaded contour enclosed by a white edge corresponds to our detection criterion, $\bar{\rho} \geq 10$. This region is largely separate from the collected background.

uniform distribution with bounds ±0.1, ±1, ±5, ±3, ±2, ±15, ±5, ±10, ±50, and ±30 for the $\delta\varphi_{-2}$, $\delta\varphi_0$, $\delta\varphi_1$, $\delta\varphi_2$, $\delta\varphi_3$, $\delta\varphi_4$, $\delta\varphi_{5l}$, $\delta\varphi_6$, $\delta\varphi_{6l}$, and $\delta\varphi_7$, respectively, where the "l" subscript denotes the logarithmic phase terms. The bounds are chosen such that the inferred deviations from individual events are entirely covered by the selection. We only vary one coefficient at a time to match the analysis usually applied to actual data [15]. This procedure results in one BBH injection set per PN order, each containing the same number of BBHs with identical GR parameters, differing only in the order and strength of the random GR deviations. After specifying injection parameters, we generate a corresponding waveform using the IMRPhenomPv2 approximant andadd it to the data stream of a single detector. We space the simulated signals 7 seconds apart through a single stretch of data collected in the LIGO Livingston detector during April of 2019 with global-positioning-system (GPS) times in the range [1239641219 s, 1240334066 s] [365].

Detection criterion and efficiency

We analyze the injection sets with a simplified infrastructure based on GstLAL, one of the matched-filter-based search pipelines presently used to search for GWs from compact binaries [151, 139, 138, 363, 366, 367, 368, 369, 148, 143, 150, 370, 371]. Matched-filter based search pipelines discretely sample the GR-based signal manifold to create template banks of possible signals. The discretization results in a 1%−3% loss of SNR over the parameter space covered by the bank [372,

$\delta\varphi_1$ therefore represent absolute deviations.

373]. Pipelines presently restrict their searches to emission from sources with spin angular momenta aligned with the orbital angular momenta, and therefore neglect the impact of precession or higher-order angular modes; the signal loss incurred for these systems is, therefore, larger. We specifically consider the GstLAL-based matched filtering pipeline for its signal consistency check and because it most densely sampled the signal space in LIGO-Virgo's third observing run (O3), and thus had the minimum expected SNR loss from discreteness. For BBHs, the GstLAL bank used an effective-one-body model of the GW emission, SEOBNRv4_ROM [374]. The specific structure and maximum SNR loss of GstLAL's template bank is described in Table II of the GWTC-2 publication [4].

Pipelines correlate waveforms from the template bank with the data collected in each detector to produce an SNR time series. Peaks in the SNR time series, called triggers, are checked for coincidence across detectors, and are then ranked according to the pipeline's detection statistic. GstLAL's ranking statistic is the likelihood ratio $\mathcal{L}$, defined in [375, 363], which relates the probability of observing a set of parameters under the signal hypothesis to that of the instrumental-noise hypothesis. This quantity is a function of a number of factors: the set of instruments participating in a detection, the matched-filter SNR, a signal-based-veto parameter, the event time and phase in the frame of each detector, and the masses and spins of the identifying template. In general, it is computationally expensive to accurately estimate the background of the search and recover simulated signals via $\mathcal{L}$. Since no background for O3 is publicly available, and to minimize the analysis cost, we instead employ an approximate detection statistic $\bar{\rho}$ that weights the measured SNR by a signal consistency check [376, 377], namely

$$\bar{\rho} = \frac{\rho}{\left[\frac{1}{2}\left(1 + \max(1, \xi^2)^3\right)\right]^{1/5}}\,, \tag{5.1}$$

where ρ is the matched-filter SNR and ξ^2 is a signal consistency test defined from the autocorrelation as

$$\xi_j^2 = \frac{\int_{-\delta t}^{\delta t} \mathrm{d}t\, |z_j(t) - z_j(0)\, R_j(t)|^2}{\int_{-\delta t}^{\delta t} \mathrm{d}t\, \left(2 - 2\left|R_j(t)\right|^2\right)}\,, \tag{5.2}$$

where z_j and R_j denote the complex SNR and autocorrelation of template j, respectively, and the integrand in the denominator is the expectation value in Gaussian noise [138]. We compute a value of ξ^2 for each trigger by integrating Eq. (5.2) over a small window of time $\pm\delta t$, centered about the trigger. We use $\delta t = 0.17\,\mathrm{s}$

($\delta t = 0.34\,\mathrm{s}$) for templates with chirp masses greater (less) than 15 $M_\odot$, which was also done in production by the full GstLAL pipeline. When the observed strain data closely matches the template j, then $\bar{\rho} = \rho$. For each BBH injection, we compute the matched-filter SNR ρ and ξ^2 value against the GstLAL template bank. Since signals generally match with multiple templates in a bank, we perform the same data reduction clustering as the GstLAL pipeline does in GWTC-3. We discard triggers within 0.1 s of other triggers with a larger $\bar{\rho}$ value, breaking ties by ρ.

Since we consider the response in only a single detector, we conservatively set a detection threshold of $\bar{\rho} \geq 10$. This choice is motivated by the fact that significant candidates from GWTC-2 and GWTC-3 were identified for network SNR $\rho_{\mathrm{net}} \gtrsim 10$, which typically corresponded to events with single detector SNRs $\rho_{\mathrm{H}} \sim \rho_{\mathrm{L}} \sim 7$. As we only filter a single detector, we assert that a signal in a single detector with $\rho = 10$ will have approximately the same significance as a signal observed in multiple detectors with $\rho_{\mathrm{net}} = 10$. We further assert that our proxy detection statistic threshold is approximately equivalent to the false-alarm-rate (FAR) threshold of $\mathcal{O}(10^{-3}/\mathrm{yr})$ adopted in past tests of GR [13, 14, 15]. This choice is conservative for our study in that a weaker detection criterion could only reduce the *detection* bias, i.e., it could only increase the fraction of signals that are detected by the pipeline.

Although we use an abbreviated version of the detection pipeline, we argue that the resulting selection function is a good approximation for the full selection effect for the following reasons:

1. The threshold of $\bar{\rho} > 10$ selects triggers that are disjoint from the background typically collected by the search. Triggers that meet this criterion exist in the shaded contour shown in Fig. 5.1, which is cleanly off a representative background observed by the search. In other words, $\bar{\rho} > 10$ implies vanishing support from the background.

2. In addition to the background, $\mathcal{L}$ contains a signal term that we do not explicitly take into account here. This is justified because, in the $\bar{\rho} > 10$ region, the noise distribution varies significantly more rapidly than the signal distribution (see Figs. 9 and 10 in [138]). Therefore, the contribution of the signal term to $\mathcal{L}$ is approximately constant over this region, and the FAR is mostly determined by the noise distribution.

3. Finally, although $\mathcal{L}$ depends on parameters beyond ρ and ξ, namely the event time, phase, mass, and spin, those should be minimally affected by the kinds

of GR deviations that we consider here. Since the polarizations are unaffected by phasing corrections and the signals still propagate at the speed of light, the expected distribution of time delays and phase differences across detectors will remain the same. Regarding masses and spins, it is possible for non-GR signals to be identified by GR templates with masses and spins that differ from the source. Though this would change the population model's contribution, the model itself is broad (see Section IVB of the GWTC-2 publication [4]) and contributes weakly to the overall value of $\mathcal{L}$.

These three reasons justify our $\bar{\rho}$ criterion as a proxy for detecting signals with high significance.

5.3 Impact on detection efficiency

To develop intuition for how deviations in the PN parameters affect the detection statistic, $\bar{\rho}$ in Eq. (5.1), Fig. 5.2 shows the SNR and autocorrelation time series with (right) and without (left) a deviation applied to the -1PN coefficient, $\delta\varphi_{-2}$, for a high (top) and low (bottom) injected SNR. We examine these two ingredients of the total detection statistic $\bar{\rho}$ for a characteristic BBH with redshifted masses $30{-}30\,M_\odot$ in the detector frame. The two components of $\bar{\rho}$, ρ and ξ^2, are represented in these plots by, respectively, the peak of the SNR time series (black) and the integrated area between it and the scaled autocorrelation time series (blue). Mismatches between a signal and the template bank induced by a GR deviation will impact detection efficiency due to both a loss in the recovered SNR ρ (reduction in the peak height) and increase in the signal consistency check value ξ^2 (increased disagreement between blue and black curves).

Indeed, the beyond-GR deviation causes a reduction in the recovered SNR, seen through a reduced peak between the left and right panels of Fig. 5.2, thus directly affecting $\bar{\rho}$. Moreover, the introduction of beyond-GR effects creates secondary peaks in the SNR time series obtained from filtering with a GR waveform. The oscillations in SNR further reduce the signal consistency check, ξ^2—that is, the square difference between the measured SNR and the scaled autocorrelation, per Eq. (5.2). These oscillations become harder to discern from the Gaussian background with decreasing SNR, thus minimizing the effect of ξ^2 on the detectability of the signal. Figure 5.2 is helpful in understanding the interplay between ρ and ξ^2 in the presence of a deviation from GR. However, it is not sufficient to determine the degree of selection bias against beyond-GR signals, as it only shows the effect of a single

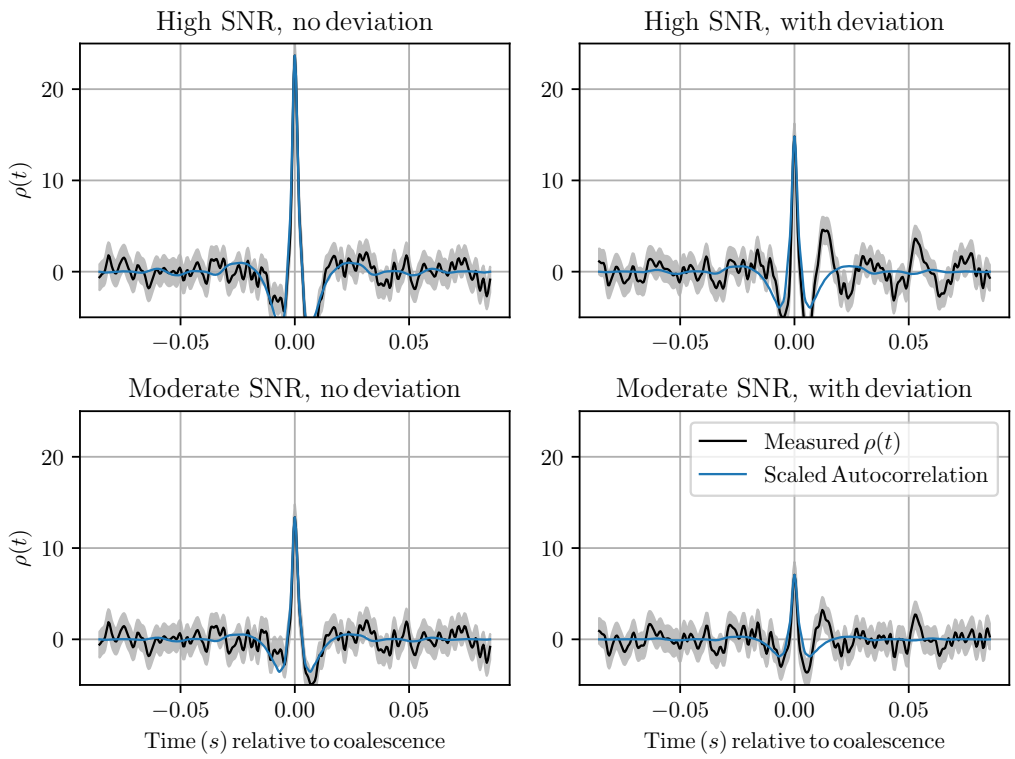

Figure 5.2: The response of a single search template to a $30M_\odot - 30M_\odot$ BBH without (left) and with (right) deviations to $\delta\varphi_{-2}$ for SNR $\sim$ 24 (top) and $\sim$ 15 (bottom) injections in Gaussian noise colored to O3 sensitivities. The injections that deviate from GR use $\delta\varphi_{-2} = -0.1$. The black line shows the measured SNR time series for a single template waveform, with the gray band denoting the 1σ measurement uncertainty. The beyond-GR phasing results in an SNR loss of $\sim 40\%$ between the left and right columns. Additionally, there is a mismatch between the measured SNR time series and the SNR scaled autocorrelation that weakens the signal consistency test, ξ^2. Both effects lead to a reduction of our detection statistic $\bar{\rho}$, Eq. (5.1), and thus a loss in sensitivity.

injection relative to the corresponding GR template with the same parameters. In an actual search, we compare a beyond-GR injection against the entire bank, and the detection statistic is based on the best match.

To quantify the actual impact of GR deviations on the detection efficiency, we study the distribution of parameters of the signals that made it through our simplified detection pipeline, i.e., those that returned a value of $\bar{\rho} > 10$ when compared against *any* template in the GR bank. This amounts to measuring the detectable fraction,

$$\hat{\mathcal{E}}(\Lambda) = \int \mathrm{d}\theta \, p_{\mathrm{det}}(\theta) \, \pi(\theta|\Lambda) \,, \tag{5.3}$$

where Λ is the set of hyperparameters that describe the underlying population dis-

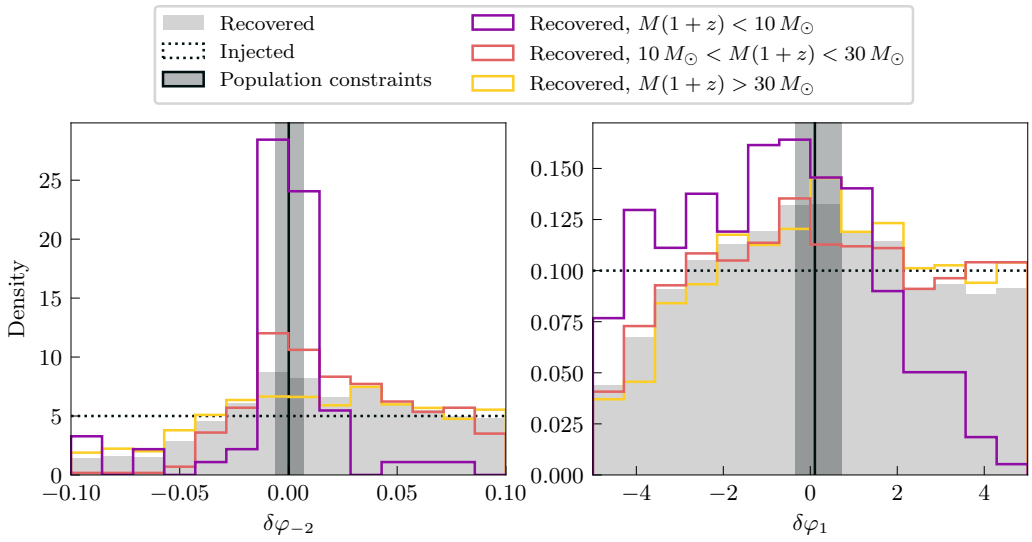

Figure 5.3: Histograms of recovered injections with deviations from GR in the -1PN ($\delta\varphi_{-2}$, left) and 0.5PN ($\delta\varphi_1$, right) coefficients. Although the initial injection set was assigned deviations from a uniform distribution (dotted black), the pipeline selects against large negative values of the deviation parameters, as indicated by the dearth of detections in the leftmost bins (gray histograms). Besides the total set of injections, we show sub-distributions corresponding to different injected mass bins in the detector frame (colored histograms). The distributions of recovered injections are largely flat over the span of values allowed by the analysis of the 12 events considered in Sec. 5.4 (which are ~4× broader than GWTC-3 constraints [15]; vertical gray band, median and 90% CL), suggesting that the selection bias is not strong enough to affect the population constraints.

tribution, $\pi(\theta|\Lambda)$, and $p_{\rm det}(\theta)$ is the selection function that describes the probability of detecting a system with parameters θ.

Figure 5.3 shows the marginal selection function, $p_{\rm det}(\delta\varphi)$, for the -1PN coefficient ($\delta\varphi_{-2}$, left) and the 0.5PN coefficient ($\delta\varphi_1$, right), over the whole mass space (gray) as well as subsections for different BBH mass bins (colors). For both parameters, the distribution of detected signals departs from the uniform intrinsic distribution that we injected (black): there is a dearth of detected signals with large negative values of the deviation parameters, indicating that such signals are selected against. This can be explained by the fact that a negative value for these parameters will shorten the inspiral, which in turn reduces the SNR of the signal. This effect is more pronounced for the -1PN coefficient, which is consistent with the intuition that this coefficient should have a larger impact on the GW phase than the 0.5PN coefficient over the duration of an inspiral because it is associated with a correction entering

at a lower power of the frequency. The drop in detection efficiency is also sharper for lower masses, as expected given the scaling of the inspiral length with the BBH mass.

In spite of the drop in sensitivity observed at the edges of the histograms in Fig. 5.3, the recovered distributions are generally flat in the region that is allowed by the population constraints from GWTC-3 (gray band). Lower detector-frame masses demonstrate a larger gradient across these regions (e.g. $M(1+z) < 10\,M_\odot$; purple). However, the observed events considered here do not reside in this region of the mass parameter-space. Since there is no gradient in the region allowed by the observations, there is no preference for any particular value of the deviation parameter in the range still consistent with current data. This suggests that the selection bias is not strong enough to affect the population constraints, which are more sensitive to GR deviations than the detection pipeline. We confirm this below by repeating catalog analysis of GR deviations with and without the selection effects.

5.4 Updated population estimates

We incorporate the selection function computed from Sec. 5.3 into population-level inference for inspiral tests of GR. By computing the astrophysical distribution of beyond-GR parameters, we can now make statements about the types of GR deviations consistent with an observed set of detections. In practice, computing the astrophysical distribution requires incorporating knowledge of the detection efficiency over parameter space to deconvolve the instrument's selection function from the set of observed measurements.

We evaluate the consistency of a set of observations with GR through a hierarchical analysis without imposing strong assumptions about the nature of the deviation across events. As a null test, we follow [262, 265, 84, 15] in parameterizing the intrinsic distribution of individual-event values for some deviation parameter $\delta\phi$ as a Gaussian $\delta\phi \sim \mathcal{N}(\mu, \sigma)$. This model targets the mean μ and variance σ^2 of GR deviations, regardless of the true shape of the underlying distribution. Beyond-GR parameters are typically defined to vanish in GR, so that the null hypothesis that GR is valid for all events predicts $\mu = \sigma = 0$. If GR is not correct, then the deviation parameters may take different (nonzero) values as a function of source parameters, resulting in nonvanishing μ or σ. We apply the approach in [84] to simultaneously model the distribution of astrophysical parameters.

Existing implementations of this hierarchical analysis characterize the set of *ob-*

served events but do not inform about possible *intrinsic* deviation distributions that predict events with such large deviations that are undetectable. To factor this in, we use the result of Sec. 5.3 following the techniques used in the context of astrophysical inference to study the astrophysical distribution of within-GR parameters, such as masses and spins. The key additional step is to incorporate the detection efficiency into the hierarchical likelihood through a term that can be approximated as the Monte-Carlo sum population weights over a set of m detected injections with parameters θ_k [289, 282, 332],

$$\hat{\mathcal{E}}(\Lambda) = \frac{1}{M}\sum_{k}^{m}\frac{\pi(\theta_k|\Lambda)}{p(\theta_k|\text{draw})}\,, \tag{5.4}$$

where M is the total number of drawn injections (out of which m were detected), $p(\theta_k|\text{draw})$ is the probability of drawing parameters θ_k from the population adopted in the injection campaign, with $\Lambda = \{\mu, \sigma\}$, in addition to the parameters describing the astrophysical population of GR quantities (like masses and spins). The hierachical likelihood, $p(\{d\}|\Lambda)$, governing the inferred astrophysical population from N observations with dataset $\{d\}$ is

$$p(\{d\}|\Lambda) = \frac{1}{\hat{\mathcal{E}}(\Lambda)^N}\prod_{i}^{N}\int \mathrm{d}\theta_i\, p(d_i|\theta_i)\pi(\theta_i|\Lambda)\,, \tag{5.5}$$

where $p(d_i|\theta_i)$ are the individual event likelihoods. The selection function influences the inferred hyperparameters through its inclusion in Eq. (5.5).

In order to include an injection in the "detected" sum of Eq. (5.4), besides GstLAL's detection threshold of $\bar{\rho} > 10$ from Sec. 5.2, we additionally require that the measured SNR in the inspiral satisfy $\rho_{\text{insp}} > 6$. The latter corresponds to the selection criterion for estimating the inspiral PN coefficients in [13, 14, 15]. In order to avoid computing the inspiral SNR for each injection in the set, we approximate the fraction of SNR in the inspiral as a linear function of the detector frame total mass as in [84].

In addition to hierarchically modeling the beyond-GR astrophysical distribution, we incorporate population models for the within-GR population distributions. Due to a lower number of recovered injections than the standard set of injections used in population studies [332], we only infer the primary mass and mass ratio distributions jointly with the beyond-GR population, using the models outlined in Ref. [84]. We fix the spin distribution to be uniform in spin-magnitude and isotropic about all

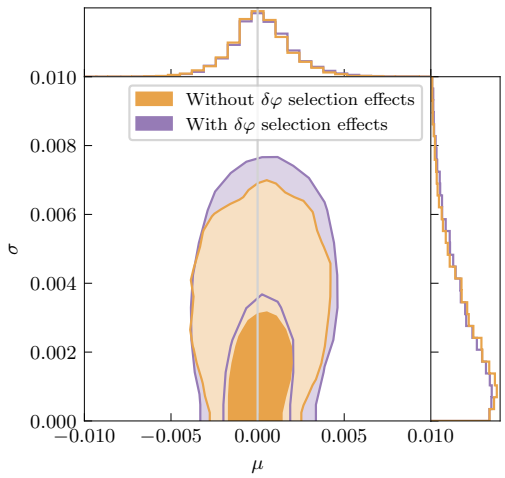

Figure 5.4: Inference on the mean and standard deviation of the −1PN coefficient, $\delta\varphi_{-2}$. The orange contours show the result of the hierarchical analysis without accounting for selection effects, while the purple contours show the result when the selection function is included. The two results are consistent with each other, with the selection function widening the population only slightly. We find no difference in the coupling between μ and σ and the parameters controlling the mass distribution either (not shown).

possible spin orientations; the redshift distribution is consistent with the *maximum a posteriori* power-law found in Ref. [8].

With the setup described above, we repeat the hierarchical analysis in [262, 14, 15, 84] applied to 12 events in O3a, to be consistent with times over which the selection function is estimated. A list of the included events can be found in Table I of Ref. [84]. Figure 5.4 shows the resulting inference on μ and σ for the −1PN coefficient, $\delta\varphi_{-2}$, compared to the result that does not account for selection biases in the beyond-GR parameters. Although this was the coefficient with the strongest detection bias as evaluated in the previous section (Fig. 5.3), this effect is very small, and the two results, with and without selection, are consistent with each other up to a slight widening of the population when selection is factored in. This is consistent with the expectation from Fig. 5.3, which suggested the impact of selection should be minimal in light of the accuracy of the constraint from parameter estimation. Figure 5.5 shows that this is the case for all coefficients, none of which show significant differences between the two results.

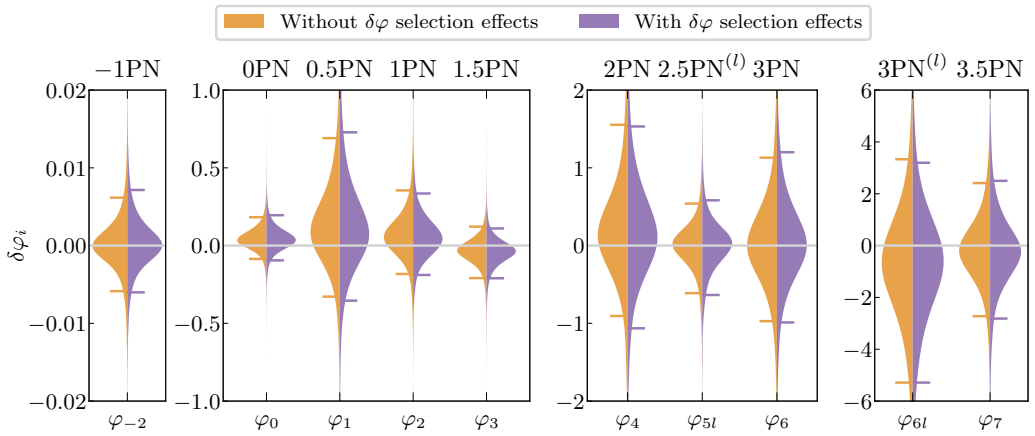

Figure 5.5: Posterior predictive distributions (also known as the population-marginalized expectation) for deviations at all PN orders we consider, without (orange) and with (purple) selection effects factored in. No coefficient shows a significant impact when factoring in the selection: the $\delta\varphi_{-2}$ displays the strongest effect, with a slight broadening of the inferred distribution at the level of ∼10%.

5.5 Conclusions

In this study, we revisited tests of GR from the inspiral GW phase by accounting for the selection effect of templated searches against signals with GR deviations. We estimated the selection function by considering the performance of a simplified version of the GstLAL search pipeline against simulated signals with beyond-GR effects affecting the PN evolution of a BBH inspiral. Since GstLAL detects signals by comparing them to a template bank constructed with GR waveforms, its detection efficiency decreases under sufficiently large deviations from GR. However, we found that this threshold for deviations is less stringent than the precision of GWTC-3 constraints, suggesting that population inference on the inspiral deviation parameters is minimally affected by selection effects. In other words, existing constraints are already a very good approximation to the full astrophysical population of deviation parameters, apart from the possibility of a disconnected subpopulation of sources with very high deviations.

This finding can be understood by noting that the sensitivity of parameter estimation to deviations from GR scales inversely with the SNR of the signal, while the detection threshold imposed by the search pipelines is best represented as a hard SNR cutoff. A deviation $\delta\varphi$ that induces a mismatch $\mathcal{M}$ relative to the best-fitting GR template will result in an SNR loss of order $\rho \to \mathcal{M}\rho$; accordingly, the measurement precision in parameter estimation will scale as $\Delta(\delta\varphi) \sim 1/\rho$. For a given SNR, the

mismatch tolerated by the search pipeline will be much higher than the sensitivity of the parameter estimation. Therefore, signals that incur an SNR penalty would still be detectable as long as they remain above the search's threshold; meanwhile, given a GR signal in the data, parameter estimation will constrain the magnitude of a deviation tightly around zero, with much better precision than would be directly associated with the pipeline's detection threshold.

In other words, the tolerance for detection is much larger than the tolerance for parameter estimation, and the latter is what determines the population constraints. Since the population of observed deviations is extremely narrow (a delta function at zero if GR is correct), the hierarchical measurement is minimally affected by selection effects, as we have shown in Fig. 5.5. This argument does not apply to other parameters, such as the BH masses, since their distribution is intrinsically broad.

Our main conclusion is that the deviation population is already narrower than the extent of the selection effects, and thus the latter do not impact the former. However, this assumes that deviations form a single, compact population whose mean and standard deviation we constrain. Since no observed events are inconsistent with GR, the inferred width of this population grows smaller as the catalog increases. We are therefore not considering, and thus not ruling out, disjoint populations with a subset of events that have extremely large (and potentially undetectable) deviations or a mass-dependent deviation population model. It remains conceivable that a subpopulation of signals with extremely high deviations could exist and remain hidden from GR-based pipelines, motivating dedicated searches [342, 343, 344]. However, that does not translate into selection biases for the components of the population that are already constrained by the existing catalog.

This distinction also suggests that there is no contradiction between our results and those of Ref. [342, 343, 344]: we both find appreciable selection effects for sufficiently large values of the deviation parameters, c.f., Fig. 5.3. Our study, however, highlights that under the assumption of a single, unimodal population distribution of the deviation parameters, such large values of the deviation parameters are already ruled out.

As Essick and Fishbach [349] recently pointed out, the existence of prominent selection biases would complicate the interpretation of hierarchical constraints that do not factor in selection effects, as the inferred population would not be strictly representative of neither the true astrophysical distribution nor the observed distribution

of parameters. However, in the absence of strong selection effects, hierarchical inference *without* a selection term remains a valid tool to constrain the population of beyond-GR parameters, as we have shown here for PN tests of the BBH inspiral. This, of course, may not be the case for other tests or implementations.

Our results are subject to a number of caveats, and selection effects might be stronger for different GR tests or population models. First, to mitigate computational costs, we have used an approximate ranking statistic that only incorporates information from a single detector. We impose a detection threshold of $\rho \geq \bar{\rho} \geq 10$ to maximize purity in accordance with the FAR threshold adopted in past GR tests [13, 14, 15]. We do not expect a full injection campaign utilizing the complete ranking statistic described in [363] would yield more precise results at this threshold and for the inspiral deviation test considered here. However, our results do not obviate the need for a full injection campaign for other tests of GR or other pipelines.

Besides the adopted threshold, the $\bar{\rho}$ ranking statistic differs from the full likelihood ratio also on the information it considers. The latter also includes information about the phase and time of the signal in different detectors. Though we do not expect those terms to be important for the inspiral deviation parameters we consider here, they could become important for other tests of GR, such as those considering propagation effects or the signal polarization. Quantifying selection effects for such tests would require a full multi-detector and likelihood ratio calculation.

We produce injected signals with GR deviations using standard infrastructure [133, 134, 107, 69, 350, 351], and choose parameter ranges consistent with priors used in LIGO-Virgo-KAGRA publications. However, for some of these extreme values, the resulting waveform could become pathological [378], and may not represent a physically meaningful configuration [324]. Although this might affect the overall applicability and physical interpretation of the tests, it does not affect the interpretation of our results that relate to the selection effects of the tests as formulated. Reformulations of the inspiral tests to ensure the GW phase calculation remains in the convergent series expansion regime [324, 267] would likely be affected by selection effects even less, as they restrict the allowed range of possible deviations.

Among the compact-binary pipelines, we restrict to a simplified version of GstLAL. We expect the impact of this assumption to be small, as we only consider the most confidently-detected BBHs with single detector SNRs $\gtrsim 10$, all of which are detectable by GstLAL. If we decreased the SNR threshold, we might encounter events detected by other compact-binary pipelines, in which case we would need

to quantify their selection effects. However, we expect that relaxing SNR or FAR thresholds should only make pipelines more tolerant to signals beyond GR.

Extending beyond matched-filter pipelines, we expect weakly-modeled search methods [140, 117] to surpass template-based ones for sufficiently large GR deviations. However, it is the case that both all events we consider here and all events that have been detected in general are detected significantly by at least one template-based search. Ultimately, the sensitivity of weakly-modeled searches should also be quantified and taken into account, though some have started to explore the biases this would introduce [344].

As the sensitivity of GW detectors improves, so does the number and quality of detections, leading to increasing sensitivity to both subtle deviations from GR and systematics in our models. While here we have focused on tests of GR based on GW inspiral phases and single-Gaussian populations, exploring the effect of selection biases in other tests or under other population models will also become important. As both our detectors and techniques evolve, future studies need to evaluate this and other potential systematics.

Chapter 6

THE CURVATURE DEPENDENCE OF GRAVITATIONAL-WAVE TESTS OF GENERAL RELATIVITY

E. Payne, M. Isi, K. Chatziioannou, L. Lehner, Y. Chen, and W. M. Farr. "Curvature Dependence of Gravitational-Wave Tests of General Relativity". In: *Phys. Rev. Lett.* 133.25 (2024), p. 251401. DOI: 10.1103/PhysRevLett.133.251401. arXiv: 2407.07043 [gr-qc].

E.P. helped conceive the idea, undertook all calculations presented, and led the writing of the manuscript.

6.1 Introduction

Searches for deviations from General Relativity (GR) with gravitational waves (GWs) are hampered by the vast landscape of alternative theories [379, 380] and the scarcity of detailed predictions under any specific theory. Faced with these challenges, most tests are framed as theory-agnostic searches for generic deviations [249, 12, 13, 14, 15]. Although this approach has provided increasingly precise null tests, it forgoes physical expectations for the likely behavior of realistic deviations, making constraints harder to interpret and potentially less sensitive [324, 254]. However, even without reference to a specific beyond-GR theory, general arguments limit how deviations may manifest under broad theory classes. This presents an opportunity for improving tests of GR with GWs.

In this paper, we exploit one such argument arising from effective field theory: the magnitude of the beyond-GR effect should scale with the spacetime curvature of the source, e.g., [87, 88]. Since curvature is proxied by mass, *lighter* systems should manifest *larger*—and hence more measurable—deviations. In the context of binaries observed with GWs, this expectation has been folded in (though not directly inferred from) post-merger (ringdown) constraints [381, 382, 383] and simulations of residual cross-correlated power between detectors [384]. Here, we go beyond folding in fixed values of the curvature scaling [381, 382, 383]: we exploit the fact that the curvature scaling will manifest within an ensemble of observations to *directly infer* the curvature dependence of GR extensions without resorting to

theory-specific assumptions. A direct measurement of the curvature scaling will then provide insights on the modification to the Einstein-Hilbert action.

We propose a search for deviations from GR in a catalog of GW observations that leverages this effective field theory insight. Instead of committing to a specific theory, we constrain expected morphologies from a large set of potential theories at once. Deviations from GR are linked to the leading power correction in the Einstein-Hilbert action, e.g., [385, 386, 58, 59, 60, 387], with a Lagrangian

$$\mathcal{L} = \mathcal{L}_{\rm GR} + \lambda\, F_\gamma(\mathcal{R}, \phi)\,, \tag{6.1}$$

where $\mathcal{L}_{\rm GR}$ is the GR term, $F_\gamma(\mathcal{R}, \phi)$ is some functional of the curvature $\mathcal{R}$ (and potentially other degrees of freedom ϕ) scaling as $\mathcal{R}^\gamma$, and λ is the dimensionful coupling coefficient.[1] Dimensional analysis reveals that $\lambda \sim \ell^{2(\gamma-1)}$, representing a theory-specific coupling governed by a theory-specific length-scale, ℓ. Importantly, this scaling is imprinted in (dimensionless) deviations from GR *regardless of the physical mechanism* they induce.

For instance, consider beyond-GR theories with cubic or quartic curvature corrections, $\gamma = \{3, 4\}$, and no further degrees of freedom. Such theories introduce tidal effects in black hole binaries, under the assumption that the theory-specific length scale is smaller than the lightest black hole. Such deviations first appear at the 5th post-Newtonian (PN) order through tidal Love numbers whose magnitude depends on the specific correction.[2] Crucially, these deviations scale as $\lambda/M^{2(\gamma-1)} \propto M^{-\{4,6\}}$ [388, 60, 389], with M the binary total mass. Additional degrees of freedom impact these scalings. For instance, quadratic theories, $\gamma = 2$, with additional degrees of freedom yield hairy black holes with deviations in the inspiraling GW signal at either -1 or 2PN order depending on the parity of the correction [390]. Now the dimensionless deviation is $(\lambda/M^{2(\gamma-1)})^2 \propto M^{-4}$ with the additional square power coming from the coupling of the scalar degree of freedom and the metric tensor. In either case, constraining the value of the mass exponent has tremendous power in narrowing viable corrections.

In this chapter, we exploit the expected curvature/mass scaling in the context of deviations in the post-Newtonian inspiral phase of binary black-hole coalescences. We incorporate the mass dependence into hierarchical tests of GR and infer both the

[1] The functional $F_\gamma(\mathcal{R}, \phi)$ could include any combination of curvature tensors, and/or their derivatives (not just the Ricci scalar R), derivatives of ϕ, and couplings that scale as $\ell^{-2\gamma}$ [386].

[2] Nominally 2PN effects are also introduced, but their contribution is subleading to tidal effects if $\ell \lesssim 5\,$km.

magnitude and the curvature dependence of measured deviations. Using observations from the third LIGO-Virgo-KAGRA [1, 16, 362] GW transient catalog [4, 77], we confirm the validity of GR. We further highlight the method's ability to constrain the curvature order at which a modification appears with simulated observations.

6.2 Constraining the curvature dependence with gravitational waves

Combining information from a catalog of GW observations in a theory-agnostic way amounts to characterizing the distribution of putative deviations [264, 262, 265], with GR recovered under vanishing deviations for all sources. This hierarchical framework can incorporate arbitrary numbers of deviation parameters [391], astrophysical parameters [84], and selection effects [85]. In all cases so far, the deviation population has been modeled as a (potentially multidimensional) Gaussian, whose mean μ and variance σ^2 are global parameters, independent of source properties. This framework provides a powerful null test of GR, which is recovered for $\mu = \sigma = 0$, see Ref. [392] for inference caveats; however, it does not impose any structure on the scale of the deviations as a function of source parameters.

We extend the hierarchical framework to incorporate the expectation that the magnitude of deviations scales with source curvature by anchoring the deviation distribution to the total binary (source-frame) mass M. We achieve this by reparametrizing μ and σ as

$$\mu = \mu_0 \left(\frac{M}{10\,M_\odot}\right)^{-p}, \quad \sigma = \sigma_0 \left(\frac{M}{10\,M_\odot}\right)^{-p}, \tag{6.2}$$

where μ_0 and σ_0 control the magnitude of the conditional mean and spread of the GR deviation at $M = 10\,M_\odot$. The curvature scaling order, p, is directly related to the index, γ in Eq. (6.1), as either $p = 2(\gamma - 1)$ in the absence of additional fields or $p = 4(\gamma - 1)$ in their presence. In this notation, $p = 4$ corresponds to quadratic curvature corrections with additional degrees of freedom or cubic corrections in their absence while $p = 6$ implies quartic corrections. Propagation effects, such as modifications to the GW dispersion relation [255, 256], impose source-independent deviations and, thus, $p = 0$. Even when the deviation distribution is not Gaussian, this method will still identify a violation of GR [265] and, if the shape of the distribution is unchanged over the binary mass range, also identify the scaling, p. Irrespective of p, GR corresponds to $\mu_0 = \sigma_0 = 0$.

While this framework can be applied to any test that infers both the deviation and the system total mass (such as correlated power among detectors [384, 393]), we turn our attention to the post-Newtonian phase deviation test [66, 354, 69, 70, 254].

Deviations at the $(k/2)$PN order are inferred by varying the respective phasing coefficient by some (dimensionless) fractional deviation, $\delta\varphi_k$ [70]. Since GR is recovered when $\delta\varphi_k = 0$ and the parameter is dimensionless, this is a null test that should follow a curvature scaling as we have described. Below, we consider deviations from -1PN up to 3.5PN order, including logarithmic terms [4, 77].

Following Refs. [4, 77, 15, 84], we consider the 20 observations from the third LIGO-Virgo observing run with a false-alarm-rate of less than $1/1000\,\mathrm{yr}$ and with an estimated inspiral signal-to-noise ratio greater than 6 (Tab. I of Ref. [84]). We do not consider data from the first and second observing periods as semi-analytic simulations for sensitivity estimation are not available. Individual-event posteriors were computed in Refs. [4, 77, 15] with a modified form of the SEOBNRv4 waveform [271, 272, 274, 273, 254] and released in Refs. [298, 275]. To mitigate against systematic bias due to incorrect astrophysical assumptions, we jointly model the distribution of the GR deviation parameter and the system masses and spins with the astrophysical population models and selection function [84]. Based on Ref. [85], we assume that there are no direct selection effects for the magnitude of the deviation. We infer the population distribution of each post-Newtonian term separately with uniform hyperpriors on $\mu_0 \in [-30, 30]$, $\sigma_0 \in [0, 100]$, and $p \in [-1, 8]$, chosen so as to remain agnostic on the magnitude and character of the curvature scaling.

The expected inference structure depends on a number of considerations. Observations of signals spanning $10-100\,M_\odot$ in total mass [4, 77] have yielded no evidence for a violation of GR [15]. Among those, constraints are generally stronger for lighter signals with more inspiral cycles [394], however, there are more observed signals at $M \sim 60\,M_\odot$ [8]. Importantly, for $p \geq 4$ lighter systems are expected to manifest larger deviations. Absent a detected deviation, we expect those systems to provide the overall strongest constraints. This expectation plays out in the results below.

Figure 6.1 shows results for the -1PN deviation, related to deviations due to a scalar field coupling to the Gauss-Bonnet invariant, i.e., Einstein-scalar Gauss-Bonnet [390]. We discuss this order in detail, but obtain qualitatively similar results for other PN orders. The constraints are consistent with $(\mu_0, \sigma_0) = (0, 0)$ and, thus, GR.

To further understand the posterior, consider that, in the absence of inferred deviations and denoting the most informative mass range as M_I, the allowed values of

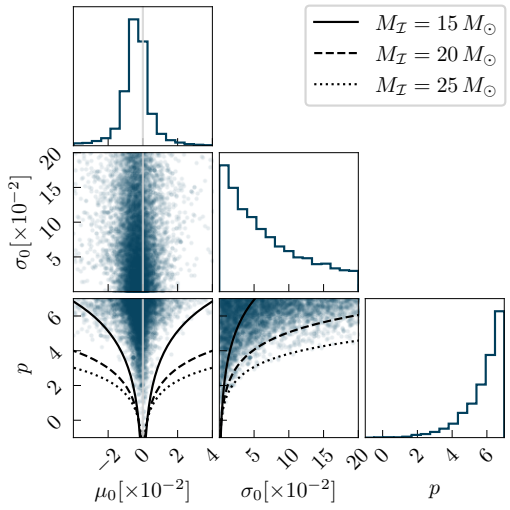

Figure 6.1: Posterior distribution for the −1PN deviation population parameters inferred from the 20 GW observations in GWTC-2 and GWTC-3 which pass the threshold criteria [15, 4, 77, 84], confirming consistency with GR, $(\mu_0, \sigma_0) = (0, 0)$. Due to the non-detection of a violation, the constraint is dominated by $M_{\mathcal{I}} \in [15, 25]\,M_\odot$ and the posterior is bounded per Eq. (6.3) (lines). While the marginal posterior for the scaling parameter, p, indicates preference for larger values, it is a product of this bounded structure.

$\{\mu_0, \sigma_0, p\}$ correspond to deviations that would be undetectable at $M_{\mathcal{I}}$:

$$\{\mu_0, \sigma_0\} \left(\frac{M_{\mathcal{I}}}{10\,M_\odot}\right)^{-p} \sim \text{const.}\,, \tag{6.3}$$

where the constant represents the test sensitivity. To determine $M_{\mathcal{I}}$, we split events based on their total mass into $5\,M_\odot$ bins and compute the *precision*

$$\mathcal{P}(M, p) \equiv \frac{1}{\Sigma^2(M, p)} = \sum_{i=1}^{N_\mathrm{b}} \frac{1}{\Sigma_i^2 \left(\frac{M_i}{10\,M_\odot}\right)^{2p}}\,, \tag{6.4}$$

where i runs over the N_b events within the bin with central mass M, Σ_i^2 is the variance of the GR deviation of an individual event marginalized over all other parameters, and M_i is the median total mass. The precision corresponds to the total inverse variance scaled by the expected value of the deviation in each mass bin; it therefore quantifies which mass range is more constraining. For the −1PN order, the precision is maximized when $M_{\mathcal{I}} \in [15, 25]\,M_\odot$, resulting in the black lines in Fig. 6.1, which track the general shape of the posterior.

