Robustly measuring the spins of binary black holes with gravitational waves

Thesis by
Simona J. Miller

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Physics

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended May 20, 2026

© 2026

Simona J. Miller
ORCID: 0000-0001-5670-7046

Some rights reserved. This thesis is distributed under a Creative Commons Attribution.

ACKNOWLEDGEMENTS

I almost certainly would not be getting a Ph.D in gravitational-wave astrophysics were it not for my incredible experience as a LIGO SURF student in 2018, and in particular, had I not been mentored by Tom Callister. Tom, I will never forget how at the end of that summer, you took me out to lunch to discuss my career options (the first time anyone had done this with me) and said "Physics needs you." Thank you for always believing in me; for taking the time to mentor me through my undergraduate senior thesis and first-ever publication while you were a first-year postdoc; and for being a sounding board for research questions, career decisions, and personal challenges over the past eight years. Your kind, patient guidance has helped me through many bumps in the road, and has fueled my confidence, sense of belonging, and excitement about physics since day one.

Similarly, it is difficult to find words to adequately express my gratitude for the physics department at my alma mater, Smith College. I owe so much to my undergraduate professors Travis Norsen, Will Raven, Gary Felder, Nalini Easwar, Doreen Weinberger, Courtney Lannert, and Joyce Palmer-Fortune. Thank you for creating the endlessly supportive foundation upon which I flourished; for shaping me into a physicist who prioritizes education and the well-being of our world; and for modeling the type of department I dream of teaching in someday. In particular, I want to extend my ecstatic thanks to Travis for taking me under his wing and for providing continued support and guidance in the years since I left Smith; to Will for cultivating the most supportive and exciting classroom and office hours I can imagine; and to Gary for sparking my love of General Relativity. Thank you all for taking my (many) questions seriously, thus teaching me that asking for help is never something to be ashamed of. Additionally, I want to thank my Smith Physics class of 2020 comrades for getting through the loneliest months with me—in particular, Olivia Cooper and Michaela Guzetti, who I still delight in seeing at the occasional conference.

Next, I am immensely thankful for my graduate research mentors. First and foremost, I offer a profound thank you to my advisor, Katerina Chatziioannou, for sharing her vast well of knowledge, teaching me how to be a strong researcher, and providing me with countless opportunities to travel and connect with other educators and scientists across the world. Additionally, thank

you to Max Isi for hosting my formative fall 2023 semester at the Flatiron Institute Center for Computational Astrophysics and for teaching me everything I know about time-domain inference. Thank you to Alan Weinstein and Jonah Kanner for all the helpful feedback and laughs at LIGO group meetings over the years, and to Alan Weinstein, Katerina Chatziioannou, Saul Teukolsky, and Lee McCuller for serving on my candidacy and doctoral committees, including giving immensely helpful feedback on drafts of this dissertation.

Coming full circle from being a LIGO SURF student to mentor was unquestionably one of the most rewarding aspects of my PhD. I extend my wholehearted thanks to my SURF students Zoe Ko, Karen Kang, Sophia Winney, Andres Nava, and Serena Fink for their eagerness, dedication, and patience as I learned (and continue to learn) how to be a good mentor. I feel so lucky to be a part of each of your physics journeys. Thank you.

In a similar vein, thank you so much to the Pomona College Physics & Astronomy department for taking a chance and hiring me as a Lecturer to teach undergraduate cosmology during my final semester of graduate school. In particular, I wish to sincerely thank Janice Hudgings, Phil Choi, Jorge Moreno-Garcia, and Tom Moore for the priceless mentorship and opportunities, and of course, my brilliant ASTR 121 students: Sage, Izzy, Leia, Brian, Nico, Lauren, and Gada. I could not have asked for a better first class to teach.

Next, thank you to all the friends and research collaborators with whom I have had the pleasure of sharing an office at Caltech with over the years: Sophie Hourihane, Rhiannon Udall, Isaac Legred, Alvin Li, Jacob Golomb, Ethan Payne, Sophie Bini, Lucy Thomas, and Jane Glanzer. In particular, thank you to those in the third-floor West Bridge grad office for answering my $\mathcal{O}(10^6)$ questions about `bilby`. Additionally, I extend thanks to my fellow past and present graduate students and postdocs in the Caltech LIGO Lab and/or Chatziioannou group: Pat Meyers, Eliot Finch, Derek Davis, Aaron Johnson, Javier Roulet, Ryan Magee, Taylor Knapp, Lisa Drummond, Kristen Aloh, Rico Lo, Arianna Renzini, Ian MacMillan, Shruti Maliakal, Anchal Gupta, and Radhika Bhatt.

Thank you to my fellow paper-writing team members for *GWTC-4.0: Population Properties of Merging Compact Binaries*: Aditya Vijaykumar, Ben Farr, Jack Heinzel, Jaxen Godfrey, Shio Sakon, Christian Adamcewicz, and Debatri Chattopadhyay.

Thank you to the myriad of people who have ever reviewed my papers or code, been excited to discuss research with me, and/or taught me something new. This includes (but is certainly not limited to) Sylvia Biscoveneau, Sharan Banagiri, Cailin Plunkett, Sofia Alvarez Lopez, Harrison Siegel, Jess McIver, Noah Wolfe, Will Farr, Snehal Tibrewal, Aasim Jan, Matthew Mould, Colm Talbot, Salvatore Vitale, Maya Fishbach, Reed Essick, Amanda Farah, Utkarsh Mali, and Michele Mancarella. Thank you to Saavik Ford on this front, and in general for being such a delight—I can't wait to join you at CUNY!

On a more personal note, certain periods of graduate school were really lonely for me, and I'm brimming with love and gratitude for those who made the experience less so. First, thank you to Sophie Hourihane for understanding me like no one else, for teaching me the meaning of patient and steadfast friendship, and for growing together with me. Second, thank you to Sam Ponnada for bringing me into union organizing at Caltech and thus changing my life, for always making me feel less alone, and for countless lunch and coffee yaps. I am so lucky to have such kind, funny, selfless, passionate friends.

Thank you with my whole heart to my many intrepid UAW 2478 comrades, including (in no particular order) Sophie, Sam, Jasmine, Pat, Jack, Harshda, Nadia, Natasha, Abdullah, Rohan, Varun, Jess, David, Richard, Alex, Aditi, Tom, Ashay, Dee, Jakob, George, Olivia, Ingrid, Emma, Korbi, Nicolas, Ranjani, Marina, Mike, Rahma, Ruolin, Ruth, Tessa, and Quinn. I am particularly grateful for my fellow bargaining team members: I'm so proud of us for fighting through many long, demoralizing hours and coming out on top with one of the strongest first contract in higher education. We've transformed Caltech into a place I barely thought possible my first year here. Similarly, thank you to my LVK friends and fellow organizers Amanda, Storm, Aditya, Ana, Michele, Greg, and Graeme. You all fill me with hope; I am so grateful to have found people in my academic circles who bravely act upon our shared vision for a better world.

I am endlessly thankful for incredible friends I have made in Los Angeles: Emily D., Serena, Kelly, E, Jeremy, Cristy, Lucas, Lieven, Emily M., John, Nahuel, Jamie, Evie, Ari, Lance, and so many others. Your friendship has provided a much-needed respite from grad school; thank you for continually reminding me that the world is big, that I am so much happier when I have a

work-life balance, and that there is endless fun to be had in this silly life. You all are the reason LA is going to be so hard to leave.

Thank you to the North Figueroa Bookshop in Highland Park for cultivating my favorite community space in LA.

Thank you to my true rocks and oldest friends: Logan, Juliana, Sophia, and Izabella. You remind me who I am and I love you forever.

Thank you to Lou Gardner and Adam Guerin.

Thank you to the Freedmans: Lauri, Doug, Jack, Henry, and Lily.

Thank you Mom, Dad, Ansel, Theo, Ozzy, Ivan, and to my late Grandma Jane. Your support means the world to me, and I hope I've made you proud. Thank you for teaching me the importance of education for as long as I can remember.

Most of all, thank you to the love of my life, Lucy. To say that I could not have done it without you is the understatement of a lifetime. I am so incredibly lucky.

$\sim\sim\sim$

This material was supported by NSF Grants PHY-2308770 and PHY-2409001, and is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. I am grateful for computing resources provided by the LIGO Laboratory (National Science Foundation Grants PHY-0757058 and PHY-0823459), Cardiff University (STFC grants ST/I006285/1 and ST/V005618/1), and the Flatiron Institute (Simons Foundation). Additional acknowledgments are included in relevant chapters.

ABSTRACT

Over the past decade, the gravitational-wave (GW) detector network of Advanced LIGO, Advanced Virgo, and KAGRA (LVK) has advanced from the first groundbreaking observation of a merging binary black hole (BBH) to the production of a catalog of hundreds of GW signals from the astrophysical population of compact binaries. I focus on the measurement of one fundamental property of black holes: their spin, or intrinsic angular momentum. Spin is a unique probe of astrophysical processes across scales, from fluid dynamics inside stellar cores to the large-scale evolutionary history of our universe, and is the most promising means of disentangling which of the many proposed BBH formation and evolutionary mechanisms dominate the observed population. GWs remain the only way to directly measure black hole spin, yet spin remains poorly constrained: it has a comparatively weak imprint on GW signals, can mimic other physical effects like eccentricity, and is highly susceptible to features in detector noise and systematic uncertainty in waveform models. Without careful safeguards, we risk spurious spin inference and false astrophysical conclusions. To this end, I develop a suite of computational and statistical methods to produce accurate, precise, and unbiased black hole spin inference at every stage of the LVK's data analysis pipeline. To make spin measurements robust in *individual* BBH detections, I find that we must connect measured parameters with their phenomenology in GW signals. Then, for the astrophysical BBH *population*, we need to thoroughly test model behavior against simulated populations, understand the role of Monte Carlo uncertainty, and use data-level parameters to probe model misspecification.

LIGO-VIRGO-KAGRA COLLABORATION MEMBER STATEMENT

The following work does not reflect the scientific opinion of the LIGO-Virgo-KAGRA (LVK) Collaboration. This dissertation as a whole was not reviewed by the LVK Collaboration, although chapters that are additionally published elsewhere went through internal LVK review prior to publication.

PUBLISHED CONTENT AND CONTRIBUTIONS

Presented in chronological order

[1] Thomas A. Callister, **Simona J. Miller**, Katerina Chatziioannou, and Will M. Farr. No Evidence that the Majority of Black Holes in Binaries Have Zero Spin. *Astrophys. J. Lett.*, 937(1):L13, 9 2022. ***SJM*** *conceptualized the project jointly with co-authors, ran all spin magnitude and tilt population analyses, wrote around half of the paper's text, and made Figures 5-9, 12, 14, and 15.*

[2] **Simona J. Miller**, Maximiliano Isi, Katerina Chatziioannou, Vijay Varma, and Ilya Mandel. GW190521: Tracing imprints of spin-precession on the most massive black hole binary. *Phys. Rev. D*, 109(2):024024, 2 2024. ***SJM*** *assisted in the project's conceptualization, conducted all analyses, wrote the text, and made all figures.*

[3] **Simona J. Miller**, Zoe Ko, Thomas Callister, and Katerina Chatziioannou. Gravitational waves carry information beyond effective spin parameters but it is hard to extract. *Phys. Rev. D*, 109(10):104036, 10 2024. ***SJM*** *conceptualized the project jointly with co-authors, mentored Z. Ko (undergraduate), conducted most of the analyses, wrote the majority of the text, and made all of the figures except Figures 13 and 14.*

[4] Rhiannon Udall, Sophie Hourihane, **Simona J. Miller**, Derek Davis, Katerina Chatziioannou, Maximiliano Isi, and Howard Deshong. Antialigned spin of GW191109: Glitch mitigation and its implications. *Phys. Rev. D*, 111(2):024046, 2 2025. ***SJM*** *conducted the time-domain inference analysis presented in Section 3b and wrote the associated text.*

[5] Karen Kang, **Simona J. Miller**, Katerina Chatziioannou, and Deborah Ferguson. Mapping parameter correlations in spinning binary black hole mergers. *Phys. Rev. D*, 112(6):064020, 6 2025. ***SJM*** *conceptualized the project jointly with co-authors, mentored K. Kang (undergraduate), made Figure 13, and extensively edited the text.*

[6] **Simona J. Miller**, Sophie Hourihane, Maximiliano Isi, Rhiannon Udall, and Katerina Chatziioannou. `tdinf`: time domain parameter estimation for gravitational-wave signals, August 2025. ***SJM*** *lead code development, writing, testing, review, and release.*

[7] **Simona J. Miller**, Maximiliano Isi, Katerina Chatziioannou, Vijay Varma, and Sophie Hourihane. Measuring spin precession from massive black hole binaries with gravitational waves: insights from time-domain signal morphology. *Phys. Rev. D*, 112(10):104046, 10 2025. ***SJM*** *conceptualized the project jointly with co-authors, conducted all analyses, wrote all text, and made all figures.*

[8] Reed Essick et. al. incl. **Simona J. Miller**. Compact Binary Coalescence Sensitivity Estimates with Injection Campaigns during the LIGO-Virgo-KAGRA Collaborations' Fourth Observing Run. *Phys. Rev. D*, 112(10):102001, 10 2025. ***SJM*** *assisted in data generation: produced two months of software injections and reviewed another two months' injections.*

[9] LIGO-Virgo-KAGRA Scientific Collaboration incl. **Simona J. Miller**. GW250114: Testing Hawking's Area Law and the Kerr Nature of Black Holes. *Phys. Rev. Lett.*, 135(11):111403, 11 2025. ***SJM*** *was one of two key analysts who generated the data leading to the first-ever high-confidence observational confirmation of Hawking's Area Law.*

[10] LIGO-Virgo-KAGRA Scientific Collaboration incl. **Simona J. Miller**. GWTC-4.0: Updating the Gravitational-Wave Transient Catalog with Observations from the First Part of the Fourth LIGO-Virgo-KAGRA Observing Run. 8 2025. ***SJM*** *produced the population-reweighted posterior distributions, made Figure 2, and liaised between the GWTC-4.0 Catalog and GWTC-4.0 Population Properties paper-writing teams.*

[11] LIGO-Virgo-KAGRA Scientific Collaboration incl. **Simona J. Miller**. GWTC-4.0: Population Properties of Merging Compact Binaries. 8 2025. *As an invited member of the 8-person paper writing team,* ***SJM*** *wrote and made figures for everything spin-related (Section 6.3; Appendices B.5, D.2, and D.3; Figures 7-9, 17-19, and 21) and provided detailed feedback and edits on the full paper throughout the writing process. Additionally,* ***SJM*** *was a key analyst, producing effective spin measurements shown in Section 6.3.2 and Appendix D.3, and was the central coordinator of all production, result review, and synthesis spin analyses in the paper.*

[12] Rhiannon Udall, Sophie Bini, Katerina Chatziioannou, Derek Davis, Sophie Hourihane, Yannick Lecoeuche, Jess McIver, and **Simona J. Miller**. Inferring the spins of merging black holes in the presence of data-quality issues. *Phys. Rev. D*, 113(4):042005, 4 2026. ***SJM*** *assisted in the project's conceptualization, generated waveform configurations to be used in the analysis, and extensively edited the text.*

[13] Hui Tong et. al. incl. **Simona J. Miller**. Evidence of the pair-instability gap from black-hole masses. *Nature*, 4 2026. ***SJM*** *assisted in the project's conceptualization and model development.*

[14] **Simona J. Miller**, Sophia Winney, Katerina Chatziioannou, and Patrick M. Meyers. Posterior Predictive Checks for Gravitational-wave Populations: Limitations and Improvements. 4 2026. ***SJM*** *conceptualized the project, mentored S. Winney (undergraduate) and collaboratively conducted analysis together, wrote the majority of the text, and made Figures 1-3, 5, 6, and 10-18.*

[15] Yannick Lecoeuche, Jess McIver, Alan M. Knee, Rhiannon Udall, Katie Rink, Sophie Hourihane, **S. J. Miller**, Katerina Chatziioannou, TJ Massinger, and Derek Davis. Coalescing Compact Binary Parameter Estimation with Gravitational Waves in the Presence of non-Gaussian Transient Noise. 4 2026. ***SJM*** *assisted in the project's conceptualization, generated waveform configurations to be used in the analysis, and extensively edited the text.*

[16] LIGO-Virgo-KAGRA Scientific Collaboration incl. **Simona J. Miller**. GWTC-5.0: Observations from the Second Part of the Fourth LIGO-Virgo-KAGRA Observing Run and Updates to the Gravitational-Wave Transient Catalog. 5 2026. ***SJM*** *contributed population-reweighted posterior distributions and served as a figure reviewer.*

[17] LIGO-Virgo-KAGRA Scientific Collaboration incl. **Simona J. Miller**. GWTC-5.0: Population Properties of Merging Compact Binaries. 5 2026. ***SJM*** *was a central analyst and developer of spin population models, and extensively edited the text.*

[18] **Simona J. Miller**, Maximiliano Isi, and Katerina Chatziioannou. The imprint of spin precession on GW231123 is morphologically distinct from GW190521. (In preparation) ***SJM*** *conceptualized the project, conducted all analyses, wrote all text, and made all figures.*

TABLE OF CONTENTS

Acknowledgements . . . iii
Abstract . . . vii
LIGO-Virgo-KAGRA Collaboration Member Statement . . . viii
Published Content and Contributions . . . ix
Table of Contents . . . xi
List of Illustrations . . . xvi
List of Tables . . . xxiv
Nomenclature . . . xxv

Chapter I: **Introduction** . . . 1
1.1 A decade of gravitational-wave observation . . . 1
1.2 Black hole spin . . . 5
1.3 High-mass binary black holes . . . 8
1.4 Overview of dissertation . . . 11

I What spin parameters are measurable? 16

Chapter II: **The imprint of black hole spin on a gravitational-wave signal** . . . 17
2.1 Newtonian versus Einsteinian formulations of gravity . . . 17
2.2 Gravitational waves 101 . . . 19
2.3 Post-Newtonian expansions . . . 22
2.4 Effective inspiral spin parameters . . . 25
2.5 Spins in the ringdown . . . 33

Chapter III: **Bayesian inference** . . . 35
3.1 Gravitational wave parameter estimation . . . 36
3.2 Population inference: from one detection to many . . . 40

Chapter IV: **Mapping parameter correlations in spinning binary black hole mergers** . . . 46
4.1 Introduction . . . 47
4.2 Methodology . . . 51
4.3 Recovering Spin-Aligned Inspiral correlations at 30 $M_\odot$ . . . 58
4.4 Aligned-spin correlations beyond the inspiral . . . 62
4.5 Precessing degrees of freedom . . . 67
4.6 Conclusions . . . 71

Chapter V: **Gravitational waves carry information beyond effective spin parameters but it is hard to extract** 75
5.1 Introduction 76
5.2 Spin magnitudes and tilts versus effective spin parameters . . 78
5.3 Methodology 79
5.4 Different spin magnitude and tilt distributions can be distinguished 81
5.5 Difficulties of Measuring Component Spin Distributions 83
5.6 Identifying bias is difficult: limitations of posterior predictive checks 87
5.7 Comparison to past work on Hierarchical Inference with Spin-Precession 93
5.8 Conclusions 94
5.A Simulated Populations 97
5.B Individual-event Parameter Estimation 98
5.C Hierarchical Inference 100
5.D Spin population models 102
5.E Detailed Hierarchical Inference Results 105
5.F Reweighting individual event posteriors 106
5.G Verification of methods 106

II Where in gravitational-wave signals do measurements of binary black hole spin originate? 124

Chapter VI: **Time-domain inference** 125
6.1 Time-domain likelihood 125
6.2 Time-truncated analyses 126

Chapter VII: **Tracing imprints of spin-precession on the massive black hole binary GW190521** 129
7.1 Introduction 130
7.2 Methods 132
7.3 Results 133
7.4 Conclusions 139
7.A Parameter estimation settings and priors 141
7.B Whitening Waveforms 142
7.C Quantifying precession 144
7.D Signal strength over time 149

Chapter VIII: **Insights from time-domain signal morphology on measuring spin precession from massive black hole binaries** 151
8.1 Introduction 152
8.2 Methods 156
8.3 Analyzing the GW190521 maximum likelihood waveform . . . 162
8.4 Exploring Spin Measurability 166

8.5 Conclusions . . . 183
8.A Measuring more mass and spin parameters in the post-peak data and exploring conditional priors . . . 189
8.B Extracting the merger frequency . . . 192
8.C Measuring the mass parameters for systems with different M . 192
8.D Impact of extrinsic angles: Equations and waveforms . . . 193

Chapter IX: **Exploring eccentric and intermediate mass ratio interpretations of GW190521 with time-domain inference** . . 198
9.1 Introduction . . . 198
9.2 Eccentricity e . . . 199
9.3 Mass ratio q . . . 204
9.4 Conclusion . . . 209

Chapter X: **Using time-domain inference to link the anti-aligned spin of GW191109 with a glitch** . . . 211
10.1 Introduction . . . 211
10.2 Modeling the Compact Binary Signal . . . 214
10.3 Understanding the GW191109 Inference . . . 215

Chapter XI: **Testing Hawking's Area Law with GW250114** . . . 222
11.1 The Area Law . . . 223
11.2 Testing the Area Law with a Merging Binary Black Hole . . . 223
11.3 GW250114 . . . 224
11.4 Results . . . 225
11.A Inspiral Analysis Details . . . 231

III What do we know about the spins of the binary black hole population? 234

Chapter XII: **No evidence that the majority of black holes in binaries have zero spin** . . . 235
12.1 Introduction . . . 236
12.2 The spins of black holes in binaries . . . 237
12.3 A counting experiment . . . 242
12.4 No sharp features in the effective spin distribution . . . 245
12.5 The population of spin magnitudes and tilts . . . 250
12.6 Discussion and Conclusions . . . 258
12.A Spin population models . . . 264
12.B Hierarchical inference and data analyzed . . . 268
12.C Model Checking . . . 270
12.D Effective samples . . . 272
12.E Avoiding samples with likelihood KDEs . . . 275
12.F Mock injection campaign . . . 278
12.G Bayes factors through the Savage Dickey density ratio . . . 281

12.H Different Catalogs and Model Variations 285

Chapter XIII: **The astrophysical distribution of binary black hole spins in GWTC-4.0** 288
13.1 Spin Magnitudes and Tilts 289
13.2 Effective Spins 295
13.A Spin Models 298
13.B Model Comparison Study: Spin Magnitudes and Tilt Angles . 302
13.C Dependence of the Effective Spin Distribution on the Likelihood Variance Cut 303
13.D Comparison of Weakly Modeled Approach Results 307

Chapter XIV: **Limitations and improvements for posterior predictive checks** 310
14.1 Introduction 311
14.2 Summary of Main Results 313
14.3 Methods 317
14.4 Event vs. Data Level PPCs 325
14.5 Partial PPCs 338
14.6 Split Predictive Checks 342
14.7 PPCs applied to GWTC-4.0 344
14.8 Recommendations and Future Work 347
14.A Simulated Population and Inference 350
14.B Derivation of posterior predictive p-values 351
14.C Optimizing the Full 15-dimensional BBH GW likelihood . . . 353
14.D Simple Toy Model 356

Chapter XV: **Conclusion** 362
15.1 Summary of dissertation 362
15.2 Future plans 363
Bibliography 368
Pocket Material: Time-domain inference on GW231123
Pocket Material: GW231123 Infographic

LIST OF ILLUSTRATIONS

Number *Page*

1.1 Schematic of the inspiral, merger, and ringdown phases of a binary black hole coalescence. 2
1.2 Mass and spin measurements for all new gravitational wave signals in GWTC-4.0. 3
1.3 Schematic of masses, spin magnitudes, and spin tilt angles of a black hole binary 6
1.4 Illustration of the effect of several proposed binary black hole formation and evolution scenarios on spin magnitudes and tilts 7
1.5 Comparison of gravitational-wave signals from lighter vs. heavier binary black holes 9
2.1 Schematic of the plus and cross GW polarizations 21
2.2 Leading order spin coefficients in post-Newtonian expansion. . . 26
2.3 Impact of aligned spin on BBH waveforms. 27
2.4 Location of innermost stable circular orbit around a spinning BH versus its spin magnitude 29
2.5 Diagram of spin and orbital angular momenta summing to the total angular momentum $\vec{J}$. 29
2.6 Impact of precessing spin on BBH waveforms. 31
2.7 QNM frequencies as a function of spin 34
3.1 Schematic example of population inference inputs from individual events under some idealized scenarios. 41
4.1 Example waveforms for equal-mass BBHs with zero spins at different total masses ($30\,M_\odot$, $90\,M_\odot$, and $270\,M_\odot$). 54
4.2 Workflow diagram for the BBH intrinsic parameter correlation mapping algorithm. 55
4.3 Correlation path in $\eta, \chi_{\rm eff}$ for BBHs with a total mass of $30\,M_\odot$ (inspiral-dominated). 59
4.4 Waveforms demonstrating the $\chi_{\rm eff} - \eta$ correlation from Fig. 4.3. 60
4.5 Correlation path in q, χ_{1z}, χ_{2z} for BBHs with a total mass of $30\,M_\odot$. 61
4.6 Comparison of correlation paths in $\eta, \chi_{\rm eff}$ mapped directly in that space versus projected from mapping in q, χ_{1z}, χ_{2z}. 62

4.7 Similar to Fig. 4.3 but for BBHs with a total mass of $90\,M_\odot$ where both the inspiral and the merger phase are observable.. . 63
4.8 Similar to Fig. 4.3 but for BBHs with a total mass of $270\,M_\odot$ where only the merger phase is observable. 64
4.9 Correlation paths that have been constructed in the 3D aligned-spin parameter space of q, χ_{1z}, χ_{2z} and then projected in pairwise 2D subspaces for BBHs with total masses of $30\,M_\odot$, $90\,M_\odot$, and $270\,M_\odot$. 65
4.10 Comparison of the $\chi_{\rm eff} - \eta$ correlation for signals with total mass $30\,M_\odot$, $90\,M_\odot$, and $270\,M_\odot$. 65
4.11 Comparison of precessing correlations for BBHs with a total mass of $30\,M_\odot$, $90\,M_\odot$, and $270\,M_\odot$ at two points in η, χ_1 space. . . . 68
4.12 Comparison of precessing correlations for BBHs with a total mass of $30\,M_\odot$, $90\,M_\odot$, and $270\,M_\odot$ at 12 points in η, χ_1 space. . . . 69
4.13 A demonstrative example of how small variations in η and in-plane χ_1 affect signal morphology in the late inspiral and merger. 70
5.1 Three simulated BBH populations with the same effective spin ($\chi_{\rm eff}$) distribution but different underlying spin magnitude and tilt distributions. 79
5.2 Inferred spin distributions from the true populations shown in Fig. 5.1 using a Beta distribution for the spin magnitude model and a mixture of two Gaussians for the cosine-tilt model. . . . 81
5.3 Posterior predictive checks for the inferred populations in Fig. 5.2. 88
5.4 Posterior predictive checks for spin tilts with inrceasingly large single-event measurement uncertainty, using a Gaussian single-event likelihood. 90
5.5 Observed and detected mass and redshift distributions for the simulated populations of Fig. 5.1. 99
5.6 Inferred hyper-parameters corresponding to the HighSpinPrecessing population in Fig. 5.2. 107
5.7 Inferred hyper-parameters corresponding to the MediumSpin population in Fig. 5.2. 108
5.8 Inferred hyper-parameters corresponding to the LowSpinAligned population in Fig. 5.2. 109
5.9 Probability-probability (P-P) plots for masses, spins, and redshifts for the three simulated populations in Figs. 5.1 and 5.5. . 111

5.10 Similar to Fig. 5.2 but for Gaussian single-event likelihoods. . . 113
5.11 Similar to Fig. 5.2 but with various methods of reducing complexity when sampling of individual-event posteriors. 115
5.12 Inferred spin magnitude distribution for the LowSpinAligned population where the spin tilt distribution is fixed to the truth, and vice versa. 116
5.13 Similar to Fig. 5.2, but now additionally inferring the mass and redshift distributions of the three simulated populations in conjunction with their spin distributions. 118
5.14 Similar to Fig. 5.2, but now fitting for the absolute rate of black hole mergers as a function of spin, rather than component spin probability distributions. 119
5.15 Inferred spin magnitude distribution for the LowSpinAligned population under different sorting schema within the bimodal Gaussian model, and with vs. without the inclusion of spin selection effects. 122
5.16 Similar to Fig. 5.2 but with different subsets of events. 123
6.1 Cartoon explaining pre/post-cutoff signal analysis, enabled by time-domain inference 127
7.1 Evolution of GW190521 inference for representative cutoff times $t \in \{-40, -20, -10, 20, 30\}\, M_{\rm ref}$ from the time of coalescence. . 134
7.2 Comparison of GW190521 results from the full and post-$t = -10\, M_{\rm ref}$ analysis that are informative and uninformative about precession, respectively. 136
7.3 GW190521 posteriors for $\chi_{\rm p}$, M (detector frame total mass), and q (mass ratio) for the same cutoff times as Fig. 7.1. 137
7.4 Comparison of GW190521 posteriors when time of coalescence and sky position are fixed versus sampled over. 143
7.5 Comparison of GW190521 posteriors from LVK analysis versus Miller et. al. 143
7.6 Posteriors for $\chi_{\rm p}$ and $\chi_{\rm eff}$ for GW190521 informed by data after five representative cutoff times: $t = \{-40, -20, -10, 20, 30\}\, M_{\rm ref}$. 145
7.7 GW190521 posteriors for $\chi_{\rm p}$, generalized $\chi_{\rm p}$, and $||\vec{\chi}_\perp||$ for the full, post-$t = -40\, M_{\rm ref}$, and post-$t = -10\, M_{\rm ref}$ analyses. 146
7.8 Network matched-filter SNR of GW190521 for the pre- and post-cutoff analyses as a function of the cutoff time. 148

7.9 Regions in time during which the 50% credible intervals of the reconstructed waveform strain are inconsistent between the full and post-$t = -10\,M_{\rm ref}$ analyses of GW190521. 149
8.1 Evolution of posteriors for representative cutoff times for GW190521 and its max. $\mathcal{L}$ waveform, exemplifying that the max. $\mathcal{L}$ waveform behaves similarly to the real data. 163
8.2 Evolution of the $\chi_{\rm p}$ posterior for the GW190521 max. $\mathcal{L}$ waveform at different full-signal optimal network SNRs. 166
8.3 Information about $\chi_{\rm p}$ in the ringdown of GW190521's max. $\mathcal{L}$ waveform at different SNRs. 167
8.4 Waveforms and inference results for GW190521's max. $\mathcal{L}$ signal at different total masses. 171
8.5 For GW190521's max. $\mathcal{L}$ injection, the Jenson Shannon Divergence (JSD) between the pre-cutoff $\chi_{\rm p}$ posterior and the prior versus the pre-cutoff SNR for different total masses and a range of cutoff times. 174
8.6 Schematic diagram explaining the effect of changing inclination, polarization, and phase on a binary's configuration and the resultant gravitational waveform. 175
8.7 Results from full-duration signal analyses where one extrinsic angle at a time is changed from GW190521's max. $\mathcal{L}$ configuration. 179
8.8 Results from post-cutoff analyses for GW190521's max. $\mathcal{L}$ signal oriented at various extrinsic angles. 182
8.9 For GW190521's max. $\mathcal{L}$ injection, the JSD between the prior and posterior for effective spins and masses from data after different cutoff times and SNRs. 190
8.10 Comparison of JSD for conditioned versus un-conditioned priors. 192
8.11 Amplitude spectral density versus frequency for five signals with different total masses. 193
8.12 Evolution of the total mass M and mass ratio q as a function of cutoff time for GW190521's max. $\mathcal{L}$ injection at different total masses. 194
8.13 Effect on a gravitational waveform of varying the polarization, inclination, and phase angle at a fixed SNR. 197
9.1 Time-cut comparison of eccentric and precessing signals that are consistent with GW190521 202

9.2 Time-cut comparison of equal- and unequal-mass signals that are consistent with GW190521 205
9.3 Pre-merger strain versus $\chi_{\rm p}$ for different interpretations of GW190521 207
9.4 Pre-merger strain versus mass parameters for signals consistent with GW190521 209
10.1 Spectrograms of LIGO Livingston data (before glitch mitigation) centered around the time of GW191109, overlaid with time-frequency tracks of the scattered light glitch. 213
10.2 Tracing the $\chi_{\rm eff}$ inference of GW191109 across frequencies and times. 217
10.3 Shifts in the probability of $\chi_{\rm eff} \leq 0$ for 100 simulated signals consistent with GW191109 in Gaussian noise. 221
11.1 Schemata for the "radius" of a black hole and Hawking's Area Law. 224
11.2 GW250114 strain data 226
11.3 Component and remnant masses and spins of GW250114 for inspiral-only, ringdown-only, and full-signal analyses. 227
11.4 Area Law test on GW250114. 228
11.5 Measurements for the initial and final black hole areas of GW250114 as a function of fit time. 229
11.6 The significance of a non-detection of a violation of Hawking's Area Law as a function of fit time. 230
11.7 Comparison of time- and frequency-domain mass and spin posteriors for GW250114 full-signal analysis with two waveform models. 233
11.8 Testing Area Law method with GR and non-GR injections. . . 233
12.1 Posterior on the fraction of BBHs in GWTC-2 in which both components have zero spin, as obtained in the counting experiment of Sec. 12.3. 244
12.2 Comparison of the marginal population distributions for $\chi_{\rm eff}$ from that have no excess of zero-spin events and versus those that imply that $\sim 50\%$ of binaries have zero spins. 246
12.3 Marginalized posterior on the fraction of BBHs comprising a distinct sub-population with $\chi_{\rm eff} = 0$ for GWTC-3 and mock catalogs. 247
12.4 The $\chi_{\rm eff}$ distribution for BBHs in GWTC-3, as inferred by model consisting of the sum of two Gaussians. 249

12.5 Inferred distributions of spin magnitudes, tilts, and $\chi_{\rm eff}$ in GWTC-3 under various component-spin models. 253
12.6 Inferred spin magnitude and tilts hyper-parameters GWTC-3 under various component-spin models. 254
12.7 Posterior distributions for the 1st percentile of the $\chi_{\rm eff}$ distributions implied by the component spin models of Fig. 12.5. 257
12.8 Posterior predictive checks of the three effective spin models explored in Sec. 12.4. 272
12.9 Posterior predictive checks of the component spin models studied in Sec. 12.5. 273
12.10 Exploration of per-event effective samples with regards to a zero-spin subpopulation. 275
12.11 Spin magnitude posterior samples demonstrating events with and without support for a zero-spin subpopulation. 276
12.12 Validation of the Kernel-Density-Estimate likelihood approach discussed in Appendix 12.E 279
12.13 Follow-up exploration of an injected mock catalog that most strongly (and incorrectly) finds support for the existance of a zero-spin subpopulation. 280
12.14 Exploration of probabilities that individual events in GWTC-2 are non-spinning. 282
12.15 Comparing the fraction of BBHs that are non-spinning in GWTC-2 vs. GWTC-3. 283
12.16 Comparing component-spin hyperparameters for GWTC-2 and GWTC-3. 284
12.17 Comparison of component spin models on GWTC-2 and GWTC-3. 285
13.1 Spin magnitude χ (left) and cosine tilt distributions under the Gaussian Component Spins and B-Spline models on GWTC-4.0. 290
13.2 Spin-sorted spin magnitude distributions for GWTC-4.0. 292
13.3 Marginal $\chi_{\rm eff}$ and $\chi_{\rm p}$ distributions for GWTC-4.0. 294
13.4 Gaussian Component Spins hyperparameters when the spin magnitudes and tilts of the primary and secondary BHs are assumed identically versus non-identically distributed. 304

13.5 Gaussian Effective Spins hyperparameters for GWTC-4.0 under two methods of cutting out samples that may lead to an unconverged likelihood. 306
13.6 Marginal and joint $\chi_{\rm eff}$ and $\chi_{\rm p}$ distributions for GWTC-4.0 under the Gaussian Effective Spins model using two methods of cutting out hyperparameter samples that may lead to an unconverged likelihood. 307
13.7 Comparison of many models for $\chi_{\rm eff}$ on GWTC-4.0. 309
14.1 Summary figure of posterior predictive p-values for all PPCs, test statistics, and single-event likelihoods explored in Ch. 14. . . . 314
14.2 Schematic showing the steps for conducting a posterior predictive check (PPC). 321
14.3 Schematic showing the steps for calculating a posterior predictive p-value. 321
14.4 Results of population inference and event-level versus data-level PPCs for $\cos\theta$ using a Gaussian population model for simulated 70-event catalogs from a bimodal population using four likelihood variants. 329
14.5 Distributions and associated posterior predictive p-values for four test statistics calculated on $\cos\theta$ for simulated 70-event catalogs from a bimodal population using four likelihood variants. . . . 335
14.6 Fraction of simulated universes where we are able to discern model misspecification using p-values as a function of single-event measurement uncertainty (using a toy model). 336
14.7 Partial PPCs and associated p-values where test statistic fixed between predicted and observed catalogs is the ratio of tilts with $\cos\theta$ between $-1/3$ and $1/3$ to those with $\cos\theta > 1/3$. 340
14.8 Partial PPCs and associated p-values where test statistic fixed between predicted and observed catalogs is the standard deviation of a catalog's $\cos\theta$ draws. 341
14.9 Same as Fig. 14.7 but for the SPC, rather than pPPC. 343
14.10 Event and data-level PPCs applied to the GWTC-4.0 spin magnitude and tilt distributions. 345
14.11 Underlying and detected distributions of all BBH parameters for the LowSpinAligned population. 351

14.12 (a.) Histogram of max. $\mathcal{L}$ samples from stochastically sampled posteriors versus max. $\mathcal{L}$ values obtained via full likelihood optimization for the same underlying population and (b.) The CDFs for three example populations' true and corresponding max. $\mathcal{L}$ $\cos\theta$ distributions. 354
14.13 Predicted vs. observed max. $\mathcal{L}$ $\cos\theta$ distributions for four different likelihoods models. 355
14.14 Observed versus predicted distributions for max. $\mathcal{L}$ and true values of spin magnitudes and tilts for GWTC-4.0. 355
14.15 Results from hierarchical inference performed analytically with a Gaussian population model and Gaussian individual-event likelihoods at three different measurement uncertainties. 357
14.16 Posterior predictive p-values corresponding to Fig. 14.15. 358
14.17 Similar to Fig. 14.15 but with a true underlying distribution that is bimodal, and therefore cannot be accurately recovered by the Gaussian population model. 361
14.18 Posterior predictive p-values corresponding to Fig. 14.17. 361
15.1 Comparison of spin precession measurements before and after various cutoff times for GW2311223 and GW190521 420
15.2 "Meet GW231123" Infographic 421

LIST OF TABLES

Number *Page*

5.1 Spin magnitude and tilt population model details for the hierarchical inference of Ch. 5. . . . 103
7.1 Priors used in time-domain parameter estimation on GW190521. 142
7.2 Settings for the time-domain parameter estimation for GW190521. . . . 144
8.1 The maximum likelihood parameters from two time-domain inference analyses conducted on GW190521: one where all parameters are defined at $f_{\rm ref} = 11\,{\rm Hz}$ and another at $t_{\rm ref} = -100\,M$. 161
8.2 Merger frequency and SNR information of the signals with different total masses shown in Fig. 8.4. . . . 174
11.1 Reference parameters for the GW250114 inspiral and ringdown analyses. . . . 231
12.1 Physical quantities of interest derived from component spin models applied to GWTC-3, including the fraction of BHs with $\chi \sim 0$ and the minima and maxima of the tilt and $\chi_{\rm eff}$ distributions. . 252
12.2 Summary of effective spin models of Ch. 12. . . . 265
12.3 Summary of component spin models of Ch. 12. . . . 267
13.1 Summary of probes of spin misalignment in GWTC-4.0, measured by $\chi_{\rm eff}$. . . . 296
13.2 Summary of Gaussian Component Spins model parameters for spin magnitudes and tilts for GWTC-4.0. . . . 299
13.3 Summary of Gaussian Effective Spins (Equation 13.11) and Skew-normal Effective Spin (Equation 13.12) spin parameters. . . . 301
13.4 Comparison of different parametric population models for spin magnitudes and tilts on GWTC-4.0. . . . 303

NOMENCLATURE

Acronyms

LIGO Laser Interferometer Gravitational Wave Observatory

LVK LIGO-Virgo-KAGRA

BH Black hole

BBH Binary black hole

CBC Compact binary coalescence (includes binary black holes, binary neutron stars, and neutron-star black holes)

PPC Posterior predictive check

GR General relativity

NR Numerical relativity

PN Post-Newtonian

Binary black hole parameters

Note: A subscript of 1 *indicates that the quantity describes the more-massive black hole in the binary, while* 2 *corresponds to the less-massive.*

$m_{1,2}$ Black hole mass

M Binary black hole total mass; $M = m_1 + m_2$

q Binary black hole mass ratio; $q = m_2/m_1 \leq 1$

$\chi_{1,2}$ Dimensionless black hole spin magnitude; $\chi_i \in [0, 1]$

$\theta_{1,2}$ Angle between black hole spin and orbital angular momentum; often parameterized with the cosine of the angle; $\theta_i \in [0, \pi] \Rightarrow \cos\theta_i \in [-1, 1]$.

z Redshift

Chapter 1

INTRODUCTION

1.1 A decade of gravitational-wave observation

Throughout the cosmos, pairs of stellar mass black holes, i.e., $\mathcal{O}(1{-}100\,M_\odot)$, orbit each other for millions of years, spiraling inwards as they emit ripples in space-time called gravitational waves. Under finely-tuned astrophysical conditions [e.g., 1, 2], these binary black holes can lose sufficient energy via gravitational radiation to merge within a Hubble time, resulting in our cosmos' most extreme collisions: in a fraction of a second, one merger releases over ten times more energy than all starlight in the observable universe [3].

Over the past ten years, the global detector network of Advanced LIGO [4], Advanced Virgo [5], and KAGRA [6] (LVK) has observed hundreds of gravitational-wave signals from binary black hole mergers [7–9].[1] Each signal (cf., Fig. 1.1) encodes information about its compact binary origins, including the masses and spins[2] of the component and remnant black holes, the distance to the source, its orientation in the sky relative to the observer's line of sight, etc. As gravitational-wave astronomers and data analysts, we are responsible for modeling these source properties, robustly constraining them in the data, and leveraging them towards our collective understanding of the cosmos.

Our (publicized) detections so far[3]—some of whose source properties are showcased in Fig. 1.2—are bookended by the groundbreaking GW150914 [17] and its much louder yet nearly identical successor GW250114 [18], the latter of which was recently used to test the nature of gravity [19] and fundamental laws of energy and entropy [Ch. 11] with high precision. In the interim decade, we have observed a wide range of binary black hole signals, including those from highly-massive sources like GW190521 [20] and GW231123 [21]—each

[1] We have also observed binary neutron stars [10] and neutron-star-black-holes [11, 12], but the focus of this dissertation is binary black holes, the most abundant source class.

[2] Having no material extent, black holes do not *rotate* but rather spin, analogously to electrons (albeit non-quantized).

[3] This dissertation is being prepared concurrently with the release of GWTC-5.0, the LVK gravitational-wave transient catalog containing compact binaries detected through the O4b observing run. The GWTC-5.0 data are planned to be made public at the end of May 2026; their discussion is excluded from this dissertation.

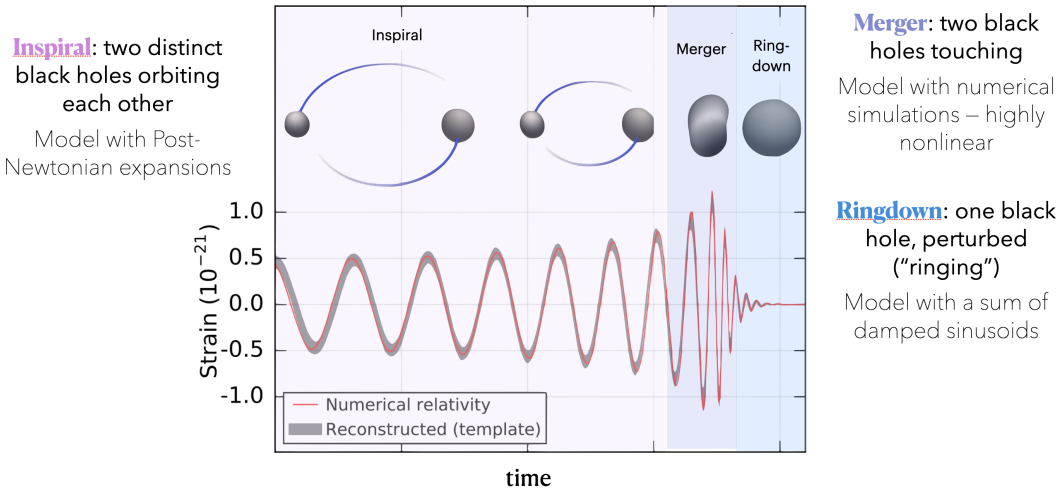


Figure 1.1: A schematic example of strain versus time for a gravitational-wave signal sourced by a binary black hole coalescence. The signal can be split into three regimes: the inspiral (pink), where the system contains two distinct black holes orbiting each other at a distance; the merger (purple), where the event horizons of the black holes start to overlap and the two bodies become one; and the ringdown (blue), consisting of quasi-normal modes excited by the perturbed remnant black hole. Each of these regimes requires different modeling techniques: the inspiral and ringdown can be expressed analytically, with the former relying on post-Newtonian expansions [e.g., 13, 14] and the latter expressed as a superposition of damped sinusoids [e.g., 15]. The merger phase is highly non-linear and must be modeled with numerical relativity simulations [e.g., 16]. Waveform modeling approaches are further described in Ch. 2.

weighing over 100 times our Sun (Sec. 1.3)—as well as rapidly spinning and asymmetric systems like GW241011 and GW241110 [22]. The ever-growing Gravitational-wave Transient Catalog (GWTC) allows us to measure the astrophysical population of compact binary mergers [e.g., 23, Ch. 13], test general relativity [e.g., 24], constrain our universe's cosmology [e.g., 25], and more.

This dissertation is being written at a time when gravitational-wave astronomy, while still young, has exited its true infancy. However, the field remains data-limited with many exciting "firsts" yet to come.[4] Our computational toolbox for studying gravitational waves is rapidly expanding, rising to the challenge of digging deeper into our rich dataset, as well as that predicted for next-generation detectors like Cosmic Explorer [27], the Einstein Telescope [28], and LISA [29]. This is to say, gravitational-wave astronomers have only scratched

[4] We haven't even (believably) detected orbital eccentricity yet! (Although it is possible that I am operating under an overly-strict prior for what counts as "believable.")

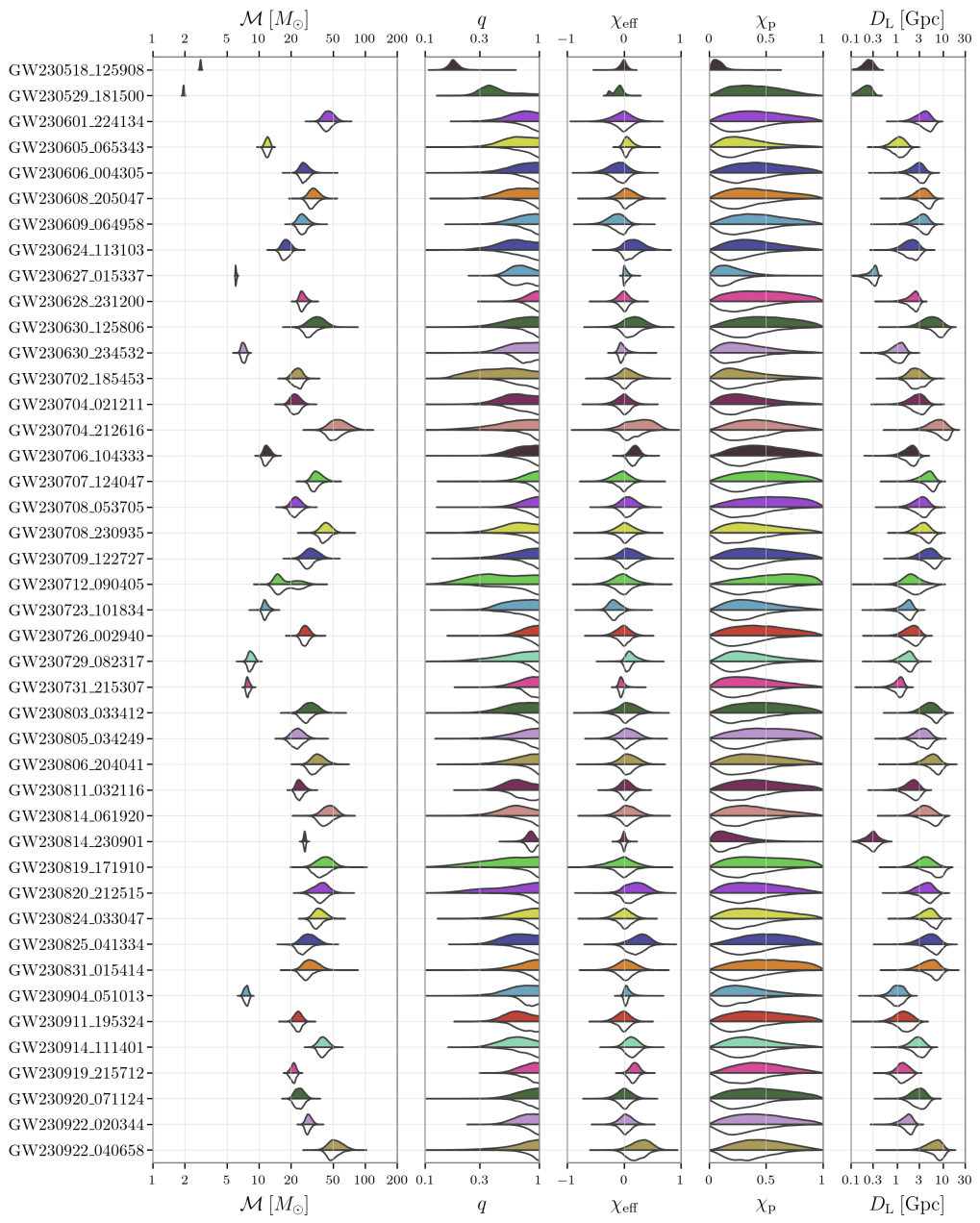


Figure 1.2: The marginal probability distributions for the source frame chirp mass $\mathcal{M}$, mass ratio q, effective inspiral spin $\chi_{\rm eff}$, effective precession spin $\chi_{\rm p}$, and luminosity distance D_L (see Ch. 2 for a discussion of these parameters) for O4a candidates with a detection false alarm rate FAR $< 1\,{\rm yr}^{-1}$. The colored upper half of the plot shows the marginal posterior distributions using our default agnostic priors [8], while the white lower halves show these marginal distributions after reweighting according to the inferred population model [23] for each binary black hole. The two neutron star black hole candidates have only the non-reweighted posterior distributions. *(Figure continued on following page.)*

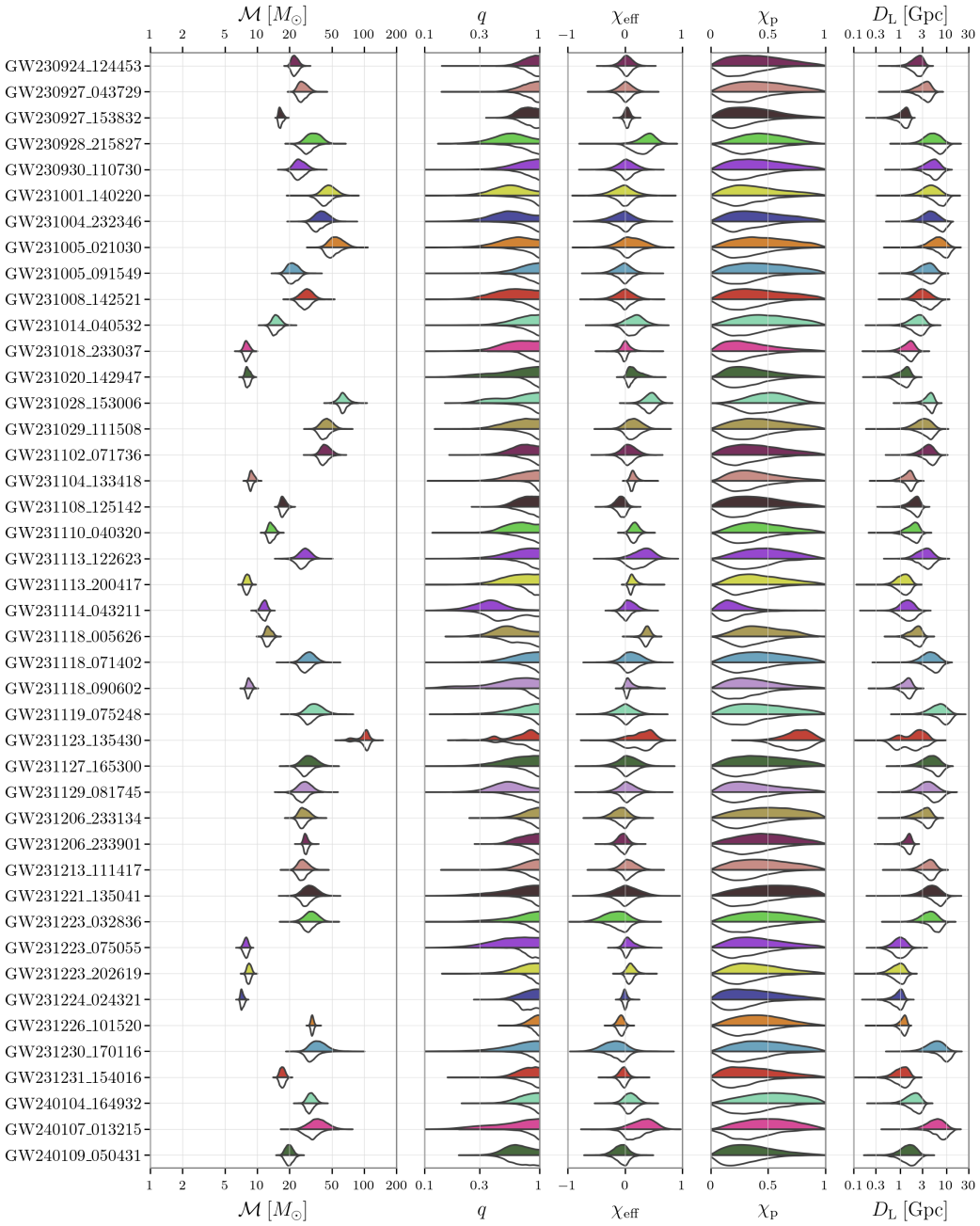


Figure 1.2: *(Continuation of previous page. This figure was re-produced from Ref. [26]. The population-informed posteriors were produced by* ***SJM****.)*

the surface of a truly robust understanding of many elements of our current and future data. Two such aspects—black hole spin (Sec. 1.2) and high masses (Sec. 1.3)—serve as the primary and secondary foci of my research.