Equation (6.4) qualitatively characterizes the inference: constraints are improved either with more observations or with better measurements. The M^{-2p} term further

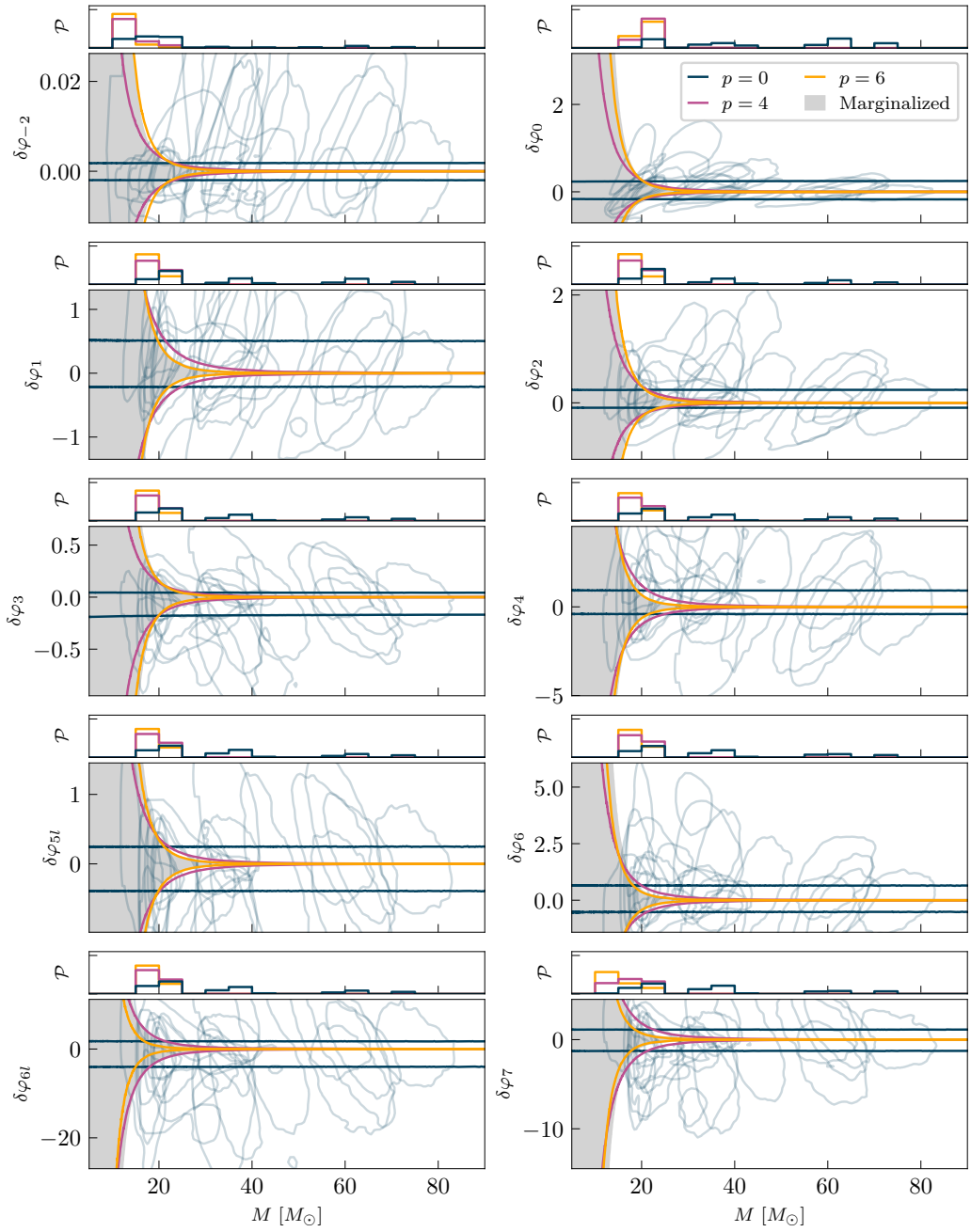

Figure 6.2: Posterior predictive distributions, Eq. (6.5), for deviations across all PN orders (main panel) and the precision, $\mathcal{P}$ (top sub-panel), as a function of binary total mass. We show results for fixed values of $p = 0, 4, 6$ indicative of different theoretical models and when marginalizing over p. The 90% credible regions of the 20 individual-event posteriors are shown in faint blue. The precision indicates the relative contribution on the constraints for the different curvature orders, generally maximized at $\sim 20\,M_\odot$; it is normalized for each p.

indicates that upper limits from heavier systems are less informative than numerically similar upper limits from lighter systems for $p > 0$. The funnel-like structure in $\mu_0 - p$ and $\sigma_0 - p$ is driven by Eq. (6.4) and leads the marginal for p to prefer higher values (since lower values are disallowed by the data). This feature is, however, prior-dominated and will remain so until a deviation is detected and $(\mu_0, \sigma_0) = (0, 0)$ is excluded.

In Fig. 6.2 we plot the distribution of deviations that are consistent with observations for each PN order, i.e., the posterior predictive distribution,

$$p(\delta\varphi_k|M, d) = \int \mathrm{d}\Lambda\, p(\Lambda|d)\, \pi(\delta\varphi_k|\Lambda, M)\,, \tag{6.5}$$

where $\Lambda \equiv \{\mu_0, \sigma_0, p\}$, $\pi(\delta\varphi_k|\Lambda, M)$ is the deviation Gaussian distribution, $p(\Lambda|d)$ is the posterior on Λ (cf., Fig. 6.1), and d is the data. The integral is computed by averaging Gaussian distributions $\pi(\delta\varphi_k|\Lambda, M)$ over the posterior $p(\Lambda|d)$. We present Eq. (6.5) for $p = 0, 4, 6$ (blue, purple, and orange) as well as integrating over p (shaded). The distributions areconsistent with GR, $\mu_0 = \sigma_0 = 0$, constrained within the $Q_{\rm GR} < 46\%$ credible intervals for all PN orders. We find overall similar behavior across orders. In each panel, the upper sub-panel shows the precision $\mathcal{P}(M, p)$, normalized independently for each index p. For all cases of p, the precision is maximized at $\sim 20\,M_\odot$. For $p \geq 4$, corresponding to corrections for gravity in 4-dimensional spacetimes [395, 58, 396], constraints are dominated by lower total mass binaries.

6.3 Detectability of simulated violations

The current GW catalog does not exhibit evidence of a deviation from GR, we therefore explore inference in the presence of deviations with a simulated catalog of $N = 5000$ observations. We consider the 0PN order and simulate data with $\mu_0 = 0$, $\sigma_0 = 0.3$, $p = 4$ per Eq. (6.2), which is consistent with current constraints. For simplicity, we adopt a mass distribution that matches the current observations and apply no selection effects. With this simulated catalog, we repeat the analysis and present 90% constraints on σ_0 and p for varying numbers of detections in Fig. 6.3 (blue). For reference, we compare to an analysis that fixes $p = 0$ (orange), corresponding to the standard procedure of Refs. [84, 15].

Fewer observations are required to identify a deviation from GR ($\sigma_0 > 0$) than to constrain its curvature scaling. For these simulations, $\sigma_0 = 0$ is excluded at the 90% level after $\sim$100 observations, whereas data-driven (as opposed to prior-dominated—c.f., discussion of Fig. 6.1) constraints on p require $O(500)$ observations (blue). A

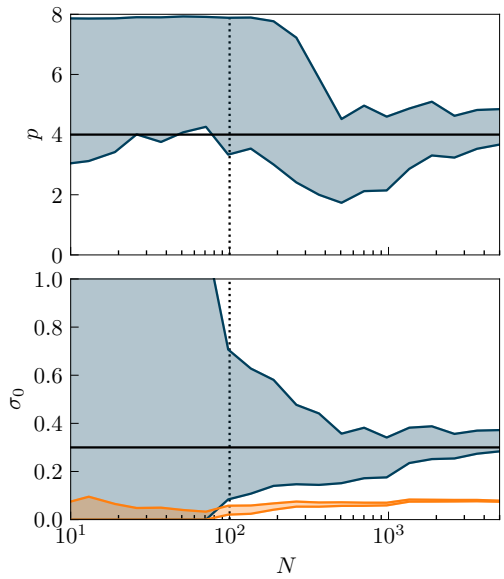

Figure 6.3: Inferred curvature scaling p (top) and standard deviation σ_0 (bottom) at the 90% level as a function of the number of simulated GW observations. The blue bounds correspond to an analysis that infers the curvature index, p, whereas the orange corresponds to fixing $p = 0$. The true values are shown in solid black horizontal lines. For this population we infer a violation of GR, i.e., $\sigma_0 > 0$, starting at $N \sim 100$ (dotted black vertical line), while $p = 0$ and 6 are ruled out by the data after $N \sim 500$ observations. Fixing $p = 0$ misestimates the deviation.

model without curvature scaling (fixing $p = 0$) identifies a violation of GR with a similar number of observations but provides no information about its curvature scaling and infers a lower value of $\sigma_0 \sim 0.08$ (orange). The addition of the curvature dependence in the inference unlocks the capability to infer the curvature structure and characterize the properties of a putative deviation. In this example, we would be able to rule out propagation effects ($p = 0$) and quartic curvature corrections ($p = 6$) after ~500 observations. Although these exact numbers depend on the mass distribution and simulated deviation, we expect the general trends to be robust.

6.4 Conclusions

In this chapter, we have extended tests of GR with GW inspirals to incorporate physical expectations for the curvature dependence of extensions of GR. This approach not only incorporates more physically realistic—albeit still theory-agnostic—models, but also allows us to better characterize the nature of the deviation by inferring its scaling with spacetime curvature. We applied this method to existing LIGO-Virgo-KAGRA observations, finding consistency with GR. We also demonstrated, with

simulated signals, how the curvature dependence can be constrained and thus provide clues about the properties of the beyond-GR theory. Although we focused on PN inspiral deviations, this method can be applied to any test with a dimensionless deviation parameter. More broadly, the key realization of this work, namely that the curvature scaling can be agnostically inferred from data, can be leveraged across all tests of GR—beyond the field of GW astronomy. Beyond GR, this agnostic approach can be tailored to any effective-field-theory treatment, e.g., to analyze the temperature-dependent scaling of transport coefficients affecting viscous effects in hydrodynamics [397].

Beyond constraining the curvature dependence, our physical arguments suggest ways to either strengthen confidence in a detected deviation or safeguard against systematics, e.g., [398]. Firstly, if GR is found to be incorrect and the curvature scaling p is inferred to be an integer, it will immediately inform on viable theories. Further, extraction of p at different PN orders would not only allow for consistency checks, but also—in case of differences—to draw key information on potential theories. For example, it is possible that some specific PN corrections are subleading, displaying a higher curvature scaling than the majority of the other PN corrections due to the underlying model under consideration (e.g., no dipole radiation for equal-mass objects in scalar-tensor theories). In all cases, different tests (e.g., PN phase and ringdown) should give compatible results. This idea can further be extended to multiparameter tests [321, 325].

Secondly, false deviations could be induced by missing physics [398], waveform systematics [41, 312], or detector glitches [399]. These effects often have specific mass-dependent behavior, e.g., $M^{-5/6}$ for eccentricity [314, 400] or large deviations only present for heavy binary masses due to glitches [98]. Therefore constraining the mass dependence would help distinguish between such systematics and a genuine GR deviation under the effective-field-theory framework.

Finally, the expectation that $p \geq 4$ suggests that the highest-curvature black holes, i.e., the *lightest* black holes, will yield the strongest constraints[3]. For example, a $\mathcal{O}(0.1)$ deviation constraint from a $10\,M_\odot$ binary is equivalent to a $\mathcal{O}(10^{-21})$ constraint from a $10^6\,M_\odot$ system if $p = 4$—the expectation for cubic or quadratic corrections with an additional degree of freedom. This suggests that ground-based GW detectors, including the next-generation Einstein Telescope [401, 47] and Cos-

[3]For non-vacuum systems, the expected value is $p = 2$ as quadratic curvature corrections would be the leading modification in the EFT action. However, such events represent a more difficult challenge for testing GR with lower observational rates and the introduction of matter effects.

mic Explorer [45, 46] detectors, will provide deeper probes of GR than observations of supermassive black holes with pulsar timing arrays [402, 403], the Event Horizon Telescope [404, 405] or LISA (beyond the extreme mass ratio regime) [406, 407]. Modeling the curvature dependence within these GW tests allows us to more deeply probe the fundamental nature of gravity and/or invalidate whole families of theories without resorting to theory-specific models.

Chapter 7

MODEL EXPLORATION IN GRAVITATIONAL-WAVE ASTRONOMY WITH THE MAXIMUM POPULATION LIKELIHOOD

E. Payne and E. Thrane. "Model exploration in gravitational-wave astronomy with the maximum population likelihood". In: *Phys. Rev. Res.* 5.2 (2023), p. 023013. DOI: 10.1103/PhysRevResearch.5.023013. arXiv: 2210.11641 [astro-ph.IM].

E.P. helped with the conception of the idea, developed and implemented all the of analysis methods, and led the writing of the manuscript.

7.1 Motivation

Bayesian inference has become a mainstay of modern scientific data analysis as a means of analysing signals in noisy observations. This procedure determines the posterior distributions for parameters given one or more model. In order to study the *population properties* of a set of uncertain observations, a hierarchical Bayesian framework can be employed. The basic idea is to model the population using a conditional prior $\pi(\theta|\Lambda, M)$, which describes, for example, the distribution of black hole masses $\{m_1, m_2\} \in \theta$ given some hyper-parameters Λ, which determine the shape of the prior distribution. Here, M denotes the choice of model. One then carries out Bayesian inference using a "population likelihood"

$$\mathcal{L}(d|\Lambda, M) = \prod_i^N \frac{1}{\xi(\Lambda)} \int \mathrm{d}\theta_i \, \mathcal{L}(d_i|\theta_i)\pi(\theta_i|\Lambda, M), \tag{7.1}$$

where $\mathcal{L}(d_i|\theta_i)$ is the likelihood for data associated with event i given parameters θ_i, and $\xi(\Lambda)$ is the detected fraction for a choice of hyper-parameters. Meanwhile, N is the total number of observations. For an overview of hierarchical modeling in gravitational-wave astronomy including selection effects, see Refs. [74, 277, 276].

The LIGO-Virgo-KAGRA (LVK) Collaboration's third gravitational-wave transient catalog (GWTC-3) [77] contains the cumulative set of observations of $N = 69$ confident binary black-hole mergers[1] detected by the LVK [1, 16, 362]. Additional detection candidates have been put forward by independent groups [179, 408, 409,

[1] We adopt the threshold utilized in [8] of a false-alarm-rate $< 1\,\mathrm{yr}^{-1}$.

410, 411]. Hierarchical inference is employed to study the population properties these merging binary black holes; see, e.g., Refs. [6, 7, 8, 278, 279, 280, 281, 195, 224, 412, 202, 174, 172, 282, 283, 284, 285, 286, 287]. These analyses have revealed a number of exciting results, such as the surprising excess rate of mergers with a primary black hole mass of $\sim 35\,M_\odot$ [7], and the evolution of the binary merger rate with redshift [8], to name just two.

However, Bayesian inference has its limitations. One can use Eq. (7.1) in order to infer the distribution of binary black hole parameters—*given some model*; and one can compare the marginal likelihoods of two models to see which one better describes the data. However, Bayesian inference does not tell us if any of the models we are using are suitable descriptions of the data. While all models for the distribution of binary black hole parameters are likely to be imperfect, some may be adequate for describing our current dataset[2]. When a model fails to capture some salient feature of the data, it is said to be "misspecified" [318, 319]. Some effort has been made to assess the suitability of gravitational-wave models, both qualitatively and quantitatively; see, e.g., [7, 8, 318, 413]. However, the idea of "model criticism"—testing the suitability of Bayesian models—is still being developed within the context of gravitational-wave astronomy and beyond.

Hierarchical Bayesian inference studies often depend upon parametric models. Modelers design parameterizations in order to capture the key features of the astrophysical distributions. However, one must still worry about "unknown unknowns"—features which do not occur to the modeler to add. For example, recent studies [7, 8, 297, 414] find a sub-population of binary black holes merge with spin vectors that are misaligned with respect to the orbital angular momentum axis. However, the degree to which the spins are misaligned might be model dependent. In Refs. [7, 8, 297], the inferred minimum spin tilt is confidently $\gtrsim 90^\circ$. In contrast, Refs. [278, 283, 414] argue this signature could be due to a lack of flexibility in LVK models to account for a sub-population of black holes with negligible spin magnitude, finding support for misalignment at smaller minimum tilt angles. The inferred population distribution of spin misalignment has important consequences for understanding the formation channels of binary black-hole channels. This debate highlights how astrophysical inferences can be affected by model design.

In order to help alleviate some of the issues arising from model misspecification in

[2]Here, we paraphrase the aphorism attributed to statistician, George Box: "all models are wrong, but some are useful."

Bayesian inference, we present a framework for assessing the suitability of a model. This framework is built around the concept of the *maximum population likelihood* $\mathpzc{L}$ (pronounced "L stroke")—the largest possible value of $\mathcal{L}(d|\Lambda)$ in Eq. (7.1), maximized over all possible choices of population model $\pi(\theta|\Lambda)$ *independent of the choice of parameterization*. The "prior" distribution, which yields this maximum is $\pi(\theta)$ (pronounced "pi stroke"). It is not a true prior because it is determined by the data. The theory behind the maximization of population likelihoods has been studied previously in optimization and statistics literature [415, 416, 417, 418, 419, 420]. This work is underpinned by Carathéodory's theorem [421] and the mathematics of convex hulls [419]. However, its application to observational science has been somewhat limited as far as we can tell.

The $\mathpzc{L}$ framework is useful for several reasons. First, the numerical value of $\mathpzc{L}$ is an upper bound on the population likelihood. We can compare the maximum likelihood for a specific model

$$\mathcal{L}_{\max}(M) = \max_{\Lambda \sim p(\Lambda|d)} \mathcal{L}(d|\Lambda, M) \tag{7.2}$$

to $\mathpzc{L}$. Often in Bayesian model selection, the Bayesian evidence values ($\mathcal{Z}_i$) of two hypotheses can be used to determine the extent to which one model is preferred over the other. A typical threshold chosen to rule out one model in favor of another is that $\ln(\mathcal{Z}_1/\mathcal{Z}_2) > 8$ [422]. In a similar vein, if $\ln(\mathpzc{L}/\mathcal{L}_{\max}(M)) \lesssim 8$, we can be sure the model M is not badly misspecified since there is no second model M' that can be written down with that will yield a statistically significant improvement. We emphasize that a model which does not satisfy this condition is not necessarily misspecified.

Second, the $\mathpzc{L}$ framework can be used to quantitatively assess if a model M is misspecified. By generating synthetic data from M, one can generate the expected distribution of $(\mathpzc{L}, \mathcal{L}_{\max}(M))$. In this paper, we show how one can compare the observed values of $(\mathpzc{L}, \mathcal{L}_{\max}(M))$ to the expected distribution in order to determine the extent to which M is misspecified—and the *way* in which it is misspecified.

Third, the $\mathpzc{L}$ framework can be used for "model exploration"—providing clues of *where* in parameter space unmodeled features might be lurking. By comparing $\pi(\theta)$ with the prior from our phenomenological model $\pi(\theta|M)$, one can see if the phenomenological model is capturing key structure present in π and use the comparison to design new models to test on forthcoming datasets.

The remainder of this paper is organized as follows. In Sec. 7.2, we introduce the $\mathcal{£}$ formalism, illustrating key features with a simple toy model. In Sec. 7.3, we show how the formalism can be used for model criticism. In Sec. 7.4, we apply the formalism to study the population properties of merging binary black holes observed by the LVK. Our concluding remarks are presented in Sec. 7.5.

7.2 The maximum population likelihood $\mathcal{£}$

Preliminaries

We begin with a brief review of Bayesian hierarchical inference with a parametric model. Our starting point is the population likelihood (copied here from Eq. (7.1)):

$$\mathcal{L}(d|\Lambda, M) = \prod_i^N \frac{1}{\xi(\Lambda)} \int \mathrm{d}\theta_i \, \mathcal{L}(d_i|\theta_i)\pi(\theta_i|\Lambda, M). \tag{7.3}$$

Here, $\mathcal{L}(d_i|\theta_i)$ is the likelihood of event-i data d_i given parameters θ_i. The quantity $\pi(\theta_i|\Lambda, M)$ is a conditional prior for θ_i given hyper-parameters for some population model M, which describes the shape of the prior distribution. The term $\xi(\Lambda)$ accounts for selection effects; for example, high-mass systems are typically easier to detect than low-mass systems. It is the detectable fraction of the population given the model given hyper-parameters Λ

$$\xi(\Lambda) = \int \mathrm{d}\theta \, p_{\mathrm{det}}(\theta)\pi(\theta|\Lambda, M). \tag{7.4}$$

Here, $p_{\mathrm{det}}(\theta)$ is the detection probability of an observation with parameters θ.

The maximum population likelihood $\mathcal{£}$

The maximum population likelihood $\mathcal{£}$ is obtained by taking Eq. (7.3) and maximizing over all possible prior distributions $\pi(\theta)$. Thus, $\mathcal{£}$ is an upper bound (or supremum) on the set of likelihoods from all possible choices of models for $\pi(\theta)$ such that

$$\mathcal{£} \equiv \mathcal{L}(d|\not{M}) \geq \mathcal{L}(d|\Lambda, M), \tag{7.5}$$

for all models M. The "prior" distribution that yields $\mathcal{£}$ is denoted

$$\not{\pi}(\theta) \tag{7.6}$$

(pronounced "pi stroke"). It is not a true prior because the distribution which maximizes the population likelihood in Eq. (7.3) depends on the data. One should

therefore refer to π as a pseudo-prior. The associated model is denoted $\not{M}$ (pronounced "M stroke"). Combining this notation into a single equation, we have

$$\not{\mathcal{L}} \equiv \prod_{i=1}^{N} \frac{1}{\xi(\not{M})} \int \mathrm{d}\theta_i \, \mathcal{L}(d_i|\theta_i)\pi(\theta_i). \tag{7.7}$$

Calculating π: special cases

Having introduced the concept of $\not{\mathcal{L}}$ and π, the natural next question is: given data d, how does one calculate these quantities? Before answering this question, we study three special cases where we can work out π from intuition. This discussion will help sharpen our instincts for the more general solution that follows. Readers looking to skip to the punchline may wish to skip this subsection.

A single measurement

For the first case, we consider a single measurement ($N = 1$) with a unimodal likelihood function $\mathcal{L}(d|\theta)$, which is maximal when the parameter θ is equal to the maximum likelihood value $\widehat{\theta}$. For the sake of simplicity, we ignore selection effects so that $\xi(\not{M}) = 1$. In this case, $\not{\mathcal{L}}$ in Eq. (7.7) is clearly maximized if the prior support is entirely concentrated at $\widehat{\theta}$. Thus, π is a delta function

$$\pi(\theta) = \delta(\theta - \widehat{\theta}), \tag{7.8}$$

which yields

$$\begin{aligned} \not{\mathcal{L}} &= \int \mathrm{d}\theta \, \mathcal{L}(d|\theta) \, \delta(\theta - \widehat{\theta}) \\ &= \mathcal{L}(d|\widehat{\theta}). \end{aligned} \tag{7.9}$$

This result is intuitive: the prior that maximizes the population likelihood is the one that concentrates all its support at the maximum-likelihood value of θ.

N signals in the high-SNR Limit

For the second case, we consider a scenario in which the data consists of N observations carried out in the high-SNR limit. In this limit, the likelihood of the data for each measurement d_i given some parameter θ approaches a delta function

$$\mathcal{L}(d_i|\theta_i) = \delta(\theta_i - \widehat{\theta}_i), \tag{7.10}$$

located at the maximum-likelihood value $\widehat{\theta}_i$. We assume that each measurement is distinct so that no two maximum-likelihood values $\widehat{\theta}_i$ are exactly the same. Again,

for the sake of simplicity, we ignore selection effects so that $\xi(M) = 1$, though, the argument here holds even if we relax this assumption. Equation (7.7) becomes

$$\mathcal{L} = \prod_{i=1}^{N} \int \mathrm{d}\theta_i \, \delta(\theta_i - \widehat{\theta}_i) \pi(\theta_i). \tag{7.11}$$

The population likelihood is maximized when π is a sum of delta functions peaking at the set of $\{\widehat{\theta}_i\}$:

$$\pi(\theta) = \sum_{k=1}^{N} w_k \, \delta(\theta - \widehat{\theta}_k) \tag{7.12}$$

$$w_k = 1/N. \tag{7.13}$$

This solution for π ensures that there is maximal prior support at every likelihood peak. Obviously, the population likelihood is not maximized if any prior probability density is wasted to values of θ where all the likelihood functions are zero. Choosing an equal weight for each delta function $w_i = 1/N$ produces the largest possible population likelihood[3].

We illustrate this case in Fig. 6.1(a) using high-SNR, toy-model data drawn from a mean-zero, unit-variance Gaussian distribution. In the top-panel, we plot the set of $N = 10$ maximum likelihood points $\{\widehat{\theta}_i\}$ and the position of the delta functions (blue). In the lower panel, we "plot" the $\pi(\theta)$ for these ten data points. We put the word "plot" in quotation marks because, technically, we are not plotting $\pi(\theta)$, which goes to infinity, but rather we are plotting the weights w_k (Eq. (7.13)), which allows us to see the relative weight given to each delta function—something that will prove useful below. Throughout the paper, when we refer to plots of $\pi(\theta)$, it should be understood that we are actually plotting *representations* of $\pi(\theta)$ using the weights w_k. Finally, note that each peak in the distribution of $\pi(\theta)$ matches up with one of the maximum likelihood points in the upper panel.

N identical measurements

For the third case, we consider a set of N observations. This time, we do not assume the high-SNR limit, but we assume that every measurement has the same maximum-likelihood value of $\widehat{\theta}$. This case is highly contrived—one does not typically work with multiple identical measurements—but the example is nonetheless helpful for illustrative purposes. In this case, the integral in Eq. (7.7) is maximized when the

[3]This is a well-known result known as the empirical distribution function [417].

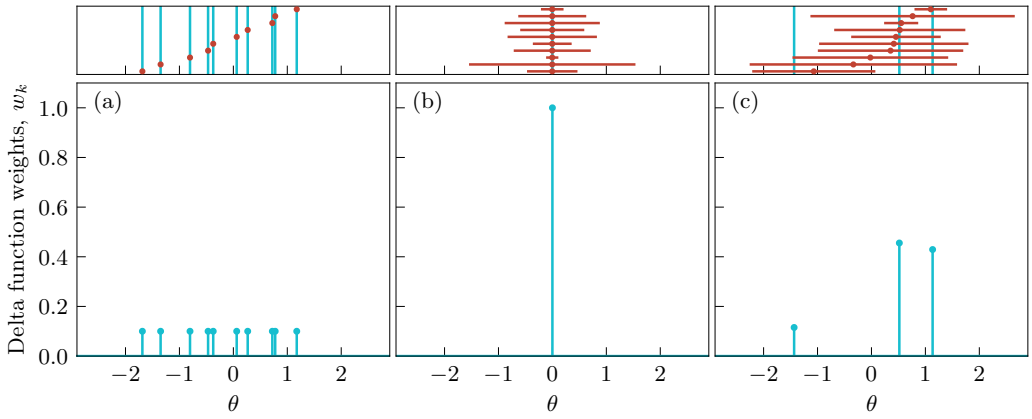

Figure 7.1: Examples of the distribution $\pi(\theta)$ described in Subsections 7.2-7.2. Each column represents a different dataset. The top-panel dots show the set of $N = 10$ maximum-likelihood estimates $\{\widehat{\theta}_i\}$. The top-panel horizontal lines represent error bars; (in the first column they are too small to see), and the vertical lines (blue) indicate the inferred delta function locations. The bottom panels show the distribution of $\pi(\theta)$ associated with each data set. The left-hand column (a) represents data in the high-SNR limit so that the likelihood functions for each measurement approach delta functions (this is why the error bars are not visible). In this case, $\pi(\theta)$ consists of N delta functions, each associated with one of the maximum likelihood points $\widehat{\theta}_i$. In the middle column (b), we are no longer in the high-SNR limit, but the maximum likelihood points are all assumed to be identical with $\widehat{\theta}_i = 0$. In this case, $\pi(\theta)$ consists of one delta function peaking at $\theta = 0$. In the right-hand column (c), the data are not in the high-SNR limit, and each $\widehat{\theta}_i$ is random. In this case, $\pi(\theta)$ consists of $n = 3$ delta functions, each with different heights.

prior support is entirely concentrated at $\widehat{\theta}$ (where all of the likelihood functions peak), so that π is a single delta function:

$$\pi(\theta) = \delta(\theta - \widehat{\theta}), \tag{7.14}$$

while

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{L}(d_i|\widehat{\theta}). \tag{7.15}$$

This scenario is demonstrated in Fig. 6.1(b). The top panel shows the set of $N = 10$ maximum-likelihood points $\{\widehat{\theta}_i\}$, all with the same value. The horizontal lines represent the error bars for each measurement, which we draw from a uniform distribution on the interval $(0.01, 1)$. In the lower panel, we plot $\pi(\theta)$ for these ten data points. This time, since every measurement is identical, $\pi(\theta)$ is a single delta function peaking at $\theta = 0$.

From these three examples, we observe a pattern: in each case, $\pi(\theta)$ can be written as a weighted sum of delta functions. Indeed, it has been proven that this is in fact the case [415, 416, 417, 418, 419, 420]. We refer readers interested in an explanation of the delta function structure of π to Appendix 7.6, where we summarize the key concepts surrounding the proof outlined in Ref. [419] using the mathematics of convex hulls. We do not reproduce the proof in its entirety, but rather we use visualisations to explain how it works with $N = 2$ observations, before providing a qualitative explanation for how it generalizes to arbitrary values of N. We explore this general structure and the consequences thereof in the next subsection.

The general form of π

We proceed with the knowledge that Eq. (7.7) is true in general, regardless of the form of the likelihood $\mathcal{L}(d|\theta)$ and the selection effect term $p_{\rm det}(\theta)$. *For any set of observations, $\pi(\theta)$ is always of the form,*

$$\pi(\theta) = \sum_{k=1}^{n} w_k \, \delta(\theta - \theta_k), \tag{7.16}$$

where w_k are weights which sum to unity

$$\sum_{k=1}^{n} w_k = 1. \tag{7.17}$$

The number of delta function is always less than or equal to the number of measurements and the solution is unique in all but the most pathological of cases (e.g., multimodal distributions with regions of equivalent maximum likelihoods) so that

$$n \leq N. \tag{7.18}$$

The ratio

$$\mathcal{I} \equiv n/N, \tag{7.19}$$

is a measure of the "informativeness" of the data. It compares the typical likelihood width to the scatter in the astrophysical distribution. In the high-SNR limit, $\mathcal{I} = 1$, since a delta function is required for every data point (see Fig. 6.1(a)). The other limiting case is, $\mathcal{I} = 1/N$, which happens when the likelihood for each measurement completely overlaps (see Fig. 7.1(b)).

Using this insight into the structure of $\pi(\theta)$, we now consider a variation on the toy-model problems discussed in the earlier subsections. In particular, we consider

finite-SNR data drawn from our Gaussian, toy-model distribution. Using Eqs. (7.16-7.17) as an ansatz, we calculate $\pi(\theta)$ for $N = 10$ random data points. The maximum likelihood values $\widehat{\theta}_i$ are drawn from a mean-zero, unit-variance Gaussian and the error bars are drawn from a uniform distribution on the interval $(0.01, 1)$. The results of this calculation are shown in Fig. 6.1(c). The top panel shows the data, represented by the maximum-likelihood values $\{\widehat{\theta}_i\}$, which are arranged from bottom to top in increasing order. The horizontal lines show the uncertainty for each measurement and the vertical blue lines indicate the positions of the delta functions. In the bottom panel, we show $\pi(\theta)$ for this dataset. It consists of just $n = 3$ delta functions of varying heights ($\mathcal{I} = 0.3$). The exact weights, locations, and number of delta functions are not obvious; we obtain them numerically by maximising Eq. (7.16) subject to Eq. (7.17) using the "combined" method described below in Subsection 7.2. Comparing the red data points with error bars to the turquoise representation of $\pi(\theta)$, one can see that every data point can be plausibly associated with at least one of the delta functions.

Given the form of $\pi(\theta)$ described by Eq. (7.16), we can write down a general expression for $\mathcal{L}$:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\xi(M)} \sum_{k=1}^{n} w_k \, \mathcal{L}(d_i|\theta_k), \tag{7.20}$$

where

$$\xi(M) = \sum_{k=1}^{n} w_k \, p_{\text{det}}(\theta_k). \tag{7.21}$$

Given Eqs. (7.20) and (7.21), the problem of calculating $\mathcal{L}, \pi$ reduces to the problem of simply finding the locations and weights of n delta functions. In Section 7.2, we explore three different approaches to this problem.

Computing π

In this subsection, we consider three techniques that can be applied to compute $\mathcal{L}, \pi$: optimization, iterative grid, and stochastic methods. We show that a combined approach, which uses a grid-based approach to guess a solution, which is subsequently refined through optimization performs the best out of the algorithms we tried. Meanwhile, the stochastic approach allows us to illustrate the existence of the delta function structure proven in Ref. [419], but with minimal assumptions.

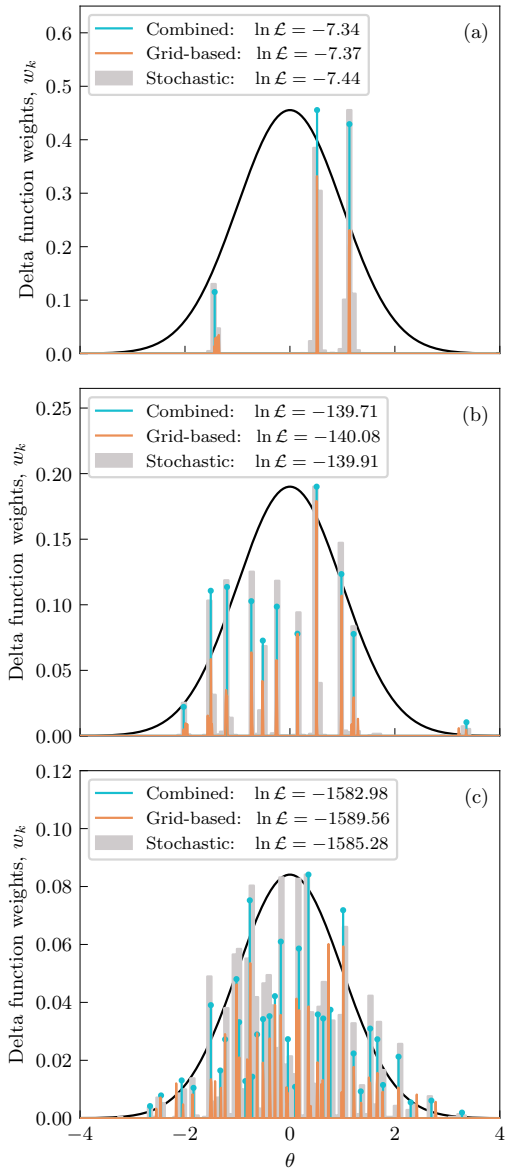

Figure 7.2: Demonstration of different methods for calculating $\pi, \mathcal{L}$. Each panel shows the results for a different number of measurements with (a) $N = 10$, (b) $N = 100$, and (c) $N = 1000$. The black distribution is the true distribution $\pi(\theta)$ used to generate the data. The colored spikes show the reconstructed distribution $\pi(\theta)$ as determined by different methods. Cyan is for the "combined" technique, which uses the iterative grid to obtain a first guess that is refined with the optimization method. Meanwhile, orange is for the grid-based technique by itself and gray is for the stochastic method.

Optimization

The first approach we consider is to use an optimization algorithm subject to the constraint in Eq. (7.17) [4]. We use SCIPY's `trust-constr` optimization implementation [423, 424]. We find this approach fails to find the correct global maximum of Eq. (7.20) once the number of peaks n becomes large. However, this issue can be resolved if a sufficiently close guess to the true shape of $\pi(\theta)$ can be made. Fortunately, the iterative-grid approach can be used to supply this initial guess.

Iterative grid

The second approach we consider is to iteratively place delta functions on a fixed grid. There are two steps: the greedy addition of many delta functions, and the removal of no-longer-useful delta functions. In the first step, we first attempt to place a delta function with a fixed height at each grid point and evaluate Eq. (7.20) (with appropriate normalization of the distribution). We determine which of all possible delta function additions produces the highest population likelihood. We then vary the height of this delta function between zero and twice the initial height in order to obtain an updated guess for $\pi(\theta)$. The addition of delta functions is repeated, reducing the initial height by a factor at each iteration. After many iterations, we then attempt to remove no-longer-useful delta functions to further increase the population likelihood. We repeat this procedure five times, iteratively adding 30 delta functions with varying heights at each iteration. After these iterations, $\mathcal{L}$ is usually well-converged for the problems we are studying. In some iterations, this procedure adds support to preexisting delta functions. This is how the approach "corrects" under-supported delta functions.

This method has a significant advantage over generic constrained optimization techniques as the procedure does not require the optimization of individual parameters governing the delta functions through the $\{\theta_k, w_k\}$ space. However, we find that this method is improved by pairing it with optimization. The most accurate optimization of the maximum population likelihood and structure of the distribution occurs when we utilize grid-based approximation to inform the starting location and weights

[4] In theory, the constraint condition does not need to be enforced during the analysis. The normalization appears in the selection function term and π. However, since any multiple of the weights (without normalization) would produce an identical likelihood, many numerical optimization methods can falter at these likelihood "plateaus". Therefore, we enforce the constraint to ensure a more robust analysis.

for the constrained optimization. This allows for the grid-based approximation to find the region of parameter space where $\mathcal{L}$ is nearly maximal. The constrained optimization then purifies the delta function structure and slightly increases the maximum population likelihood. The *combined* method is used for all the maximum population likelihood computations in Sec. 7.4.

Stochastic construction

Our final approach is to stochastically generate samples for $\pi(\theta)$, which are accepted/rejected depending on whether the new samples increases the population likelihood. This is a form of importance sampling in which an arbitrary "proposal distribution" is used to generate proposal samples. When a proposal sample is generated, we add it to a list of previously accepted points and evaluate $\mathcal{L}$ as a Monte Carlo integral,

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\xi(M)} \Big\langle \mathcal{L}(d_i|\theta_i) \Big\rangle_{\theta_i \sim \pi(\theta_i)}, \tag{7.22}$$

where

$$\xi(M) = \Big\langle p_{\rm det}(\theta) \Big\rangle_{\theta \sim \pi(\theta)}. \tag{7.23}$$

Here, the angled brackets indicate averaging over the samples. If the addition of the new sample increases $\mathcal{L}$, we retain the sample in the list of samples from π. As the process is repeated, the set of samples produces an ever-improving representation of π.

This method can be extended to employ an additional burn-in phase and/or a thinning phase to ensure more rapid convergence by removing unfavorable samples that sometimes get accepted early on before the distribution is well-converged. While this approach converges more slowly than the other two methods, *it does not employ any assumptions about the structure of the distribution*. Thus, this method can be used to validate the structure put forward in Eqs. (7.20-7.21), that $\pi(\theta)$ is a sum of delta functions.

Numerical study

We demonstrate each method using our Gaussian, toy-model distribution described in the last subsection: true maximum likelihood values $\widehat{\theta}_i$ drawn from zero-mean, unit-variance Gaussian with error bars drawn from a uniform distribution on the interval (0.01, 1). The observed maximum likelihood values are then shifted from the true value by an offset generated from each individual observation's uncertainty.

The results of this demonstration are compiled in Fig. 7.2. The three panels of Fig. 7.2 represent tests performed with $N = 10$, 100, and 1000 observations. In each panel, the black curve represents the true distribution $\pi(\theta)$. The colored spikes illustrate different numerical solutions for $\boldsymbol{\pi}(\theta)$: cyan is the "combined" approach, which uses the iterative grid to obtain an initial guess that is subsequently refined using the optimization method. Meanwhile, orange represents the iterative grid approach by itself. For the grid-based approach we run 30 iterations of adding peaks with variable but decreasing weights, before repeating this process an additional ten times. Finally, gray represents the stochastic approach. For the stochastic method, we generate 3000 samples with 1000 samples for burn-in.

We see that the combined approach better estimates $\mathcal{L}$ relative to the other techniques considered [5]. We observe that, as N increases, $\boldsymbol{\pi}(\theta)$ increasingly resembles the true Gaussian distribution $\pi(\theta)$ (shown in Fig. 7.2 as a black curve). To illustrate this more clearly, we take the inferred delta function locations from the $N = 1000$ "combined" result in Fig. 7.2(c) and compute the weighted histogram. This result is directly compared to the true distribution in Fig. 7.3, from which we see that indeed the inferred distribution is (albeit slowly) approaching the true distribution. *We conjecture that, in general, $\boldsymbol{\pi}(\theta)$ approaches the true distribution in the infinite-data limit:*

$$\lim_{N\to\infty} \boldsymbol{\pi}(\theta) \to \pi_{\text{true}}(\theta). \tag{7.24}$$

Computational challenges

Before continuing, we discuss two computational challenges. First, we note that the examples illustrative above are all one-dimensional. The discussion above generalizes to ≥ 2 dimensions; $\boldsymbol{\pi}(\theta)$ is still a sum of delta functions in ≥ 2 dimensions. However, it becomes increasingly challenging to determine the location and height of these peaks in higher dimensions. Furthermore, by increasing the dimensionality of the problem, constructing continuous representations of the individual-event likelihoods and the detection probability, $p_{\text{det}}(\theta)$, becomes increasing difficult. Recent developments in using Gaussian mixture models to produce continuous representations of these distributions might alleviate these concerns [425, 426]. Second, even

[5] A method is "better" if it yields a larger value of $\mathcal{L}$ than another approach.

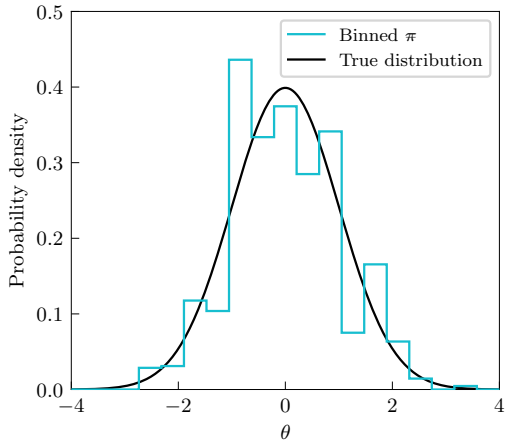

Figure 7.3: Comparison between a binned representation of π as computed for the toy model data set with $N = 1000$ observations and the true underlying population distribution. This representation more clearly shows that π is approaching the true distribution in the limit of many observations.

if we stay in one dimension, the computational cost of calculating π, $\mathcal{L}$ grows with N [6].

7.3 Model criticism with $\mathcal{L}$

In this section, we show how the $\mathcal{L}$ formalism can be used to determine if a model M is an adequate description of data. The first step is to generate synthetic datasets based on the posterior distribution for the model hyper-parameters $p(\Lambda|d)$. For each data set, we calculate the maximum population likelihood $\mathcal{L}$ (Eq. (7.7)) as well as the maximum likelihood for M, which we denote

$$\mathcal{L}_{\max}(M) = \max_{\Lambda \sim p(\Lambda|d)} \mathcal{L}(d|\Lambda, M), \tag{7.25}$$

where $\mathcal{L}(d|\Lambda, M)$ is the population likelihood defined in Eq. (7.3). In this way we can estimate

$$p(\mathcal{L}, \mathcal{L}_{\max}(M)), \tag{7.26}$$

the joint distribution for $\mathcal{L}$ and $\mathcal{L}_{\max}(M)$ given model M. By comparing the *measured* values of $(\mathcal{L}, \mathcal{L}_{\max}(M))$, to this distribution of *expected* values, one can see if the dataset is typical of what one would expect given M. If the measured values

[6] For the results in Fig. 7.2, the computation time of the "combined" approach was the following: 10 observations required only 5.3 seconds, 100 observations required 65 seconds and 10^3 observations required 2.78×10^3 seconds. Generally, more data tends to require more delta functions (each with a location and a height), meaning the computational difficulty grows with N.

of $(\mathcal{L}, \mathcal{L}_{\max}(M))$ are atypical, one can conclude that M is misspecified. Moreover, one may determine the nature of the misspecification by noting the location of the observed value of $(\mathcal{L}, \mathcal{L}_{\max}(M))$ relative to the typical values of $(\mathcal{L}, \mathcal{L}_{\max}(M))$. This is best illustrated with an example.

In our example, we imagine that an observer measures $N = 100$ values of some parameter θ. Their model M for the distribution of θ consists of a Gaussian distribution with mean $\mu = 0$ and width $\sigma = 1$:

$$\pi(\theta|M) \sim \mathcal{N}(\mu = 0, \sigma = 1). \tag{7.27}$$

However, their model may be misspecified so that θ is not really distributed according to M. We consider five "possible worlds" [7], one in which the observer's model is correctly specified and four in which it is not. Each world is assigned a color:

- Black: model is correctly specified $(\mu = 0, \sigma = 1)$.
- Purple: model is too wide because the true distribution is $(\mu = 0, \sigma = 0.6)$.
- Blue: model is too narrow because the true distribution is $(\mu = 0, \sigma = 1.4)$.
- Salmon: model is shifted to one side because the true distribution is $(\mu = 1, \sigma = 1)$.
- Yellow: model is too wide *and* shifted to one side because the true distribution is $(\mu = 0.8, \sigma = 0.6)$.

We create ten mock datasets for each of the five possible worlds (black, purple, blue, salmon, and yellow) and 5000 mock datasets from the model M (grey contours). For each dataset, we compute $(\mathcal{L}, \mathcal{L}_{\max}(M))$—always using model M (Eq. 7.27) even if the data are generated according to, say, the blue-world distribution. This is because we are studying the case where our observer might apply a misspecified model.

The results are shown in Fig. 7.4. The vertical axis is $\ln \mathcal{L}$ while the horizontal axis is $\ln \mathcal{L}_{\max}(M)$. The dark-grey region in the bottom-right corner is forbidden since $\mathcal{L} \geq \mathcal{L}_{\max}(M)$ by construction. The grey contours show the one, two, and three-sigma contours for the expected distribution from the model. Only the

[7] We borrow the language of "possible worlds" from the philosopher, David Lewis, who invokes them in his account of counterfactuals and necessity [427].

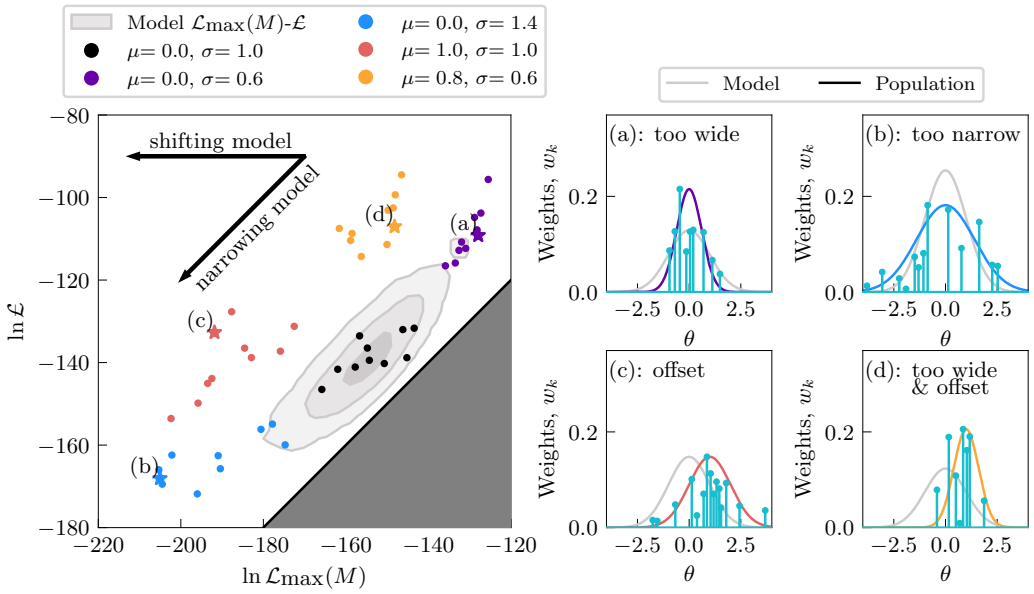

Figure 7.4: An illustration of model criticism with the $\mathcal{L}$ formalism. In the left-hand panel, we plot $(\mathcal{L}, \mathcal{L}_{\max}(M))$ for five different underlying populations (each with ten different realizations), analyzed a toy-model with a mean of $\mu = 0$ and standard deviation $\sigma = 1$. Each population is represented by a different color. The gray contours show the 1, 2, and 3-sigma credible intervals for the expected distribution of $p(\mathcal{L}, \mathcal{L}_{\max}(M))$ from the toy-model. By comparing the measured values of $(\mathcal{L}, \mathcal{L}_{\max}(M))$ from an observed population to the expected distribution from our choice of model, one may determine if the dataset is typical of what one would expect given the model. If the measured values of $(\mathcal{L}, \mathcal{L}_{\max}(M))$ fall outside these intervals, one may conclude that the toy-model is misspecified (does not accurately model the data). Moreover, the location of a point on this plot relative to the expected distribution, conveys information about the way in which a model is misspecified. The right-hand panel shows the toy-model (grey), the true population distribution for the starred and labeled datapoint (a-d), and the respective π for the observed data (turquoise). This demonstrates that shifts away from the expected distribution (left-hand panel; grey) in $(\mathcal{L}, \mathcal{L}_{\max}(M))$ can be visually identifiable to the reconstruction of π.

black world datasets are consistent with the expected distribution, as the model is correctly specified in the black world. The colored dots, meanwhile, show ten random realizations of $(\ln \mathcal{L}, \ln \mathcal{L}_{\max}(M))$ in colored worlds where the model is misspecified in various ways. This is fundamentally different from a typical Bayesian inference plot where the data are fixed and the model is varied. Here, the model is fixed to M (Eq. 7.27), and we consider different datasets, which may or may not be misspecified depending on the world of our observer.

When the model M is sufficiently misspecified with respect to the true distribution,

it becomes unlikely for our observer to obtain values of $(\pounds, \mathcal{L}_{\rm max}(M))$ that reside within the expected three-sigma interval—a sign of misspecification. Interestingly, the different colored dots cluster in different regions. For example, in the world where the model M is too broad (purple), the dots cluster above-right of the gray contours. In the world where the model M is shifted away from the true peak (salmon), the dots cluster to the left of the gray contours. By studying *where* one's observed values of $(\pounds, \mathcal{L}_{\rm max}(M))$ fall on this diagram, one can gain some insight into the way in which one's model is misspecified. This example focuses on relatively simple forms of misspecification involving the mean and variance. Other forms of misspecification (e.g., involving skewness and kurtosis) are, of course possible as well. Given all the ways that a model can be misspecified, the "shifting model" / "narrowing model" arrows on Fig. 7.4 should be taken as rule-of-thumb signposts.

In practice, it is computationally challenging to create plots like Fig. 7.4 for population studies in gravitational-wave astronomy. While it is easy to create mock datasets, it is time-consuming to calculate individual-event likelihoods for one dataset, let alone thousands. There may be workarounds. We discuss this possibility in greater detail below.

7.4 Application to gravitational-wave astronomy

In this section, we apply the $\pounds$ formalism to results from gravitational-wave astronomy to stress-test models for the population of merging binary black holes. We analyze data from the second gravitational-wave transient catalog (GWTC-3) [77, 428], which includes 69 confidently detected binary black hole mergers with false alarm rates $< 1\,{\rm yr}^{-1}$. To ensure similarity to the GWTC-3 LVK population analysis [8, 429], we utilize the same individual-event posterior samples—constructed from equally weighted samples generated from effective-one-body (SEOBNRv3 [430, 431], SEOBNRv4PHM [271, 94]) and phenomenological (IMRPhenomPv2 [107], IMRPhenomXPHM [241]) waveform results (see [8] for more details). To construct the lower-dimensional individual-event likelihoods, we utilize the same samples while marginalizing over all other "*nuisance*" parameters. For these "nuisance" parameters, we chose the distributions associated with the *maximum a posteriori* hyper-parameters from the LVK's GWTC-3 population analysis with the Power Law+Peak-Default-Power Law model [8].

We divide out the sampling prior to convert the one-dimensional posterior to a likelihood. The likelihood normalization is computed using the Bayesian evidence of

each event. The normalization is not important for the calculation of π, but it affects the misspecification tests demonstrated in Sec. 7.4. We calculate the hyperparmeter distributions and $\mathcal{L}_{\max}(M)$ using GWPopulation [432], which employs Bilby [36, 95] and Dynesty [162]. We utilize the combined injection set from Ref. [433] to compute the estimated detectable fraction of binary black-hole mergers over the first three observing runs.

Model inspiration through visual inspection

One straightforward application of the $\mathcal{L}$ formalism is to visually compare the reconstructed population distribution (obtained using a phenomenological model) with $\pi(\theta)$. By comparing these two distributions, it is possible to see which features in the phenomenological model reconstruction are due to prior assumptions, which features are due to real trends in the data, and which features might be missing from the phenomenological model. Formally, we compare $\pi(\theta)$ to the population predictive distribution (PPD)

$$\mathrm{PPD}(\theta|d, M) = \int \mathrm{d}\Lambda \, p(\Lambda|d)\pi(\theta|\Lambda, M), \tag{7.28}$$

which describes the astrophysical distribution of θ given a phenomenological model M with hyper-parameters Λ.

In Fig. 7.5, we present $\pi(\theta)$ with the PPDs from the LVK analysis of GWTC-3 [8, 429] for source-frame primary mass m_1 (top), the effective inspiral spin parameter χ_{eff} (middle), and redshift z (bottom). Each row contains two sub-panels; the small upper panel shows the maximum-likelihood estimate for each gravitational-wave event and the 90% confidence interval while the larger lower panel compares π with the PPD. The PPD is plotted as a thick band to show the 90% credibility region at each value of θ.

We first turn our attention to the primary mass distribution in the top row. There are $M = 10$ delta function peaks, implying an informativeness of $\mathcal{I} = 0.15$ (see Eq. (7.19)). This result is computed in 169.3 seconds. The gray band is the Power Law + Peak model from [280] while the orange band is a (more flexible) semi-parametric power-law-spline model denoted Spline from [285]. The agreement between π and the two PPDs is striking, with cyan spikes closely matching several of the features in both models including the turn-over at low masses near $\approx 12M_\odot$ and the bump at $30M_\odot$. Furthermore, we see that π also recovers some of the finer detail features found only by the Spline model. In particular, the shift in the low-mass peak and the dips in posterior support at $\sim 16\,M_\odot$ and $\sim 25\,M_\odot$ are present in

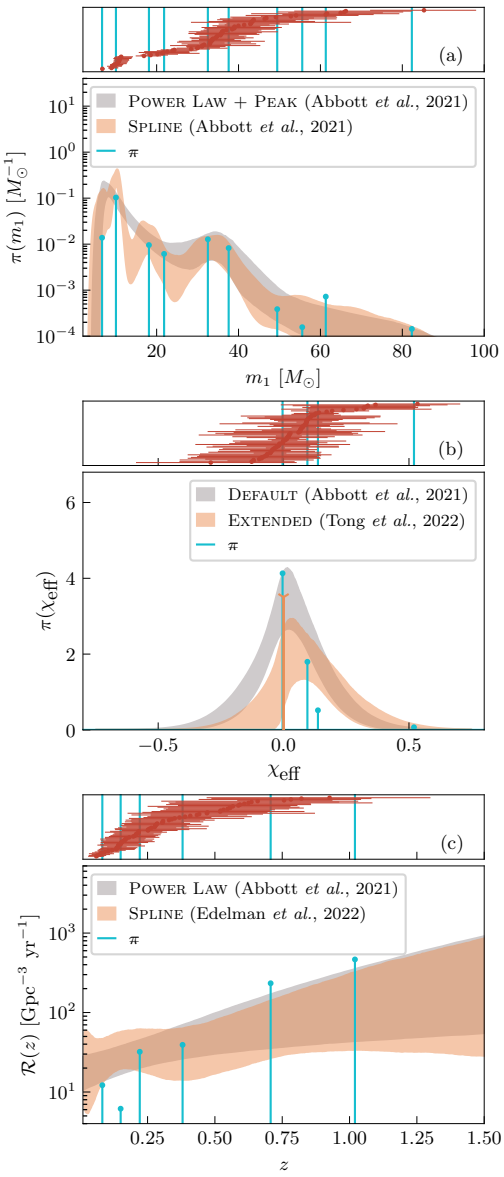

Figure 7.5: Population predictive distributions (90% credibility) and π for (a) the primary black-hole mass (m_1), (b) effective inspiral parameter ($\chi_{\rm eff}$), and (c) redshift (z) distributions. For the redshift, we divide by the evolution of the comoving volume and time delay as a function of redshift to plot the merger rate, $\mathcal{R}(z)$. Comparison of the different models with π highlights which features are present in the data and which are due to assumptions in the model.

the structure of π. Based on our visual inspection, it appears that current models are capturing much if not all of the structure present in π.

Turning our attention to the middle row, we study the distribution of effective inspiral

spin parameter [218],

$$\chi_{\rm eff} \equiv \frac{\chi_1 \cos\theta_1 + q\chi_2 \cos\theta_2}{1+q}, \tag{7.29}$$

which measures the mass-weighted black hole spin projected along the orbital angular momentum [8]. This time, only $n = 4$ delta function spikes are required to fit the data ($\mathcal{I} = 0.06$), showing how much harder it is to measure $\chi_{\rm eff}$ than m_1. Computing $\pi(\chi_{\rm eff})$ requires 71.3 seconds. The quicker computation time is likely a result of the lower number of delta functions required. In gray, we plot the PPD for the Default model from Refs. [7, 8], which draws on work from Refs. [281, 296]. In orange we plot the PPD for the Extended model from Refs. [283, 414], which only analyse 68 binary black-hole events in the population due to data quality concerns regarding one event [75]. To plot the Extended model results, which incorporates a delta function at $\chi_{\rm eff} = 0$, we plot the 90% interval for the delta function height, δ, multiplied by the same scale factor as π. The continuous contribution to the Extended model is then scaled by the ratio of the PPD evaluated at only the non-zero $\chi_{\rm eff}$ π delta functions to the previously computed scaling.

The data-driven π includes a delta function at $\chi_{\rm eff} \approx 0$ and three smaller peaks in the $\chi_{\rm eff} > 0$ region, but no peaks with $\chi_{\rm eff} < 0$. The lack of support for $\chi_{\rm eff} < 0$ is in contrast to Refs [7, 8], which find support for a sub-population of binary black holes with $\chi_{\rm eff} < 0$. The strong delta function at $\chi_{\rm eff} = 0$ lends support to the argument put forward in Refs. [282, 278, 283] that the data can be well-modeled with a sub-population of "non-spinning" $\chi_{\rm eff} = 0$ binaries, even if there is not strong statistical support for the existence of such a peak [297, 434, 414]. However, our visual comparison suggests that the Extended model may over-predict the abundance of binaries with $\chi_{\rm eff} \approx 0.3$. Moreover, we note that the distribution of $\chi_{\rm eff} = 0$ appears to also be consistent with a smooth, one-sided distribution, maximal at $\chi_{\rm eff} = 0$, and slowly decaying at larger positive values of $\chi_{\rm eff} = 0$—that is, a single population.

Turning our attention to the bottom row of Fig. 7.5, we consider the case of redshift. For this parameter, $n = 6$ ($\mathcal{I} = 0.09$), and takes 116 seconds to compute. Here we plot the merger rate as a function of redshift, $\mathcal{R}(z)$ by dividing the posterior predictive distribution by the PPD by the evolution of the comoving volume and time delay with respect to redshift. The merger rate is more commonly utilized for interpreting the redshift evolution. The π distribution fits a decrease in the merger

[8] In Eq. (7.29), $q \equiv m_2/m_1$ is mass ratio, $\chi_{1,2}$ are the dimensionless black hole spins, and $\theta_{1,2}$ are the spin vector tilt angles relative to the orbital angular momentum.

rate at a redshift of $z \sim 0.13$. While we caution that π is purely data-informed, and such a feature might diminish with additional observations, the POWER LAW model utilized in Refs. [7, 8] lacks the flexibility to resolve such a feature. Comparing our results to Ref. [286], we observe that π is qualitatively different from the "non-parametric" model [9] used in that paper. Our best guess is that the reconstruction from Ref. [286] is reasonable, and that the different features in π are due to noise fluctuations, though, it is possible that the smooth spline structure imposed by the [286] model is misspecified or that the prior on "knot location" is somehow subtly influencing the fit. As more gravitational-wave observations are made, finer structure may emerge in the redshift evolution of the binary merger rate. These differences between the parametric reconstructions and π might present the first hints of such structure. We suggest that future redshift models include additional flexibility to study the possibility of a deficit of mergers in the nearby Universe.

By using the iterative "grid-based" method (without further constrained optimization), we also demonstrate the computation of a two-dimensional π distribution. In particular, we study the joint distribution of mass ratio q and effective spin inspiral parameter $\chi_{\rm eff}$. Recent studies have explored the possibility of astrophysical correlations between q and $\chi_{\rm eff}$ [195, 200, 8], finding an anticorrelation, i.e. more unequal mass systems typically possess a effective spin inspiral parameter. The presence of an anticorrelation in the q-$\chi_{\rm eff}$ distribution has implications for the formation environments of binary black holes. Ref. [199], for example, propose that such an anticorrelation could be due to assembly of binary black holes in active galactic nuclei.

In Fig. 7.6 we plot $\pi(q, \chi_{\rm eff})$ as eight colored pixels. It is easier to digest this π plot than the superposition of single-event, 90% credible intervals for all 69 events (gray). In order to compare π to recent models, we plot the 90% contours of *maximum a posteriori* distribution estimates for the DEFAULT model in Ref. [8] which assumes no correlation (black curve), the CORRELATED model from Ref. [195] (blue curve) and the COPULA model from Ref. [200]. From visual examination of π , it is clear that the anticorrelation identified in Ref. [195] is based on actual features in the data: the pixels corresponding to the delta functions π are consistent with anticorrelation between $(q, \chi_{\rm eff})$. However, π is also consistent with they hypothesis that there are separate sub-populations located at different regions in the q-$\chi_{\rm eff}$ space (an instance of Simpson's reversal [435]).

[9] Ref. [286]'s spline model is probably better described as "ultra-parameterized".

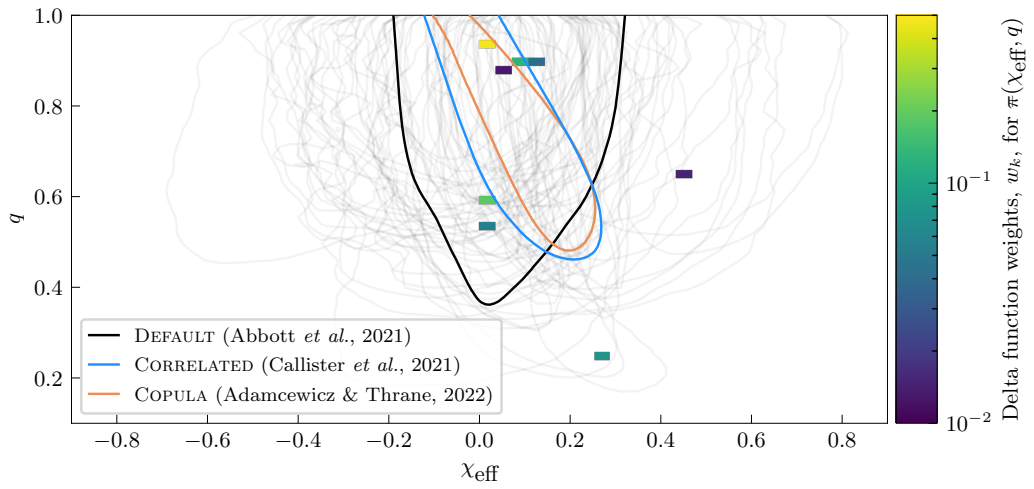

Figure 7.6: The joint distribution $\pi(q, \chi_{\rm eff})$ represented by eight colored pixels. The pixel color is related to the delta-function weight. The purely data-derived π can be compared to the 90% contours of *maximum a posteriori* distribution estimates for three specific models. The black curve shows the reconstructed population given the Default model from Ref. [8] (which does not allow for correlation) while the blue and orange curves show the reconstructed population given by the Correlated model from Ref. [195] and the Copula model from Ref. [200], respectively. The grey contours correspond to the 90% credible intervals of the 69 events in GWTC-3 [77, 8].

Upper bounds on population model likelihoods

In Table 7.1 we report the difference in natural log likelihood comparing the various population models to the maximum population likelihood $\mathcal{L}$:

$$\ln \mathcal{B} \equiv \ln \mathcal{L} - \ln \mathcal{L}_{\rm max}(M). \tag{7.30}$$

The $\ln \mathcal{B}$ values in Table 7.1 measure the fit of population models relative to the best possible fit. Motivated by the typical threshold for model selection in terms of Bayes factors [422], a value of $\ln \mathcal{B} \lesssim 8$ indicates that the population model is very close to the maximum population likelihood [74], which would imply that the fit cannot be dramatically improved. A large value of $\ln \mathcal{B}$ by itself does not imply that a model is "wrong" or unsuitable to describe the data, but it does quantify the extent to which an alternative model can in-principle improve over the current offerings.

Returning to Table 7.1, the Power Law + Peak model for m_1 shows the most potential room for improvement. This may be due to structure identified using the Spline model, which is missing from the less flexible Power Law + Peak. However, the m_1 measurements are also the most informative in Table 7.1 (with the

Parameter	$\mathcal{I}$	Model	$\ln \mathcal{B}$
m_1	0.15	Power Law + Peak	14.89
		Spline	6.66
$\chi_{\rm eff}$	0.06	Default	7.70
		Extended	3.53
z	0.09	Power Law	8.93
		Spline	6.59

Table 7.1: The performance of different population models relative to the $\mathscr{M}$. The quantity $\mathcal{B}$ (Eq. (7.30)) is a measure of the population likelihood of each model relative the maximum possible population likelihood $\mathcal{L}$. The "informativeness" $\mathcal{I}$ (Eq. (7.19)) is a measure of the information available about the distribution of each parameter.

largest value of $\mathcal{I}$). With more information, it is probably easier to concoct an *a posteriori* model with a large population likelihood that explains various features in the distribution of m_1 through over-fitting. The Default and Extended spin models both exhibit $\ln \mathcal{B} < 8$, which implies that neither model can be unequivocally ruled out, though, the Extended model provides a somewhat better fit with a natural log likelihood difference of 4.17. We also note that the $\chi_{\rm eff}$ and z observations are noticeably less informative, and simultaneously the associated values of $\mathcal{L}_{\rm max}(M)$ are closer to $\mathcal{L}$. This might indicate that, while there are features present in π that are present in the data, they are not statistically significant.

Model criticism in gravitational-wave astronomy

It would be interesting to make a version of the left-hand panel of Fig. 7.4 using the population models from gravitational-wave astronomy discussed in the previous subsection. Unfortunately, this is quite computationally difficult. First, we would need to run single-event parameter estimation of $N \approx 69$ events drawn from a random realization of the population fit to the observed gravitational-wave events. This needs to be repeated $O(1000)$ times to produce the refined contours as those shown in the toy-model example (Fig. 7.4). However, as an initial demonstration, we generate three simulated catalogs of 69 events using three draws from the Power Law + Peak - Default - Power Law hyperposterior informed by observations from GWTC-3 [8]. These simulated observations were produced with injections of the

IMRPhenomXPHM [241] waveform into simulated Gaussian noise colored by the power spectral density from the first half of the third LVK observing run.

We then run Bayesian hierarchical inference to determine the posterior predictive distributions from the parameterized model. Using the posterior predictive distributions, following the calculation undertaken for the collection of real gravitational-wave observations, we produce the one-dimensional marginal likelihoods which are then used to compute $\mathcal{L}$ and $\mathcal{L}_{\rm max}(M)$. Unlike in Sec. 7.4, where enough simulated catalogs are produced to construct an expected distribution in the $(\mathcal{L}, \mathcal{L}_{\rm max}(M))$ plane, here we are required to model and fit the distribution. We employ Bayesian inference and a simple multivariate Gaussian distribution model to estimate the structure in the expected $(\mathcal{L}, \mathcal{L}_{\rm max}(M))$ distribution. We use a Wishart prior on the covariance matrix [436]. We use the posterior predictive distribution of fitted Gaussian distributions to estimate whether the models utilized in Ref. [8] are inadequate for the observations.

The results are shown in Fig. 7.7 for the primary black-hole mass, effective inspiral parameter, and redshift. The blue dots correspond to the three simulated gravitational-wave catalogs, whereas the black star corresponds to the observed values from GWTC-3. The gray ellipses are 3σ intervals for $(\mathcal{L}, \mathcal{L}_{\rm max}(M))$, each associated with a different realisation of our Gaussian fit. (The large amount of scatter is due to the fact that we are attempting to fit a Gaussian to just three points.) The dashed blue curve corresponds to the *maximum a posteriori* (MAP) estimate. The value of $\mathcal{L}_{\rm max}(M)$ has been normalized to the value found for GWTC-3. The inferred points in $(\mathcal{L}, \mathcal{L}_{\rm max}(M))$ for GWTC-3 typically reside beyond the 3σ confidence interval, which we use as our criteria for misspecification.

We calculate a p-value for each panel, which quantifies the probability of observing the GWTC-3 values for $(\mathcal{L}, \mathcal{L}_{\rm max}(M))$ given our fit; small p-values are indicative of misspecification. For the Power Law + Peak primary black-hole mass model is misspecified we find $p = 47\%$, for the Default $\chi_{\rm eff}$ model we find $p = 44\%$, and for the redshift Power Law model we find $p = 10\%$. None of the models we consider are clearly ruled out as misspecified, as the sensitivity of this test is somewhat hamstrung by the small number of simulated catalogs. It would not surprise us if a more aggressive follow-up study $O(1000)$ simulations identified one or more models as more obviously misspecified.

One important caveat to these results is that the overall normalization of the likelihood depends on the computation of the individual observation Bayesian evidences.

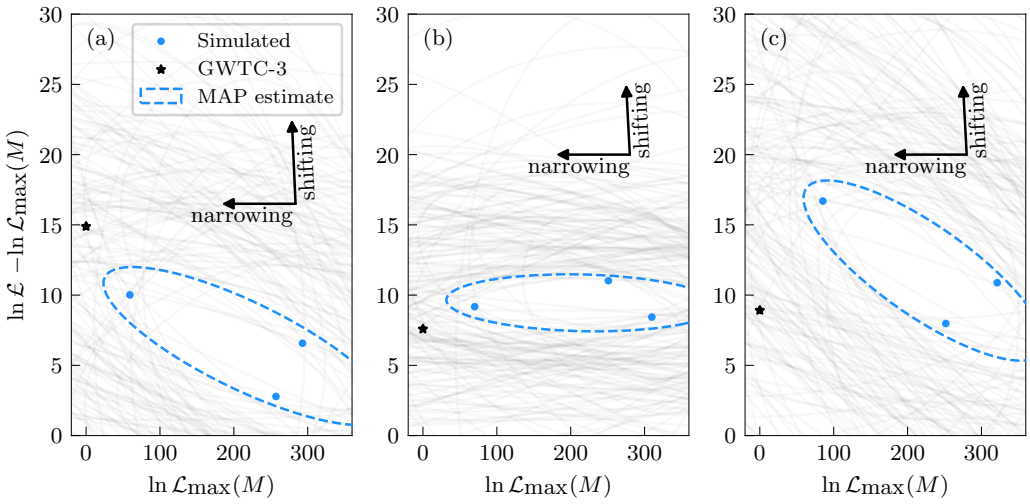

Figure 7.7: Demonstration of the $(\pounds, \mathcal{L}_{\max}(M))$ model misspecification test for three parameterized models used in Ref. [8]—(a) the Power Law + Peak model for the primary black-hole mass distribution, (b) the Default for the χ_{eff} distribution, and (c) the Power Law redshift distribution. Due to the limited number of simulated gravitational-wave catalogs, we model the expected distribution $p(\pounds, \mathcal{L}_{\max}(M))$ as a multivariate Gaussian distribution and infer the possible mean and covariance matrix from the three simulated values (blue). The grey ellipses correspond to the 3σ confidence intervals for 100 different realizations of the possible distribution. The dashed blue ellipses correspond to the *maximum a posteriori* (MAP) predictive distributions. The inferred values of $(\pounds, \mathcal{L}_{\max}(M))$ from the 69 events in GWTC-3 are shown by the black point. The likelihoods are normalized by the maximum likelihood inferred from the GWTC-3 model. From the inferred ellipses, we can conclude that there is a possibility that some or all models used are inadequate for the observations. Further studies with larger simulated catalogs are required to truly determine whether these models are misspecified.

With stark differences between the analyses made in Refs. [77, 8], it is difficult to accurately emulate the correct overall normalization of the likelihood. This globally impacts in the scale of $\mathcal{L}_{\max}(M)$ for the simulated catalog—potentially shifting the distributions closer or further from the inferred GWTC-3 result. In addition, the robustness of the evidences computed within Ref. [77] are not guaranteed (see e.g. Ref. [297]).

There are a number of solutions to address the computational cost of this analysis. While probably not realistic in the near future, it may be possible to represent the likelihood functions of simulated events using a Fisher matrix approximation, which would speed up the calculation significantly. However, verifying that this approximation produces adequately estimates for $\pounds, \mathcal{L}_{\max}(M)$ could remain a chal-

lenge. Another possibility worthy of investigation is the idea that the distribution of $\pounds, \mathcal{L}_{\rm max}(M)$ might have some quasi-universal properties. If it can be shown that a large class of problems produce a similarly-shaped distribution of $\pounds, \mathcal{L}_{\rm max}(M)$, perhaps a relatively small number of simulations can be used to work out the shape of $p(\pounds, \mathcal{L}_{\rm max}(M))$. We leave this for future work. Perhaps most promising are efforts to speed up inference with various machine learning schemes; see, e.g., Ref. [437]. As these tools become more reliable, it may become possible to estimate $(\pounds, \mathcal{L}_{\rm max}(M))$ in a matter of seconds, which would in turn enable precision tests of misspecification.

7.5 Conclusion

The $\pounds$ formalism provides a useful lens through which to view population studies in gravitational-wave astronomy. It provides an upper bound on the Bayesian evidence for population models, $\pounds$. The associated pseudo-prior distribution π is a sum of delta functions. The π distribution can be used to see which features in a reconstructed distribution are model-dependent, and which are genuinely present in the data. The π distribution can also draw attention to features in the data that are not fit by current models, providing a tool for the design of new models. Finally, the $\pounds$ formalism can be used to determine if a model is misspecified, by comparing the values of $(\pounds, \mathcal{L}_{\rm max}(M))$ to the expected distribution of these quantities given the model M. This comparison can be made quantitatively with a p-value. And, by comparing the measured values of $(\pounds, \mathcal{L}_{\rm max}(M))$ to the distribution expected given the model, it is possible to see the way in which the model is misspecified. Constructing a distribution of $\pounds, \mathcal{L}_{\rm max}(M)$ may be computationally prohibitive in gravitational-wave astronomy, though, future work is required to investigate simplifying assumptions that might bring down the cost.

While we have introduced the $\pounds$ formalism within the context of gravitational-wave astronomy, the framework is general, and we expect it can be applied to a broad range of problems in astronomy and beyond where one seeks to infer the distribution of parameters θ with potentially unreliable hierarchical models.

7.6 Appendix: Outline of π structure proof

Overview

In this appendix we outline the basic ideas underpinning the proof from Ref. [419] by Lindsay that π consists of a sum of $\leq N$ delta functions:

$$\pi(\theta) = \sum_{k=1}^{n} w_k \, \delta(\theta - \theta_k). \tag{7.31}$$

Our aim is to provide readers with a qualitative understanding. To this end, we consider a simple example of $N = 2$ measurements, each characterized by a Gaussian likelihood functions. Our example measurements are depicted in the right-hand column of Fig. 7.8, which shows two single-event likelihoods (one in purple, the other in red), both conditioned on some parameter θ. In each row of Fig. 7.8, we vary the separation of these two single-event likelihood functions relative to their width: far apart in the top row, becoming closer together in the two subsequent rows. We show below how π consists of either one or two delta functions, depending on this relative separation and explain how this generalizes to $N > 2$.

Lindsay's proof relies on the mathematics of *convex hulls*, geometric shapes which can be defined in arbitrarily high dimensions. If one draws a line between any two points on a convex hull, all the points on that line are also part of the hull. (The gray shaded regions in the left-hand column of Fig. 7.8 are all examples of convex hulls.) Convex hulls are often used in optimization problems with constraints where the optimal solution occurs on the boundary of the hull, which is determined by the constraints. In Lindsay's proof, the relevant constraint equation is the unitarity of the $\pi(\theta)$:

$$\int d\theta \, \pi(\theta) = 1. \tag{7.32}$$

The unitarity constraint means that the form of $\pi(\theta)$ that maximizes the population likelihood exists on the boundary of a complex hull.

A geometric picture

For the sake of simplicity, we ignore the impact of the selection function [10]. We represent the observations using what Lindsay refers to as an *atomic likelihood vector*,

$$L(\widehat{\theta}) \equiv \{\mathcal{L}(d_1|\widehat{\theta}), \mathcal{L}(d_2|\widehat{\theta}), ..., \mathcal{L}(d_N|\widehat{\theta})\}. \tag{7.33}$$

[10] The selection function term, $p_{\rm det}(\theta)$, can be absorbed into the prior to determine π on the observed population before correcting the detection probability afterwards.

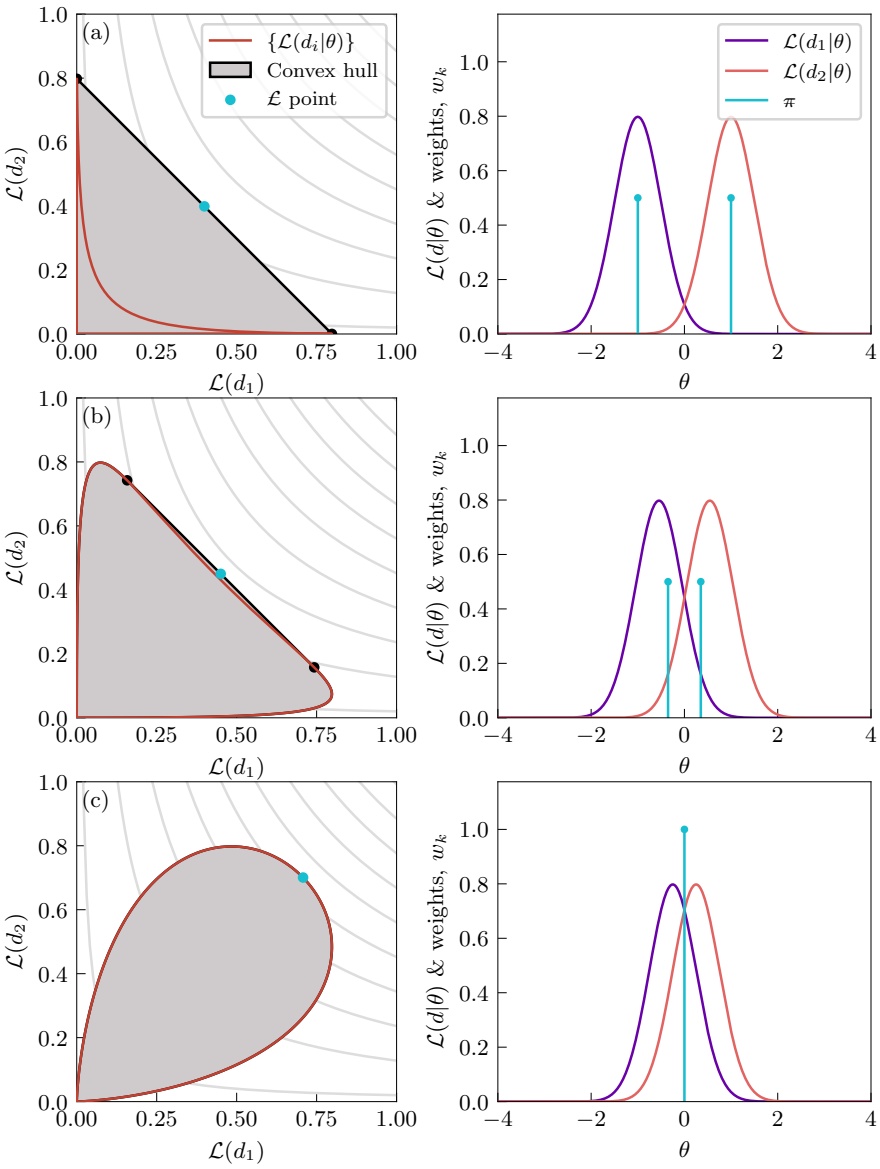

Figure 7.8: Visual illustrations of the proof in Ref. [419]. The left-hand column panels show the atomic likelihood vectors (red), the convex hull produced from the red curve (grey with black outline), and the cyan point on the convex-hull boundary with the maximum population likelihood $\mathcal{L}$. The black points correspond to the points from the set of atomic likelihood vectors which generate the maximum population likelihood. The right-hand column panels show three examples of $N = 2$ single-event likelihood functions (purple and red). The distribution of π is indicated with one or more cyan spikes. These spikes correspond to the $\mathcal{L}$ solution (cyan dot) in the corresponding left-hand panel. In (a), the two single-event likelihoods are mostly disjoint and so two delta functions are required to maximize the population likelihood (cf. Fig. 1 in Ref. [419]). As the two single-event likelihoods begin to overlap further, these two delta functions move closer together as shown in (b). Moving the single-event likelihoods closer still, the set of atomic likelihood vectors becomes the boundary of the convex hull, at which point only one delta function is required to maximize the likelihood as shown in (c).

Each element of this vector is a single-event likelihood marginalised over a delta-function prior peaking at $\widehat{\theta}$:

$$\mathcal{L}(d_i|\widehat{\theta}) = \int d\theta_i \, \mathcal{L}(d_i|\theta_i) \, \delta(\theta - \widehat{\theta}). \tag{7.34}$$

This allows us to represent the problem in an abstract N-dimensional likelihood space. The left-hand column of Fig. 7.8 provides a visualization of such a two-dimensional atomic likelihood vector space. Scanning over all possible values of $\widehat{\theta}$ traces out the red curve in the atomic likelihood vector space, which represents all possible values of the atomic likelihood vector $L(\theta)$. By varying $\widehat{\theta}$, we can make an individual element of the atomic likelihood vector large, but doing may make other elements of the vector small as we see in the top row with widely separated single-event likelihood functions.

The weighted sum of atomic likelihood vectors

$$L(\vec{w}) = \sum_k w_k \, L(\widehat{\theta}_k) \tag{7.35}$$

yields a vector of likelihoods with elements

$$\mathcal{L}(d_i|\vec{w}) = \sum_k w_k \, \mathcal{L}(d_i|\widehat{\theta}_k), \tag{7.36}$$

corresponding to the marginal likelihood given a prior of delta functions

$$\pi(\theta) = \sum_k w_k \, \delta(\theta - \widehat{\theta}_k), \tag{7.37}$$

where

$$\sum_k w_k = 1. \tag{7.38}$$

This means we can construct more general *marginal likelihood vectors* with a linear combination of atomic vectors. Furthermore, in the continuum limit, *any* prior can be used to *marginalize* over the atomic likelihood vectors. Elements of the marginal likelihood vector in the continuum limit take the form,

$$\mathcal{L}(d_i|M) = \int d\widehat{\theta}_i \, \mathcal{L}(d_i|\widehat{\theta}_i) \, \pi(\widehat{\theta}_i|M). \tag{7.39}$$

Let us consider again the $N = 2$ example illustrated in Fig. 7.8. If we pick any two points on the red curve, each corresponding to some value of $\widehat{\theta}$, which we denote

A and B, we can define two basis vectors: $\hat{e}_A$ and $\hat{e}_B$. The linear combinations of these two basis vectors forms a line connecting A and B. All of the points along this line represent likelihood vectors constructed from $N = 2$ delta functions. By connecting together every possible pair of points on the red atomic likelihood points, we map out the gray region—the convex hull. Every possible marginal likelihood vector (for *any* choice of prior) is part of the hull. That is, the set of all possible summations is the convex hull and is a representation of all possible probability distributions in the likelihood space. This result is profound—our original problem is reduced from an infinite set of possible population distributions to a closed region in an N-dimensional likelihood space. The construction of the convex hull is unique [419], except in pathological cases further discussed in Sec. 7.6.

Now that we have studied the geometry of the atomic likelihood vector space, we ask the question: what point in our convex hull corresponds to the maximum population likelihood? The population likelihood can be written as a product of the marginal likelihood vector elements:

$$\mathcal{L}_{\rm pop}(\vec{d}|M) = \prod_{i=1}^{N} \mathcal{L}(d_i|M). \tag{7.40}$$

In $N = 2$ dimensions, we can fix $\mathcal{L}_{\rm pop}(\vec{d})$ and identify hyperbolic curves of the form

$$\mathcal{L}(d_2) = \mathcal{L}(\vec{d})/\mathcal{L}(d_1), \tag{7.41}$$

represented in the left-hand column of Fig. 7.8 by gray curves. All the points on one of these curves have the same population likelihood. If we jump up and to the right from one gray curve to another, the population likelihood increases. These constant-population-likelihood, hyperbolic curves do not depend on any population model. The population likelihood is then maximized by finding the point on the boundary of the hull tangent to the gray curve with the largest population likelihood (the most up-and-to-right gray curve). In general, the maximum population likelihood point lies on the boundary of the hull [438, 419]. Our maximization problem can therefore be rewritten as a geometry problem.

We now turn our attention to the different rows of Fig. 7.8. In the top row, the two single-event likelihoods (right) are widely separated. The cyan dot on the left-hand plot shows the maximum population likelihood point on the surface of the hull. This is where the population likelihood has a value of $\pounds$. It falls on a straight black surface of the hull, but not on the red atomic likelihood vector curve. This means

that the cyan point is a linear combination of two atomic likelihood vectors, which are indicated by the two black points (cf. Fig. 1 in Ref. [419]). Thus, the maximum population likelihood solution consists of two delta functions, each corresponding to a different atomic vector. This linear combination of delta functions is shown in the right-hand panel with cyan spikes. Unsurprisingly, they coincide with the two single-event likelihood function peaks.

Moving down to the second row, the single-event likelihood functions (right) are now closer together. The shape of the hull changes accordingly (left). The hull boundary point that maximizes the population likelihood still does not fall on the red curve of atomic vectors. Again, it is a linear combination of two black points. However, since the shape of the hull has changed, the black points have moved relative to the top row. The corresponding delta function spikes (right) therefore shift toward $\theta = 0$ and no longer correspond to the maximimum likelihood points of the single-event likelihoods.

In the bottom row, the single-event likelihood functions (right) are closer still. The hull (left) has now changed shape so that the cyan point marking the maximum population likelihood falls on the red curve denoting the set of atomic vectors (left). This means that the likelihood can be maximized with a single delta function at $\theta = 0$ (right). In each case (and almost all scenarios, see Sec. 7.6) the convex hull is unique, and so the cyan point of maximum population likelihood is unique as well. In all but the most pathological cases, Carathéodory's theorem [421, 439] states that all points on the boundary of a convex hull can be constructed by, at most, N points that were used to initially construct the hull (in our problem these are the atomic likelihood vectors). The relative weight of each delta function corresponds to the position along the boundary of the hull [419]. Thus, the population prior corresponding to the maximum population likelihood is a construction of a finite set of, at most, N delta functions.