1.2 Black hole spin

Gravitational waves remain the only direct[5] way to measure black hole **spin**, or intrinsic angular momentum—the primary topic of this dissertation. Spin is a crucial yet notoriously difficult-to-constrain probe of astrophysical processes across many scales, ranging from fluid dynamics inside stars [e.g., 31], to the cosmic environments where compact binaries form [e.g., 32], to the large-scale evolutionary history of our universe [e.g., 33].

Despite ongoing improvements to both LVK detector sensitivity [34] and theoretical astrophysical models, the cosmic origins of binary black hole mergers remain uncertain [e.g., 23, 35]. Several theorized evolutionary pathways can produce binary black holes that coalesce within the age of the universe; disentangling which of these dominate the observed population is an open question at the forefront of astrophysics.

Broadly, binary black hole formation channels can be sorted into two categories: *(i.)* **isolated binary formation** in the galactic field where the components were already in a binary system *before* becoming black holes, and *(ii.)* **dynamical formation** in dense environments like globular clusters in which the components become gravitationally bound *after* black hole formation. Isolated binary formation encompasses pathways like double common envelope evolution [e.g., 36], stable mass transfer [e.g., 37], and chemically homogeneous evolution [e.g., 38]; dynamical formation can occur in e.g., globular [e.g., 39], nuclear [e.g., 40], or open [e.g., 41] star clusters or the disks of active galactic nuclei [e.g., 42], and includes hierarchical mergers [e.g., 43]. Alternately, some binary black holes may be non-astrophysical entirely, instead forming, e.g., from primordial black holes from the early Universe [e.g., 44]. Crucially, each different origin story *predicts different spin population distributions* [e.g., 35, 45], rendering trustworthy spin models and measurements essential.

[5] Electromagnetic astronomers indirectly measure spin through X-ray emission from in-falling gas in a black hole's accretion disk, via both X-ray reflection spectroscopy and thermal continuum fitting [e.g., 30].

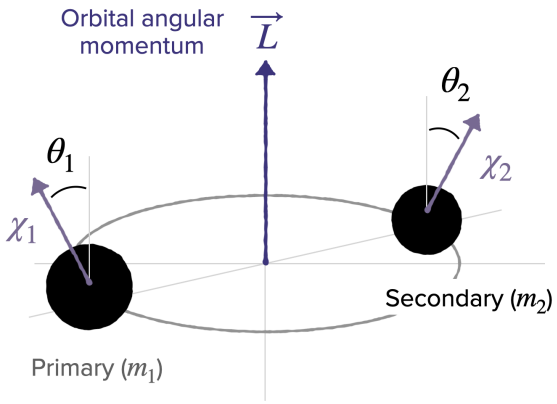


Figure 1.3: Schematic of masses m_i, spin magnitudes χ_i, and spin tilt angles θ_i of a binary black hole. A subscript of 1 indicates that the quantity describes the more-massive ("primary") black hole, while 2 corresponds to the less-massive ("secondary"). The orbital angular momentum $\vec{L}$ provides the "z-axis" for the coordinate system in which spins are defined. *Graphic by* ***SJM***.

For measurement purposes, we typically reduce a black hole's three-dimensional spin angular momentum vector $\vec{S}$ to its *dimensionless* magnitude $\chi \in [0, 1]$,[6]

$$\chi = |\vec{\chi}| \quad \text{where} \quad \vec{\chi} = \frac{\vec{S}}{m^2} \tag{1.1}$$

where a value of $\chi = 0$ intuitively represents no spin, while $\chi = 1$ is the maximal spin a black hole can support given its mass, according to the Kerr metric [46] in general relativity [Ch. 2]. We also look at the cosine of its angle $\theta \in [0, \pi]$ with the binary's orbital angular momentum $\vec{L}$ (or "tilt angle"),

$$\cos\theta = \frac{\vec{L} \cdot \vec{\chi}}{|\vec{L}||\vec{\chi}|} \ , \tag{1.2}$$

as visualized in Fig. 1.3. Each spin additionally has an azimuthal angle $\phi \in [0, 2\pi)$, defined *in* the orbital plane with respect to the line between the black holes. Spin magnitudes and angles generically evolve over a binary's lifetime, and are typically quoted at some reference time or frequency [cf., 47]. The spin azimuthal angles remain difficult to constrain in GW data [48], and aside from rare cases are not astrophysically significant [e.g., 49, 50]. Alternative spin parameterizations are introduced in Ch. 2.

The measured spins of binary black hole components near merger depend on their natal spins χ_b (i.e., spin at black hole formation or "birth," hence the

[6] Here and throughout, vector magnitudes are defined as $|\vec{x}| = \sqrt{\vec{x} \cdot \vec{x}}$. Additionally, units where $c = G = 1$ are used throughout unless otherwise specified.

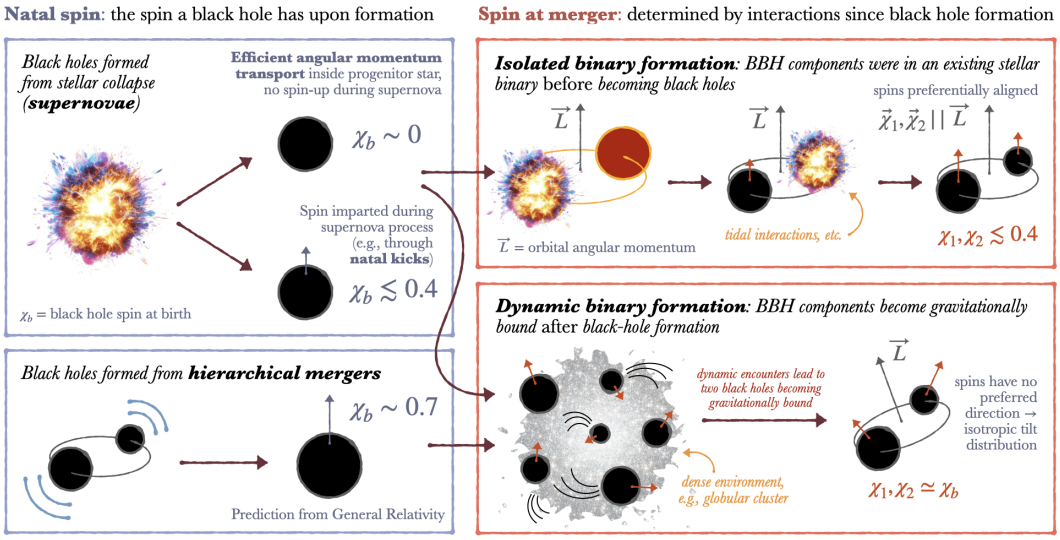


Figure 1.4: Illustration of the effect of several proposed binary black hole formation and evolution scenarios on spin magnitudes and tilts. The left-hand column compares natal spin magnitudes χ_b for black holes formed through supernovae vs. hierarchical mergers. The right-hand column shows the canonical isolated vs. dynamical formation pathways. *Graphic by* ***SJM****.*

subscript "b") and any spin-evolving interactions between birth and coalescence. Figure 1.4 visualizes the impact of astrophysical processes on binary black hole spins. If angular momentum transport inside stars is efficient, black holes born from stellar collapse are theorized to have negligible spin [31, 51]. Additionally accounting for supernovae kicks [e.g., 52] and two-body interactions like tides, accretion, or stellar winds [e.g., 53–59], stellar-collapse black holes are predicted to have spin magnitudes $\chi \lesssim 0.4$ [60]. In contrast, if the black holes we observe are instead hierarchical mergers, i.e., remnants of previous binary black hole mergers [e.g., 22, 43, 61–72], then natal spin magnitudes will be larger, with $\chi \sim 0.7$ predicted by general relativity [73] and confirmed in simulations [e.g., 64]. Binary black holes that evolve through chemically homogeneous evolution [e.g., 38, 56, 74, 75], i.e., tidally-locked tight stellar binaries that do not separate into a helium core and hydrogen envelope, may have even larger (near maximal) spin magnitudes, although these could be decreased due to magnetic braking [76].

The tilt angle θ between the spin and orbital angular momenta offers further insight into the astrophysical origins of binary black holes. A non-zero tilt leads to *precession* of the orbital plane, causing amplitude and phase modulation of the emitted gravitational-wave signal [77, 78]. In isolated binaries,

where both black holes inherit angular momentum from the same collapsing gas cloud, the spins and orbital angular momentum tend to be aligned [e.g., 79–81]. Tidal torquing and episodes of mass transfer may further align spin and orbital angular momenta [e.g., 82, 83]. Supernovae kicks may impart misalignment, but the magnitude of such kicks is debated [52, 84–90]. In comparison, dynamical formation has no *a priori* preferred spin direction, leading to an isotropic tilt distribution [39, 91–95]—with the exception that binary black holes assembled within disks of active galactic nuclei may have more finely-tuned tilts [e.g., 96, 97]. Regardless, significant spin precession is typically attributed to dynamic origins, making it an important waveform feature to robustly understand and correctly measure.

Despite their astrophysical importance, spin magnitudes and tilt angles remain poorly constrained with LVK data, especially in contrast to the other fundamental property of black holes: mass. Spin precession is particularly difficult to measure because its imprint on gravitational-wave signals is weak [98] and may mimic other physical effects like eccentricity [99, 100]. Moreover, waveform systematics, unmodeled physics, and features in detector noise all disproportionately impact spin measurements [e.g., 101–106]. Without careful safeguards, we risk spurious precession inference and false astrophysical conclusions. Any unaccounted for mismodeling, whether in single events or the population, will propagate downstream into incorrect understandings of fundamental properties of binary black holes and the universe that creates them. Additionally, the problem of spin measurement can help answer a widely-applicable, cross-disciplinary question: *how do we best extract subtle information from noisy data?*

1.3 High-mass binary black holes

Binary black holes in the high-mass tail of the population observable with the current LVK detectors, $M \sim \mathcal{O}(100\,M_\odot)$ and henceforth deemed **high-mass**,[7] are also featured as a recurrent topic in this thesis. These rare observations—making up only a few of the hundreds of signals in the current GWTC [9, 21, 107]—offer unique insight into some of the formation scenarios illustrated in Fig. 1.4. Most notable are the aforementioned hierarchical mergers [e.g., 43, 61–72], thought to primarily occur in dense, chaotic environments like clusters

[7] Although, in the grand scheme of black holes, these $100\,M_\odot$ ones are rather small. The supermassive black holes at the center of galaxies are $\mathcal{O}(10^6)\,M_\odot$.

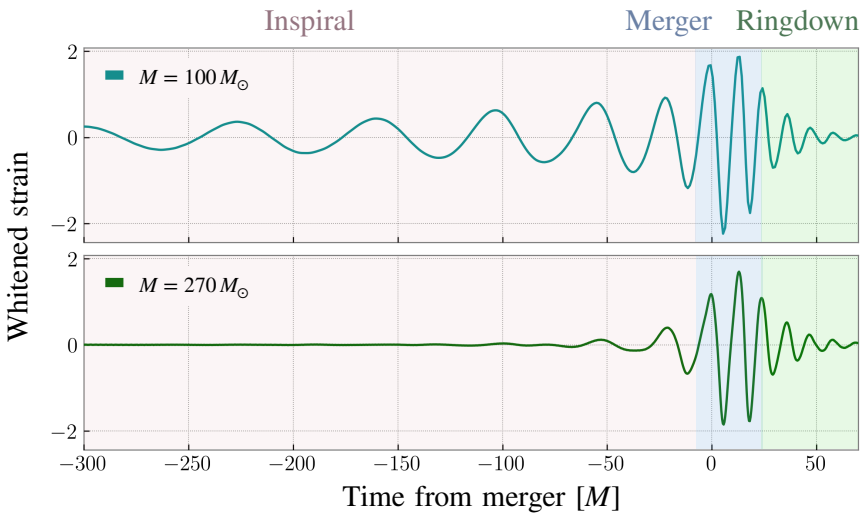


Figure 1.5: Schematic comparing time-domain gravitational-wave signals from a lighter (total mass of $100\,M_\odot$; top) versus heavier ($270\,M_\odot$; bottom) binary black hole, whitened by a representative LIGO power spectral density (see Sec. 8.2). The inspiral regime of each signal, where the black holes are well-separated, is shaded pink; the merger is shaded blue; and ringdown green. The higher-mass signal has barely any observable inspiral. *Graphic by* ***SJM****.*

[e.g. 39–41, 91, 92, 94, 108–119] or the disks of active galactic nuclei [e.g. 96, 97, 120–123]. High-mass black hole binaries also probe the pulsational pair-instability mass gap [e.g. 124–129] which promises to illuminate stellar [e.g., 31, 51] and supernovae physics [e.g., 130–134]. Moreover, as they are observable out to higher redshifts, these systems let us look further back in cosmic time [135] to better probe cosmology [e.g., 136, 137], and are even thought to be the potential "missing link" between stellar-mass and supermassive black holes [e.g., 138–140].

However, for several reasons, GW data from high-mass systems are substantially more difficult to analyze and interpret than those from lower-mass binaries. Three physical phenomena are at play: first, current ground-based detectors lose sensitivity around 20 Hz, below which noise is significantly higher [141]; second, a binary's merger frequency is inversely proportional to its total mass; and third, more-massive systems can be detected out to higher redshifts ($z \sim 1$), so the inspiral is redshifted to even lower frequencies. For instance, a binary black hole with a redshifted (or *"detector frame"*) total mass of $M = (1+z)\,M^{\rm source} = 80\,M_\odot$ merges at ~ 200 Hz, while one with $M = 250\,M_\odot$ merges at ~ 60 Hz and with $500\,M_\odot$ merges at ~ 30 Hz (see

Sec. 8.B, Tab. 8.2). Thus, the GW inspiral frequencies emitted by high-mass systems fall at or below the low edge of the LVK detectors' sensitivity bands.

Consequently, high-mass binary black holes produce only a few observable cycles (Fig. 1.5) from which we must then measure over a dozen source parameters [e.g., 142, 143]. If one of these cycles overlaps with non-Gaussian transient noise (a "glitch") [144–146], a large portion of the signal is compromised, which can drastically impact parameter estimates [105]. Moreover, the cycles which *are* observable from high-mass binaries are predominantly in the nonlinear *merger* regime, shaded blue in Fig. 1.5, where waveform models are typically least accurate, requiring the "stitching together" of regimes over which different model types are most valid [147, Fig. 1.1]. When applied to massive signals, many current data-analysis techniques and models are prone to systematic uncertainties and yield conflicting results [e.g., 21].

The difficulty measuring source properties of high-mass black hole binaries is further compounded for phenomena already hard to constrain, like individual spin components and spin-orbit precession [98]. Unlike in the inspiral (pink region of Fig. 1.5) where the black holes are well-separated and the signal can be described by analytic expansions like the post-Newtonian formalism [e.g., 148, 149], spins are not analytically tractable in the merger [150, Ch. 4]. Additionally, the precession timescale is comparable to the observable duration of these short signals themselves [78]. For many configurations, it is very difficult to constrain precession without seeing a full precession cycle [102, Ch. 8].

Despite all odds, investigations with numerical relativity and other simulated data by myself and others suggest that information about spin *does* exist in heavy binary black hole signals [e.g., 102–104, 151–157]. Even more curiously, the two most massive observed systems, GW231123 [21] and GW190521 [20, 101, Ch. 7], are both inferred to be rapidly spinning and/or precessing. GW191109 and GW200129 are other high-mass signals with unique spin configurations, yet both suffer data-quality issues: GW191109's anti-aligned spin [158, Ch. 10] and GW200129's precession [159] are degenerate with detector glitches coincident with their respective pre-merger data. In all four cases, astrophysically significant spin information originates primarily from the inspiral, which is short, quiet, and/or affected by glitches or waveform systematics, risking false astrophysical conclusions.

1.4 Overview of dissertation

In my Ph.D research, I have developed and applied a suite of **methods to produce accurate, precise, and unbiased spin inference at every stage of the LVK's data analysis pipeline**, from *individual* detections to the astrophysical *population distribution* of binary black hole mergers at large.

The analyses and results presented in the following chapters are organized around three questions central to understanding and solving challenges in binary black hole spin inference. **Part I** explores the question, *What spin parameters are measurable?* **Part II** focuses on high-mass systems and asks, *Where in gravitational-wave signals do measurements of binary black hole spin originate?* Finally, **Part III** discusses *What do we know about the spins of the binary black hole population?*

As this thesis concerns itself with measuring the spins of black holes through gravitational waves, a pedagogical overview of gravitational-wave signals from merging compact binaries is warranted. **Chapter 2** begins this discussion by describing gravitational waves sourced by binary black hole mergers and in particular, how spins show up in the equations. I describe *(i)* the post-Newtonian formalism for inspiraling two body systems and *(ii)* the quasi-normal mode spectrum for gravitational waves from black hole merger remnants, and present common mathematical parameterizations for spins that emerge from these two pictures, including the so-called "effective spin" parameters, $\chi_{\rm eff}$ and $\chi_{\rm p}$ for the former.

Next, **Chapter 3** describes the Bayesian approach used to actually extract information about the parameters described in Ch. 2 from gravitational-wave data. I present the statistical framework for single-event and population-level inference.

The remainder of Part I challenges some of the underlying assumptions put forth in Chs. 2 & 3 for individual-system measurements and the population as a whole. Spin aligned with a binary's orbital angular momentum—quantified via the $\chi_{\rm eff}$ parameter—is typically considered the best-measured spin parameter for single events and their ensemble. However, it is not always clear which spin degrees of freedom best describe binaries of different total mass, nor what is possible to probe at the population level.

To this end, in **Chapter 4** (published Ref. [150]), I present a data-driven ap-

proach to systematically map binary black hole mass-spin degeneracies and thus examine what mass-spin configurations produce similar signals, making different sources indistinguishable. The work provides limits on what combinations of mass and spin are uniquely measurable, and lays the groundwork for future work of deriving parameters that more robustly capture high-mass spin dynamics.

Chapter 5 (published as Ref. [160]) then challenges the foundational assumption in LVK population analysis that $\chi_{\rm eff}$ can be reliably decomposed into spin magnitudes and tilts. I perform end-to-end single-event and population inference on three simulated binary black hole populations with the same $\chi_{\rm eff}$ but different underlying spin magnitude and tilt distributions. While the recovered distributions are not always consistent with those injected, I find that we can distinguish between binary black hole populations with more or less extreme magnitudes and tilts at current sensitivity, even when $\chi_{\rm eff}$ distributions are identical.

Moving to Part II, I posit that a rigorous understanding of the link between binary black hole *source parameters* (masses, spins, etc.) and the *morphology* (amplitude over time) of their emitted gravitational-wave signals is crucial for robust measurements and interpretation, especially for subtle properties like spin. It is insufficient to simply state that "we measure spin;" such claims must be supported by a systematic understanding of where and in what types of signals spin is measurable. By identifying consistent spin-driven waveform features, we can determine if other phenomena, whether astrophysical or not, can mimic the same morphology. This connection is especially important for short, merger-dominated signals from high-mass binaries whose data are not well described by analytic models.

Typically, the likelihood used to infer binary parameters from gravitational-wave data is constructed in the frequency domain, as described in Ch. 3. However, when analyzing a *subset* of a signal's duration—e.g., just the inspiral (pre-merger) or ringdown (post-merger) data—crucial signal power falls at the edge of the segment and the assumptions that make frequency-domain likelihood computation possible break down. **Chapter 6** describes a remedy and the central method used in Chs. 7–11: calculating the likelihood in the *time domain*. I also discuss `tdinf`, the open-source time-domain inference Python code I developed and published (Ref. [161]).

In **Chapter 7** (published as Ref. [101]), I use time-domain inference to investigate the previously-elusive measurement of large spin precession in the high-mass signal GW190521. I localize GW190521's precession information to a short, late-inspiral segment that is suppressed relative to the loud merger, and connect this quiet pre-merger cycle to GW190521's underlying time-domain source dynamics: precessional motion causes the binary's direction of maximal GW emission to sweep away from our line of sight right before merger.

In a follow-up study presented in **Chapter 8** (published as Ref. [102]), I discover that the same morphological features do not always drive precession constraints, using a suite of simulated massive, highly-precessing signals consistent with GW190521. For precessing binaries with *identical underlying spin configurations*, whether or not precession is *observable* depends on the system's total mass and full three-dimensional angular orientation in the sky: some configurations make precession loud and clear in the inspiral or ringdown data alone, while others rely on analogous morphology to GW190521. In this chapter, I also place the first-ever constraints on spin precession in ringdown data using a full inspiral-merger-ringdown waveform model.

Furthermore, GW190521's short signal offers multiple interpretations: its source could be an intermediate-mass-ratio or eccentric binary black hole. **Chapter 9** presents in-progress work using time-domain inference to untangle these alternative scenarios, with preliminary findings indicating that GW190521's time-evolution aligns most closely with the precession hypothesis.

Transient non-Gaussian detector noise known as a "glitch" can non-trivially impact the estimated parameters of gravitational-wave sources when overlapping a signal. In **Chapter 10** (published in part as Ref. [158]), I use time-domain inference to link the intriguing inference of spin pointing below the orbital plane in the signal GW191109 to a region of data concurrent with a scattered-light glitch, drawing into question whether this spin measurement is astrophysical or caused by the glitch.

Chapter 11 (published in part as Ref. [18]) presents my continuation to high-profile work using time-domain inference in another context: testing fundamental physics. I was a key analyst in the first-ever observational confirmation of Hawking's Area Law, which states that the black hole horizon area—a simple function of mass and spin—cannot decrease over time.

Throughout my Ph.D, I have played a key role in spin population analyses within the LVK collaboration, described in Part III. My work began with Ref. [162], where I proposed the simple truncated Gaussian model now used as the basis for all $\chi_{\rm eff}$ population analyses produced by the LVK.[8] In my foundational study, I discovered that the $\chi_{\rm eff}$ distribution is narrow and centered near zero—a conclusion in tension with BHs observed via X-rays [e.g., 163]—but that has persisted even as more subtle structure has become apparent with the advent of new gravitational-wave data. In Ref. [162], I additionally explored the implications of ensemble properties on individual spin measurements. By re-analyzing the ten events in GWTC-1,[9] I showed that without population-informed single-event inference we consistently overestimate $\chi_{\rm eff}$.

Chapter 12 (published as [164]) presents subsequent work tackling a hotly debated yet astrophysically imperative question: *what fraction of binary black holes have zero spin?* I find that the majority of BHs in binaries have non-negligible spin, in stark contrast to previous results [165], which we showed to contain a technical error.

Due to my expertise, I was invited to serve on the writing team and as lead analyst for the spin results section of the LVK's paper reporting astrophysical population measurements through the O4a observing run (GWTC-4.0), Ref. [23]. My contributions to Ref. [23] are presented in **Chapter 13**. I report that the binary black hole population has predominantly small spin magnitudes and a wide range of tilt angles (i.e., precession exists in the population), and conclude that it likely originates from a superposition of formation channels, e.g., those in Fig. 1.4. In Ch. 13, I also discuss technical details related to the non-trivial interplay between our method of choice for mitigating the impact of Monte Carlo approximation in the hierarchical likelihood and the resultant inferred effective-spin population.

My most recent research, presented in **Chapter 14** (published as Ref. [166]), centers tests of model fit and bias in binary black hole population inference. Large single-event measurement uncertainty—as in the case for spins—allows different population models to yield disparate astrophysical conclusions [Ch. 5].

[8] Ref. [162] is not included as a separate chapter in this dissertation, as it was published prior to my being a graduate student. I summarize its contents here in the introduction to provide important context for my graduate contributions.

[9] It is incredible to note that within the six years I have been publishing papers about gravitational-wave astrophysics, our catalog size has increased by over a factor of ten.

Moreover, standard goodness-of-fit tests called "posterior predictive tests" fail to identify mismatch between inferred and true populations when single-event uncertainty is high, posing a significant problem: with existing and commonly-utilized statistical tools, it may be impossible to discern whether claims drawn from the binary black hole spin population reflect real information in the data or just arbitrary model choices. In Ch. 14, I rigorously discuss canonical and alternative posterior predictive checks, describe where they go wrong, and suggest more robust model checking procedures for use in gravitational-wave population inference.

Chapter 15 presents concluding thoughts and plans for future work. Finally, in the **Pocket Material** (i.e., following the bibliography) I briefly present preliminary results from time-domain analysis of GW231123, the most massive binary black hole detected thus far. I also show an infographic that I made about GW231123 as a part of LVK science communication and outreach efforts. The infographic has been translated into over ten languages and has been featured in the press worldwide, warranting inclusion in this thesis as one of my important contributions to the field of gravitational-wave science.

$\sim\sim\sim$

A note on repeated information: Chapters 4, 5, 7, 8, 10, 12, 13, & 14 are self-contained papers (or parts of self-contained papers) and accordingly contain their own tailored introductions. Information may be repeated between these and the general thesis background.

A note on units: Throughout this dissertation, I use units where the gravitational constant G and the speed of light c are all equal to unity, i.e., $G = c = 1$, as is common to general-relativity-based analyses. In this system, distance, time, mass, energy, and momentum all have the same units. Unless otherwise specified, factors of G and c are not explicitly written in formulae.

Part I

What spin parameters are measurable?

Chapter 2

THE IMPRINT OF BLACK HOLE SPIN ON A GRAVITATIONAL-WAVE SIGNAL

The central aim of this thesis is the robust measurement of black hole (BH) spins using gravitational-wave (GW) data. In this and the following introductory chapters, I lay some groundwork necessary to pursue this aim. To begin, Secs. 2.1 & 2.2 present a pedagogic overview of GWs sourced by binary black hole (BBH) mergers. Following this, I describe how spin enters under various analytic descriptions of GW signals. In Sec. 2.3, I introduce the post-Newtonian (PN) framework for describing inspiraling two-body systems. I then discuss how the PN formalism motivates the so-called "effective spin" parameter $\chi_{\rm eff}$ in Sec. 2.4, and describe various parameters that capture the effects of spin-induced orbital precession. Finally, in Sec. 2.5, I describe how spins imprint on the post-merger "ringdown" signal emitted by the system's perturbed remnant BH.

Note: *With the goal of keeping this introduction (relatively) concise and out of the weeds, I will refrain from providing detailed derivations of any formulae, and will assume that readers are familiar with general relativity (GR)—specifically tensor notation, spacetime metrics, and gauge symmetry. I refer those seeking more detail to my Bachelor's thesis (Ref. [167]), containing fifty detailed pages of GR derivations related to GWs and spinning BHs.*

Throughout the years, I have learned about GR from Maggiore [168, 169], Hartle [170], Moore [171], Carroll [172], Thorne & Blandford [173], and Misner, Thorne, & Wheeler [174]. Section 2.2 draws upon all of these textbooks, as well as notes I prepared for the Cosmology & Extragalactic Astrophysics course I taught at Pomona College in Spring 2026.

2.1 Newtonian versus Einsteinian formulations of gravity

The story many of us are told from our earliest days in a physics classroom is that "gravity is a force." In this so-called Newtonian formulation, mass generates a gravitational force, and that force determines how objects accelerate. We first learn that this gravitational force F felt by an object with mass m

is equivalent to "$F = mg$," and may measure the time it takes objects to fall to the ground from different heights to empirically construct the acceleration g at the surface of the Earth. Later on, we are introduced to Newton's Law of Universal Gravitation, perhaps applying it to describe the orbital motion of planets around the Sun and so on. Mathematically, Newtonian gravity can be summarized with the Poisson equation, i.e., at every point $\vec{x}$ in space the *gravitational potential* Φ obeys

$$\nabla^2\Phi(\vec{x}) = 4\pi G\rho(\vec{x}) \ , \tag{2.1}$$

where $G = 6.67 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ is the gravitational constant and ρ is the mass density.

The theory of general relativity (GR)—thus far empirically unimpeachable—instead posits that *gravity is the geometry of spacetime,* with Newton's theory only offering an approximation of gravity's true, underlying behavior. In contrast to the Newtonian formalism, GR (or, gravity's *Einsteinian formalism*) instead tells us that mass and energy curve spacetime and this warped spacetime sets the trajectories along which bodies move. Mathematically, we represent spacetime with a four-dimensional *metric tensor*, $g_{\mu\nu}$. The analog to Eq. (2.1) in GR is the Einstein field equation:

$$G_{\mu\nu} = 8\pi\, G\, T_{\mu\nu} \ , \tag{2.2}$$

where, conceptually, the mass density ρ is replaced by the *stress-energy tensor* $T_{\mu\nu}$ and the gravitational potential Φ is replaced by the *Einstein tensor* $G_{\mu\nu}$ containing information about space-time curvature.[1] Equation (2.2) actually represents 10 coupled nonlinear partial differential equations, as $G_{\mu\nu}$ and $T_{\mu\nu}$ are four-by-four symmetric tensors, and can in general not be solved analytically.[2]

[1] In full generality, Eq. (2.2) reads $G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi\, G\, T_{\mu\nu}$, where Λ accounts for the expansion or contraction of the universe. Observational evidence [e.g., 25, 175–179] make clear that our universe is expanding at an accelerating rate, meaning $\Lambda > 0$. However, the impact of Λ is negligible on sub-cosmological scales (i.e., $\lesssim 100\,\mathrm{Mpc}$) and therefore can be ignored when calculating GW morphology in, e.g., Eq. (2.5). We instead account for cosmological effects through the redshift z, as discussed later.

[2] Some important exceptions where Einstein's equation *can* be solved analytically are Schwarzschild spacetime (spherical symmetry), Kerr spacetime (azimuthal symmetry), and Friedmann-Robertson-Walker spacetime (homogeneity and isotropy).

2.2 Gravitational waves 101

Gravitational waves (GWs) are a natural consequence of the relativistic picture of gravity: special relativity puts a limit of $c = 3 \times 10^8 \, \frac{\mathrm{m}}{\mathrm{s}}$ on the speed at which any information can travel, and gravitational interactions are not exempt. Consider a region of spacetime where the gravitational field is weak, e.g., far away from any significant mass or energy density. In this *weak field limit*, we can write the metric as

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{2.3}$$

where $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$ is the Minkowski metric describing flat spacetime, and $h_{\mu\nu}$ is a small symmetric *perturbation*, where "small" means $|h_{\mu\nu}| \ll 1$. Our goal is to solve the Einstein equation (Eq. 2.2) in vacuum ($T_{\mu\nu} = 0$) for this perturbation $h_{\mu\nu}$. Doing so involves the repeated invocation of gauge freedom: the ability to map the same physical spacetime to different coordinate systems. I provide an overview of this derivation below, highlighting select key equations along the way.

The transverse traceless gauge

It is useful to express the metric perturbation $h_{\mu\nu}$ as

$$h_{\mu\nu} = H_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}H \tag{2.4}$$

where $H_{\mu\nu}$ is the *trace-reversed* form of $h_{\mu\nu}$.[3] Combining Eqs. (2.3) & (2.4) with the Einstein equation in vacuum, one can derive the following linearized ($h \ll 1$) differential equation:

$$\frac{1}{2}\left[\partial^\mu\partial_\mu H^{\gamma\sigma} - \partial^\gamma\partial_\mu H^{\mu\sigma} - \partial^\sigma\partial_\mu H^{\mu\gamma} + \eta^{\gamma\sigma}\partial_\beta\partial_\mu H^{\mu\beta}\right] = 0. \tag{2.5}$$

Moreover, we can always find a gauge (called the Lorentz gauge) such that $\partial_\mu H^{\mu\nu} = 0$, under which Eq. (2.5) simplifies to the wave equation:

$$\partial^\alpha\partial_\alpha H^{\mu\nu} = \left(-\frac{\partial^2}{\partial t^2} + \nabla^2\right) H^{\mu\nu} = 0\,, \tag{2.6}$$

with the general plane wave solution,

$$H^{\mu\nu} = A^{\mu\nu} \, \exp[ik_\alpha x^\alpha] \tag{2.7}$$

[3] The trace of some tensor $x_{\mu\nu}$ is defined as $x = \eta^{\mu\nu} \, x_{\mu\nu}$. Here, $H = -h$.

at a spacetime location $x^\alpha = (t, x, y, z)$, composed of an amplitude tensor $A^{\mu\nu}$ and wave four-vector k^α. One can show that k^α must be *null*, implying that GWs propagate at the speed of light.[4]

Exploiting further gauge freedom, it is typical to transform to the *transverse traceless* ("TT") gauge—where any time components of $A^{\mu\nu}$ are zero, as is the trace of $A^{\mu\nu}$—and work in a frame where the GW is propagating along the $\hat{z}$-axis, i.e., $k^\alpha = (\omega, 0, 0, \omega)$ for some frequency ω. Notably, in the TT gauge, $H^{TT}_{\mu\nu} = h^{TT}_{\mu\nu}$, as each is traceless: $H^{TT} = h^{TT} = 0$ (see Eq. 2.4). In such a coordinate system, we can succinctly express the GW strain $h^{TT}_{\mu\nu}$ through its *plus and cross polarizations* [e.g., 180, 181],

$$h^{TT}_{\mu\nu} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & h_+ & h_\times & 0 \\ 0 & h_\times & -h_+ & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = h_+ \hat{e}^+_{\mu\nu} + h_\times \hat{e}^\times_{\mu\nu} \tag{2.8}$$

where $\hat{e}^{+,\times}_{\mu\nu}$ are the polarization basis tensors and

$$h_{+,\times} = A_{+,\times} \cos(\omega t + \phi_{+,\times}). \tag{2.9}$$

Figure 2.1 demonstrates the impact of these polarizations on a ring of particles—and accordingly helps visualize the origin of their "plus" and "cross" namesakes.

Sourcing gravitational waves

We now have the general solution for $h_{\mu\nu}$ involving some unknowns $A_{+,\times}$ and $\phi_{+,\times}$. The specific forms of $A_{+,\times}$ and $\phi_{+,\times}$ depend on the source generating the GWs. Their calculation requires looking at the general linearized Einstein equation,

$$\partial^\alpha \partial_\alpha H^{\mu\nu} = 16\pi\, T^{\mu\nu}\,, \tag{2.10}$$

for a non-zero stress-energy tensor $T^{\mu\nu}$. For a nearly-Newtonian (slow-moving) source with a spatial extent much smaller than the luminosity distance d_L [183] to the observer, the perturbation in the TT gauge takes the form

$$h^{TT}_{jk} = \frac{2}{d_L}\, \ddot{I}^{TT}_{jk}(t - r) \tag{2.11}$$

[4] The GR prediction that GWs propagate at the speed of light has been observationally supported at high precision by the binary neutron star merger GW170817 [10].

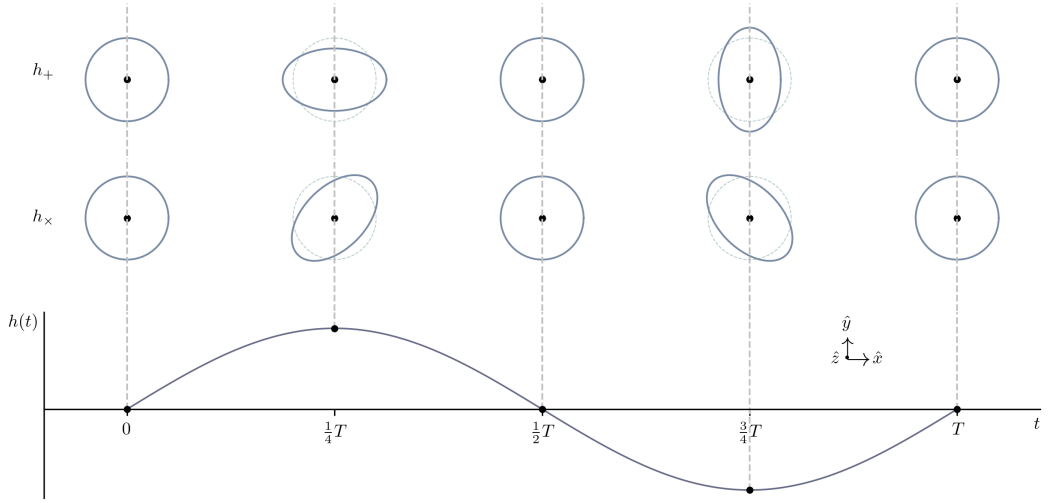


Figure 2.1: The effect of the plus (*top*) and cross (*bottom*) polarizations of a GW with period T propagating in the $\hat{z}$ direction, passing through a circular ring of particles in the xy-plane. *Graphic adapted by* ***SJM*** *from Ref. [182].*

where I_{jk}^{TT} is the transverse traceless reduced quadrupole moment and t is the time at observation. The jk indices indicate a purely spatial (three-by-three) tensor, as the TT projection has already removed the time components. Equation (2.11) is known as the *quadrupole formula* and is the leading-order approximation for a generalized h_{jk}^{TT}.

Gravitational waves from an orbiting binary

If the GW source is an orbiting binary with a total mass $M = m_1 + m_2$, reduced mass $\mu = \frac{m_1 m_2}{M}$, and angular velocity Ω, then one can calculate I_{jk}^{TT} explicitly and show that

$$h_+ = \frac{-2\,\mu(M\,\Omega)^{2/3}}{d_L}\left(1+\cos^2\iota\right)\cos\left[2(\Omega\,t-\varphi_0)\right] \text{ and} \tag{2.12}$$

$$h_\times = \frac{-4\,\mu(M\,\Omega)^{2/3}}{d_L}\cos\iota\,\sin\left[2(\Omega\,t-\varphi_0)\right]\,, \tag{2.13}$$

where ι is the orientation of the binary's orbital angular momentum with respect to the observer and φ_0 is the orbital phase at $t = 0$.

The leading-order expressions for the strain amplitudes h_+ and $h_\times$ in Eqs. (2.12) & (2.13) already illuminate a good deal about the GWs we now routinely observe from merging BBHs with the LVK detectors, and provide intuition about how various source parameters imprint on the data. First, we can see that the strain h depends on the binary's mass, and decreases with the luminosity dis-

tance d_L to the source $\propto 1/d_L$. Second, the relative amplitudes of the two GW polarizations depend on the orientation ι of the binary with respect to the observer (see Fig. 8.6 for a visualization). Third, the GW frequency is twice the orbital frequency Ω.

With some additional algebraic manipulation, we can further show that as the binary loses energy to GW emission, its orbital frequency Ω evolves, following

$$\Omega(t) = \left[\frac{256}{5}\,\mu\,M^{2/3}\,(t_0 - t)\right]^{-3/8}, \tag{2.14}$$

where t_0 is the time at which $\Omega \to \infty$, a proxy for when the separation between the two orbiting objects becomes zero, i.e., they merge. The best-measured source property for compact binaries directly falls out of Eq. (2.14): the so-called *chirp mass*, $\mathcal{M} = \mu^{3/5} M^{2/5}$.

Higher-order multipolar modes

In full generality, a GW signal can be represented as a complex time series which can be decomposed into the sum of spin -2 weighted spherical harmonics $Y^{(-2)}_{\ell m}$ which depend on the inclination angle ι and some reference phase φ and are labeled by polar and azimuthal indices ℓ and m [148, 184–186]:

$$h = h_+ - i h_\times = \frac{M}{d_L} \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} Y^{(-2)}_{\ell m}(\iota, \varphi)\, h_{\ell m}(\lambda, t)\,. \tag{2.15}$$

The modes $h_{\ell m}$ are each a function of time t and the BBH's intrinsic parameters λ (masses and spins), and can be expressed in terms of their amplitude $A_{\ell m}$ and phase $\Phi_{\ell m}$ evolution

$$h_{\ell m}(\lambda, t) = A_{\ell m}(\lambda, t) e^{-i\Phi_{\ell m}(\lambda, t)}\,, \tag{2.16}$$

which are in turn approximated by post-Newtonian theory in the inspiral, as described in the subsequent section.

2.3 Post-Newtonian expansions

The quadrupole formula (Eq. 2.11) is a leading-order description of GWs sourced by the inspiral phase of BBHs, and offers insight into how GWs encode information about their astrophysical sources. However, to *accurately* estimate GW source properties from real data, we must move substantially beyond this approximation. For example, Eqs. (2.12) & (2.13) entirely ignore spin effects

from the orbiting bodies, among other simplifications. In this section, I introduce post-Newtonian (PN) expansions—one possible formalism for describing GW signals from inspiraling binaries beyond leading order—and examine how they illuminate the imprint of BH spin on GWs.

The linearized gravity equation (Eq. 2.10) is a universal wave equation for $h_{\mu\nu}$ under the assumption that $|h_{\mu\nu}| \ll 1$. The distinction between the Newtonian and PN approximations then lies in how we calculate $T_{\mu\nu}$ and thus solve Eq. (2.10) to obtain $h_{\mu\nu}$. The quadrupole formula is derived by calculating the binary's orbital motion (and thus $T_{\mu\nu}$) to lowest order, while PN continues at higher orders.

Specifically, in the case of compact binaries, the PN formalism is an expansion in orders of the speed of the orbiting bodies with respect to the speed of light, v/c.[5] PN theory is used to obtain an analytic expression for the GW *phase* Ψ as a function of time or frequency under the assumptions that *(i)* $v/c \ll 1$, *(ii)* the binary components are far away from each another, and *(iii)* the GW amplitude evolves much slower than its phase (stationary phase approximation). In the frequency domain [187],

$$h \propto \frac{1}{r}\frac{\mathcal{M}^{5/6}}{f^{7/6}}\exp[i\Psi(f)]\ , \tag{2.17}$$

following the notation of Sec. 2.2. Somewhat counterintuitively, the "n^{th} PN order" adds a correction of order $(v/c)^{2n}$ to the expression for Ψ. For reference, Newtonian gravity is the 0^{th} PN order, perihelion precession enters at 1PN, and radiation reaction at 2.5PN. For non-spinning binaries, the GW flux and phase evolution have been calculated up to the 4.5PN order. Spin effects have been incorporated to 3.5PN in commonly-used waveform models, with recent analytic results extending to 4PN and beyond; see e.g., Refs. [14, 188–190] and the references therein .

In the case of a binary comprised of *non-spinning* BHs, the first several terms of the PN expansion can be written [14]

$$\Psi^{\text{no spin}}(f) = \Psi_0 + \frac{3}{128\,\eta\,v^5}\Big[\,1 + C_2(\eta)\,v^2 - 16\pi\,v^3 + C_4(\eta)\,v^4 + \ \ldots\,\Big] \tag{2.18}$$

where $\eta = m_1 m_2/M^2$ is the symmetric mass ratio, C_i are mass-dependent constants, $v = (\pi M f)^{1/3}$ is the characteristic orbital velocity (written here

[5] For a review of the PN formalism, see Ref. [148].

implicitly in units of the speed of light, i.e., $c = 1$), and

$$\Psi_0 = 2\pi f t_c - \phi_c - \pi/4$$

for a time t_c and phase ϕ_c at coalescence. With *spinning* BHs, additional terms are added [188], such that

$$\Psi^{\rm spin}(f) = \Psi^{\rm no\,spin} + \frac{3}{128\,\eta\,v^5}\left[\,4\beta\,v^3 - 10\,\sigma\,v^4 + \,\ldots\,\right] . \tag{2.19}$$

Spin-orbit effects first enter at the 1.5PN (v^3) order through the β parameter, while *spin-spin effects* show up at 2PN (v^4) through σ. Contrasting Eq. (2.19) with Eq. (2.18), we can immediately see why spins are harder to measure in GW data than masses: through η, mass appears at the leading PN order, while spins do not begin to appear until multiple orders higher, corresponding to an orders-of-magnitude weaker imprint on the data.

In full generality, a BH has a three-dimensional spin vector pointing along its spin axis with a magnitude equal to its intrinsic angular momentum. Such a BH is described by the Kerr metric [46], the most general axisymmetric spacetime.[6] Previously in Sec. 1.2, I presented the dimensionless magnitude (χ_i; Eq. 1.1) and cosine tilt angle with respect to the binary's orbital angular momentum ($\cos\theta_i$; Eq. 1.2) as astrophysically meaningful expressions of BBH spins (see Fig. 1.3). However, looking at the PN equations, we can intuit that these are *not* the best-measured spin-related quantities—at least not for inspiral-dominated signals. BH spin magnitudes and tilts never appear *directly* in Eq. (2.19), instead entering via the convoluted β and σ parameters, which are respectively equivalent to[7]

$$\begin{aligned}\beta =& \frac{1}{2}\left(\frac{113}{12} - \frac{19}{3}\eta\right)\left(\chi_1\cos\theta_1 + \chi_2\cos\theta_2\right) \\ &+ \frac{113}{24}\delta\left(\chi_1\cos\theta_1 - \chi_2\cos\theta_2\right)\end{aligned} \tag{2.20}$$

[6] More precisely, the Kerr metric describes the most general axisymmetric spacetime *in the astrophysically relevant regime where BH electric charge is negligible.*

[7] The expressions in Eqs. (2.20) & (2.21) are not given directly in any paper, and were instead derived by me based off of alternative expressions in Ref. [188].

where $\delta = (m_1 - m_2)/M$, and

$$\begin{aligned}\sigma =& \frac{1}{48}\eta \left[721\,\chi_1\chi_2 \cos\theta_1 \cos\theta_2 + 247\,\vec{\chi}_1 \cdot \vec{\chi}_2 \right] \\ &+ \frac{1}{192}(1-2\eta) \left[719\left(\chi_1^2 \cos^2\theta_1 + \chi_2^2 \cos^2\theta_2\right) - 233\left(\chi_1^2 + \chi_2^2\right) \right] \\ &+ \frac{1}{192}\delta \left[719\left(\chi_1^2 \cos^2\theta_1 - \chi_2^2 \cos^2\theta_2\right) - 233\left(\chi_1^2 - \chi_2^2\right) \right] .\end{aligned} \tag{2.21}$$

Eq. (2.20) and especially Eq. (2.21) are not immediately intuitive. Looking at β, we see that the lowest-order spin contribution to the PN expansion involves both the sum and difference between each spin vector projected onto the orbital angular momentum, i.e., $\chi_i \cos\theta_i$, each weighted by a different combination of the component masses, η and δ. For reference, the *full* 1.5PN coefficient, $C_3 = 4\beta - 16\pi$, is plotted as a function of η and $\chi_1 \cos\theta_1$ in the right panel of Fig. 2.2. The 2PN parameter σ involves various quadratic combinations of χ_i and $\cos\theta_i$, also weighted by η and δ. In the subsequent Sec. 2.4, I present more readily interpretable spin parameters derived from the above PN expressions and other considerations.

2.4 Effective inspiral spin parameters

The two main impacts of spin on a GW inspiral can be broken down into those stemming from the components of spin *aligned* with the BBH's orbital angular momentum, and those *perpendicular* to it, the latter of which induces *precession* of the orbital plane. The aligned and precessing (perpendicular/misaligned) degrees of freedom are described below in turn.

Aligned degrees of freedom

The leading order β parameter encoding spin in BBH inspirals can be written more succinctly in terms of the quantities χ_+ and χ_-, respectively defined as the sum of and difference between the components of the primary and secondary BHs' spins aligned with the orbital angular momentum:

$$\chi_\pm = \frac{1}{2}\Big(\chi_1 \cos\theta_1 \pm \chi_2 \cos\theta_2\Big) . \tag{2.22}$$

Starting from Eq. (2.20), we can write

$$\beta = \frac{113 - 76\,\eta}{12}\,\chi_+ + \frac{113\,\delta}{12}\,\chi_- \tag{2.23}$$

which, after some algebra, can be further expanded to [13, 187, 191, 192]

$$\beta = \frac{113 - 76\,\eta}{12}\Big(\chi_+ + \delta\chi_-\Big) + \frac{76\,\delta\,\eta}{12}\chi_- . \tag{2.24}$$

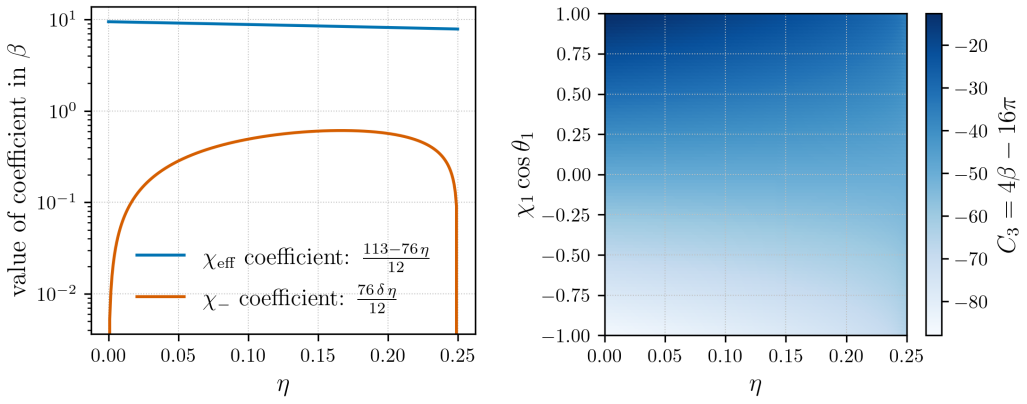


Figure 2.2: *Left:* Comparison of the coefficients of $\chi_{\rm eff}$ (blue) and χ_- (orange) in Eq. (2.24) as a function of symmetric mass ratio η. The $\chi_{\rm eff}$ coefficient is always at least an order of magnitude larger than that for χ_-. *Right:* The full 1.5PN coefficient $C_3 = 4\beta - 16\pi$ as a function of η and $\chi_1 \cos\theta_1$. Here, $\chi_2 = 0$. *Graphic by* ***SJM****.*

From Eq. (2.24) emerges the *effective inspiral spin* parameter, $\chi_{\rm eff}$ [193, 194]:

$$\chi_{\rm eff} = \chi_+ + \delta\chi_- = \frac{m_1\,\chi_1\,\cos\theta_1 + m_2\,\chi_2\,\cos\theta_2}{m_1 + m_2}\,, \tag{2.25}$$

the mass-weighted projection of each BH spin onto the orbital angular momentum $\vec{L}$—a far more intuitive quantity than β alone.

The $\chi_{\rm eff}$ parameter encompasses the leading-order effect of spin on GWs from an inspiraling BBH. This can be seen by comparing the mass-dependent coefficients for $\chi_{\rm eff}$ and χ_- in Eq. (2.24).[8] The symmetric mass ratio η has a maximal value of $1/4$ for an equal-mass binary and approaches 0 as the difference between masses grows. It follows directly from the definitions of η and δ that $\delta = \sqrt{1-4\eta}$. Then, the $\chi_{\rm eff}$ coefficient $\left(\frac{113-76\,\eta}{12}\right)$ will always be larger than the χ_- coefficient $\left(\frac{76\,\delta\,\eta}{12}\right)$—by at least an order of magnitude, as visualized in the left panel of Fig. 2.2. Consequently, $\chi_{\rm eff}$ is typically the best-measured spin parameter in GW data [98].

Armed with the knowledge of *which* spin degrees of freedom most impact the GW inspiral phase, the next logical question to ask is *how* do they impact this phasing, i.e., what is the *morphological* imprint of $\chi_{\rm eff}$ on an actual GW signal? Figure 2.3 displays waveforms sourced by BBHs with spins fully parallel (or anti-parallel) to $\vec{L}$, i.e., $\theta_1 = 0$ or π. The systems have

[8] Themselves, $\chi_{\rm eff}$ and χ_- are each bounded on $[-1, 1]$, so to look at their effect we must only compare the relative coefficients.