The transition from two delta functions to one delta function occurs when the red curve passes through the black one (when the set of atomic likelihood vectors becomes convex). During this transition, the cyan point changes from residing on a straight line connecting two atomic vectors to residing on a single atomic vector point. This picture generalizes to higher dimensions. Solutions with three delta functions (which can only exist when $N \geq 3$) reside on two-dimensional planes. Solutions with four delta delta functions (which can only exist when $N \geq 4$) reside on three-dimensional hyper-planes. And so on.

Pathological cases

While we see that the maximum population likelihood almost always corresponds to a finite, unique set of N or fewer delta functions, there are pathological cases (not likely to come up in real-world data analysis) where this is not the case. Such cases stem from the maximum population likelihood point not being unique. So while the maximum population likelihood point is still found, multiple distributions can map to the same point in likelihood space. This requires artificial degeneracies in the measurements. In Fig. 7.9, we demonstrate one such example with two likelihood functions perfectly symmetric about $\theta = 0$ and one of which is bimodal. In the likelihood space, the $\mathcal{L}$ point corresponds to two possible positions of the delta function. However, unlike in Fig. 7.8(a) where the two possible delta function positions are separated, here they correspond to same point in likelihood space. Therefore, any normalized combination of the two delta functions produces the maximum population likelihood. This is emphasized by the dashed blue lines in the right column of Fig. 7.9(a), indicating that any combination of the two delta functions here is a permissible solution. However, we emphasize that this pathology arises from an artificial degeneracy, which is immediately broken if the likelihood functions are not precisely symmetric as demonstrated in Fig. 7.9(b). Other, even more pathological, situations can be constructed where infinitely many atomic likelihood vectors reside at the maximum population likelihood point, allowing for arbitrarily structured π distributions. However, all such situations require regions of perfectly uniform likelihood functions, which we do not expect in realistic observations—at least, not in gravitational-wave astronomy.

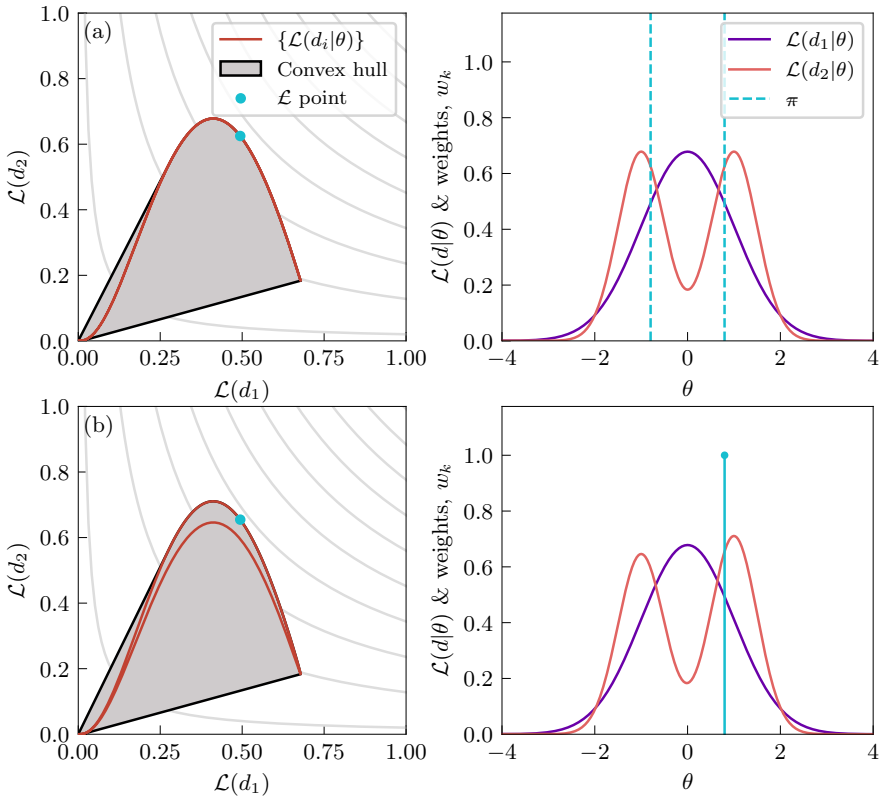

Figure 7.9: Demonstration of a pathological failure of the uniqueness of π. This occurs when multiple distributions map to exactly the same point on the convex hull. In (a), a perfectly symmetric, bimodal single-event likelihood has two delta functions with produce the same population likelihood. Therefore, any combination of the two is a valid π. However, such perfectly symmetric multi-modal distributions do not typically occur in gravitational-wave data analysis. We see here we can break this degeneracy by only slightly breaking the symmetry, shown in (b).

Chapter 8

NEUTRON STAR POST-MERGER GRAVITATIONAL-WAVE INFERENCE WITH PHOTON COUNTING READOUT SCHEMES

8.1 Motivation

An important scientific goal for future generations of gravitational-wave detectors is the study of dense nuclear matter through the detection of binary neutron-star (BNS) mergers and the characterization of their post-merger remnant through gravitational-wave radiation [89, 45, 90]. Since binary neutron stars exhibit densities beyond the nuclear saturation density, they exist as astrophysical laboratories for which no terrestrial experiment or astrophysical object can emulate. Therefore, direct observations of binary neutron stars and their post-merger signals stand to provide provide unprecedented insights into the dense matter equation-of-state [440, 441, 442] (including the possibility of phase transitions [443, 444]), the nature of hyper- and supramassive neutron stars [445], and the production of electromagnetic radiation from BNS merger remnants [446, 447].

Post-merger signals present a rich scientific opportunity, though detecting these signals with future generations of gravitational-wave detectors remains a significant challenge when using standard interferometric methods [45, 448]. Redshifted post-merger signals typically fall within the $\sim 500\,\mathrm{Hz}$ to $4\,\mathrm{kHz}$ frequency range, though the most detectable post-merger signals will fall between $\sim 1.5\,\mathrm{kHz}$ to $4\,\mathrm{kHz}$; see Fig. 8.1 from Ref. [449] for examples of Fourier-domain strains from BNS post-mergers described by different equations-of-state. For ground-based detectors such as LIGO [1], Virgo [16], KAGRA [362] and future observatories like the Cosmic Explorer (CE) [45] and Einstein Telescope [47], sensitivity at these frequencies is primarily limited by quantum measurement noise that results from photon shot noise on the interferometer fringe light that is recorded [91]. Modulations of that light are calibrated into a strain time-series [31, 33]. The quantum shot noise is now being reduced by almost 6 dB by injecting squeezed light (see Fig. 17 from Ref. [17] for current sensitivity improvements in O4 due to squeezing) [450, 451, 17], and future observatories plan to achieve 10 dB of quantum noise power suppression [45, 47].

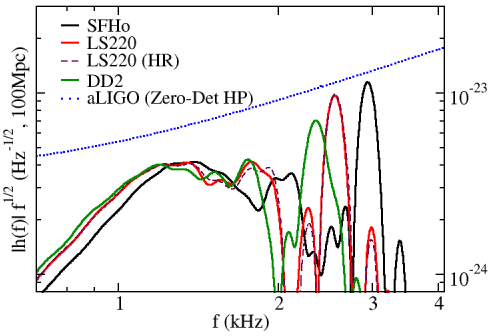

Figure 8.1: Fig. 5 from Ref. [449]. The post-merger strain amplitude for a source at a distance of 100 Mpc multiplied by $f^{1/2}$, is shown for three different equations-of-state. These different equations-of-state present different fundamental frequency peaks that may be resolved. The dashed blue curve indicates the design sensitivity of aLIGO [1].

However, even with improvements in squeezing, third-generation detectors like CE are only expected to detect post-merger signals with a signal-to-noise ratio (SNR) greater than five at a rate of approximately one per year [45] (also see the upper panel of Fig. 8.7 for a summary of the SNR distribution from 10^4 observations). Post-merger waveform signals with SNRs below this threshold result in almost entirely uninformative posteriors [440]. Given that CE is projected to observe hundreds of thousands of BNS mergers annually [452, 453, 454], this threshold results in a low post-merger detection rate and prevents the complete utilization of future detector designs towards their scientific goals. Future observatories are assumed to operate using the same measurement process as existing detectors, where the interferometer generates a time-series by recording modulations fringe light. In quantum measurement literature, this measurement process is known as homodyne readout [455]. The sensitivity of Gravitational-Wave interferometers with this readout is quantified through the power spectral density of their strain readout channel, where the total density is the sum of quantum and classical noise power contributions. The time-series data product and quantum shot noise are both a result of using homodyne readout as the typical measurement process for interferometer observatories.

The quantum noise contribution is unlike the classical noise contributions, because its impact can be influenced by the choice of quantum measurement process. Alternate quantum measurements on the output of interferometers are physically possible and modify the influence of both signals and noise on data distributions. Alter-

nate readout methods are still limited by randomness from quantum mechanical indeterminacy, but can have more favorable statistical figures of merit than that assumed of additive quantum shot noise power, even with squeezing applied [456, 457, 458]. For this work, we assume the eventual existence of hardware that performs matched-template computations directly on the electro-magnetic field emitted from an interferometer, and we choose a quantum measurement of power rather than of amplitude. Power measurements are *discretized into photon counts* [92]. We present the statistical framework and observational methodology required to predict and interpret post-merger detections using photon counting techniques, and we compare it to the standard time-series readout that provides a continuous measure of electromagnetic amplitude through homodyne readout.

A heuristic argument why photon counting may benefit the observation of a post-merger signal follows. If post-merger inference requires marginalization of nuisance parameters such as the difficult-to-predict ringdown waveform phase, then inference becomes similar to an excess power search. To detect excess power of an event population below detection threshold requires a number of events that scales inversely to the square of the total noise power spectral density. If photon counting is used, one can compute that the rate that signal waveforms emit photons scales inversely to the power spectrum of shot noise, due to common calibration factors between the emission rate and noise process [92, 459]. In the regime that classical noise is small and emits many fewer photons per event than signals, even a single photon detection event provides a significant detection. The detection rate for counting thus has a better scaling than excess power search with homodyne readout. In reality, the rate-constants are important, the classical noise is non-zero, and the complete inference process must be compared to determine which method is superior. The potential for improved detection and complexity of this statistical argument motivates our work to implement a fully Bayesian analysis and directly compare quantum measurement methods for this important astrophysics science goal.

In this manuscript, we explore the possibility of using a photon counting readout in-depth. In Sec. 8.2, we outline the photon counting methodology and its translation into the standard Bayesian formulation that has become the gold-standard for inference techniques in gravitational-wave astronomy. In Sec. 8.3, we then present the scientific value of photon counting for individual post-merger signals and address some of the quirks that manifest in inferences which rely on this discrete observational data, as opposed to the more continuous nature of homodyne readout-

based observations of the gravitational-wave detector's strain. We find that photon counting measurements are able to gain meaningful information from post-merger observations with SNRs$\lesssim 1$. Then, in Sec. 8.4 we demonstrate the resolving power of photon counting to place meaningful constraints on equation-of-state parameters through observations of the post-merger. We find that, in a hierarchical context, photon counting will perform as well as a homodyne readout with 10 dB of squeezing. Furthermore, a reduction of classical noise by an order of magnitude allows for significantly greater constraining power. Implications of the construction of such an readout scheme and concluding remarks are presented in Sec. 8.5. For the remainder of the manuscript, while this analysis can be extended to multiple detectors, we will focus on the implementation of such a method in a single CE interferometer.

8.2 Photon Counting for Post-merger detection and inference

In this section, we present the statistical framework and physical instrumentation elements required to realize photon counting as an alternative to the standard time-series readout of homodyne detection. We first summarize the experimental setup and statistical properties of the more familiar homodyne readout in Sec. 8.2. We then detail the photon counting readout to detect and measure discretized signals from simulated BNS post-merger gravitational-wave signals in Sec. 8.2, as well as how the underlying distributions follow from the measured photons from signal and noise sources. Finally, in Sec. 8.2, we outline the full likelihood function describing the photon counting measurement scheme.

Summary of the homodyne readout

Here we provide simple expressions that describe the physical process of detecting waveforms with an interferometer. This helps establish the statistical framework for the standard homodyne readout [460], which we then translate to the framework of the less familiar photon counting approach.

Michelson interferometers are generally described as directing laser light to a beamsplitter, where it splits and then simultaneously travels down both arms to accumulate a phase shift, reflects at respective end-mirrors, and then is recombined upon a second pass through the beamsplitter. The combined fields are emitted into constructive and destructive interference outputs. The differential phase accumulation that car-

ries strain waveform information is encoded in the optical field $E_h(t)$ and measured from the modulations of an otherwise-constant fringe light power at the destructive interference output. The output field is expressed as a sum of the time varying waveform component and the static fringe field,

$$E_{\text{out}}(t) = E_{\text{h}}(t) + E_{\text{fringe}}. \tag{8.1}$$

The output-port power that is monitored is then

$$P_{\text{out}}(t) = |E_{\text{out}}(t)|^2 = P_{\text{fringe}} + \delta P(t), \tag{8.2}$$

where

$$\delta P(t) = 2\,\text{Re}\{E_{\text{h}}(t) \cdot E_{\text{fringe}}\} + P_q(t). \tag{8.3}$$

In this expression, $\delta P(t)$ purely captures the time-varying component (ignoring the minuscule $|E_h(t)|^2$ contribution). $P_q(t)$ represents the mean-0 quantum shot noise fluctuation term, discussed below. The optical field, $E_h(t)$, encodes the differential length signal of the interferometer which includes the strain waveform, $h(t)$ as well as a noise contribution from classical noise sources, $n(t)$. Classical noise source encompass almost all sources such as seismic, thermal, electronics noise [34, 17]—all noise sources that aren't inherent to the nature of the readout scheme itself.

Both the waveform and classical noise are calibrated from their units of strain into optical field by applying an optical detector gain, $g(f)$, expressed in the frequency domain as a linear time-invariant filter. Gravitational wave detectors use recycling cavities optimized for a detection bandwidth of ΔF_{det} (typically around 450 Hz [1]) and circulating arm power P_{arm} (around 400 kW in O4 LIGO [17] to 1.5 MW in future detectors [45, 47]). An idealized but accurate expression for the optical calibration function is [461]

$$g(f) = ik\sqrt{\frac{cLP_{\text{arm}}}{\pi\Delta F_{\text{det}}}}\left(1 + \frac{if}{\Delta F_{\text{det}}}\right)^{-1}, \tag{8.4}$$

from which we can define the sensing function,

$$C(f) = 2g(f)\sqrt{P_{\text{fringe}}}. \tag{8.5}$$

In the expression for the optical gain, k is the wavenumber of the interferometer's source laser, $2\pi/1064\,\text{nm}$ for LIGO and Cosmic Explorer [1, 45]. The factor of i on the calibration applies the convention that signals are imprinted in the phase component of the light. It makes $E_h(t)$ purely imaginary and E_{fringe} is also purely

imaginary in this convention. Applying the calibration gives the frequency domain expressions for the emitted optical field that carries the signal and classical noise components,

$$E_h(f) = g(f)(h(f) + n(f)). \tag{8.6}$$

We can directly apply Eq. (8.3) to write

$$\delta P(f) = C(f)(h(f) + n(f) + q(f)). \tag{8.7}$$

The new term $q(f)$ represents the quantum noise process in units of strain. Specifically, $P_q(f) = C(f)q(f)$. The classical noise, $n(f)$, is assumed to be mean-0, and its power spectral density obeys the relation

$$\langle n(f)n^*(f')|n(f)n^*(f')\rangle = S_n(f)\delta(f - f')/2. \tag{8.8}$$

Note that the angled brackets indicate the time-averaged expectation over many realizations. The classical noise spectrum $S_n(f)$ will be used extensively with photon counting, even though it is much smaller than the quantum noise process at high frequencies where post-merger waveforms are most informative.

The quantum noise process imposes a noise power on the measurement of the fringe light into photons,

$$S_{P_q}(f) = 2\hbar kcP_{\text{fringe}} \cdot 10^{-\text{dB}_{\text{sqz}}/10}. \tag{8.9}$$

This spectral density represents the white-noise fluctuations of photo-power at the readout detector. Improved (lower) noise from injecting squeezed states can be applied through the parameter dB_{sqz} in decibels of observed quantum noise power reduction. Without squeezing ($\text{dB}_{\text{sqz}} = 0$), this spectral density represents shot noise from Poissonian statistical fluctuations in the arrival of photons from the fringe light.

The shot noise power spectral density (PSD) calibrated in units of strain is given by $S_q(f)$ with the expression [451],

$$S_q(f) = \frac{S_{P_q}(f)}{|C(f)|^2} = \frac{\hbar kc}{2|g(f)|^2} \cdot 10^{-\text{dB}_{\text{sqz}}/10}. \tag{8.10}$$

Note that the shot noise, after calibration, no longer depends on the specific level of P_{fringe}, and only depends on the amount of power in the interferometer arms, the interferometer bandwidth, and the amount of observed squeezing. Although shot noise is Poissonian in nature, the large number of photons in P_{fringe} allows for it to be well-approximated as a Gaussian distribution.

Altogether, after calibrating into units of strain, the strain measured using the fringe light as a homodyne readout follows from Eq. (8.7),

$$h_{\mathrm{HD}}(f) = h(f) + n(f) + q(f), \tag{8.11}$$

with a noise backgreound given by

$$S_{\mathrm{HD}}(f) = S_n(f) + S_q(f). \tag{8.12}$$

In summary, the astrophysical strain waveform, $h(f)$ adds to two separate noise processes to result in the measured strain $h_{\mathrm{HD}}(f)$. The first noise process is classical noise $n(f), S_n(f)$ which arises from random thermal, mechanical, electrical and laser processes in the interferometer that mask signal waveforms [34]. The second is quantum noise $q(f)$ with power spectral desnity $S_q(f)$, which results from the Poissonian arrival of photons from the fringe light [91]. From this result, a Gaussian likelihood can be constructed to infer the properties of the observed gravitational-wave signal with waveform model, $h(f;\theta)$.

Photon counting readout

Due to the dominating effect of quantum shot noise at frequencies above $\sim 500\,\mathrm{Hz}$, an alternative readout scheme has been proposed using single photon sideband detection, also known as a *photon counting* readout [92]. In broad strokes, the implementation of photon counting requires additional infrastructure in the beam-path between the gravitational-wave detectors signal recycling mirror and the DC readout photodiodes. The principle is that one can build an apparatus to interact with and filter specific photonic signal temporal modes present in the output. By then reading out the occupation number in each temporal mode basis used during the filtering, the underlying signal can be detected and interpreted. In this readout scheme, rather than generating shot noise, vacuum states lead to simply reducing the rate of signal photons. This implies that shot noise no longer imposes the fundamental limit on the observability of a signal with this readout. Instead the background rate of photons from other noise sources provides this limit. It is important to note that shot noise in a form is still present—now signal photons have shot noise associated with their observation and therefore follow an appropriate statistical distribution.

The crucial difference between the incremental improvements that have been made to the typical readout of gravitational-wave detectors and the inclusion of photon

counting, is that photon counting fundamentally changes the output measurement data. Rather than recording a continuous representation of the strain present on the detector, instead a discrete set of photon counts are recorded for a discrete set of filters in series. This leads to a measurement described by fundamentally different statistics. This difference in readout methodology leads to a reduced sensitivity to louder signals (demonstrated further in Sec. 8.3) however leads to increased sensitivity in the very low signal-to-noise ratio (SNR) regime.

In order to formulate the expected number of photons from a gravitational-wave signal with a background of classical noise sources, we can return to the definition of the emitted optical field in Eq. (8.6). From this expression, we see that quantum noise effects are not present in the measurement. Labelling the bases of temporal modes as $\{d_k(f)\}$, we can write down the electric field in each mode as

$$E_k = \int_{-\infty}^{\infty} \mathrm{d}f \; d_k(f) E_h^*(f) \; . \quad [\text{unitless}] \; . \tag{8.13}$$

In order to undertake this analysis, these optical mode filters, $\{d_k(f)\}$, correspond to an orthonormal basis which should mimic the signals of interest (i.e. have the same amplitude and phase behavior as a function of time/frequency). The condition of orthonormality is to simplify the experimental design. With an orthonormal filter basis, the filters can be placed in series along the beampath (which maintains a higher sensitivity than the parallel array of filters) without the complication of tracking the order of each filter in the process. Note that $d_k(f)$ has the units $1/\sqrt{\mathrm{Hz}}$. Within the experimental design this template bank can be implemented via filter cavities or—more likely for sophisticated signal inference – quantum memory technologies which are not currently developed to a level of maturity required for such an application [92].

Then the expected number of photons in a template given a particular realization of the optical field, which is again directly related to the strain in via Eq. (8.6),

$$\bar{n}_k = \langle E_k E_k^* \rangle = |E_k|^2 \; , \quad [\text{unitless or counts}] \; . \tag{8.14}$$

Thus, following Eq. (8.6) through to Eq. (8.14), will lead to the expected photon count for a particular realization of a strain signal. Experimentally, when an observation is then made, some *discrete* number of photons are generated in each mode basis filter. With this photon count per template, we are able to construct the likelihood function and perform statistical inference using this result. However, the specifics of the photon count depend on the origin of the signal. The exact functions

are discussed in detail below, though it is important to note that the underlying photon count distributions differ between signal-sourced photons, classical noise source photons, and single photon count readout photons. This allows for different photon sources to be disentangled in hierarchical studies without exact knowledge of the noise background. For the remainder of the manuscript, we assume perfect efficiency in our single photon readout devices.

Classical noise photons

The classical photons contribute a *background* of photons which will diminish the resolving power of counting individual signal photons. To compute the expected photon count from a classical noise background, we can compute expected number following Eqs. (8.13) and (8.14) to write

$$\bar{n}_{\mathrm{cl},k} = \left\langle \left(\int_{-\infty}^{\infty} \mathrm{d}f \; d_k(f) E_n^*(f) \right) \left(\int_{-\infty}^{\infty} \mathrm{d}f' \; d_k^*(f') E_n(f') \right) \right\rangle , \tag{8.15}$$

where $E_n(f) = g(f)n(f)$. The above expression can be rearranged, noting that the basis filters are unchanging functions so that $\langle E_n(f) d_k(f) \rangle = \langle E_n(f) \rangle d_k(f)$, as

$$\bar{n}_{\mathrm{cl},k} = \int_{-\infty}^{\infty} \mathrm{d}f \int_{-\infty}^{\infty} \mathrm{d}f' \; d_k(f) d_k^*(f') \langle E_n^*(f) E_n(f') \rangle \, . \tag{8.16}$$

Using the definition of $E_n(f)$, we can relate the expected number of classical noise photons to the classical noise power strain power spectral density, $S_n(f)$,

$$\bar{n}_{\mathrm{cl},k} = \int \mathrm{d}f \; |d_k(f)|^2 |g(f)|^2 S_n(f) \, , \quad [\mathrm{counts}] \, . \tag{8.17}$$

With the expected number of classical noise photons calculated, we can also outline the underlying distribution from which their count is generated. This fundamentally determines how many photons are actually observed due to classical noises. The corresponding distribution differs for different noise sources as a result of different occupation of the state in phase space. For the majority of noise sources, and all noises that we will consider here, the classical noises purely manifest in the phase quadrature. This leads to a specific distribution on the discrete photon count in a particular basis due to this occupation. The distribution is [458],

$$\mathcal{N}_{\varphi}(n_{\mathrm{cl},k} | \bar{n}_{\mathrm{cl},k}) = \frac{(2n_{\mathrm{cl},k})!}{2^{n_{\mathrm{cl},k}} (n_{\mathrm{cl},k}!)^2} \frac{\bar{n}_{\mathrm{cl},k}^{n_{\mathrm{cl},k}}}{(2\bar{n}_{\mathrm{cl},k} + 1)^{n_{\mathrm{cl},k}+1/2}} \, . \tag{8.18}$$

Note that broadly speaking, squeezing will complicate this further since it introduces additional photons due to the squeezed states, and such photons *will not* follow the above distribution. Therefore, it is actually more reasonable to operate a photon counting readout in the absence of squeezing [92, 459].

Signal photons

The expected number of signal photons in a particular basis mode, denoted $\bar{n}_{\mathrm{sig},k}(\theta)$, where θ are the parameters of the signal model, are computed nearly similarly to Eqs. (8.13) and (8.14),

$$\bar{n}_{\mathrm{sig},k}(\theta) = \left| \int_{-\infty}^{\infty} \mathrm{d}f \, d_k(f) g_h(f) h^*_{\mathrm{sig}}(f;\theta) \right|^2 , \quad [\mathrm{counts}] \, . \tag{8.19}$$

Note that the expected signal photon count increases quadratically with the signal amplitude. In particular, it is expected that $\sum_k \bar{n}_{\mathrm{sig,k}}(\theta) \sim \mathrm{SNR}^2/2$, provided the template bases can perfectly match the observed signals. While this is not feasible in practice it provides a useful baseline to compare observations.

Now, unlike the classical noise photons which will typically follow the distribution in Eq. (8.18), the signal photons exist as a coherent state that is filtered via the basis filters from the strain signal. Therefore, the number of signal photons follows a Poisson distribution,

$$P\left(n_{\mathrm{sig},k} \middle| \bar{n}_{\mathrm{sig},k}(\theta)\right) = e^{-\bar{n}_{\mathrm{sig},k}(\theta)} \frac{\bar{n}_{\mathrm{sig},k}(\theta)^{n_{\mathrm{sig},k}}}{n_{\mathrm{sig},k}!} \, . \tag{8.20}$$

The fundamental reason for these different distributions is that the photons from background originate from a stochastic process which leads to a geometric-like distribution (with a mean $\sim \bar{n}$ and variance $\sim \bar{n}^2$). In contrast, the signal photons follow Poisson distribution with mean $\sim \bar{n}$ and variance $\sim \bar{n}$.

Photon counting likelihood for transient signals

Now that we have the underlying distribution which we expect both signal and noise photons to follow, we can construct the likelihood for an observation of $\{n_k\}$ photons from N basis modes. Since, when we observe a set of photons, we will have no information on whether the photon is a signal or a noise photon and therefore need to convolve the uncertainties – marginalizing over whether each of the n_k photons are associated with a transient signal or classical noise processes. To do this, we rely on a measurement of the classical noise PSD to calculate $\bar{n}_{\mathrm{cl},k}$ for noise photons following Eq. (8.17). For a single filter template, we can write the likelihood as

$$p(n_k|\theta) = \sum_{m=0}^{n_k} P(m|\bar{n}_{\mathrm{sig},k}(\theta)) \mathcal{N}_\varphi(n_k - m|\bar{n}_{\mathrm{cl},k}) \, , \tag{8.21}$$

This convolution is simply a mixing of the two distributions.

It helps to think of this with a single photon. When a single photon is detected in a filter, Eq. (8.21) evaluates the likelihood as the sum of the probabilities that the photon was a signal photon multiplied by the probability that the photon does not originate from a noise process or vice versa (signal photon and not a noise photon). While the probabilities are fixed for the classical noise photons (under the assumption that the PSD is known), varying the parameters of the signal θ will lead to different probabilities of $\mathcal{P}(m|\bar{n}_{\mathrm{sig},k}(\theta))$, thereby leading to the inference of the model parameters. Additionally, note that as $\bar{n}_{\mathrm{cl},k} \to 0$, Eq. (8.21) reduces to the Poisson distribution.

To write the full likelihood for a photon counting readout with N filters $\{d_k(f)\}$ for $k = 1, ..., N$, we can multiply the likelihoods from each filter together,

$$p(\{n_k\}|\theta) = \prod_{k=0}^{N} p(n_k|\theta)\,, \tag{8.22}$$

since all the N templates are orthonormal by construction. This is analogous to the construction of the Whittle likelihood for homodyne readout analyses, where the likelihood functions from individual frequency bins are multiplied together [462, 463]. This therefore lays out the foundation for conducting gravitational-wave inference calculations with a photon counting readout. For the remainder of the manuscript, we will simulate photon counting readout observations for BNS post-merger detection and inference, and compare to standard homodyne methods.

8.3 Individual-event post-merger inference

Having laid out the statistical background for the photon counting readout and its relation to the homodyne observations, we turn our attention to its utility for observations of BNS post-merger remnants. Since these signals are anticipated to be high-frequency observations (with the most detectable between 1.5 and 4 kHz), the benefit of photon counting is anticipated to be greater since $S_q(f) \gg S_n(f)$. In this section, we explore a photon counting readout's role as a viable readout alternative in the context of individual BNS post-merger observations. We also outline the designed set of basis filters used for the analysis. We then continue on to explore the impact of different noise backgrounds, different photon counts, and different SNRs on the constraints placed on the dominant behavior of the BNS post-merger signals' fundamental modes.

Filter design for post-merger signals

To optimize the information gained from photon counting measurements for specific signals, we need to model the temporal mode filter responses to closely mimic the signal to be observed. In the context considered here, the temporal modes should mimic the structure of post-merger signals. This can be understood, figuratively, as in-situ hardware-based matched filtering where the basis mode must closely match the signal (i.e. maximize Eq. (8.19)) in order for the filter to lead to a non-negligible photon count expectation.

The salient features of a post-merger signal are governed by peak features within their frequency spectrum which can be closely modeled as damped sinusoids. The general function for the damped sinusoid in the frequency domain is given by

$$L(f|f_0,\gamma,A,\phi_0,t_0) = \frac{A\gamma e^{-2i\pi f t_0}}{2\sqrt{2}\pi^{3/2}}\times \left(\frac{e^{-i\phi_0}}{\gamma^2+(f-f_0)^2}+\frac{e^{i\phi_0}}{\gamma^2+(f+f_0)^2}\right). \tag{8.23}$$

The amplitude profiles of this function capture Lorentzian line-shape structures in the data. Here, f_0 is the peak frequency of the damped sinusoid, γ is the half-width at half-maximum (HWHM), A is the amplitude (in the time domain), and ϕ_0 and t_0 are phase and time offsets of the function.

To construct our simulated photon counting readout observations, we need to define the set of bases to consider. The general expectation from the signal duration, ΔT, and the bandwidth of the observation, ΔF, is that the number of filters required to adequately cover the space is given by

$$N \sim 2\Delta F \Delta T\,. \tag{8.24}$$

This follows directly from the number of terms present in a Fourier transform of post-merger signal over a ΔT duration. The factor of two accounts for both sine and cosine components. Since we are designing the temporal basis modes to target post-merger signals of larger SNR, we construct the bases to span 1.5 to 4 kHz, and the duration of the window under consideration is $\sim$ 40 ms, we anticipate having $\sim$ 200 filters (with 100 sine and 100 cosine filters). The filters have the following parameters:

1. A is set according to the orthonormalization (discussed shortly),

2. $\gamma = 100\,\mathrm{Hz}$ to mimic the rough width of a post-merger signal's frequency behavior.

3. f_0 spans 1.5 to 4 kHz with 10 equally spaced frequencies,

4. t_0 spans -20 to 20 ms with 10 equally spaced points.

5. ϕ_0 is set to either 0 or $\pi/2$ for the cosine or sine component.

These choices lead to 200 filters to be used in the simulated photon counting readout design. This of course may not be optimal since the damped sinusoidal basis does not exactly mimic the post-merger signal. The resolving power of the basis design could be improved by constructing filters that mimic the full structure of the post-merger signal. Finally, these modes are then shuffled into a random order and the Gram Schmidt method is applied to orthonormalize the set of filters, $\{d_k(f)\}$. This ends up generating basis modes which are orthonormal, though some of the properties of the mode are warped to accommodate this.

As an example of these temporal mode bases in action, we simulate a BNS post-merger signal generated from a supervised learning model trained on numerical relativity simulations [464]. This model assumes an APR4 equation of state. The simulated post-merger signal was generated with a total mass of $2.4\,M_\odot$, at a redshift of 5×10^{-2}, and has an optimal SNR of 1.10 (in an unsqueezed version of CE). The corresponding peak frequency of the signal is $\sim 3100\,\mathrm{Hz}$ (see upper panel of Fig. 8.2). To demonstrate the measurement, we plot the strain, as well as the temporal mode basis functions colored by their expected number of signal photons in Fig. 8.2. In the upper panel, we show the amplitudes of these filters as well as the strain amplitude as a function of frequency. The basis filters possess non-regular, "warped" patterns and features as a result of the orthonormalization procedure. In the lower panel, we represent the time, frequency, and phase (sine or cosine) of the signal, where the half circles correspond to the cosine (right) and sine (left) sides. The filters will not exactly correspond to these time and frequency grid points due to the orthonormalization. However, it is a useful visual indicator of which basis templates are leading to possible photon generation. Unsurprisingly, the greatest number of expected photons are originate from the templates which overlap the greatest with the signal.

Single-event inference of damped sinusoids

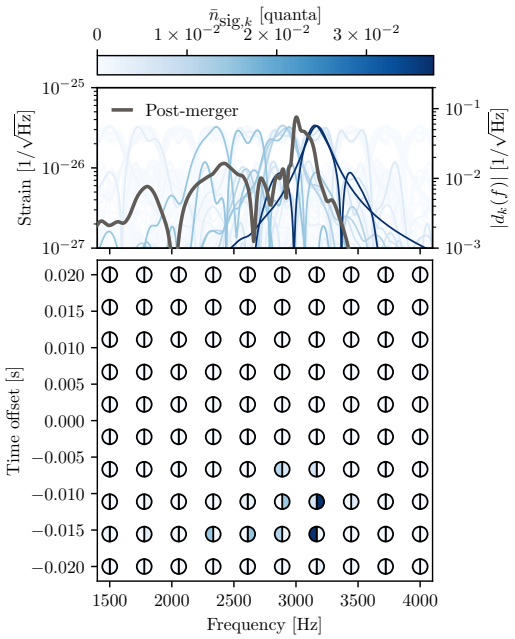

Figure 8.2: Demonstration of the interaction between the temporal mode basis that was constructed and the measurement of a binary neutron star post-merger signal. In the upper panel, the strain amplitude and basis filter amplitudes are shown as a function of frequency. The basis modes are colored according to their expected number of signal photons. While each basis mode is initially constructed according to Eq. (8.23), and the parameters laid-out above, the process of orthonormalization leads to unexpected basis filter structures. In the lower panel, the time, frequency, and phase (sine or cosine) of the temporal basis are presented. This grid summarizes the 200 basis filters which are present in the observational strategy.

With this understanding of the photon count readout design from an analysis point-of-view, and in broad strokes how it interacts with the post-merger signal, we can now move forward with inferring the signal properties. Utilizing the photon counting likelihood presented in Eq. (8.22), we model the post-merger signal approximately as a damped sinusoid as expressed in Eq. (8.23), and sample the posterior distribution of the model parameters, $\theta = \{f_0, \gamma, A, \phi_0, t_0\}$. While we have chosen the same functional form for the basis mode construction (prior to orthonormalization) and the signal model, these do not need to be same. In fact, in future studies with a photon counting readout method, it is plausible a more fine-tuned model for the signal can be applied. The priors on all parameters are flat. The peak frequency ranges from 1.5 to 4 kHz, γ range from 0 to 400 Hz, A from 0 to 10^{-19}, and ϕ_0 from 0 to 2π. Due to the highly oscillatory nature of the signal, and therefore sharp and frequent peaks in the t_0 posterior, the t_0 samples are marginalized numerically between -20 and 20 ms. To demonstrate the utility of the photon counting readout, we consider

three different scenarios, and compare the observations to an unsqueezed version of Cosmic Explorer. Doing so allows for a direct comparison between homodyne and photon counting readouts, and the optimal SNRs computed. In the remaining results in this section, we consider no noise contamination in either the homodyne or photon counting readouts corresponding to no Gaussian noise fluctuations for the homodyne, and no background noise photons for the photon counting readout. In order to mitigate the impact of model misspecification between the simulated signals being observed and recovered, we use our simple damped sinusoid model for both injection and recovery.

The first scenario to consider is a super-threshold observation—a post-merger signal where it would comfortably be detected with a standard homodyne readout. We simulate a damped sinusoid with an optimal SNR of 5, a peak frequency of 2.75 kHz, $\gamma = 50$ Hz, $\phi_0 = \pi/2$, and $t_0 = 0$ ms. This injection has an expectation of generating 3.13 signal photons (this differs from the optimal 12.5 expected photons due to mismatches between the mode basis and the signal[1]). In this simulated observation, 3 signal photons were detected. In Fig. 8.3, the posterior distributions of the model parameters from both the homodyne readout (orange) and the photon counting readout (orange) are shown. The shades correspond to the 50% and 90% credible intervals. Overall, we find that both readouts are able to confidently able to constrain A, f_0, γ, and t_0 well, with less certainty in the phase of the signal. In the superthreshold case of an SNR 5 post-merger-like signal, the homodyne outperforms the photon counting readout—the homodyne readout constraining the the amplitude 1.5 times and peak frequency 3 times more tightly at the 68% credible interval.

However, such a loud post-merger signal is not expected, even for the vast majority of signals observed with a network of third-generation detectors. For a more realistic example, we now consider a subthreshold observation of an SNR 1 post-merger-like signal—the same damped sinusoid with a smaller amplitude. The expected number of photons for the SNR 1 signal is 0.125, corresponding to a 11.8% chance that at least a single photon is generated. The result presented here corresponds to one photon being generated. The posterior distributions for this analysis are depicted in Fig. 8.4. This result demonstrates the broad benefit of the photon counting readout scheme. While the homodyne readout is incapable of providing

[1] Specifically, in this case the filter mode basis has a broader $\gamma = 100$ Hz, does not have a filter that lies on the exact peak frequency, and are orthonormalized. All these contribute to the reduced photon count expectation.

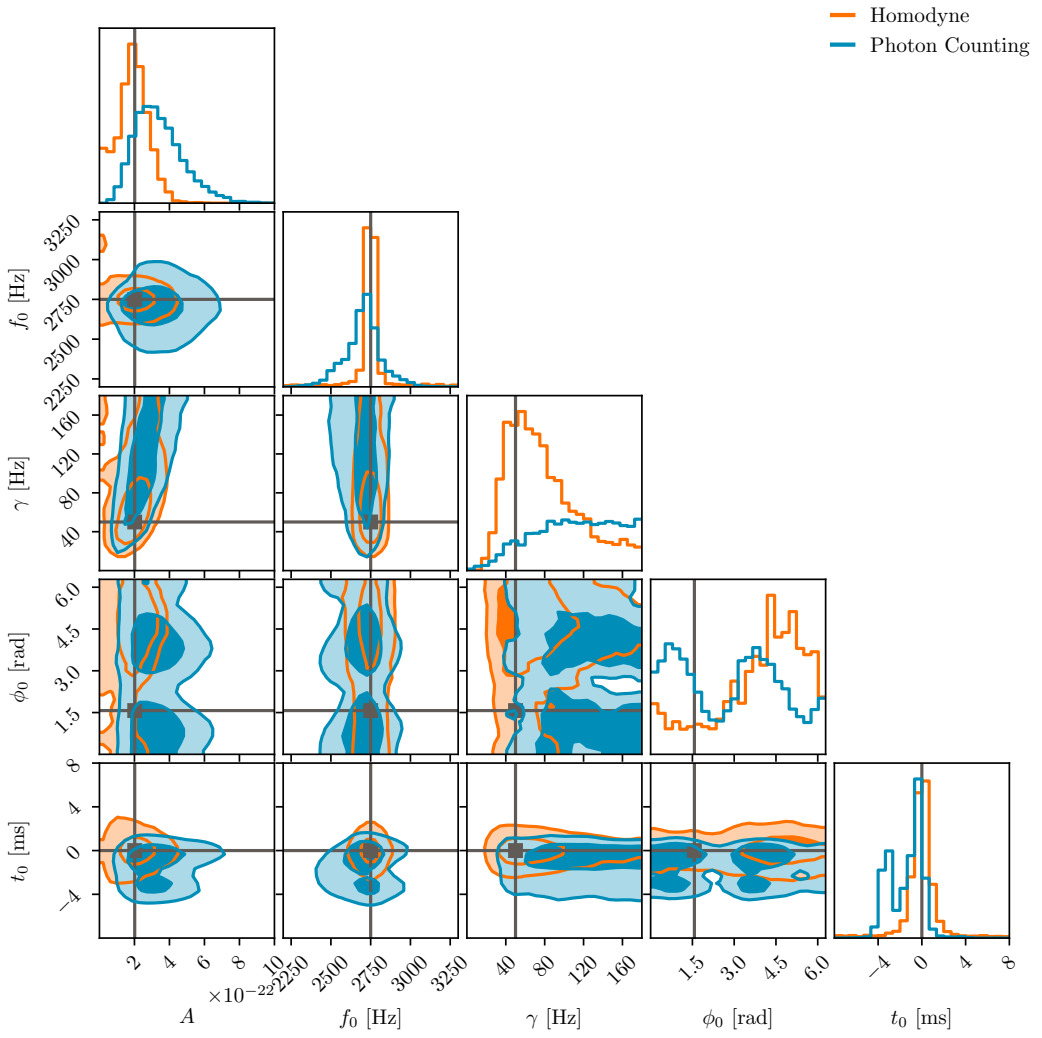

Figure 8.3: Corner plot representation of the posterior distributions for the inferred damped sinusoidal parameters using measurements from the homodyne (orange) and photon counting (blue) readouts' data outputs for an SNR 5 damped sinusoid. The true values for the simulated damped sinusoid are shown in grey. The contours correspond to the 50% and 90% credible levels. Overall, the homodyne constraints on the observations are more stringent in this relatively higher SNR regime—as expected. The jagged structure in the ϕ_0 homodyne readout posterior distributions originates from the rapid oscillatory nature of a 2.75 kHz signal. This is not present in the photon counting readout, since the resolution of the time offset is only ~ 4 ms. While more photons in individual filters can shrink the time posterior, it is less drastic than the homodyne readout result.

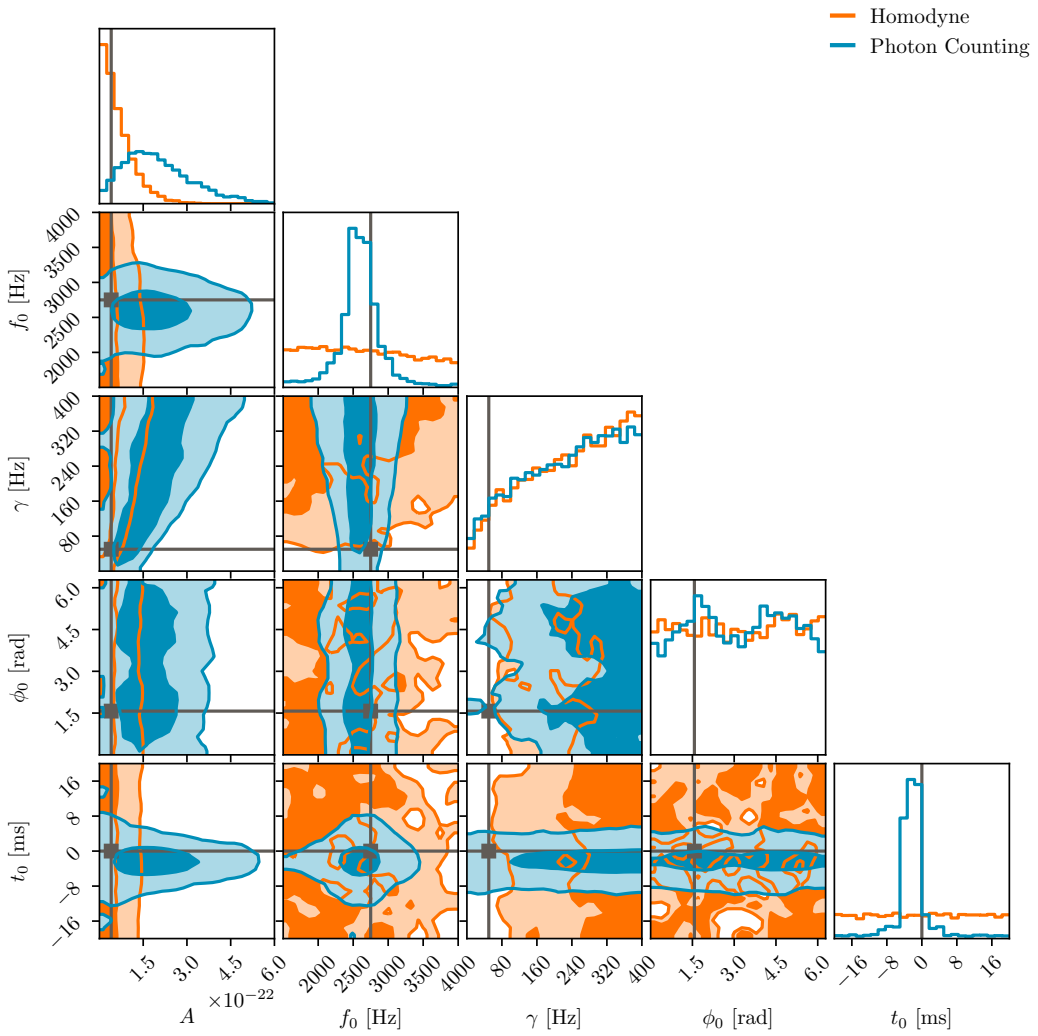

Figure 8.4: Corner plot representation of the posterior distributions for the inferred damped sinusoidal parameters using measurements from the homodyne (orange) and photon counting (blue) readouts' data outputs for an SNR 1 damped sinusoid. The true values for the simulated damped sinusoid are shown in grey. The contours correspond to the 50% and 90% credible levels. We see that such a low SNR signal is not resolved by the homodyne readout. However, in the case of photon counting, $\bar{n}_{\text{sig}} = 0.125$, and so there is a 11.8% chance that at least one photon is generated by such a signal. The posterior distribution for the photon counting readout can be informed by an individual photon, leading to meaningful constraints.

information from such a low SNR signal observation, the photon counting readout is able to constrain both the peak frequency and the time offset well. In the context of post-merger signals, gaining this information regarding the peak frequency is useful for understanding the equation-of-state and NS properties and will be the primary target for future discussions in this manuscript.

Role of noise backgrounds and photon counts

Having seen the straight-forward examples demonstrating *how* photon counting can lead to improved constraints, there are two unintuitive behaviors of a photon counting readout scheme that need to be conveyed. Unlike the homodyne readout where the width of parameter posterior can be directly related to the inverse of the total PSD, $S_{\rm HD}(f)$, the inferred posterior distributions from photon counting depend on *both* the relative relationship between the signal, $h(f)$, quantum shot noise, $S_q(f)$, and the other noises, $S_n(f)$, *and* the number of signal photons generated which is an inherently random process.

For the former behavior relating the noise sources, as made apparent in Eqs. (8.17) and (8.19), $\bar{n}_{\rm cl} \sim S_n(f)/S_q(f)$, and $\bar{n}_{\rm sig} \sim |h(f)|^2/S_q(f)$. With these relevant ratios in mind, we can consider a number of different cases.

1. $|h(f)|^2 \gg S_q(f)$ & $S_n(f)$ corresponds to the high SNR regime, where the signal is comfortably above the noise backgrounds. In this situation, the relative sensitivity loss due to quantum effects and classical noises are unimportant. This is analogous to the result in Fig. 8.3 with and SNR 5 post-merger-like signal. In this regime, a homodyne readout is expected to outperform photon counting.

2. $S_n(f) \gg |h(f)|^2$ corresponds to a classical noise background that buries the signal in the data. For a homodyne readout, this results in significant Gaussian noise due to classical effects (such as controls noise in LIGO detectors at lower frequencies). For a photon counting readout, this leads to too many noise photons to distinguish the signal photons. In the photon counting case, if $S_q(f) \gg S_n(f)$, the fewer photons are seen each observation, but the classical noise photons still dominate over an ensemble. Here, no readout choice will be able to lead an observable signal. Conversely, when $S_n(f) \gg S_q(f)$, many noise photons will appear in even a single observation.

3. $S_q(f) \gg |h(f)|^2 \gg S_n(f)$ corresponds to a low SNR observation, that

will still produce more signal photons than background noise photons. This is analogous to the observations shown in Fig. 8.4 with an SNR 1 signal. In this low SNR regime with the specific hierarchy of noise sources, photon counting will show the most promise, especially in regions of parameter space where quantum shot noise dominates the homodyne readout PSD. This will correspond to signals in the frequency band above ~ 1 kHz and motivates this study focusing on post-merger signals.

To understand this quantitatively, we simulate the signal presented in Fig. 8.2 which corresponds to an optimal SNR of 1.10 in an unsqueezed CE in Fig. 8.5. For the photon counting readout, this observation has an expected photon count rate of 0.19, and therefore a 16% chance to generate a single photon (and a 1.5% chance to generate two photons). We then vary the overall sensitivity of the readouts by scaling the shot noise in the case of the homodyne readout, or by scaling the classical noise background in the case of the photon counting readout. The scaling of the homodyne readout is approximately equivalent to different dB of squeezing from 0 to 20 dB. The thicker lines in the figure correspond to the expected total PSD for CE [45] (including squeezing; 10 dB) for the homodyne results, and to the expected classical noise background for the photon counting results. In the left panels we show both the simulated post-merger signal, as well as the different PSDs, and in the right panels we show the marginal posterior distributions on the inferred peak frequency. For all results shown, a single photon is detected in the most likely basis filter to observe a photon following Eq. (8.19).

The key observation from these posterior distributions is that their behavior is fundamentally different between the homodyne and photon counting readouts. For a homodyne readout, as the SNR increases, the overall width of the posterior decreases. However, this is only true up to a point for photon counting. Once $\bar{n}_{\rm sig} \gtrsim \bar{n}_{\rm cl}$ (which is closely related to the third scenario), the width of the posterior is unchanging. This is due to the resolution attainable from a single photon. Finally, note that once the SNR exceeds ~ 5, the homodyne readout provides significantly better constraints than photon counting.

In order to generate tighter constraints on the post-merger parameters with photon counting, it is then necessary to observe more photons for a given observation. We explore this in Fig. 8.6, where 4 realizations are generated with no photons through to 3 photons. While more photons could in principle be used, higher photon counts suffer from the model misspecification between the post-merger signal simulated

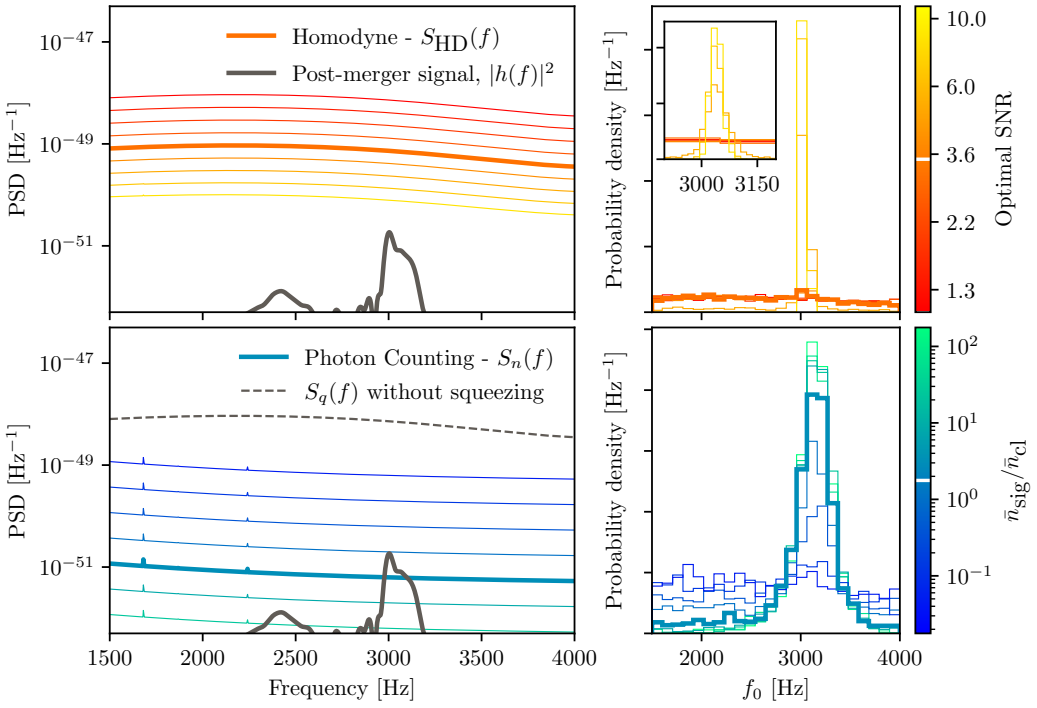

Figure 8.5: Impact on the change in the noise backgrounds for the homodyne (upper; orange) and photon counting (lower; blue) readouts. The left panel shows the post-merger strain simulated, as well as $S_{\rm HD}(f)$ for the homodyne and $S_n(f)$ for the photon counting at various different levels. The relevant statistic for the homodyne readout is SNR $\sim 1/S_{\rm HD}(f)$, while the relevant statistic for the photon counting is $\bar{n}_{\rm sig}/\bar{n}_{\rm cl} \sim |h(f)|^2/S_{\rm n}(f)$. These quantities, for their respective readout schemes, control the constraints placed on the parameters such as the peak frequency, as seen in the right panels. The thicker lines correspond to the expected results with CE's designed squeezing level (10 dB; for the homodyne), or classical noise realization (for the photon counting), and the white lines on the colorbars indicate their corresponding values.

and the damped sinusoid model used for recovery. Note that this leads to generating photons in multiple different bases for $n_{\rm sig} \geq 2$. The amplitude of the signal is scaled such that a realization with the desired number of photons is generated. This is a subtlety of a photon counting readout. An observation has the *total* Poisson probability $P(n_{\rm sig}|\bar{n}_{\rm sig})$ of generating $n_{\rm sig}$ photons. Therefore, there are two options for generating a desired photon count; either increase the signal amplitude such that $\bar{n}_{\rm sig}$ increases, or simply wait until a fortuitous photon count realization occurs. In these results, we see that as the photon count increases, constraints on all parameters shrink in a similar manner to homodyne constraints in e.g. Fig. 8.5. While the behavior of other parameters indicates the improved constraints since

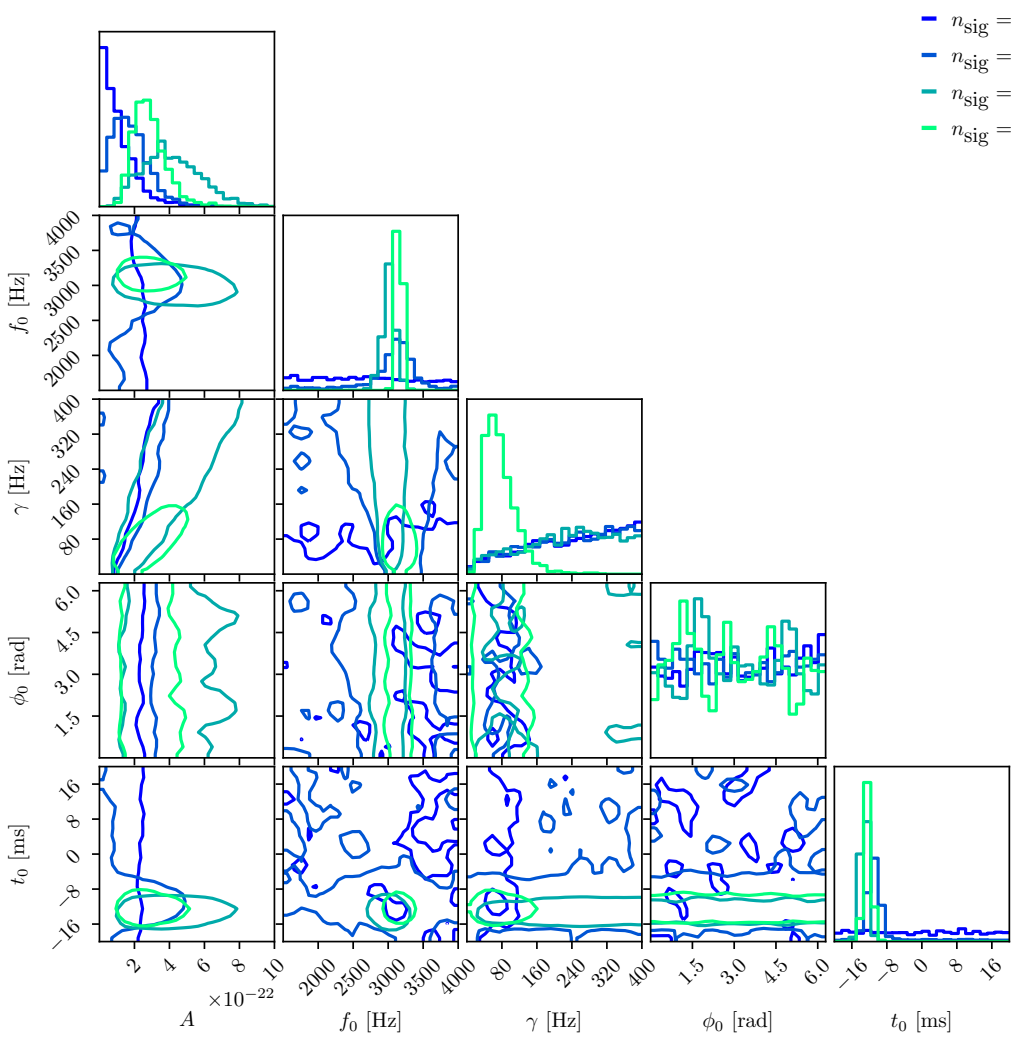

Figure 8.6: 90% credible levels of simulated post-merger signal posteriors with varying photon counts. As the photon count increases, constraints on all the parameters of interest narrow in a manner similar to a homodyne readout result. Note that the different photon counts origin from different basis modes (not all photons are in the same basis). Therefore this is only one plausible realization of each posterior with $n_{\rm sig}$ photons. This is an exercise in understanding how the readout behaves from an analysis point-of-view. For the SNRs anticipated for post-merger signal observations, $n_{\rm sig} > 1$ will be highly unlikely for the majority of observed signals.

these parameters are fixed as $n_{\rm sig}$ increases, the amplitude is scaled in such a way to generate the desired photon count.

To summarize, a photon counting readout can lead to significantly improved constraints in the low SNR regime when, and only when, quantum noise dominates both the signal strain and the classical noise. Rather than having a single noise term like a homodyne readout, the overall result from photon counting is determined by both

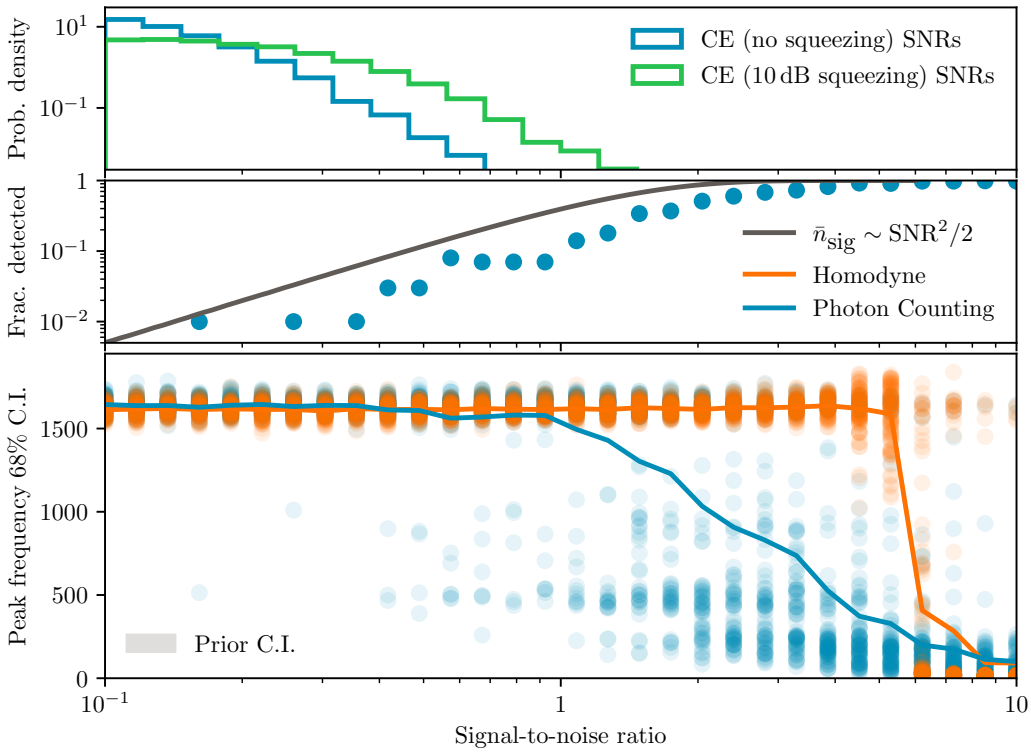

Figure 8.7: Summary of the capabilities of a photon counting readout scheme for detecting and measuring BNS post-merger signals. The peak frequency 68% credible intervals are shown for 100 observations at each SNR from 0.1 to 10 in the lower panel for the homodyne readout (orange) and photon counting readout (blue). The solid lines correspond to the mean at each SNR value. In the middle panel, the fraction of detected signals according to the photon counting readout are shown, as well as the theoretical expectation if $\bar{n}_{\rm sig}$ follows ${\rm SNR}^2/2$. The top panel highlights the expected distributions of SNRs from 10^4 observations both in an unsqueezed CE (blue), as well as at design sensitivity (green).

the quantum noise, $S_q(f)$, and the classical noise $S_n(f)$. The relative background floor in the posterior distribution is determined by the ratio of the signal strength to the classical noise, $|h(f)|^2/S_n(f)$. While the number of photons, and hence the tightness of the constraints, is given by the ratio of the signal strength to the quantum noise, $|h(f)|^2/S_q(f)$.

Summary of individual constraints

Equipped with this understanding of inferences made with a photon counting readout, we can put these observations in the context of ensembles of realistic post-merger signals. For SNRs ranging from 0.1 to 10, we generate 100 different post-merger signals for each SNR using the KNN APR4 EoS model [464] with a peak frequency

between 1.5 and 4 kHz. We simulate the detection of this signal in both a homodyne readout (CE with no squeezing) and a photon counting readout. We assume no background noise from either the homodyne noise PSD, or the classical background photon counts. In Fig. 8.7, we then show the different 68% credible intervals on the peak frequency from the two readout schemes, as well as the mean from the 100 observations (solid curve). We find, as expected, photon counting readout observations are capable of making meaningful measurements from very low SNR signals. We find that ~ 1 in $O(100)$ observations of an SNR 0.1 post-merger signal lead to an f_0 68% credible interval $\lesssim 500\,\mathrm{Hz}$. In the middle panel, we highlight what fraction of observations resulted in a detection, and how this relates to the theoretical result when $\bar{n}_{\mathrm{sig}} \sim \mathrm{SNR}^2/2$. The detected fraction in our simulated dataset is typically below this theoretical curve, since there are some inefficiencies due to the mismatches between the signal and templates bases. Finally, in the top panel, we highlight the distribution of expected SNRs in CE from post-merger signals. Presenting both the SNR distributions when there is no squeezing present (relevant for photon counting), or when 10 dB of squeezing is present (relevant for design sensitivity CE), it is clear that the vast majority of signals will fall in the low SNR regime where photon counting can provide tighter constraints due to the serendipitous detection of the occasional photon. The rate of single photon detection is somewhat surprisingly high even for low SNR signals.

8.4 Hierarchical Equation-of-state constraints

Thus far, we have explored the analysis of individual post-merger signals. However, since many observations will be in the low SNR regime, it prudent to discuss strategies to combine information across multiple observations. In particular, when discussing whether photon counting is a viable alternative for high frequency readout schemes in third generation detectors such as CE, it is appropriate to study its role on a hierarchical level.

To demonstrate how the overall impact of the readout scheme on hierarchical constraints, we follow Ref. [465] relating the peak frequency of a post-merger signal to the chirp mass of the binary progenitor and an equation-of-state (EoS) parameter—the radius of a $1.6\,M_{\odot}$ neutron star, denoted $R_{1.6}$. The underlying expression follows

$$\begin{aligned} f_{\mathrm{peak}} = \beta_0 + \beta_1 \mathcal{M} + \beta_2 \mathcal{M}^2 + \beta_3 R_{1.6} \mathcal{M} \\ + \beta_4 R_{1.6} \mathcal{M}^2 + \beta_5 R_{1.6}^2 \mathcal{M} + \epsilon, \end{aligned} \tag{8.25}$$

where β_1 through β_5 are fit parameters to numerical simulations. Their exact values can be found in Table I of Ref. [465]. The additional error term ϵ is assumed to follow a Gaussian distribution with a standard deviation of 61 Hz. While this expression depends on the binary chirp mass (which we do not simulate), we assume all the mergers are equal mass binaries, since the EoS model used here does not take mass ratio into consideration. The KNN model approximates the APR4 equation-of-state $R_{1.6} = 11.07\,\mathrm{km}$—slightly lower than the standard APR4 equation-of-state value of $R_{1.6} = 11.27\,\mathrm{km}$ [466]—and thus this will be considered the true value for the population.

Again somewhat following Ref. [465], we can then construct the population likelihood for the hierarchical analysis as

$$p(\{d\}|R_{1.6}) = \prod_{i=1}^{N_{\mathrm{events}}} p\left(d_i \middle| f_{0,i} = \frac{f_{\mathrm{peak}}(M_i|R_{1.6})}{1+z_i}\right). \tag{8.26}$$

This expression assumes we know the total mass and redshift from the inspiral of the BNS. While this is only approximate, it provides a "best-case scenario" that can be used to compare the photon counting result and the homodyne result.

Equation (8.26) can be evaluated with either the marginal likelihood from either the photon counting or homodyne readout inference results. However, in order to construct the marginal likelihood efficiently (for both the homodyne and photon counting cases), we take the following steps. Before we can start the analysis of an individual signal, we need to ensure that there are no prior mismatches between the simulated population and the population models used during the recovery. This would be not be as problematic if we both generating and recovering the properties of the signal with the same model. However, in a more realistic scenario, we do not know the underlying post-merger model, and so would need to assume some functional form, such as a damped sinusoid, for a simplistic model. This means that we do not have an understanding of the appropriate amplitude or HWHM prior of the Lorentzian model when fitting the signal. Therefore, to simulate this, we generate 10^4 simulated signals and extract the best-fit (maximum likelihood) amplitude, and width γ to construct the "astrophysical" prior on the damped sinusoid parameters. These signals follow the APR4 EoS. The primary and secondary masses of the BNS are drawn from a uniform distribution ranging from 1.2 to 1.4 $M_\odot$, the redshift from the Madau-Dickinson star formation rate, the sky location isotropically across the sky, and rotation angles uniformly through all possible angles. This is then used in the marginalization procedure outlined below.

To produce the simulated ensemble of observations, we generate 10^4 strain signals following the same distributions as above. Due to the broad uncertainty in the BNS merger rate density [8], this corresponds to duration anywhere between 5×10^{-3} and 0.75 years. From each signal, we can compute the observed gravitational strain for the homodyne readout and the photon count (per basis) for the photon counting readout. We note that 95% of these signals possess SNRs $\lesssim 0.47$ in design sensitivity CE (with squeezing), and only a total of 14 signal photons were observed. In contrast, 1145 background noise photons were measured with the design sensitivity classical noise, and 93 background noise photons were measured when the classical noise was reduced by an order of magnitude. In order for the hierarchical inference to succeed in measuring the hierarchical EoS parameter from the ensemble, it needs to overcome the noise background present.

To compute the individual event marginal likelihoods, $p(d_i|f_{0,i})$, we then compute the likelihood for $f_{0,i}$ given the "known" values of M_i and z_i and a choice of $R_{1.6}$. The expected amplitude is marginalized over a uniform distribution from zero to the fitted value from the maximum-likelihooe estimate. The other parameters for the damped sinusoid are drawn from their population distributions as discussed above. The error term on the peak frequency, ϵ, is also randomly drawn from its appropriate distribution. We can compute the summation over the likelihood values corresponding to these draws to generate the marginalized likelihood at a particular EoS parameter value. We can then sweep over values of $R_{1.6}$ to construct the marginal likelihood each individual event. Finally, we then employ Eq. (8.26) to construct the final hierarchical likelihood, $p(\{d\}|R_{1.6})$. Ultimately, this is an approximation to what would be necessary in future studies, where the underlying hierarchical model should provide information both about the peak frequency and the amplitude of post-merger signal.

For a fair comparison of the hierarchical analysis, we present four different results. Until now in the manuscript, we have presenting results comparing the photon counting to an *unsqueezed* homodyne readout. While this is appropriate for a comparison between readout methods, to understand the viability of photon counting in future detectors, it needs to be compared to their design sensitivity, which involves squeezing. For CE, the design sensitivity is achieved with 10 dB of squeezing. Therefore for the hierarchical analysis comparison, we compare photon counting with no squeezing and design sensitivity classical noise PSD to a homodyne readout for CE with 10 dB of squeezing. We also present inferences for when the classical

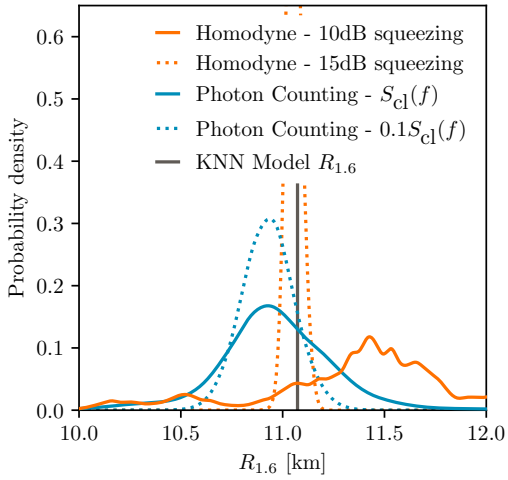

Figure 8.8: One-dimensional posterior distributions of the simulated inference of the radius of a 1.6 $M_\odot$ neutron star with both a homodyne (orange) and photon counting (blue) readout schemes. While the homodyne result with 10 dB of squeezing has a more localized mode near the true value of $R_{1.6} = 11.07$ km, it finds additional possible viable features in the distribution which increase the inferred credible intervals. As the SNRs of the ensemble increase, this mode vanishes. Finally, as the detector is improved for either a homodyne readout through increased squeezing or for a photon counting readout with a lower classical noise, the overall constraint on $R_{1.6}$ improves to a similar degree.

noise PSD is reduced by an order of magnitude, and when squeezing is increased to 15 dB.

The one-dimensional posterior distributions on the EoS parameter $R_{1.6}$ in Fig. 8.8 for the hierarchical inference analysis with 10^4 observations. From these results, we find that the current design sensitivity would lead to photon counting outperforming a homodyne readout by about a factor of 2.5, with the total 68% credible region width for $R_{1.6}$ found to be 1.25 km and 0.51 km for the homodyne with 10 dB of squeezing and photon counting with design sensitivity (and no squeezing), respectively. Interestingly, the 10 dB squeezing result also finds additional viable radii given the observed data. Their significance decreases with more observations or higher SNRs, and are likely a product of the mismatch between the simulated waveform model and the recovery with the damped sinusoidal model. Furthermore, either increasing squeezing or reducing the classical noises leads to improved constraints. The homodyne readout result with 15 dB of squeezing leads to the best constraints, however such a degree of squeezing is not expected in CE, and the constraint is dominated by a few loud observations.

8.5 Implications

We have explored the possibility of utilizing a photon counting readout to aid future gravitational-wave detectors in the task of detecting BNS post-merger signals. We have demonstrated at photon counting readout methods can plausibly outperform a homodyne readout in CE for BNS post-merger detection. Future studies will be required to further investigate the design of such a readout scheme to optimally detect post-merger signals and continue to compare with homodyne readout methods.

Photon counting clearly has a number of advantages and disadvantages. In terms of disadvantages, the current technology required for these temporal basis mode filters has not reached maturity and such a detector design will crucially rely in near-future development of quantum memory and metrology apparatuses. Furthermore, its utility is applicable to a very specific domain where the signal models can be relatively straightforward to lead to simple interpretation of single- or few-photon count measurements, and that the quantum noises dominate over both the signal and the classical noise background. As studied here, the salient, detectable details of BNS post-mergers are relatively simple and are sufficiently quiet in a gravitational-wave detector for the photon counting to potentially outperform a homodyne readout. The clear benefits of the photon counting readout are that when these criteria are met, information can be gained from large ensembles of low SNR observations. We find that with post-merger signals, photon counting could plausibly detect 1 BNS post-merger out of only 100 SNR 0.1 signals which fall within its designed bandwidth. Furthermore for hierarchical ensembles of observations, this translates into photon counting providing a competitive alternative for inferring EoS parameters from the post-merger. Additionally, it will allow for more clear classification of which signals are contributing information to the hierarchical constraints.

Beyond BNS post-merger signals, there are other possible applications for such a readout scheme. One such possibility is the use of photon counting readouts for more optimal, experimentally-enhanced strategies for stochastic gravitational-wave signal searches. Another intriguing possibility is the use of photon counting readout methods for probing deviations from general relativity, which will be inherently low SNR. Such applications will require careful construction of basis filters and many design studies. While we have demonstrated a straightforward application of photon counting readouts for gravitational-wave detection and inference for BNS post-merger signals, there are likely many prospective science cases for this potential

experimental readout scheme.

Chapter 9

SUMMARY AND FUTURE OUTLOOK

In this thesis, I have presented my research aimed at improving field of gravitational-wave astronomy through careful statistical analysis. I have looked at making robust statements about the spin-effects of individual gravitational-wave observations (Chapter 2) and the implications of such statements for formation channels of compact binary mergers (Chapter 3). I then formulated the analysis framework required to undertake hierarchical tests of gravity to account and correct for both the astrophysical population model (Chapter 4) and selection biases that originate from searches only targeting observations consistent with Einstein's theory (Chapter 5). The framework I developed was then extended to incorporate theoretical expectations for the behavior of plausible deviations from general relativity through the expected curvature dependence (Chapter 6). In addition to these improvements to specific hierarchical tests, I developed a summary statistic that is useful for quantifying model misspecification in hierarchical population studies (Chapter 7). Finally, looking toward future development of third-generation gravitational-wave detectors, I proposed how such detectors can utilize a photon counting readout scheme to potentially observe the post-merger signal from binary neutron star mergers for effectively (Chapter 8).

While I have presented complete, published works on these topics in this thesis, no research is ever truly finished. Below, I will summarize possible future avenues for the work presented here.

9.1 Outlook for theoretically motivated tests of gravity

In Chapter 6, I exploited the theoretical expectation that higher-order corrections to general relativity will necessarily appear as higher-order curvature terms in the action [87, 88]. For gravitational-wave observations, this leads to a deviation that scales inversely with the total mass of the binary (to an integer power greater than one) [388]. While I demonstrated that this information can be incorporated into the theory-agnostic hierarchical gravitational-wave tests, there is more theoretical knowledge that can be gained from theoretical calculations. In particular, in the

absence of additional fields such as in Einstein-dilaton Gauss-Bonnet [55, 467, 468] or Chern-Simons extensions [469, 470], the leading order correction occurs specifically at the 5 post-Newtonian expansion order. The term is analogous to tidal effects in neutron-star mergers [471]. In the presence of these additional fields, the expected post-Newtonian order decreases to either the -1, 2, or 3.5 order depending on the relevant lengthscales in the system [472]. Future extensions to the framework presented in Chapters 4 and 6 could be implemented to both infer the relevant curvature dependence of any possible deviations, and distinguish any possible preferred post-Newtonian order. This would allow for significantly stronger tests of gravity, and would strengthen any claims of a possible deviation if the inferred orders agree with existing possible extensions.

9.2 Outlook for hierarchical model misspecification tests

In Chapter 7, I outlined the population-level summary statistic—the maximum population likelihood—as data driven measure of the goodness-of-fit of a hierarchical population model to an ensemble of gravitational-wave observations. In this work, I presented a number of different approaches to compute this summary statistic. However, their extendability to an increased number of observations and increased dimensionality was not optimal. With the rapid increase in the number of gravitational-wave observations [5, 18], more robust computation methods are needed to make this statistic a viable tool for future analysis diagnostics. Furthermore, as studies begin to probe correlations in the population distributions [195, 202, 200, 473], the maximum population likelihood needs to be computed over more dimensions. While this is computationally difficult, work is ongoing to realize these future goals.

9.3 Future applications of photon counting readout schemes

I presented the analysis framework required for a novel photon counting readout for gravitational-wave astronomy in Chapter 8. In this chapter, I framed the demonstration of the experimental design difference in the context of neutron-star post-merger remnants [89]. While such observations are an important science case for future generations of detectors [45], and will likely require these novel technological improvements, post-merger signals are not the only science case. The criteria required for photon counting to be a viable observational strategy are that the signals anticipated are very low signal-to-noise ratio, and they reside in a frequency range where third-generation detectors will be dominated by quantum shot noise. One

incredibly powerful use case for such a readout, I personally think, is for excess power-related tests of general relativity. While this is similar to signal-to-noise ratio residual tests presently undertaken [13, 14, 15], by leveraging photon counting the presence of violations from general relativity can be teased out from an ensemble of signals. The exact details of this calculation depend crucially on the structure of the basis design—requiring careful design to ensure exotic phenomena are not confused with GR deviations—though far fewer bases may be required. Future studies into the construction of such a readout for general relativistic deviations may prove as the most promising avenue to search for such violations in the third generation of detectors.

Over the past decade, gravitational-wave astronomy has evolved from an improbable observational breakthrough into a routine—yet still profoundly transformative—scientific endeavor. The regular detection of gravitational waves from compact binary coalescences has provided deep insights into the astrophysical processes shaping our Universe. At the same time, this rapid influx of data has created a pressing need for robust, scalable methods of analysis, both at the level of individual events and for population-wide inference. As the field continues to expand, the statistical techniques and approaches developed in this thesis will serve as a foundation for numerous different analyses.

BIBLIOGRAPHY

[1] J. Aasi et al. "Advanced LIGO". In: *Class. Quant. Grav.* 32 (2015), p. 074001. DOI: 10.1088/0264-9381/32/7/074001. arXiv: 1411.4547 [gr-qc].

[2] B. P. Abbott et al. "Properties of the Binary Black Hole Merger GW150914". In: *Phys. Rev. Lett.* 116.24 (2016), p. 241102. DOI: 10.1103/PhysRevLett.116.241102. arXiv: 1602.03840 [gr-qc].

[3] B. P. Abbott et al. "GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs". In: *Phys. Rev. X* 9.3 (2019), p. 031040. DOI: 10.1103/PhysRevX.9.031040. arXiv: 1811.12907 [astro-ph.HE].

[4] R. Abbott et al. "GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo during the First Half of the Third Observing Run". In: *Physical Review X* 11.2, 021053 (Apr. 2021), p. 021053. DOI: 10.1103/PhysRevX.11.021053. arXiv: 2010.14527 [gr-qc].

[5] R. Abbott et al. "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run". In: *Phys. Rev. X* 13.4 (2023), p. 041039. DOI: 10.1103/PhysRevX.13.041039. arXiv: 2111.03606 [gr-qc].

[6] B. P. Abbott et al. "Binary Black Hole Population Properties Inferred from the First and Second Observing Runs of Advanced LIGO and Advanced Virgo". In: *Astrophys. J. Lett.* 882.2 (2019), p. L24. DOI: 10.3847/2041-8213/ab3800. arXiv: 1811.12940 [astro-ph.HE].

[7] R. Abbott et al. "Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog". In: *Astrophys. J. Lett.* 913.1 (2021), p. L7. DOI: 10.3847/2041-8213/abe949. arXiv: 2010.14533 [astro-ph.HE].

[8] R. Abbott et al. "The population of merging compact binaries inferred using gravitational waves through GWTC-3". In: *arXiv e-prints*, arXiv:2111.03634 (Nov. 2021), arXiv:2111.03634. DOI: 10.48550/arXiv.2111.03634. arXiv: 2111.03634 [astro-ph.HE].

[9] B. P. Abbott et al. "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral". In: *Phys. Rev. Lett.* 119.16 (2017), p. 161101. DOI: 10.1103/PhysRevLett.119.161101. arXiv: 1710.05832 [gr-qc].

[10] B. P. Abbott et al. "GW170817: Measurements of neutron star radii and equation of state". In: *Phys. Rev. Lett.* 121.16 (2018), p. 161101. DOI: 10.1103/PhysRevLett.121.161101. arXiv: 1805.11581 [gr-qc].

[11] B. P. Abbott et al. “GW190425: Observation of a Compact Binary Coalescence with Total Mass 3.4 M ”. In: *The Astrophysical Journal* 892.1 (Mar. 2020), p. L3. ISSN: 2041-8213. DOI: 10.3847/2041-8213/ab75f5. URL: http://dx.doi.org/10.3847/2041-8213/ab75f5.

[12] B. P. Abbott et al. “Tests of General Relativity with GW170817”. In: 123.1, 011102 (July 2019), p. 011102. DOI: 10.1103/PhysRevLett.123.011102. arXiv: 1811.00364 [gr-qc].

[13] B. P. Abbott et al. “Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1”. In: 100.10, 104036 (Nov. 2019), p. 104036. DOI: 10.1103/PhysRevD.100.104036. arXiv: 1903.04467 [gr-qc].

[14] R. Abbott et al. “Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog”. In: 103.12, 122002 (June 2021), p. 122002. DOI: 10.1103/PhysRevD.103.122002. arXiv: 2010.14529 [gr-qc].

[15] R. Abbott et al. “Tests of General Relativity with GWTC-3”. In: *arXiv e-prints*, arXiv:2112.06861 (Dec. 2021), arXiv:2112.06861. DOI: 10.48550/arXiv.2112.06861. arXiv: 2112.06861 [gr-qc].

[16] F. Acernese et al. “Advanced Virgo: a second-generation interferometric gravitational wave detector”. In: *Class. Quant. Grav.* 32.2 (2015), p. 024001. DOI: 10.1088/0264-9381/32/2/024001. arXiv: 1408.3978 [gr-qc].

[17] E. Capote et al. “Advanced LIGO detector performance in the fourth observing run”. In: *Phys. Rev. D* 111.6 (2025), p. 062002. DOI: 10.1103/PhysRevD.111.062002. arXiv: 2411.14607 [gr-qc].

[18] R. Abbott et al. *GraceDB | LVK Public Alerts —gracedb.ligo.org*. https://gracedb.ligo.org/superevents/public/O4/. [Accessed 13-06-2025]. 2025.

[19] F. S. Broekgaarden, S. Banagiri, and E. Payne. “Visualizing the Number of Existing and Future Gravitational-wave Detections from Merging Double Compact Objects”. In: *Astrophys. J.* 969.2 (2024), p. 108. DOI: 10.3847/1538-4357/ad4709. arXiv: 2303.17628 [astro-ph.HE].

[20] K. Chatziioannou, T. Dent, M. Fishbach, F. Ohme, M. Pürrer, V. Raymond, and J. Veitch. “Compact binary coalescences: gravitational-wave astronomy with ground-based detectors”. In: (Sept. 2024). arXiv: 2409.02037 [gr-qc].

[21] A. Einstein. “The foundation of the general theory of relativity.” In: *Annalen Phys.* 49.7 (1916). Ed. by J.-P. Hsu and D. Fine, pp. 769–822. DOI: 10.1002/andp.19163540702.

[22] C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation*. San Francisco: W. H. Freeman, 1973. ISBN: 978-0-7167-0344-0, 978-0-691-17779-3.

[23] F. W. Dyson, A. S. Eddington, and C. Davidson. "A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919". In: *Phil. Trans. Roy. Soc. Lond. A* 220 (1920), pp. 291–333. DOI: 10.1098/rsta.1920.0009.

[24] I. I. Shapiro. "Solar system tests of general relativity: Recent results and present plans". In: *General relativity and gravitation* (1990), pp. 313–330.

[25] B. Mashhoon, F. W. Hehl, and D. S. Theiss. "On the gravitational effects of rotating masses: the Thirring-Lense papers." In: *General Relativity and Gravitation* 16.8 (Aug. 1984), pp. 711–750. DOI: 10.1007/BF00762913.

[26] A. Einstein. "Über Gravitationswellen". In: *Sitzungsber. K. Preuss. Akad. Wiss.* 1 (1918), pp. 154–167.

[27] M. Maggiore. *Gravitational Waves. Vol. 1: Theory and Experiments*. Oxford University Press, 2007. ISBN: 978-0-19-171766-6, 978-0-19-852074-0. DOI: 10.1093/acprof:oso/9780198570745.001.0001.

[28] K. S. Thorne. "Multipole Expansions of Gravitational Radiation". In: *Rev. Mod. Phys.* 52 (1980), pp. 299–339. DOI: 10.1103/RevModPhys.52.299.

[29] R. Abbott et al. "GW190412: Observation of a Binary-Black-Hole Coalescence with Asymmetric Masses". In: *Phys. Rev. D* 102.4 (2020), p. 043015. DOI: 10.1103/PhysRevD.102.043015. arXiv: 2004.08342 [astro-ph.HE].

[30] P. Linsay, P. Saulson, R. Weiss, and S. Whitcomb. "A study of a long baseline gravitational wave antenna system". In: *National Science Foundation Document* (1983).

[31] J. Abadie et al. "Calibration of the LIGO Gravitational Wave Detectors in the Fifth Science Run". In: *Nucl. Instrum. Meth. A* 624 (2010), pp. 223–240. DOI: 10.1016/j.nima.2010.07.089. arXiv: 1007.3973 [gr-qc].

[32] C. Cahillane et al. "Calibration uncertainty for Advanced LIGO's first and second observing runs". In: 96.10, 102001 (Nov. 2017), p. 102001. DOI: 10.1103/PhysRevD.96.102001. arXiv: 1708.03023 [astro-ph.IM].

[33] L. Sun et al. "Characterization of systematic error in Advanced LIGO calibration". In: *Class. Quant. Grav.* 37.22 (2020), p. 225008. DOI: 10.1088/1361-6382/abb14e. arXiv: 2005.02531 [astro-ph.IM].

[34] C. Cahillane and G. Mansell. "Review of the Advanced LIGO Gravitational Wave Observatories Leading to Observing Run Four". In: *Galaxies* 10.1 (2022), p. 36. DOI: 10.3390/galaxies10010036. arXiv: 2202.00847 [gr-qc].

[35] D. Davis and M. Walker. "Detector Characterization and Mitigation of Noise in Ground-Based Gravitational-Wave Interferometers". In: *Galaxies* 10.1 (2022), p. 12. DOI: 10.3390/galaxies10010012.

[36] G. Ashton et al. “BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy”. In: *Astrophys. J. Suppl.* 241.2 (2019), p. 27. DOI: 10.3847/1538-4365/ab06fc. arXiv: 1811.02042 [astro-ph.IM].

[37] J. Roulet and T. Venumadhav. “Inferring Binary Properties from Gravitational Wave Signals”. In: (Feb. 2024). DOI: 10.1146/annurev-nucl-121423-100725. arXiv: 2402.11439 [gr-qc].