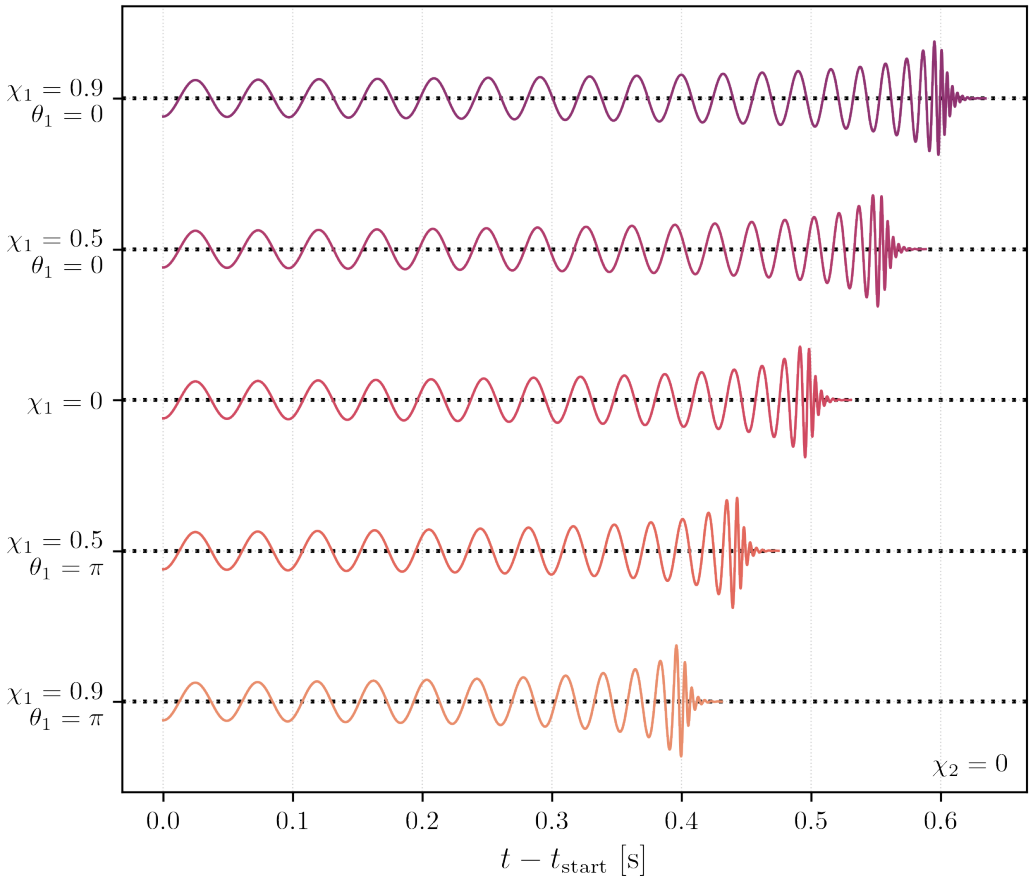


Figure 2.3: GW signals from BBH mergers with spin fully aligned with $\vec{L}$. When $\theta = 0$ ($\theta = \pi$), the primary spin $\vec{\chi}_1$ is parallel (antiparallel) to $\vec{L}$. From top to bottom, the waveforms are sourced by systems with large aligned spin to no spin to large anti-aligned spin. All other parameters are the same between each signal; the source BBHs are equal-mass with $m_1 = m_2 = 40\,M_\odot$ (in the detector frame), have non-spinning secondaries, $\chi_2 = 0$, and an inclination angle of $\iota = 0$. The start time $t_{\rm start}$ is when frequency of the dominant $\ell = 2, m = 2$ mode is 20 Hz. Waveforms are generated with the IMRPhenomT model [195]. *Graphic by* ***SJM****.*

$\chi_1 \cos\theta_1 = [0.9, 0.5, 0, -0.5, -0.9]$ (from top to bottom) and $\chi_2 = 0$. As all waveforms shown in Fig. 2.3 begin at the same GW frequency of $f = 20\,\text{Hz}$, we can see that aligned spin primarily changes how long the BBH orbits before merging. Equivalently, *aligned spin changes the GW merger frequency*: when spins are aligned (anti-aligned) with the orbital angular momentum, the BBH merges at a higher (lower) frequency.

To understand why (physically) the aligned-spin degrees of freedom primarily impact the merger frequency, we look to the effect of spin on orbital dynamics.

Systems where the spins and orbital angular momentum are aligned have more total angular momentum than those with no spin, which in turn have more than those with anti-aligned spin. The binary must radiate away that angular momentum in GWs before the BHs can merge into a remnant BH with $\chi \leq 1$, leading to what is known as the *orbital hangup effect*: when the spins of the BH and the orbital angular momentum point in the same direction, spin-orbit coupling acts like a "gravitational brake" causing the BBH to complete more orbits before merging [196, 197].

One way to quantify the impact of orbital hangup is to use the Kerr metric to calculate the innermost stable circular orbit (ISCO) for objects orbiting a spinning BH. The radius at ISCO is a function of the BH's mass m and spin magnitude χ,

$$\begin{aligned} r_{\rm ISCO} &= m\Big(3 + A_2 \pm \sqrt{(3 - A_1)(3 + A_1 + 2A_2)}\Big) \\ \text{where } A_1 &= 1 + \left(1 - \chi^2\right)^{1/3}\Big[\left(1 + \chi\right)^{1/3} + \left(1 - \chi\right)^{1/3}\Big] \\ \text{and } A_2 &= \sqrt{3\chi^2 + {A_1}^2}\ , \end{aligned} \tag{2.26}$$

and is plotted in Fig. 2.4. The "−" solutions are for orbits moving *with* the BH's spin, while the "+" solutions are for orbits moving against.[9] For spin aligned with the orbit, the ISCO shrinks relative to the non-spinning case. Orbiting objects can therefore approach more closely before plunging into the BH, leading to higher merger frequencies. Conversely, for anti-aligned spin, the ISCO expands and so the merger happens at a larger separation and corresponding lower merger frequency.

Precessing degrees of freedom

All preceding discussion has been restricted to the scenario in which BBH spins $\vec{\chi}_1, \vec{\chi}_2$ are parallel or anti-parallel to the binary's orbital angular momentum $\vec{L}$. Under these aligned/anti-aligned configurations, the system's orbital motion occurs in a fixed two-dimensional plane defined by $\vec{L}$. However, generically, BH spins are *not* aligned with the orbital angular momentum. Generic spins have an arbitrary angle θ with respect to $\vec{L}$, causing $\vec{L}, \vec{\chi}_1, \vec{\chi}_2$ to each precess

[9] *A note on terminology:* The phrases "aligned spin," "prograde orbit," and "moving with the BH's spin" are all synonymous and used interchangeably, as are "anti-aligned," "retrograde orbit," and "moving against the BH's spin." The former involves spins pointing *above* the orbital plane, yielding $\chi_{\rm eff} > 0$, while the latter refers to spins pointing *below* the orbital plane, with $\chi_{\rm eff} < 0$.

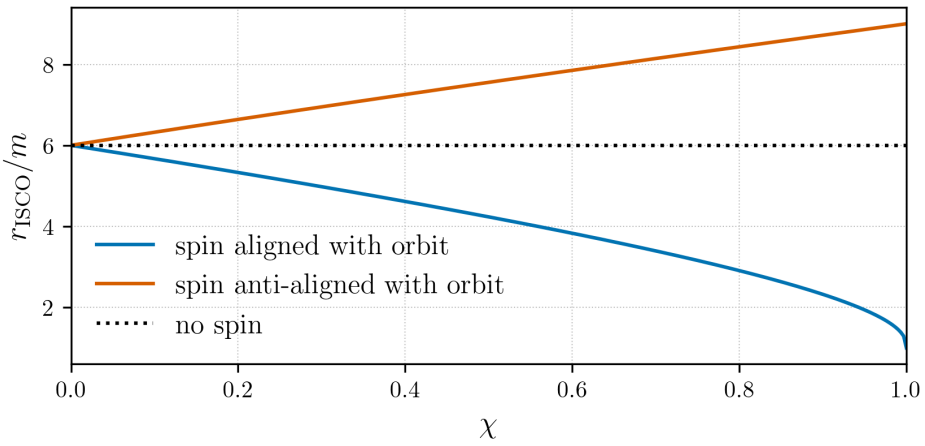


Figure 2.4: ISCO radius $r_{\rm ISCO}$ versus spin magnitude χ for a BH with mass m. In the case that $\chi = 0$, $r_{\rm ISCO} = 6\,m$. At $\chi = 1$, the inner (prograde) ISCO has $r = m$ while the outer (retrograde) ISCO has $r = 9\,m$. *Graphic by* ***SJM****.*

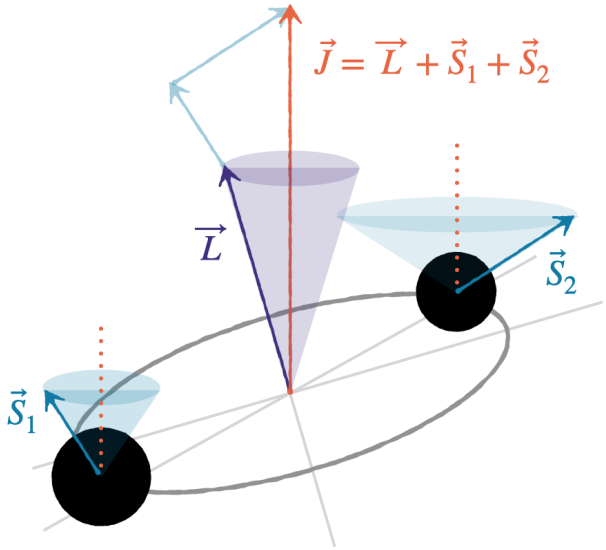


Figure 2.5: Spin ($\vec{S}_i = \vec{\chi}_i\, m_i^2$, blue) and orbital ($\vec{L}$, purple) angular momenta sum to the total angular momentum ($\vec{J}$, red). As visualized with the shaded cones, $\vec{S}_i$ and $\vec{L}$ each precess around $\vec{J}$, modulating the amplitude and phase of maximal GW emission (cf., Fig. 2.6) as the orbital plane sweeps towards and away from our line of sight. *Graphic by* ***SJM****.*

around the *total* angular momentum $\vec{J}$, as visualized in Fig. 2.5. The vector $\vec{J}$ is approximately conserved on timescales short compared to that of radiation reaction. Physically, this leads to precession of the orbital plane itself, inducing amplitude, phase, and polarization modulations in the emitted GWs [77, 198–200] and contributions from higher-order GW harmonics [e.g., 201].

Figure 2.6 shows examples of GW signals from BBHs with varying degrees of misalignment between the primary spin $\vec{\chi}_1$ and $\vec{L}$. Two features immediately jump out when comparing the precessing vs. non-precessing waveforms. First, as well as introducing phase modulation (i.e., time of merger changes), the precessing degrees of freedom drastically impact the over-all amplitude envelope of the GW signal, which aligned degrees of freedom do not (Fig. 2.3). Second, the timescale over which precession occurs is much longer than the orbital timescale but much shorter than the radiation reaction timescale over which the BBH orbit shrinks as energy is lost to GW emission [78].

Spin-orbit coupling causes precessional PN effects on top of the expansions to the GW phase described previously in Sec. 2.3. The lowest-order *orbit-averaged* PN precession equation for the evolution of $\vec{L}$ is [78, 199]

$$\begin{aligned} \frac{d\vec{L}}{dt} &= \frac{d\hat{L}}{dt}L + \frac{dL}{dt}\hat{L} \\ &= (\vec{\Omega}_L \times \vec{L}) + \frac{dL}{dt}\hat{L} \end{aligned} \tag{2.27}$$

where the first term describes orbital plane precession (with a precession frequency $\vec{\Omega}_L$) and the second radiation reaction. The first term, $\vec{\Omega}_L \times \vec{L}$, can be split into spin-orbit and spin-spin contributions, such that

$$\begin{aligned} \vec{\Omega}_L \times \vec{L} = \frac{1}{r^3}\left[\alpha_1 m_1^2 \vec{\chi}_1 + \alpha_2 m_2^2 \vec{\chi}_2\right] \times \vec{L} \\ - \frac{3}{2}\frac{m_1^2 m_2^2}{r^3}\left[(\vec{\chi}_2 \cdot \hat{L})\vec{\chi}_1 + (\vec{\chi}_1 \cdot \hat{L})\vec{\chi}_2\right] \times \vec{L} \end{aligned} \tag{2.28}$$

where r is the separation between the two BHs and

$$\alpha_1 = 2 + \frac{3q}{2} \quad \text{and} \quad \alpha_2 = 2 + \frac{3}{2q} \tag{2.29}$$

for a mass ratio $q = m_2/m_1 \leq 1$.

The PN precession equations have motivated several parameters to capture leading-order effects of in-plane spins on the emitted GW signal. All of these parameters are defined in a frame where the z-axis points along $\vec{L}$. Then, the *in-plane spin vectors* can be written $\vec{\chi}_{i\perp} = (\chi_{i,x}, \chi_{i,y}, 0)$ such that $\vec{\chi}_i = \vec{\chi}_{i\perp} + \chi_i \cos\theta_i \, \hat{z}$. The in-plane spin magnitudes are correspondingly $|\vec{\chi}_{i\perp}| \equiv \chi_{i\perp} = \chi_i \sin\theta_i$. The most commonly-utilized precession parameter in contemporary GW data analysis is χ_p, constructed to capture the average effect of spin-orbit coupling (the first line of Eq. 2.28) over many precession cycles [203, 204],

$$\chi_p = \frac{1}{2\alpha_1 m_1^2}\left(\alpha_1 m_1^2 \chi_{1,\perp} + \alpha_2 m_2^2 \chi_{2,\perp} + |\alpha_1 m_1^2 \chi_{1,\perp} - \alpha_2 m_2^2 \chi_{2,\perp}|\right) \tag{2.30}$$

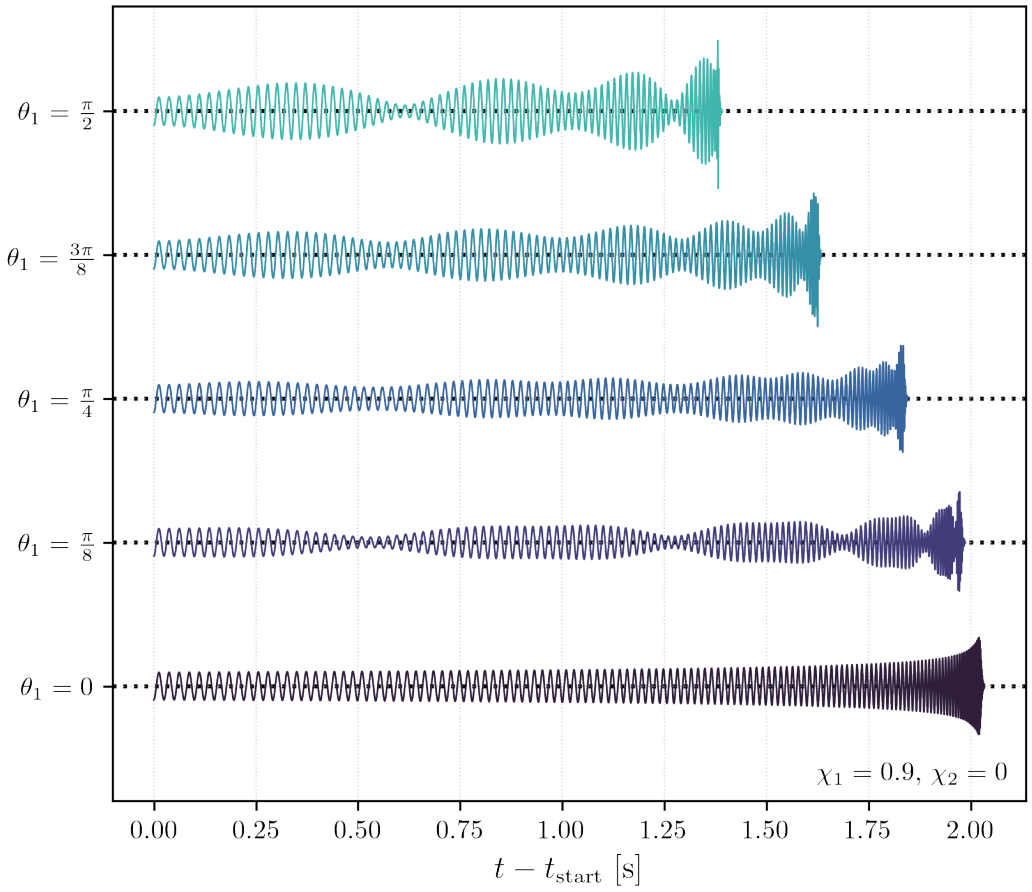


Figure 2.6: GW signals from BBH mergers with different angles $\cos\theta_1$ between the primary spin and orbital angular momentum $\vec{L}$, leading to different degrees of orbital-plane precession. From bottom to top, the primary spin vector goes from being parallel with $\vec{L}$ to perpendicular to $\vec{L}$. All other parameters are the same between each signal; the source BBHs have $m_1 = 30\,M_\odot, m_2 = 2\,M_\odot$ (in the detector frame), an inclination angle of $\iota = \pi/2$, a primary spin magnitude of $\chi_1 = 0.9$, and non-spinning secondaries, $\chi_2 = 0$. Relatively small and unequal masses, large spin magnitudes, and edge on inclinations are intentionally chosen as an extreme case to clearly visualize the imprint of spin precession. The start time t_{start} is when the frequency of the dominant $\ell = 2, m = 2$ mode is 40 Hz; the inclination is also defined at this time, and varies over time leading to the observed amplitude modulation. Waveforms are generated with the IMRPhenomTP model [202]. *Graphic by* ***SJM***.

or more succinctly,

$$\chi_{\rm p} = \max\left[\ \chi_{1\perp},\ \left(\frac{3+4q}{4+3q}\right) q\,\chi_{2\perp}\ \right]\ , \tag{2.31}$$

where I have substituted for α_i and m_i in favor of q, and again $\chi_{i\perp} = \chi_i \sin\theta_i$. Often, $\chi_{\rm p}$ is simply equivalent to the in-plane spin of the primary BH. The parameter is defined between 0 and 1, where $\chi_{\rm p} = 0$ means there is no in-plane spin and $\chi_{\rm p} = 1$ represents the case where the primary BH has a spin with a magnitude of $\chi_1 = 1$ and lies exactly in the orbital plane ($\theta_1 = \pi/2$).

Additional precessing spin parameters are defined in Refs. [205] & [204].[10] The former notes that in the canonical definition of $\chi_{\rm p}$ (Eq. 2.31), only *some* of the features that occur on the precession timescale are averaged over; spin-spin coupling (second line of Eq. 2.28), for instance, is ignored. Ref. [205] defines a "generalized $\chi_{\rm p}$", defined as

$$\chi_{\rm p}^{\rm gen} = \left|\ \vec{\chi}_{1\perp} + \tilde{\Omega}\,\vec{\chi}_{2\perp}\ \right| = \sqrt{\chi_{1\perp}^2 + \chi_{2\perp}^2 + 2\,\tilde{\Omega}\,\chi_{1\perp}\,\chi_{2\perp}\cos\phi_{12}} \tag{2.32}$$

where ϕ_{12} is the in-plane angle between the spins and $\tilde{\Omega}$ takes spin-orbit and spin-spin coupling into account,

$$\tilde{\Omega} \equiv \left(\frac{3+4q}{4+3q}\right) q - 3\,\chi_{\rm eff}\left[\frac{q^2(1-q)}{(4+3q)^2(1+q)}\right]\frac{M^2}{L} + \mathcal{O}\left(\frac{M^4}{L^2}\right) \tag{2.33}$$

for a total mass $M = m_1 + m_2$ [205]. The first term in Eq. (2.33) is that which appears in the traditional $\chi_{\rm p}$. The value of $\chi_{\rm p}^{\rm gen}$ is defined on the range $[0, 2]$, where $\chi_{\rm p}^{\rm gen} > 1$ only arises if *both* BHs in the binary have in-plane spin.

Finally, Ref. [204] addresses the additional fact that $\chi_{\rm p}$ alone does not accurately account for *higher order multi-polar modes* in precessing binaries; see Eq. (2.15). They define the "effective precession spin vector" $\vec{\chi}_\perp$,

$$\vec{\chi}_\perp = \frac{1}{A}\left(\vec{\chi}_{1\perp} + q^2\,\vec{\chi}_{2\perp}\right) \quad \text{where} \quad A = \begin{cases} 1 + q^2\,\chi_{2\perp} & \text{if } \chi_{1\perp} \geq q^2\,\chi_{2\perp} \\ q^2 + \chi_{1\perp} & \text{else} \end{cases} \tag{2.34}$$

which, like $\chi_{\rm p}^{\rm gen}$, accounts for both the primary and secondary spins rather than whichever has a maximal contribution to precession. The magnitude $|\vec{\chi}_\perp|$ is then a scalar quantity analogous to the traditional and generalized $\chi_{\rm p}$. I compare these three precession-encoding parameters on the GW event GW190521 in Sec. 7.C.

[10] The expressions in Eqs. (2.32) & (2.34) are not given directly in their source papers, and were instead derived by me from alternative expressions with the purpose of expressing them in terms of consistent notation with the remainder of this dissertation chapter.

2.5 Spins in the ringdown

The description of spin parameters in this chapter thus far has been limited to the inspiral phase of coalescence, when the two orbiting BHs are well separated. However, the approximations that let us model the inspiral with PN expansions break down when the BHs get sufficiently close to each other. The subsequent merger phase of coalescence is highly nonlinear and can only be well described via numerical techniques [e.g., 206]. Chapter 4 explores how the imprint of aligned and precessing spin changes as GW signals become less and less inspiral dominated.

Once the BHs have merged, the signal enters the ringdown phase, where we are concerned with modeling the perturbed spacetime surrounding the remnant BH. As it equilibrates into its final Kerr state, the remnant radiates GWs in the form of damped sinusoids, known as "quasi-normal modes" (QNMs) [15, 207–209]. After some time t_0, the GW signal can be represented as a superposition of these QNMs, i.e., each $h_{\ell m}$ in Eq. (2.15) can be written [e.g., 15, 153, 155, 157, 209–218]

$$h_{\ell m} = \sum_{n=0}^{N} C_{\ell m n} \exp\left[-i\omega_{\ell m n}(t - t_0)\right] \tag{2.35}$$

where n indexes overtones [e.g., 216, 219–221], such that a higher n is a mode with a shorter decay time.[11] The complex frequencies $\omega_{\ell m n} = 2\pi f_{\ell m n} - i/\tau_{\ell m n}$ of these QNMs depend only on the remnant BH's mass and spin—a consequence of the "no-hair" theorem [e.g., 222]. The complex amplitudes $C_{\ell m n} = A_{\ell m n} \exp[i\phi_{\ell m n}]$ depend on whatever excites the perturbations; in this case, they encode the inspiral dynamics [153, 155–157, 215, 223–225, 225].

The QNM complex frequencies and amplitudes do not have simple closed-form analytic expressions, and therefore must be computed numerically [226]. For Kerr BHs, the dependence on final mass M_f and final spin χ_f is separable, with $\omega_{\ell m n} = \hat{\omega}_{\ell m n}(\chi_f)/M_f$. The separability follows from dimensional analysis; in the Kerr metric, the mass M_f only sets the length/time scale. As seen in Fig. 2.7, the (real) frequency $f_{\ell m n}$ increases (decreases) with spin for $m > 0$ ($m < 0$) modes, while the damping time $\tau_{\ell m n}$ always increases with spin meaning the modes live longer [15, 211, 213, 226].

[11] *Note on notation:* The angular distribution $Y_{\ell m}$ coefficients are given in Eq. (2.15) and are therefore excluded in Eq. (2.35) to avoid double-counting.

GR predicts that BBH remnants for near-equal-mass, low-spin mergers will have $\chi_f \sim 0.7$ [73]. As mass ratio becomes more extreme, the binary's orbital angular momentum decreases, leading to smaller remnant spins; as the component black hole spin magnitudes increase, so too will the remnant spin. For the extremal case of $\chi_1 = \chi_2 = 1$, the remnant spin > 0.9 for all mass ratios [227].

The imprint of inspiral dynamics on the QNM amplitudes is an area of active research. For the aligned-spin case, there exist fitting formulae [223] and surrogate models [16, 228–231]. For the general case including spin precession, analogous tools are being developed. However, because precession breaks the azimuthal symmetry assumed in the non-precessing ringdown, the $(2, 1, 0)$ mode is predicted to be excited in remnants of precessing BBHs [157], making it a key signature to search for in ringdown data. There is mounting observational evidence that observed precessing systems indeed excite the $(2, 1, 0)$ mode [e.g., 214, 232].

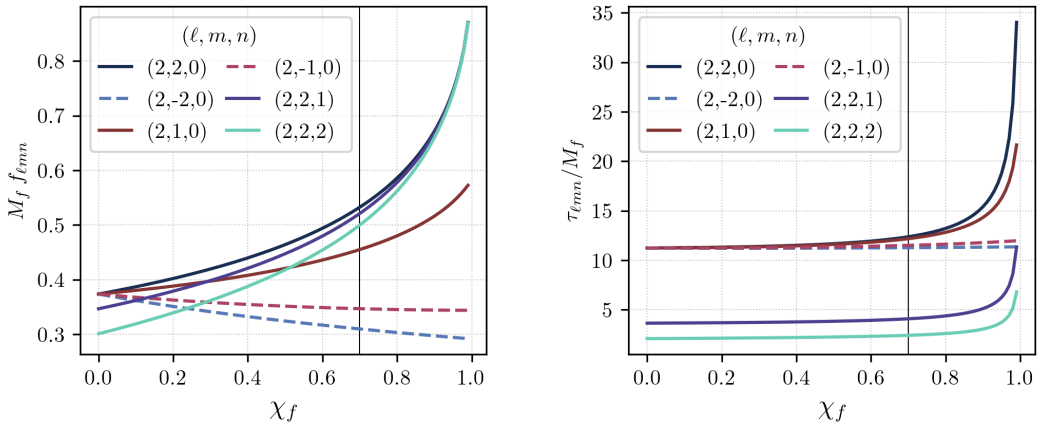


Figure 2.7: The (real) frequency $f_{\ell mn}$ (left panel) and damping time $\tau_{\ell mn}$ (right panel) for various $\ell = 2$ QNMs [233]. The vertical black line is at $\chi_f = 0.7$, the prediction for a typical remnant BH spin [73]. *Graphic by* ***SJM****.*

Chapter 3

BAYESIAN INFERENCE

Chapter 2 described how spins imprint on the gravitational waves emitted by the inspiral and eventual ringdown of two orbiting and merging black holes. In this chapter, I describe how information about black hole spins and other source properties is then actually extracted from GW data using Bayesian inference, the framework underlying nearly all analyses presented in this dissertation.

Broadly speaking, statistics can be split into two camps based on how one defines "probability:" frequentist versus Bayesian. Frequentist statistics defines probability as how *frequently* something occurs over many repeated trials. To a frequentist, a 95% confidence interval means that if you repeated some experiment many times (in the presence of random statistical fluctuations associated with experimental measurement errors), that interval would contain the truth 95% of the time. The p-value is then the probability that over many trials, you would get a result more "extreme" than the one actually measured. Bayesian statistics instead defines probability as a measure of *belief*. To a Bayesian, a 95% credible interval means there is a 95% probability that the truth falls within that interval. Bayesian inference on its own does not involve p-values, as it does not deal with repeated trials and is concerned only with the results that were *actually* obtained. Synergies between the Bayesian and frequentist frameworks are being increasingly explored [e.g., 234] through things like "posterior predictive p-values," a focus of Ch. 14.

The Bayesian method hinges on the basic fact that the probability of A and B being true, $p(A, B)$, can equivalently be written $p(A, B) = p(A|B)\ p(B)$ and $p(A, B) = p(B|A)\ p(A)$, where the former is the probability that B is true and that A is true given B, and the latter is the probability that A is true and that B is true given A, i.e., conditional probabilities. *Bayes' Theorem* [235] equates the two expressions for $p(A, B)$, essentially providing a formula for inverting conditional probabilities:

$$p(A|B) = \frac{p(B|A)\ p(A)}{p(B)}\ . \tag{3.1}$$

3.1 Gravitational wave parameter estimation

In the case of GW parameter estimation (PE), we have some model $\mathcal{H}$ for our data d. This $\mathcal{H}$ includes things such as the parametrized signal model, how we estimate detector noise, and what (if anything [236]) we assume about the astrophysical source population. We want to infer the probability that some parameters λ describe the data given our model, including astrophysical source parameters like masses and spins, noise parameters, etc. Accordingly, we can write Bayes' theorem as [142, 237]

$$p(\lambda|d,\mathcal{H}) = \frac{\mathcal{L}(d|\lambda,\mathcal{H})\ \pi(\lambda|\mathcal{H})}{\mathcal{Z}(d|\mathcal{H})}, \tag{3.2}$$

where for clarity, I use different symbols for each Bayesian "ingredient:"

1. The *posterior* probability $p(\lambda|d,\mathcal{H})$ is our target: the probability that parameters λ describe the data d assuming our model $\mathcal{H}$.

2. The *likelihood* $\mathcal{L}(d|\lambda,\mathcal{H})$ is the probability of observing d given the parameters λ, again under the model $\mathcal{H}$. The likelihood is discussed in further detail below and in Ch. 6.

3. The *prior* $\pi(\lambda|\mathcal{H})$ encodes our beliefs about what λ should be, before observing d. Much of Bayesian inference comes down to choosing an appropriate prior. In typical GW PE, we try to use "agnostic" priors, e.g., uniform or isotropic in quantities of astrophysical importance.[1] Population inference (Sec. 3.2) seeks to generate priors informed by our observed astrophysical GW source distribution.

4. The *evidence* $\mathcal{Z}(d|\mathcal{H})$ is the marginal likelihood,

$$\mathcal{Z}(d|\mathcal{H}) = \int \mathcal{L}(d|\lambda,\mathcal{H})\ \pi(\lambda|\mathcal{H})\ d\lambda \tag{3.3}$$

   which ensures that the posterior probability density is normalized. Given two models $\mathcal{H}_1$ and $\mathcal{H}_2$, the evidence tells us which model the data d prefer,

[1] As the statistician Andrew Gelman reportedly said, "There is no such thing as an uninformative prior." A uniform prior in one parameterization is not uniform in others—an issue especially pertinent to estimating BBH spins. The uniform priors in χ and $\cos\theta$ commonly used in PE are nontrivial in $\chi_{\rm eff}$ and $\chi_{\rm p}$ [238, 239]. See Sec. 13.C and, e.g., Refs. [162, 239–241] for a discussion of how these priors impact single event and population inference for effective spins.

through the so-called *Bayes factor*,

$$\mathcal{B} = \frac{\mathcal{Z}_1}{\mathcal{Z}_2} = \frac{\int \mathcal{L}(d|\lambda, \mathcal{H}_1)\ \pi(\lambda|\mathcal{H}_1)\ d\lambda}{\int \mathcal{L}(d|\lambda, \mathcal{H}_2)\ \pi(\lambda|\mathcal{H}_2)\ d\lambda}\ . \tag{3.4}$$

A Bayes' factor of $\mathcal{B} = 2$, for example, is interpreted as the data being twice as likely under $\mathcal{H}_1$ than $\mathcal{H}_2$.[2]

Defining the likelihood

As formulated in Eq. (3.2), Bayes' theorem allows us to estimate the parameters λ of some GW source in our data d. One of the two main ingredients needed to calculate the posterior probability with Bayes' theorem is the likelihood $\mathcal{L}$ (the other being the prior π, discussed above; the evidence $\mathcal{Z}$ is derived from these two and serves as a normalization constant). Going forward, I will assume we are operating under a single model $\mathcal{H}$ (i.e., not doing any model comparison) and so will exclude $\mathcal{H}$ from formulae for notational simplicity.

To construct the likelihood, we assume that d consists of an astrophysical GW signal h and stationary, Gaussian noise in our instruments n. I will assume we are only estimating the parameters of the astrophysical signal (masses, spins, etc.), and not noise properties.[3] Let λ_0 be the true parameters of the astrophysical source of h. Under these assumptions,

$$d(t) = h(t; \lambda_0) + n(t), \tag{3.5}$$

noting that in this representation the data are a function of time.

We assume that LVK instrumental noise is well described by stationary, Gaussian processes with a known covariance matrix.[4] Then, taking a long segment of data surrounding the signal and applying periodic boundary conditions through windowing, we can transform our data into the Fourier domain,

$$\tilde{d}(f) = \tilde{h}(f; \lambda_0) + \tilde{n}(f). \tag{3.6}$$

[2] Note that Bayes factors tend to be highly prior-dependent and only provide an aggregate test of model performance without demonstrating how a non-preferred model fails [242]. Chapter 14 discusses model-testing procedures that go beyond simply ranking models and seek to provide information about how to improve them to better fit the data.

[3] See Ch. 10 for a case where signal and noise parameters are simultaneously inferred.

[4] All of these assumptions are only approximate for real data, and sometimes they are very poor approximations. For instance, the data are definitely not stationary nor Gaussian in the presence of a glitch.

We additionally (implicitly) assume that there is only one signal in the segment of interest. In O4a, the LVK instruments detected one GW signal from a compact binary every few days [7], validating this treatment of the data. For next generation observatories, the assumption will break down [27–29].

Under the assumptions of stationarity, Gaussianity, and periodicity, the noise-covariance matrix becomes diagonal and is fully represented by a power spectral density (PSD), $\mathcal{S}_n(f)$. The Fourier amplitudes of a given noise *realization* are then independent Gaussian random variables with variance determined by $\mathcal{S}_n(f)$. [5] Any random noise realization $\tilde{x}(f)$ is drawn from a Gaussian distribution, according to the probability density

$$p(x) \propto \exp\left[-\tfrac{1}{2}(x|x)\right] \implies \log p(x) = -\tfrac{1}{2}(x|x) + \text{const.} \tag{3.7}$$

where $(a|b)$ is the noise-weighted inner product [187]

$$(a|b) = 4\text{Re}\int_0^\infty \frac{\tilde{a}(f)\,\tilde{b}^*(f)}{\mathcal{S}_n(f)}\,df\ . \tag{3.8}$$

Taking the inner product of the noise realization with itself yields

$$(x|x) = 4\int_0^\infty \frac{|\tilde{x}(f)|^2}{\mathcal{S}_n(f)}\,df\ . \tag{3.9}$$

Given data defined in N frequency bins f_i with uniform spacing Δf, the inner product becomes

$$(x|x) = 4\Delta f \sum_{i=0}^{N-1} \frac{|\tilde{x}(f_i)|^2}{\mathcal{S}_n(f_i)}\ . \tag{3.10}$$

Using the above, we can formulate our likelihood as the probability that the *residuals* of our data d minus the signal model $h(\lambda)$ are Gaussian-distributed according to the PSD:

$$\begin{aligned}\log\mathcal{L}(d|\lambda) &= -\tfrac{1}{2}(d-h|d-h) + \text{const.}\\ &= -2\Delta f\sum_{i=0}^{N-1}\frac{|\tilde{d}(f_i)-\tilde{h}(f_i;\lambda)|^2}{\mathcal{S}_n(f_i)} + \text{const.}\end{aligned} \tag{3.11}$$

[5] It is possible, and sometimes favorable, to compute the likelihood in the time domain directly; this is discussed in detail in Ch. 6 and forms the basis of Part II. However, it is often fully equivalent and computationally advantageous to calculate the likelihood in the frequency domain due to the diagonal noise-covariance matrix, and as the Fourier-domain representation remains the standard in GW PE [e.g., 142, 243].

Since additive constants do not change the shape of the log-likelihood, they are typically ignored unless one is concerned with computing the evidence $\mathcal{Z}$ (e.g., for Bayes factors) and/or comparing PSDs.

Equation (3.11) assumes that the model for h is a correct representation of the signal in the data; otherwise, the likelihood will be reduced. For this reason, the GW signals in LVK catalogs are typically analyzed with multiple waveform models and results are compared. Each model may rely on a different approach, ranging from the effective one-body formalism [e.g., 244], to phenomenological models [e.g., 245, 246], to building a surrogate of NR simulations [e.g., 16]. Signal mis-modeling ("waveform systematics") can drastically impact parameter estimates, especially in high-mass systems [e.g., 247, 248]. Currently, a substantial and persistent systematic error in waveform modeling is exclusion of orbital eccentricity, especially concurrently with spin precession [e.g., 100]. In future detectors with increased sensitivity, even very minor waveform systematics will have significant impact [e.g., 249–251].

Sampling the posterior

Given infinite time and computational resources, we could evaluate Eq. (3.11) over a grid of λ and get an exact representation of $\mathcal{L}(d|\lambda)$ (at the grid points) and therefore our desired $p(\lambda|d)$. However, in the case of a binary black hole merger, λ is 15-dimensional for a quasi-circular binary,[6] consisting of:

1. The *intrinsic parameters*
   a) Component masses, m_1 and m_2
   b) Component spins, $\vec{\chi}_1$ and $\vec{\chi}_2$
2. The *extrinsic parameters*
   a) Luminosity distance, d_L
   b) Inclination angle, ι
   c) Polarization angle, ψ
   d) Time at coalescence, t_0
   e) Phase at coalescence, φ_0
   f) Right ascension and declination, α and δ

[6] 17-dimensional if the orbit is eccentric, in which case eccentricity and mean anomaly are added.

Thus, evaluating $\mathcal{L}(d|\lambda)$ over a grid of even 10 points in each dimension is already at least 10^{15} calculations.

One solution is to use a *stochastic sampler*, which produces *samples* from the underlying posterior distribution such that the number of samples in some interval is proportional to $p(\lambda|d)$. Common sampling algorithms (in GW science and generally) are Markov Chain Monte Carlo (MCMC) sampling [252] and nested sampling [253].

Many sampling code packages exist, both stochastic or non-stochastic in nature. In GW data analysis, the most widely-used is likely `bilby`, which includes both MCMC and nested samplers [243, 254, 255]. The code `cogwheel` uses symmetries specific to compact binary signals to remove multi-modality in the posterior and speed up sampling [256]. `RIFT` alternatively uses a combination of grid-based and stochastic sampling [257]. Hamiltonian Monte Carlo (HMC) sampling is also often utilized for problems where the gradient of the log-posterior is computable (often done through automatic differentiation), allowing for major speedup on GPUs [258–260]. Innovations beyond stochastic sampling are also on the rise, such as simulation-based inference as encoded in, e.g., `dingo` [261] and `labrador` [262].

3.2 Population inference: from one detection to many

In addition to analyzing each signal in isolation, our abundance of GW observations allows us to probe the astrophysical *population* of merging binary black holes and neutron stars. Doing so reveals insights into how, when, and where compact binaries form and evolve beyond what any single event alone can illuminate [263].

The distribution of λ across the astrophysical population defines a single-event Bayesian prior $\pi_{\rm pop}(\lambda|\Lambda)$ where Λ are *hyper-parameters* governing the shape of the population distribution. The goal of population inference (also called hierarchical inference) is to obtain a posterior on Λ.

The hierarchical likelihood

Using Bayes' theorem (Eq. 3.2), the *hyper-posterior* on population-level parameters Λ given some *hyper-prior* $\pi(\Lambda)$ and a GW catalog $\vec{d} = \{d_i\}$ consisting of $N_{\rm obs}$ independent events is

$$p(\Lambda|\vec{d}) \propto \mathcal{L}(\vec{d}|\Lambda)\pi(\Lambda)\,. \tag{3.12}$$

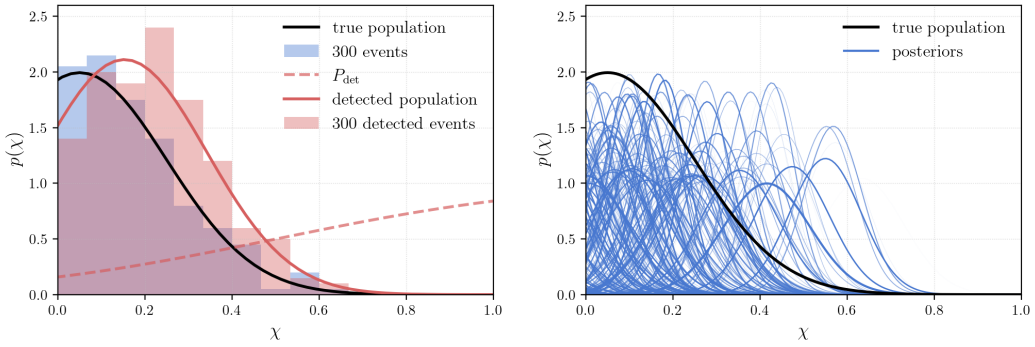


Figure 3.1: Schematic demonstration of the single-event "inputs" to population inference (Eq. 3.15) for one parameter χ in some idealized cases. *(Left)* Each source has no measurement uncertainty and is equally likely to be detected, meaning that the population distribution (blue histogram) is just a series of random draws from the true population (black line). If selection effects are at play, i.e., not every source is equally detectable—here shown with a fiducial $P_{\rm det}(\chi)$ (pink dashed)—then the detected (pink) and underlying (blue) populations are different. *(Right)* Each source is equally likely to be detected (no selection effects) but has measurement uncertainty. Each blue trace is the posterior for χ for one event in the observed population.

To build intuition, let us begin by constructing the population likelihood (or *hierarchical likelihood*) $\mathcal{L}(\vec{d}|\Lambda)$ in an idealized scenario where every GW event in the astrophysical population has an equal probability of being detected. The population likelihood is then a product over the individual events in the catalog

$$\mathcal{L}(\vec{d}|\Lambda) = \prod_{i}^{N_{\rm obs}} \mathcal{L}(d_i|\Lambda) \tag{3.13}$$

whose individual likelihoods $\mathcal{L}(d_i|\Lambda)$ are each calculated by *marginalizing* over the single-event parameters λ,

$$\mathcal{L}(d_i|\Lambda) = \int \mathcal{L}(d_i|\lambda)\, \pi_{\rm pop}(\lambda|\Lambda)\, d\lambda \ , \tag{3.14}$$

where $\mathcal{L}(d_i|\lambda)$ is given in Eq. (3.11) and $\pi_{\rm pop}(\lambda|\Lambda)$ is our population model of choice.

However, in real life, compact binaries with different source properties are more or less likely to be detected. For example, mergers that are closer to Earth and/or more massive are more detectable. As such, we must include *selection effects* in our hierarchical likelihood. Combining the above two equations and

adding selection effects yields

$$\boxed{\mathcal{L}(\vec{d}|\Lambda) \propto \prod_{i}^{N_{\rm obs}} \frac{\int \mathcal{L}(d_i|\lambda)\,\pi_{\rm pop}(\lambda|\Lambda)\,d\lambda}{\xi(\Lambda)}} \tag{3.15}$$

where the *detection efficiency* $\xi(\Lambda)$ is equivalent to the detectable fraction of the population described by Λ. It is calculated by

$$\xi(\Lambda) = \int \pi_{\rm pop}(\lambda|\Lambda)\,P_{\rm det}(\lambda)\,d\lambda\,, \tag{3.16}$$

where $P_{\rm det}(\lambda)$ is the probability (between 0 and 1) that an individual event with parameters λ is detected.

In contrast to nearly every other sub-field in astronomy, GW selection effects are conceptually straightforward to estimate (albeit a massive computational undertaking involving many people). We robustly estimate $P_{\rm det}(\lambda)$ using software injections from some known, simulated reference population. Simulated events are drawn from the reference population, inserted into real GW data, and sent through the same pipelines used to detect real signals, i.e., utilizing search statistics and signal recognition [264, 265]. Once complete, the processed injections provide a near-exact map from "detected population" to "true underlying astrophysical population."

Note: Equation (3.15) implicitly conditions on the observed number of detections $N_{\rm obs}$. More generally, we can instead *explicitly* model the merger rate. Canonically, it is assumed that an inhomogeneous Poisson process generates realizations from the source population with a differential merger rate density,

$$\frac{dN}{d\lambda} = \mathcal{R}\,\pi_{\rm pop}(\lambda|\Lambda)\ , \tag{3.17}$$

where N is the total number of mergers (detected and not) over some period of time, $\mathcal{R}$ is the overall rate normalization, and $\pi_{\rm pop}(\lambda|\Lambda)$ is the population distribution (normalized to unity). Including the rate, the joint posterior density for $\vec{d}$, Λ, and $\mathcal{R}$ is

$$p(\vec{d},\mathcal{R},\Lambda) \propto \pi(\Lambda)\,\pi(\mathcal{R})\,\mathcal{R}^{N_{\rm obs}}\ e^{-\mathcal{R}\,\xi(\Lambda)} \prod_{i}^{N_{\rm obs}} \int \mathcal{L}(d_i|\lambda)\,\pi_{\rm pop}(\lambda|\Lambda)\,d\lambda\ . \tag{3.18}$$

Equation (3.15) is recovered by first analytically marginalizing over the rate normalization with a scale-invariant prior $\pi(\mathcal{R}) \propto 1/\mathcal{R}$:

$$\begin{aligned}\int_0^\infty \pi(\mathcal{R})\, \mathcal{R}^{N_{\rm obs}} e^{-\mathcal{R}\,\xi(\Lambda)} d\mathcal{R} &= \int_0^\infty \mathcal{R}^{N_{\rm obs}-1} e^{-\mathcal{R}\,\xi(\Lambda)} d\mathcal{R} \\ &= \xi(\Lambda)^{-N_{\rm obs}} \int_0^\infty u^{N_{\rm obs}-1}\, e^{-u} du = \xi(\Lambda)^{-N_{\rm obs}}\, (N_{\rm obs}-1)!\end{aligned} \tag{3.19}$$

where the factorial is an overall normalization constant that can be ignored. Thus, Eq. (3.18) becomes

$$\begin{aligned}p(\Lambda|\vec{d}) &\propto \pi(\Lambda)\, \xi(\Lambda)^{-N_{\rm obs}} \prod_i^{N_{\rm obs}} \int \mathcal{L}(d_i|\lambda)\, \pi_{\rm pop}(\lambda|\Lambda)\, d\lambda \\ &\propto \pi(\Lambda)\, \mathcal{L}(\vec{d}|\Lambda)\end{aligned} \tag{3.20}$$

where $\mathcal{L}(\vec{d}|\Lambda)$ is that of Eq. (3.15). In practice, one can directly sample $\mathcal{R}$ along with Λ. The analyses in this dissertation instead sample from Eq. (3.15) and, when needed, reconstruct the merger rate in post-processing.

Choosing a population model

Population modeling refers to choosing the functional form of $\pi_{\rm pop}(\lambda|\Lambda)$, and is often broken down into two complementary approaches [e.g., 23]. The *strongly modeled approach* (also called the parametric approach) assumes that $\pi_{\rm pop}(\lambda|\Lambda)$ takes a fixed functional form that is characterized by a finite number of interpretable parameters Λ that are (often) linked to a specific, underlying astrophysical interpretation. Examples of strong models include power laws, Gaussians, notch filters, etc. or some mixture between them. In the case of a Gaussian, for instance, the hyper-parameters Λ are the distribution's mean and standard deviation, while for a power law distribution, Λ includes the power-law index.

On the other hand, the *weakly modeled approach* (also called data-driven or non-parametric approach) assumes that $\pi_{\rm pop}(\lambda|\Lambda)$ has a more flexible functional form characterized by a large number of parameters that are not linked to a specific astrophysical interpretation. Weakly-parametrized models include histograms, splines, Gaussian processes, and autoregressive processes [e.g., 266–271]. The two approaches can also be combined, leading to a population model that consists of a superposition of some baseline strongly-parametrized function with a weakly-parametrized function on top to capture deviations from the baseline [e.g., 272].

Monte Carlo estimation

In practice, instead of working with the single-event likelihood function $\mathcal{L}(d_i|\lambda)$ itself, we have discrete samples from single-event posteriors, as discussed in Sec. 3.1. Similarly, we don't have an analytic form for the detection efficiency $\xi(\Lambda)$, but rather discrete examples of events from our simulated reference population that pass the detection threshold. Thus, by necessity, when computing the hierarchical likelihood (Eq. 3.15), we nearly always use *Monte Carlo estimation*.[7]

Monte Carlo estimation is a technique for approximating intractable integrals by taking many random draws from some known and easy-to-sample probability density, reweighting those samples to the integrand, and averaging over them [e.g., 279]. In the following, a tilde over a quantity indicates a Monte Carlo estimator for that quantity. For the single-event likelihoods,

$$\mathcal{L}(d_i|\Lambda) = \int \mathcal{L}(d_i|\lambda)\,\pi_{\rm pop}(\lambda|\Lambda)\,d\lambda$$
$$\rightarrow \quad \tilde{\mathcal{L}}(d_i|\Lambda) = \frac{1}{N_{{\rm PE},i}} \sum_{j=1}^{N_{{\rm PE},i}} \frac{\pi_{\rm pop}(\lambda_{ij}|\Lambda)}{\pi_{\rm PE}(\lambda_{ij})} \tag{3.21}$$

where $N_{{\rm PE},i}$ is the number of PE samples for the $i^{\rm th}$ event in the catalog and $\pi_{\rm PE}$ is whatever prior was used when obtaining those PE samples. Here we use Bayes' theorem to relate the single-event likelihood, posterior, and prior, i.e., $\mathcal{L}(d_i|\lambda) \propto p(\lambda|d_i)/\pi_{\rm PE}(\lambda)$, leaving the reweighting shown. Similarly, for the detection efficiency,

$$\xi(\Lambda) = \int \pi_{\rm pop}(\lambda|\Lambda)\,P_{\rm det}(\lambda)\,d\lambda$$
$$\rightarrow \quad \tilde{\xi}(\Lambda) = \frac{1}{N_{\rm draw}} \sum_{j=1}^{N_{\rm found}} \frac{\pi_{\rm pop}(\lambda_j|\Lambda)}{p_{\rm draw}(\lambda_j)} \tag{3.22}$$

where $N_{\rm draw}$ is the total number of events simulated in the injection study, $N_{\rm found}$ is the number of those events that are detected, and $p_{\rm draw}$ is the reference population the injections were drawn from.

Monte Carlo integration leads to intrinsic variance in the estimation of the hierarchical likelihood. Several statistical quantities have been proposed to assess whether a given Monte Carlo estimate is trustworthy [279–281]. Some

[7] See, e.g., Refs. [273–278] for analyses that make use of other approaches.

propose excising regions of parameter-space afflicted by a certain level of Monte Carlo uncertainty; doing so acts as a non-trivial hyperprior. More details are discussed in, e.g., Sec. 12.D and Sec. 13.C.

Chapter 4

MAPPING PARAMETER CORRELATIONS IN SPINNING BINARY BLACK HOLE MERGERS

This chapter is published as Ref. [150]:

Karen Kang, **Simona J. Miller**, Katerina Chatziioannou, and Deborah Ferguson. Mapping parameter correlations in spinning binary black hole mergers. *Phys. Rev. D*, 112(6):064020, 6 2025. ***SJM*** *conceptualized the project jointly with co-authors, mentored K. Kang (undergraduate), made Figure 13, and extensively edited the text.*

The Appendices of Ref. [150] excluded from this thesis are those that do not reflect ***SJM****'s notable contribution to the work.*

Abstract

The spins of binary black holes measured with gravitational waves provide insights about the formation, evolution, and dynamics of these systems. However, interpreting these measurements—especially for heavy black holes— remains an open problem. While the imprint of spin during the inspiral phase, where the black holes are well-separated, is understood through analytic descriptions of the dynamics, no such expressions exist for the merger. Though numerical relativity simulations provide an exact solution (to within numerical error), the imprint of the full six spin degrees of freedom on the signal is not transparent. In the absence of analytic expressions for the merger and to advance our ability to interpret massive binary black hole spin measurements, here we propose a waveform-based approach. Leveraging a neural network to efficiently calculate mismatches between waveforms, we identify regions in the parameter space of spins and mass ratio that result in low mismatches and thus similar waveforms. We map these regions with a Gaussian fit, thus identifying correlations and quantifying their strength. For low-mass, inspiral-dominated systems, we recover the known physical imprint: larger aligned spins are correlated with more equal masses as they have opposite effects on the inspiral length. For high-mass, merger-dominated signals, a qualitatively similar correlation is present, though its shape is altered and strength decreases with larger total mass. Correlations between in-plane spins and mass ratio follow

a similar trend, with their shape and strength altered as the mass increases. Our new methodology of waveform-based correlation mapping provides a first step toward systematically modeling spin effects in merger-dominated signals across the full intrinsic parameter-space and motivates future effective spin parameters beyond the reach of analytic methods.