[38] D. Davis, T. B. Littenberg, I. M. Romero-Shaw, M. Millhouse, J. McIver, F. Di Renzo, and G. Ashton. “Subtracting glitches from gravitational-wave detector data during the third LIGO-Virgo observing run”. In: *Class. Quant. Grav.* 39.24 (2022), p. 245013. DOI: 10.1088/1361-6382/aca238. arXiv: 2207.03429 [astro-ph.IM].

[39] S. Hourihane, K. Chatziioannou, M. Wijngaarden, D. Davis, T. Littenberg, and N. Cornish. “Accurate modeling and mitigation of overlapping signals and glitches in gravitational-wave data”. In: (May 2022). arXiv: 2205.13580 [gr-qc].

[40] S. Ghonge, J. Brandt, J. M. Sullivan, M. Millhouse, K. Chatziioannou, J. A. Clark, T. Littenberg, N. Cornish, S. Hourihane, and L. Cadonati. “Assessing and mitigating the impact of glitches on gravitational-wave parameter estimation: A model agnostic approach”. In: *Phys. Rev. D* 110.12 (2024), p. 122002. DOI: 10.1103/PhysRevD.110.122002. arXiv: 2311.09159 [gr-qc].

[41] M. Pürrer and C.-J. Haster. “Gravitational waveform accuracy requirements for future ground-based detectors”. In: *Phys. Rev. Res.* 2.2 (2020), p. 023151. DOI: 10.1103/PhysRevResearch.2.023151. arXiv: 1912.10055 [gr-qc].

[42] R. Gamba, M. Breschi, S. Bernuzzi, M. Agathos, and A. Nagar. “Waveform systematics in the gravitational-wave inference of tidal parameters and equation of state from binary neutron star signals”. In: *Phys. Rev. D* 103.12 (2021), p. 124015. DOI: 10.1103/PhysRevD.103.124015. arXiv: 2009.08467 [gr-qc].

[43] A. Puecher, A. Samajdar, G. Ashton, C. Van Den Broeck, and T. Dietrich. “Comparing gravitational waveform models for binary black hole mergers through a hypermodels approach”. In: *Phys. Rev. D* 109.2 (2024), p. 023019. DOI: 10.1103/PhysRevD.109.023019. arXiv: 2310.03555 [gr-qc].

[44] J. Golomb, I. Legred, K. Chatziioannou, and P. Landry. “Interplay of astrophysics and nuclear physics in determining the properties of neutron stars”. In: *Phys. Rev. D* 111.2 (2025), p. 023029. DOI: 10.1103/PhysRevD.111.023029. arXiv: 2410.14597 [astro-ph.HE].

[45] M. Evans et al. “A Horizon Study for Cosmic Explorer: Science, Observatories, and Community”. In: *arXiv eprints* (Sept. 2021). arXiv: 2109.09882 [astro-ph.IM].

[46] M. Evans et al. “Cosmic Explorer: A Submission to the NSF MPSAC ngGW Subcommittee”. In: *arXiv eprints* (June 2023). arXiv: 2306.13745 [astro-ph.IM].

[47] M. Branchesi et al. “Science with the Einstein Telescope: a comparison of different designs”. In: *JCAP* 2023.07 (2023), p. 068. DOI: 10.1088/1475-7516/2023/07/068. arXiv: 2303.15923 [gr-qc].

[48] D. Wands. “Extended gravity theories and the Einstein-Hilbert action”. In: *Class. Quant. Grav.* 11 (1994), pp. 269–280. DOI: 10.1088/0264-9381/11/1/025. arXiv: gr-qc/9307034.

[49] C. Brans and R. H. Dicke. “Mach’s principle and a relativistic theory of gravitation”. In: *Phys. Rev.* 124 (1961). Ed. by J.-P. Hsu and D. Fine, pp. 925–935. DOI: 10.1103/PhysRev.124.925.

[50] C. de Rham, G. Gabadadze, and A. J. Tolley. “Resummation of Massive Gravity”. In: *Phys. Rev. Lett.* 106 (2011), p. 231101. DOI: 10.1103/PhysRevLett.106.231101. arXiv: 1011.1232 [hep-th].

[51] C. de Rham, G. Gabadadze, and A. J. Tolley. “Helicity decomposition of ghost-free massive gravity”. In: *JHEP* 11 (2011), p. 093. DOI: 10.1007/JHEP11(2011)093. arXiv: 1108.4521 [hep-th].

[52] C. M. Will. “Bounding the mass of the graviton using gravitational wave observations of inspiralling compact binaries”. In: *Phys. Rev. D* 57 (1998), pp. 2061–2068. DOI: 10.1103/PhysRevD.57.2061. arXiv: gr-qc/9709011.

[53] C. de Rham, J. T. Deskins, A. J. Tolley, and S.-Y. Zhou. “Graviton Mass Bounds”. In: *Rev. Mod. Phys.* 89.2 (2017), p. 025004. DOI: 10.1103/RevModPhys.89.025004. arXiv: 1606.08462 [astro-ph.CO].

[54] P. G. S. Fernandes, P. Carrilho, T. Clifton, and D. J. Mulryne. “The 4D Einstein–Gauss–Bonnet theory of gravity: a review”. In: *Class. Quant. Grav.* 39.6 (2022), p. 063001. DOI: 10.1088/1361-6382/ac500a. arXiv: 2202.13908 [gr-qc].

[55] P. Kanti, N. E. Mavromatos, J. Rizos, K. Tamvakis, and E. Winstanley. “Dilatonic black holes in higher curvature string gravity”. In: *Phys. Rev. D* 54 (1996), pp. 5049–5058. DOI: 10.1103/PhysRevD.54.5049. arXiv: hep-th/9511071.

[56] N. Yunes and F. Pretorius. “Dynamical Chern-Simons Modified Gravity. I. Spinning Black Holes in the Slow-Rotation Approximation”. In: *Phys. Rev. D* 79 (2009), p. 084043. DOI: 10.1103/PhysRevD.79.084043. arXiv: 0902.4669 [gr-qc].

[57] D. Lovelock. “The Einstein tensor and its generalizations”. In: *J. Math. Phys.* 12 (1971), pp. 498–501. DOI: 10.1063/1.1665613.

[58] S. Endlich, V. Gorbenko, J. Huang, and L. Senatore. "An effective formalism for testing extensions to General Relativity with gravitational waves". In: *JHEP* 2017.09 (2017), p. 122. DOI: 10.1007/JHEP09(2017)122. arXiv: 1704.01590 [gr-qc].

[59] P. Bueno, P. A. Cano, V. S. Min, and M. R. Visser. "Aspects of general higher-order gravities". In: *Physical Review D* 95.4 (Feb. 2017). ISSN: 2470-0029. DOI: 10.1103/physrevd.95.044010. URL: http://dx.doi.org/10.1103/PhysRevD.95.044010.

[60] P. A. Cano, K. Fransen, T. Hertog, and S. Maenaut. "Gravitational ringing of rotating black holes in higher-derivative gravity". In: *Phys. Rev. D* 105.2 (2022), p. 024064. DOI: 10.1103/PhysRevD.105.024064. arXiv: 2110.11378 [gr-qc].

[61] D. Psaltis, C. Talbot, E. Payne, and I. Mandel. "Probing the Black Hole Metric. I. Black Hole Shadows and Binary Black-Hole Inspirals". In: *Phys. Rev. D* 103 (2021), p. 104036. DOI: 10.1103/PhysRevD.103.104036. arXiv: 2012.02117 [gr-qc].

[62] C. M. Will. "The Confrontation between General Relativity and Experiment". In: *Living Rev. Rel.* 17 (2014), p. 4. DOI: 10.12942/lrr-2014-4. arXiv: 1403.7377 [gr-qc].

[63] K. Akiyama et al. "First Sagittarius A* Event Horizon Telescope Results. VI. Testing the Black Hole Metric". In: *Astrophys. J. Lett.* 930.2 (2022), p. L17. DOI: 10.3847/2041-8213/ac6756. arXiv: 2311.09484 [astro-ph.HE].

[64] K. Akiyama et al. "First M87 Event Horizon Telescope Results. VI. The Shadow and Mass of the Central Black Hole". In: *Astrophys. J. Lett.* 875.1 (2019), p. L6. DOI: 10.3847/2041-8213/ab1141. arXiv: 1906.11243 [astro-ph.GA].

[65] M. Okounkova, M. Isi, K. Chatziioannou, and W. M. Farr. "Gravitational wave inference on a numerical-relativity simulation of a black hole merger beyond general relativity". In: *Phys. Rev. D* 107.2 (2023), p. 024046. DOI: 10.1103/PhysRevD.107.024046. arXiv: 2208.02805 [gr-qc].

[66] N. Yunes and F. Pretorius. "Fundamental Theoretical Bias in Gravitational Wave Astrophysics and the Parameterized Post-Einsteinian Framework". In: *Phys. Rev. D* 80 (2009), p. 122003. DOI: 10.1103/PhysRevD.80.122003. arXiv: 0909.3328 [gr-qc].

[67] L. Blanchet. "Gravitational Radiation from Post-Newtonian Sources and Inspiralling Compact Binaries". In: *Living Reviews in Relativity* 17.1, 2 (Dec. 2014), p. 2. DOI: 10.12942/lrr-2014-2. arXiv: 1310.1528 [gr-qc].

[68] K. G. Arun, B. R. Iyer, B. S. Sathyaprakash, and P. A. Sundararajan. "Parameter estimation of inspiralling compact binaries using 3.5 post-Newtonian gravitational wave phasing: The nonspinning case". In: 71.8, 084008 (Apr. 2005), p. 084008. DOI: 10.1103/PhysRevD.71.084008. arXiv: gr-qc/0411146 [gr-qc].

[69] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio. "Towards a generic test of the strong field dynamics of general relativity using compact binary coalescence". In: *Phys. Rev. D* 85 (2012), p. 082003. DOI: 10.1103/PhysRevD.85.082003. arXiv: 1110.0530 [gr-qc].

[70] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale. "TIGER: A data analysis pipeline for testing the strong-field dynamics of general relativity with gravitational wave signals from coalescing compact binaries". In: 89.8, 082001 (Apr. 2014), p. 082001. DOI: 10.1103/PhysRevD.89.082001. arXiv: 1311.0420 [gr-qc].

[71] K. Chatziioannou, A. Klein, N. Yunes, and N. Cornish. "Constructing gravitational waves from generic spin-precessing compact binary inspirals". In: 95.10, 104004 (May 2017), p. 104004. DOI: 10.1103/PhysRevD.95.104004. arXiv: 1703.03967 [gr-qc].

[72] K. Chatziioannou, M. Isi, C.-J. Haster, and T. B. Littenberg. "Morphology-independent test of the mixed polarization content of transient gravitational wave signals". In: *Phys. Rev. D* 104.4 (2021), p. 044005. DOI: 10.1103/PhysRevD.104.044005. arXiv: 2105.01521 [gr-qc].

[73] T. C. K. Ng, M. Isi, K. W. K. Wong, and W. M. Farr. *Constraining gravitational wave amplitude birefringence with GWTC-3*. May 2023. arXiv: 2305.05844 [gr-qc].

[74] E. Thrane and C. Talbot. "An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models". In: 36, e010 (Mar. 2019), e010. DOI: 10.1017/pasa.2019.2. arXiv: 1809.02293 [astro-ph.IM].

[75] E. Payne, S. Hourihane, J. Golomb, R. Udall, R. Udall, D. Davis, and K. Chatziioannou. "Curious case of GW200129: Interplay between spin-precession inference and data-quality issues". In: *Phys. Rev. D* 106.10 (2022), p. 104017. DOI: 10.1103/PhysRevD.106.104017. arXiv: 2206.11932 [gr-qc].

[76] M. Hannam et al. "General-relativistic precession in a black-hole binary". In: *Nature* 610.7933 (2022), pp. 652–655. DOI: 10.1038/s41586-022-05212-z. arXiv: 2112.11300 [gr-qc].

[77] R. Abbott et al. "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run". In:

Phys. Rev. X 13.4 (2023), p. 041039. DOI: 10.1103/PhysRevX.13.041039. arXiv: 2111.03606 [gr-qc].

[78] E. Payne, K. Kremer, and M. Zevin. "Spin Doctors: How to Diagnose a Hierarchical Merger Origin". In: *Astrophys. J. Lett.* 966.1 (2024), p. L16. DOI: 10.3847/2041-8213/ad3e82. arXiv: 2402.15066 [gr-qc].

[79] A. Buonanno, L. E. Kidder, and L. Lehner. "Estimating the final spin of a binary black hole coalescence". In: 77.2, 026004 (Jan. 2008), p. 026004. DOI: 10.1103/PhysRevD.77.026004. arXiv: 0709.3839 [astro-ph].

[80] M. Fishbach, D. E. Holz, and B. Farr. "Are LIGO's Black Holes Made from Smaller Black Holes?" In: 840.2, L24 (May 2017), p. L24. DOI: 10.3847/2041-8213/aa7045. arXiv: 1703.06869 [astro-ph.HE].

[81] C. L. Rodriguez, M. Zevin, P. Amaro-Seoane, S. Chatterjee, K. Kremer, F. A. Rasio, and C. S. Ye. "Black holes: The next generation—repeated mergers in dense star clusters and their gravitational-wave properties". In: 100.4, 043027 (Aug. 2019), p. 043027. DOI: 10.1103/PhysRevD.100.043027. arXiv: 1906.10260 [astro-ph.HE].

[82] K. Kremer, C. S. Ye, N. Z. Rui, N. C. Weatherford, S. Chatterjee, G. Fragione, C. L. Rodriguez, M. Spera, and F. A. Rasio. "Modeling Dense Star Clusters in the Milky Way and Beyond with the CMC Cluster Catalog". In: 247.2, 48 (Apr. 2020), p. 48. DOI: 10.3847/1538-4365/ab7919. arXiv: 1911.00018 [astro-ph.HE].

[83] C. L. Rodriguez et al. "Modeling Dense Star Clusters in the Milky Way and beyond with the Cluster Monte Carlo Code". In: 258.2, 22 (Feb. 2022), p. 22. DOI: 10.3847/1538-4365/ac2edf. arXiv: 2106.02643 [astro-ph.GA].

[84] E. Payne, M. Isi, K. Chatziioannou, and W. M. Farr. "Fortifying gravitational-wave tests of general relativity against astrophysical assumptions". In: *Phys. Rev. D* 108.12 (2023), p. 124060. DOI: 10.1103/PhysRevD.108.124060. arXiv: 2309.04528 [gr-qc].

[85] R. Magee, M. Isi, E. Payne, K. Chatziioannou, W. M. Farr, G. Pratten, and S. Vitale. "Impact of selection biases on tests of general relativity with gravitational-wave inspirals". In: *Phys. Rev. D* 109.2 (2024), p. 023014. DOI: 10.1103/PhysRevD.109.023014. arXiv: 2311.03656 [gr-qc].

[86] E. Payne, M. Isi, K. Chatziioannou, L. Lehner, Y. Chen, and W. M. Farr. "Curvature Dependence of Gravitational-Wave Tests of General Relativity". In: *Phys. Rev. Lett.* 133.25 (2024), p. 251401. DOI: 10.1103/PhysRevLett.133.251401. arXiv: 2407.07043 [gr-qc].

[87] J. Donoghue. "Quantum gravity as a low energy effective field theory". In: *Scholarpedia* 12.4 (2017), p. 32997. DOI: 10.4249/scholarpedia.32997.

[88] C. P. Burgess. *Introduction to Effective Field Theory*. Cambridge University Press, Dec. 2020. ISBN: 978-1-139-04804-0, 978-0-521-19547-8. DOI: 10.1017/9781139048040.

[89] J. A. Clark, A. Bauswein, N. Stergioulas, and D. Shoemaker. "Observing Gravitational Waves From The Post-Merger Phase Of Binary Neutron Star Coalescence". In: *Class. Quant. Grav.* 33.8 (2016), p. 085003. DOI: 10.1088/0264-9381/33/8/085003. arXiv: 1509.08522 [astro-ph.HE].

[90] M. Breschi, R. Gamba, S. Borhanian, G. Carullo, and S. Bernuzzi. "Kilohertz Gravitational Waves from Binary Neutron Star Mergers: Inference of Postmerger Signals with the Einstein Telescope". In: (May 2022). arXiv: 2205.09979 [gr-qc].

[91] C. M. Caves. "Quantum-mechanical noise in an interferometer". In: *Phys. Rev. D* 23 (8 Apr. 1981), pp. 1693–1708. DOI: 10.1103/PhysRevD.23.1693. URL: https://link.aps.org/doi/10.1103/PhysRevD.23.1693.

[92] L. McCuller. "Single-Photon Signal Sideband Detection for High-Power Michelson Interferometers". In: (Nov. 2022). arXiv: 2211.04016 [physics.ins-det].

[93] G. Pratten et al. "Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes". In: *Phys. Rev. D* 103.10 (2021), p. 104056. DOI: 10.1103/PhysRevD.103.104056. arXiv: 2004.06503 [gr-qc].

[94] S. Ossokine et al. "Multipolar Effective-One-Body Waveforms for Precessing Binary Black Holes: Construction and Validation". In: *Phys. Rev. D* 102.4 (2020), p. 044055. DOI: 10.1103/PhysRevD.102.044055. arXiv: 2004.09442 [gr-qc].

[95] I. M. Romero-Shaw et al. "Bayesian inference for compact binary coalescences with bilby: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue". In: *Mon. Not. Roy. Astron. Soc.* 499.3 (2020), pp. 3295–3319. DOI: 10.1093/mnras/staa2850. arXiv: 2006.00714 [astro-ph.IM].

[96] D. Wysocki, R. O'Shaughnessy, Y. L. Fang, and J. Lange. "Accelerating parameter inference with graphics processing units". In: *arXiv e-prints* (Feb. 2019). arXiv: 1902.04934 [astro-ph.IM].

[97] G. Pratten, P. Schmidt, R. Buscicchio, and L. M. Thomas. "Measuring precession in asymmetric compact binaries". In: *Phys. Rev. Res.* 2.4 (2020), p. 043096. DOI: 10.1103/PhysRevResearch.2.043096. arXiv: 2006.16153 [gr-qc].

[98] G. Ashton, S. Thiele, Y. Lecoeuche, J. McIver, and L. K. Nuttall. "Parameterised population models of transient non-Gaussian noise in the LIGO gravitational-wave detectors". In: *Class. Quant. Grav.* 39.17 (2022), p. 175004. DOI: 10.1088/1361-6382/ac8094. arXiv: 2110.02689 [gr-qc].

[99] V. Varma, S. Biscoveanu, T. Islam, F. H. Shaik, C.-J. Haster, M. Isi, W. M. Farr, S. E. Field, and S. Vitale. "Evidence of large recoil velocity from a black hole merger signal". In: (Jan. 2022). arXiv: 2201.01302 [astro-ph.HE].

[100] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer. "Surrogate models for precessing binary black hole simulations with unequal masses". In: *Phys. Rev. Research.* 1 (2019), p. 033015. DOI: 10.1103/PhysRevResearch.1.033015. arXiv: 1905.09300 [gr-qc].

[101] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne. "Spin induced orbital precession and its modulation of the gravitational wave forms from merging binaries". In: *Phys.Rev.* D49 (1994), pp. 6274–6297. DOI: 10.1103/PhysRevD.49.6274.

[102] L. E. Kidder. "Coalescing binary systems of compact objects to postNewtonian 5/2 order. 5. Spin effects". In: *Phys. Rev. D* 52 (1995), pp. 821–847. DOI: 10.1103/PhysRevD.52.821. arXiv: gr-qc/9506022.

[103] A. Buonanno, Y.-b. Chen, and M. Vallisneri. "Detecting gravitational waves from precessing binaries of spinning compact objects: Adiabatic limit". In: *Phys. Rev.* D67 (2003). [Erratum: Phys. Rev.D74,029904(2006)], p. 104025. DOI: 10.1103/PhysRevD.67.104025,10.1103/PhysRevD.74.029904. arXiv: gr-qc/0211087 [gr-qc].

[104] A. Buonanno, Y.-b. Chen, Y. Pan, and M. Vallisneri. "A Quasi-physical family of gravity-wave templates for precessing binaries of spinning compact objects. 2. Application to double-spin precessing binaries". In: *Phys. Rev. D* 70 (2004). [Erratum: Phys.Rev.D 74, 029902 (2006)], p. 104003. DOI: 10.1103/PhysRevD.74.029902. arXiv: gr-qc/0405090.

[105] P. Schmidt, M. Hannam, S. Husa, and P. Ajith. "Tracking the precession of compact binaries from their gravitational-wave signal". In: *Phys. Rev. D* 84 (2011), p. 024046. DOI: 10.1103/PhysRevD.84.024046. arXiv: 1012.2879 [gr-qc].

[106] P. Schmidt, M. Hannam, and S. Husa. "Towards models of gravitational waveforms from generic binaries: A simple approximate mapping between precessing and non-precessing inspiral signals". In: *Phys. Rev. D* 86 (2012), p. 104063. DOI: 10.1103/PhysRevD.86.104063. arXiv: 1207.3088 [gr-qc].

[107] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer. "Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms". In: 113.15, 151101 (Oct. 2014), p. 151101. DOI: 10.1103/PhysRevLett.113.151101. arXiv: 1308.3271 [gr-qc].

[108] K. Chatziioannou, A. Klein, N. Cornish, and N. Yunes. "Analytic Gravitational Waveforms for Generic Precessing Binary Inspirals". In: *Phys. Rev.*

Lett. 118.5 (2017), p. 051101. DOI: 10.1103/PhysRevLett.118.051101. arXiv: 1606.03117 [gr-qc].

[109] D. Gerosa, M. Kesden, U. Sperhake, E. Berti, and R. O'Shaughnessy. "Multi-timescale analysis of phase transitions in precessing black-hole binaries". In: *Phys. Rev. D* 92 (2015), p. 064016. DOI: 10.1103/PhysRevD.92.064016. arXiv: 1506.03492 [gr-qc].

[110] M. Kesden, D. Gerosa, R. O'Shaughnessy, E. Berti, and U. Sperhake. "Effective potentials and morphological transitions for binary black-hole spin precession". In: *Phys. Rev. Lett.* 114.8 (2015), p. 081103. DOI: 10.1103/PhysRevLett.114.081103. arXiv: 1411.0674 [gr-qc].

[111] A. Ramos-Buades, P. Schmidt, G. Pratten, and S. Husa. "Validity of common modeling approximations for precessing binary black holes with higher-order modes". In: *Phys. Rev. D* 101.10 (2020), p. 103014. DOI: 10.1103/PhysRevD.101.103014. arXiv: 2001.10936 [gr-qc].

[112] S. Fairhurst, R. Green, C. Hoy, M. Hannam, and A. Muir. "Two-harmonic approximation for gravitational waveforms from precessing binaries". In: *Phys. Rev. D* 102.2 (2020), p. 024055. DOI: 10.1103/PhysRevD.102.024055. arXiv: 1908.05707 [gr-qc].

[113] R. O'Shaughnessy, L. London, J. Healy, and D. Shoemaker. "Precession during merger: Strong polarization changes are observationally accessible features of strong-field gravity during binary black hole merger". In: *Phys. Rev. D* 87.4 (2013), p. 044038. DOI: 10.1103/PhysRevD.87.044038. arXiv: 1209.3712 [gr-qc].

[114] S. Biscoveanu, M. Isi, V. Varma, and S. Vitale. "Measuring the spins of heavy binary black holes". In: *Phys. Rev. D* 104.10 (2021), p. 103018. DOI: 10.1103/PhysRevD.104.103018. arXiv: 2106.06492 [gr-qc].

[115] N. J. Cornish and T. B. Littenberg. "BayesWave: Bayesian Inference for Gravitational Wave Bursts and Instrument Glitches". In: *Class. Quant. Grav.* 32.13 (2015), p. 135012. DOI: 10.1088/0264-9381/32/13/135012. arXiv: 1410.3835 [gr-qc].

[116] T. B. Littenberg and N. J. Cornish. "Bayesian inference for spectral estimation of gravitational wave detector noise". In: *Phys. Rev.* D91.8 (2015), p. 084034. DOI: 10.1103/PhysRevD.91.084034. arXiv: 1410.3852 [gr-qc].

[117] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse. "BayesWave analysis pipeline in the era of gravitational wave observations". In: *Phys. Rev. D* 103.4 (2021), p. 044006. DOI: 10.1103/PhysRevD.103.044006. arXiv: 2011.09494 [gr-qc].

[118] K. Chatziioannou, N. Cornish, M. Wijngaarden, and T. B. Littenberg. "Modeling compact binary signals and instrumental glitches in gravitational wave data". In: *Phys. Rev. D* 103.4 (2021), p. 044013. DOI: 10.1103/PhysRevD.103.044013. arXiv: 2101.01200 [gr-qc].

[119] D. Davis, T. J. Massinger, A. P. Lundgren, J. C. Driggers, A. L. Urban, and L. K. Nuttall. "Improving the Sensitivity of Advanced LIGO Using Noise Subtraction". In: *Class. Quant. Grav.* 36.5 (2019), p. 055011. DOI: 10.1088/1361-6382/ab01c5. arXiv: 1809.05348 [astro-ph.IM].

[120] L. C. Stein. "qnm: A Python package for calculating Kerr quasinormal modes, separation constants, and spherical-spheroidal mixing coefficients". In: *J. Open Source Softw.* 4.42 (2019), p. 1683. DOI: 10.21105/joss.01683. arXiv: 1908.10377 [gr-qc].

[121] V. Varma, D. Gerosa, L. C. Stein, F. Hébert, and H. Zhang. "High-accuracy mass, spin, and recoil predictions of generic black-hole merger remnants". In: *Phys. Rev. Lett.* 122.1 (2019), p. 011101. DOI: 10.1103/PhysRevLett.122.011101. arXiv: 1809.09125 [gr-qc].

[122] C. Cutler and E. E. Flanagan. "Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral wave form?" In: *Phys. Rev. D* 49 (1994), pp. 2658–2697. DOI: 10.1103/PhysRevD.49.2658. arXiv: gr-qc/9402014.

[123] K. Haris, A. K. Mehta, S. Kumar, T. Venumadhav, and P. Ajith. "Identifying strongly lensed gravitational wave signals from binary black hole mergers". In: (July 2018). arXiv: 1807.07062 [gr-qc].

[124] O. A. Hannuksela, K. Haris, K. K. Y. Ng, S. Kumar, A. K. Mehta, D. Keitel, T. G. F. Li, and P. Ajith. "Search for gravitational lensing signatures in LIGO-Virgo binary black hole events". In: *Astrophys. J. Lett.* 874.1 (2019), p. L2. DOI: 10.3847/2041-8213/ab0c0f. arXiv: 1901.02674 [gr-qc].

[125] F. Robinet, N. Arnaud, N. Leroy, A. Lundgren, D. Macleod, and J. McIver. "Omicron: a tool to characterize transient noise in gravitational-wave detectors". In: *SoftwareX* 12 (2020), p. 100620. DOI: 10.1016/j.softx.2020.100620. arXiv: 2007.11374 [astro-ph.IM].

[126] S. Chatterji, L. Blackburn, G. Martin, and E. Katsavounidis. "Multiresolution techniques for the detection of gravitational-wave bursts". In: *Class. Quant. Grav.* 21 (2004), S1809–S1818. DOI: 10.1088/0264-9381/21/20/024. arXiv: gr-qc/0412119.

[127] D. M. Macleod, J. S. Areeda, S. B. Coughlin, T. J. Massinger, and A. L. Urban. "GWpy: A Python package for gravitational-wave astrophysics". In: *SoftwareX* 13, 100657 (Jan. 2021), p. 100657. DOI: 10.1016/j.softx.2021.100657.

[128] T. Accadia et al. “Noise from scattered light in Virgo’s second science run data”. In: *Class. Quant. Grav.* 27 (2010). Ed. by F. Ricci, p. 194011. DOI: 10.1088/0264-9381/27/19/194011.

[129] A. Longo, S. Bianchi, W. Plastino, N. Arnaud, A. Chiummo, I. Fiori, B. Swinkels, and M. Was. “Scattered light noise characterisation at the Virgo interferometer with tvf-EMD adaptive algorithm”. In: *Class. Quant. Grav.* 37.14 (2020), p. 145011. DOI: 10.1088/1361-6382/ab9719. arXiv: 2002.10529 [astro-ph.IM].

[130] A. Longo, S. Bianchi, G. Valdes, N. Arnaud, and W. Plastino. “Daily monitoring of scattered light noise due to microseismic variability at the Virgo interferometer”. In: *Class. Quant. Grav.* 39.3 (2022), p. 035001. DOI: 10.1088/1361-6382/ac4117. arXiv: 2112.06046 [astro-ph.IM].

[131] F. Acernese et al. “Virgo Detector Characterization and Data Quality during the O3 run”. In: (May 2022). arXiv: 2205.01555 [gr-qc].

[132] S. Soni et al. “Reducing scattered light in LIGO’s third observing run”. In: *arXiv e-prints*, arXiv:2007.14876 (July 2020), arXiv:2007.14876. arXiv: 2007.14876 [astro-ph.IM].

[133] 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”. In: *Phys. Rev. D* 93.4 (2016), p. 044006. DOI: 10.1103/PhysRevD.93.044006. arXiv: 1508.07250 [gr-qc].

[134] 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”. In: *Phys. Rev. D* 93.4 (2016), p. 044007. DOI: 10.1103/PhysRevD.93.044007. arXiv: 1508.07253 [gr-qc].

[135] E. Payne, S. Hourihane, J. Golomb, R. Udall, D. Davis, and K. Chatziioannou. *Data Release for the curious case of GW200129: interplay between spin-precession inference and data-quality issues*. 2022. DOI: 10.5281/zenodo.7259655. URL: https://zenodo.org/record/7259655.

[136] R. Abbott et al. “Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo”. In: (Dec. 2019). arXiv: 1912.11716 [gr-qc].

[137] LIGO Scientific Collaboration and Virgo Collaboration. In: *GCN* 26926 (2020). URL: https://gcn.gsfc.nasa.gov/other/S200129m.gcn3.

[138] C. Messick et al. “Analysis Framework for the Prompt Discovery of Compact Binary Mergers in Gravitational-wave Data”. In: *Phys. Rev. D* 95.4 (2017), p. 042001. DOI: 10.1103/PhysRevD.95.042001. arXiv: 1604.04324 [astro-ph.IM].

[139] C. Hanna et al. "Fast evaluation of multidetector consistency for real-time gravitational wave searches". In: *Phys. Rev. D* 101.2 (2020), p. 022003. DOI: 10.1103/PhysRevD.101.022003. arXiv: 1901.02227 [gr-qc].

[140] S. Klimenko et al. "Method for detection and reconstruction of gravitational wave transients with networks of advanced detectors". In: *Phys. Rev. D* 93.4 (2016), p. 042004. DOI: 10.1103/PhysRevD.93.042004. arXiv: 1511.05999 [gr-qc].

[141] A. H. Nitz, T. Dal Canton, D. Davis, and S. Reyes. "Rapid detection of gravitational waves from compact binary mergers with PyCBC Live". In: *Phys. Rev. D* 98.2 (2018), p. 024050. DOI: 10.1103/PhysRevD.98.024050. arXiv: 1805.11174 [gr-qc].

[142] T. Dal Canton, A. H. Nitz, B. Gadre, G. S. Cabourn Davies, V. Villa-Ortega, T. Dent, I. Harry, and L. Xiao. "Real-time Search for Compact Binary Mergers in Advanced LIGO and Virgo's Third Observing Run Using PyCBC Live". In: *Astrophys. J.* 923.2 (2021), p. 254. DOI: 10.3847/1538-4357/ac2f9a. arXiv: 2008.07494 [astro-ph.HE].

[143] T. Adams, D. Buskulic, V. Germain, G. M. Guidi, F. Marion, M. Montani, B. Mours, F. Piergiovanni, and G. Wang. "Low-latency analysis pipeline for compact binary coalescences in the advanced gravitational wave detector era". In: *Class. Quant. Grav.* 33.17 (2016), p. 175012. DOI: 10.1088/0264-9381/33/17/175012. arXiv: 1512.02864 [gr-qc].

[144] Q. Chu et al. "SPIIR online coherent pipeline to search for gravitational waves from compact binary coalescences". In: *Phys. Rev. D* 105.2 (2022), p. 024023. DOI: 10.1103/PhysRevD.105.024023. arXiv: 2011.06787 [gr-qc].

[145] B. P. Abbott et al. "Characterization of transient noise in Advanced LIGO relevant to gravitational wave signal GW150914". In: *Class. Quant. Grav.* 33.13 (2016), p. 134001. DOI: 10.1088/0264-9381/33/13/134001. arXiv: 1602.03844 [gr-qc].

[146] D. Davis et al. *Data Quality Vetoes Applied to the Analysis of LIGO Data from the Third Observing Run*. Tech. rep. DCC-T2100045. LIGO, 2021. URL: https://dcc.ligo.org/LIGO-T2100045/public.

[147] D. Davis et al. "LIGO detector characterization in the second and third observing runs". In: *Class. Quant. Grav.* 38.13 (2021), p. 135014. DOI: 10.1088/1361-6382/abfd85. arXiv: 2101.11673 [astro-ph.IM].

[148] A. H. Nitz, T. Dent, T. D. Canton, S. Fairhurst, and D. A. Brown. "Detecting Binary Compact-object Mergers with Gravitational Waves: Understanding and Improving the Sensitivity of the PyCBC Search". In: *The Astrophysical Journal* 849.2 (Nov. 2017), p. 118. DOI: 10.3847/1538-4357/aa8f50. URL: https://doi.org/10.3847%2F1538-4357%2Faa8f50.

[149] G. S. Davies, T. Dent, M. Tápai, I. Harry, C. McIsaac, and A. H. Nitz. "Extending the PyCBC search for gravitational waves from compact binary mergers to a global network". In: *Phys. Rev. D* 102.2 (2020), p. 022004. DOI: 10.1103/PhysRevD.102.022004. arXiv: 2002.08291 [astro-ph.HE].

[150] F. Aubin et al. "The MBTA pipeline for detecting compact binary coalescences in the third LIGO–Virgo observing run". In: *Class. Quant. Grav.* 38.9 (2021), p. 095004. DOI: 10.1088/1361-6382/abe913. arXiv: 2012.11512 [gr-qc].

[151] S. Sachdev et al. "The GstLAL Search Analysis Methods for Compact Binary Mergers in Advanced LIGO's Second and Advanced Virgo's First Observing Runs". In: (Jan. 2019). arXiv: 1901.08580 [gr-qc].

[152] K. Cannon et al. "GstLAL: A software framework for gravitational wave discovery". In: (Oct. 2020). arXiv: 2010.05082 [astro-ph.IM].

[153] B. Allen, W.-s. Hua, and A. C. Ottewill. "Automatic cross talk removal from multichannel data". In: (Sept. 1999). arXiv: gr-qc/9909083.

[154] E. Poisson and C. M. Will. "Gravitational waves from inspiraling compact binaries: Parameter estimation using second postNewtonian wave forms". In: *Phys. Rev. D* 52 (1995), pp. 848–855. DOI: 10.1103/PhysRevD.52.848. arXiv: gr-qc/9502040.

[155] L. Blanchet, T. Damour, B. R. Iyer, C. M. Will, and A. G. Wiseman. "Gravitational radiation damping of compact binary systems to second postNewtonian order". In: *Phys. Rev. Lett.* 74 (1995), pp. 3515–3518. DOI: 10.1103/PhysRevLett.74.3515. arXiv: gr-qc/9501027.

[156] L. S. Finn and D. F. Chernoff. "Observing binary inspiral in gravitational radiation: One interferometer". In: *Phys. Rev. D* 47 (1993), pp. 2198–2219. DOI: 10.1103/PhysRevD.47.2198. arXiv: gr-qc/9301003.

[157] E. Racine. "Analysis of spin precession in binary black hole systems including quadrupole-monopole interaction". In: *Phys. Rev.* D78 (2008), p. 044021. DOI: 10.1103/PhysRevD.78.044021. arXiv: 0803.1820 [gr-qc].

[158] L. Santamaria et al. "Matching post-Newtonian and numerical relativity waveforms: systematic errors and a new phenomenological model for non-precessing black hole binaries". In: *Phys. Rev.* D82 (2010), p. 064016. DOI: 10.1103/PhysRevD.82.064016. arXiv: 1005.3306 [gr-qc].

[159] P. Ajith et al. "Inspiral-merger-ringdown waveforms for black-hole binaries with non-precessing spins". In: *Phys. Rev. Lett.* 106 (2011), p. 241101. DOI: 10.1103/PhysRevLett.106.241101. arXiv: 0909.2867 [gr-qc].

[160] P. Schmidt, F. Ohme, and M. Hannam. "Towards models of gravitational waveforms from generic binaries II: Modelling precession effects with a single effective precession parameter". In: *Phys. Rev.* D91.2 (2015), p. 024043. DOI: 10.1103/PhysRevD.91.024043. arXiv: 1408.1810 [gr-qc].

[161] R. J. E. Smith, G. Ashton, A. Vajpeyi, and C. Talbot. "Massively parallel Bayesian inference for transient gravitational-wave astronomy". In: *Mon. Not. Roy. Astron. Soc.* 498.3 (2020), pp. 4492–4502. DOI: 10.1093/mnras/staa2483. arXiv: 1909.11873 [gr-qc].

[162] J. S. Speagle. "DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences". In: 493.3 (Apr. 2020), pp. 3132–3158. DOI: 10.1093/mnras/staa278. arXiv: 1904.02180 [astro-ph.IM].

[163] M. Boyle et al. "The SXS Collaboration catalog of binary black hole simulations". In: *Class. Quant. Grav.* 36.19 (2019), p. 195006. DOI: 10.1088/1361-6382/ab34e2. arXiv: 1904.04831 [gr-qc].

[164] S. Vitale, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, I. Mandel, B. Aylott, and J. Veitch. "Effect of calibration errors on Bayesian parameter estimation for gravitational wave signals from inspiral binary systems in the Advanced Detectors era". In: *Phys. Rev.* D85 (2012), p. 064034. DOI: 10.1103/PhysRevD.85.064034. arXiv: 1111.3044 [gr-qc].

[165] E. Payne, C. Talbot, P. D. Lasky, E. Thrane, and J. S. Kissel. "Gravitational-wave astronomy with a physical calibration model". In: *Phys. Rev. D* 102 (2020), p. 122004. DOI: 10.1103/PhysRevD.102.122004. arXiv: 2009.10193 [astro-ph.IM].

[166] S. Vitale, C.-J. Haster, L. Sun, B. Farr, E. Goetz, J. Kissel, and C. Cahillane. "Physical approach to the marginalization of LIGO calibration uncertainties". In: *Phys. Rev. D* 103.6 (2021), p. 063016. DOI: 10.1103/PhysRevD.103.063016. arXiv: 2009.10192 [gr-qc].

[167] R. Essick. "Calibration uncertainty's impact on gravitational-wave observations". In: *Phys. Rev. D* 105.8 (2022), p. 082002. DOI: 10.1103/PhysRevD.105.082002. arXiv: 2202.00823 [astro-ph.IM].

[168] M. Wijngaarden, K. Chatziioannou, A. Bauswein, J. A. Clark, and N. J. Cornish. "Probing neutron stars with the full premerger and postmerger gravitational wave signal from binary coalescences". In: *Phys. Rev. D* 105.10 (2022), p. 104019. DOI: 10.1103/PhysRevD.105.104019. arXiv: 2202.09382 [gr-qc].

[169] Q. Hu and J. Veitch. "Assessing the model waveform accuracy of gravitational waves". In: (May 2022). arXiv: 2205.08448 [gr-qc].

[170] M. Zevin, C. Pankow, C. L. Rodriguez, L. Sampson, E. Chase, V. Kalogera, and F. A. Rasio. "Constraining Formation Models of Binary Black Holes with Gravitational-wave Observations". In: 846.1, 82 (Sept. 2017), p. 82. DOI: 10.3847/1538-4357/aa8408. arXiv: 1704.07379 [astro-ph.HE].

[171] S. Stevenson, F. Ohme, and S. Fairhurst. "Distinguishing compact binary population synthesis models using gravitational-wave observations of coalescing binary black holes". In: *Astrophys. J.* 810.1 (2015), p. 58. DOI: 10.1088/0004-637X/810/1/58. arXiv: 1504.07802 [astro-ph.HE].

[172] S. Stevenson, C. P. L. Berry, and I. Mandel. "Hierarchical analysis of gravitational-wave measurements of binary black hole spin-orbit misalignments". In: 471.3 (Nov. 2017), pp. 2801–2811. DOI: 10.1093/mnras/stx1764. arXiv: 1703.06873 [astro-ph.HE].

[173] M. Fishbach and D. E. Holz. "Where Are LIGO's Big Black Holes?" In: 851.2, L25 (Dec. 2017), p. L25. DOI: 10.3847/2041-8213/aa9bf6. arXiv: 1709.08584 [astro-ph.HE].

[174] S. Vitale, R. Lynch, R. Sturani, and P. Graff. "Use of gravitational waves to probe the formation channels of compact binaries". In: *Class. Quant. Grav.* 34.3 (2017), 03LT01. DOI: 10.1088/1361-6382/aa552e. arXiv: 1503.04307 [gr-qc].

[175] D. Gerosa and E. Berti. "Are merging black holes born from stellar collapse or previous mergers?" In: *Phys. Rev. D* 95.12 (2017), p. 124046. DOI: 10.1103/PhysRevD.95.124046. arXiv: 1703.06223 [gr-qc].

[176] M. Arca Sedda and M. Benacquista. "Using final black hole spins and masses to infer the formation history of the observed population of gravitational wave sources". In: 482.3 (Jan. 2019), pp. 2991–3010. DOI: 10.1093/mnras/sty2764. arXiv: 1806.01285 [astro-ph.GA].

[177] M. Safarzadeh. "The Branching Ratio of LIGO Binary Black Holes". In: 892.1, L8 (Mar. 2020), p. L8. DOI: 10.3847/2041-8213/ab7cdc. arXiv: 2003.02764 [astro-ph.HE].

[178] D. Gerosa and M. Fishbach. "Hierarchical mergers of stellar-mass black holes and their gravitational-wave signatures". In: *Nature Astron.* 5.8 (2021), pp. 749–760. DOI: 10.1038/s41550-021-01398-w. arXiv: 2105.03439 [astro-ph.HE].

[179] S. Olsen, T. Venumadhav, J. Mushkin, J. Roulet, B. Zackay, and M. Zaldarriaga. "New binary black hole mergers in the LIGO-Virgo O3a data". In: *Phys. Rev. D* 106.4 (2022), p. 043009. DOI: 10.1103/PhysRevD.106.043009. arXiv: 2201.02252 [astro-ph.HE].

[180] I. Mandel and A. Farmer. "Merging stellar-mass binary black holes". In: 955 (Apr. 2022), pp. 1–24. DOI: 10.1016/j.physrep.2022.01.003. arXiv: 1806.05820 [astro-ph.HE].

[181] A. Q. Cheng, M. Zevin, and S. Vitale. "What You Don't Know Can Hurt You: Use and Abuse of Astrophysical Models in Gravitational-wave Population Analyses". In: 955.2, 127 (Oct. 2023), p. 127. DOI: 10.3847/1538-4357/aced98. arXiv: 2307.03129 [astro-ph.HE].

[182] M. Mapelli. "Formation Channels of Single and Binary Stellar-Mass Black Holes". In: *Handbook of Gravitational Wave Astronomy*. 2021, 16, p. 16. DOI: 10.1007/978-981-15-4702-7_16-1.

[183] M. Spera, A. A. Trani, and M. Mencagli. "Compact Binary Coalescences: Astrophysical Processes and Lessons Learned". In: *Galaxies* 10.4, 76 (June 2022), p. 76. DOI: 10.3390/galaxies10040076. arXiv: 2206.15392 [astro-ph.HE].

[184] J. Liu, J. E. McClintock, R. Narayan, S. W. Davis, and J. A. Orosz. "Precise Measurement of the Spin Parameter of the Stellar-Mass Black Hole M33 X-7". In: 679.1 (May 2008), p. L37. DOI: 10.1086/588840. arXiv: 0803.1834 [astro-ph].

[185] J. C. A. Miller-Jones et al. "Cygnus X-1 contains a 21-solar mass black hole—Implications for massive star winds". In: *Science* 371.6533 (Mar. 2021), pp. 1046–1049. DOI: 10.1126/science.abb3363. arXiv: 2102.09091 [astro-ph.HE].

[186] C. S. Reynolds. "Observational Constraints on Black Hole Spin". In: 59 (Sept. 2021), pp. 117–154. DOI: 10.1146/annurev-astro-112420-035022. arXiv: 2011.08948 [astro-ph.HE].

[187] M. Fishbach, Z. Doctor, T. Callister, B. Edelman, J. Ye, R. Essick, W. M. Farr, B. Farr, and D. E. Holz. "When Are LIGO/Virgo's Big Black Hole Mergers?" In: *Astrophys. J.* 912.2 (2021), p. 98. DOI: 10.3847/1538-4357/abee11. arXiv: 2101.07699 [astro-ph.HE].

[188] K. Belczynski, Z. Doctor, M. Zevin, A. Olejak, S. Banerje, and D. Chattopadhyay. "Black Hole-Black Hole Total Merger Mass and the Origin of LIGO/Virgo Sources". In: 935.2, 126 (Aug. 2022), p. 126. DOI: 10.3847/1538-4357/ac8167. arXiv: 2204.11730 [astro-ph.HE].

[189] P. Mahapatra, A. Gupta, M. Favata, K. G. Arun, and B. S. Sathyaprakash. "Black hole hierarchical growth efficiency and mass spectrum predictions". In: (Sept. 2022). arXiv: 2209.05766 [astro-ph.HE].

[190] L. A. C. van Son, S. E. de Mink, M. Chruślińska, C. Conroy, R. Pakmor, and L. Hernquist. "The Locations of Features in the Mass Distribution of Merging Binary Black Holes Are Robust against Uncertainties in the Metallicity-dependent Cosmic Star Formation History". In: 948.2, 105 (May 2023), p. 105. DOI: 10.3847/1538-4357/acbf51. arXiv: 2209.03385 [astro-ph.GA].

[191] C. L. Rodriguez and A. Loeb. "Redshift Evolution of the Black Hole Merger Rate from Globular Clusters". In: 866.1, L5 (Oct. 2018), p. L5. DOI: 10.3847/2041-8213/aae377. arXiv: 1809.01152 [astro-ph.HE].

[192] L. A. C. van Son, S. E. de Mink, T. Callister, S. Justham, M. Renzo, T. Wagg, F. S. Broekgaarden, F. Kummer, R. Pakmor, and I. Mandel. "The Redshift Evolution of the Binary Black Hole Merger Rate: A Weighty Matter". In: 931.1, 17 (May 2022), p. 17. DOI: 10.3847/1538-4357/ac64a3. arXiv: 2110.01634 [astro-ph.HE].

[193] M. Fishbach and L. van Son. "LIGO-Virgo-KAGRA's Oldest Black Holes: Probing Star Formation at Cosmic Noon With GWTC-3". In: 957.2, L31 (Nov. 2023), p. L31. DOI: 10.3847/2041-8213/ad0560. arXiv: 2307.15824 [astro-ph.GA].

[194] M. Zevin, I. M. Romero-Shaw, K. Kremer, E. Thrane, and P. D. Lasky. "Implications of Eccentric Observations on Binary Black Hole Formation Channels". In: 921.2, L43 (Nov. 2021), p. L43. DOI: 10.3847/2041-8213/ac32dc. arXiv: 2106.09042 [astro-ph.HE].

[195] T. A. Callister, C.-J. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr. "Who Ordered That? Unequal-mass Binary Black Hole Mergers Have Larger Effective Spins". In: *Astrophys. J. Lett.* 922.1 (2021), p. L5. DOI: 10.3847/2041-8213/ac2ccc. arXiv: 2106.00521 [astro-ph.HE].

[196] V. Tiwari. "Exploring Features in the Binary Black Hole Population". In: *Astrophys. J.* 928.2 (2022), p. 155. DOI: 10.3847/1538-4357/ac589a. arXiv: 2111.13991 [astro-ph.HE].

[197] M. Zevin and S. S. Bavera. "Suspicious Siblings: The Distribution of Mass and Spin across Component Black Holes in Isolated Binary Evolution". In: 933.1, 86 (July 2022), p. 86. DOI: 10.3847/1538-4357/ac6f5d. arXiv: 2203.02515 [astro-ph.HE].

[198] F. S. Broekgaarden, S. Stevenson, and E. Thrane. "Signatures of Mass Ratio Reversal in Gravitational Waves from Merging Binary Black Holes". In: 938.1, 45 (Oct. 2022), p. 45. DOI: 10.3847/1538-4357/ac8879. arXiv: 2205.01693 [astro-ph.HE].

[199] B. McKernan, K. E. S. Ford, T. Callister, W. M. Farr, R. O'Shaughnessy, R. Smith, E. Thrane, and A. Vajpeyi. "LIGO–Virgo correlations between mass ratio and effective inspiral spin: testing the active galactic nuclei channel". In: *Mon. Not. Roy. Astron. Soc.* 514.3 (2022), pp. 3886–3893. DOI: 10.1093/mnras/stac1570. arXiv: 2107.07551 [astro-ph.HE].

[200] C. Adamcewicz and E. Thrane. "Do unequal-mass binary black hole systems have larger χeff? Probing correlations with copulas in gravitational-wave astronomy". In: *Mon. Not. Roy. Astron. Soc.* 517.3 (2022), pp. 3928–3937. DOI: 10.1093/mnras/stac2961. arXiv: 2208.03405 [astro-ph.HE].

[201] V. Baibhav, Z. Doctor, and V. Kalogera. "Dropping Anchor: Understanding the Populations of Binary Black Holes with Random and Aligned-spin Orientations". In: *Astrophys. J.* 946.1 (2023), p. 50. DOI: 10.3847/1538-4357/acbf4c. arXiv: 2212.12113 [astro-ph.HE].

[202] S. Biscoveanu, T. A. Callister, C.-J. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr. "The Binary Black Hole Spin Distribution Likely Broadens with Redshift". In: *Astrophys. J. Lett.* 932.2 (2022), p. L19. DOI: 10.3847/2041-8213/ac71a8. arXiv: 2204.01578 [astro-ph.HE].

[203] A. Ray, I. Magaña Hernandez, S. Mohite, J. Creighton, and S. Kapadia. "Nonparametric Inference of the Population of Compact Binaries from Gravitational-wave Observations Using Binned Gaussian Processes". In: *Astrophys. J.* 957.1 (2023), p. 37. DOI: 10.3847/1538-4357/acf452. arXiv: 2304.08046 [gr-qc].

[204] 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". In: 910.2, 152 (Apr. 2021), p. 152. DOI: 10.3847/1538-4357/abe40e. arXiv: 2011.10057 [astro-ph.HE].

[205] M. E. Lower, E. Thrane, P. D. Lasky, and R. Smith. "Measuring eccentricity in binary black hole inspirals with gravitational waves". In: 98.8, 083028 (Oct. 2018), p. 083028. DOI: 10.1103/PhysRevD.98.083028. arXiv: 1806.05350 [astro-ph.HE].

[206] I. M. Romero-Shaw, P. D. Lasky, and E. Thrane. "Searching for eccentricity: signatures of dynamical formation in the first gravitational-wave transient catalogue of LIGO and Virgo". In: 490.4 (Dec. 2019), pp. 5210–5216. DOI: 10.1093/mnras/stz2996. arXiv: 1909.05466 [astro-ph.HE].

[207] I. Romero-Shaw, P. D. Lasky, and E. Thrane. "Four Eccentric Mergers Increase the Evidence that LIGO-Virgo-KAGRA's Binary Black Holes Form Dynamically". In: 940.2, 171 (Dec. 2022), p. 171. DOI: 10.3847/1538-4357/ac9798. arXiv: 2206.14695 [astro-ph.HE].

[208] C. Kimball, C. P. L. Berry, and V. Kalogera. "What GW170729's exceptional mass and spin tells us about its family tree". In: *Res. Notes AAS* 4.1 (2020), p. 2. DOI: 10.3847/2515-5172/ab66be. arXiv: 1903.07813 [astro-ph.HE].

[209] C. Kimball, C. Talbot, C. P. L. Berry, M. Carney, M. Zevin, E. Thrane, and V. Kalogera. "Black Hole Genealogy: Identifying Hierarchical Mergers with Gravitational Waves". In: 900.2, 177 (Sept. 2020), p. 177. DOI: 10.3847/1538-4357/aba518. arXiv: 2005.00023 [astro-ph.HE].

[210] C. Kimball et al. "Evidence for Hierarchical Black Hole Mergers in the Second LIGO-Virgo Gravitational Wave Catalog". In: 915.2, L35 (July 2021), p. L35. DOI: 10.3847/2041-8213/ac0aef. arXiv: 2011.05332 [astro-ph.HE].

[211] P. Mahapatra, A. Gupta, M. Favata, K. G. Arun, and B. S. Sathyaprakash. "Remnant Black Hole Kicks and Implications for Hierarchical Mergers". In: *Astrophys. J. Lett.* 918.2 (2021), p. L31. DOI: 10.3847/2041-8213/ac20db. arXiv: 2106.07179 [astro-ph.HE].

[212] R. Farmer, M. Renzo, S. E. de Mink, P. Marchant, and S. Justham. "Mind the Gap: The Location of the Lower Edge of the Pair-instability Supernova Black

Hole Mass Gap". In: 887.1, 53 (Dec. 2019), p. 53. DOI: 10.3847/1538-4357/ab518b. arXiv: 1910.12874 [astro-ph.SR].

[213] B. Edelman, Z. Doctor, and B. Farr. "Poking Holes: Looking for Gaps in LIGO/Virgo's Black Hole Population". In: *Astrophys. J. Lett.* 913.2 (2021), p. L23. DOI: 10.3847/2041-8213/abfdb3. arXiv: 2104.07783 [astro-ph.HE].

[214] R. Abbott et al. "GW190521: A Binary Black Hole Merger with a Total Mass of 150 $M_\odot$". In: 125.10, 101102 (Sept. 2020), p. 101102. DOI: 10.1103/PhysRevLett.125.101102. arXiv: 2009.01075 [gr-qc].

[215] M. Fishbach and D. E. Holz. "Minding the Gap: GW190521 as a Straddling Binary". In: 904.2, L26 (Dec. 2020), p. L26. DOI: 10.3847/2041-8213/abc827. arXiv: 2009.05472 [astro-ph.HE].

[216] A. H. Nitz and C. D. Capano. "GW190521 May Be an Intermediate-mass Ratio Inspiral". In: 907.1, L9 (Jan. 2021), p. L9. DOI: 10.3847/2041-8213/abccc5. arXiv: 2010.12558 [astro-ph.HE].

[217] M. Mould, D. Gerosa, M. Dall'Amico, and M. Mapelli. "One to many: comparing single gravitational-wave events to astrophysical populations". In: *Mon. Not. Roy. Astron. Soc.* 525.3 (2023), pp. 3986–3997. DOI: 10.1093/mnras/stad2502. arXiv: 2305.18539 [astro-ph.HE].

[218] T. Damour. "Coalescence of two spinning black holes: An effective one-body approach". In: 64.12 (Dec. 2001), p. 124013. DOI: 10.1103/PhysRevD.64.124013. arXiv: gr-qc/0103018 [gr-qc].

[219] É. Racine. "Analysis of spin precession in binary black hole systems including quadrupole-monopole interaction". In: 78.4, 044021 (Aug. 2008), p. 044021. DOI: 10.1103/PhysRevD.78.044021. arXiv: 0803.1820 [gr-qc].

[220] P. Schmidt, F. Ohme, and M. Hannam. "Towards models of gravitational waveforms from generic binaries: II. Modelling precession effects with a single effective precession parameter". In: 91.2, 024043 (Jan. 2015), p. 024043. DOI: 10.1103/PhysRevD.91.024043. arXiv: 1408.1810 [gr-qc].

[221] C. L. Rodriguez, M. Zevin, C. Pankow, V. Kalogera, and F. A. Rasio. "Illuminating Black Hole Binary Formation Channels with Spins in Advanced LIGO". In: 832.1, L2 (Nov. 2016), p. L2. DOI: 10.3847/2041-8205/832/1/L2. arXiv: 1609.05916 [astro-ph.HE].

[222] K. K. Y. Ng, S. Vitale, A. Zimmerman, K. Chatziioannou, D. Gerosa, and C.-J. Haster. "Gravitational-wave astrophysics with effective-spin measurements: asymmetries and selection biases". In: *Phys. Rev. D* 98.8 (2018), p. 083007. DOI: 10.1103/PhysRevD.98.083007. arXiv: 1805.03046 [gr-qc].

[223] V. Baibhav, D. Gerosa, E. Berti, K. W. K. Wong, T. Helfer, and M. Mould. "The mass gap, the spin gap, and the origin of merging binary black holes". In: *Phys. Rev. D* 102.4 (2020), p. 043002. DOI: 10.1103/PhysRevD.102.043002. arXiv: 2004.00650 [astro-ph.HE].

[224] M. Fishbach, C. Kimball, and V. Kalogera. "Limits on Hierarchical Black Hole Mergers from the Most Negative χ_{eff} Systems". In: *Astrophys. J. Lett.* 935.2 (2022), p. L26. DOI: 10.3847/2041-8213/ac86c4. arXiv: 2207.02924 [astro-ph.HE].

[225] R. C. Zhang, G. Fragione, C. Kimball, and V. Kalogera. "On the Likely Dynamical Origin of GW191109 and Binary Black Hole Mergers with Negative Effective Spin". In: 954.1, 23 (Sept. 2023), p. 23. DOI: 10.3847/1538-4357/ace4c1. arXiv: 2302.07284 [astro-ph.HE].

[226] K. Breivik et al. "COSMIC Variance in Binary Population Synthesis". In: 898.1, 71 (July 2020), p. 71. DOI: 10.3847/1538-4357/ab9d85. arXiv: 1911.00903 [astro-ph.HE].

[227] C. L. Rodriguez, P. Amaro-Seoane, S. Chatterjee, K. Kremer, F. A. Rasio, J. Samsing, C. S. Ye, and M. Zevin. "Post-Newtonian dynamics in dense star clusters: Formation, masses, and merger rates of highly-eccentric black hole binaries". In: 98.12, 123005 (Dec. 2018), p. 123005. DOI: 10.1103/PhysRevD.98.123005. arXiv: 1811.04926 [astro-ph.HE].

[228] W. E. Harris. "A Catalog of Parameters for Globular Clusters in the Milky Way". In: 112 (Oct. 1996), p. 1487. DOI: 10.1086/118116.

[229] J. P. Brodie and J. Strader. "Extragalactic Globular Clusters and Galaxy Formation". In: 44.1 (Sept. 2006), pp. 193–267. DOI: 10.1146/annurev.astro.44.051905.092441. arXiv: astro-ph/0602601 [astro-ph].

[230] C. J. Lada and E. A. Lada. "Embedded Clusters in Molecular Clouds". In: 41 (Jan. 2003), pp. 57–115. DOI: 10.1146/annurev.astro.41.011802.094844. arXiv: astro-ph/0301540 [astro-ph].

[231] K. El-Badry, E. Quataert, D. R. Weisz, N. Choksi, and M. Boylan-Kolchin. "The formation and hierarchical assembly of globular cluster populations". In: 482.4 (Feb. 2019), pp. 4528–4552. DOI: 10.1093/mnras/sty3007. arXiv: 1805.03652 [astro-ph.GA].

[232] Y. Qin, T. Fragos, G. Meynet, J. Andrews, M. Sørensen, and H. F. Song. "The spin of the second-born black hole in coalescing binary black holes". In: *Astron. Astrophys.* 616 (2018), A28. DOI: 10.1051/0004-6361/201832839. arXiv: 1802.05738 [astro-ph.SR].

[233] J. Fuller and L. Ma. "Most Black Holes Are Born Very Slowly Rotating". In: 881.1, L1 (Aug. 2019), p. L1. DOI: 10.3847/2041-8213/ab339b. arXiv: 1907.03714 [astro-ph.SR].

[234] M. Zevin and D. E. Holz. "Avoiding a Cluster Catastrophe: Retention Efficiency and the Binary Black Hole Mass Spectrum". In: *Astrophys. J. Lett.* 935 (2022), p. L20. DOI: 10.3847/2041-8213/ac853d. arXiv: 2205.08549 [astro-ph.HE].

[235] V. Baibhav, E. Berti, D. Gerosa, M. Mould, and K. W. K. Wong. "Looking for the parents of LIGO's black holes". In: *Phys. Rev. D* 104.8 (2021), p. 084002. DOI: 10.1103/PhysRevD.104.084002. arXiv: 2105.12140 [gr-qc].

[236] H. Jeffreys. *Theory of Probability*. 3rd ed. Oxford, England: Oxford, 1961.

[237] S. Fairhurst, R. Green, M. Hannam, and C. Hoy. "When will we observe binary black holes precessing?" In: *Phys. Rev. D* 102.4 (2020), p. 041302. DOI: 10.1103/PhysRevD.102.041302. arXiv: 1908.00555 [gr-qc].

[238] D. Gerosa, M. Mould, D. Gangardt, P. Schmidt, G. Pratten, and L. M. Thomas. "A generalized precession parameter $\chi_{\rm p}$ to interpret gravitational-wave data". In: *Phys. Rev. D* 103.6 (2021), p. 064067. DOI: 10.1103/PhysRevD.103.064067. arXiv: 2011.11948 [gr-qc].

[239] L. M. Thomas, P. Schmidt, and G. Pratten. "New effective precession spin for modeling multimodal gravitational waveforms in the strong-field regime". In: *Phys. Rev. D* 103.8 (2021), p. 083022. DOI: 10.1103/PhysRevD.103.083022. arXiv: 2012.02209 [gr-qc].

[240] L. Ma and J. Fuller. "Tidal Spin-up of Black Hole Progenitor Stars". In: *Astrophys. J.* 952.1 (2023), p. 53. DOI: 10.3847/1538-4357/acdb74. arXiv: 2305.08356 [astro-ph.HE].

[241] G. Pratten et al. "Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes". In: *Phys. Rev. D* 103.10 (2021), p. 104056. DOI: 10.1103/PhysRevD.103.104056. arXiv: 2004.06503 [gr-qc].

[242] R. Abbott et al. "Properties and Astrophysical Implications of the 150 $M_\odot$ Binary Black Hole Merger GW190521". In: *Astrophys. J. Lett.* 900.1 (2020), p. L13. DOI: 10.3847/2041-8213/aba493. arXiv: 2009.01190 [astro-ph.HE].

[243] R. Abbott et al. "GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run". In: *Phys. Rev. D* 109.2 (2024), p. 022001. DOI: 10.1103/PhysRevD.109.022001. arXiv: 2108.01045 [gr-qc].

[244] S. J. Miller, Z. Ko, T. A. Callister, and K. Chatziioannou. "Gravitational waves carry information beyond effective spin parameters but it is hard to extract". In: (Jan. 2024). arXiv: 2401.05613 [gr-qc].

[245] M. Hannam et al. “General-relativistic precession in a black-hole binary”. In: *Nature* 610.7933 (2022), pp. 652–655. DOI: 10.1038/s41586-022-05212-z. arXiv: 2112.11300 [gr-qc].

[246] R. Udall and D. Davis. “Bayesian modeling of scattered light in the LIGO interferometers”. In: *Appl. Phys. Lett.* 122.9 (2023), p. 094103. DOI: 10.1063/5.0136896. arXiv: 2211.15867 [astro-ph.IM].

[247] R. Macas, A. Lundgren, and G. Ashton. “Revisiting GW200129 with machine learning noise mitigation: it is (still) precessing”. In: (Nov. 2023). arXiv: 2311.09921 [gr-qc].

[248] R. W. Kiendrebeogo et al. “Updated Observing Scenarios and Multimessenger Implications for the International Gravitational-wave Networks O4 and O5”. In: *Astrophys. J.* 958.2 (2023), p. 158. DOI: 10.3847/1538-4357/acfcb1. arXiv: 2306.09234 [astro-ph.HE].

[249] B. P. Abbott et al. “Tests of General Relativity with GW150914”. In: *Phys. Rev. Lett.* 116.22, 221101 (June 2016), p. 221101. DOI: 10.1103/PhysRevLett.116.221101. arXiv: 1602.03841 [gr-qc].

[250] B. P. Abbott et al. “GW170104: Observation of a 50-Solar-Mass Binary Black Hole Coalescence at Redshift 0.2”. In: 118.22, 221101 (June 2017), p. 221101. DOI: 10.1103/PhysRevLett.118.221101. arXiv: 1706.01812 [gr-qc].

[251] B. P. Abbott et al. “GW170814: A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence”. In: 119.14, 141101 (Oct. 2017), p. 141101. DOI: 10.1103/PhysRevLett.119.141101. arXiv: 1709.09660 [gr-qc].

[252] A. Ghosh et al. “Testing general relativity using golden black-hole binaries”. In: 94.2, 021101 (July 2016), p. 021101. DOI: 10.1103/PhysRevD.94.021101. arXiv: 1602.02453 [gr-qc].

[253] A. Ghosh, N. K. Johnson-McDaniel, A. Ghosh, C. Kant Mishra, P. Ajith, W. Del Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London. “Testing general relativity using gravitational wave signals from the inspiral, merger and ringdown of binary black holes”. In: *Classical and Quantum Gravity* 35.1, 014002 (Jan. 2018), p. 014002. DOI: 10.1088/1361-6382/aa972e. arXiv: 1704.06784 [gr-qc].

[254] A. K. Mehta, A. Buonanno, R. Cotesta, A. Ghosh, N. Sennett, and J. Steinhoff. “Tests of general relativity with gravitational-wave observations using a flexible theory-independent method”. In: *Phys. Rev. D* 107.4 (2023), p. 044020. DOI: 10.1103/PhysRevD.107.044020. arXiv: 2203.13937 [gr-qc].

[255] C. M. Will. "Bounding the mass of the graviton using gravitational wave observations of inspiralling compact binaries". In: *Phys. Rev. D* 57 (1998), pp. 2061–2068. DOI: 10.1103/PhysRevD.57.2061. arXiv: gr-qc/9709011.

[256] S. Mirshekari, N. Yunes, and C. M. Will. "Constraining Generic Lorentz Violation and the Speed of the Graviton with Gravitational Waves". In: *Phys. Rev. D* 85 (2012), p. 024041. DOI: 10.1103/PhysRevD.85.024041. arXiv: 1110.2720 [gr-qc].

[257] T. Zhu, W. Zhao, J.-M. Yan, C. Gong, and A. Wang. "Tests of modified gravitational wave propagations with gravitational waves". Apr. 2023. arXiv: 2304.09025 [gr-qc].

[258] D. M. Eardley, D. L. Lee, A. P. Lightman, R. V. Wagoner, and C. M. Will. "Gravitational-Wave Observations as a Tool for Testing Relativistic Gravity". In: *Phys. Rev. Lett.* 30 (18 Apr. 1973), pp. 884–886. DOI: 10.1103/PhysRevLett.30.884. URL: https://link.aps.org/doi/10.1103/PhysRevLett.30.884.

[259] M. Isi and A. J. Weinstein. "Probing gravitational wave polarizations with signals from compact binary coalescences". In: *arXiv e-prints*, arXiv:1710.03794 (2017), arXiv:1710.03794. DOI: 10.48550/arXiv.1710.03794. arXiv: 1710.03794 [gr-qc].

[260] P. T. H. Pang, R. K. L. Lo, I. C. F. Wong, T. G. F. Li, and C. Van Den Broeck. "Generic searches for alternative gravitational wave polarizations with networks of interferometric detectors". In: *Phys. Rev. D* 101.10 (2020), p. 104055. DOI: 10.1103/PhysRevD.101.104055. arXiv: 2003.07375 [gr-qc].

[261] K. Chatziioannou, M. Isi, C.-J. Haster, and T. B. Littenberg. "Morphology-independent test of the mixed polarization content of transient gravitational wave signals". In: *Phys. Rev. D* 104.4 (2021), p. 044005. DOI: 10.1103/PhysRevD.104.044005. arXiv: 2105.01521 [gr-qc].

[262] M. Isi, K. Chatziioannou, and W. M. Farr. "Hierarchical Test of General Relativity with Gravitational Waves". In: 123.12, 121101 (Sept. 2019), p. 121101. DOI: 10.1103/PhysRevLett.123.121101. arXiv: 1904.08011 [gr-qc].

[263] M. Saleem, N. V. Krishnendu, A. Ghosh, A. Gupta, W. Del Pozzo, A. Ghosh, and K. G. Arun. "Population inference of spin-induced quadrupole moments as a probe for nonblack hole compact binaries". In: 105.10, 104066 (May 2022), p. 104066. DOI: 10.1103/PhysRevD.105.104066. arXiv: 2111.04135 [gr-qc].

[264] A. Zimmerman, C.-J. Haster, and K. Chatziioannou. "On combining information from multiple gravitational wave sources". In: 99.12, 124044 (June 2019), p. 124044. DOI: 10.1103/PhysRevD.99.124044. arXiv: 1903.11008 [astro-ph.IM].

[265] M. Isi, W. M. Farr, and K. Chatziioannou. "Comparing Bayes factors and hierarchical inference for testing general relativity with gravitational waves". In: *Phys. Rev. D* 106.2 (2022), p. 024048. DOI: 10.1103/PhysRevD.106.024048. arXiv: 2204.10742 [gr-qc].

[266] D. Psaltis, C. Talbot, E. Payne, and I. Mandel. "Probing the black hole metric: Black hole shadows and binary black-hole inspirals". In: 103.10, 104036 (May 2021), p. 104036. DOI: 10.1103/PhysRevD.103.104036. arXiv: 2012.02117 [gr-qc].

[267] N. E. Wolfe, C. Talbot, and J. Golomb. "Accelerating Tests of General Relativity with Gravitational-Wave Signals using Hybrid Sampling". In: *arXiv e-prints*, arXiv:2208.12872 (Aug. 2022), arXiv:2208.12872. DOI: 10.48550/arXiv.2208.12872. arXiv: 2208.12872 [gr-qc].

[268] N. Loutrel, T. Tanaka, and N. Yunes. "Spin-Precessing Black Hole Binaries in Dynamical Chern-Simons Gravity". In: *Phys. Rev. D* 98.6 (2018), p. 064020. DOI: 10.1103/PhysRevD.98.064020. arXiv: 1806.07431 [gr-qc].

[269] S. E. Perkins, R. Nair, H. O. Silva, and N. Yunes. "Improved gravitational-wave constraints on higher-order curvature theories of gravity". In: *Phys. Rev. D* 104.2 (2021), p. 024060. DOI: 10.1103/PhysRevD.104.024060. arXiv: 2104.11189 [gr-qc].

[270] N. Loutrel and N. Yunes. "Parity violation in spin-precessing binaries: Gravitational waves from the inspiral of black holes in dynamical Chern-Simons gravity". In: *Phys. Rev. D* 106.6 (2022), p. 064009. DOI: 10.1103/PhysRevD.106.064009. arXiv: 2205.02675 [gr-qc].

[271] A. Bohé et al. "Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors". In: 95.4, 044028 (Feb. 2017), p. 044028. DOI: 10.1103/PhysRevD.95.044028. arXiv: 1611.03703 [gr-qc].

[272] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine. "Enriching the symphony of gravitational waves from binary black holes by tuning higher harmonics". In: 98.8, 084028 (Oct. 2018), p. 084028. DOI: 10.1103/PhysRevD.98.084028. arXiv: 1803.10701 [gr-qc].

[273] R. Cotesta, S. Marsat, and M. Pürrer. "Frequency-domain reduced-order model of aligned-spin effective-one-body waveforms with higher-order modes". In: 101.12, 124040 (June 2020), p. 124040. DOI: 10.1103/PhysRevD.101.124040. arXiv: 2003.12079 [gr-qc].

[274] R. Brito, A. Buonanno, and V. Raymond. "Black-hole spectroscopy by making full use of gravitational-wave modeling". In: 98.8, 084038 (Oct. 2018), p. 084038. DOI: 10.1103/PhysRevD.98.084038. arXiv: 1805.00293 [gr-qc].

[275] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. *Data release for Tests of General Relativity with GWTC-3*. 2022. DOI: 10.5281/zenodo.7007370. URL: %7B%5Curl%7Bhttps://zenodo.org/record/7007370%7D%7D.

[276] I. Mandel, W. M. Farr, and J. R. Gair. "Extracting distribution parameters from multiple uncertain observations with selection biases". In: *Mon. Not. Roy. Astron. Soc.* 486.1 (2019), pp. 1086–1093. DOI: 10.1093/mnras/stz896. arXiv: 1809.02063 [physics.data-an].

[277] S. Vitale, D. Gerosa, W. M. Farr, and S. R. Taylor. "Inferring the properties of a population of compact binaries in presence of selection effects". In: (July 2020). DOI: 10.1007/978-981-15-4702-7_45-1. arXiv: 2007.05579 [astro-ph.IM].

[278] J. Roulet, H. S. Chia, S. Olsen, L. Dai, T. Venumadhav, B. Zackay, and M. Zaldarriaga. "Distribution of effective spins and masses of binary black holes from the LIGO and Virgo O1–O3a observing runs". In: *Phys. Rev. D* 104.8 (2021), p. 083010. DOI: 10.1103/PhysRevD.104.083010. arXiv: 2105.10580 [astro-ph.HE].

[279] W. M. Farr, S. Stevenson, M. Coleman Miller, I. Mandel, B. Farr, and A. Vecchio. "Distinguishing Spin-Aligned and Isotropic Black Hole Populations With Gravitational Waves". In: *Nature* 548 (2017), p. 426. DOI: 10.1038/nature23453. arXiv: 1706.01385 [astro-ph.HE].

[280] C. Talbot and E. Thrane. "Measuring the binary black hole mass spectrum with an astrophysically motivated parameterization". In: *Astrophys. J.* 856.2 (2018), p. 173. DOI: 10.3847/1538-4357/aab34c. arXiv: 1801.02699 [astro-ph.HE].

[281] C. Talbot and E. Thrane. "Determining the population properties of spinning black holes". In: *Phys. Rev. D* 96.2 (2017), p. 023012. DOI: 10.1103/PhysRevD.96.023012. arXiv: 1704.08370 [astro-ph.HE].

[282] S. Miller, T. A. Callister, and W. Farr. "The Low Effective Spin of Binary Black Holes and Implications for Individual Gravitational-Wave Events". In: *Astrophys. J.* 895.2 (2020), p. 128. DOI: 10.3847/1538-4357/ab80c0. arXiv: 2001.06051 [astro-ph.HE].

[283] S. Galaudage et al. "Building Better Spin Models for Merging Binary Black Holes: Evidence for Nonspinning and Rapidly Spinning Nearly Aligned Subpopulations". In: *Astrophys. J. Lett.* 921.1 (2021). [Erratum: Astrophys.J.Lett. 936, L18 (2022), Erratum: Astrophys.J. 936, L18 (2022)], p. L15. DOI: 10.3847/2041-8213/ac2f3c. arXiv: 2109.02424 [gr-qc].

[284] M. Fishbach, D. E. Holz, and W. M. Farr. "Does the Black Hole Merger Rate Evolve with Redshift?" In: *Astrophys. J. Lett.* 863.2 (2018), p. L41. DOI: 10.3847/2041-8213/aad800. arXiv: 1805.10270 [astro-ph.HE].

[285] B. Edelman, Z. Doctor, J. Godfrey, and B. Farr. "Ain't No Mountain High Enough: Semiparametric Modeling of LIGO–Virgo's Binary Black Hole Mass Distribution". In: *Astrophys. J.* 924.2 (2022), p. 101. DOI: 10.3847/1538-4357/ac3667. arXiv: 2109.06137 [astro-ph.HE].

[286] B. Edelman, B. Farr, and Z. Doctor. "Cover Your Basis: Comprehensive Data-driven Characterization of the Binary Black Hole Population". In: 946.1, 16 (Mar. 2023), p. 16. DOI: 10.3847/1538-4357/acb5ed. arXiv: 2210.12834 [astro-ph.HE].

[287] J. Golomb and C. Talbot. "Searching for structure in the binary black hole spin distribution". In: *arXiv e-prints*, arXiv:2210.12287 (Oct. 2022), arXiv:2210.12287. DOI: 10.48550/arXiv.2210.12287. arXiv: 2210.12287 [astro-ph.HE].

[288] T. A. Callister and W. M. Farr. "A Parameter-Free Tour of the Binary Black Hole Population". In: *arXiv e-prints*, arXiv:2302.07289 (Feb. 2023), arXiv:2302.07289. DOI: 10.48550/arXiv.2302.07289. arXiv: 2302.07289 [astro-ph.HE].

[289] W. M. Farr. "Accuracy Requirements for Empirically Measured Selection Functions". In: *Research Notes of the American Astronomical Society* 3.5, 66 (May 2019), p. 66. DOI: 10.3847/2515-5172/ab1d5f. arXiv: 1904.10879 [astro-ph.IM].

[290] C. J. Moore and D. Gerosa. "Population-informed priors in gravitational-wave astronomy". In: *Phys. Rev. D* 104 (2021), p. 083008. DOI: 10.1103/PhysRevD.104.083008. arXiv: 2108.02462 [gr-qc].

[291] W. M. Farr and T. A. Callister. *Re-Weighting Existing Samples to a Population Analysis*. Tech. rep. 2021. URL: https://github.com/farr/Reweighting/blob/master-pdf/note/reweighting.pdf.

[292] T. A. Callister. *Reweighting Single Event Posteriors with Hyperparameter Marginalization*. Tech. rep. LIGO-T2100301. 2021. URL: https://dcc.ligo.org/LIGO-T2100301/public.

[293] W. Del Pozzo, J. Veitch, and A. Vecchio. "Testing general relativity using Bayesian model selection: Applications to observations of gravitational waves from compact binary systems". In: 83.8, 082002 (Apr. 2011), p. 082002. DOI: 10.1103/PhysRevD.83.082002. arXiv: 1101.1391 [gr-qc].

[294] J. Meidam, M. Agathos, C. Van Den Broeck, J. Veitch, and B. S. Sathyaprakash. "Testing the no-hair theorem with black hole ringdowns using TIGER". In: 90.6, 064009 (Sept. 2014), p. 064009. DOI: 10.1103/PhysRevD.90.064009. arXiv: 1406.3201 [gr-qc].

[295] J. Meidam et al. "Parametrized tests of the strong-field dynamics of general relativity using gravitational wave signals from coalescing binary black holes: Fast likelihood calculations and sensitivity of the method". In: 97.4, 044033 (Feb. 2018), p. 044033. DOI: 10.1103/PhysRevD.97.044033. arXiv: 1712.08772 [gr-qc].

[296] D. Wysocki, J. Lange, and R. O'Shaughnessy. "Reconstructing phenomenological distributions of compact binaries via gravitational wave observations". In: *Phys. Rev. D* 100.4 (2019), p. 043012. DOI: 10.1103/PhysRevD.100.043012. arXiv: 1805.06442 [gr-qc].

[297] T. A. Callister, S. J. Miller, K. Chatziioannou, and W. M. Farr. "No Evidence that the Majority of Black Holes in Binaries Have Zero Spin". In: *Astrophys. J. Lett.* 937.1 (2022), p. L13. DOI: 10.3847/2041-8213/ac847e. arXiv: 2205.08574 [astro-ph.HE].

[298] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. *Tests of General Relativity with Binary Black Holes from the second LIGO–Virgo Gravitational-Wave Transient Catalog - Full Posterior Sample Data Release*. 2021. DOI: 10.7935/903s-gx73. URL: %7B%5Curl%7Bhttps://zenodo.org/record/5172704#.YTOaSC1h2Zw%7D%7D.

[299] R. Abbott et al. "GW190814: Gravitational Waves from the Coalescence of a 23 Solar Mass Black Hole with a 2.6 Solar Mass Compact Object". In: *Astrophys. J. Lett.* 896.2 (2020), p. L44. DOI: 10.3847/2041-8213/ab960f. arXiv: 2006.12611 [astro-ph.HE].

[300] R. Abbott et al. "Observation of Gravitational Waves from Two Neutron Star–Black Hole Coalescences". In: *Astrophys. J. Lett.* 915.1 (2021), p. L5. DOI: 10.3847/2041-8213/ac082e. arXiv: 2106.15163 [astro-ph.HE].

[301] D. Phan, N. Pradhan, and M. Jankowiak. "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro". In: *arXiv e-prints*, arXiv:1912.11554 (Dec. 2019), arXiv:1912.11554. DOI: 10.48550/arXiv.1912.11554. arXiv: 1912.11554 [stat.ML].

[302] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman. "Pyro: Deep Universal Probabilistic Programming". In: *J. Mach. Learn. Res.* 20 (2019), 28:1–28:6. URL: http://jmlr.org/papers/v20/18-403.html.

[303] J. Bradbury et al. *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. 2018. URL: http://github.com/google/jax.

[304] T. P. Robitaille et al. "Astropy: A community Python package for astronomy". In: *Astron. & Astrophys.* 558, A33 (2013), A33. DOI: 10.1051/0004-6361/201322068. arXiv: 1307.6212 [astro-ph.IM].

[305] A. M. Price-Whelan et al. “The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package”. In: *Astron. J.* 156.3, 123 (Sept. 2018), p. 123. DOI: 10.3847/1538-3881/aabc4f. arXiv: 1801.02634 [astro-ph.IM].

[306] A. M. Price-Whelan et al. “The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package”. In: *apj* 935.2, 167 (Aug. 2022), p. 167. DOI: 10.3847/1538-4357/ac7c74. arXiv: 2206.14220 [astro-ph.IM].

[307] P. Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[308] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[309] R. Kumar, C. Carroll, A. Hartikainen, and O. Martin. “ArviZ a unified library for exploratory analysis of Bayesian models in Python”. In: *Journal of Open Source Software* 4.33 (2019), p. 1143. DOI: 10.21105/joss.01143. URL: https://doi.org/10.21105/joss.01143.

[310] D. Foreman-Mackey. “corner.py: Scatterplot matrices in Python”. In: *The Journal of Open Source Software* 1.2 (June 2016), p. 24. DOI: 10.21105/joss.00024. URL: https://doi.org/10.21105/joss.00024.

[311] E. Payne, M. Isi, K. Chatziioannou, and W. M. Farr. *Code Release for “Fortifying gravitational-wave tests of general relativity against astrophysical assumption”*. 2023. URL: https://github.com/ethanpayne42/testingGR_astro.git.

[312] C. J. Moore, E. Finch, R. Buscicchio, and D. Gerosa. “Testing general relativity with gravitational-wave catalogs: The insidious nature of waveform systematics”. In: *iScience* 24.6 (2021), p. 102577. ISSN: 2589-0042. DOI: https://doi.org/10.1016/j.isci.2021.102577. URL: https://www.sciencedirect.com/science/article/pii/S2589004221005459.

[313] Q. Hu and J. Veitch. “Accumulating Errors in Tests of General Relativity with Gravitational Waves: Overlapping Signals and Inaccurate Waveforms”. In: *Astrophys. J.* 945.2 (2023), p. 103. DOI: 10.3847/1538-4357/acbc18. arXiv: 2210.04769 [gr-qc].

[314] P. Saini, M. Favata, and K. G. Arun. “Systematic bias on parametrized tests of general relativity due to neglect of orbital eccentricity”. In: *Phys. Rev. D* 106.8 (2022), p. 084031. DOI: 10.1103/PhysRevD.106.084031. arXiv: 2203.04634 [gr-qc].

[315] S. A. Bhat, P. Saini, M. Favata, and K. G. Arun. "Systematic bias on the inspiral-merger-ringdown consistency test due to neglect of orbital eccentricity". In: *Phys. Rev. D* 107.2 (2023), p. 024009. DOI: 10.1103/PhysRevD.107.024009. arXiv: 2207.13761 [gr-qc].

[316] R. Abbott et al. "Constraints on the Cosmic Expansion History from GWTC–3". In: *Astrophys. J.* 949.2 (2023), p. 76. DOI: 10.3847/1538-4357/ac74bb. arXiv: 2111.03604 [astro-ph.CO].

[317] D. Wysocki, R. O'Shaughnessy, L. Wade, and J. Lange. "Inferring the neutron star equation of state simultaneously with the population of merging neutron stars". In: (Jan. 2020). arXiv: 2001.01747 [gr-qc].

[318] I. M. Romero-Shaw, E. Thrane, and P. D. Lasky. "When models fail: An introduction to posterior predictive checks and model misspecification in gravitational-wave astronomy". In: *Publ. Astron. Soc. Austral.* 39 (2022), e025. DOI: 10.1017/pasa.2022.24. arXiv: 2202.05479 [astro-ph.IM].

[319] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN: 9781439840955.

[320] E. Payne and E. Thrane. "Model exploration in gravitational-wave astronomy with the maximum population likelihood". In: *Phys. Rev. Res.* 5 (2 Apr. 2023), p. 023013. DOI: 10.1103/PhysRevResearch.5.023013. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.5.023013.

[321] A. Gupta, S. Datta, S. Kastha, S. Borhanian, K. G. Arun, and B. S. Sathyaprakash. "Multiparameter tests of general relativity using multiband gravitational-wave observations". In: *Phys. Rev. Lett.* 125.20 (2020), p. 201101. DOI: 10.1103/PhysRevLett.125.201101. arXiv: 2005.09607 [gr-qc].

[322] S. Datta, A. Gupta, S. Kastha, K. G. Arun, and B. S. Sathyaprakash. "Tests of general relativity using multiband observations of intermediate mass binary black hole mergers". In: *Phys. Rev. D* 103.2 (2021), p. 024036. DOI: 10.1103/PhysRevD.103.024036. arXiv: 2006.12137 [gr-qc].

[323] A. A. Shoom, P. K. Gupta, B. Krishnan, A. B. Nielsen, and C. D. Capano. "Testing the post-Newtonian expansion with GW170817". In: *General Relativity and Gravitation* 55.4, 55 (2023), p. 55. DOI: 10.1007/s10714-023-03100-z. arXiv: 2105.02191 [gr-qc].

[324] S. Perkins and N. Yunes. "Are parametrized tests of general relativity with gravitational waves robust to unknown higher post-Newtonian order effects?" In: *Phys. Rev. D* 105.12 (2022), p. 124047. DOI: 10.1103/PhysRevD.105.124047. arXiv: 2201.02542 [gr-qc].

[325] S. Datta, M. Saleem, K. G. Arun, and B. S. Sathyaprakash. "Multiparameter tests of general relativity using principal component analysis with next-generation gravitational wave detectors". In: *arXiv e-prints*, arXiv:2208.07757 (2022), arXiv:2208.07757. DOI: 10.48550/arXiv.2208.07757. arXiv: 2208.07757 [gr-qc].

[326] P. A. R. Ade et al. "Planck 2015 results. XIII. Cosmological parameters". In: *Astron. Astrophys.* 594 (2016), A13. DOI: 10.1051/0004-6361/201525830. arXiv: 1502.01589 [astro-ph.CO].