4.1 Introduction

Quasicircular black hole binaries (BBHs) observed through gravitational waves (GWs) by LIGO [4], Virgo [5], and KAGRA [6] are characterized by the component masses and spins. Though less well measured than the masses, the magnitudes and directions of the BH spins carry imprints of the evolutionary history of the binary and its constituents, e.g., Ref. [1]. Measurements of BH spins from both individual events and the population as a whole are therefore central to the astrophysical interpretation of the observed signals, e.g., Refs. [282, 283].

Inference of BBH parameters hinges on waveform models for the GW signal as observed in the detectors [284–287]. During the inspiral phase, these models are based on analytic equations that are valid in the regime of low-orbital-velocity (compared to the speed of light): the post-Newtonian (PN) expansion [148]. The subsequent merger phase is not analytically tractable in any perturbative scheme. Instead, waveform models are based on numerical relativity (NR) simulations of the binary's full dynamics either by calibrating phenomenological terms, e.g., Refs. [245, 288] or by directly interpolating the simulations [16].

A general analytic understanding of a problem helps identify the most appropriate parameters for studying its behavior. For example, the first term in the PN expansion of the inspiral GW phase introduces a certain mass combination, the chirp mass [14, 148]. On the spin front, constraints are expressed through the effective aligned and precessing spins, both again motivated by PN equations. The effective aligned spin $\chi_{\rm eff}$ is related (but not exactly equal) to the leading-order 1.5PN spin-orbit term in the GW phase [188, 289]. It characterizes the spin components in the direction of the Newtonian orbital angular momentum and is conserved under precession to at least the 2PN order [193]. The effective precessing spin $\chi_{\rm p}$ is instead motivated by the 2PN precession equations, and captures the precessional motion of the binary orbit

around the direction of the total angular momentum [199] through in-plane spin components [77, 200, 203].

Beyond inspiral and PN considerations, spin dynamics also affect the merger regime which can dominate the observed signal from massive BBHs. NR simulations show that the direction of GW emission keeps changing, i.e., precessing, during the merger [151] and further affects the ringdown mode content [155–157, 214]. Parameter inference based on simulated data further shows that spin constraints can be achieved from massive binaries with merger-dominated signals though they are typically weaker than mass constraints [103, 104].

Although the majority of GW signals observed are consistent with vanishing $\chi_{\rm eff}$ and uninformative $\chi_{\rm p}$ [285, 287], notable exceptions exist. These include some of the most massive BBHs detected, with a significant portion of the signal corresponding to the merger. GW190521 [20, 107] is the most massive confidently detected BBH to date. While consistent with $\chi_{\rm eff} = 0$, its $\chi_{\rm p}$ is constrained to large values, corresponding to large in-plane spins [20, 107]. Under the interpretation of a quasicircular BBH, the measurement is attributed to the suppression of the final pre-merger cycle due to precessional motion [101]. However, the short duration of the signal makes it subject to alternative explanations, like an eccentric [290, 291] or highly-unequal mass binary [292], or a head-on collision [293]. GW200129 displays similar spin properties [287, 294], but its interpretation is complicated by the fact that an instrumental glitch overlapped the portion of the signal that drives the $\chi_{\rm p}$ measurement [159, 295]. GW191109, notable for its negative $\chi_{\rm eff}$ [287], also has an informative $\chi_{\rm p}$ measurement, but also overlaps with a glitch [158]. Further events have $\chi_{\rm p}$ values that are informatively *small* [287], while a total of 10 events exclude $\chi_{\rm eff} = 0$ at 99% credibility [285–287, 296, 297] At the population level, the distributions of $\chi_{\rm eff}$ and $\chi_{\rm p}$ further point to the existence of BBHs that are precessing [282, 283].

The above discussion and most BBH spin constraints in the literature are based on $\chi_{\rm eff}$ and $\chi_{\rm p}$ (and its variants [204, 205]). However, both parameters are motivated by PN dynamics and their interpretation is only fully valid in the inspiral regime. Though this does not invalidate their use to express constraints from merger-dominated events, it does suggest that they might not fully capture the imprint of spins on merger signals. For example, in the case of non-precessing BBHs, perturbation theory suggests that post-merger emission is characterized by the remnant's mass and spin, for which there exist

approximate fitting formulae that depend on the pre-merger masses and spin amplitudes [197, 223]. Alternative merger-motivated spin parametrizations would aid in theoretically understanding the merger dynamics and devise more efficient parameter estimation methods, for example through re-parametrized sampling, e.g., Ref. [298]. Moreover, since spin constraints drive influential astrophysical conclusions, morphologically understanding their imprint on BBH mergers could safeguard against systematics from waveform models [48], instrumental glitches [158, 159], or alternative interpretations [99].

In the absence of simple analytic expressions for the merger dynamics, we propose an alternative approach to study the correlation structure of BBH parameter space. Efficient parametrizations typically simplify a problem by tracking symmetries and parameter correlations. For example, the fact that the chirp mass is the best measured mass combination is equivalent to the fact that a small change in the chirp mass results in a very different signal. Two signals with different chirp masses therefore have a low match, i.e., normalized inner product. If we had no knowledge of the chirp mass, its existence and formula could be identified by studying how signals differ as their parameters change or by examining the structure of the posterior for the component masses. Such an approach is based on the ability to generate signals for various parameters, a situation that exactly corresponds to merging BBHs: using waveform models, we can simulate signals for any parameter combination even though we lack simple equations for the binary dynamics.

In this work, we apply the above idea to spin parameters. Restricting to the seven-dimensional space of mass ratio and spins, we compute waveform mismatches between signals generated by BBHs with different parameters. Mismatches are evaluated by extensions of the neural network constructed in Ref. [299] that maps binary parameters directly to signal mismatches, bypassing the computationally-inefficient steps of waveform generation and inner-product calculation. Compared to that study, we enhance the network training data to incorporate higher-order modes, and additionally train networks at different values for the binary total mass so as to target different parts of the signal (inspiral versus merger). In this first study, we restrict to systems with fixed values of the total mass. We consider the spin degrees of freedom that are aligned with the orbital angular momentum and those in the orbital plane separately, and use a Gaussian fit to map directions in the parameter space

along which mismatches remain low. We interpret these directions as "spin correlation" directions, and study how they change for systems with different binary total mass, i.e., signals that are more or less merger-dominated.

We verify that this approach can recover the known imprint of $\chi_{\rm eff}$ and mass ratio on low-mass, inspiral-dominated systems. Increasing the aligned spin or decreasing the mass ratio leads to longer signals [196, 197], therefore the two parameters are positively correlated; there exists a prescription for increasing both such that the signal remains approximately unaltered. This correlation is distinct from the $\chi_{\rm eff}$-mass ratio correlation observed in parameter estimation [13, 187, 191, 300, 301], since we fix the system total mass. **As the total mass of the system increases and less of the inspiral is observable, both the shape and the strength of the $\chi_{\rm eff}$-mass ratio correlation are altered.** The slope of the correlation in $\chi_{\rm eff}$-mass ratio space decreases. A smaller change in spin is required to cancel out a change in the mass ratio and thus the former's imprint on the signal is stronger. Simultaneously, the strength of the correlation, quantified by the Gaussian fit covariance matrix, is weakened. An analytic fit to the $\chi_{\rm eff}$-mass ratio correlation as a function of the total mass provides a phenomenological quantitative description of the spin imprint that extends to merger-dominated signals. **Correlations between the in-plane spins and mass ratio also become less steep and strong as the total mass increases, effectively disappearing at a total mass of $270\,M_\odot$.**

Traditionally, BBH parameter-space exploration is done *locally* around a given parameter point, either in full generality using parameter estimation (PE) with sampling [243] or approximately through the use of the Fisher information matrix (FIM) [302]. Our method is technically related to the FIM, which is approximately the *inverse* covariance matrix, whereas we study the eigen-directions of the covariance matrix. More importantly, we extend the local analysis by tracing across a given parameter space to reveal *non-local* correlation paths.[1] Unlike traditional sampling, which probes correlations through computationally intensive posterior sampling, our approach efficiently explores a structured subset of parameter space informed by the local correlations. Each point in one of our correlation maps approximates what would

[1] By "non-local" we mean not restricted to be close to the original point of interest. For example, some paths span the full allowed region of possible spins and/or mass ratio.

otherwise require a full PE analysis, making the otherwise computationally prohibitive exploration of the BBH intrinsic parameter space feasible through mismatches [303] and a neural network [299].

In full generality, all 15 parameters of a quasi-circular BBH affect how the signal appears in a detector and thus if spins can be measured. To keep the problem manageable, we break down the full, complicated imprint of spins into its influence on reduced-dimension parameter-spaces. We restrict to intrinsic parameters only, fixing all extrinsic degrees of freedom (e.g., inclination, sky position, and distance). We further break down the space by fixing total mass and analyzing aligned-spin and precessing-spin systems separately. For precessing systems, we restrict to single-spin configurations. This structured decomposition breaks the problem into manageable pieces and provides a first step toward characterizing mass–spin correlations in the merger regime, where such correlations remain poorly understood quantitatively. Our methodology thusly sets the groundwork for future studies in fuller parameter spaces lacking analytical descriptions.

The rest of the paper is organized as follows. In Sec. 4.2, we present the details of the methodology including notation, the mismatch-predicting networks, and the correlation mapping algorithm. In Sec. 4.3, we apply the mapping algorithm to low-mass systems and aligned-spin degrees of freedom, while the aligned-spins of more massive systems (merger correlations) are considered in Sec. 4.4. In Sec. 4.5 we consider in-plane degrees of freedom for all masses. Finally, our conclusions are given in Sec. 4.6.

4.2 Methodology

In this section, we describe the approach to exploring correlations in the parameter space of BBH signals. In Sec. 4.2a, we introduce parametrizations used for a binary's masses and spins and establish notation. In Sec. 4.2b, we discuss the neural networks that map binary parameters to mismatches. We detail the mapping algorithm that tracks low-mismatch directions along the parameter space in Sec. 4.2c.

(a) Parametrizing quasicircular binaries

Quasi-circular BBHs are characterized by 8 intrinsic parameters: 2 masses and 6 components of the spin vectors.[2] Masses can be expressed through the total mass $M = m_1 + m_2$, the mass ratio $q = m_1/m_2 \geq 1$, and the symmetric mass ratio $\eta = q/(q+1)^2 \in (0, 0.25]$, where m_i, $i \in \{1, 2\}$ are the component masses. A symmetric mass ratio of $\eta = 0.25$ corresponds to an equal-mass BBH; increasingly large q map to increasingly small η.

The dimensionless component spin vectors $\vec{\chi}_1, \vec{\chi}_2$ are typically expressed in a frame whose z-axis denotes the direction of the Newtonian orbital angular momentum $\hat{L}$. The frame is defined at specific reference time or frequency; here we use a reference of 20 Hz, and fix the inclination to be face-on ($\iota = 0$), which maximizes the amplitude of the dominant $(2, 2)$ mode, corresponding to the loudest and most likely observed signals. We note that care must be taken when interpreting results for precessing systems: since precession enters through inclination-dependent mode mixing [304], correlations may vary at other inclinations [102]. Each vector has a magnitude χ_i, polar (tilt) angle θ_i relative to $\hat{L}$, and azimuthal angle ϕ_i with respect to the x-axis, defined as the line of separation between the two BH. Equivalently, $\vec{\chi}_i$ can be expressed in Cartesian components $(\chi_{ix}, \chi_{iy}, \chi_{iz})$.

Motivated by PN equations, the 6 spin degrees of freedom can be re-packaged into two effective parameters. The effective aligned spin $\chi_{\rm eff}$ is the mass-weighted average of the component spins along $\hat{L}$ [193, 289]:[3]

$$\chi_{\rm eff} = \frac{q\chi_1 \cos\theta_1 + \chi_2 \cos\theta_2}{q+1} \in [-1, 1]\,. \tag{4.1}$$

This quantity reduces to the leading-order spin-orbit term in the GW phase in the equal-mass limit and it is conserved under precession to at least the 2PN order [193]. When the BH spins are misaligned with $\hat{L}$, the orbital plane precesses about the direction of the total angular momentum [199]. The effective precessing spin $\chi_{\rm p}$ is the mass-weighted in-plane spin component motivated by

[2] An additional 7 extrinsic parameters describe a binary's location and orientation. As these affect the waveform through known analytic expressions (e.g., antenna patterns) [168], we work in the geocenter frame and focus on intrinsic parameters.

[3] *Note on notation:* In this chapter, we define $q = m_1/m_2 \geq 1$, which is the inverse from other sections of this dissertation. Thus, Eqs. (4.1) & (4.2) look slightly different than, say, Eqs. (7.3) & (7.1). Both mass ratio conventions are both widely-used in GW astronomy, with $q \geq$ being somewhat more common in waveform modeling and NR, while $q \leq 1$ is more widely used in parameter estimation.

the precession equations [77, 200, 203]

$$\chi_{\rm p} = \max\left(\chi_1 \sin\theta_1, \left(\frac{4+3q}{4q^2+3q}\right)\chi_2 \sin\theta_2\right) \in [0,1]\,. \tag{4.2}$$

Since we are interested in the imprint of binary parameters on the detected signal, all parameters are quoted in the detector frame.

(b) Waveform mismatch with neural networks

We quantify the similarity between the signals emitted by two BBHs h_1, h_2 with the *mismatch* [302, 305]

$$\mathcal{MM} \equiv 1 - \max_{t,\phi} \frac{\langle h_1 \mid h_2\rangle}{\sqrt{\langle h_1 \mid h_1\rangle \langle h_2 \mid h_2\rangle}} \in [0,1]\,, \tag{4.3}$$

where maximization is over relative time t and phase ϕ shifts and the inner product is

$$\langle h_1 \mid h_2\rangle = 2\int_{f_0}^{\infty} \frac{\tilde{h}_1^* \tilde{h}_2 + \tilde{h}_1 \tilde{h}_2^*}{S_n}\, df\,, \tag{4.4}$$

where tildes denote Fourier transforms, S_n is the one-sided noise power spectral density, and $f_0 = 20\,\mathrm{Hz}$ is the starting frequency that corresponds to the detector low-frequency cutoff. In what follows, we take $S_n \sim$ constant, i.e., a flat noise spectrum. The mismatch is normalized such that $\mathcal{MM} = 0$ represents identical waveforms and $\mathcal{MM} = 1$ indicates orthogonal waveforms.

Directly calculating enough mismatches to map parameter correlations in a high-dimensional parameter space is computationally expensive. Instead we turn to Ref. [299] in which a neural network is trained on the Simulating Extreme Spacetimes (SXS) NR waveform catalog [206]. Originally constructed to identify regions of the parameter space that are sparsely covered by the catalog, the network provides a map between two sets of intrinsic BBH parameters $\boldsymbol{\lambda} \equiv \{\eta, \vec{\chi}_1, \vec{\chi}_2\}$ and the mismatch between their waveforms. Mismatches refer to geocenter waveforms h_i that have not been projected onto a detector network and thus focus on the intrinsic binary dynamics. All signals are evaluated at a constant total mass determined by the length of the simulations.

We extend this network [299] as detailed in Appendix A of Ref. [150]. Firstly, we include higher-order radiation modes and precession effects. Secondly, as we are interested in exploring the impact of spins on different portions of the signal, we consider three values of the total mass: $30\,M_\odot$, $90\,M_\odot$, and $270\,M_\odot$,

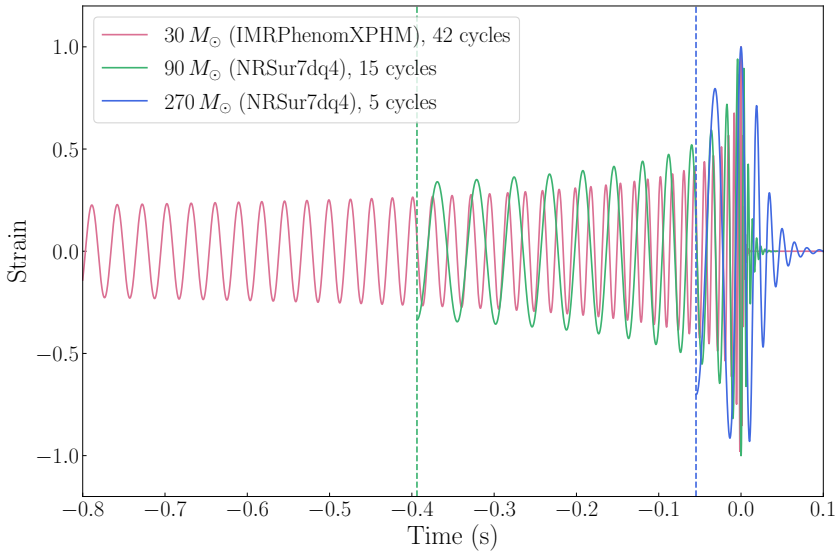


Figure 4.1: Scaled waveforms for equal-mass BBHs with zero spins at different total masses: 30 $M_\odot$ (red), 90 $M_\odot$ (green), and 270 $M_\odot$ (blue) evaluated from 20 Hz. Dashed lines indicate the start of the waveform and the number of cycles is given in the legend.

training a separate network for each value. The former corresponds to low-mass BBHs whose observed signal is dominated by the inspiral phase, the second to signals where both the late inspiral and the merger are observable, and the third to GW190521-like signals where only a handful of merger cycles are observed. For reference, Fig. 4.1 shows equal-mass nonspinning waveforms for each total mass. Thirdly, rather than the SXS catalog, we train the network with waveform models. NRSur7dq4 is a surrogate to NR with accuracy comparable to NR simulations [16]. We therefore use it for the 90 $M_\odot$, and 270 $M_\odot$ networks. Due to its finite length, however, NRSur7dq4 cannot model the full signal from 30 $M_\odot$ BBHs, for which we instead use the phenomenological IMRPhenomXPHM [245] model. Based on NRSur7dq4's regime of validity, all networks are trained in the region $q \leq 6$ (equivalently $\eta > 6/49 \simeq 0.12$), $\chi_i < 1$, $\theta_i \in [0, \pi]$, and $\phi_i \in [0, 2\pi]$.

(c) Mapping parameter correlations

We define parameter correlations as regions in the parameter space that result in similar waveforms. To identify these regions, we define the *parameter space* that consists of N draws of BBH parameters uniformly in symmetric mass

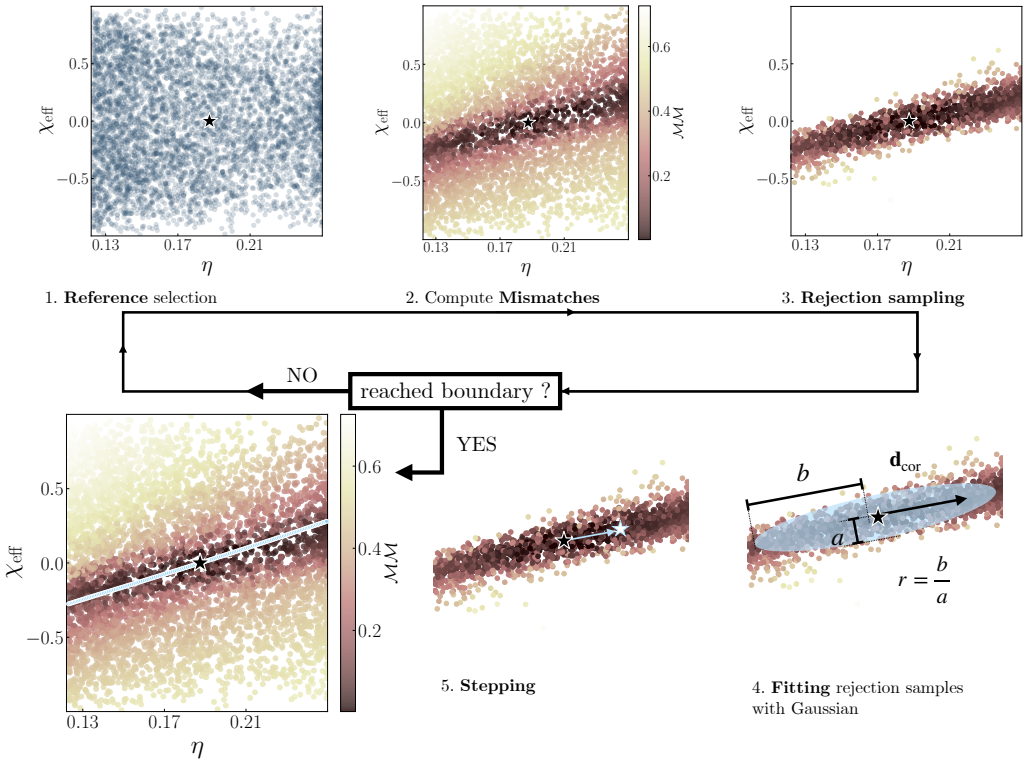


Figure 4.2: Workflow diagram for the correlation mapping algorithm described in Sec. 4.2. This demonstration corresponds to mapping correlations in the 2D effective space of $\boldsymbol{x}^{(\mathbf{eff})} = [\eta, \chi_{\rm eff}]$ and the $90\,M_\odot$ network. 1. Reference point (black star) in a randomly sampled parameter space (blue dots); 2. Mismatch with respect to the reference point (color map); 3. Rejection sampling based on Eq. (4.7); 4. The fitted 1-σ contour, the maximum variance eigendirection of the covariance matrix (vector) $\boldsymbol{d}_{\rm cor}$, and the correlation strength r; 5. Update the reference point along $\boldsymbol{d}_{\rm cor}$ per Eq. (4.8). We iterate through steps 1–5 until we reach a parameter space boundary in one of the dimensions over which we are mapping (in this visualization, the upper limit of η). The total path taken in the $\chi_{\rm eff} - \eta$ space is shown in the final panel (bottom left) together with the mismatches against the reference point.

ratio, spin magnitude, and spin directions within their training region:

$$\mathcal{PS} = \{\boldsymbol{\lambda}_i \mid \boldsymbol{\lambda}_i = \{\eta_i, \vec{\chi}_{1i}, \vec{\chi}_{2i}\}\,, \forall\, i \in N\}\ . \tag{4.5}$$

Parameter correlations can be traced with different parameter subsets and we consider the following.

1. **Effective parameter space**: Knowledge of the form of $\chi_{\rm eff}$ and that η (rather than q) appears in the GW phase, suggests directly mapping correlations in the reduced 2-dimensional $\boldsymbol{x}^{\mathbf{(eff)}} = [\eta, \chi_{\rm eff}]$ space. In this space, the full parameter set $\boldsymbol{\lambda}$ is not fully determined; we select $\boldsymbol{\lambda} = \{\eta, \vec{\chi}_1 = (0, 0, \chi_{\rm eff}), \vec{\chi}_2 = (0, 0, \chi_{\rm eff})\}$.

2. **Aligned parameter space**: When we instead consider generic aligned-spin correlations (e.g., a generalization of $\chi_{\rm eff}$), we map the 3-dimensional space of $\boldsymbol{x}^{\mathbf{(al)}} = [q, \chi_{1z}, \chi_{2z}]$. The full parameter set $\boldsymbol{\lambda}$ is now fully determined by $\boldsymbol{x}^{\mathbf{(al)}}$.

3. **Precessing parameter space**: Correlations between the in-plane spin degrees of freedom correspond to the 5−dimensional space of $[\eta, \chi_{1x}, \chi_{2x}, \chi_{1y}, \chi_{2y}]$. However, since two-spin precession effects [306, 307] and the imprint of the in-plane azimuthal angle [308] are weak,[4] we restrict to a single precessing spin along the binary x-axis and the 2-dimensional space of $\boldsymbol{x}^{\mathbf{(pre)}} = [\eta, \chi_1]$, thus setting the secondary spin and other components of the primary spin to zero. The full parameter set $\boldsymbol{\lambda}$ is then fully determined by $\boldsymbol{x}^{\mathbf{(pre)}}$.

In summary, each $\boldsymbol{x}$ (effective, aligned, precessing) is a subset of $\boldsymbol{\lambda}$ with some parameters fixed to a constant value.

Given a choice of parameters $\boldsymbol{x}$ within which we look for correlations and the appropriate neural network `NN`, we map the correlation directions iteratively. The process is shown schematically in Fig. 4.2 for the $\chi_{\rm eff} - \eta$ correlation and the $90\,M_\odot$ network, and consists of the following steps:

1. **Reference**: We select a reference point in the appropriate parameter space (effective, aligned, precessing) $\boldsymbol{\lambda}_j$ and corresponding $\boldsymbol{x}_j$ around which we look for correlations.

[4] We have verified both conclusions by finding no correlations in preliminary results involving $\vec{\chi}_2$ and ϕ_1.

2. **Mismatches**: Using the neural network, we compute the mismatch $\mathcal{MM}$ between the reference and all other points in $\mathcal{PS}$:

$$\mathcal{MM} = \{\mathcal{MM}_i \mid \mathcal{MM}_i = \texttt{NN}\left(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j\right), \forall \boldsymbol{\lambda}_i \in \mathcal{PS}\} \ . \tag{4.6}$$

   A colormap of the mismatches reveals the correlation direction in Fig. 4.2; here the region of lowest mismatch (darkest color) is a diagonal path from negative $\chi_{\rm eff}$-low η to positive $\chi_{\rm eff}$-high η.

3. **Rejection Sampling**: Each point in $\boldsymbol{\lambda}_i$ receives a weight w_i inspired by the dependence of the likelihood on the mismatch [142, 243, 309–311]

$$w_i = \exp\left(-\frac{\rho^2 \mathcal{MM}_i^2}{2}\right) , \tag{4.7}$$

   where ρ corresponds to the signal-to-noise ratio (SNR) in a parameter estimation setting. Higher SNR signals result in posteriors with mismatches more tightly clustered around 0. Given these weights, we perform rejection sampling to retain points with low mismatch.

4. **Fitting**: We define the correlation direction $\boldsymbol{d}_{\rm cor}$ (which has the same dimensionality as $\boldsymbol{x}$) as the direction of maximum variance–equivalently, maximum eigenvalue, i.e., mismatch values remain close to 0. The direction $\boldsymbol{d}_{\rm cor}$ and its eigenvalue $\lambda_{\rm cor}$ are computed with a Gaussian fit; see Appendix 2 of Ref. [150] for details, where we also confirm that we obtain similar results with a principal component analysis.

   To compare the correlation strength between different paths, we use the ratio of the semi-major to the semi-minor axis of the 1-σ contour, r. The light blue ellipse in Fig. 4.2 represents this contour. The correlation strength is the square root of the condition number of the Gaussian covariance matrix, defined as the ratio of its largest and its smallest eigenvalue. We caution against over-interpreting the exact numerical value of r: it should be viewed as an order-of-magnitude estimate of the correlation strength, i.e., $r \sim \mathcal{O}(100)$ represents a strong correlation, $r \sim \mathcal{O}(10)$ is moderate, while $r \sim \mathcal{O}(1)$ is weak. The specific value of r also depends on the dimensionality of the parameter-space in which correlations are measured.

5. **Stepping**: We define the next reference point as

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \frac{\alpha \lambda_{\rm cor}}{\sqrt{D}} \boldsymbol{d}_{\rm cor} \ , \tag{4.8}$$

where α is a dimensionless free parameter that controls the size of the steps, and D is the dimensionality of $\boldsymbol{x}$.

6. **Repeat** from #1 and in the opposite direction until a parameter space boundary is reached. The list of points $\boldsymbol{x}_j$ map the correlation direction for the initial reference point $\boldsymbol{x}_0$.

We tune free parameters to ensure that correlation directions are reliably recovered. Specifically, the total number of points N in $\mathcal{PS}$ and the SNR ρ are determined empirically such that we have a sufficient number of samples remaining after rejection sampling. We typically use $N = 100,000$, while ρ is increased with the total mass and the number of dimensions D to ensure at least 1% efficiency in rejection sampling.[5] The choice of the step size α depends on the bounds of the parameter space we are mapping. We typically take $\alpha = 1$ in the effective space and $\alpha = 0.1$ for the aligned and precessing spaces.

The correlation mapping process is looking for *paths*, i.e., 1-dimensional lines, across the parameter space. Network uncertainties may broaden the recovered correlations but do not impact their direction; Appendix B in Ref. [150] shows percent level errors in network prediction, while the correlations we are recovering below are at the $\mathcal{O}(10\%)$. Since we restrict to a relatively small number of parameters (up to 3 for the aligned parameter space), this process still recovers non-trivial structure. Exploring correlations with more parameters (for example all 8 intrinsic degrees of freedom) can be achieved by generalizing the above method to higher-dimensional correlation hypersurfaces. Appendix C of Ref. [150] further discusses this issue for the 3D aligned spin space.

4.3 Recovering Spin-Aligned Inspiral correlations at $30\,M_\odot$

We begin by exploring correlations in systems with a total mass of $30\,M_\odot$ where a major portion of the observed signal corresponds to the inspiral, c.f. Fig. 4.1. Analytical knowledge from PN equations suggests that an appropriate spin combination for the aligned spin dynamics is $\chi_{\rm eff}$, e.g., Refs. [188, 191]. We therefore start by mapping correlations in the 2D space of $\boldsymbol{x}^{(\mathbf{eff})} = [\eta, \chi_{\rm eff}]$. Figure 4.3 demonstrates how the correlation mapping algorithm of Sec. 4.2

[5] The SNR cut focuses sampling on the low-mismatch region and provides a more robust and consistent selection than a hard mismatch threshold, which would otherwise require arbitrary initial configuration and mass-dependent tuning.

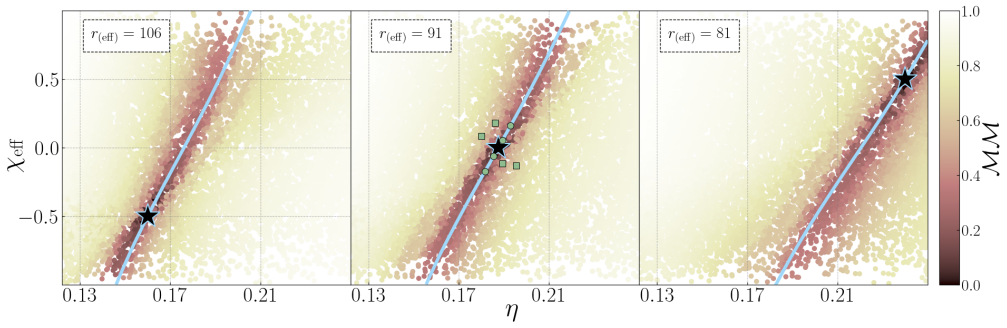


Figure 4.3: Correlation path (blue line) in $[\eta, \chi_{\mathrm{eff}}]$ for binaries with a total mass of $30\,M_\odot$ and different initial reference points: $(\eta, \chi_{\mathrm{eff}}) = (0.16, -0.5), (0.19, 0), (0.24, 0.5)$ (left to right, star). The colormap indicates the mismatch between the initial reference point and samples in the parameter space. The correlation is visible by eye in the colormap and the computed paths track it consistently. The correlation strength $r_{(\mathrm{eff})}$, corresponding to the ratio of the semi-major to the semi-minor axis of the 1-σ contour is given in-plot, confirming that the right-most system has the weakest correlation, as also evident from the colormap. Green circles and boxes in the middle panel denote the waveforms plotted in Fig. 4.4.

recovers correlations in the $\chi_{\mathrm{eff}} - \eta$ space. Each panel corresponds to a different initial reference point $\boldsymbol{x}^{(\mathbf{eff})}{}_0$. We present 3 systems with $(\eta, \chi_{\mathrm{eff}}) = (0.16, -0.5), (0.19, 0), (0.24, 0.5)$ (mass ratios of 4, 3, and 1.5 respectively), but obtain similar results for other configurations. The colormap denotes the mismatch across the parameter space with respect to the *initial* reference point (star), where the white/yellow region indicates a region of high mismatch (dissimilar waveforms). In all cases the recovered correlation path (light blue) agrees well with the region of low mismatches (maroon). Moreover, regions of low mismatch are tightly concentrated around the correlation path, showing that the correlation itself is strong. For Fig. 4.3, the correlation strength, i.e., the ratio of the semi-major to the semi-minor axis of the 1-σ contour $r_{(\mathrm{eff})}$ is 106, 91, 81 from left to right. These numbers confirm the by-eye observation that the correlation is the least tight in the right panel, albeit not significantly.

A waveform demonstration of the correlation is provided in Fig. 4.4 for the reference point with $(\eta, \chi_{\mathrm{eff}}) = (0.19, 0)$. The top panel shows waveforms along the correlation path, while the bottom panel shows waveforms outside the correlation, all labeled by their mismatch and denoted in Fig. 4.3 with green dots and boxes respectively. In the top panel, waveforms are by-eye

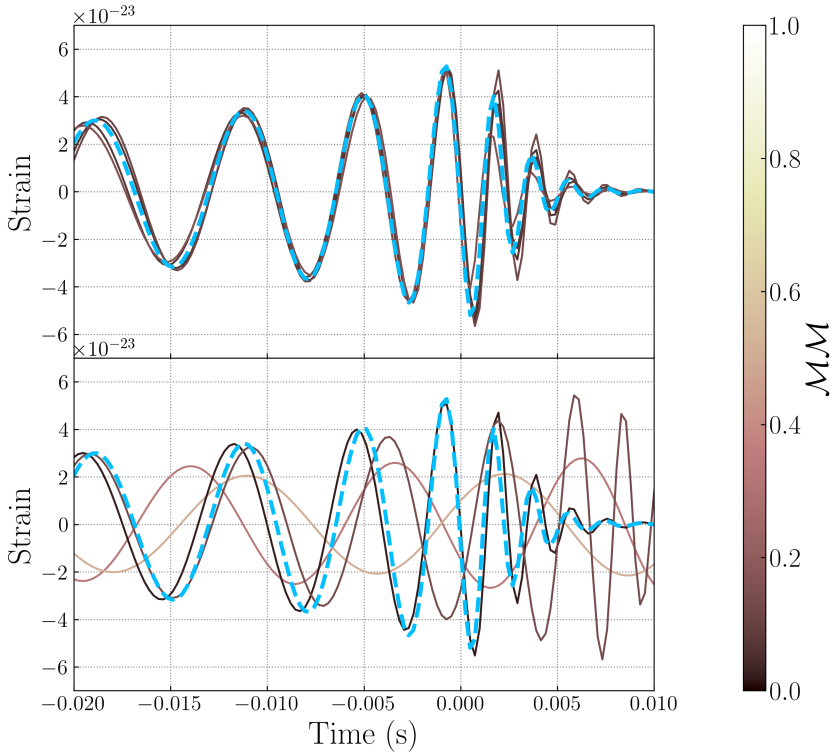


Figure 4.4: IMRPhenomXPHM waveforms demonstrating the $\chi_{\rm eff} - \eta$ correlation for the reference point $(\eta, \chi_{\rm eff}) = (0.19, 0)$ from Fig. 4.3. The blue dashed line is the reference point waveform. The top panel shows waveforms along the correlation path of the reference point, colored by their mismatch with respect to the reference (green circles in Fig. 4.3). These waveforms are visually similar and all have small mismatches with the reference waveform. The bottom panel shows random waveforms near the reference injection (green boxes in Fig. 4.3). We plot the waveforms with the minimum mismatch, i.e., minimized over time and phase shifts, with respect to the reference.

similar to the reference waveform as expected from the low mismatch. Our analysis leverages this similarity to identify a correlation between $\chi_{\rm eff}$ and η.

Having validated the correlation mapping algorithm in the effective space, we consider the more complex scenario in which we pretend that *a priori* we do not know which aligned spin combination (i.e., $\chi_{\rm eff}$) is most relevant. We therefore consider the full 3D aligned spin space with $\boldsymbol{x}^{(\rm al)} = [q, \chi_{1z}, \chi_{2z}]$. Figure 4.5 shows the correlation paths in $[q, \chi_{1z}, \chi_{2z}]$, mapped directly in the 3D aligned spin space. The smoothness of the paths indicates that a well-defined correlation does exist, though from this projection the lines do not appear to align with the darkest (smallest) mismatches, see App. C of Ref. [150] for an explanation. The recovered correlation has a reduced correlation strength

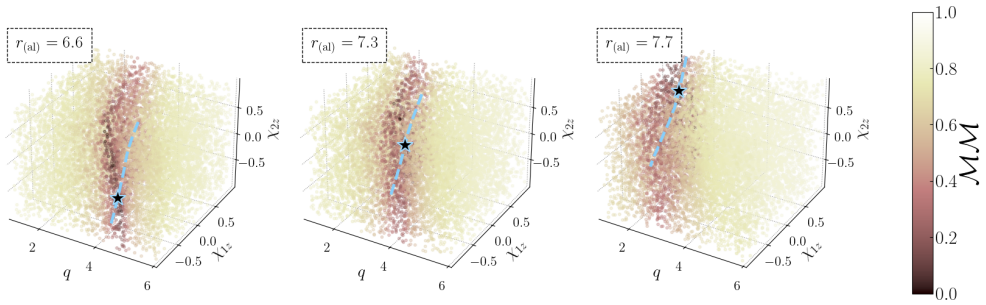


Figure 4.5: Correlation paths in $[q, \chi_{1z}, \chi_{2z}]$ for binaries with a total mass of $30\,M_\odot$ and the same initial reference points (stars) as Fig. 4.3: $(q, \chi_{1z}, \chi_{2z}) = (4, -0.5, -0.5), (3, 0, 0), (1.5, 0.5, 0.5)$. Correlations are directly mapped in the 3D aligned spin space of $\boldsymbol{x}^{(\mathbf{al})} = [q, \chi_{1z}, \chi_{2z}]$. A coherent path is tracked across the parameter space, even though it has a higher dimensionality than strictly necessary. As a consequence the correlation strengths are lower than those of Fig. 4.3.

of 6.6, 7.3, 7.7 from left to right. The left panel of Fig. 4.9 (discussed more later) shows the correlation paths projected into the 2D subspaces formed from $[q, \chi_{1z}, \chi_{2z}]$. *If we had no knowledge of the analytical form of* χ_{eff}*, these correlation paths could aid us in motivating its formula.* For example, as the correlation paths are smoother in $[q, \chi_{1z}]$ than the subspaces including χ_{2z}, we would know that our effective spin parameter would more strongly depend on χ_{1z} than χ_{2z}. This is true for χ_{eff}: the primary spin is upweighted by q; see Eq. (4.1).

Figure 4.6 compares the correlation paths in $[\chi_{\mathrm{eff}}, \eta]$ computed in different ways: 1. directly from the 2D effective space $\boldsymbol{x}^{(\mathbf{eff})} = [\eta, \chi_{\mathrm{eff}}]$, and 2. from the 3D aligned spin space $\boldsymbol{x}^{(\mathbf{al})} = [q, \chi_{1z}, \chi_{2z}]$ and then *projected* to $[\chi_{\mathrm{eff}}, \eta]$. For each reference system, the effective ($\boldsymbol{x}^{(\mathbf{eff})}$; solid lines) and aligned ($\boldsymbol{x}^{(\mathbf{al})}$; dashed lines) correlation mappings return consistent paths when projected to 2D. Even when mapping a correlation in more dimensions than strictly necessary, our algorithm can recover equivalent structure.

The morphology of the correlation we recover, c.f., Fig. 4.3, can be explained as follows. For binaries with fixed total mass, both a larger χ_{eff} and more unequal masses increase the length of the waveform. The former is because the binary has more angular momentum and energy to emit before merger, and the latter because energy emission is weaker. Therefore the two are positively correlated: increasing both together leads to their effect approximately canceling out and

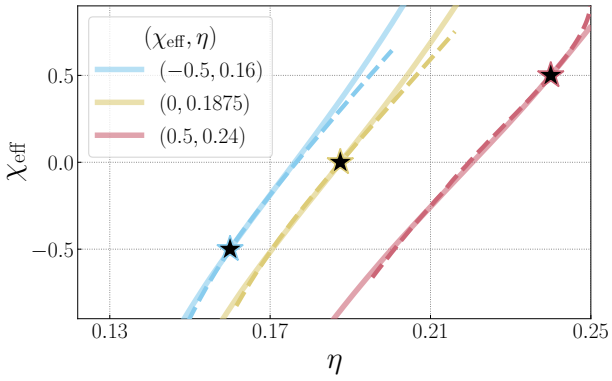


Figure 4.6: Correlation paths in $[\eta, \chi_{\rm eff}]$ for binaries with a total mass of $30\,M_\odot$ and the same initial reference points (stars) as Fig. 4.3. We plot the 2D effective space of $\boldsymbol{x}^{(\mathbf{eff})} = [\eta, \chi_{\rm eff}]$ (solid; same as Fig. 4.3) and the 3D aligned space of $\boldsymbol{x}^{(\mathbf{al})} = [q, \chi_{1z}, \chi_{2z}]$ (dashed; same as Fig. 4.5) mapping that has then been projected to the $[\eta, \chi_{\rm eff}]$ space. Correlation paths are consistent, thus validating the higher-dimensional correlation mapping.

the waveform remains approximately unchanged. The result of Fig. 4.3 would have aided us in deducing this physical imprint of $\chi_{\rm eff}$ and η on the waveform, if we did not already know about it through PN equations. Notably, this is not the same as the "inspiral spin-mass ratio correlation" [13, 187, 191, 300, 301], which instead leads to *anti*-correlated $\chi_{\rm eff}$ and η. That correlation is instead driven by the functional form of the 1.5PN phase term. The compounding effect on the waveform length of changing $\chi_{\rm eff}$ and η in an anti-correlated fashion is counteracted by a change in the total mass. Fixing the total mass instead completely alters the shape of the correlation. Preliminary results with an expanded network that varies the total mass suggest that our methodology can recover the spin-mass ratio correlation of Ref. [187]; we leave further such explorations to future work.

4.4 Aligned-spin correlations beyond the inspiral

As a binary approaches merger, the orbital velocity increases and the PN expansion becomes inaccurate. Late inspiral and merger dynamics alter the signal evolution and change the spin correlations. In this section, we explore spin correlations beyond inspiral-dominated signals. We consider binaries with a total mass of $90\,M_\odot$ whose observable signal includes both the late inspiral and the merger, and $270\,M_\odot$ whose observable signal is overwhelmingly dominated by the merger. Figure 4.1 shows example signals for reference.

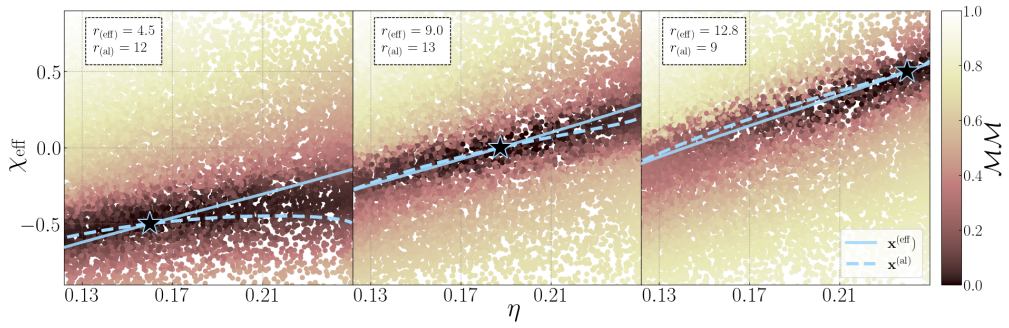


Figure 4.7: Similar to Fig. 4.3 but for the $90\,M_\odot$ network and signals where both the inspiral and the merger phase are observable. We also include in dashed the correlation paths from the 3D aligned space of $\boldsymbol{x}^{(\mathbf{al})} = [q, \chi_{1z}, \chi_{2z}]$ mapping that have then been projected to the $[\eta, \chi_{\mathrm{eff}}]$ space. The correlation paths are generally consistent, suggesting that χ_{eff} is still an appropriate aligned spin parameter at this total mass. The effective mapping correlation strength numbers are lower than those of Fig. 4.3, reflecting the weakening of the $\chi_{\mathrm{eff}}-\eta$ correlation for heavier systems. The aligned spin mapping numbers are higher than the effective mapping ones, likely suggesting that a better spin-aligned parameter than χ_{eff} might exist.

With the PN expansion breaking down close to merger, it is no longer clear whether χ_{eff} is the appropriate spin aligned parameter. However, as a first approximation it is reasonable to explore whether a correlation in $\chi_{\mathrm{eff}}-\eta$ is still present, albeit in a weakened and/or altered form. Figure 4.7 shows correlations recovered in the 2D effective space of $\boldsymbol{x}^{(\mathbf{eff})} = [\eta, \chi_{\mathrm{eff}}]$ with the $90\,M_\odot$ network. Firstly, some degree of correlation between χ_{eff} and η exists: the colormap shows a clear structure and the mapping algorithm recovers coherent correlation paths. Secondly, compared to the equivalent Fig. 4.3 for $30\,M_\odot$ signals, the correlation here is much *weaker*, evident by the fact that low mismatches are spread out over a larger region around the correlation path (solid lines). Correspondingly, the correlation strength reduces to 4.5, 9.0, 12.8 from left to right. Third, and again comparing to Fig. 4.3, the correlation structure is *altered*, i.e., has a different "slope" in the $\chi_{\mathrm{eff}}-\eta$ plane.

The altered and weakened $\chi_{\mathrm{eff}} - \eta$ correlation for $90\,M_\odot$ signals motivates exploring correlations in the 3D aligned space of $\boldsymbol{x}^{(\mathbf{al})} = [q, \chi_{1z}, \chi_{2z}]$. The correlation in the $\boldsymbol{x}^{(\mathbf{al})}$ space projected to $[\eta, \chi_{\mathrm{eff}}]$ is included in Fig. 4.7 (dashed lines). The two paths, computed from the 3D aligned and the 2D effective spaces, are largely consistent, especially in the vicinity of the reference point.

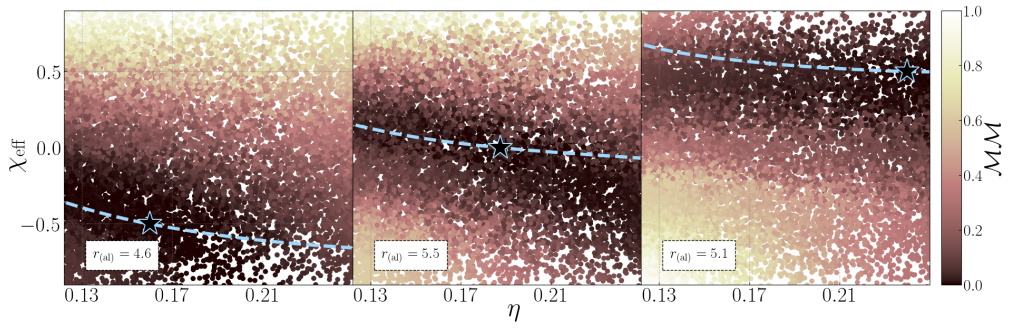


Figure 4.8: Similar to Fig. 4.7 but for the 270 $M_\odot$ network and signals where only the merger phase is observable. In contrast to Fig. 4.7, we only include the correlation paths from the 3D aligned space of $\boldsymbol{x}^{\rm (al)} = [q, \chi_{1z}, \chi_{2z}]$ mapping that have then been projected to the $[\eta, \chi_{\rm eff}]$ space. The correlation strength is lower than Fig. 4.7, showing only weak trends in $[\eta, \chi_{\rm eff}]$ for merger-dominated systems, and works in the opposite direction: $\chi_{\rm eff}$ and η are here *anti-correlated*, albeit weakly.

Differences emerge away from the reference point with the 3D paths being more curved, likely due to the larger flexibility of an extra dimension. This discrepancy is the largest for the negative $\chi_{\rm eff}$ unequal-mass test case (left panel). Overall, $\chi_{\rm eff}$ is still a relevant aligned spin parameter for systems with a total mass of 90 $M_\odot$. The 2D effective mapping correlation strength numbers are lower than those of Fig. 4.3 for 30 $M_\odot$ systems, reflecting the weakening of the $\chi_{\rm eff}-\eta$ correlation that is also visible in the colormaps. However, contrary to the 30 $M_\odot$ case, for 90 $M_\odot$ binaries the 3D aligned spin correlation has a comparable (or in same cases larger) strength than the 2D effective one ($r_{\rm (al)} > r_{\rm (eff)}$). This likely indicates that although $\chi_{\rm eff}$ is an appropriate aligned spin parameter for signals without much visible inspiral, there might exist an even better one.

The situation is different for the 270 $M_\odot$ signals. Initial comparisons between the 2D effective and 3D aligned spaces revealed very different correlation paths. As such, we only present 3D results, which are more generic. Figure 4.8 shows results in the $[\eta, \chi_{\rm eff}]$ space with correlation paths computed in the 3D aligned space and then projected. By eye the correlation seems to be almost completely gone with little structure apparent in the colormap. The correlation path still roughly follows the region of low mismatch with a marginally negative slope. The large extent of the low-mismatch region is reflected on the correlation strength: we report $r_{\rm (al)} = 4.6$, 5.5, 5.1 compared to 12, 13, 9 for systems with

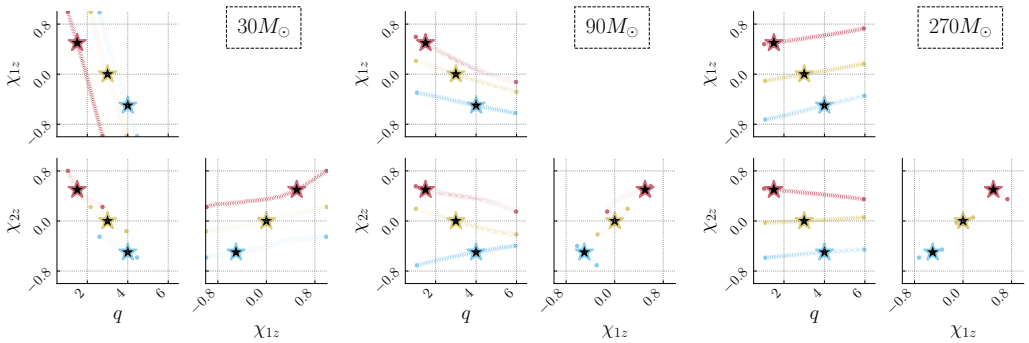


Figure 4.9: Correlation paths that have been constructed in the 3D aligned spin space of $[q, \chi_{1z}, \chi_{2z}]$ and then projected in the 2D subspaces for the 30 $M_\odot$ (left), 90 $M_\odot$ (middle), and 270 $M_\odot$ (right) networks. We show the same initial reference points as Fig. 4.5, indicated by different colors.

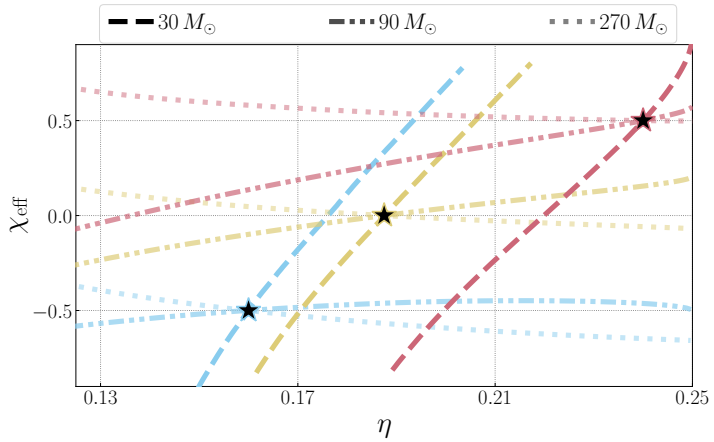


Figure 4.10: Comparison of the $\chi_{\rm eff} - \eta$ correlation for signals with total mass 30 $M_\odot$, 90 $M_\odot$, 270 $M_\odot$ (darker to lighter shades of the same color for each reference point). We show results for the same three reference points as Fig. 4.6 but only include the 3D correlation paths computed from the aligned space of $\boldsymbol{x}^{(\mathbf{al})} = [q, \chi_{1z}, \chi_{2z}]$. The slope of the correlation dramatically decreases as the total mass increases. Fits to the correlation paths are provided in the text.

identical spins and mass ratio, but a total mass of 90 $M_\odot$.

Having recovered spin-aligned correlations with the different networks, we now turn to cross-network comparisons. Such comparisons reveal how correlations change as the observed signal becomes increasingly merger-dominated. Figure 4.9 shows correlation paths recovered in the 3D aligned spin space and then projected into the 2D subspaces formed from $[q, \chi_{1z}, \chi_{2z}]$ for all three different total mass test cases. Figure 4.10 further projects down to the $[\eta, \chi_{\rm eff}]$ space and demonstrates how the $\chi_{\rm eff}-\eta$ correlation is altered as the signal's

total mass increases.

For $90\,M_\odot$ and $30\,M_\odot$ signals, the slope of the correlation path in $\chi_{\rm eff} - \eta$ is positive: a larger aligned spin is degenerate with more equal masses. We interpret the sign of the correlation again through the effect of aligned spin and mass ratio on the signal length. The slope of the correlation paths decreases with increasing total mass from $30\,M_\odot$ to $90\,M_\odot$. For signals where more of the merger is observable, a change in the mass ratio can be "counteracted" by a smaller change in the aligned spin. Stated differently, at a fixed total mass, the aligned spin is better measured than the mass ratio for $90\,M_\odot$ signals, while the opposite is true for $30\,M_\odot$ ones. This can be seen in Fig. 4.10: due to the correlations' "steep slope," the paths from the $30\,M_\odot$ signals span the full $\chi_{\rm eff}$ regime of -1 to 1, but only a subset of the allowed η values; the inverse is true for the $90\,M_\odot$ case.

At $270\,M_\odot$, the correlation is mostly flat with a small *negative* slope, i.e., larger aligned spins correspond to more unequal masses, the inverse of the correlation direction found for the lower masses cases. However, the trend is weak, suggesting that at a fixed total mass, the aligned spin is well-measured, while the mass ratio remains unmeasurable. We interpret this as follows: when the inspiral is no longer observable, $\chi_{\rm eff}$ is no longer obtained from the inspiral length, but from the merger frequency. If the total mass, which also affects the merger frequency, is fixed, $\chi_{\rm eff}$ can be measured accurately. The mass ratio, on the other hand, has a subdominant effect on the merger frequency and is thus more challenging to measure without seeing the inspiral.

Finally, we fit the $\chi_{\rm eff}-\eta$ correlation paths of Fig. 4.10 with the ansatz

$$\begin{aligned}\chi_{\rm eff} = &(a_2 M^2 + a_1 M + a_0)\eta^2 \\ &+ (b_2 M^2 + b_1 M + b_0)\eta \\ &+ (c_2 M^2 + c_1 M + c_0)\,. \end{aligned} \tag{4.9}$$

The fitting formula allows for a quadratic relation between $\chi_{\rm eff}$ and η where the coefficients are themselves quadratic functions of the total mass M. We obtain

$$\begin{aligned} a_2 &= -5.6\times10^{-3}, \quad a_1 = 2.2, \quad a_0 = -1.7\times10^{2}, \\ b_2 &= 3.9\times10^{-3}, \quad b_1 = -1.5, \quad b_0 = 1.1\times10^{2}, \\ c_2 &= -4.9\times10^{-4}, \quad c_1 = 0.2, \quad c_0 = -14\,, \end{aligned}$$

for the reference system $(q, \chi_{1z}, \chi_{2z}) = (4, -0.5, -0.5)$. Other systems can be approximately fit by shifting this fit vertically by an amount that depends on the new reference value. Though this fit is based on a handful of reference systems and values of the total mass,[6] it represents a preliminary quantitative (albeit phenomenological) description of the imprint of spins on GWs across total masses. Generalizing the fits to more dimensions and reference points is left for future work.

4.5 Precessing degrees of freedom

We now turn to the precessing parameter space, $\boldsymbol{x}^{(\mathbf{pre})}$. As a reminder, we here set $\vec{\chi}_2$ and ϕ_1, in addition to the aligned components, to zero. Under this restriction, the effective precessing parameter $\chi_{\rm p}$ reduces to the primary spin magnitude χ_1, which is the same as χ_{1x}; see Eq. (4.2). Figure 4.11 compares precessing correlations for two unequal-mass reference systems across total masses: one with near-maximal χ_1 (top row) and one with a smaller value (bottom row). The mismatches (whose plotted colormap ranges have been restricted for readability) show clear structure for the $30\,M_\odot$ and $90\,M_\odot$ cases, but little structure for $270\,M_\odot$. The latter is consistent with the fact that precession can only be measured for specific, fine-tuned configurations for merger-dominated signals [103].[7] As such, for the rest of the section we only consider $30\,M_\odot$ and $90\,M_\odot$ binaries.

We map correlations in the 2D precessing space of $\boldsymbol{x}^{(\mathbf{pre})} = [\eta, \chi_1]$. For $30\,M_\odot$, the correlation strength is $r_{(\rm pre)} = \mathcal{O}(100)$.[8] Given that the correlation path is almost vertical in the χ_1–η plane, this large r value reflects that the mass ratio can be measured accurately. The slight tilt of the $30\,M_\odot$ correlation path indicates, however, that there is some degree of correlation between the in-plane spin and the mass ratio: either drastically decreasing the in-plane spin magnitude or making the masses slightly more unequal both alter the waveform in similar ways. The $90\,M_\odot$ correlation path tilts further from the vertical and the correlation strength reduces to $r_{(\rm pre)} = \mathcal{O}(10)$. Moreover,

[6] For this reason we also do not comment on its accuracy. As we are fitting three values of the total mass with three parameters, the fit's error or residual might be deceptive.

[7] Additionally, over the plotted parameter space, mismatches vary by ∼0.1 which is within a factor of 10 of the expected network error.

[8] The precessing correlation strength $r_{(\rm pre)}$ is not comparable to the effective aligned correlation $r_{(\rm eff)}$ presented in Sec. 4.3. Even though both refer to correlations in the same number of dimensions, they vary or set to zero different parameters.

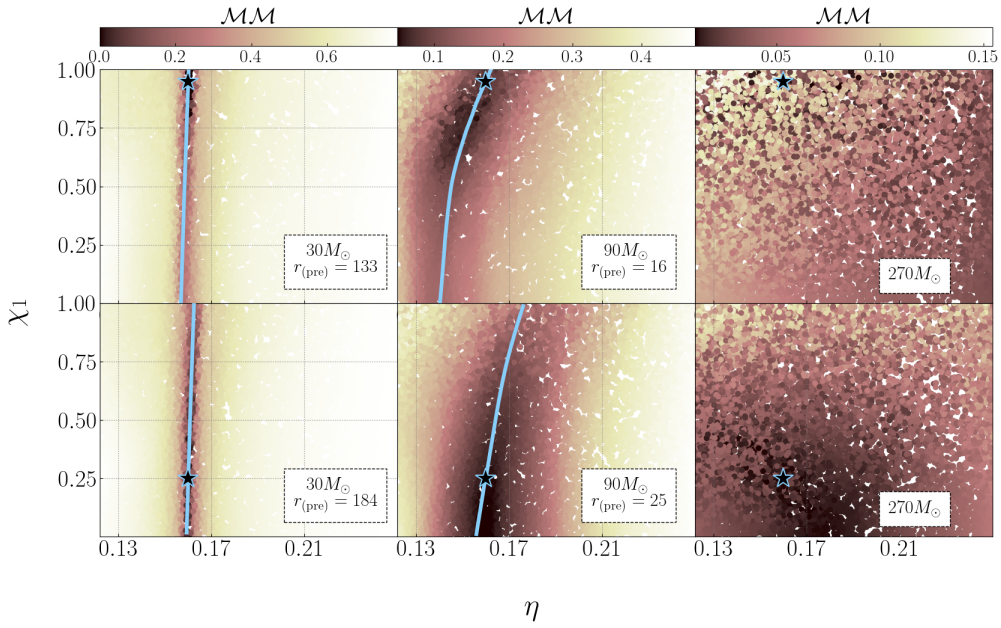


Figure 4.11: Comparison of precessing correlations for systems with a total mass of $30\,M_\odot$ (left), $90\,M_\odot$ (middle), and $270\,M_\odot$ (right) for a reference point with $(\eta, \chi_1) = (0.16, 0.95)$ (top) and $(0.16, 0.25)$ (bottom). The colormap denotes the mismatch (spanning a reduced range compared to previous results for readability) with respect to the reference point (star). Correlations are mapped in $\boldsymbol{x}^{(\mathbf{pre})} = [\eta, \chi_1]$ with the secondary spin set to zero and the primary spin along the binary x-axis. Blue lines are the correlation path. The correlation strengths are given in-plot for the region of the reference point; the correlation strength varies along the correlation path. No correlation path or strength is provided for the $270\,M_\odot$ case as no structure exists in the mismatches. Similar to the aligned-spin case, both the strength and the shape of the correlation between in-plane spins and mass ratio changes with the binary total mass.

the correlation strength changes along the correlation path, becoming weaker (lower mismatches on average) away from the reference point.

Similar to the aligned-spin results of Secs. 4.3 and 4.4, the correlation between the in-plane spin and the mass ratio both changes in shape and reduces in strength as the total mass increases. Put differently, as the total mass increases, the impact of mass ratio on the waveform morphology diminishes, making η harder to measure. Since χ_1 is comparably constrained at $30\,M_\odot$ and $90\,M_\odot$, this increased uncertainty on η makes the over-all correlation *strength* weaker. However it also reveals more correlational *structure* in the χ_1–η plane: a smaller change in χ_1 is now sufficient to mimic the same

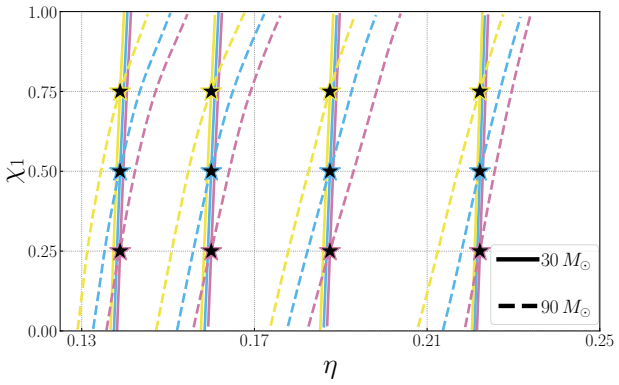


Figure 4.12: Comparison of the precessing correlation for signals with total mass $30\,M_\odot$ (solid) and $90\,M_\odot$ (dashed). We show results for various reference points (stars) and correlations mapped in $\boldsymbol{x}^{(\mathbf{pre})} = [\eta, \chi_1]$. Colors indicate the value of χ_1 at reference, with pink, blue, and yellow representing $\chi_1 = 0.25, 0.5, 0.75$ respectively. From left to right, we consider $\eta = 0.14, 0.16, 0.19, 0.22$ (equivalently, $q = 5, 4, 3, 2$). The change in correlation shape shown in Fig. 4.11 is a generic feature across multiple reference points.

variation in η. The shape of the correlation across more reference points is explored in Fig. 4.12. We plot correlation paths for different reference points (stars and different colors) and values of the total mass (solid vs dashed lines).[9] Consistent, with Fig. 4.11, the correlation paths for inspiral-dominated $30\,M_\odot$ systems are mostly vertical, while the tilt of the path increases at $90\,M_\odot$ for all reference points. Our results suggest that when less inspiral is observable, an appropriate combination of η and χ_1 should be better measured than either parameter alone.

We explore the underlying cause of thesecorrelations by looking directly at waveform morphology in the late inspiral and merger. Figure 4.13 shows $90\,M_\odot$ waveforms with fixed versus varied η and χ_1. Empirically, both η and χ_1 subtly change the waveform phase evolution and alter the structure of the ringdown. At a fixed χ_1, increasing η (making the masses more equal; lighter to darker colors) slightly *increases* the merger frequency (top panel). To identify this, draw your eye to the peak cycle where waveforms are aligned (between -0.01 and 0 s), and compare these to the location of the preceding peaks/troughs. A lower merger frequency corresponds to more time between these local ex-

[9] Since we only have results for two values of the total mass, we do not present fits of the paths of Fig. 4.12, leaving them to future extensions.

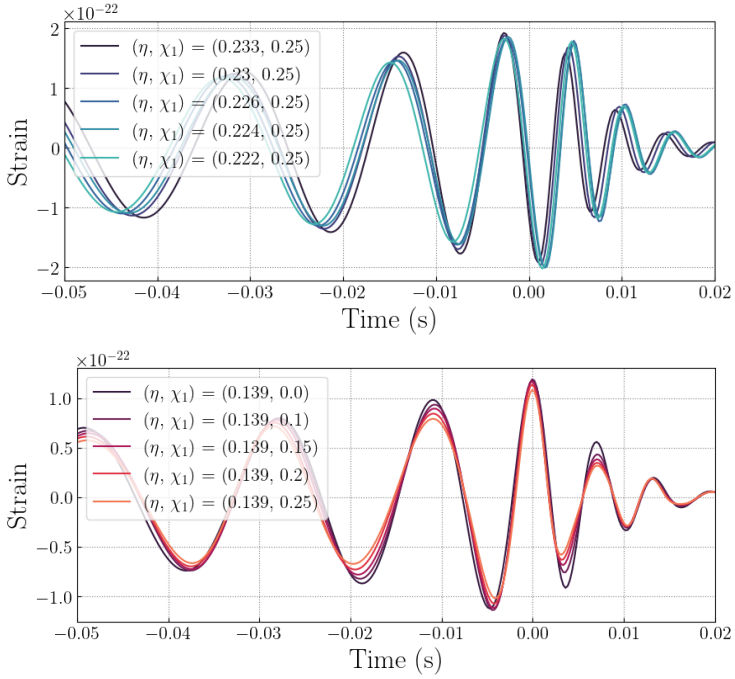


Figure 4.13: A demonstrative example of how small variations in η and in-plane χ_1 effect signal morphology in the late inspiral and merger. We show sets of waveforms with fixed $\chi_1 = 0.25$ and varied η (top), compared to fixed $\eta = 0.139$ ($q = 5$) and varied χ_1 (bottom). All waveforms are generated using NRSur7dq4 with a total mass of $90\,M_\odot$ and all spin degrees of freedom set to 0 aside from the primary's in-plane spin along the x-axis.

trema. At a fixed mass ratio, increasing χ_1 (darker to lighter colors) slightly *decreases* the merger frequency (bottom panel). This effect is stronger for more unequal masses. Combining these observations suggests that increasing both parameters along the line of correlation in the χ_1–η plane leaves the merger frequency essentially unchanged. In agreement with Fig. 4.12, a small change in η affects the waveform much more than a corresponding change in χ_1. Given that precession-driven amplitude fluctuations [312] are often unobservable in short, merger-dominated signals, the merger frequency shift could become a primary precession observable. The question of how exactly precession is measured in the merger remains open, but observations like this offer a qualitative understanding of the link between parameter correlations and waveform morphology. Correlation mapping is a systematic approach to identifying and understanding these relations.

4.6 Conclusions

In this study, we proposed a waveform-based approach to study the imprint of spin degrees of freedom on GWs emitted by BBHs. Our approach does not require analytic expressions for the binary dynamics; it is instead based on the availability of forward models for the emitted signal given the binary parameters. It is therefore applicable even to the merger for which solutions can only be obtained via full numerical relativity simulations. The method is based on identifying parameter correlations, i.e., paths in parameter space that leave the signal approximately unaltered.

Rather than focusing on *a single point in the parameter space*, as in traditional parameter estimation, our framework traces *correlation paths* across a deliberately restricted subset of the intrinsic parameter space. This simplification provides a systematic way to map spin correlations across broad regions of parameter-space and study merger dynamics without relying on analytic models. Studying the spin aligned and precessing degrees of freedom for inspiral-dominated (total mass $30\,M_\odot$), late-inspiral-plus-merger ($90\,M_\odot$), and merger-dominated ($270\,M_\odot$) signals, c.f. Fig. 4.1, we find the following.

- Mapping correlations in the generic 3D aligned space of $\boldsymbol{x}^{(\mathbf{al})} = [q, \chi_{1z}, \chi_{2z}]$ can reveal the known lower-dimensional correlations of the 2D effective space of $\boldsymbol{x}^{(\mathbf{eff})} = [\eta, \chi_{\rm eff}]$ for inspiral-dominated signals; see Fig. 4.6.

- The effective aligned spin $\chi_{\rm eff}$ captures the aligned spin dynamics for systems up to $\sim 90\,M_\odot$. However, at $90\,M_\odot$, the aligned spin-mass ratio correlation is comparable or stronger in the 3D aligned space of $\boldsymbol{x}^{(\mathbf{al})} = [q, \chi_{1z}, \chi_{2z}]$ than the 2D effective space of $\boldsymbol{x}^{(\mathbf{eff})} = [\chi_{\rm eff}, \eta]$; see Fig. 4.7. This suggests that a more appropriate parameter might exist.

- For systems with a total mass of $270\,M_\odot$, correlation mapping using the full aligned spin degrees of freedom does not yield the same results as working directly with $\chi_{\rm eff}$; see Fig. 4.8. This strongly suggests that $\chi_{\rm eff}$ is a suboptimal aligned effective spin for merger-dominated signals [197, 313].

- As the binary total mass increases, the correlation between the the aligned spins and the mass ratio becomes less steep, see Figs. 4.9 and 4.10. The corresponding fit of Eq. (4.9) is a phenomenological description of aligned spin dynamics that is applicable even for systems with a (detector frame) total mass of $270\,M_\odot$ and for which only the merger is observable.

- Similarly to the aligned-spin case, the correlation between the in-plane spin and the mass ratio becomes less steep and reduces in strength as the total mass increases. The correlation has effectively disappeared by a total mass of $270\,M_\odot$, supporting known difficulties of measuring spin precession in merger-dominated signals [78, 98, 103, 104, 314].

The correlation mapping procedure can be extended along various directions. Firstly, though we explore results at different values of the total mass, its value remains fixed within each mapping. This restriction ignores correlations between the total mass and other parameters. Preliminary results with a neural network that includes the total mass are promising in that they can reproduce the spin-mass ratio correlation known from parameter estimation, see Sec. 4.3. However, technical challenges remain related to the boundaries in the mass parameter space where we need to switch between the two waveform models considered, IMRPhenomXPHM and NRSur7dq4. Secondly, with the total mass varying, it would be preferable to switch to defining the spin directions at a fixed point in the waveform, e.g., $t = -100\,M$ rather than a fixed frequency [48]. Thirdly, in this study we only consider the intrinsic degrees of freedom. The measurability of spins also depends on extrinsic binary parameters [201]; for example, χ_p and the binary's orientation in the sky have known observational degeneracies, e.g., [102, 301], making correlations between spins and extrinsic degrees of freedom interesting to consider, particularly the aforementioned effects of inclination-dependent mode mixing in precessing systems. Fourthly, we are looking for correlation *paths*. As we look for correlations in higher dimensions, these should be correspondingly generalized to higher-dimensional correlations, e.g., correlation *surfaces*. Finally, Eq. (4.9) provides a first quantitative result on the imprint of spins applicable even to merger-dominated signals, but it is still restricted to χ_eff and η. Higher dimensional fits of the 3D aligned spin parameters or the precessing parameters could motivate better effective parameters. In particular, we aim to derive better-measured spin parameters for high-mass signals–especially to capture the precessing degrees of freedom.

The results of Figs. 4.7 and 4.8 indeed suggest that a different effective aligned spin parameter than χ_eff might be more appropriate for merger-dominated signals. Studies of the ringdown and final-state of spin-aligned systems show that post-merger properties and phenomenology can be well-captured by functions

of the pre-merger masses and spins [197, 223, 227]. This implies that combinations of mass ratio and aligned spin that lead to the same *remnant* spin may, at a fixed total mass, be a better measured parameter combination for merger/ringdown dominated signals. Additionally, studies of the final-state of waveforms in the IMRPhenom family show that the late inspiral and merger are better characterized by an effective *total* spin parameter [313]. Even in the inspiral phase, there are effective-aligned spin parameters that may capture dynamics better than $\chi_{\rm eff}$; e.g., a normalized version of the exact 1.5 PN coefficient, which directly contributes to the signals' phase evolution [313, 315]. We leave exploring these alternate parametrizations, and extending beyond them, to future work.

Insights from correlations paths can aid in waveform modeling through identification of further effective spin parameters like $\chi_{\rm eff}$, $\chi_{\rm p}$, and others [204, 205, 313, 315]. Using fits like Eq. (4.9) (after extending to more values for the total mass) can lead to phenomenological definition of effective parameters that are appropriate for heavy BBHs. Such effective parameters can improve the ansatz for waveforms calibration to numerical relativity simulations [245] or help simplify the dynamics [316]. Additionally, better parameters to describe the late inspiral and merger regime can lead to more accurate predictions for remnant BH properties [204]. Finally, while we have focused on correlation paths via the covariance eigendirection with the largest eigenvalue, the full matrix contains information about the parameter structure. Specifically, studying the orthogonal directions along the correlation path would indicate where the mismatch grows most rapidly and thus point to dominant modeling effects.

Correlation paths can also help interpret and improve parameter estimation results. First, Fig. 4.10 shows how the $\chi_{\rm eff}$-η degeneracy changes at high mass. This is echoed in parameter estimation results, c.f., Fig. 5 of Ref. [26] where low-mass binaries have a steeper slope in the $\chi_{\rm eff}$-q plane than the high-mass ones. Such "consistency with expectations" tests can help us safeguard against biases: a posterior that does *not* follow these known paths could indicate that the signal may be contaminated by non-Gaussian noise [317] or that waveform systematics are at play. Second, $\chi_{\rm eff}$ and $\chi_{\rm p}$ are not just beneficial for waveform modeling, but are also used in parameter estimation–for both individual BBHs and their population at large. The tilt angles, for example, of BBHs are poorly constrained on a population level [160, 318], but the $\chi_{\rm eff}$ distribution–also used

to probe anti-aligned spin–is much more precisely constrained [283]. Lastly, as shown in Fig. 4.6, the mismatch still varies along the correlation path, indicating that these are only approximate degeneracies. A sufficiently high SNR will eventually resolve them. Though not analyzed here, our framework provides a structure to quantify this SNR.

Measurements of spins from massive systems have inspired exciting astrophysical interpretations. Understanding how precession is imprinted on these signals and what parameter combinations drive the signal morphology will remain important as more events are detected. Waveform-based correlation mapping provides a principled way to motivate the form of aligned- and precessing-spin effective parameters across masses.

Ch. 4 Data and Code Availability

The neural networks used to generate all results presented in Ch. 4 are found at `https://github.com/yixinkang/MappingBBHParameters`, as are notebooks and scripts to run our correlation mapping algorithm and visualize results.

Ch. 4 Acknowledgments

We thank Nicole Khusid, Max Isi, and Geraint Pratten for helpful comments and suggestions, as well as Lucy Thomas, Aaron Johnson, and Rhiannon Udall for their collaboration and insights about waveform modeling and glitches. This work was supported by NSF Grant PHY-2150027 as part of the LIGO Caltech REU Program which funded KK. SM and KC were supported by NSF Grant PHY-2308770 and PHY-2409001. DF was supported by NSF Grants OAC-2004879, PHY-2207780, and PHY-2114581. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation.

Software: `LALsuite` [319], `matplotlib` [320], `sklearn` [321], `Tensorflow` [322], `numpy` [323], `pandas` [324, 325], `sympy` [326].

Chapter 5

GRAVITATIONAL WAVES CARRY INFORMATION BEYOND EFFECTIVE SPIN PARAMETERS BUT IT IS HARD TO EXTRACT

This chapter is published as Ref. [160]:

Simona J. Miller, Zoe Ko, Thomas Callister, and Katerina Chatziioannou. Gravitational waves carry information beyond effective spin parameters but it is hard to extract. *Phys. Rev. D*, 109(10):104036, 10 2024. ***SJM*** *conceptualized the project jointly with co-authors, mentored Z. Ko (undergraduate), conducted most of the analyses, wrote the majority of the text, and made all of the figures except Figures 13 and 14.*

Abstract

Gravitational wave observations of binary black hole mergers probe their astrophysical origins via the binary spin, namely the spin magnitudes and directions of each component black hole, together described by six degrees of freedom. However, the emitted signals primarily depend on two effective spin parameters that condense the spin degrees of freedom to those parallel and those perpendicular to the orbital plane. Given this reduction in dimensionality between the physically relevant problem and what is typically measurable, we revisit the question of whether information about the component spin magnitudes and directions can successfully be recovered via gravitational-wave observations, or if we simply extrapolate information about the distributions of effective spin parameters. To this end, we simulate three astrophysical populations with the same underlying effective-spin distribution but different spin magnitude and tilt distributions, on which we conduct full individual-event and population-level parameter estimation. We find that parameterized population models can indeed qualitatively distinguish between populations with different spin magnitude and tilt distributions at current sensitivity. However, it remains challenging to either accurately recover the *true* distribution or to diagnose biases due to model misspecification. We attribute the former to practical challenges of dealing with high-dimensional posterior distributions, and the latter to the fact that each individual event carries very little information about the full six spin degrees of freedom.

5.1 Introduction

The spins of black holes (BHs) in binaries (BBHs) are a unique probe of physics on multiple scales, from fundamental BH properties to stellar interiors and the astrophysical environments in which compact binaries form. Each binary possesses six spin degrees of freedom: the spin magnitudes, polar angles (tilts) and azimuthal angles of each binary component [148]. BH spins are encoded in the gravitational waves (GWs) the binary emits and can, at least in principle, be constrained from observation [282, 283, 327] by the LIGO [4] and Virgo [5] detectors. The magnitudes and directions of the spins at merger are determined by the spin each BH has upon formation as well as the binary's evolutionary history, e.g. [61, 67, 83]. Spin measurements are therefore a promising way to determine whether BBHs form dynamically or in the field, e.g. [1, 45, 85, 93, 328, 329], and answer questions such as the role of angular momentum transfer in stars, tidal interactions, and mass transfer, e.g. [51, 58, 59, 81, 86].

Despite their astrophysical importance, spins remain poorly constrained in GW data. Their imprint on the signal is typically subdominant to other intrinsic effects such as the BH masses, e.g. [101, 103, 143, 306, 307]. Furthermore, not all six spin degrees of freedom affect the signal equally. Though waveform models formally depend on the full spin vectors [16, 245, 330, 331], analytical post-Newtonian calculations indicate that the dominant spin effect is captured by two effective parameters: the effective aligned spin $\chi_{\rm eff}$ that includes the spin components parallel to the Newtonian orbital angular momentum [193], and the effective precessing parameter $\chi_{\rm p}$ that includes the perpendicular components [203]. The former primarily affects the length of the signal while the latter describes spin-precession, the change in binary orientation due to spin-orbit and spin-spin interactions [199]. Unsurprisingly, then, constraints on the astrophysical distributions of $\chi_{\rm eff}$ and $\chi_{\rm p}$ can be typically obtained with fewer observations and are less prone to population model systematics than the spin components [162, 164, 332].

Although less well measurable, it is instead the underlying spin components that are of prime astrophysical interest. GW signals contain *some* information about component spins. However, unlike $\chi_{\rm eff}$ and $\chi_{\rm p}$ that appear prominently in the GW phase and amplitude and whose measurability can be predicted with analytic arguments [191, 333], individual spin components have a signif-

icantly subdominant effect on the waveform. The resulting constraints on the *astrophysical* distribution of spin components are correspondingly weaker and in many cases subject to uncertainties about the role of population models [164, 318]. Indeed, even though it is widely accepted that BBHs have a range of $\chi_{\rm eff}$ values that are not symmetric about zero [282, 283, 327] and that not all BBHs have a vanishing $\chi_{\rm p}$ [282, 283], the exact shape of the inferred distribution for spin magnitudes and directions depends on the parameterization of the corresponding population model. For example, different parametrizations for the angle between the spins and the Newtonian orbital angular momentum lead to varied conclusions about where the distribution peaks and the degree of spin-orbit misalignment [164, 266, 267, 283, 318, 334].

Central to this discussion are the questions of how much information GW signals actually contain about the BH spin components versus $\chi_{\rm eff}$ alone, how feasible it practically is to reliably extract this information, and the extent to which conclusions are driven by informative data or simply by overly restrictive models. In this paper, we approach these issues by posing three questions, from which we conclude:

1. **Do GWs carry information about spin components, or are we just extrapolating the effective aligned spin $\chi_{\rm eff}$?** (§5.4)

   Yes, we can distinguish between populations with low, moderate, and high spins even when they have identical effective spin distributions.

2. **Can component spin distributions be *accurately* measured?** (§5.5)

   Even though we can qualitatively tell apart BH populations with different spin distributions, characterizing them accurately is practically challenging.

3. **Can we tell when measurements of component spin distributions are biased?** (§5.6)

   Common tests based on posterior predictive checks cannot identify modeling biases in component spin distributions due to the fact that individual-event posteriors are extremely weakly informative about spin components.

The remainder of this paper presents our analysis in support of these conclusions. We discuss spin degrees of freedom, effective spin parameters, and the notation used throughout in Sec. 5.2. Our Methods are briefly described in

Sec. 5.3, and are expanded upon in the Appendices. Results about measuring the component spin distributions are presented in Sec. 5.4. Section 5.5 introduces the extensive series of verification methods – both population and individual-event level – we use to ensure the robustness of our results, all of which are further elaborated upon in the Appendices. In Sec. 5.6, we identify limitations of the traditional method of using posterior predictive checks to assess biased population measurements, and identify the sources of this bias. We compare our findings to those of past work in Sec. 5.7, and then conclude in Sec. 5.8.

5.2 Spin magnitudes and tilts versus effective spin parameters

Each BH in the binary is described by a dimensionless spin vector $\vec{\chi}_i$, $i \in \{1, 2\}$. In a coordinate system where the z-axis is aligned with the binary Newtonian orbital angular momentum $\vec{L}$, the spin vector is characterized by a magnitude $\chi_i \in [0, 1]$, polar angle $\theta_i \in [0, \pi]$, and azimuthal angle $\phi_i \in [0, 2\pi]$. Modulo horizon absorption effects, the spin magnitude is constant throughout the binary evolution [335, 336], while the spin angles evolve due to spin-orbit and spin-spin interactions causing the spin vector to precess [78, 199].

The full six spin degrees of freedom remain relatively poorly constrained by GW signals. Rather, the dominant spin effects are expressed by two effective parameters. The mass-weighted average spin projected onto $\vec{L}$

$$\chi_{\rm eff} = \frac{\chi_1 \cos\theta_1 + q\chi_2 \cos\theta_2}{1+q} \in (-1, 1)\,, \tag{5.1}$$

is referred to as *effective aligned spin*, where we have defined the binary mass ratio $q \equiv m_2/m_1$, where $m_1 \geq m_2$ are the BH masses. The effective aligned spin is, in general, better constrained as it is related (in the equal mass limit) to the leading-order spin contribution in the post-Newtonian expansion for the GW inspiral phase [148]. Additionally, $\chi_{\rm eff}$ is conserved under spin-precession and radiation reaction to at least the second post-Newtonian order [193].

Spin-precession effects are captured with the *effective precessing parameter*

$$\chi_{\rm p} = \max\left[\chi_1 \sin\theta_1, \left(\frac{3+4q}{4+3q}\right) q\, \chi_2 \sin\theta_2\right] \in [0, 1)\ . \tag{5.2}$$

This parameter and its extensions [204, 205] are motivated by the fact that spin-orbit precession (and the GW amplitude and phase modulations it induces) are driven by in-plane spin components [200, 203]. Constraints on $\chi_{\rm p}$

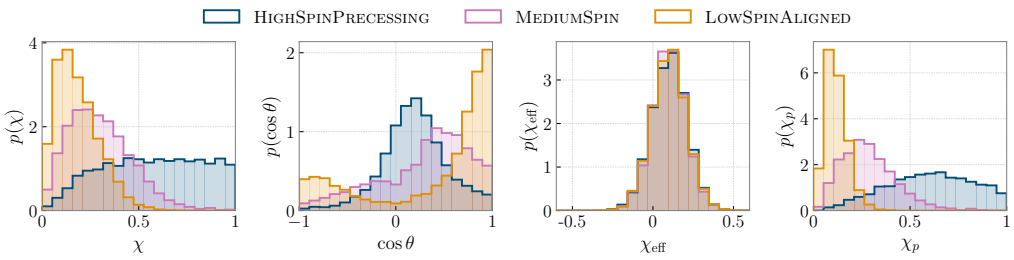


Figure 5.1: Spin distributions for the three simulated BBH populations we use to assess the amount of recoverable information GW signals contain about component spin distributions. The three populations share the same $\chi_{\rm eff}$ distribution (chosen for consistency with current data), but differ in their underlying component spin and $\chi_{\rm p}$ distributions. From left to right, panels show the spin magnitude χ, spin tilt $\cos\theta$, the effective spin parameter $\chi_{\rm eff}$, and the effective precession parameter $\chi_{\rm p}$. (*Navy*) HighSpinPrecessing: a BBH population with few vanishing spins that are preferentially oriented close to the orbital plane. (*Pink*) MediumSpin: a BBH population with moderate spin magnitudes peaking at $\chi = 0.25$ and preferentially aligned with the Newtonian orbital angular momentum. (*Orange*) LowSpinAligned: a BBH population with low spin magnitudes peaking at $\chi = 0.10$ with both strongly aligned and anti-aligned sub-populations.

are typically much weaker than $\chi_{\rm eff}$ especially given the observed absence of large spin-precession in BBHs [282, 283]. In what follows, we therefore focus on $\chi_{\rm eff}$.

5.3 Methodology

In order to isolate the amount of information included in GW signals about the spin components relative to the effective spin, we simulate astrophysical populations with identical $\chi_{\rm eff}$ distributions but different underlying component spin distributions. We choose a $\chi_{\rm eff}$ distribution that is qualitatively similar to current constraints [164, 268, 283] and decompose it into three populations with distinct spin magnitudes and tilt angle distributions. The azimuthal angles are uniformly distributed. These distributions are not astrophysically motivated, but rather selected as distinct test cases of potential distributions.

The three simulated astrophysical distributions are shown in Fig. 5.1, with further details given in Appendix 5.A:

- The HighSpinPrecessing population contains BHs with the most extremal spins and tilts: the majority of the population has $\chi > 0.5$ and tilts nearly in-plane, corresponding to significant spin-precession.

- The MediumSpin population is most similar to current constraints: preferentially small to moderate spin magnitudes peaking at 0.25, and a wide range of tilts with a preference for alignment compared to anti-alignment.

- The LowSpinAligned population has the smallest spin magnitudes, with nearly all BHs having $\chi < 0.5$. Uniquely, this population has a bimodal spin tilt angle distribution, with a larger peak at $\cos\theta = 1$ (perfect alignment) and a smaller peak at $\cos\theta = -1$ (perfect anti-alignment). It is therefore a test case of sensitivity to mixture models.

With these three populations, we conduct a full end-to-end injection/recovery campaign. We draw parameters describing individual GW events from each distribution, restrict to detectable events with a network optimal signal-to-noise ratio (SNR) above 10 in the LIGO Livingston, LIGO Hanford, and Virgo detectors, simulate data assuming O3 sensitivity [337], and obtain samples from the multidimensional posterior distribution of the binary parameters for each event individually. We then hierarchically model the population distribution of the simulated posteriors with parametrized population models.

The individual-event posterior sampling is conducted with the nested sampler Dynesty [253] as implemented in Bilby [243, 254]. We use the IMRPhenomXPHM waveform model [245] both for simulation and recovery as it models all six spin degrees of freedom, contains higher order radiation modes, and is the least computationally expensive option available. Although more computationally expensive than approximate parameter estimation [301, 338, 339], it is essential to use full stochastic sampling for this work. As we are trying to discern subtle effects in the signals, we must properly characterize the individual-event likelihoods. Full details about parameter estimation settings are given in Appendix 5.B. For hierarchical inference, we primarily use the Markov Chain Monte Carlo sampler Emcee [252], with some follow-up studies run with Numpyro [259, 260]. The full hierarchical inference procedure is outlined in Appendix 5.C, with the parameterized population models detailed in Appendix 5.D.

For simplicity, the hierarchical inference ignores the azimuthal angles and in what follows use the term "spin components" to refer to the spin magnitudes and tilt angles. The parameter estimation prior and the population distribu-

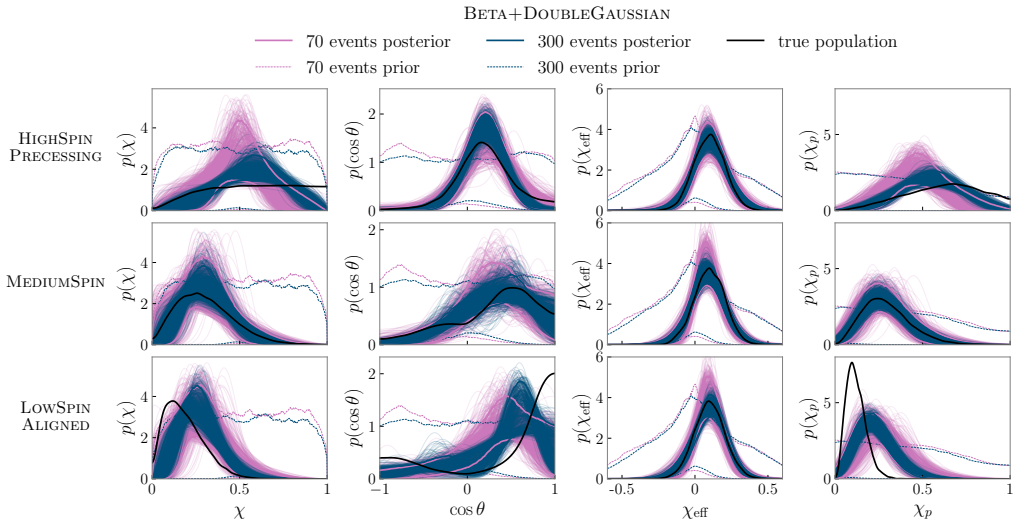


Figure 5.2: Inferred distributions for various spin parameters (from left to right: spin magnitude χ, spin tilt $\cos\theta$, effective $\chi_{\rm eff}$ spin, effective precessing spin $\chi_{\rm p}$) and the three simulated populations (top to bottom). All results are obtained with the BETA+DOUBLEGAUSSIAN model using 70 (pink) and 300 (navy) events. Traces correspond to draws from the population posterior and solid lines enclose 90% of the probability. The black solid line corresponds to the true, underlying population. The dashed lines show the 90% credible intervals inferred by sampling the prior on the population-level parameters, including the effective sample cut as defined in Eq. (5.9).

tion for the azimuthal angles coincide, therefore fixing their distribution (to truth) does not incur a bias, more details are available in Appendix 5.D.

5.4 Different spin magnitude and tilt distributions can be distinguished

Using the results of the signal injection and parameter estimation campaign, we perform hierarchical inference [340–342] on events drawn from each population shown in Fig. 5.1 in order to reconstruct their underlying spin distributions. Population inference requires the adoption of a model for the component spin and tilt distributions. We select an analytic model in which spin magnitudes follow a non-singular Beta distribution and the spin tilt angles follow a bimodal Gaussian distribution, which we hereby refer to as the BETA+DOUBLEGAUSSIAN; see Appendix 5.D for a full description. We choose this population model for its simplicity and similarity to common models in the literature, albeit with small modifications to target our questions of interest. While the true, underlying distributions shown in Fig. 5.1 were *not* explicitly

drawn from the BETA+DOUBLEGAUSSIAN model, this model is expected to agree with each simulated population to within statistical uncertainties.[1] We do not model the spin azimuthal angle, effectively (and correctly) assuming that it is distributed according to its uniform prior.

Fig. 5.2 shows the inferred distributions for various spin parameters for the three simulated populations under the BETA+DOUBLEGAUSSIAN population model. The results from a 70-event catalog are plotted in pink, and from a 300-event catalog in navy, chosen to mimic O3 and projected O4 catalog size respectively. Black, bold traces show the true underlying populations for comparison. The χ (first column) and $\cos\theta$ (second column) distributions are generated by random draws from the posteriors on the population parameters, shown in Figs. 5.6, 5.7, and 5.8 in Appendix 5.E.

The $\chi_{\rm eff}$ (third column) and $\chi_{\rm p}$ (fourth column) distributions are generated by *(i)* randomly drawing from the inferred χ, $\cos\theta$, and q distributions, *(ii)* calculating the effective spin parameters from these draws, and *(iii)* generating a Gaussian kernel density estimate. We find that *the $\chi_{\rm eff}$ distributions are reconstructed accurately across all three populations and for both catalog sizes.*

Switching to the component spins, we can qualitatively distinguish between their distributions for each population. The spin magnitude inferred for the HIGHSPINPRECESSING population is the widest and has the largest mean. From HIGHSPINPRECESSING to MEDIUMSPIN to LOWSPINALIGNED, the means and widths of the inferred distributions get progressively smaller, as is the case for the true, underlying populations. The mass ratio distributions are also successfully recovered for all three populations and two catalog sizes, as shown in Appendix 5.E. We, therefore, *can distinguish between populations with low, moderate, and high spins when they have identical $\chi_{\rm eff}$ distributions.*

Although we can qualitatively characterize the spin magnitude and tilt distributions among these three populations, in some cases we cannot reliably characterize their properties accurately. Specifically, the true underlying distribution of the LOWSPINALIGNED population does not lie within the 90% credible

[1] We confirm that the BETA+DOUBLEGAUSSIAN model is a good fit to the underlying populations through a least-squares fit. The Kullback-Leibler divergences [343] between the best fit and true underlying distributions are $< 10^{-4}$ for all spin magnitude distributions and < 0.08 for all cosine tilt distributions. As a further check, we also run hierarchical inference on catalogs of simulated, Gaussian individual-event spin posteriors, with results discussed in Appendix 5.G.

reconstructed region as measured with the BETA+DOUBLEGAUSSIAN model. The bimodality of this population's tilt angle distribution is not recovered and the inferred spin magnitude distribution has a higher mean than truth. We confirm that this mismatch between the true and inferred distributions is not *solely* driven by overly restrictive priors on the population-level parameters, plotted with dashed lines in Fig. 5.2. For example, the bias in the inferred spin magnitude distribution for the HIGHSPINPRECESSING population occurs over a region where the prior generates a flat distribution.

Population measurements of the effective spin are more robust against bias than component spins. Even a notable mismatch between the true and recovered spin magnitude and tilt distributions for the LOWSPINALIGNED population results in a precisely and accurately constrained $\chi_{\rm eff}$ distribution. The $\chi_{\rm p}$ distributions are more susceptible to inaccurate recovery in the component spins, effectively inheriting their biases. For example, for the $\chi_{\rm p}$ distribution of the LOWSPINALIGNED population: the inference of more in-plane spin ($\cos\theta \sim 0$) combined with the slight over-estimation of the mean of the spin magnitude distribution, leads to a corresponding over-estimate of the bulk of the $\chi_{\rm p}$ distribution. This means that $\chi_{\rm eff}$, but not necessarily $\chi_{\rm p}$ can be reliably characterized on a population level by component spin measurements. We additionally look at alternative definitions of $\chi_{\rm p}$ (see e.g., Ref. [205]) and obtain over-all consistent results with the standard $\chi_{\rm p}$ given in Eq. (5.2).

5.5 Difficulties of Measuring Component Spin Distributions

The biased reconstruction of the LOWSPINALIGNED population is unexpected. The injection and recovery campaign was performed using the same waveform model, and selection effects were self-consistently handled in both signal selection and parameter estimation. Under these conditions, there is no *a priori* reason why population recovery should fail. As such, our results instead suggest a shortcoming in either the parameter estimation or population recovery stages of the analysis.

To diagnose this shortcoming, we employ a slew of checks, all of which are further elaborated upon in Appendix 5.G. To ensure that the problem is not our hierarchical inference framework and implementation we do the following:

- **Simulate Gaussian individual-event spin posteriors** (§5.G.ii) – For each of the 300 events per population, we generate a series of simulated

Gaussian individual-event spin posteriors with a range of measurement errors and underlying correlations, and use these as input to hierarchical inference. In these cases, the BETA+DOUBLEGAUSSIAN population model *is* able to recover the underlying populations, as seen in Fig. 5.10.

Implications: The hierarchical inference and selection effect framework is algorithmically robust, and the BETA+DOUBLEGAUSSIAN model is able to recover the true population distributions to within statistical uncertainties.

- **Fix either the spin magnitude or tilt angle distribution to the truth** (§5.G.iii) – When only fitting for the χ or $\cos\theta$ population and not the other, we are still unable to recover the correct distribution for the LOWSPINALIGNED population; see Fig. 5.12.

  Implications: The observed bias in the spin magnitude and tilt angle distributions is not related to correlations between the two distributions.

- **Use a different sampler for the hierarchical likelihood** (§5.G.iv) – We repeat the analysis of Fig. 5.2 with an independently-implemented hierarchical inference code that is based on NUMPYRO instead of EMCEE. We obtain essentially identical results, shown in Fig. 5.13.

  Implications: The hierarchical inference and selection effect framework is algorithmically robust.

- **Fit for the mass and redshift distributions instead of fixing it to truth** (§5.G.v) – Our main results fit for the spin magnitude, spin tilt, and mass ratio distributions, while fixing the distributions of the primary mass and redshift to their true population values, given in Appendix 5.A. Figure 5.13 extends these results to also fit for the mass and redshift distributions and shows the corresponding spin population posteriors, which remain unchanged. The mass and redshift distributions are recovered with no bias.

  Implications: We have not misspecified the mass or redshift distributions when fixing them to truth during hierarchical inference, nor biased results of our spin inference by neglecting to simultaneously fit for the mass and redshift distributions.

- **Plot rates instead of probability distributions** (§5.G.vi) – Apparent disagreement between injected and recovered probability distributions can

sometimes be caused by comparing injected and recovered probability distributions, rather than differential merger rate densities. In the main text we do not infer the overall rates of black hole mergers, but only the shapes of their spin distributions. To check if neglecting the merger rate contributes to apparent disagreement between injected and recovered populations, we repeat our hierarchical inference while also fitting for the rate of black hole mergers as function of spin magnitude and tilt. This yields the results shown in Fig. 5.14, which remain qualitatively similar to those in Fig. 5.2.

Implications: We have not biased results of our spin measurements by failing to measure or plot absolute merger rates, rather than probability distributions.

- **Exclude spin selection effects** (§5.G.vii) – The selection function only negligibly affects spin magnitudes and tilt angles. To ensure that we are not incorrectly implementing the selection function in the hierarchical likelihood, we conduct hierarchical inference without including selection effects in spin. This does not impact our results, as can be seen in Fig. 5.15.

  Implications: The implementation of the selection function is algorithmically robust.

- **Employ different methods of breaking the degeneracy in the bimodal Gaussian model** (§5.G.vii) – For a bimodal distribution, some method must be imposed to break the degeneracy between the two components of the model. For the BETA+DOUBLEGAUSSIAN model, this can be done in one of three ways: imposing an ordering of the means, the widths, or limiting the mixing fraction be ≤ 0.5. Sometimes one method of breaking the degeneracy converges better than another. As shown in Fig. 5.15, we find that this is not the case here and different methods perform comparably.

  Implications: Our choice of degeneracy-breaking between the two Gaussian components in our population model is not causing convergence issues.

- **Run hierarchical inference on different 70-event catalog instantiations** (§5.G.vii) – Finally, to get a sense of how much the specific 70 events we select from the underlying population affect hierarchical inference, we repeat the procedure with several different catalog instantiations. While there is expected variance in the results –see Fig. 5.16– it cannot account for the degree of mismatch seen in the bottom row of Fig. 5.2. Additionally,

each catalog instantiation leads to a different number of per-event effective samples, which we find are not correlated to the goodness of fit.

Implications: The observed bias does not arise from an insufficient number of per-event effective samples.

We then move towards investigating the underlying individual-event parameter estimation with the following checks:

- **Sampler settings** – We run parameter estimation with a large variety of sampler settings in BILBY, and eventually adopt the standard, reviewed settings for our headline results of Fig. 5.2.

  Implications: Running with more aggressive sampler settings in BILBY may fix convergence problems, but this was not the case for any configurations we employed.

- **Probability-probability (P-P) plots** (§5.G.i) – We generate P-P plots [344, 345] for reweighted individual-event BILBY posteriors. As seen in Fig. 5.9, the test passes.

  Implications: Either the BILBY individual-event posterior samples are unbiased, or the biases are subtle enough to not be detectable by a P-P test, as warned against in Ref. [346].

- **Use a different waveform model** (§5.G.iii) – We re-run individual-event inference on the same sets of events with BILBY using the IMRPHENOMXP waveform model instead of IMRPHENOMXPHM both for injection and recovery. Results with this waveform model are comparable or worse to that presented in the main text with IMRPHENOMXPHM, although the bimodality of the LOWSPINALIGNED population is slightly better constrained; see Fig. 5.11.

  Implications: The *existence* of bias in the measured spin magnitude and tilt distributions is not driven by our choice of waveform model, although the specific details of how that bias manifests appear to be, i.e. different waveforms yield different population-level results. This indicates that the bias may be due to individual-event sampling issues.

- **Fix non-spin parameters to truth in individual-event sampling** (§5.G.iii) – Finally, we conduct individual-event inference with IMRPHE-

NOMXPHM fixing all parameters aside from the spin magnitudes to tilt angles to truth (i.e. use delta function priors at their injected values). In this case, the BETA+DOUBLEGAUSSIAN population model *is* able to successfully recover the truth for all three populations; see Fig. 5.11.

Implications: The added complexity going from sampling just spins to all fifteen binary parameters is a likely culprit for the biased spin magnitude and tilt angle distributions.

Although we can qualitatively tell apart the different populations in Fig. 5.1, our results indicate that the spin distribution of all possible BBH populations cannot necessarily be *accurately* measured under the range of analyses considered in this work. Full parameter estimation with spin-precession is a technically challenging analysis. Despite conducting tens of model checking procedures, we cannot fully identify the driving source of the bias observed in Fig. 5.2. We hypothesize that the error is due to issues related to sampling from the high-dimensional posterior for individual events, as suggested by the final bulletpoint above. If the issue with unbiased recovery is indeed due to poor convergence of parameter estimation, then it is possible that future algorithmic improvements in parameter estimation will resolve things and allow for accurate recovery.

5.6 Identifying bias is difficult: limitations of posterior predictive checks

While the BETA+DOUBLEGAUSSIAN model was able to produce qualitatively correct results for each of the three distinct populations, it was not able to successfully recover the true underlying populations. This leads us to the question: *can commonly-used modeling diagnostics successfully identify poorly performing fits to component spin distributions?* There are multiple avenues through which a population model can fail: either the model is theoretically a good fit and for any number of reasons (e.g. those discussed in Sec. 5.5) cannot find the truth, or the model is intrinsically a poor fit, i.e. it does not have enough flexibility to find the shape of the true, underlying distribution. Both cases induce mismatch between the true distribution and the inferred distribution, which we hope to diagnose using only the information available to us. In this section, we begin by discussing the first scenario (Fig. 5.3) and then the second (Fig. 5.4).

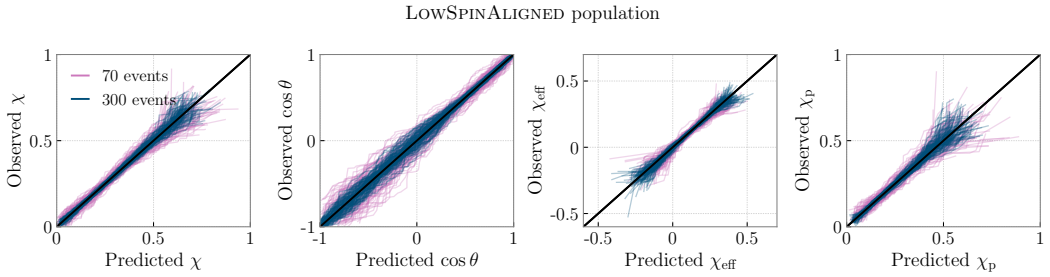


Figure 5.3: Posterior predictive checks (PCCs) for spin parameters (left to right: component spin magnitudes χ, component spin tilts $\cos\theta$, effective spin $\chi_{\rm eff}$, and effective precessing spin $\chi_{\rm p}$) of the LowSpinAligned population under the Beta+DoubleGaussian population model. Each trace is one catalog from the observed and the predicted populations. 100 catalogs are shown, for results with 70 (pink) and 300 (navy) events. In the absence of discrepancy between the inferred and true population distributions, the traces should on average follow the diagonal. However, here there *is* such a discrepancy, as can be seen in the bottom row of Fig. 5.2, and the traces do on average follow the diagonal, meaning that this PPC is *not necessarily a good diagnostic tool* for component spins.

In reality, given the complexities of astrophysical BH spin evolution, it is almost certain that our measured distributions are in some other way discrepant with the truth; phenomenological models likely cannot perfectly reflect the underlying populations. Model checks on current data sets are then used to motivate more complicated parametric models that do not suffer from identifiable deficiencies. In parallel, nonparametric inference introduces more flexible models that are based on a large number of parameters, however those are also subject to model uncertainties and impose correlations across the population parameter space [266–268]. Detailed model checking remains an essential ingredient of population constraints.

For end-to-end event simulation and population recovery such as Fig. 5.2, we *a priori* know what the "true" underlying astrophysical distribution is. However, when dealing with real GW observations, this is, of course, not the case. We therefore diagnose the bias seen in Fig. 5.2 using only information available to us when dealing with real observations. To do so, we use posterior predictive checks (PPCs) that examine the predictive accuracy of the inferred models via its ability to predict future data that are consistent with current observations. PPCs are ubiquitous in the field of GW population analyses [164, 282, 283, 338, 347].

We now look at the results from the BETA+DOUBLEGAUSSIAN model presented in Sec. 5.4: a case in which a population model is theoretically a good fit, but cannot find the underlying distribution accurately. A PPC for the LOWSPINALIGNED population[2] is plotted in Fig. 5.3. Specifically, we plot the spin parameters predicted by the fitted model against those of the observed events. The "predicted" (horizontal axis) and "observed" (vertical axis) draws and are generated as follows:

1. Draw one sample from the posterior for the BETA+DOUBLEGAUSSIAN hyper-parameters.
2. Draw one sample from the *detectable* [264, 268] χ_i and $\cos\theta_i$ distribution corresponding to this hyper-parameter. This is the *predicted* draw.
3. Draw one sample from one individual-event posterior in the catalog, reweighted to the population from Step 1, as described in Appendix 5.F. This is the *observed* draw.
4. Repeat 70 times for the O3-like catalog, or 300 times for the O4-like catalog.