[327] B. S. Sathyaprakash and S. V. Dhurandhar. "Choice of filters for the detection of gravitational waves from coalescing binaries". In: 44.12 (Dec. 1991), pp. 3819–3834. DOI: 10.1103/PhysRevD.44.3819.

[328] L. Blanchet and G. Schafer. "Gravitational wave tails and binary star systems". In: *Classical and Quantum Gravity* 10.12 (Dec. 1993), pp. 2699–2721. DOI: 10.1088/0264-9381/10/12/026.

[329] L. Blanchet and B. S. Sathyaprakash. "Signal analysis of gravitational wave tails". In: *Classical and Quantum Gravity* 11.11 (Nov. 1994), pp. 2807–2831. DOI: 10.1088/0264-9381/11/11/020.

[330] C. M. Will. "The Confrontation between General Relativity and Experiment". In: *Living Reviews in Relativity* 17.1, 4 (Dec. 2014), p. 4. DOI: 10.12942/lrr-2014-4. arXiv: 1403.7377 [gr-qc].

[331] M. Campanelli, C. O. Lousto, and Y. Zlochower. "Spinning-black-hole binaries: The orbital hang up". In: *Phys. Rev. D* 74 (2006), p. 041501. DOI: 10.1103/PhysRevD.74.041501. arXiv: gr-qc/0604012.

[332] R. Essick and W. Farr. "Precision Requirements for Monte Carlo Sums within Hierarchical Bayesian Inference". In: *arXiv e-prints*, arXiv:2204.00461 (Apr. 2022), arXiv:2204.00461. DOI: 10.48550/arXiv.2204.00461. arXiv: 2204.00461 [astro-ph.IM].

[333] C. Talbot and J. Golomb. "Growing Pains: Understanding the Impact of Likelihood Uncertainty on Hierarchical Bayesian Inference for Gravitational-Wave Astronomy". In: *arXiv e-prints*, arXiv:2304.06138 (Apr. 2023), arXiv:2304.06138. DOI: 10.48550/arXiv.2304.06138. arXiv: 2304.06138 [astro-ph.IM].

[334] D. W. SCOTT. "On optimal and data-based histograms". In: *Biometrika* 66.3 (Dec. 1979), pp. 605–610. ISSN: 0006-3444. DOI: 10.1093/biomet/66.3.605. eprint: https://academic.oup.com/biomet/article-pdf/66/3/605/632347/66-3-605.pdf. URL: https://doi.org/10.1093/biomet/66.3.605.

[335] L. Kish. *Survey Sampling*. Third. Oxford, England: Wiley-Interscience, 1995.

[336] V. Elvira, L. Martino, and C. P. Robert. "Rethinking the Effective Sample Size". In: *arXiv e-prints*, arXiv:1809.04129 (Sept. 2018), arXiv:1809.04129. DOI: 10.48550/arXiv.1809.04129. arXiv: 1809.04129 [stat.CO].

[337] D. W. Hogg, A. M. Price-Whelan, and B. Leistedt. "Data Analysis Recipes: Products of multivariate Gaussians in Bayesian inferences". In: *arXiv e-prints*, arXiv:2005.14199 (May 2020), arXiv:2005.14199. DOI: 10.48550/arXiv.2005.14199. arXiv: 2005.14199 [stat.CO].

[338] N. Yunes, K. Yagi, and F. Pretorius. "Theoretical Physics Implications of the Binary Black-Hole Mergers GW150914 and GW151226". In: *Phys. Rev. D* 94.8 (2016), p. 084002. DOI: 10.1103/PhysRevD.94.084002. arXiv: 1603.08955 [gr-qc].

[339] B. P. Abbott et al. "Observing gravitational-wave transient GW150914 with minimal assumptions". In: *Phys. Rev. D* 93.12 (2016). [Addendum: Phys.Rev.D 94, 069903 (2016)], p. 122004. DOI: 10.1103/PhysRevD.93.122004. arXiv: 1602.03843 [gr-qc].

[340] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, et al. "All-sky search for short gravitational-wave bursts in the first Advanced LIGO run". In: *Phys. Rev. D* 95 (4 Feb. 2017), p. 042003. DOI: 10.1103/PhysRevD.95.042003. URL: https://link.aps.org/doi/10.1103/PhysRevD.95.042003.

[341] B. P. Abbott et al. "All-Sky Search for Short Gravitational-Wave Bursts in the Second Advanced LIGO and Advanced Virgo Run". In: *Phys. Rev. D* 100.2 (2019), p. 024017. DOI: 10.1103/PhysRevD.100.024017. arXiv: 1905.03457 [gr-qc].

[342] H. S. Chia and T. D. P. Edwards. "Searching for General Binary Inspirals with Gravitational Waves". In: *JCAP* 11 (2020), p. 033. DOI: 10.1088/1475-7516/2020/11/033. arXiv: 2004.06729 [astro-ph.HE].

[343] H. S. Chia, T. D. P. Edwards, D. Wadekar, A. Zimmerman, S. Olsen, J. Roulet, T. Venumadhav, B. Zackay, and M. Zaldarriaga. "In Pursuit of Love: First Templated Search for Compact Objects with Large Tidal Deformabilities in the LIGO-Virgo Data". May 2023. arXiv: 2306.00050 [gr-qc].

[344] H. Narola, S. Roy, and A. S. Sengupta. "Beyond general relativity: Designing a template-based search for exotic gravitational wave signals". In: *Phys. Rev. D* 107.2 (2023), p. 024017. DOI: 10.1103/PhysRevD.107.024017. arXiv: 2207.10410 [gr-qc].

[345] W. James and C. Stein. "Estimation with quadratic loss". In: *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*. Univ. California Press, Berkeley, Calif., 1961, pp. 361–379.

[346] D. V. Lindley and A. F. M. Smith. "Bayes Estimates for the Linear Model". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.1 (1972), pp. 1–41. ISSN: 00359246. URL: http://www.jstor.org/stable/2985048 (visited on 04/21/2022).

[347] B. Efron and C. Morris. "Stein's Paradox in Statistics". In: *Scientific American* 236.5 (May 1977), pp. 119–127. DOI: 10.1038/scientificamerican0577-119.

[348] D. B. Rubin. "Estimation in Parallel Randomized Experiments". In: *Journal of Educational Statistics* 6.4 (1981), pp. 377–401. ISSN: 03629791. URL: http://www.jstor.org/stable/1164617 (visited on 04/21/2022).

[349] R. Essick and M. Fishbach. "DAGnabbit! Ensuring Consistency between Noise and Detection in Hierarchical Bayesian Inference". Oct. 2023. arXiv: 2310.02017 [gr-qc].

[350] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio. "Towards a generic test of the strong field dynamics of general relativity using compact binary coalescence: Further investigations". In: *J. Phys. Conf. Ser.* 363 (2012). Ed. by M. Hannam, P. Sutton, S. Hild, and C. van den Broeck, p. 012028. DOI: 10.1088/1742-6596/363/1/012028. arXiv: 1111.5274 [gr-qc].

[351] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale. "TIGER: A data analysis pipeline for testing the strong-field dynamics of general relativity with gravitational wave signals from coalescing compact binaries". In: *Phys. Rev. D* 89 (2014), p. 082001. DOI: 10.1103/PhysRevD.89.082001. arXiv: 1311.0420 [gr-qc].

[352] N. Cornish, L. Sampson, N. Yunes, and F. Pretorius. "Gravitational Wave Tests of General Relativity with the Parameterized Post-Einsteinian Framework". In: 84 (2011), p. 062003. DOI: 10.1103/PhysRevD.84.062003. arXiv: 1105.2088 [gr-qc].

[353] L. Sampson, N. Cornish, and N. Yunes. "Mismodeling in gravitational-wave astronomy: The trouble with templates". In: *Phys. Rev. D* 89.6 (2014), p. 064037. DOI: 10.1103/PhysRevD.89.064037. arXiv: 1311.4898 [gr-qc].

[354] L. Sampson, N. Cornish, and N. Yunes. "Gravitational Wave Tests of Strong Field General Relativity with Binary Inspirals: Realistic Injections and Optimal Model Selection". In: *Phys. Rev. D* 87.10 (2013), p. 102001. DOI: 10.1103/PhysRevD.87.102001. arXiv: 1303.1185 [gr-qc].

[355] J. Meidam, M. Agathos, C. Van Den Broeck, J. Veitch, and B. S. Sathyaprakash. "Testing the no-hair theorem with black hole ringdowns using TIGER". In: *Phys. Rev.* D90.6 (2014), p. 064009. DOI: 10.1103/PhysRevD.90.064009. arXiv: 1406.3201 [gr-qc].

[356] W. Del Pozzo, J. Veitch, and A. Vecchio. "Testing General Relativity using Bayesian model selection: Applications to observations of gravitational waves from compact binary systems". In: *Phys. Rev.* D83 (2011), p. 082002. DOI: 10.1103/PhysRevD.83.082002. arXiv: 1101.1391 [gr-qc].

[357] A. Ghosh et al. "Testing general relativity using golden black-hole binaries". In: *Phys. Rev. D* 94.2 (2016), 021101(R). DOI: 10.1103/PhysRevD.94.021101. arXiv: 1602.02453 [gr-qc].

[358] B. P. Abbott et al. "Binary Black Hole Mergers in the first Advanced LIGO Observing Run". In: *Phys. Rev.* X6.4 (2016). [erratum: Phys. Rev.X8,no.3,039903(2018)], p. 041015. DOI: 10.1103/PhysRevX.6.041015,10.1103/PhysRevX.8.039903. arXiv: 1606.04856 [gr-qc].

[359] J. Meidam et al. "Parametrized tests of the strong-field dynamics of general relativity using gravitational wave signals from coalescing binary black holes: Fast likelihood calculations and sensitivity of the method". In: *Phys. Rev. D* 97.4 (2018), p. 044033. DOI: 10.1103/PhysRevD.97.044033. arXiv: 1712.08772 [gr-qc].

[360] A. Ghosh, N. K. Johnson-McDaniel, A. Ghosh, C. K. Mishra, P. Ajith, W. Del Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London. "Testing general relativity using gravitational wave signals from the inspiral, merger and ringdown of binary black holes". In: *Classical Quantum Gravity* 35.1 (2018), p. 014002. DOI: 10.1088/1361-6382/aa972e. arXiv: 1704.06784 [gr-qc].

[361] R. Brito, A. Buonanno, and V. Raymond. "Black-hole Spectroscopy by Making Full Use of Gravitational-Wave Modeling". In: *Phys. Rev.* D98.8 (2018), p. 084038. DOI: 10.1103/PhysRevD.98.084038. arXiv: 1805.00293 [gr-qc].

[362] T. Akutsu et al. "Overview of KAGRA: Detector design and construction history". In: *PTEP* 2021.5 (2021), 05A101. DOI: 10.1093/ptep/ptaa125. arXiv: 2005.05574 [physics.ins-det].

[363] L. Tsukada et al. "Improved ranking statistics of the GstLAL inspiral search for compact binary coalescences". May 2023. arXiv: 2305.06286 [astro-ph.IM].

[364] R. Abbott et al. *GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run — O3 search sensitivity estimates*. Zenodo, Nov. 2021. DOI: 10.5281/zenodo.5546676. URL: https://doi.org/10.5281/zenodo.5546676.

[365] R. Abbott et al. "Open data from the third observing run of LIGO, Virgo, KAGRA and GEO". In: (Feb. 2023). arXiv: 2302.03676 [gr-qc].

[366] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton. "FINDCHIRP: An Algorithm for detection of gravitational waves from inspiraling compact binaries". In: *Phys. Rev. D* 85 (2012), p. 122006. DOI: 10.1103/PhysRevD.85.122006. arXiv: 0509116 [gr-qc].

[367] B. Allen. "A chi**2 time-frequency discriminator for gravitational wave detection". In: *Phys. Rev. D* 71 (2005), p. 062001. DOI: 10.1103/PhysRevD.71.062001. arXiv: gr-qc/0405045 [gr-qc].

[368] T. Dal Canton et al. "Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors". In: *Phys. Rev. D* 90.8 (2014), p. 082004. DOI: 10.1103/PhysRevD.90.082004. arXiv: 1405.6731 [gr-qc].

[369] S. A. Usman et al. "The PyCBC search for gravitational waves from compact binary coalescence". In: *Class. Quant. Grav.* 33.21 (2016), p. 215004. DOI: 10.1088/0264-9381/33/21/215004. arXiv: 1508.02357 [gr-qc].

[370] Q. Chu. "Low-latency detection and localization of gravitational waves from compact binary coalescences". PhD thesis. The University of Western Australia, 2017. DOI: 10.4225/23/5987feb0a789c.

[371] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga. "New search pipeline for compact binary mergers: Results for binary black holes in the first observing run of Advanced LIGO". In: *Phys. Rev. D* 100.2 (2019), p. 023011. DOI: 10.1103/PhysRevD.100.023011. arXiv: 1902.10341 [astro-ph.IM].

[372] D. Mukherjee et al. "Template bank for spinning compact binary mergers in the second observation run of Advanced LIGO and the first observation run of Advanced Virgo". In: 103.8, 084047 (Apr. 2021), p. 084047. DOI: 10.1103/PhysRevD.103.084047. arXiv: 1812.05121 [astro-ph.IM].

[373] S. Sakon et al. "Template bank for compact binary mergers in the fourth observing run of Advanced LIGO, Advanced Virgo, and KAGRA". Nov. 2022. arXiv: 2211.16674 [gr-qc].

[374] A. Bohé et al. "Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors". In: *Phys. Rev.* D95.4 (2017), p. 044028. DOI: 10.1103/PhysRevD.95.044028. arXiv: 1611.03703 [gr-qc].

[375] K. Cannon, C. Hanna, and J. Peoples. "Likelihood-Ratio Ranking Statistic for Compact Binary Coalescence Candidates with Rate Estimation". 2015. arXiv: 1504.04632 [astro-ph.IM].

[376] J. Abadie et al. "Search for Gravitational Waves from Low Mass Compact Binary Coalescence in LIGO's Sixth Science Run and Virgo's Science Runs 2 and 3". In: *Phys. Rev. D* 85 (2012), p. 082002. DOI: 10.1103/PhysRevD.85.082002. arXiv: 1111.7314 [gr-qc].

[377] S. Babak, R. Biswas, P. Brady, D. Brown, K. Cannon, et al. "Searching for gravitational waves from binary coalescence". In: *Phys. Rev. D* 87 (2013), p. 024033. DOI: 10.1103/PhysRevD.87.024033. arXiv: 1208.3491 [gr-qc].

[378] N. K. Johnson-McDaniel, A. Ghosh, S. Ghonge, M. Saleem, N. V. Krishnendu, and J. A. Clark. "Investigating the relation between gravitational wave tests of general relativity". In: *Phys. Rev. D* 105.4 (2022), p. 044020. DOI: 10.1103/PhysRevD.105.044020. arXiv: 2109.06988 [gr-qc].

[379] N. Yunes and X. Siemens. "Gravitational-Wave Tests of General Relativity with Ground-Based Detectors and Pulsar Timing-Arrays". In: *Living Rev. Rel.* 16 (2013), p. 9. DOI: 10.12942/lrr-2013-9. arXiv: 1304.3473 [gr-qc].

[380] S. Shankaranarayanan and J. P. Johnson. "Modified theories of gravity: Why, how and what?" In: *Gen. Rel. Grav.* 54.5 (2022), p. 44. DOI: 10.1007/s10714-022-02927-2. arXiv: 2204.06533 [gr-qc].

[381] G. Carullo. "Enhancing modified gravity detection from gravitational-wave observations using the parametrized ringdown spin expansion coeffcients formalism". In: *Physical Review D* 103.12 (June 2021). ISSN: 2470-0029. DOI: 10.1103/physrevd.103.124043. URL: http://dx.doi.org/10.1103/PhysRevD.103.124043.

[382] A. Maselli, P. Pani, L. Gualtieri, and E. Berti. "Parametrized ringdown spin expansion coefficients: a data-analysis framework for black-hole spectroscopy with multiple events". In: *Phys. Rev. D* 101.2 (2020), p. 024043. DOI: 10.1103/PhysRevD.101.024043. arXiv: 1910.12893 [gr-qc].

[383] A. Maselli, S. Yi, L. Pierini, V. Vellucci, L. Reali, L. Gualtieri, and E. Berti. "Black hole spectroscopy beyond Kerr: Agnostic and theory-based tests with next-generation interferometers". In: *Phys. Rev. D* 109.6 (2024), p. 064060. DOI: 10.1103/PhysRevD.109.064060. arXiv: 2311.14803 [gr-qc].

[384] G. Dideron, S. Mukherjee, and L. Lehner. "New framework to study unmodeled physics from gravitational wave data". In: *Phys. Rev. D* 107.10 (2023), p. 104023. DOI: 10.1103/PhysRevD.107.104023. arXiv: 2209.14321 [gr-qc].

[385] S. Weinberg. "Effective field theory for inflation". In: *Physical Review D* 77.12 (June 2008). ISSN: 1550-2368. DOI: 10.1103/physrevd.77.123541. URL: http://dx.doi.org/10.1103/PhysRevD.77.123541.

[386] L. C. Stein and K. Yagi. "Parametrizing and constraining scalar corrections to general relativity". In: *Phys. Rev. D* 89.4 (2014), p. 044026. DOI: 10.1103/PhysRevD.89.044026. arXiv: 1310.6743 [gr-qc].

[387] P. A. Cano, B. Ganchev, D. R. Mayerson, and A. Ruipérez. "Black hole multipoles in higher-derivative gravity". In: *JHEP* 2022.12 (2022), p. 120. DOI: 10.1007/JHEP12(2022)120. arXiv: 2208.01044 [gr-qc].

[388] V. Cardoso, M. Kimura, A. Maselli, and L. Senatore. "Black Holes in an Effective Field Theory Extension of General Relativity". In: *Phys. Rev. Lett.* 121 (25 Dec. 2018), p. 251105. DOI: 10.1103/PhysRevLett.121.251105. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.251105.

[389] R. Cayuso, P. Figueras, T. França, and L. Lehner. "Self-Consistent Modeling of Gravitational Theories beyond General Relativity". In: *Phys. Rev. Lett.* 131.11 (2023), p. 111403. DOI: 10.1103/PhysRevLett.131.111403.

[390] K. Yagi, L. C. Stein, N. Yunes, and T. Tanaka. "Post-Newtonian, Quasi-Circular Binary Inspirals in Quadratic Modified Gravity". In: *Phys. Rev. D* 85 (2012). [Erratum: Phys.Rev.D 93, 029902 (2016)], p. 064022. DOI: 10.1103/PhysRevD.85.064022. arXiv: 1110.5950 [gr-qc].

[391] H. Zhong, M. Isi, K. Chatziioannou, and W. M. Farr. "Multidimensional hierarchical tests of general relativity with gravitational waves". In: *arXiv e-prints*, arXiv:2405.19556 (May 2024), arXiv:2405.19556. DOI: 10.48550/arXiv.2405.19556. arXiv: 2405.19556 [gr-qc].

[392] C. Pacilio, D. Gerosa, and S. Bhagwat. "Catalog variance of testing general relativity with gravitational-wave data". In: *Phys. Rev. D* 109.8 (2024), p. L081302. DOI: 10.1103/PhysRevD.109.L081302. arXiv: 2310.03811 [gr-qc].

[393] G. Dideron, S. Mukherjee, and L. Lehner. "SCORE forecast for 3G". In preparation. 2024.

[394] C. K. Mishra, K. G. Arun, B. R. Iyer, and B. S. Sathyaprakash. "Parametrized tests of post-Newtonian theory using Advanced LIGO and Einstein Telescope". In: *Phys. Rev. D* 82 (2010), p. 064010. DOI: 10.1103/PhysRevD.82.064010. arXiv: 1005.0304 [gr-qc].

[395] T. P. Sotiriou and S.-Y. Zhou. "Black Hole Hair in Generalized Scalar-Tensor Gravity". In: *Physical Review Letters* 112.25 (June 2014). ISSN: 1079-7114. DOI: 10.1103/physrevlett.112.251102. URL: http://dx.doi.org/10.1103/PhysRevLett.112.251102.

[396] C. de Rham, A. J. Tolley, and J. Zhang. "Causality Constraints on Gravitational Effective Field Theories". In: *Phys. Rev. Lett.* 128.13 (2022), p. 131102. DOI: 10.1103/PhysRevLett.128.131102. arXiv: 2112.05054 [gr-qc].

[397] F. S. Bemfica, M. M. Disconzi, and J. Noronha. "Nonlinear causality of general first-order relativistic viscous hydrodynamics". In: *Physical Review D* 100.10 (Nov. 2019). ISSN: 2470-0029. DOI: 10.1103/physrevd.100.104020. URL: http://dx.doi.org/10.1103/PhysRevD.100.104020.

[398] A. Gupta et al. "Possible Causes of False General Relativity Violations in Gravitational Wave Observations". In: *arXiv eprints* (May 2024). arXiv: 2405.02197 [gr-qc].

[399] J. Y. L. Kwok, R. K. L. Lo, A. J. Weinstein, and T. G. F. Li. "Investigation of the effects of non-Gaussian noise transients and their mitigation in parameterized gravitational-wave tests of general relativity". In: *Phys. Rev. D* 105.2 (2022), p. 024066. DOI: 10.1103/PhysRevD.105.024066. arXiv: 2109.07642 [gr-qc].

[400] P. Saini, S. A. Bhat, M. Favata, and K. G. Arun. "Eccentricity-induced systematic error on parametrized tests of general relativity: Hierarchical Bayesian inference applied to a binary black hole population". In: *Phys. Rev. D* 109.8 (2024), p. 084056. DOI: 10.1103/PhysRevD.109.084056. arXiv: 2311.08033 [gr-qc].

[401] M. Maggiore et al. "Science Case for the Einstein Telescope". In: *JCAP* 2020.03 (2020), p. 050. DOI: 10.1088/1475-7516/2020/03/050. arXiv: 1912.02622 [astro-ph.CO].

[402] Z. Arzoumanian et al. "The NANOGrav 12.5-year Data Set: Search for Non-Einsteinian Polarization Modes in the Gravitational-wave Background". In: *Astrophys. J. Lett.* 923.2 (2021), p. L22. DOI: 10.3847/2041-8213/ac401c. arXiv: 2109.14706 [gr-qc].

[403] G. Agazie et al. "The NANOGrav 15 yr Data Set: Search for Transverse Polarization Modes in the Gravitational-wave Background". In: *Astrophys. J. Lett.* 964.1 (2024), p. L14. DOI: 10.3847/2041-8213/ad2a51. arXiv: 2310.12138 [gr-qc].

[404] S. Doeleman et al. "Imaging an Event Horizon: submm-VLBI of a Super Massive Black Hole". In: *astro2010: The Astronomy and Astrophysics Decadal Survey*. Vol. 2010. Jan. 2009, p. 68. DOI: 10.48550/arXiv.0906.3899. arXiv: 0906.3899 [astro-ph.CO].

[405] S. S. Doeleman, V. L. Fish, A. E. Broderick, A. Loeb, and A. E. E. Rogers. "Detecting Flaring Structures in Sagittarius A* with High-Frequency VLBI". In: 695.1 (Apr. 2009), pp. 59–74. DOI: 10.1088/0004-637X/695/1/59. arXiv: 0809.3424 [astro-ph].

[406] P. Amaro-Seoane et al. "Laser Interferometer Space Antenna". In: *arXiv e-prints*, arXiv:1702.00786 (Feb. 2017), arXiv:1702.00786. DOI: 10.48550/arXiv.1702.00786. arXiv: 1702.00786 [astro-ph.IM].

[407] J. R. Gair, M. Vallisneri, S. L. Larson, and J. G. Baker. "Testing General Relativity with Low-Frequency, Space-Based Gravitational-Wave Detectors". In: *Living Rev. Rel.* 16 (2013), p. 7. DOI: 10.12942/lrr-2013-7. arXiv: 1212.5575 [gr-qc].

[408] A. H. Nitz, C. D. Capano, S. Kumar, Y.-F. Wang, S. Kastha, M. Schäfer, R. Dhurkunde, and M. Cabero. "3-OGC: Catalog of Gravitational Waves from Compact-binary Mergers". In: *Astrophys. J.* 922.1 (2021), p. 76. DOI: 10.3847/1538-4357/ac1c03. arXiv: 2105.09151 [astro-ph.HE].

[409] B. Zackay, L. Dai, T. Venumadhav, J. Roulet, and M. Zaldarriaga. "Detecting gravitational waves with disparate detector responses: Two new binary black hole mergers". In: *Phys. Rev. D* 104.6 (2021), p. 063030. DOI: 10.1103/PhysRevD.104.063030. arXiv: 1910.09528 [astro-ph.HE].

[410] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga. "New binary black hole mergers in the second observing run of Advanced LIGO and Advanced Virgo". In: *Phys. Rev. D* 101.8 (2020), p. 083030. DOI: 10.1103/PhysRevD.101.083030. arXiv: 1904.07214 [astro-ph.HE].

[411] B. Zackay, T. Venumadhav, L. Dai, J. Roulet, and M. Zaldarriaga. "Highly spinning and aligned binary black hole merger in the Advanced LIGO first observing run". In: *Phys. Rev. D* 100.2 (2019), p. 023007. DOI: 10.1103/PhysRevD.100.023007. arXiv: 1902.10331 [astro-ph.HE].

[412] S. Biscoveanu, M. Isi, S. Vitale, and V. Varma. "New Spin on LIGO-Virgo Binary Black Holes". In: *Phys. Rev. Lett.* 126.17 (2021), p. 171103. DOI: 10.1103/PhysRevLett.126.171103. arXiv: 2007.09156 [astro-ph.HE].

[413] R. Essick, A. Farah, S. Galaudage, C. Talbot, M. Fishbach, E. Thrane, and D. E. Holz. "Probing Extremal Gravitational-wave Events with Coarse-grained Likelihoods". In: *Astrophys. J.* 926.1 (2022), p. 34. DOI: 10.3847/1538-4357/ac3978. arXiv: 2109.00418 [astro-ph.HE].

[414] H. Tong, S. Galaudage, and E. Thrane. "Population properties of spinning black holes using the gravitational-wave transient catalog 3". In: *Phys. Rev. D* 106.10 (2022), p. 103019. DOI: 10.1103/PhysRevD.106.103019. arXiv: 2209.02206 [astro-ph.HE].

[415] J. Kiefer and J. Wolfowitz. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters". In: *The Annals of Mathematical Statistics* 27.4 (1956), pp. 887–906. DOI: 10.1214/aoms/1177728066. URL: https://doi.org/10.1214/aoms/1177728066.

[416] L. Simar. "Maximum Likelihood Estimation of a Compound Poisson Process". In: *The Annals of Statistics* 4.6 (1976), pp. 1200–1209. DOI: 10.1214/aos/1176343651. URL: https://doi.org/10.1214/aos/1176343651.

[417] N. Laird. "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution". In: *Journal of the American Statistical Association* 73.364 (1978), pp. 805–811. ISSN: 01621459. URL: http://www.jstor.org/stable/2286284 (visited on 12/07/2022).

[418] D. Bohning. "Convergence of Simar's Algorithm for Finding the Maximum Likelihood Estimate of a Compound Poisson Process". In: *The Annals of Statistics* 10.3 (1982), pp. 1006–1008. DOI: 10.1214/aos/1176345890. URL: https://doi.org/10.1214/aos/1176345890.

[419] B. G. Lindsay. "The Geometry of Mixture Likelihoods: A General Theory". In: *The Annals of Statistics* 11.1 (1983), pp. 86–94. DOI: 10.1214/aos/1176346059. URL: https://doi.org/10.1214/aos/1176346059.

[420] W. Jiang and C. Zhang. "General maximum likelihood empirical Bayes estimation of normal means". In: *The Annals of Statistics* 37.4 (2009), pp. 1647–1684. DOI: 10.1214/08-AOS638. URL: https://doi.org/10.1214/08-AOS638.

[421] C. Carathéodory. "Über den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen". In: *Rendiconti del Circolo Matematico di Palermo* 32 (1911), p. 193.

[422] H. Jeffreys. *Theory of Probability*. 3rd ed. Oxford, England: Oxford, 1961.

[423] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[424] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000. DOI: 10.1137/1.9780898719857. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9780898719857. URL: https://epubs.siam.org/doi/abs/10.1137/1.9780898719857.

[425] C. Talbot and E. Thrane. "Flexible and Accurate Evaluation of Gravitational-wave Malmquist Bias with Machine Learning". In: *Astrophys. J.* 927.1 (2022), p. 76. DOI: 10.3847/1538-4357/ac4bc0. arXiv: 2012.01317 [gr-qc].

[426] J. Golomb and C. Talbot. "Hierarchical Inference of Binary Neutron Star Mass Distribution and Equation of State with Gravitational Waves". In: *Astrophys. J.* 926.1 (2022), p. 79. DOI: 10.3847/1538-4357/ac43bc. arXiv: 2106.15745 [astro-ph.HE].

[427] D. K. Lewis. *Counterfactuals*. Cambridge, MA, USA: Blackwell, 1973.

[428] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. *GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run — Parameter estimation data release*. 2021. DOI: 10.5281/zenodo.5546663. URL: %7B%5Curl%7Bhttps://zenodo.org/record/5546663%7D%7D.

[429] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. *The population of merging compact binaries inferred using gravitational waves through GWTC-3 - Data release*. 2021. DOI: 10.5281/zenodo.5655785. URL: %7B%5Curl%7Bhttps://zenodo.org/record/5655785%7D%7D.

[430] Y. Pan, A. Buonanno, A. Taracchini, L. E. Kidder, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi. "Inspiral-merger-ringdown waveforms of spinning, precessing black-hole binaries in the effective-one-body formalism". In: *Phys. Rev. D* 89.8 (2014), p. 084006. DOI: 10.1103/PhysRevD.89.084006. arXiv: 1307.6232 [gr-qc].

[431] A. Taracchini et al. "Effective-one-body model for black-hole binaries with generic mass ratios and spins". In: *Phys. Rev. D* 89.6 (2014), p. 061502. DOI: 10.1103/PhysRevD.89.061502. arXiv: 1311.2544 [gr-qc].

[432] C. Talbot, R. Smith, E. Thrane, and G. B. Poole. "Parallelized Inference for Gravitational-Wave Astronomy". In: *Phys. Rev. D* 100.4 (2019), p. 043030. DOI: 10.1103/PhysRevD.100.043030. arXiv: 1904.02863 [astro-ph.IM].

[433] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. *GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run — O3 search sensitivity estimates*. 2023. DOI: 10.5281/zenodo.7890437. URL: %7B%5Curl%7Bhttps://zenodo.org/record/7890437%7D%7D.

[434] M. Mould, D. Gerosa, F. S. Broekgaarden, and N. Steinle. "Which black hole formed first? Mass-ratio reversal in massive binary stars from gravitational-wave data". In: *Mon. Not. Roy. Astron. Soc.* 517.2 (2022), pp. 2738–2745. DOI: 10.1093/mnras/stac2859. arXiv: 2205.12329 [astro-ph.HE].

[435] E. H. Simpson. "The Interpretation of Interaction in Contingency Tables". In: *Journal of the Royal Statistical Society* 13 (1951), p. 238.

[436] Y. Chung, A. Gelman, S. Rabe-Hesketh, J. Liu, and V. Dorie. "Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models". In: *Journal of Educational and Behavioral Statistics* 40.2 (2015), pp. 136–157. DOI: 10.3102/1076998615570945. eprint: https://doi.org/10.3102/1076998615570945. URL: https://doi.org/10.3102/1076998615570945.

[437] M. Dax, S. R. Green, J. Gair, M. Pürrer, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf. "Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference". In: *Phys. Rev. Lett.* 130.17 (2023), p. 171403. DOI: 10.1103/PhysRevLett.130.171403. arXiv: 2210.05686 [gr-qc].

[438] S. Silvey. *Optimal design: an introduction to the theory for parameter estimation*. Vol. 1. Springer Science & Business Media, 1980.

[439] A. Roberts and D. Varberg. *Convex Functions*. New York: Academic Press, 1973.

[440] K. Chatziioannou, J. A. Clark, A. Bauswein, M. Millhouse, T. B. Littenberg, and N. Cornish. "Inferring the post-merger gravitational wave emission from binary neutron star coalescences". In: *Phys. Rev. D* 96.12 (2017), p. 124035. DOI: 10.1103/PhysRevD.96.124035. arXiv: 1711.00040 [gr-qc].

[441] P. J. Easter, P. D. Lasky, A. R. Casey, L. Rezzolla, and K. Takami. "Computing Fast and Reliable Gravitational Waveforms of Binary Neutron Star Merger Remnants". In: *Phys. Rev. D* 100.4 (2019), p. 043005. DOI: 10.1103/PhysRevD.100.043005. arXiv: 1811.11183 [gr-qc].

[442] K. W. Tsang, T. Dietrich, and C. Van Den Broeck. "Modeling the postmerger gravitational wave signal and extracting binary properties from future binary neutron star detections". In: *Phys. Rev. D* 100.4 (2019), p. 044047. DOI: 10.1103/PhysRevD.100.044047. arXiv: 1907.02424 [gr-qc].

[443] E. R. Most, L. J. Papenfort, V. Dexheimer, M. Hanauske, S. Schramm, H. Stöcker, and L. Rezzolla. "Signatures of quark-hadron phase transitions in general-relativistic neutron-star mergers". In: *Phys. Rev. Lett.* 122.6 (2019), p. 061101. DOI: 10.1103/PhysRevLett.122.061101. arXiv: 1807.03684 [astro-ph.HE].

[444] A. Bauswein, N.-U. F. Bastian, D. B. Blaschke, K. Chatziioannou, J. A. Clark, T. Fischer, and M. Oertel. "Identifying a first-order phase transition in neutron star mergers through gravitational waves". In: *Phys. Rev. Lett.* 122.6 (2019), p. 061102. DOI: 10.1103/PhysRevLett.122.061102. arXiv: 1809.01116 [astro-ph.HE].

[445] P. Beniamini and W. Lu. "Survival Times of Supramassive Neutron Stars Resulting from Binary Neutron Star Mergers". In: *Astrophys. J.* 920.2 (2021), p. 109. DOI: 10.3847/1538-4357/ac1678. arXiv: 2104.01181 [astro-ph.HE].

[446] L.-X. Li and B. Paczynski. "Transient Events from Neutron Star Mergers". In: 507 (1998), p. L59. DOI: 10.1086/311680. arXiv: astro-ph/9807272.

[447] B. D. Metzger, G. Martínez-Pinedo, S. Darbha, E. Quataert, A. Arcones, D. Kasen, R. Thomas, P. Nugent, I. V. Panov, and N. T. Zinner. "Electromagnetic counterparts of compact object mergers powered by the radioactive decay of r-process nuclei". In: *Monthly Notices of the Royal Astronomical Society* 406.4 (Aug. 2010), pp. 2650–2662. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2010.16864.x. eprint: https://academic.oup.com/mnras/article-pdf/406/4/2650/3356185/mnras0406-2650.pdf. URL: https://doi.org/10.1111/j.1365-2966.2010.16864.x.

[448] A. Torres-Rivas, K. Chatziioannou, A. Bauswein, and J. A. Clark. "Observing the post-merger signal of GW170817-like events with improved gravitational-wave detectors". In: *Phys. Rev. D* 99.4 (2019), p. 044014. DOI: 10.1103/PhysRevD.99.044014. arXiv: 1811.08931 [gr-qc].

[449] F. Foucart, R. Haas, M. D. Duez, E. O'Connor, C. D. Ott, L. Roberts, L. E. Kidder, J. Lippuner, H. P. Pfeiffer, and M. A. Scheel. "Low mass binary neutron star mergers : gravitational waves and neutrino emission". In: *Phys. Rev. D* 93.4 (2016), p. 044019. DOI: 10.1103/PhysRevD.93.044019. arXiv: 1510.06398 [astro-ph.HE].

[450] J. Lough et al. “First Demonstration of 6 dB Quantum Noise Reduction in a Kilometer Scale Gravitational Wave Observatory”. In: *Phys. Rev. Lett.* 126.4 (2021), p. 041102. DOI: 10.1103/PhysRevLett.126.041102. arXiv: 2005.10292 [physics.ins-det].

[451] D. Ganapathy et al. “Broadband Quantum Enhancement of the LIGO Detectors with Frequency-Dependent Squeezing”. In: *Phys. Rev. X* 13 (4 Oct. 2023), p. 041021. DOI: 10.1103/PhysRevX.13.041021. URL: https://link.aps.org/doi/10.1103/PhysRevX.13.041021.

[452] V. Baibhav, E. Berti, D. Gerosa, M. Mapelli, N. Giacobbo, Y. Bouffanais, and U. N. Di Carlo. “Gravitational-wave detection rates for compact binaries formed in isolation: LIGO/Virgo O3 and beyond”. In: *Phys. Rev. D* 100.6 (2019), p. 064060. DOI: 10.1103/PhysRevD.100.064060. arXiv: 1906.04197 [gr-qc].

[453] F. Iacovelli, M. Mancarella, S. Foffa, and M. Maggiore. “Forecasting the Detection Capabilities of Third-generation Gravitational-wave Detectors Using GWFAST”. In: *Astrophys. J.* 941.2 (2022), p. 208. DOI: 10.3847/1538-4357/ac9cd4. arXiv: 2207.02771 [gr-qc].

[454] I. Gupta et al. “Characterizing gravitational wave detector networks: from A$^\sharp$ to cosmic explorer”. In: *Class. Quant. Grav.* 41.24 (2024), p. 245001. DOI: 10.1088/1361-6382/ad7b99. arXiv: 2307.10421 [gr-qc].

[455] P. Fritschel, M. Evans, and V. Frolov. “Balanced homodyne readout for quantum limited gravitational wave detectors”. In: *Opt. Express* 22.4 (Feb. 2014), pp. 4224–4234. DOI: 10.1364/OE.22.004224. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-22-4-4224.

[456] S. Ng, S. Z. Ang, T. A. Wheatley, H. Yonezawa, A. Furusawa, E. H. Huntington, and M. Tsang. “Spectrum analysis with quantum dynamical systems”. In: *Phys. Rev. A* 93 (4 Apr. 2016), p. 042121. DOI: 10.1103/PhysRevA.93.042121. URL: https://link.aps.org/doi/10.1103/PhysRevA.93.042121.

[457] M. Tsang. “Quantum noise spectroscopy as an incoherent imaging problem”. In: *Phys. Rev. A* 107 (1 Jan. 2023), p. 012611. DOI: 10.1103/PhysRevA.107.012611. URL: https://link.aps.org/doi/10.1103/PhysRevA.107.012611.

[458] J. W. Gardner, T. Gefen, S. A. Haine, J. J. Hope, J. Preskill, Y. Chen, and L. McCuller. “Stochastic waveform estimation at the fundamental quantum limit”. In: (Apr. 2024). arXiv: 2404.13867 [quant-ph].

[459] S. M. Vermeulen et al. “Photon-Counting Interferometry to Detect Geontropic Space-Time Fluctuations with GQuEST”. In: *Phys. Rev. X* 15.1 (2025), p. 011034. DOI: 10.1103/PhysRevX.15.011034. arXiv: 2404.07524 [gr-qc].

[460] T. T. Fricke et al. “DC readout experiment in Enhanced LIGO”. In: *Class. Quant. Grav.* 29 (2012), p. 065005. DOI: 10.1088/0264-9381/29/6/065005. arXiv: 1110.2815 [physics.ins-det].

[461] M. Rakhmanov, J. D. Romano, and J. T. Whelan. “High-frequency corrections to the detector response and their effect on searches for gravitational waves”. In: *Class. Quant. Grav.* 25 (2008). Ed. by S. Hughes and E. Katsavounidis, p. 184017. DOI: 10.1088/0264-9381/25/18/184017. arXiv: 0808.3805 [gr-qc].

[462] M. van der Sluys et al. “Parameter estimation of spinning binary inspirals using Markov-chain Monte Carlo”. In: *Class. Quantum Grav.* 25 (2008), p. 184011. DOI: 10.1088/0264-9381/25/18/184011. arXiv: 0805.1689 [gr-qc].

[463] J. Veitch and A. Vecchio. “A Bayesian approach to the follow-up of candidate gravitational wave signals”. In: *Phys. Rev. D* 78 (2008), p. 022001. DOI: 10.1103/PhysRevD.78.022001. arXiv: 0801.4313 [gr-qc].

[464] T. Soultanis, K. Maltsev, A. Bauswein, K. Chatziioannou, F. K. Roepke, and N. Stergioulas. “Gravitational-wave model for neutron star merger remnants with supervised learning”. In: *Phys. Rev. D* 111.2 (2025), p. 023002. DOI: 10.1103/PhysRevD.111.023002. arXiv: 2405.09513 [astro-ph.HE].

[465] A. W. Criswell, J. Miller, N. Woldemariam, T. Soultanis, A. Bauswein, K. Chatziioannou, M. W. Coughlin, G. Jones, and V. Mandic. “Hierarchical Bayesian method for constraining the neutron star equation of state with an ensemble of binary neutron star postmerger remnants”. In: *Phys. Rev. D* 107.4 (2023), p. 043021. DOI: 10.1103/PhysRevD.107.043021. arXiv: 2211.05250 [astro-ph.HE].

[466] A. Akmal, V. R. Pandharipande, and D. G. Ravenhall. “The Equation of state of nucleon matter and neutron star structure”. In: *Phys. Rev. C* 58 (1998), pp. 1804–1828. DOI: 10.1103/PhysRevC.58.1804. arXiv: nucl-th/9804027.

[467] K.-i. Maeda, N. Ohta, and Y. Sasagawa. “Black Hole Solutions in String Theory with Gauss-Bonnet Curvature Correction”. In: *Phys. Rev. D* 80 (2009), p. 104032. DOI: 10.1103/PhysRevD.80.104032. arXiv: 0908.4151 [hep-th].

[468] T. P. Sotiriou and S.-Y. Zhou. “Black hole hair in generalized scalar-tensor gravity”. In: *Phys. Rev. Lett.* 112 (2014), p. 251102. DOI: 10.1103/PhysRevLett.112.251102. arXiv: 1312.3622 [gr-qc].

[469] R. Jackiw and S. Y. Pi. “Chern-Simons modification of general relativity”. In: *Phys. Rev. D* 68 (2003), p. 104012. DOI: 10.1103/PhysRevD.68.104012. arXiv: gr-qc/0308071.

[470] S. Alexander and N. Yunes. "Chern-Simons Modified General Relativity". In: *Phys. Rept.* 480 (2009), pp. 1–55. DOI: 10.1016/j.physrep.2009.07.002. arXiv: 0907.2562 [hep-th].

[471] É. É. Flanagan and T. Hinderer. "Constraining neutron star tidal Love numbers with gravitational wave detectors". In: *Phys. Rev. D* 77 (2008), p. 021502. DOI: 10.1103/PhysRevD.77.021502. arXiv: 0709.1915 [astro-ph].

[472] S. Tahura and K. Yagi. "Parameterized Post-Einsteinian Gravitational Waveforms in Various Modified Theories of Gravity". In: *Phys. Rev. D* 98.8 (2018). [Erratum: Phys.Rev.D 101, 109902 (2020)], p. 084042. DOI: 10.1103/PhysRevD.98.084042. arXiv: 1809.00259 [gr-qc].

[473] F. Antonini, I. M. Romero-Shaw, and T. Callister. "Star Cluster Population of High Mass Black Hole Mergers in Gravitational Wave Data". In: *Phys. Rev. Lett.* 134.1 (2025), p. 011401. DOI: 10.1103/PhysRevLett.134.011401. arXiv: 2406.19044 [astro-ph.HE].