The predicted and observed values are sorted and plotted against each other, generating one trace in Fig. 5.3. We repeat this procedure 100 times to generate a collection of traces. If we have perfectly measured the true underlying distribution and in the limit of infinitely many observations, the traces should be an exact diagonal. For a number of finite observations, the *average* of the traces should be diagonal [162, 338, 348–350]. As the number of observed events increases, the spread of the traces around the diagonal should decrease.

For all spin parameters shown, the traces on average *do* follow the diagonal, *even though the measured population does not match the truth.* The 300 event case (navy) traces are more tightly clustered around the diagonal than the 70 event case (pink), as expected. That the traces average to the diagonal but we know the fit is poor indicates that this class of PPC, although widely used in GW population analyses, is not a sufficient diagnostic of model mismatch or inaccurate population inference in this case.

[2]We highlight the LOWSPINALIGNED population throughout Sec. 5.6 as it displays the largest bias under the BETA+DOUBLEGAUSSIAN model.

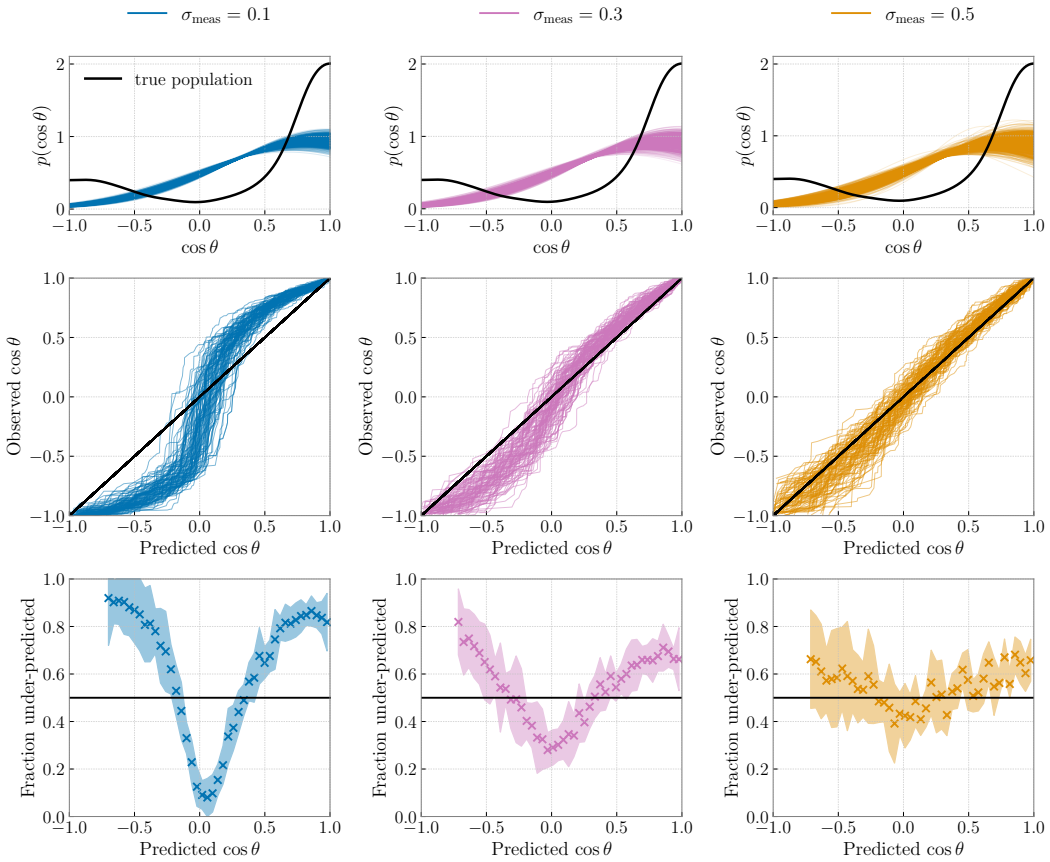


Figure 5.4: Results for diagnosing model misspecification for the LowSpinAligned population's $\cos\theta$ distribution under the Beta+Gaussian model. Results from three different simulated measurement uncertainties are shown: individual-event spin measurement error of $\sigma_{\rm meas} = 0.1$ (blue), 0.3 (pink), and 0.5 (yellow). The "true" measurement error from the Bilby runs averages to $\sigma_{\rm meas} = 0.48$. All results are shown from runs done on 70-event catalogs. *(Top Row)* Traces corresponding to draws from the population posterior, compared to the true population (black). *(Middle Row)* PPCs from 100 catalogs, each with 70 events. In the absence of model misspecification, the traces should on average follow the diagonal. *(Bottom Row)* Fraction of events from the posterior predictive checks with $\cos\theta$ under-predicted. The shaded regions indicate three-sigma uncertainty on each average value, marked by the crosses. In the absence of model misspecification, the error bars should encompass a horizontal line at 0.50.

We continue to investigate the conditions under which PPCs succeed or fail by next turning to the second case discussed previously: that in which a population model is an *intrinsically* bad fit to the underlying astrophysical distribution. We again perform population inference of simulated data, now deliberately using a model that *cannot* reproduce the injected spin distributions. In particular, we will adopt a model in which cosine tilts are described only as a single Gaussian, which we call the BETA+GAUSSIAN model and is given analytically in Appendix 5.D. This population model *is* capable of reproducing the $\chi_{\rm eff}$ distribution, but at the level of component spins cannot capture the bimodality present in the MEDIUMSPIN and LOWSPINALIGNED populations.

We here wish to isolate the effect of a bad model, *without* having to worry about the shortcomings of inference per Fig. 5.2. As such, this time we do not perform full signal injection and parameter estimation, but instead produce mock spin magnitude and cosine tilt posteriors, allowing us to better control and understand the interplay between individual-event and population-level measurement uncertainty. We assume these mock-posteriors to be Gaussian distributed with width $\sigma_{\rm meas}$. Our procedure for generating these mock-posteriors is detailed in Appendix 5.G.

Results from conducting hierarchical inference using the Gaussian mock posteriors are shown in Fig. 5.4. The left-hand column (blue), shows results from individual event spin posteriors with $\sigma_{\rm meas} = 0.1$, between $\sim$1-5 times more informative than the BILBY-produced $\cos\theta$ posteriors, which averages to $\sigma_{\rm meas} = 0.48$. In the right-hand column (orange) are plotted results from more-realistic individual-event measurement error of $\sigma_{\rm meas} = 0.5$; the center column (pink) is an intermediate case of $\sigma_{\rm meas} = 0.3$.

The inferred $\cos\theta$ distributions for the LOWSPINALIGNED population under the BETA+GAUSSIAN model are plotted in the top row of Fig. 5.4. For each of the three different individual-event measurement errors, the traces are clustered tightly, meaning that the inferred population is precisely measured, even though the model is not a good fit to the underlying population. The BETA+GAUSSIAN model is, in essence, doing its job: even with large individual event uncertainty, it identifies the mean and the over-all width of the distribution very well, even though it cannot capture the full underlying bimodal structure.

We again ask: if we did *not* know the injected distribution, would we have been able to tell that this model is insufficient? Going further, in the case that we *can* tell a PPC fails, we are looking for an estimate of how we should amend our population model to better fit the truth. Beyond just inspecting the diagonality of PPCs by eye, we can calculate the fraction of events over/under-predicted by our model across parameter space using the slopes of the PPC traces. If the slope of a PPC trace is steeper (shallower) than the diagonal, then the model is predicting more (fewer) events in that region of parameter space than are observed. To find the slopes of each trace as a function of each parameter of interest, we perform linear regression in a small region around each point on a grid spanning that parameter. The fractions of each spin parameters under-predicted for each simulated population is then the fraction of traces with slopes shallower than the diagonal (i.e. < 1). If the model is a good fit to the data, the fraction under-predicted should be consistent with 0.5.

PPCs and the corresponding fraction of events under-predicted are shown in the middle and bottom rows of Fig. 5.4 respectively. Errors on the fraction under-predicted are calculated by repeating the PPC procedure ten times, and calculating the mean (crosses) and variance (shaded region) of the results. For the $\sigma_{\rm meas} = 0.1$ case, the PPC is inconsistent with the diagonal, meaning that here we *can* identify that inferred distribution under the Beta+Gaussian model is not a good fit. The fraction of events under-predicted is correspondingly inconsistent with 0.5. For $\cos\theta \lesssim 0.25$ and $\cos\theta \gtrsim 0.75$, the fraction under-predicted is greater than 0.5, meaning that the model ubiquitously under-predicts the population in this region of parameter space. Between $0.25 \lesssim \cos\theta \lesssim 0.75$, the fraction is less than 0.5, meaning here the model over-predicts. Looking at the top left corner of Fig. 5.4, we can see that this is exactly the case. These results hint at how we could improve the $\cos\theta$ model: to find the truth we should allow the model to predict more events at alignment and anti-alignment, i.e., include a bimodality.

As the individual-event measurement uncertainty increases (left to right), the PPCs become more consistent with the diagonal, and correspondingly the fractions become consistent with 0.5. By realistic measurement uncertainty, we lose our ability to diagnose inconsistency between the underlying and measured populations using PPCs. A crucial step in generating PPCs is the reweighting

of individual-event posteriors to the inferred population. If individual-event posteriors are sufficiently uninformative, then this process yields reweighted posteriors that are all essentially identical to the measured population. Thus, the "observed" and "predicted" draws will be the same, and the PPCs will be on average diagonal. *This type of PPC is therefore insufficient for weakly informative parameters as the reweighted posterior is dominated by the population rather than the individual-event likelihood.* We propose that alternative model-checking procedures must, therefore, be developed and utilized for diagnosing model bias and misspecification for poorly-measured BBH parameters such as spin components.

5.7 Comparison to past work on Hierarchical Inference with Spin-Precession

To our knowledge, our study includes the first full, end-to-end individual-event and population-level GW injection campaign for multiple distinct populations of spin magnitudes *and* tilt angles of BBHs. Past studies performing injection-recovery campaigns for spin populations either use different waveform models and sampling implementations, and/or consider cases with reduced complexity compared to ours. Our work is consistent with past findings, as described below.

Ref. [351] investigated the measurability of the spin tilt angle distributions alone using an astrophysically-motivated model assuming that some fraction of BBHs form in isolated binaries, while the rest form dynamically. They performed an injection and recovery campaign for spin tilts where they measured the fraction of binary mergers with preferentially-aligned versus isotropically distributed tilts, and the typical degree of spin misalignment for each BH. In their study, all simulated binaries share the same masses, distance, and spin magnitudes, chosen to be similar to LIGO's first event GW150914 [17]. Using the waveform model IMRPhenomPv2 [316] and nested sampling implemented in LALInference [311], they found that they *are* able to constrain the parameters of the tilt-angle distributions for five different populations. Although they *do* sample over all fifteen BBH parameters during individual-event inference, Ref. [351] is most similar to the follow-up studies we present in Appendix 5.G: we too are able to better recover the underlying distributions for all three of our populations when the complexity of the explored parameter space is reduced (see, e.g. Fig. 5.12), on either an individual-event or

population level.

In the context of searching for unresolved binary signals, Ref. [352] also simulated and recovered BBH spin magnitude and tilt distributions. They looked at a single population, consistent with the LIGO/Virgo O1 and O2 observations [327], and used the IMRPhenomPv2 waveform model [316] implemented in Bilby for individual-event inference. A crucial difference between this study and ours is the use of a selection function: as Ref. [352] are looking at resolved *and* unresolved binaries, they ignore selection effects entirely. Under these conditions, Ref. [352] find that the spin magnitude and tilt angle distributions are both accurately measurable. Given that their simulated populations are most similar to our MediumSpin population, this is in agreement with our findings, as we are able to well constrain the MediumSpin population for both 70 and 300 events. It is only when more complex distributions are introduced that our inference fails.

Another point of comparison between our work and others' is on the subject of biased measurements from population model misspecification. In particular, other authors have also identified shortcomings of traditionally and widely used model checking techniques such as probability-probability plots (Appendix 5.G) and posterior predictive checks (Sec. 5.6). Ref. [346] discussed population model bias in the mass distribution of binary neutron star populations arising from misspecification of spin priors. As part of their work, they show that a P-P check on individual-event posteriors can pass but still lead to highly biased population inference. This is in agreement with our findings.

5.8 Conclusions

In this work, we investigated the measurability of the spin magnitude and tilt angle distributions of BBH populations via GW observations. To see if realistic GW populations contain information about spin components or just the effective spin, we simulated three BBH populations that have the same underlying effective-spin distributions, but deliberately distinct spin magnitude and tilt distributions, and on them conducted individual-event and population-level parameter estimation. We then turned to the question of whether mismatch between the injected and recovered spin magnitude and tilt distributions can be identified using only the individual-event and population-level data available to us, without knowledge of the true underlying population. Our work

focuses on the three questions posed in Sec. 5.1, the answers to which we summarize below.

(1) ***There is information in gravitational-wave signals beyond the effective-spin.*** As discussed in Sec. 5.4, we can tell that our three different populations have different spin magnitude and tilt distributions despite their having identical $\chi_{\rm eff}$ distributions.

(2) ***Measuring component spin distributions accurately is practically challenging.*** Under standard, reviewed parameter estimation settings, we were able to accurately measure the spin magnitude and/or tilt angle distributions for some of the populations, but not all three. The bimodal tilt distribution of the LowSpinAligned population proved especially resistant to being accurately constrained, even when using a population model that was inherently bimodal. We employed a suite of verification methods to ensure the robustness of these results, which are enumerated in Sec. 5.5. Notably, we find that the effective spin distribution *is*, however, accurately measured no matter the degree of mismatch in the component spin model constraints; this is not true for the effective precessing spin $\chi_{\rm p}$ which remains susceptible to biased spin magnitude and tilt inference.

Although we cannot say for certain, we hypothesize that the root of the mismatch between the recovered and underlying distributions is related to a lack of convergence in individual-event posteriors, the specifics of which are subtle enough to not present themselves via a standard P-P check (Appendix 5.G). We do not claim that accurately recovering component spin population distributions is *impossible* at current sensitivity, just challenging. Running `Bilby` with more aggressive settings, while computationally costly, may very well fix the problems presented in this work. However, unlike our injection set, real observations do not come with an answer key. If manually tuning sampler settings is a requirement to recover truth, we must be aware what these same errors could manifest in real LIGO/Virgo events.

(3) ***At current sensitivity, we cannot tell when measurements of component spin distributions are biased via the currently widely-used method of posterior predictive checks.*** Due to the fact that individual-event posteriors are extremely weakly informative about spin components, reweighting these posteriors to the inferred population distribution–a crucial step in conducting posterior predictive checks–yields individual-event mea-

surements that are all nearly identical to the inferred population itself. Nearly *any* population model can seem like a good fit to poorly constrained data, as discussed in Sec. 5.6.

Ref. [338] detailed different categories of posterior predictive checking for GW data. The most commonly used level is what we do in this work: performing consistency checks on the *true* underlying parameters of the observed data versus predicted by the model. However, one can also conduct PPCs on the *observed* parameters (e.g. max likelihood parameters) of the data versus those predicted by the model. While checks on the true parameters are susceptible to the issues related to reweighting that we discuss in Sec. 5.3, checks on observed parameters might be more constraining. However, they are far more computationally expensive to perform, as one must generate maximum likelihood values predicted by the model: this involves either running an optimization routine or conducting mock-parameter estimation on thousands of events. While trustworthy mock-parameter estimation exists for some parameters (e.g. masses) [301, 338, 339], the imprint of spin magnitudes and tilt angles on data is more subtle and remains unincorporated into these algorithms. Developing different, more-informative methods of posterior predictive checking for poorly-constrained parameters such as spin is an essential topic of future work.

Ch. 5 Data and Code Availability

The code used to produce all results presented in Ch. 5 can be found at `https://github.com/simonajmiller/measuring-bbh-component-spin`. Posterior samples can be shared upon request.

Ch. 5 Acknowledgments

We thank Sylvia Biscoveanu, Jacob Golomb, Ethan Payne, and Colm Talbot for their extensive assistance with Bilby parameter estimation and other helpful comments on this work; Sophie Hourihane for essential insights about reweighting and probability-probability plots; and Salvatore Vitale for useful feedback and discussion Furthermore, we extend thanks to Maya Fishbach, Reed Essick, Matthew Mould, Zoheyr Doctor, and Will Farr for their suggestions related to the range of verification methods performed in this paper. Finally, we thank our anonymous referees for their feedback.

This work is supported by National Science Foundation grant No. PHY-2150027 as part of the LIGO Caltech REU Program, which funded ZK. S.M and KC were supported by NSF Grant PHY-2110111 and NSF Grant PHY-2308770. TC is supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459.

Software: `emcee` [252], `bilby` (version 2.2.2) [243, 254], `dynesty` (version 2.1.2) [253], `numpy` [323], `scipy` [353], `matplotlib` [320], `seaborn` [354], `astropy` [355, 356], `jax` [258], `numpyro` [259, 260]

Ch. 5 Appendices

5.A Simulated Populations

We simulate three populations with the same $\chi_{\rm eff}$ but different spin magnitude χ and tilt angle θ distributions. To generate the populations, we first choose distributions of the mass ratio q and spin z-component $s_{i,z} \equiv \chi_i \cos\theta_i$ to be shared in common across all three populations; this ensures the same $\chi_{\rm eff}$ distribution. The mass ratio distribution corresponds to the median posterior value inferred with the PowerLaw+Peak model in Ref. [283], while for $s_{i,z}$ we select a Gaussian with mean 0.10 and standard deviation 0.15. This yields a Gaussian-like $\chi_{\rm eff}$ distribution with mean 0.10 and standard deviation 0.11.

To decompose $\chi_{\rm eff}$ into component spins, we choose a different spin magnitude χ distribution for each population and then numerically calculate the resultant $\cos\theta$ distribution implied by $p(\chi)$ and $p(s_{i,z})$. This procedure results in the three populations shown in Fig. 5.1. The χ distribution for the HighSpinPrecessing population is uniform between $s_{i,z}$ and 1, i.e., $s_{i,z}$ values are drawn from the Gaussian described above and then a χ_i value is conditionally drawn based on each $s_{i,z}$. For the MediumSpin (LowSpinAligned) population, each χ value is drawn from a Gaussian distribution about $s_{i,z}$, truncated on $0 \leq \chi \leq 1$, with a standard deviation of 0.20 (0.05). For each population, we assume χ_1 and χ_2 are identically but independently distributed, as are $\cos\theta_1$ and $\cos\theta_2$. Finally, each spin vector's azimuthal angle ϕ_i is drawn uniformly between 0 and 2π. Due to the different χ and $\cos\theta$ distributions,

the χ_p distributions of each population differ as well.[3]

The astrophysical distribution of the remaining binary parameters is the same for all populations. We inject primary masses drawn from the PowerLaw+Peak model [357] with all parameters, except for $m_{\rm min}$, fixed to their one-dimensional median values as found in Ref. [283]: $\alpha = 3.51$, $m_{\rm max} = 88.21$, $\lambda_{\rm peak} = 0.033$, $\mu_m = 33.61$, $\sigma_m = 4.72$, and $\delta_m = 4.88$ in the notation used therein. The injected mass ratio distributions for all populations are described by a power law with slope $\beta_q = 0.96$ (see Eq. 5.12), again the median inferred value from [283]. In the parameterization of the PowerLaw+Peak model, we use a population minimum mass of $m_{\rm min} = 6\,M_\odot$ instead of $5\,M_\odot$ to set the shape of the distribution. We additionally impose a mass *cut* of $8\,M_\odot$, as restricting to higher-mass events ensures shorter analysis times. This mass cut effectively becomes the minimum mass, but we renormalize the distribution to keep the same *shape* above the cutoff mass as it would with $m_{\rm min} = 6$. Explicitly setting $m_{\rm min} = 8$ in the PowerLaw+Peak model would change the over-all shape of the distribution to be inconsistent with the desired results in Ref. [283].

Finally, the BBH merger density rate in the source frame evolves with respect to redshift z as

$$R(z) \propto \frac{dV_c}{dz}\,(1+z)^{2.7}\,, \tag{5.3}$$

where V_c is the comoving volume. Distances are calculated from redshifts assuming the cosmology reported by the Planck 2013 survey [358]. All other parameters are drawn uniformly from their respective physical range. Mass and redshift distributions are plotted in Fig. 5.5.

5.B Individual-event Parameter Estimation

From each of the three astrophysical distributions described in Appendix 5.A, we draw 10^5 events. We apply a network SNR [359–362] cut of 10 in the LIGO Livingston, LIGO Hanford, and Virgo detector network using the "O3 actual" power spectral densities (PSDs) provided in Ref. [337], and select 300 detectable events. Histograms of events from the underlying (navy) versus detectable (orange) mass and redshift distributions are shown in Fig. 5.5.

[3] Different component spin distributions given identical $\chi_{\rm eff}$ *and* $\chi_{\rm p}$ distribution can only be achieved by relaxing the assumption of identically distributed component spin magnitudes and angles.

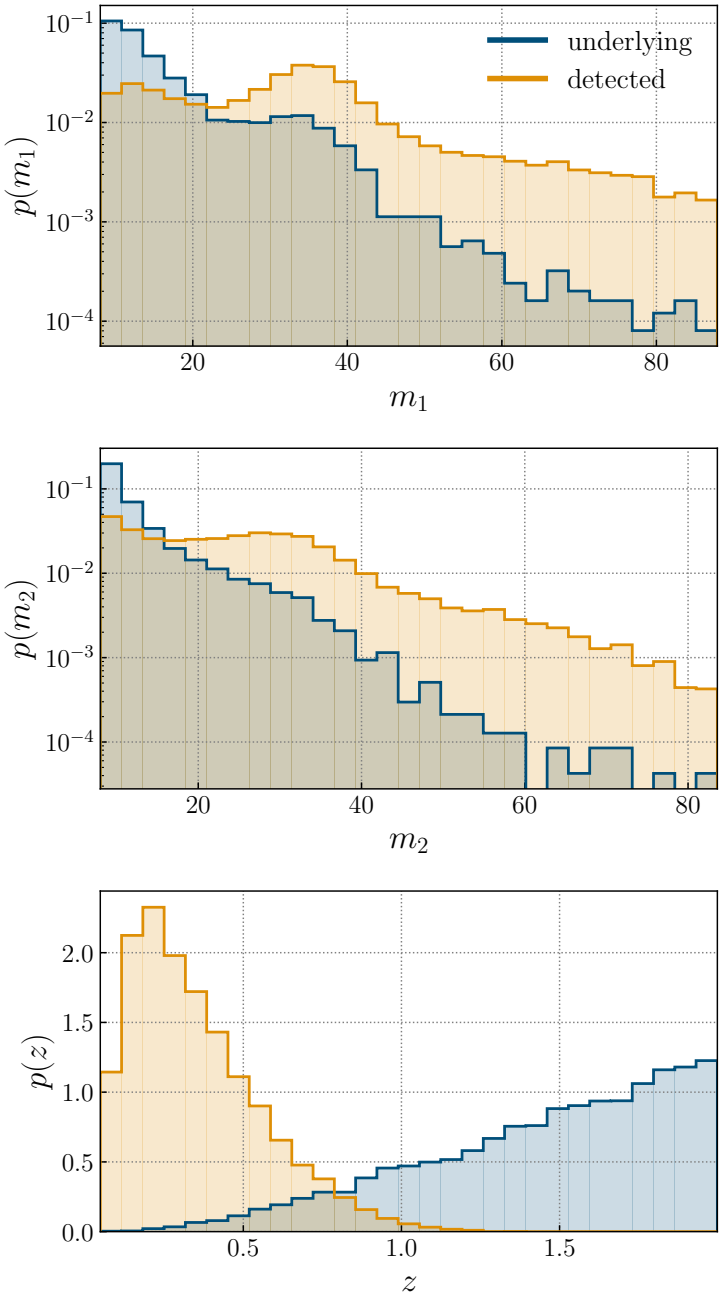


Figure 5.5: The underlying (navy) and detected (orange) distributions for source-frame primary mass m_1 (top), source-frame secondary mass m_2 (middle), and redshift z (bottom) shared between all three simulated populations.

For each event, we simulate GW data with the IMRPhenomXPHM waveform model [245] including a Gaussian noise realization and draw samples from the 15-dimensional posterior distribution for the binary parameters using the same waveform. Specifically, we sample in detector frame component masses m_1, m_2, spin magnitudes χ_1, χ_2, spin tilt angles θ_1, θ_2, the azimuthal inter-spin angle ϕ_{12}, the azimuthal cone precession angle ϕ_{JL}, the luminosity distance d_L, the inclination angle between the total angular momentum and the line of sight of the observer θ_{JN}, the right ascension α, declination δ, polarization angle ψ, and the time t and orbital phase φ at coalescence. We employ standard priors for all binary parameters [254], although we use a targeted chirp mass [363, 364] prior of $\pm 15 M_\odot$ about the injected value to reduce computational cost, which we verify does not affect results.

Simulated data assume a detector network of LIGO Hanford, LIGO Livingston [4], and Virgo [5], each at their O3 sensitivity [337] with a sampling rate of 2048 Hz. We analyze data in the $15 - 921.6 (= 0.9 \times 2048/2)$ Hz frequency range, assuming perfect knowledge of the detector calibration.

We use the nested sampler dynesty [253] as implemented in Bilby [243, 254] under reviewed settings to stochastically sample from the individual-event posteriors. For sampler settings, we use `nlive = 1000`, `naccept= 60`, and `sample = ''acceptance-walk''`. Time marginalizaton is turned on, while distance and phase marginalization remain off. Post-facto, we apply an optimal SNR cut of 10 on the posterior samples for consistency with our selection criteria [264].

5.C Hierarchical Inference

The hierarchical inference framework used in our analysis to obtain posteriors distributions on the population parameters is implemented using the Python Markov-Chain Monte-Carlo package `emcee` [252]. The likelihood $\mathcal{L}(\{d\}|\Lambda)$ that a catalog of $N_{\rm obs}$ GW events with data $\{d_i\}_{i=1}^{N_{\rm obs}}$ arises from an underlying population $\pi_{\rm pop}$ described by parameters Λ is given by [237, 340–342]

$$\mathcal{L}(\{d\}|\Lambda) \propto \prod_i^{N_{\rm obs}} \frac{\int d\lambda\, \mathcal{L}(d_i|\lambda)\pi_{\rm pop}(\lambda|\Lambda)}{\xi(\Lambda)}\,, \tag{5.4}$$

where λ_i are the parameters of the $i^{\rm th}$ event in the catalog (i.e. spins, masses, etc). In practice, we have access to the individual-event posteriors $p(\lambda_i|d_i)$ obtained with a default parameter-estimation prior $\pi_{\rm pe}(\lambda_i)$, rather than the

event likelihood $\mathcal{L}(d_i|\lambda_i)$. We thus write Eq. (5.4) as

$$\mathcal{L}(\{d\}|\Lambda) \propto \xi(\Lambda)^{-N_{\rm obs}} \prod_i \int d\lambda \, \frac{p(\lambda|d_i)}{\pi_{\rm pe}(\lambda_i)} \pi_{\rm pop}(\lambda_i|\Lambda) \,. \tag{5.5}$$

Additionally, rather than $p(\lambda_i|d_i)$ itself, we have a discrete set of N_i independent samples $\{\lambda_{i,j}\}_{j=1}^{N_i}$ drawn from $p(\lambda_i|d_i)$. Using the standard procedure, we approximate the integral of Eq. (5.5) via a Monte Carlo average,

$$\mathcal{L}(\{d\}|\Lambda) \propto \xi(\Lambda)^{-N_{\rm obs}} \prod_i \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\pi_{\rm pop}(\lambda_{i,j}|\Lambda)}{\pi_{\rm pe}(\lambda_{i,j})} \,. \tag{5.6}$$

The detection efficiency

$$\xi(\Lambda) = \int d\lambda \, \pi_{\rm pop}(\lambda|\Lambda) P_{\rm det}(\lambda) \,, \tag{5.7}$$

is the fraction of events that we would successfully detect if the population with parameters Λ is the true underlying population. Here, $P_{\rm det}(\lambda)$ is the probability that an individual event with parameters λ is detected. As with the population likelihood, we calculate the detection efficiency with a Monte Carlo average. Given $N_{\rm inj}$ injected signals drawn from some reference distribution $p_{\rm inj}(\lambda)$, the detection efficiency is

$$\xi(\Lambda) = \frac{1}{N_{\rm inj}} \sum_{i=1}^{N_{\rm fnd}} \frac{\pi_{\rm pop}(\lambda_i|\Lambda)}{p_{\rm inj}(\lambda_i)} \,, \tag{5.8}$$

where the sum is over the $N_{\rm fnd}$ injections that pass the detection criteria. We generate the set of "found" injections over which the Monte Carlo average is calculated in the same way that we produced catalogs of events in Appendix 5.B. The reference $p_{\rm inj}(\lambda)$ follows the true mass and redshift distribution (Appendix 5.A; Fig. 5.5), but is uniform in spin magnitudes and isotropic in spin tilts such that we can resolve features across the full underlying spin distribution. As in Appendix 5.B, our detection criterion is an optimal SNR greater than 10 using the waveform IMRPhenomXPHM [245] in the LIGO Livingston, LIGO Hanford, and Virgo network at O3 sensitivity [337]. We acknowledge that the optimal SNR is not a strictly accurate estimate of selection effects on real data as it is solely a function of source parameters, not detector noise. However, this approach remains formally self-consistent as long as we apply the same optimal SNR cut on posterior samples, as explained in Ref. [264].

Following Ref. [280], we account for uncertainty in the Monte Carlo integral by demanding that the effective number of independent samples

$$N_{\rm eff}(\Lambda) \equiv \frac{\left[\sum_{i=1}^{N_{\rm fnd}} w_i(\Lambda)\right]^2}{\sum_{i=1}^{N_{\rm fnd}} \left[w_i(\Lambda)\right]^2} \geq 4\, N_{\rm obs}\,, \tag{5.9}$$

where the weights w_i between the population distribution and parameter estimation prior are defined as

$$w_i(\Lambda) = \frac{\pi_{\rm pop}(\lambda_i|\Lambda)}{p_{\rm inj}(\lambda_i)}\,, \tag{5.10}$$

evaluated on the parameters of the *found* injections. This procedure rejects samples from regions of parameter space in which there are not sufficient injections to accurately probe. We use 200,000 injections to calculate ξ. Our results never rail against the $N_{\rm eff}$ cut of Eq. (5.9), so we do not believe it affects our results. For further investigation, we perform a set of analyses *without* including spins when calculating $N_{\rm eff}$ (see Appendix 5.G), under which our conclusions do not change.

In addition to including a cut on effective samples from the selection function, one can also impose a cut on the *per-event* effective samples of the posteriors used in calculating the hierarchical likelihood. Here, instead of evaluating Eq. 5.10 on the found injections, it is evaluated on the *posterior samples* for every event. If any events in the catalog have an effective sample number below some threshold, the corresponding Λ sample is tossed. In this work, we do not include any per-event $N_{\rm eff}$ cuts in the sampling of $\mathcal{L}(\{d\}|\Lambda)$, but calculate them post-facto as a check of Monte Carlo convergence (see e.g. Fig. 5.16). Other tests of Monte Carlo convergence are discussed in Ref. [281].

5.D Spin population models

We recover the simulated populations with two models. Generically, we factorize the population models as

$$\begin{aligned}\pi_{\rm pop}(\lambda|\Lambda) = p(m_1|\Lambda)\, p(m_2|m_1,\Lambda)\, p(z|\Lambda)\times \\ p(\chi_1|\Lambda)\, p(\chi_2|\Lambda)\, p(\cos\theta_1|\Lambda)\, p(\cos\theta_2|\Lambda)\,,\end{aligned} \tag{5.11}$$

meaning, aside from the masses m_1 and m_2, there are no correlations in the population. Our two parameterized models for the spin magnitude χ and tilt angle θ are described below. More details about the two models are provided

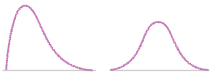
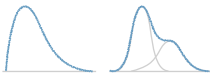
Model Name	Sketch: $p(\chi)$ and $p(\cos\theta)$	Parameter	Prior	Comments
Beta+Gaussian		μ_χ σ_χ μ_θ σ_θ	U(0,1) U(0.07,0.5) U(-1,1) U(0.16,0.8)	Component spin model that cannot reproduce the simulated populations; used to study model misspecification
Beta+DoubleGaussian		μ_χ σ_χ $\mu_{\theta,1}$ $\sigma_{\theta,1}$ $\mu_{\theta,2}$ $\sigma_{\theta,2}$ f	U(0,1) U(0.07,0.5) U(-1,1) U(0.16,0.8) U(-1,1) U(0.16,8) U(0,5)	Component spin model that can reproduce the simulated populations; used to study the amount of information available in component spins. See Appendix 5.G about methods of breaking the degeneracy between the two Gaussian components.

Table 5.1: Details about the two component spin models we employ. Columns give the model names, an example χ and $\cos\theta$ plot, the parameters they depend on, the parameter priors, and some brief comments. The notation $\mathrm{U}(a,b)$ means a prior uniform between a and b. For $p(\chi)$ in both models, we impose an additional prior requirement that the beta distribution shape parameters $\alpha, \beta > 1$, see Eq. (5.14), making the distribution non-singular at the boundaries. Full expressions for $p(\chi)$ and $p(\cos\theta)$ can be found in Eqs. (5.13), (5.16), and (5.17).

in Table 5.1. Spin magnitudes χ_i and tilt angles θ_i are assumed identically and independently distributed.

During hierarchical inference, we fix the distributions of primary mass m_1 and redshift z to truth, as described in Appendix 5.A. To account for possible correlations between spins and mass ratio, although no underlying correlation was injected, we follow Ref. [365] and simultaneously infer the distribution of binary mass ratios and spins using a secondary mass distribution of

$$p(m_2|m_1) \propto m_2^{\beta_q} \quad (m_{\rm min} \leq m_2 \leq m_1) \ , \tag{5.12}$$

where the power-law index β_q is a free parameter with a Gaussian prior of $\mathcal{N}(0,3)$. The true underlying distribution has $\beta_q = 0.96$. For all other individual-event parameters, we take population distributions identical to the priors used during the original Bilby parameter estimation. Most notably for this analysis, azimuthal spins are distributed uniformly $\phi_i \in [0, 2\pi)$. All parameters aside from masses, redshift, and spin magnitudes and tilt angles are thus excluded from hierarchical inference.

Beta+Gaussian

Following the Default model in Ref. [283], we assume that spin magnitudes χ_i are identically and independently distributed according to a Beta distribution

$$p(\chi_i|\alpha,\beta) = \frac{\chi_i^{\alpha-1}(1-\chi_i)^{\beta-1}}{B(\alpha,\beta)}\,, \tag{5.13}$$

where $B(\alpha,\beta)$ is the Beta function which ensures that the distribution is normalized to unity on $0 \leq \chi \leq 1$. Instead of sampling in the shape parameters α and β, we sample in the more familiar mean μ_χ and standard deviation σ_χ which are related to α and β by

$$\alpha = \mu_\chi \nu\,, \quad \beta = (1-\mu_\chi)\nu\,, \tag{5.14}$$

where

$$\nu = \frac{\mu_\chi(1-\mu_\chi)}{\sigma_\chi^2} - 1\,. \tag{5.15}$$

We adopt uniform priors on μ_χ and σ_χ and impose an additional cut such that $\alpha,\beta \geq 1$ to keep the distribution bounded. This cut enforces $p(\chi_i|\alpha,\beta) = 0$ at $\chi_i = 0$ and 1. Aside from the HighSpinPrecessing population at $\chi = 1$, this assumption is valid, and even there the spin model captures the over-all distribution's shape.

For the tilt-angle distribution, we adopt a truncated, normalized Gaussian distribution

$$p(\cos\theta_i|\mu_\theta,\sigma_\theta) = \mathcal{N}_{[-1,1]}(\cos\theta_i|\mu_\theta\,,\sigma_\theta) \tag{5.16}$$

on the interval $-1 \leq \cos\theta \leq 1$, and fit for the mean μ_θ and standard deviation σ_θ.

Beta+DoubleGaussian

The Beta+DoubleGaussian model uses the same spin magnitude distribution as the Beta+Gaussian, as given in Eq. (5.13) and explained thereafter. The tilt angle distribution is here instead given by a mixture of two truncated normalized Gaussians

$$\begin{aligned} p(\cos\theta_i|\mu_{\theta,1},\sigma_{\theta,1},\mu_{\theta,2},\sigma_{\theta,2},f) = \\ f\mathcal{N}_{[-1,1]}(\cos\theta_i|\mu_{\theta,1},\sigma_{\theta,1}) \\ + (1-f)\mathcal{N}_{[-1,1]}(\cos\theta_i|\mu_{\theta,2},\sigma_{\theta,2})\,, \end{aligned} \tag{5.17}$$

to capture the multimodality of the some of the underlying distributions. We measure the means $\mu_{\theta,1}$, $\mu_{\theta,2}$ and standard deviations $\sigma_{\theta,1}$, $\sigma_{\theta,2}$ of the two Gaussians, and the mixing fraction f between them. We impose $\mu_{\theta,1} \leq \mu_{\theta,2}$ to distinguish between the two components.

5.E Detailed Hierarchical Inference Results

We here show full posteriors on the hyper-parameters for the BETA+ DOUBLEGAUSSIAN component spin population model, as given in Eqs. (5.13) and (5.17), for the HIGHSPINPRECESSING (Fig. 5.6), MEDIUMSPIN (Fig. 5.7), and LOWSPINALIGNED (Fig. 5.8) populations. Results for 70 (pink) and 300 (navy) event catalogs are shown in each figure. These posteriors are compared against non-linear least-squares fit parameters (black dashed) calculated from 50,000 draws per population, representing the best possible fit for the true underlying distributions within the BETA+DOUBLEGAUSSIAN model. Population distributions generated from draws from these posteriors are plotted in Fig. 5.2. For all three populations, as expected, including more events makes hyper-parameter measurements more *precise*. However, adding more events does not necessarily make the results more *accurate*.

The hyper-parameters of the HIGHSPINPRECESSING population (Fig. 5.6) are recovered with minimal bias. While the mean μ_χ of the spin magnitude distribution of the HIGHSPINPRECESSING population is very well constrained, its width σ_χ is slightly underestimated in the case of both the 70 and 300 event catalogs. The means $\mu_{i,\cos\theta}$ are also accurately constrained. The widths of the tilt angle distributions also seem to be under-estimated, but per Fig. 5.2, the actual shape of the distribution converges on the truth. This is because – due to the allowed bimodality – different combinations of hyper-parameters can lead to the same unimodal distribution.

The MEDIUMSPIN population (Fig. 5.7), on the other hand, is reconstructed very accurately by the BETA+DOUBLEGAUSSIAN model for both the 70 and 300 event catalogs. Each "true" hyper-parameter either falls within the 90% measured credible region. This is also reflected in Fig. 5.2 – the black traces representing truth are enclosed by the 90% credible envelopes for both χ and $\cos\theta$.

Finally, the width of the spin magnitude distribution for the LOWSPINALIGNED population (Fig. 5.8) is accurately constrained, but its mean μ_χ is over-

estimated. The most striking failure of the BETA+DOUBLEGAUSSIAN model is its inability to identify the bimodality of the LOWSPINALIGNED population's tilt angle distribution, for either 70 or 300 events. Aside from the mixing fraction f, none of the tilt-angle distribution hyper-parameters' posteriors are consistent with truth at the 90% level.

For all three populations, the power law slope for the mass distribution is recovered within 90% credibility about the injected value of $\beta_q = 0.96$. Masses are recovered without bias by our hierarchical inference procedure; we only encountering biases when fitting for the spin populations.

5.F Reweighting individual event posteriors

Given a set of discrete sample from an individual-event posterior $p(\lambda|d_i)$ calculated with prior $\pi_{\rm pe}(\lambda)$ and discrete samples of hyper-parameters describing a population distribution $\pi_{\rm pop}(\lambda|\Lambda)$, we can reweight the individual-event posterior to the inferred population using a two-step algorithm [162, 350]. First, randomly select a hyper-parameter sample from the population distribution $\Lambda_i \in \{\Lambda\}$ and calculate the following weights for each individual-event posterior sample λ_j:

$$w_j \propto \frac{\pi_{\rm pop}(\lambda_j|\Lambda_i)}{\pi_{\rm pe}(\lambda_j)} \,. \tag{5.18}$$

Second, select one sample $\lambda_j \in \{\lambda_j\}$ subject to the weights w_j. Repeat this process to build up a set of samples from a reweighted individual-event posterior. This procedures ensures that events are not double-counted during weighting [350].

5.G Verification of methods

To explore the origin of the bias observed in Fig. 5.2, we perform a number of explorations that we elaborate upon in subsequent subsections. In Appendix 5.G.i, we generate probability-probability (P-P) plots for reweighted individual event BILBY posteriors, which return unbiased. In Appendix 5.G.ii, we turn to hierarchical inference and explore a range of simulated Gaussian individual-event spins posteriors, rather than ones generated via stochastic sampling with BILBY. We also reduce the complexity of the parameter spaces explored on both an individual-event and population level. In Appendix 5.G.iii, we present results where we fix various combinations of parameters to their true values in both the individual-event and hierarchical inference levels. We here also

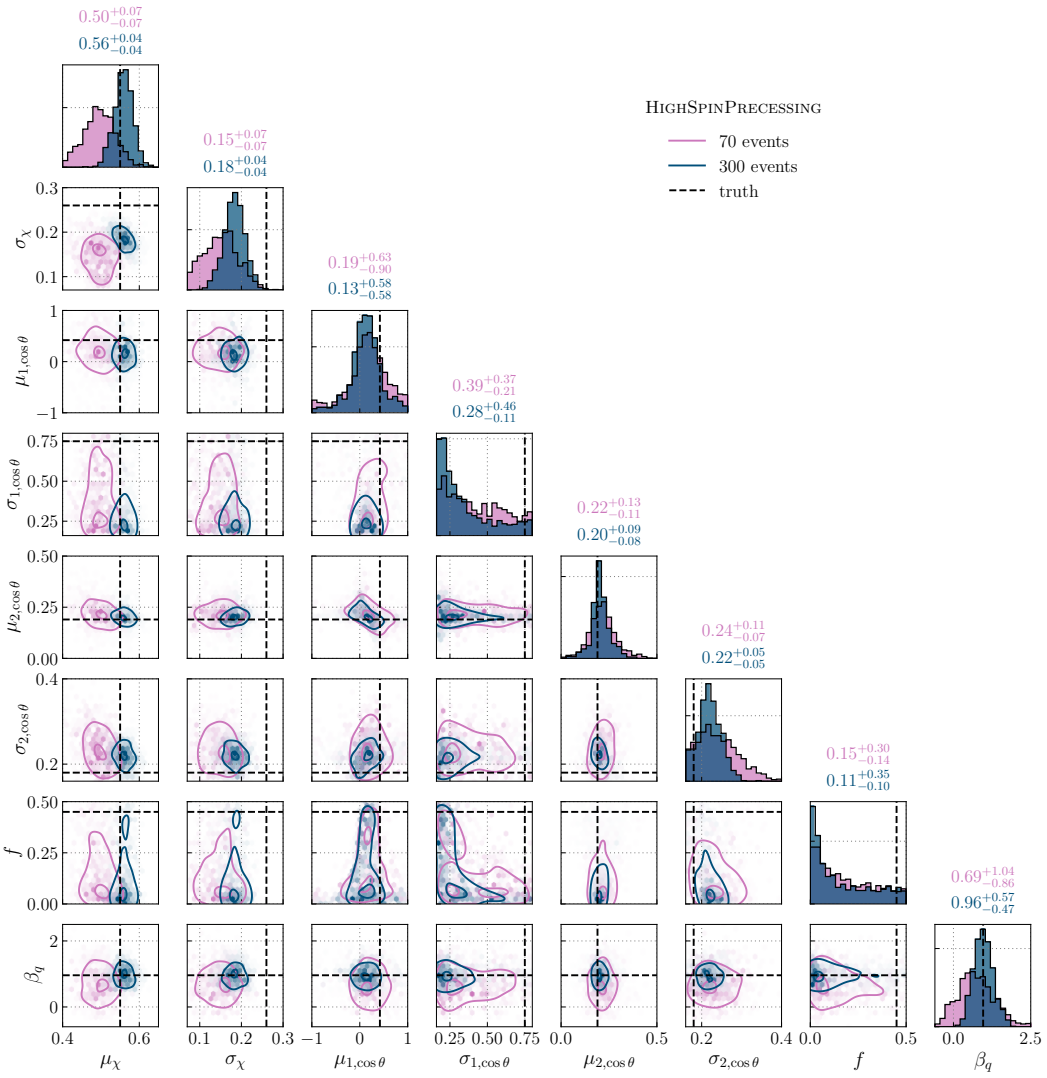


Figure 5.6: The posterior distributions on the hyper-parameters of the spin magnitude and tilt angle distributions under the Beta+DoubleGaussian model for the HighSpinPrecessing population for 70 (pink) and 300 (navy) event catalogs. The labels above each one-dimensional posterior give the medians and 90% credible intervals on each hyper-parameter for the two different catalog sizes, while the contours in each two-dimensional posterior denote the 50% and 90% credible regions. See Table 5.1 for descriptions the hyper-parameters and their priors. Black dashed lines labeled "truth" represent the theoretical best-fit parameters for the population under the Beta+DoubleGaussian model, as calculated using a least-squared fit on 50,000 draws from the population.

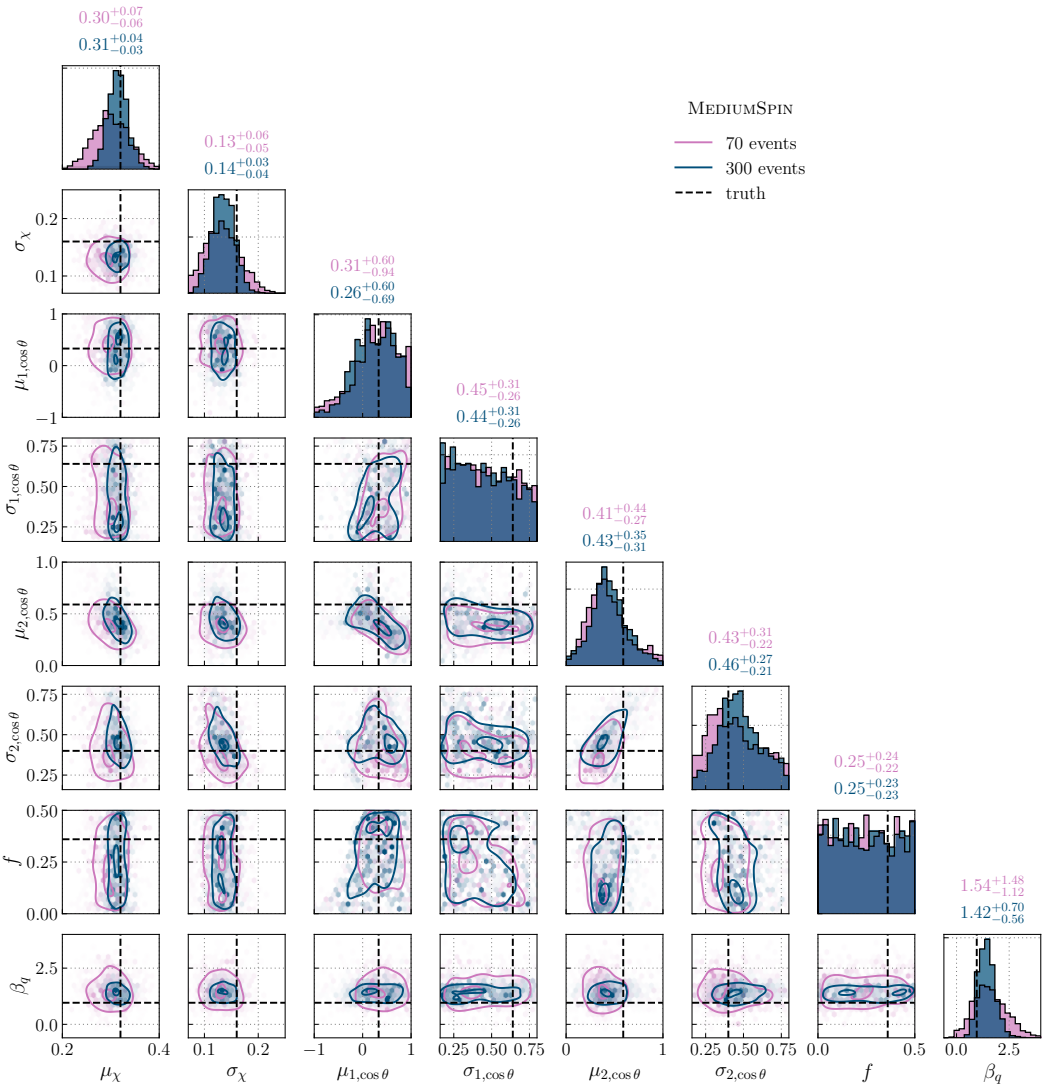


Figure 5.7: Same as Fig. 5.6 but for the MediumSpin population. This population is recovered by the Beta+DoubleGaussian model without bias.

discuss using a less complex waveform – IMRPhenomXP rather than IMRPhenomXPHM – which excludes higher order modes, in individual-event sampling. Appendix 5.G.iv uses an alternate hierarchical inference code, implemented in Numpyro instead of emcee, and Appendix 5.G.v shows results from simultaneously inferring for the mass and redshift distributions along with the spins. In Appendix 5.G.vi, we look at hierarchically-inferred rates across parameter space rather than probability density functions to ensure that the normalization is not obscuring the results. Finally, other miscellaneous checks for the hierarchical inference framework and implementation are: excluding selection effects in spin; trying different methods of breaking the

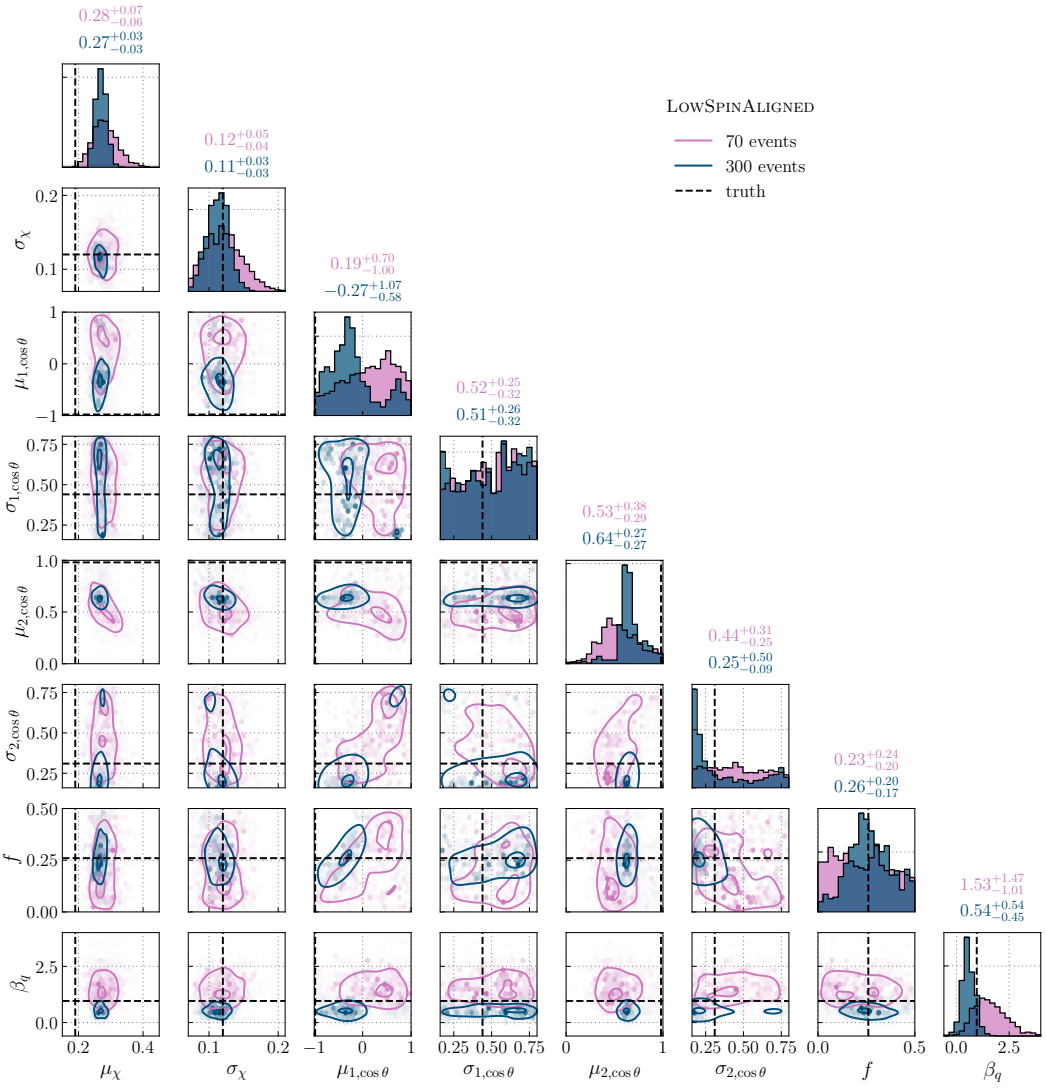


Figure 5.8: Same as Figs. 5.6 and 5.7 but for the LowSpinAligned population. This population is recovered by the Beta+DoubleGaussian model with considerable bias.

degeneracy in the double-Gaussian tilt distribution; and looking at different 70-event catalog instantiations. Plots showing these results can be found in Appendix 5.G.vii.

(i) Probability-probability (P-P) plots

A crucial assumption of hierarchical inference is that the input individual-event posteriors are themselves reliable. To test this assumption and ensure that the stochastically-sampled Bilby individual-event posteriors (see Appendix 5.B) are indeed unbiased, we perform the common diagnostic check of generating a probability-probability (P-P) plot [344, 345].

A P-P plot is generated by performing parameter estimation on events with parameters distributed according to their individual-event priors, in Gaussian noise. The percentiles, or credible intervals (CI), at which the injections fall in their resultant one-dimensional, marginalized posterior distribution shall be uniformly distributed if parameter estimation is unbiased. In our case, where the injected distribution does *not* match the priors used in parameter estimation, reweighting (see Appendix 5.F) to the injected distribution must be performed as a post-processing step. Specifically, we apply an optimal SNR cut of 10 to the posteriors, and then reweight to the underlying population; this procedure is analogous to *not* applying any SNR cut and reweighting to the detected distribution.

P-P plots for spin magnitudes, spin tilt angles, masses, and redshifts for the 300 injections per simulated population are shown in Fig. 5.9. On the horizontal axis, the (sorted) CIs are plotted. The vertical axis shows the cumulative density function (CDF) of these CIs, i.e. the frequency at which each CI occurs. This should be a diagonal line with a slope of 1 in the case of infinitely many injections: e.g. 20% of the time, the injection should fall within the lower 20% CI of its posterior. In the case of finitely many injections, these CDFs should roughly fall within a 3-σ region around the diagonal, the width of which is a function of the number of events injected, as indicated by the gray lines in Fig. 5.9.

To check if the Bilby posteriors pass the P-P test, we look at the p-values[4] that each set of y-axis values shown in Fig. 5.9 is uniform. Then, we take the p-values *of* these p-values, which should also be uniformly distributed if the

[4]We calculate p-values using a Kolmogorov-Smirnov test.

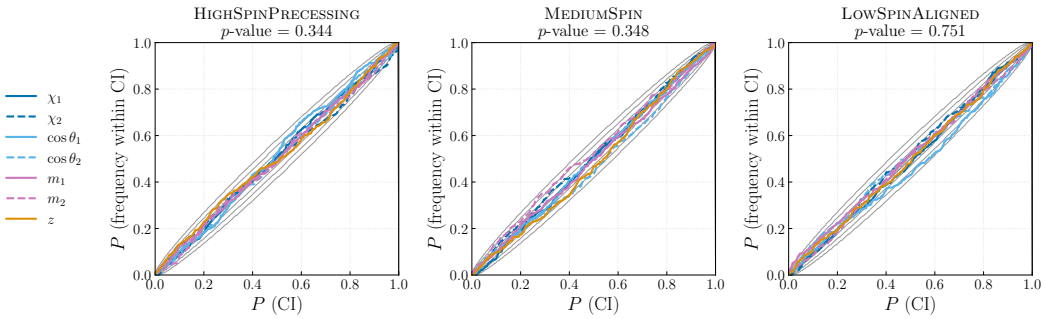


Figure 5.9: Probability-probability (P-P) plots for spin magnitudes (dark blue), spin tilt angles (light blue), masses (pink) and redshifts (orange) for each simulated population (from left to right: HighSpinPrecessing, MediumSpin, LowSpinAligned). For 300 events per population, confidence intervals (CIs; horizontal axis) are plotted against the fraction of events for which the true, injected value is recovered in that CI, stochastically sampled using the Bilby implementation of the nested sampler Dynesty. The 1-, 2-, and 3-σ regions for 300 events are plotted in gray; all parameters stay within the 3-σ region, corresponding to the outermost gray lines. The p-values for each of the three populations are greater than the threshold for 7 parameters (~ 0.143), indicating that the P-P test is passed.

sampling error is random. The p-values (listed in the titles of Fig. 5.9) for each of the three simulated populations is above the threshold of randomness expected from the 7 parameters plotted ($1/7 \sim 0.143$), indicating that the Bilby posteriors pass the P-P test.

Though a necessary check, diagonal P-P plots are not a sufficient condition for reliable individual-event posteriors. As also seen in [346], a sampling algorithm can pass a P-P test but still result in biased hierarchical inference recovery beyond the 3-σ level. In Fig. 5.8, for example, the truth for the means of the spin magnitude distribution and both modes of the spin tilt distribution all lie outside of the 90% credible interval for the recovered values. It remains unclear in our case whether such discrepancies between the true and recovered populations are due to individual-event sampling issues that are not picked up by the P-P test, as was the case in [346], or further unknown biases.

(ii) Simulated Gaussian individual-event spin posteriors

To better understand the relation between individual-event and population-level measurement uncertainty, we generate a series of simulated individual-event spin magnitude and tilt angle posteriors for each of the same 300 GW events per population that we stochastically sample in Bilby. We take these

mock posteriors to be Gaussian distributed with width $\sigma_{\rm meas}$.

First, we generate a series of mock posteriors without any underlying spin-spin correlations with the following steps. For each of the 300 injections per population,

1. Take the true, injected value of each spin parameter

$$\lambda_{\rm true} \in \{\chi_1, \chi_2, \cos\theta_1, \cos\theta_2\}\,,$$

   and from it draw an observed maximum likelihood value $\lambda_{\rm obs}$ from the Gaussian distribution $\mathcal{N}(\lambda_{\rm true}, \sigma_{\rm meas})$ with mean $\lambda_{\rm true}$ and width $\sigma_{\rm meas}$.

2. Draw N samples from $\mathcal{N}_{[a,b]}(\lambda_{\rm obs}, \sigma_{\rm meas})$ where N is the number of samples in the Bilby posterior for the injection of interest and $\mathcal{N}_{[a,b]}$ is a Gaussian distribution *truncated* on $[a, b]$. For spin magnitude this truncation is between $[0, 1]$, and for the cosine tilt angle $[-1, 1]$.

Specifically, we look at cases where $\sigma_{\rm meas} = 0.1$, 0.3, and 0.5, as shown in Fig. 5.3. In all cases, we keep the Bilby mass and redshift posteriors. Moreover, the simulated and Bilby posteriors all have the same number of samples per event.

To simulate a more realistic case, we also generate a set of mock Gaussian posteriors that do include underlying inter-spin correlations, with the same covariance as Bilby individual-event posteriors. For each injection, we first find the covariance of the corresponding four-dimensional Bilby posterior for $\{\chi_1, \chi_2, \cos\theta_1, \cos\theta_2\}$. We then generate a mock four-dimensional spin posterior with that same covariance using the procedure enumerated above: from truth, draw an observed maximum likelihood value; then generate a posterior by sampling a truncated Gaussian centered at that observed value. The only difference is, instead of separately generating each one-dimensional posterior for magnitudes and tilts, we generate a four-dimensional posterior that includes correlations.

The Beta+DoubleGaussian population model *is* able to recover the underlying populations when using the simulated Gaussian posteriors, as seen in Fig. 5.10. This is true for the most-informative individual-event mock-posteriors ($\sigma_{\rm meas} = 0.1$; light blue dotted), the least informative ($\sigma_{\rm meas} = 0.5$; blue dashed), and the most realistic (the posteriors with correlations, labeled

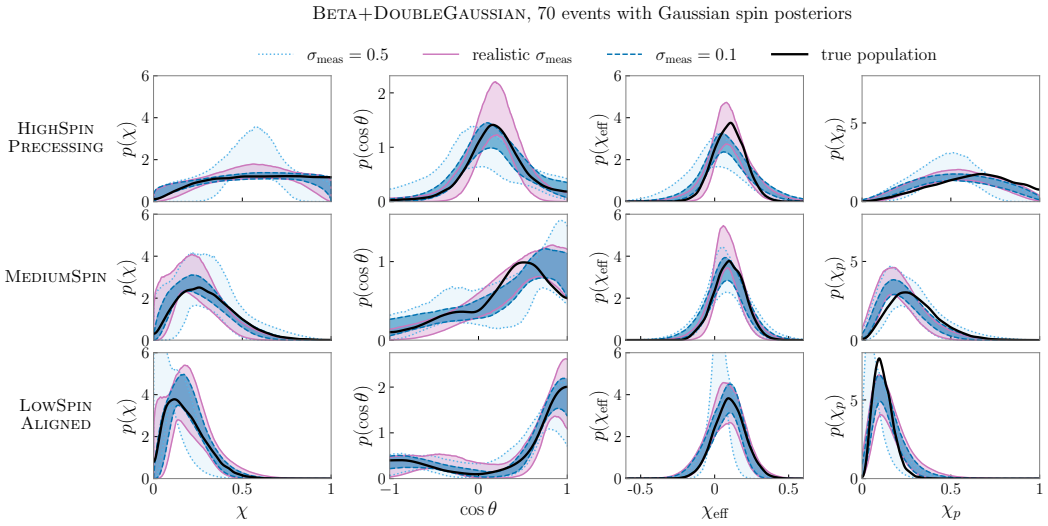


Figure 5.10: Inferred distributions obtained with the BETA+ DOUBLEGAUSSIAN model for spin magnitude χ, spin tilt $\cos\theta$, effective $\chi_{\rm eff}$ spin, effective precessing spin $\chi_{\rm p}$, for the three simulated populations with *simulated* Gaussian individual-event spin posteriors. Results are shown from 70 event catalogs with per-event spin magnitude and tilt measurement error $\sigma_{\rm meas} = 0.5$ (light blue dotted), $\sigma_{\rm meas} = 0.1$ (blue dashed), and with realistic $\sigma_{\rm meas}$ and inter-parameter correlations taken from the BILBY posteriors (pink solid). The shaded region denotes 90% of the probability, while the black solid line corresponds to the true population.

"realistic $\sigma_{\rm meas}$"; pink solid). In all three cases, the true population lies within the 90% credible interval of the recovered region. As the measurement error decreases, the constraints get tighter around truth.

(iii) Reducing complexity of the explored parameter spaces

In this section, we present two simplified scenarios for individual-event sampling in BILBY, then one simplified scenario for population inference in EMCEE.

On the side of individual-event inference, we first reduce complexity by reducing the number of parameters sampled over. The results shown in the main text use posteriors where all fifteen dimensions of parameter-space are sampled over (see Appendix 5.B). To simplify, we first conduct parameter estimation on all the same injections, but fixing their extrinsic parameters (i.e. everything aside from masses and spins) to the true, injected values. This yields the population constraints shown in the pink solid lines in Fig. 5.11 (for 100 events), labeled "fixed extrinsic parameters." We then simplify further and additionally fix the masses and spin azimuthal angles to truth, generating the orange

solid lines in Fig. 5.11 and labeled "fixed extrinsic parameters + masses." Notably, the "fixed extrinsic + masses" individual-event posteriors yield population constraints that are significantly improved from those shown in Fig. 5.2. Since sampling convergence is more challenging as the dimensionality of the explored parameter-space increases, the trend we observe suggests that convergence might at least partially contribute to the bias.

Next, we return to sampling over all parameters (masses, spins, and extrinsic), but this time with a simpler waveform model: IMRPhenomXP [245]. Coming from the same family as IMRPhenomXPHM, the IMRPhenomXP model does not contain higher order modes, which help break degeneracies between BBH parameters. The yellow (light blue) dashed lines in Fig. 5.11 show the population constraints from the same 70 (300) events as Fig. 5.2 but with individual-event posteriors sampled with IMRPhenomXP. The recovered HighSpinPrecessing and MediumSpin populations have a worse mismatch with the truth than in Fig, 5.2, but the LowSpinAligned population is recovered marginally better. Higher order modes become more important to accurately constrain BBH parameters as the degree of spin precession increases. Thus, the fact that the HighSpinPrecessing population is the worst constrained by IMRPhenomXP is consistent with our understanding of the utility of higher order modes. We emphasize that these findings are unrelated to waveform systematics: we always inject and recover with the *same* waveform model.

On the population level, to reduce the complexity of the sampling, we conduct analyses where we fit for *only* the spin magnitude *or* the tilt angle distribution, while fixing the other to it's true injected value. In theory, this could help identify if one or the other of these parameters was the driving factor for the mismatch between the true and recovered populations seen in Fig. 5.2. The inferred spin magnitude distribution for the LowSpinAligned population under the Beta+DoubleGaussian model with the tilt distribution fixed to truth is shown in blue in the top panel of Fig. 5.12; the bottom panel shows the inverse. These recoveries are indeed better than those in Fig. 5.2 (plotted in navy dashed lines for comparison), i.e. the mean of the χ distribution and mean of the larger sub-population of the $\cos\theta$ distribution are both more accurate. However, even in this much simplified version, the Beta+DoubleGaussian model again fails to recover the truth, and in particular still shows no signs of

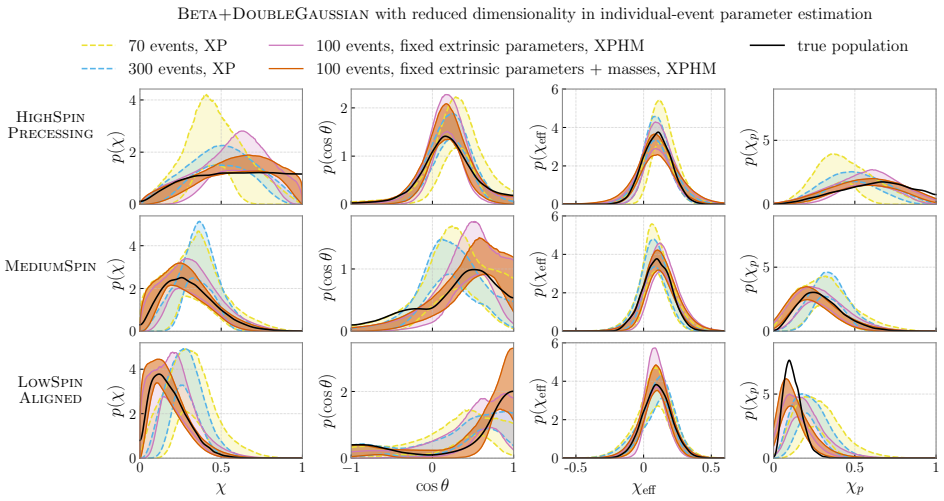


Figure 5.11: Inferred distributions obtained with the BETA+ DOUBLEGAUSSIAN model for spin magnitude χ, spin tilt $\cos\theta$, effective $\chi_{\rm eff}$ spin, effective precessing spin $\chi_{\rm p}$, for the three simulated populations with some form of reduced complexity when the sampling of individual-event posteriors. The dashed lines show population results with the IMRPHENOMXP waveform (used for both injection and recovery), which excludes higher order modes, for 70 (yellow) and 300 (blue) event catalogs. The solid lines shown populations results from 100-event catalogs using individual-event posteriors calculated with IMRPHENOMXPHM, but with various parameters fixed to their true values rather than sampled over. In pink ("fixed extrinsic"), we have fixed the extrinsic parameters (i.e. everything aside from masses and spins) to their true values. In orange ("fixed extrinsic + masses"), we further restrict by *only* sampling over spin magnitudes and tilts. Of these variations, only the "fixed extrinsic + masses" individual-event posteriors yield population constraints that are improved from those shown in Fig. 5.2.

bimodality in the tilt distribution.

(iv) Hierarchical analysis with independent codes

It is always possible that our poor recovery of component spin distributions is simply due to an unidentified error in the code used to perform hierarchical inference. As a safeguard against this possibility, we have repeated the hierarchical analysis of the HIGHSPINPRECESSING, MEDIUMSPIN, and LOWSPINALIGNED populations using a second, distinct body of code. This alternate analysis code was developed entirely independently, and furthermore relies on a different stochastic sampler: whereas our main hierarchical inference results are obtained using EMCEE, this alternative performs inference using NUMPYRO [259, 260], a probabilistic programming library implemented

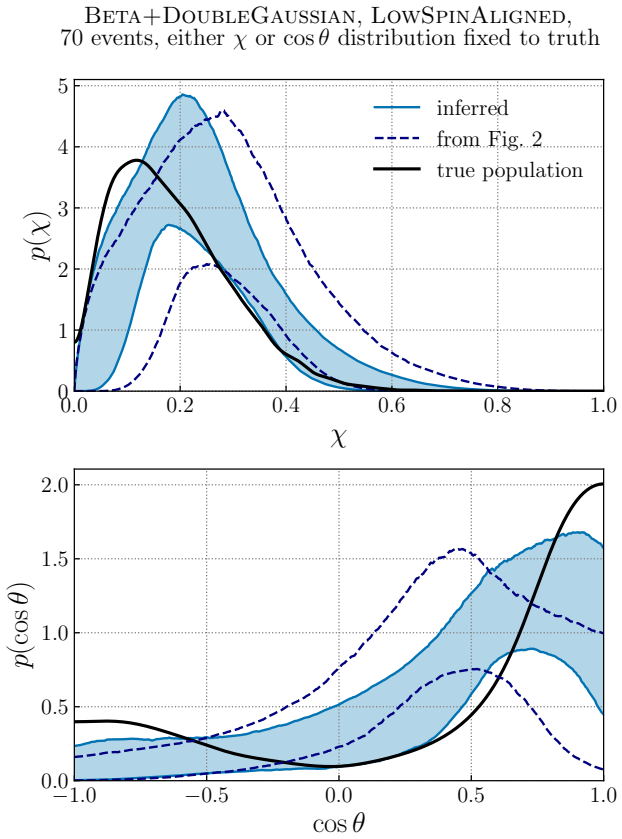


Figure 5.12: (*Top*) Inferred spin magnitude χ distribution (blue shaded) for 70 events from the LOWSPINALIGNED population under the BETA+DOUBLEGAUSSIAN population model with the tilt-angle distribution fixed to truth (black line in the bottom subplot). (*Bottom*) Inferred cosine of the tilt angle $\cos\theta$ distribution for the same 70 events and model with the spin magnitude distribution fixed to truth (black line in the top subplot). In both subplots, 90% credible intervals for the distributions inferred by fitting for *both* the spin magnitude and tilt distributions simultaneously, as shown in Fig. 5.2, are shown in navy dashed lines.

with JAX [258]. The NUMPYRO-based code produces results nearly identical to those obtained with our EMCEE-based code. This implies that our results are not attributable to an unidentified error, unless that same error was independently introduced into two bodies of code created by two different analysts.

***(v)* Simultaneously fitting the mass and redshift distributions**

Yet another source of potential bias we investigate is the choice to fix, rather than fit, the binary black hole primary mass and redshift distributions. In principle, we do not expect significant covariance between the inferred primary mass, redshift, and component spin distributions; our simulated astrophysical populations have no underlying correlations between mass, redshift, and spin. At the same time, inferred component spins are expected to correlate strongly with the *mass ratio* distribution, which in turn can depend systematically on the choice of primary mass distribution [191, 346]. Furthermore, it is known that assumptions regarding spin magnitudes can at times affect inference of the high-redshift rate of black hole mergers [282, 283]. Given these possibilities, it is possible that fixing the presumed mass and redshift distributions (even fixing them to the *correct values*, as we have done) introduces bias into our spin measurements.

To check this, we repeat our inference but now hierarchically inferring the black hole mass and redshift distributions alongside the component spin distributions. We model primary masses following the POWERLAW+PEAK model [282, 283] and assume that the merger rate density evolves with redshift as $(1+z)^{\kappa}$ for some parameter κ. This model also uses a slightly different spin magnitude model: a truncated Normal distribution instead of a Beta distribution. We perform this inference using the alternative NUMPYRO-based code introduced in Appendix 5.G above. The spin distributions inferred in this case are shown in Fig. 5.13. The results are extremely similar to those in Fig. 5.2. As before, we recover the HIGHSPINPRECESSING and MEDIUMSPIN component spin distributions reasonably well, but do not successfully measure the LOWSPINALIGNED distributions. In this latter case, we miss (or misplace) the bimodality inherent in $\cos\theta$ and, accordingly, systematically overestimate component spin magnitudes. Once more, though, the $\chi_{\rm eff}$ distribution is well-recovered in all three cases.

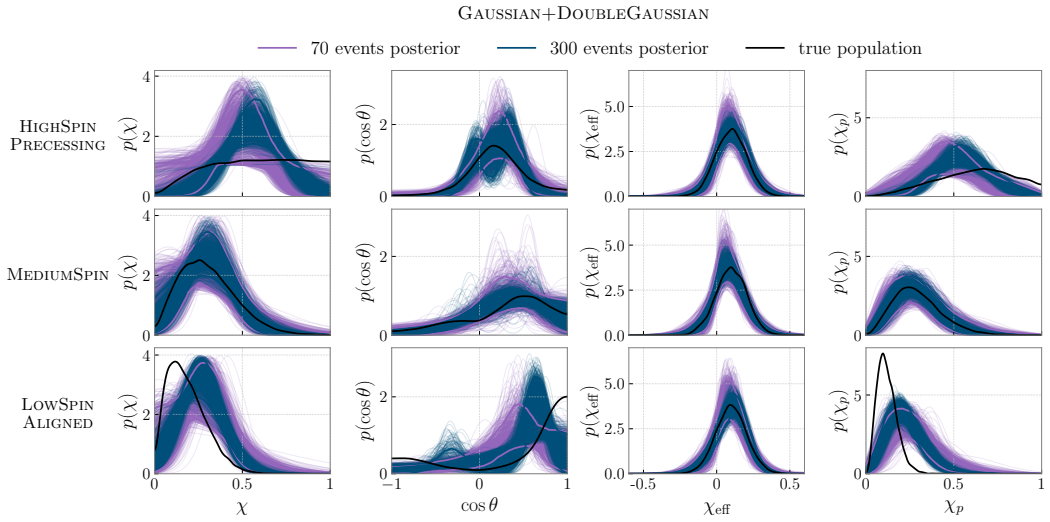


Figure 5.13: As in Fig. 5.2, but now when additionally inferring the mass and redshift distributions of the three simulated populations in conjunction with their spin distributions. These results are furthermore produced with an entirely independent body of code, using NUMPYRO rather than EMCEE to stochastically sample the population likelihood. Despite these differences, the results are nearly identical to those in Fig. 5.2, indicating that the difficult recovery of injected component spin distributions is due neither to our choice to fix the mass and redshift distributions in the main text, nor to unidentified errors in our hierarchical inference code.

(vi) Recovering rates and spins simultaneously

GW data are generated according to a Poisson point process, in which individual compact binaries stochastically trace an underlying *rate density* $dR/d\lambda$ of mergers across the space of binary parameters λ. When performing hierarchical inference over GW catalogs, we are formally reconstructing this rate density: measuring the "counts" of events occurring in different regions of parameter space. Often, however, we are concerned only about the shape of $dR/d\lambda$, not its normalization. In this case, it is common to instead study the normalized *probability distribution* $p(\lambda)$ of source parameters. This is achieved after the fact by fitting for the merger rate but only presenting $p(\lambda)$, or from the very outset by marginalizing over and subsequently ignoring the total rate, a procedure that gives Eq. (5.4).

While this procedure is usually well-behaved, there do exist instances in which the choice to present the normalized $p(\lambda)$, rather than a reconstructed rate density, can yield inadvertently misleading conclusions. In some cases, models that successfully recover the correct rate density can appear to fail in recovering

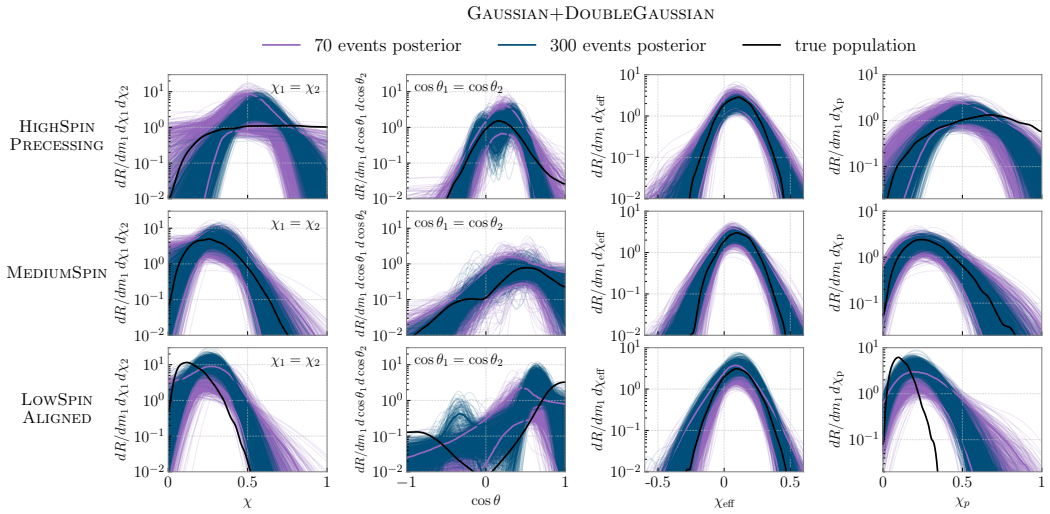


Figure 5.14: As in Fig. 5.2, but now fitting for the absolute rate of black hole mergers as a function of spin, rather than component spin probability distributions. When producing these results, we additionally infer the black hole mass and redshift distributions, as in Fig. 5.13 using NUMPYRO rather than EMCEE. The left-most column shows the differential merger rate $dR/dm_1\, d\chi_1\, \chi_2$ per unit primary and secondary spin magnitude, evaluated along the $\chi_1 = \chi_2$ line. The second column analogously shows the merger rate $dR/dm_1\, d\cos\theta_1\, d\cos\theta_2$ across the $\cos\theta_1 = \cos\theta_2$ line, and the last two columns give the rates and $dR/dm_1\, d\chi_{\rm p}$ as a function of effective spins. In all cases, rates are evaluated at $m_1 = 30\,M_\odot$ and $z = 0.2$. Even when plotting rates, we draw the same qualitative conclusions as originally identified in Fig. 5.2: the HIGHSPINPRECESSING and MEDIUMSPIN populations are recovered well, while the rate density of the LOWSPINALIGNED population is recovered poorly.

the correct probability density; see discussion in Sec. 5B of Ref. [268]. Thus, when evaluating the goodness-of-fit of a given model, the most robust results are obtained by comparing injected and recovered *merger rates*, rather than injected and recovered probability densities.

Given this discussion, does the poor agreement between injected and recovered spin probability distributions signify a true modeling and inference failure? Or is this disagreement illusory, due to our choice to compare probability distributions rather than reconstructed merger rate densities? To check this, we repeat the hierarchical analyses of the three simulated populations but now fitting for the overall merger rate alongside the hyperparameters governing the component spin distributions. As in Appendix 5.G, we simultaneously infer the primary mass and redshift distributions. Figure 5.14 shows our inferred merger rates as a function of spin for each injected population. Our initial

conclusions hold: when simultaneously fitting for and presenting differential merger rates, rather than probability densities, we still find that the HighSpinPrecessing and MediumSpin populations are recovered well, but we do not successfully recover the LowSpinAligned population. Hence our poor recovery of LowSpinAligned is a real effect, rather than a bias or misleading visualization related to our choice to marginalize over the absolute merger rate.

***(vii)* Other miscellaneous checks for hierarchical inference**

Finally, we present results from other miscellaneous verification methods for our hierarchical inference procedure. First, we investigate different methods of breaking the degeneracy between the two Gaussian components in tilt-distribution portion of the Beta+DoubleGaussian model, see Eq. (5.17). For any model defined as a mixture of multiple components, some method must be imposed to break the degeneracy these components. For a bimodal Gaussian, this can be done in three ways:

1. Imposing an ordering of the *means* – assign "distribution 1" to be that with the smaller mean and "distribution 2" to be that with the larger mean
2. Imposing an ordering of the *widths* – assign "distribution 1" to be that which is narrower, and "distribution 2" to be wider
3. Limiting the *mixing fraction* be ≤ 0.5 – assign "distribution 1" to be that which contains a smaller fraction of events, and "distribution 2" to be that which contains a larger fraction.

Sometimes one method of breaking the degeneracy converges better when used in a hierarchical inference procedure than another. We find that this is not the case in this work: different methods perform identically (within sampling error), as seen in Fig. 5.15. Using the means (mixing fraction) of the Gaussians to break the degeneracy yields the distributions plotted in blue (orange). The results are consistent with each other. We opt to use the mixing fraction to break degeneracy throughout the bulk of this work because it is more computationally efficient.

Next, to ensure there is no misspecification in the selection function for spins, see Eq. (5.8), we run hierarchical inference without any spin selection effects.

Results are shown for the LowSpinAligned population pink in Fig. 5.15 for 70 (dashed) and 300 (solid) events. Selection effects in component spins are not strong, and thus are not expected to effect population inference significantly.[5] This is indeed the case, as the distributions inferred without spin selection effects are nearly identical to those inferred with them (shown in orange). We also note that for all results shown in this work, the number of effective samples does not rail against the cut given in Eq. (5.9).

Our final check is to conduct hierarchical inference on several different random 70-event catalog realizations from the 300 total events per population. Just like different Gaussian noise instantiations of the data lead to variance in individual-event posteriors, so too can random catalog instantiation lead to variance in the recovered posteriors on the *population* parameters. Some catalogs will yield a more accurately recovered population than others, just by random chance from working with finite numbers, see, e.g. Ref. [164]. Fig. 5.16 shows some expected variance in results, but nothing corresponding to the degree of mismatch between the true and inferred tilt distributions of the LowSpinAligned population seen in Fig. 5.2. We therefore conclude that we cannot attribute bias between the injected and recovered LowSpinAligned population to be from an "unlucky" catalog realization. Additionally, each catalog instantiation leads to a different number of per-event effective samples, see Eq. (5.10) and corresponding discussion in Appendix 5.C, which we find within a given population are *not* correlated to the by-eye goodness of fit, as seen in the rightmost column of Fig. 5.16. However, the LowSpinAligned population yields, on average, the lowest N_{eff} values and is the least accurate fit. This leads us to believe that the absence of inferred bimodality is not due to the issue of our events not having enough effective samples, although the fact that the minimum event-level N_{eff} over catalog instantiations is small could be another source of imperfect recovery. As part of future work, we plan to explore the uncertainty in difference in log-likelihood formulated in Ref. [281] as another way to gauge whether our Monte Carlo estimations avoid bias.

[5] Selections effects are strong for masses and redshift, on the other hand. This can be seen when comparing the underlying and detected distributions in Fig. 5.5.

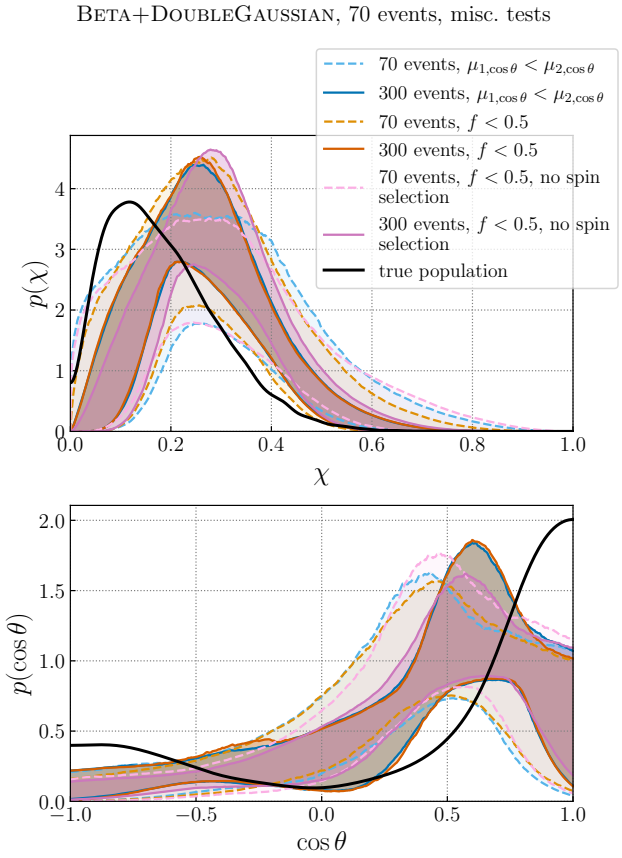


Figure 5.15: Inferred distributions (blue) for spin magnitude χ (top subplot) and cosine of the tilt angle $\cos\theta$ (bottom subplot) for 70 events from the LOWSPINALIGNED population obtained with the BETA+DOUBLEGAUSSIAN population model under various conditions. All results from 70 (300) event catalogs are plotted with dashed (solid) lines. In blue and orange we compared two methods of breaking the degeneracy between the two Gaussian sub-populations in the BETA+DOUBLEGAUSSIAN tilt distribution: first, restricting the mean of distribution 1 to be smaller than that of distribution 2 ($\mu_{1,\cos\theta} < \mu_{2,\cos\theta}$; blue), and second, restricting the mixing fraction to be less than one half, effectively assigning distribution 1 to be that containing less events ($f < 0.5$; orange). In orange and pink, we compare results where we do (orange) versus do not (pink) include spin selection effects (both using the $f < 0.5$ method of breaking degeneracy). None of these variations produce population constraints improved from those shown in Fig. 5.2; the bimodality of the tilt-angle distribution remains unrecovered.

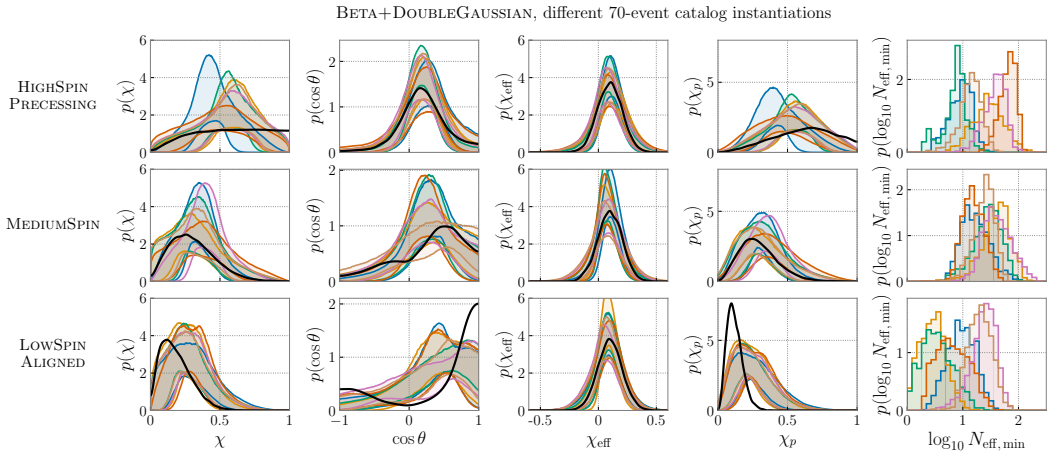


Figure 5.16: Inferred distributions obtained with the BETA+DOUBLEGAUSSIAN model for spin magnitude χ, spin tilt $\cos\theta$, effective $\chi_{\rm eff}$ spin, effective precessing spin $\chi_{\rm p}$, for various 70-event catalog instantiations of the three simulated populations, each plotted in a different color. The true, underlying populations are shown in black for comparison. While there is a small amount of variation, none of the catalog instantiations yields population constraints improved from those shown in Fig. 5.2. The rightmost column shows the minimum per-event $N_{\rm eff}$ over the catalog used for inference for that particular analysis (See Eq. (5.10) and corresponding discussion in Appendix 5.C). Within a given row, the minimum $N_{\rm eff}$ does not correlate with the by-eye goodness of fit of the distributions in the first four columns.

Part II

Where in gravitational-wave signals do measurements of binary black hole spin originate?

Chapter 6

TIME-DOMAIN INFERENCE

Most LVK transient analyses are conducted in the frequency domain, because detector noise is assumed to be well-described by a stationary, Gaussian process (in the absence of glitches). If we additionally assume periodic boundary conditions (e.g., apply windowing), then the noise covariance matrix in the frequency domain is diagonal, leading to an efficient-to-calculate likelihood, as given in Eq. (3.11).

However, there are scenarios in which the assumptions necessary for Fourier domain likelihood calculation break down. Notably, it is not possible to assume periodic boundary conditions when analyzing data with a sharp truncation in time. Here, important parts of the signal fall at the segment's edge and would be contaminated by windowing. Non-windowed data cannot be transformed to the frequency domain without spectral leakage.[1] We also cannot truncate in the frequency-domain, as there is not generically a one-to-one correspondence between time and frequency. The remedy of calculating the likelihood directly in the time-domain was piloted for ringdown studies (see Sec. 2.5) and has since been used for other applications, such as those described in the subsequent chapters in Part II.

6.1 Time-domain likelihood

The likelihood that discretely sampled time-domain (TD) strain data $d = \{d_0, \ldots, d_{N-1}\}$ where $d_i \equiv d(t_i)$ with noise described by a covariance matrix C_{ij} contains a signal $h = \{h_i\}$ is defined as follows [367]:

$$\ln P(d \mid h) = -\frac{1}{2} \sum_{i,j=0}^{N-1} (d_i - h_i)\, C_{ij}^{-1}\, (d_j - h_j). \tag{6.1}$$

Assuming the segment analyzed is bounded by $t_{\rm start}$ and $t_{\rm end}$, the index 0 corresponds to the first sample at which $t_i \geq t_{\rm start}$ and the number of data points in the time series is $N = \mathrm{int}[f_s \cdot (t_{\rm end} - t_{\rm start})]$ for a sampling frequency f_s.

[1] For a detailed discussion of windowing and the ways it can impact GW data analysis, see Ref. [366].

The above likelihood definition stems from modeling instrumental noise as a discrete-time random process, represented by a set of random variables $n = \{n_0, \ldots, n_{N-1}\}$. As in Ch. 3, the data are $d_i = h_i + n_i$. Assuming Gaussianity, the noise-generating random process is entirely defined by its mean and covariance $C_{ij} = E[n_i n_j]$, where without loss of generality the mean has been shifted to zero.

While inverting matrices is generally computationally costly, evaluating Eq. (6.1) can be sped up if the covariance matrix C_{ij} is pre-processed into a convenient form. For stationary random processes, C_{ij} depends only on the *distance* between the two points i and j, taking the form $C_{ij} = \rho(|i-j|)$ where ρ is the *auto-covariance function* (ACF). Using the ACF gives C_{ij} the following structure:

$$C_{ij} = \begin{pmatrix} \rho_0 & \rho_1 & \cdots & \rho_{N-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{N-2} \\ \vdots & & \ddots & \vdots \\ \rho_{N-1} & \cdots & \rho_1 & \rho_0 \end{pmatrix}, \tag{6.2}$$

which is a symmetric, non-circulant Toeplitz matrix that can be exploited for computational efficiency. In practice, we construct the noise covariance matrix using an aperiodic estimate of the ACF, calculated with a truncated inverse Fourier transform of a power spectral density (PSD) describing the data. The PSD must be generated from data substantially longer than the segment of interest.

As a part of this dissertation, I developed the open-source Python package `tdinf` [161] to conduct TD inference on GW signals. The code makes use of the above Toeplitz matrix decomposition to render time-domain likelihood calculation comparably efficient to the Fourier-domain approach for short-duration GWs.

6.2 Time-truncated analyses

Figure 6.1 illustrates the specific application of TD inference described in the following chapters. To isolate which portions of data inform the measurement of spin (and other BBH parameters), I *(1.)* truncate a GW signal at some time, *(2.)* separately infer source parameters from data before and after the truncation, and then *(3.)* iterate through cutoffs (i.e., truncation times) to track how constraints evolve. For the cartoon example in Fig. 6.1, the pre-

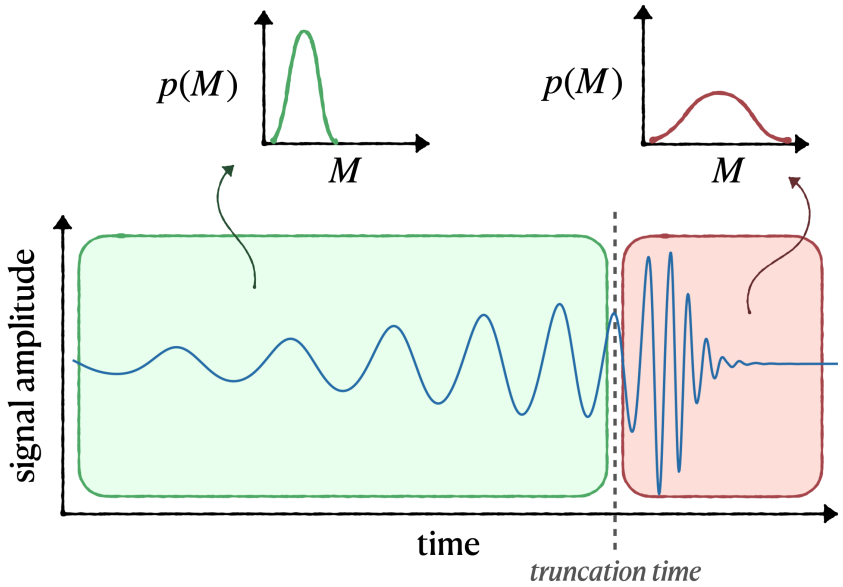


Figure 6.1: Cartoon explaining pre/post-cutoff signal analysis, enabled by time-domain inference: a GW signal (blue) is truncated in time (gray dashed line) and measurements of source parameters are made from the data before (green) or after (red) that cutoff. In this example, the pre-cutoff data contain more information about total mass M than the post-cutoff data, indicated by a narrower posterior probability distribution $p(M)$. *Graphic by* ***SJM****.*

cutoff data (green) contain more information about the binary's total mass M than the post-cutoff data (red), indicated by a narrower posterior $p(M)$.

I have used the TD inference framework towards several ends:

1. I localized the elusive spin precession of GW190521 and connected it to underlying motion of the orbital plane (Ch. 7).

2. I used GW190521-like simulations to show that *(i)* at large enough SNRs, precession can be measured in the ringdown data alone; *(ii)* the morphological imprints of precession in high-mass signals are not generic; and *(iii)* precession measurability is drastically impacted by the inclination and phase angles (Ch. 8).

3. I helped link the measurement of GW191109's spin below the orbital plane to a region of data concurrent with a scattered-light glitch (Ch. 10).

4. I helped generate the first high-confidence observational confirmation of Hawking's Area Law using the high-SNR GW250114 (Ch. 11).

I have also begun tracing the imprint of spin precession on GW231123—the most massive BBH detected to date—with preliminary findings indicating that it is localized to a different region of the signal than GW190521. A preliminary comparison of the two signals is included in the Pocket Material at the end of this dissertation.

Chapter 7

TRACING IMPRINTS OF SPIN-PRECESSION ON THE MASSIVE BLACK HOLE BINARY GW190521

This chapter is published as Ref. [101]:

Simona J. Miller, Maximiliano Isi, Katerina Chatziioannou, Vijay Varma, and Ilya Mandel. GW190521: Tracing imprints of spin-precession on the most massive black hole binary. *Phys. Rev. D*, 109(2):024024, 2 2024. ***SJM*** *assisted in the project's conceptualization, conducted all analyses, wrote the text, and made all figures.*

Abstract

GW190521 is a remarkable gravitational-wave signal on multiple fronts: its source is the most massive black hole binary identified to date and could have spins misaligned with its orbit, leading to spin-induced precession; an astrophysically consequential property linked to the binary's origin. However, due to its large mass, GW190521 was only observed during its final 3-4 cycles, making precession constraints puzzling and giving rise to alternative interpretations, such as eccentricity. Motivated by these complications, we trace the observational imprints of precession on GW190521 by dissecting the data with a novel time domain technique, allowing us to explore the morphology and interplay of the few observed cycles. We find that precession inference hinges on a quiet portion of the pre-merger data that is suppressed relative to the merger-ringdown. Neither pre-merger nor post-merger data alone are the sole driver of inference, but rather their combination; in the quasi-circular scenario, precession emerges as a mechanism to accommodate the lack of a stronger pre-merger signal in light of the observed post-merger. In terms of source dynamics, the pre-merger suppression arises from a tilting of the binary with respect to the observer. Establishing such a consistent picture between the source dynamics and the observed data is crucial for characterizing the growing number of massive binary observations and bolstering the robustness of ensuing astrophysical claims.

7.1 Introduction

At a total mass of $\sim 150\,M_\odot$, GW190521 [20, 107] is the current record-holder among massive black hole binaries confidently detected through gravitational waves by LIGO [4] and Virgo [5]. Such high-mass systems are essential probes of the role of hierarchical mergers [43, 65, 67, 368, 369] and pair-instability physics [125, 370–374] in binary formation and evolution. Observationally, massive binaries merge toward the low edge of the detectors' bandwidth and are only detectable for a short time. One third of the binaries in the latest gravitational-wave catalog have a median detector-frame total mass $> 100\,M_\odot$ [287], corresponding to $\sim$5 signal cycles at more than half a standard deviation above the noise. The short duration makes characterizing these signals and inferring astrophysical properties such as spin challenging.

Spin is a key signature of the physics behind angular momentum transport in stellar interiors, black-hole formation, black-hole retention in dense environments, and more [1, 51, 61, 85, 93, 329]. Gravitational-waves provide one of the few ways to measure spins for stellar-mass black holes directly. Spin components *parallel* to the binary's orbital angular momentum affect the signal duration [196] and are approximately conserved during the inspiral as the "effective spin" [193, 289]. Spin components *perpendicular* to the orbital angular momentum, i.e., *in the orbital plane*, cause the binary to precess, leading to signal modulations as the emission pattern varies relative to the line of sight [199, 312]. Although typically weak [98, 143, 314], this effect is highly sought-after: spin-induced precession and the associated in-plane spins could differentiate between dynamical and field binary formation, e.g., [1, 45, 62].

The elusiveness of precession is exacerbated for heavy systems. The precession timescale can be longer than the observed inspiral for large masses [78], making modulations difficult to identify. The exact imprint of precession on the ensuing merger and ringdown remains poorly understood and analytically intractable, although numerical-relativity and data-analysis studies suggest that imprints do exist [103, 151–154, 156]; for example, Ref. [103] suggests, based on simulations, that high-frequency data typically associated with the merger-ringdown can constrain precession. This uncertainty makes it difficult to distinguish precession from eccentricity, another highly-valuable binary property [99, 290, 375]. Interpretation is further complicated by the high sensitivity of the measurability of precession to the system's true param-

eters [103, 151, 201, 376] and the priors [377].

In light of this, the massive system that sourced GW190521 stands out for its informative precession constraint, when analyzed assuming a quasicircular orbit. Precession can be quantified by the effective precessing spin $\chi_{\rm p}$ [77, 200, 203] that is motivated by inspiral dynamics. A value of zero (one) indicates no (maximal) precession. Under the assumption of a quasicircular orbit, GW190521 has $\chi_{\rm p} = 0.68^{+0.25}_{-0.37}$ at 90% credibility [20, 107], the largest inferred $\chi_{\rm p}$ and the one whose posterior is the most informative to date [285–287];[1] similar conclusions are reached under alternative parametrizations for precession [50, 205].[2] The combined high mass and large in-plane spin make GW190521 an essential probe of hierarchical black hole mergers [63, 368, 378], dense stellar environments such as nuclear star clusters [61, 379, 380], active galactic nuclei disks [42, 120, 381], and more [107, 292, 347, 374, 382–384].

The high mass of GW190521 and its few observable cycles open the door to competing astrophysical interpretations. Ref. [290] and Ref. [291] find that the data are consistent with eccentricity, though this interpretation is not supported by Ref. [385] and Ref. [386]. Ref. [293] proposes a hyperbolic capture scenario. Ref. [292] suggests a highly asymmetric, but still precessing, binary interpretation. More exotic explanations include boson stars [387] and cosmic strings [388]. Any of these alternatives would have important implications if confirmed [389, 390]. Additionally, random detector noise can have an outsized impact on the inference of poorly-constrained effects, although Ref. [103] and Ref. [104] show that the inference of $\chi_{\rm p}$ away from zero in GW190521-like systems cannot be due to Gaussian noise alone. The fact that full-scale parameter estimation allows for competing interpretations suggests that different physical effects can result in similar observational imprints over GW190521's few cycles. Similarly to precession and eccentricity, these imprints are often not analytically tractable.

Toward bolstering the interpretation of massive binaries, it is essential to gain intuition about the observable imprint of physical effects of interest and how their measurability is affected by mismodeling. Lacking analytical equations

[1] GW200129, another massive binary [287], has a potentially comparable constraint. When averaged over waveform models, its $\chi_{\rm p}$ is unconstrained [287]; when restricting to the `NRSur7dq4` [16] model, GW200129 is inferred to be precessing [294]. This interpretation is however complicated by data quality issues [159].

[2] These and further ways to quantify precession are elaborated upon in Appendix 7.C.

for precession in the merger phase, we introduce a novel approach that traces its imprint along the signal and identifies the role of each *cycle* on the $\chi_{\rm p}$ constraint. We dissect the data in the *time-domain* and compare inference between different data subsets. We provide a cycle-by-cycle physical picture of source dynamics and explore the interplay of different data regions. Our work focuses on the data aspect that drives the inference of precession within a quasi-circular merger scenario. Extensions to further physical effects of interest such as orbital eccentricity can be tackled under a similar framework; we leave those to future work.

7.2 Methods

Gravitational-wave parameter estimation is typically conducted in the frequency domain for computational efficiency. Leveraging the stationarity of detector noise, noise components at different frequencies are independent which leads to a diagonal covariance matrix in likelihood calculations [284, 391]. However, frequency-domain methods are non-local in time; thus isolating temporal features of source dynamics and their imprint on the data requires nontrivial likelihood modifications [215, 392].[3] We instead adopt direct time-domain inference to isolate different signal cycles, an approach originally conceived for black hole ringdowns [222, 367, 395, 396]. We truncate data from LIGO Livingston, LIGO Hanford, and Virgo at different times ranging from $t = -50\,M_{\rm ref}$ to $50\,M_{\rm ref}$ with respect to coalescence.[4] We independently infer the signal properties solely from data before and after each cutoff as well as the full span of data.

We model the signal with the numerical relativity surrogate model `NRSur7dq4` [16], which assumes quasi-circular orbits and includes precession and higher-order modes. Within its region of validity, `NRSur7dq4` displays the lowest mismatches against numerical relativity among existing models [16]. We adapt the time-domain inference code from Ref. [395] and sample the multidimensional posterior for the binary masses, spin magnitudes and tilt angles, azimuthal

[3] Frequency truncation enables consistency checks [393, 394], investigations of data-quality issues [146, 159], or alternative studies of the measurability of precession in simulated data [103], but this is not equivalent to cuts in time.

[4] We define t with respect to geocenter GPS time 1242442967.405764 s. Under geometric units we adopt the median detector-frame remnant mass scale $M_{\rm ref} = 1.27\,{\rm ms}$ [107]; in standard units $M_{\rm ref} = 258.3\,M_\odot$. The choice of remnant rather than total mass, was inspired by ringdown analyses [214].

inter-spin angle, azimuthal precession cone angle, inclination, luminosity distance, and phase of coalescence. The time of coalescence, right ascension, declination, and polarization angle are fixed for computational efficiency.[5].

We report precession constraints using the canonical effective precessing spin, $\chi_{\rm p}$ [77, 200, 203]:

$$\chi_{\rm p} = \max\left[\chi_1 \sin\theta_1, \left(\frac{3+4q}{4+3q}\right) q\,\chi_2 \sin\theta_2\right] \in [0,1]\,. \tag{7.1}$$

Here $\chi_i \in [0,1]$ are the dimensionless spin magnitudes and θ_i are the tilt angles between the spin and orbital angular momentum vectors. Subscripts $i \in \{1,2\}$ denote each black hole with mass m_i and $q \equiv m_2/m_1 \leq 1$.

7.3 Results

Figure 7.1 shows the inferred GW190521 properties from data before (blue) and after (orange) five representative cutoff times (vertical lines) as well as the full signal (black) for comparison. Insets in the left panels visualize the truncation in LIGO Livingston, selected as the detector in which GW190521 is the loudest. The left column shows the posterior for $\chi_{\rm p}$ at a reference frequency of 11 Hz [20, 107]. For the earliest cutoff time (top row), the post-$t = -40\,M_{\rm ref}$ $\chi_{\rm p}$ posterior is almost identical to that of the full analysis, while the pre-cutoff one is identical to the prior. This is due to the fact that the post-cutoff analysis includes the full available signal, while none of it is contained in the pre-cutoff data (see inset). As the cutoff moves to later times (top to bottom), the pre- and post-cutoff posteriors gradually exchange places as the data preceding each cutoff become more informative and the data following become less so.

The ∼40 ms between $t = -40\,M_{\rm ref}$ and $-10\,M_{\rm ref}$ are crucial to constrain precession for GW190521. This region roughly corresponds to the final cycle before the onset of merger. The $\chi_{\rm p}$ posterior obtained from data after $t = -40\,M_{\rm ref}$ (first row, orange) is consistent with that from the full analysis, i.e., precession *is* constrained.[6] On the other hand, data after $t = -10\,M_{\rm ref}$ (third row, orange), result in a $\chi_{\rm p}$ posterior that is nearly identical to the prior, i.e., uninformative. Between these times, the posterior shifts smoothly between the full measurement and the prior; e.g., the post-$t = -20\,M_{\rm ref}$ analysis (second row,

[5]We have verified that these choices do not affect our conclusions. All parameter estimation settings, priors, and consistency checks are given in Appendix 7.A

[6]By "constrained," we specifically mean "visibly different from the prior."

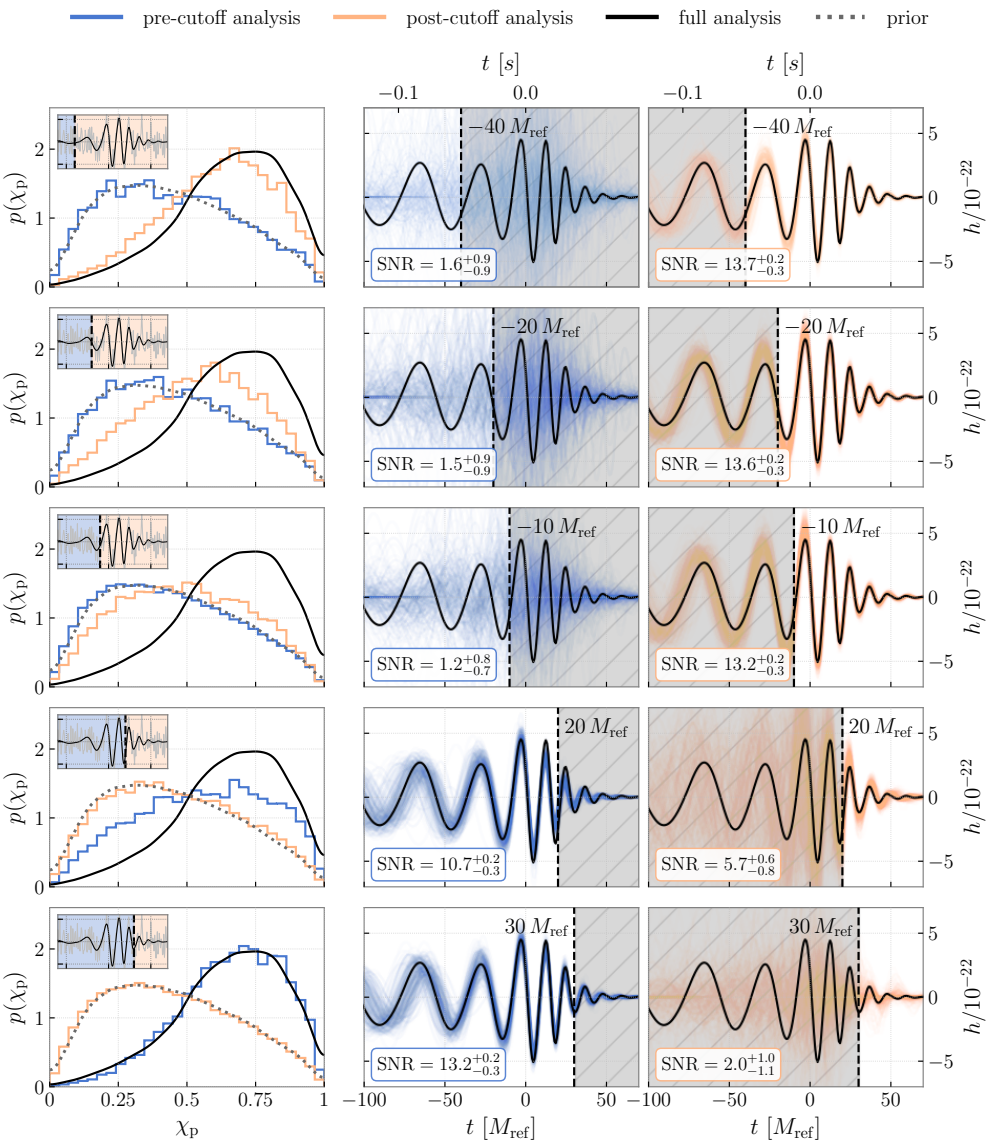


Figure 7.1: Evolution of GW190521 inference for representative cutoff times $t \in \{-40, -20, -10, 20, 30\}\, M_{\rm ref}$ from the time of coalescence (top to bottom; vertical black dashed lines where applicable). *Left:* Posterior for $\chi_{\rm p}$ from the pre- (blue), post-cutoff (orange), and full (black solid) analysis and prior (gray dotted). The inset shows the whitened maximum-posterior waveform (with $\chi_{\rm p} = 0.62$, detector-frame total mass $M = 267\, M_\odot$, and $q = 0.89$) from the full analysis (black) along with the whitened LIGO Livingston data (gray). The blue/orange shaded regions highlight the data informing the same-color $\chi_{\rm p}$ posterior. *Center and Right:* Waveform reconstruction draws for LIGO Livingston from the pre- (center, blue) and post-cutoff (right, orange) analyses and maximum-posterior waveform for the full analysis (black). Median and 90% credible intervals for the matched-filter network SNRs are given in-figure. Gray shading denotes data excluded from each analysis.

orange). The reduction in the signal-to-noise ratio (SNR) from excluding data is negligible between $-40\,M_{\rm ref}$ and $-10\,M_{\rm ref}$ suggesting this qualitative change in precession inference is not due to an SNR drop, see Fig. 7.8 in Appendix 7.D.

Neither the inspiral nor the merger/ringdown data alone are fully responsible for precession constraints in GW190521. The data both pre- and post-$t = -10\,M_{\rm ref}$ alone are uninformative about precession (third row, orange and blue). Moreover, the pre-$t = 30\,M_{\rm ref}$ analysis that excludes the final ringdown cycle (fifth row, blue) is consistent with the full analysis. It is therefore not solely the final pre-merger cycle that informs precession, but rather its combination with the subsequent 2 merger and early-ringdown cycles. This does not rule out ringdown imprints of precession that are too weak to discern at this SNR or with this waveform.

The center and right columns of Fig. 7.1 investigate *features* of the waveforms. The blue and orange waveforms are informed only by data in the unshaded regions and extended coherently into the shaded regions. As progressively less data are analyzed (center bottom to top, right top to bottom), the waveform reconstructions agree less with the full-analysis waveform, eventually becoming incoherent. The right column reveals the morphological imprint of precession on the signal during the transition from an informative $\chi_{\rm p}$ posterior (first row) to the prior (third row). When the $\chi_{\rm p}$ inference returns (close to) the prior, the final pre-merger cycle is extrapolated to be *larger* than when $\chi_{\rm p}$ is constrained to take higher values, c.f., the waveform peak at $t \sim -30\,M_{\rm ref}$ and trough at $t \sim -15\,M_{\rm ref}$. Again there is a progression: the post-$t = -40\,M_{\rm ref}$ inferred waveforms (orange) are consistent with the full analysis (black), while the final pre-merger cycle subtly increases in strength toward post-$t = -10\,M_{\rm ref}$ (top row to third row).

To further explore the pre-merger waveform suppression, we compare the full analysis (black) in which precession *is* constrained and the post-$t = -10\,M_{\rm ref}$ analysis where the data are uninformative about precession in Fig. 7.2. In order to focus on waveform features that are informative compared to the noise, we plot the whitened waveform.[7] In the top panel the grayed-out region

[7] Whitened waveforms are obtained by dividing the Fourier-domain waveform by the noise amplitude spectral density and then inverse Fourier-transforming. See Appendix 7.B for details.

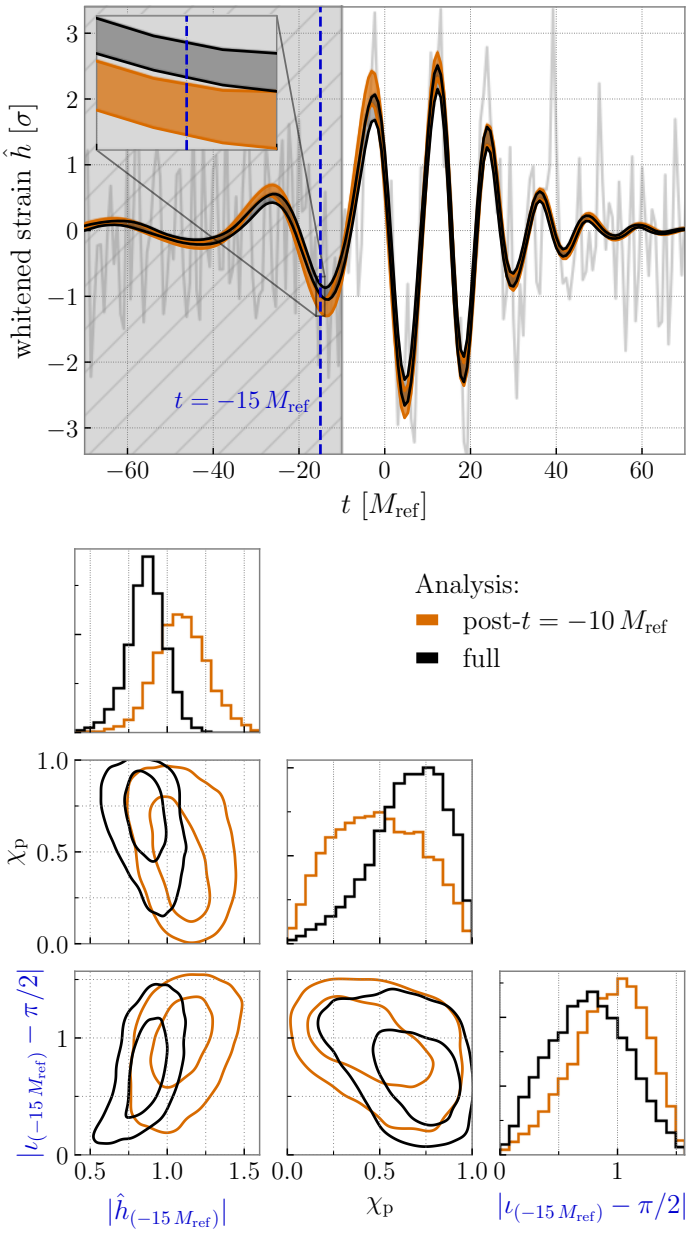


Figure 7.2: Results from the full (black) and from the post-$t = -10\,M_{\rm ref}$ analysis (orange) that are informative and uninformative about precession respectively. *Top:* 50% credible intervals for the whitened waveforms in LIGO Livingston in units of standard deviations of the noise. Data are plotted in gray. Gray shading denotes data excluded from the post-$t = -10\,M_{\rm ref}$ analysis. The inset zooms in around $t = -15\,M_{\rm ref}$ (blue dashed line), the minimum of the final pre-merger cycle. *Bottom:* Posteriors for $\chi_{\rm p}$, the absolute value of the whitened strain $|\hat{h}|$, and the inclination angle relative to edge-on configurations $|\iota - \pi/2|$. Quantities labeled in blue are plotted at $t = -15\,M_{\rm ref}$. Contours denote 50% and 90% credible regions. The whitened strain is anticorrelated with $\chi_{\rm p}$ and correlated with $|\iota - \pi/2|$. Large $\chi_{\rm p}$ is paired with smaller pre-merger signal and more edge-on configurations.

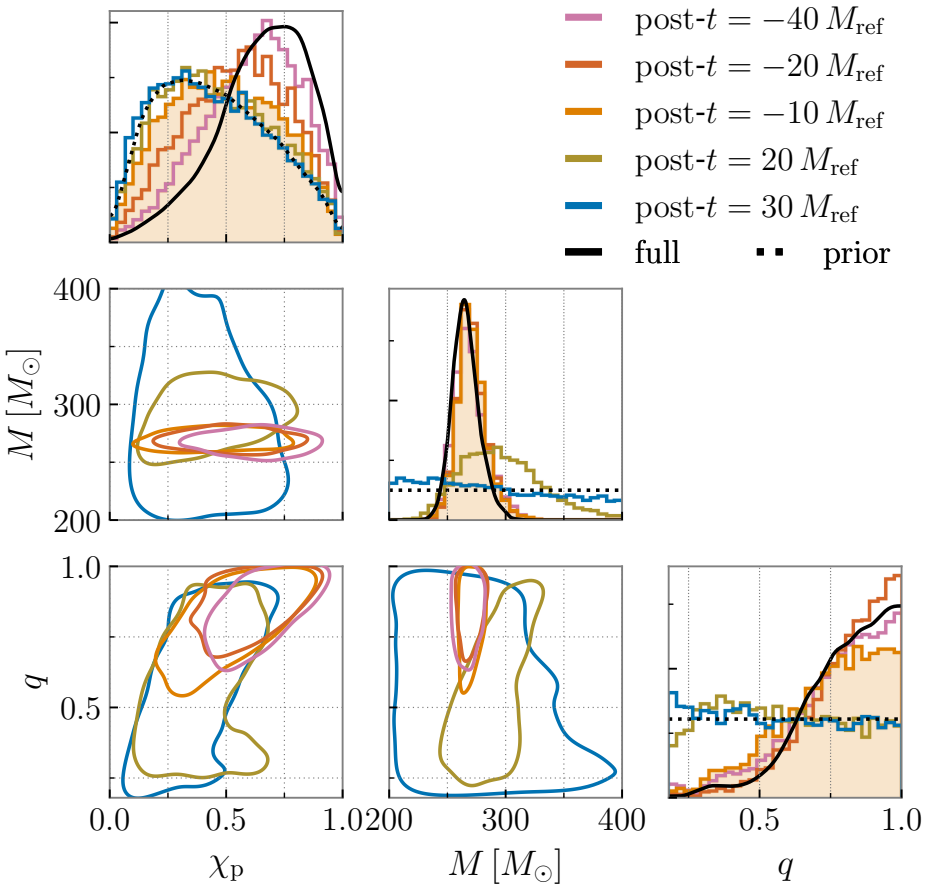


Figure 7.3: Posteriors for $\chi_{\rm p}$, M (detector frame total mass), and q (mass ratio) for the same cutoff times as Fig. 7.1. Posteriors from the full analysis and priors are plotted in black solid and dotted respectively. Contours denote 50% credible regions. Information about spin-precession is lost post-$t = -10\,M_{\rm ref}$ (shaded light orange), while the total mass and mass ratio posteriors are informative at post-$t = 20\,M_{\rm ref}$ (green).

denotes data available to the full analysis but not the post-$t = -10\,M_{\rm ref}$ one. The inset focuses around $t = -15\,M_{\rm ref}$, an extremum of the final pre-merger cycle. The reconstructions are inconsistent at the 50% credible level, with the post-$t = -10\,M_{\rm ref}$ analysis resulting in a larger amplitude (in absolute value). This inconsistency only occurs at the extrema of the final pre-merger cycle, i.e. the peak around $t \sim -30\,M_{\rm ref}$ and trough at $t \sim -15\,M_{\rm ref}$, see Fig. 7.9 in Appendix 7.D.

The bottom panel shows marginal posteriors for select quantities: the effective precessing spin $\chi_{\rm p}$, the absolute value of the whitened strain $|\hat{h}|$ (units of standard deviations σ of the noise), and the difference between ι—the angle

between the direction of maximum signal emission and the line of sight [397]—and $\pi/2$ at $t = -15\,M_{\rm ref}$. As expected from Fig. 7.1, χ_p and $|\hat{h}|$ are anticorrelated (albeit weakly): when the full data are analyzed (black), the final cycle is constrained to be weaker, resulting in a larger $\chi_{\rm p}$; when this cycle is excluded from the analysis (orange), the extrapolated waveform is not required to have such a small value, obviating the need for a higher $\chi_{\rm p}$. In summary, *precession in GW190521 is informed by a suppression of the gravitational-wave in the observed waveform's last cycle before merger*. This is also the region in which the waveform is overall quietest: the whitened signal is less than $1\,\sigma$ above the noise, compared to the subsequent merger cycles that are over $2\,\sigma$.

The origin of the signal suppression can be attributed to the evolution of the emission direction. As a binary precesses, the angle between the dominant emission direction and the line-of-sight evolves, changing the amplitude of the observed signal. Systems with $\iota \sim \pi/2$ are quieter than those with $\iota \sim 0$ or $\sim \pi$. To explore these dynamics, we plot the absolute difference between the inclination angle and $\pi/2$ at the time $t = -15\,M_{\rm ref}$ in Fig. 7.2. As expected, $\chi_{\rm p}$ and $|\iota - \pi/2|$ are anti-correlated, while $|\hat{h}|$ and $|\iota - \pi/2|$ are correlated. When precession is constrained, ι is found to be closer to edge-on at the last pre-merger cycle than when precession is unconstrained, leading to a suppressed cycle and smaller $|\hat{h}|$.

We investigate the impact of data truncation on other source parameters in Fig. 7.3. Information about $\chi_{\rm p}$, M and q is not lost or gained in lockstep as a function of cutoff time. At post-$t = -20\,M_{\rm ref}$ (dark orange), the $\chi_{\rm p}$ posterior shifts away from the full analysis posterior; however, for M and q, this does not happen until multiple cycles later. Post-$t = -10\,M_{\rm ref}$ (shaded light orange), i.e., at the end of the final pre-merger cycle, the $\chi_{\rm p}$ posterior is close to the prior, while the detector-frame total mass M and mass ratio q both resemble the posteriors from the full analysis. Thus, the lack of an informative $\chi_{\rm p}$ posterior post-$t = -10\,M_{\rm ref}$ does not simply arise from poor parameter constraints over-all due to lower SNR; rather, the suppression of the final pre-merger cycle is informative *specifically* about precession(under the assumption that the system has a quasi-circular orbit). We also confirm that the $\chi_{\rm p}$ inference is not driven by a conditional measurement based on the typically better-measured aligned spins in Appendix 7.C.

7.4 Conclusions

Precession inference for the massive, distant binary black hole signal GW190521 is subtle. It originates from contrasting a ∼40 ms slice of data from the final pre-merger cycle between $t = -40\,M_{\rm ref}$ and $-10\,M_{\rm ref}$ with the loud merger cycles following it. The merger of GW190521 is 2 loud cycles that reach $2.5\,\sigma$ above the noise and is informative about the source's masses. However, precession is only constrained away from the prior when the merger is observed in tandem with the final pre-merger cycle, which does not rise more than $1\,\sigma$ above the noise. The measurement is linked to a relative suppression of the aforementioned final cycle, caused by the binary tilting toward an edge-on configuration due to precession. This picture qualitatively agrees with the interpretation posited in Ref. [20] by comparing precessing and spin-aligned waveform reconstructions for the full signal, and is supported by simulations [375].

Ref. [214] carried out a complementary study seeking a description of the the GW190521 ringdown consistent with the `NRSur7dq4` full analysis.[8] Ringdown mode content encodes information about the preceeding binary dynamics [154, 400, 401], meaning it is (in theory) possible to identify signatures of precession in the ringdown. Ref. [214] found support for the presence of at least two modes; consistency with `NRSur7dq4` suggests a configuration including the 220 and 210 fundamental modes. A large 210 mode amplitude could be expected under strong precession [214]. The fact that past GW190521 ringdown-only analyses cannot unequivocally infer precession is consistent with our finding that a post-peak analysis is not sufficient to constrain precession.

Our study highlights the delicate nature of precessional imprints on observed signals, providing a new view of spins in massive systems beyond the frequency domain [20, 103, 104]. For GW190521, the difference between the most informative precession inference to date and the prior boils down to a single, quiet pre-merger cycle that needs to be measured to better than half a standard deviation, *cf.*, the difference between the black and orange waveforms in Fig. 7.2. This is quantitatively in agreement with the conclusions of Ref. [159] who explored the impact of data quality on our ability to obtain an unbiased measurement at that level.

Our novel time-domain approach of tracing the observational imprint of inter-

[8] Ref. [215, 398] provided an alternative interpretation based on waveform models of the PHENOM family [245, 399].

esting physical effects cycle-by-cycle can provide physical intuition about how key source properties are inferred in relation to observed data features. In anticipation of further massive observed signals, such a correspondence between source dynamics and observed data can help pinpoint the most informative data in order to assess data quality and waveform systematics, and enable us to morphologically study competing physical interpretations that are likely to keep arising.

Ch. 7 Data and Code Availability

Posterior samples from all our pre- and post-cutoff analyses – including those from additional cutoff times not included in the main text – are available on Zenodo at https://zenodo.org/record/8349582. We release posteriors for all cutoff times in intervals of $10\,M_{\rm ref}$, from $t = -50\,M_{\rm ref}$ to $50\,M_{\rm ref}$. Cutoff times in intervals of $2.5\,M_{\rm ref}$ between $t = -30\,M_{\rm ref}$ and $20\,M_{\rm ref}$ were additionally explored to more finely resolve the transition between informative and uninformative $\chi_{\rm p}$ posteriors.

Scripts to generate the waveform reconstructions and inclination angles are on Github at https://github.com/simonajmiller/gw190521-timedomain-release, as are notebooks to plot all the figures appearing in the text. The repository additionally contains animations showing results from all pre- and post-cutoff analyses in intervals of $t = 2.5\,M_{\rm ref}$:

- gif_01: Animation of $\chi_{\rm p}$ posteriors, whitened reconstructions, and colored reconstructions for all time slices; similar to Fig. 7.1.
- gif_02: Animation of a corner plot for $\chi_{\rm p}$, $\chi_{\rm eff}$, M, and q at all time slices; similar to Figs. 7.3 and 7.6.

Ch. 7 Acknowledgments

We thank Harrison Siegel for helpful discussions on ringdown analyses of GW190521, Sylvia Biscoveanu for insights about inference on high-mass GW sources, Sophie Hourihane for assistance whitening GW signals, Davide Gerosa for insight on alternative measures of precession, and Will Farr for insights about time-domain inference. We also extend thanks to Jacob Lange, Christopher Berry, Carlos Lousto, Juan Calderon Bustillo, and Salvatore Vitale for their helpful comments on our manuscript.

This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. S.M. and K.C. were supported by NSF Grant PHY-2110111. K.C. acknowledges support from the Sloan Foundation. The Flatiron Institute is funded by the Simons Foundation. V.V. acknowledges support from NSF Grant No. PHY-2309301, and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 896869. I.M. is a recipient of the Australian Research Council Future Fellowships (FT190100574). The authors are grateful for computational resources provided by Cardiff University and supported by STFC grant ST/I006285/1.

Software: `emcee` [252], `LALSuite` [319], `numpy` [323], `scipy` [353], `h5py` [402], `matplotlib` [320], `seaborn` [354], `ringdown` [222, 367], `gwtools` [403].

Ch. 7 Appendices

7.A Parameter estimation settings and priors

During parameter estimation, we sample the masses, spin magnitudes and tilt angles, azimuthal inter-spin angle, azimuthal precession cone angle, inclination angle, luminosity distance, and phase of coalescence. The time of coalescence, right ascension, declination, and polarization angle are fixed for computational efficiency; we have ensured that further sampling over these parameters does not alter our conclusions, as is shown in Fig. 7.4. Priors for all parameters are given in Table 7.1. We use slightly different mass and distance priors than Ref. [20] but – as we show in Fig. 7.5 – obtain consistent results when analyzing the full signal.[9]

Parameter estimation settings are listed in Table 7.2. All analyses make use of the three detector network of LIGO Livingston, LIGO Hanford, and Virgo, using strain and event power spectral densities publicly available on the Gravitational-wave Open Science Center [405, 406].[10] We use a lower and reference frequency of 11 Hz, a maximum frequency of 1024 Hz, and a sampling rate of 2048 Hz.[11] Each analysis conducted *before* a given cutoff time

[9] We use the `NRSur7dq4` [16] posterior samples (specifically `GW190521_posterior_samples.h5`) released by LIGO/Virgo at Ref. [404] for this comparison.

[10] Specifically we download the "32 sec, 16 KHz" strain data uploaded to Ref. [407].

[11] We opt for a minimum frequency of 11 Hz which is consistent with Refs. [286, 287], but not the more recent Ref. [285].

Parameter	Symbol	Value
Detector-frame total mass	$M(1+z)$	$U(200, 500)\,M_\odot$
Mass ratio	q	$U(0.17, 1)$
Primary spin magnitude	χ_1	$U(0, 1)$
Secondary spin magnitude	χ_2	$U(0, 1)$
Primary spin tilt	θ_1	isotropic
Secondary spin tilt	θ_2	isotropic
Azimuthal inter-spin angle	ϕ_{12}	$U(0, 2\pi)$
Azimuthal cone precession angle	ϕ_{JL}	$U(0, 2\pi)$
Inclination angle	ι	isotropic
Luminosity distance	d_L	$U(10^3, 10^4)\,\mathrm{Mpc}$
Phase of coalescence	φ	$U(-\pi, \pi)$

Table 7.1: Priors used in parameter estimation. $U(a, b)$ means uniform between a and b. An isotropic prior for an angle x means that the prior on $\cos x$ is uniform between -1 and 1. Right ascension, declination, polarization, and time of coalescence are fixed to the values given in Table 7.2.

begins at the GPS time at geocenter 1242442966.9077148 s; those conducted *after* a given time all end at 1242442967.607715 s. To translate the cutoff times from geocenter to the times at the three detectors, we use a right ascension of $\alpha = 6.075\,\mathrm{rad}$, a declination of $\delta = -0.8\,\mathrm{rad}$, and polarization angle of $\psi = 2.443\,\mathrm{rad}$.

7.B Whitening Waveforms

Whitened waveforms are obtained by dividing the Fourier-domain waveform by the noise amplitude spectral density and then inverse Fourier-transforming. In the frequency domain, the whitened waveform $\hat{h}(f)$ is obtained from the original waveform $h(f)$ by:

$$\hat{h}(f) = \sqrt{\frac{f_s}{2\,S_n(f)}}\,h(f) \tag{7.2}$$

where $S_n(f)$ is the power spectral density and f_s is the sampling rate of the data. Though we sample the data at 2048 Hz for inference, we use 1024 Hz when plotting whitened waveforms for direct comparison to Ref. [20].

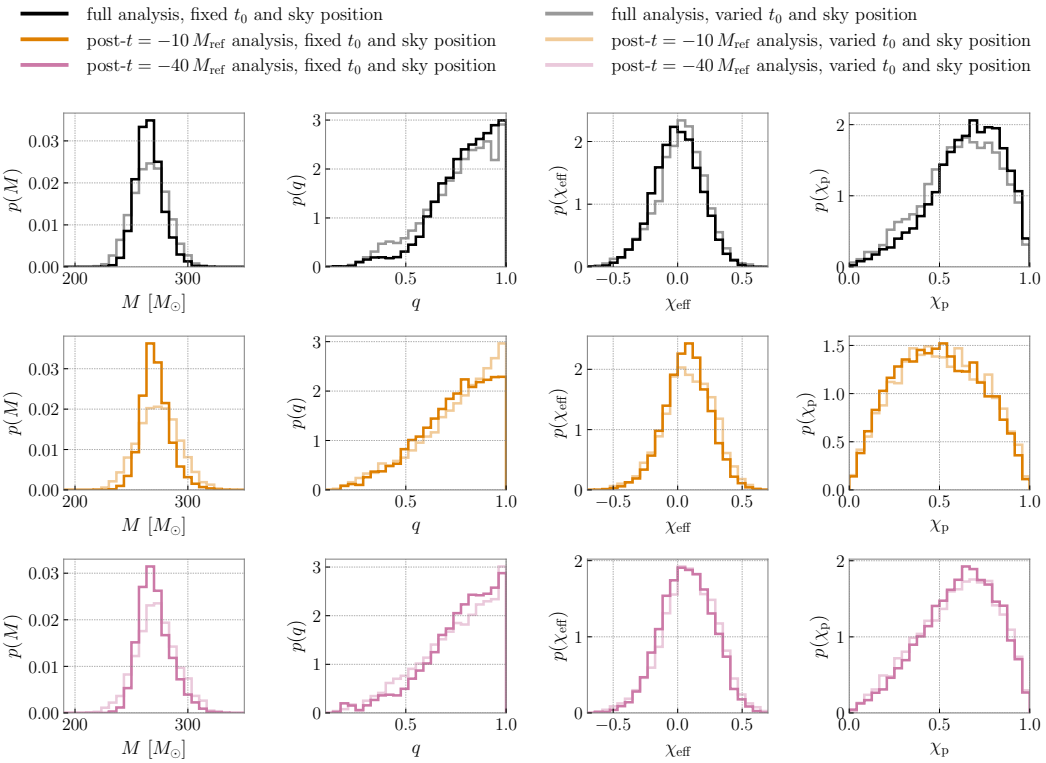


Figure 7.4: Posteriors for the detector-frame total mass M, mass ratio q, effective spin $\chi_{\rm eff}$, and effective precessing spin $\chi_{\rm p}$ for parameter estimation with (lighter color) and without (darker color) sampling over time of coalescence t_0 and sky position – referring to the right ascension, declination, and polarization angle – for the full signal (black; top row), the post-$t = -10\,M_{\rm ref}$ (orange; middle row), and the post-$t = -40\,M_{\rm ref}$ (pink; bottom row). All results are consistent, the only difference being a slightly wider total mass posterior.

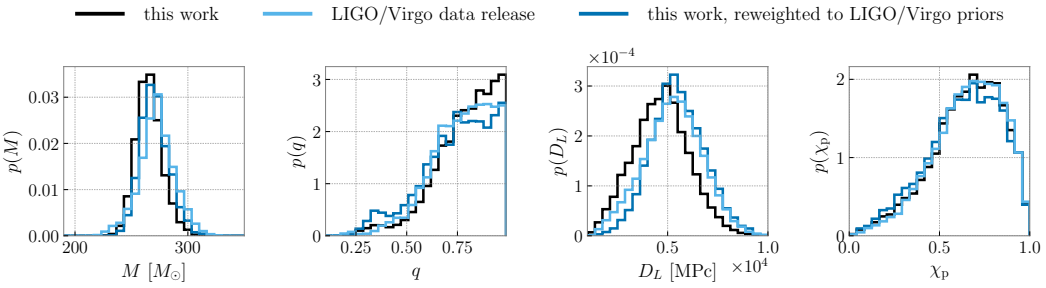


Figure 7.5: Posteriors for the detector-frame total mass M, mass ratio q, luminosity distance D_L, and effective precessing spin $\chi_{\rm p}$ from this work (black), Ref. [20] (LIGO/Virgo data release; light blue), and this work reweighted to the priors of Ref. [20] (dark blue). Our inference is consistent with Ref. [20] under the same priors.

Parameter	Symbol	Value
Start GPS time of "before" segments	$t_{\rm start}$	1242442966.907715 s
End GPS time of "after" segments	$t_{\rm end}$	1242442967.607715 s
Coalescence GPS time	t_0	1242442967.405764 s
Mass-time scaling relation	$1\,M_{\rm ref}$	0.00127 s
Right ascension	α	6.075 rad
Declination	δ	-0.800 rad
Polarization angle	ψ	2.443 rad
Minimum frequency	$f_{\rm min}$	11 Hz
Maximum frequency	$f_{\rm max}$	1024 Hz
Sampling rate	$f_{\rm samp}$	2048 Hz
Reference frequency	$f_{\rm ref}$	11 Hz

Table 7.2: Settings for the time-domain parameter estimation for GW190521. The pre-cutoff analyses all begin at $t_{\rm start}$, while the post-cutoff ones end at $t_{\rm end}$. These times are calculated with respect to the geocenter time t_0 and then shifted between detectors using the extrinsic parameters α, δ, and ψ.

7.C Quantifying precession

Spins aligned with the orbit are characterized at leading order by the effective spin $\chi_{\rm eff}$,

$$\chi_{\rm eff} = \frac{\chi_1 \cos\theta_1 + q\,\chi_2 \cos\theta_2}{1+q} \in [-1, 1]\,, \tag{7.3}$$

which is conserved to at least the second post-Newtonian order in the inspiral [193, 289]. Precession – stemming from mis-aligned spins – is typically characterized via the "effective precessing spin" parameter [203], as defined in Eq. (7.1) in the main text. Posteriors for $\chi_{\rm eff}$ versus $\chi_{\rm p}$ for five representative cutoff times (compared to the full signal posterior and the prior) are shown in Fig. 7.6. Though typically $\chi_{\rm eff}$ is better constrained than $\chi_{\rm p}$, this is not the case for GW190521: $\chi_{\rm eff}$ is never inferred to be far from its prior with only slight deviations from the prior that occur at different times from when we best constrain $\chi_{\rm p}$. Furthermore, because the $\chi_{\rm p}$ and $\chi_{\rm eff}$ are not correlated at any time slice, inferences of $\chi_{\rm p}$ – and thus our main conclusions – do not

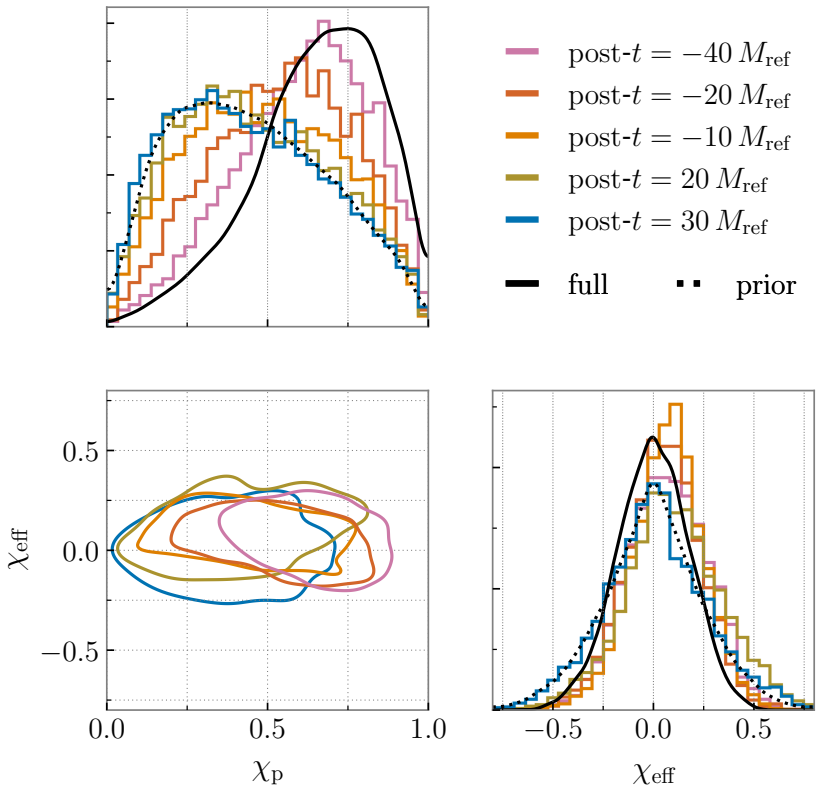


Figure 7.6: Posteriors for $\chi_{\rm p}$ and $\chi_{\rm eff}$ for GW190521 informed by data after five representative cutoff times: $t = \{-40, -20, -10, 20, 30\}\,M_{\rm ref}$ in pink, red, orange, green, blue respectively, compared to the full signal analysis (solid) and the prior (dotted) in black. Contours of the 50% credible regions are shown for the two-dimensional posteriors. The two spin parameters are not correlated, meaning that inferences on $\chi_{\rm p}$ at the various cutoff times do not simply arise from conditional priors from $\chi_{\rm eff}$.

arise from a conditional measurement driven by $\chi_{\rm eff}$ [50]. In other words, we are not just identifying differences in the spins from $\chi_{\rm eff}$ that then propagate into differences in $\chi_{\rm p}$, but rather we are directly identifying differences in the in-plane spin components' inference.

However, $\chi_{\rm p}$ is a parameter motivated by inspiral precession dynamics [77, 200, 203]. The angles θ_i associated with each black hole's spin vector – used to calculate $\chi_{\rm p}$ – are defined with respect to the binary's orbital angular momentum, a quantity that becomes meaningless during the merger as the two black holes cease orbiting each other. Indeed, the canonical equations describing precession dynamics (Eq. (11) of Ref. [199]) are only accurate through the second post-Newtonian order, an expansion which is only valid when the black holes are moving substantially slower than the speed of light, i.e. during the

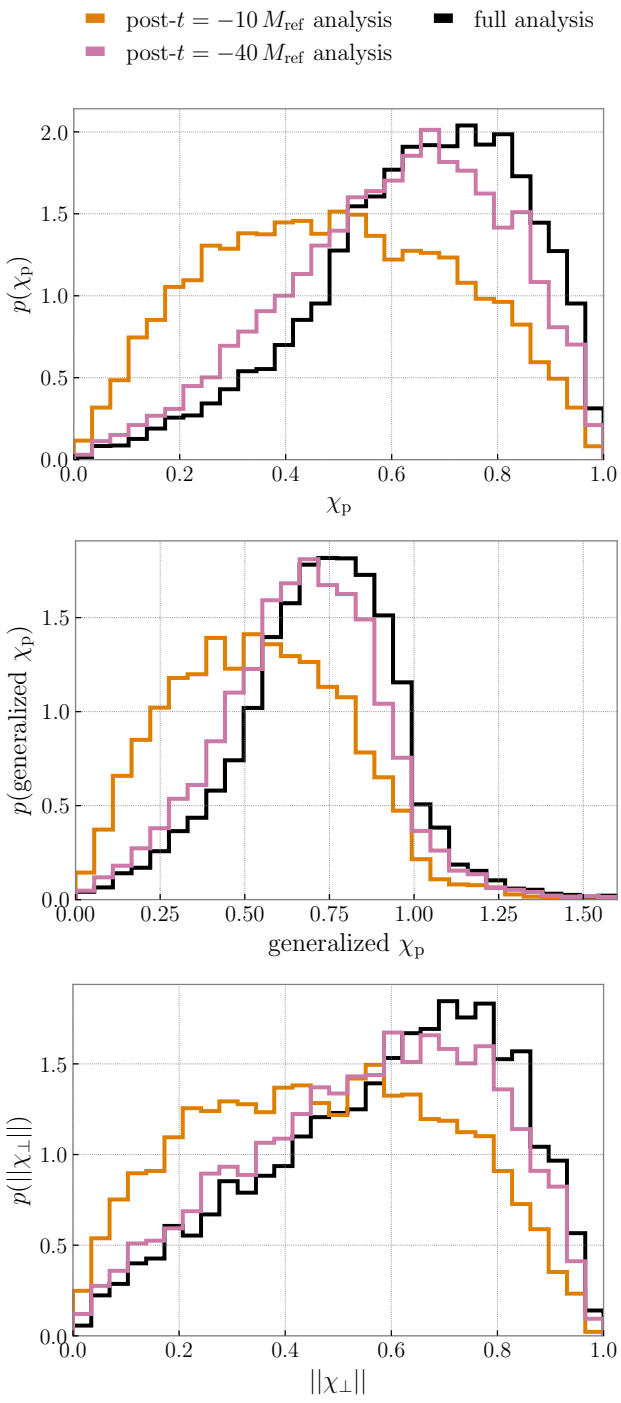


Figure 7.7: From top to bottom: posteriors for $\chi_{\rm p}$, generalized $\chi_{\rm p}$, and $||\vec{\chi}_\perp||$ for the full analysis (black), and the post-$t = -40\,M_{\rm ref}$ (pink) and post-$t = -10\,M_{\rm ref}$ (orange) analyses. All parametrizations of precession lead to qualitatively similar conclusions about spin inference and GW190521.

inspiral [14, 148, 188, 199, 312]. Once entering the merger, precession becomes no longer analytically tractable.

Alternative parameterizations of precession have been proposed, though they are all still inspired by intuition gained from the inspiral precession equations. Ref. [205] note that in the definition of $\chi_{\rm p}$, some – but not all – of the precession-timescale oscillations are averaged over. To rectify this and retain all variation over the precessional timescale [78], they define a "generalized $\chi_{\rm p}$" in their Eq. (15). Defined on the range $[0, 2)$, the generalized $\chi_{\rm p}$ has the advantage that it can help distinguish between binaries in which one versus both spins are precessing. While binaries with generalized $\chi_{\rm p} < 1$ can have one or both spins precessing, the case in which generalized $\chi_{\rm p} > 1$ only arises in the both-precessing scenario. Ref. [204] amend the fact that $\chi_{\rm p}$ does not accurately account for higher order multi-polar modes by defining the "effective precession spin vector" $\vec{\chi}_\perp$ – their Eq. (9). This parameter accounts for more degrees of freedom and could facilitate a better representation of precession in the strong-field regime. The magnitude of the effective precession spin vector $||\vec{\chi}_\perp||$ is a scalar measure of precession, analogous to the traditional $\chi_{\rm p}$. A third way to quantify the observability of spin-precession is through the precessional SNR $\rho_{\rm p}$, as defined in Fairhurst *et al.* [201, 408]. Based on the idea that precession is only inferred when two gravitational-wave harmonics are observed, $\rho_{\rm p}$ is related to the SNR in the second most significant waveform harmonic, defined in Eq. (39) of Ref. [201].

Posteriors for generalized $\chi_{\rm p}$ and $||\vec{\chi}_\perp||$ are plotted in Fig. 7.7 for the post-$t = -40\,M_{\rm ref}$ (pink) and $-10\,M_{\rm ref}$ (orange) analyses, compared to the full analysis posteriors (black). We do not plot $\rho_{\rm p}$, as the current code to compute it [409] necessitates a frequency domain waveform, which `NRSur7dq4` is not [16]; Ref. [410] presented $\rho_{\rm p}$ constraints for GW190521 using a different set of waveform models. Both gen. $\chi_{\rm p}$ and $||\vec{\chi}_\perp||$ follow the same trend as the traditional $\chi_{\rm p}$, shown in the top panel of Fig. 7.7 for comparison. The inference of precession is closer to that from the full signal in the post-$t = -40\,M_{\rm ref}$ analysis than the $t = -10\,M_{\rm ref}$.

The above quantities are all defined based on the inspiral dynamics; precession representations motivated by merger dynamics are not available to the best of our knowledge. Furthermore, all of the above quantities, as well as $\chi_{\rm p}$ itself, are quoted at the reference frequency of 11 Hz. Since quantifications of

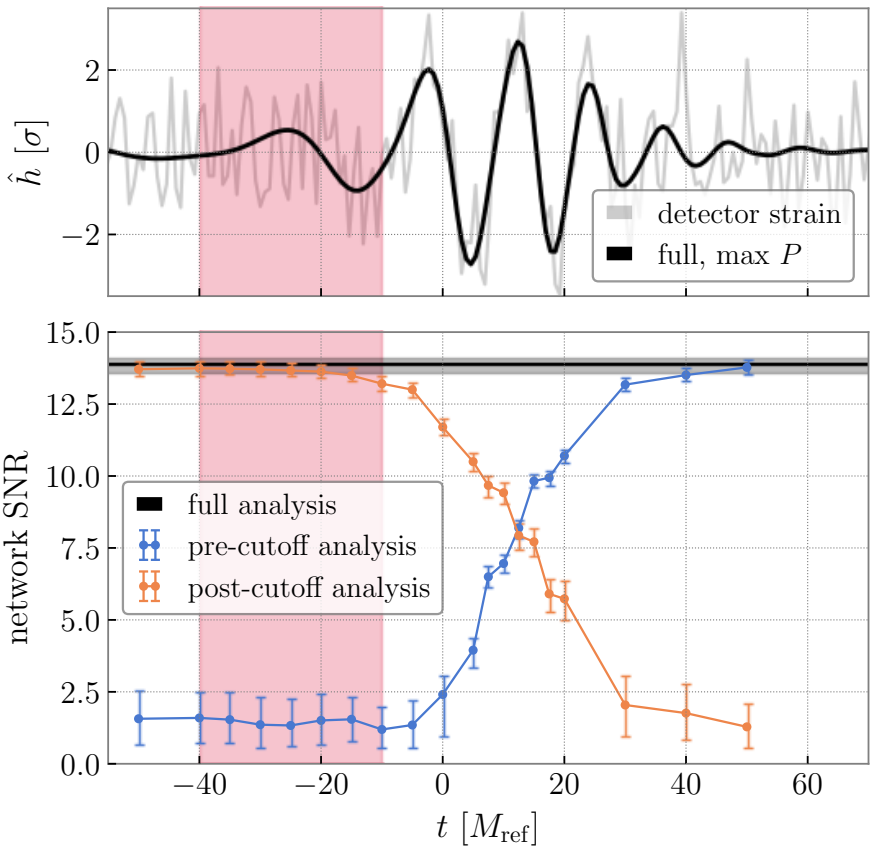


Figure 7.8: *(Lower subplot)* Network matched-filter SNR for the pre-cutoff (blue) and post-cutoff (orange) analyses as a function of the cutoff time t. Points indicate the median SNR and error bars represent the 90% credible region. The horizontal black line represents the median SNR recovered by the full analysis, with the 90% credible region shaded in gray. *(Upper subplot)* Whitened strain data in LIGO Livingston (gray), and whitened waveform reconstruction for the maximum posterior (abbr. "max P") draw from the full analysis (black), plotted to help visualize the signal alongside the SNR evolution. In both subplots, the red shaded region – between $t = -40\,M_{\rm ref}$ and $-10\,M_{\rm ref}$ – is that which is identified as crucial for constraining $\chi_{\rm p}$.

precession are time varying–albeit slowly [77]–we expect their measurement to change as a function of reference frequency. However, crucially we do not expect the choice of reference frequency to affect our conclusions. As long as the same reference frequency is used for all pre-/post-cutoff analyses, the posteriors for $\chi_{\rm p}$ (and related quantities) are expected to vary with the cutoff time in a similar way to the results presented in Fig. 7.7.

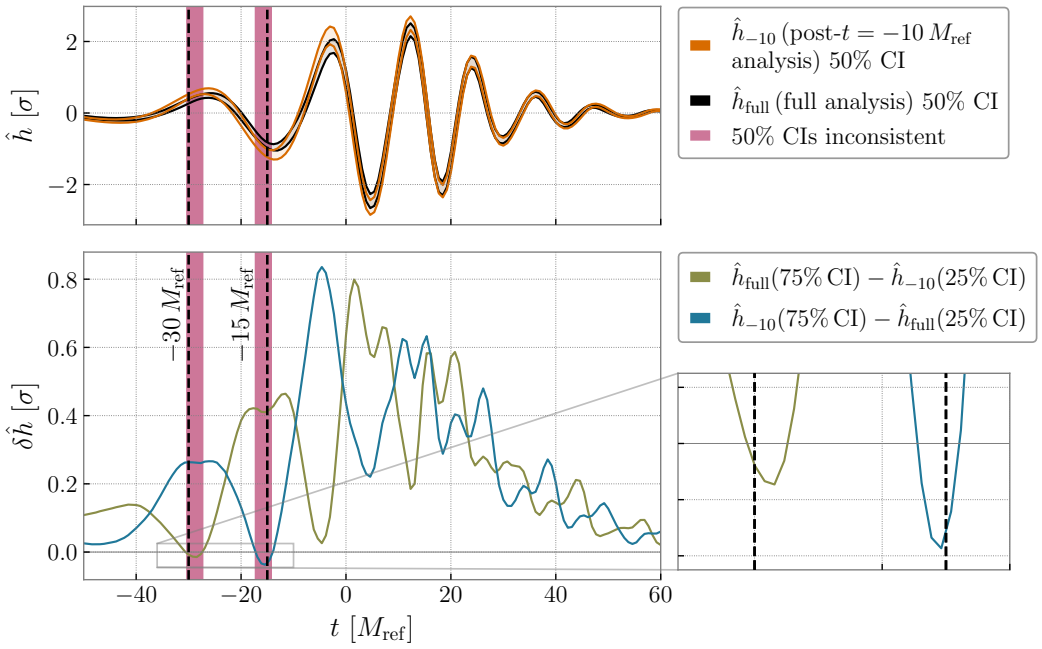


Figure 7.9: Identification of regions in time (pink shaded) during which the 50% credible intervals (CI) of the reconstructed waveform strain $\hat{h}$ are inconsistent between the full (black; labeled $\hat{h}_{\rm full}$) and post-$t = -10\,M_{\rm ref}$ (orange; labeled $\hat{h}_{-10}$) analyses. This occurs around the two extrema of the final pre-merger cycle ($t \sim -30\,M_{\rm ref}$ and $t \sim -15\,M_{\rm ref}$; indicated with vertical black dashed lines), when $\delta\hat{h}$—the difference between the 75% CI of one analysis and the 25% CI of the other (green and blue)—is negative (enlarged in the inset).

7.D Signal strength over time

We track the time-bounded matched-filter signal-to-noise ratio (SNR) in the LIGO Livingston, LIGO Hanford, and Virgo detector network recovered by our analyses, as defined in Eqs. (52) and (53) of Ref. [367]. Figure 7.8 shows the SNR evolution for the pre- (blue) and post-cutoff (orange) analyses. The change in SNR is negligible between $t = -40\,M_{\rm ref}$ and $-10\,M_{\rm ref}$ (red shaded region) for both the pre- and post-cutoff analyses, suggesting that the change in inferred $\chi_{\rm p}$ between these times is not simply due to a drop in SNR.

To investigate correlations between signal strength and $\chi_{\rm p}$ (see Fig. 7.2 in the main text), we compare the 50% credible intervals (CI) of the reconstructed waveform strain for two analyses, one in which precession is inferred (full analysis) and one in which it is not (post-$t = -10\,M_{\rm ref}$ analysis) . The top panel in Fig. 7.9 shows the 50% CIs for the strain $\hat{h}$ for these two analyses in black and orange respectively. In the bottom panel, we plot the differences $\delta\hat{h}$ between the 75% CI of one analysis and the 25% CI of the other in green and blue.

When either difference is *negative*, the two sets of waveform reconstructions are inconsistent at the 50% CI. This inconsistency *only* occurs at the extrema of the final pre-merger cycle, i.e. the peak around $t \sim -30\,M_{\rm ref}$ and trough at $t \sim -15\,M_{\rm ref}$, as indicated by the time-slices shaded in pink. These regions are enlarged in the inset.

Chapter 8

INSIGHTS FROM TIME-DOMAIN SIGNAL MORPHOLOGY ON MEASURING SPIN PRECESSION FROM MASSIVE BLACK HOLE BINARIES

This chapter is published as Ref. [102]:

Simona J. Miller, Maximiliano Isi, Katerina Chatziioannou, Vijay Varma, and Sophie Hourihane. Measuring spin precession from massive black hole binaries with gravitational waves: insights from time-domain signal morphology. *Phys. Rev. D*, 112(10):104046, 10 2025. ***SJM*** *conceptualized the project jointly with co-authors, conducted all analyses, wrote all text, and made all figures.*

Abstract

Robustly measuring binary black hole spins via gravitational waves is key to understanding these systems' astrophysical origins, but remains challenging—especially for high-mass systems, whose signals are short and dominated by the merger. Nonetheless, events like GW190521 show that strong spin precession can indeed be gleaned from high-mass systems. In this work, we track how spin precession imprints on simulated high-mass binary black hole signals cycle-by-cycle using time-domain inference. We investigate a suite of signals, all with the same spins and (near-unity) mass ratio—yielding identical spin evolution—but different signal-to-noise ratios, total masses, and extrinsic angles, all of which affect the observed waveform morphology. We truncate each signal at various times and infer source parameters using only the data before or after each cutoff. The resultant posterior allows us to identify which time segments of each signal inform its spin precession constraints. We find that at a sufficiently high post-peak signal-to-noise ratio (SNR, $\rho \sim 20$), spin precession can be constrained by the NRSur7dq4 waveform model when just the post-peak data (i.e., ringdown) are visible. Similarly, at a large enough pre-cutoff SNR ($\rho \sim 10$), spin precession can be constrained using only pre-peak data (i.e., inspiral); this occurs for signals with detector-frame total mass $\lesssim 100M_\odot$ at GW190521's full-signal SNR. Finally, we vary the inclination, polarization, and phase angles, finding that their configuration need not be fine-tuned to

measure spin precession, even for very high-mass and short signals with two to three observable cycles. We do not find that the same morphological features consistently drive precession constraints: in some signals, precession inference hinges on the relationship between a loud merger and quiet pre-merger cycle, as was the case for GW190521, but this is not generically true. Our studies enable a morphological understanding of precession that can help bolster confidence in precession inference in the presence of signal or noise systematics.

8.1 Introduction

Black hole (BH) spin is a key astrophysical observable, promised to illuminate physics on multiple scales, from understanding of stellar interiors [51] to the formation and evolution of binary black holes (BBHs) throughout cosmic history [1]. Measurable precession of the BBH orbital plane caused by misalignment between the spin and orbital angular momenta is particularly useful for distinguishing between different BBH evolutionary pathways [1, 2, 43, 45, 61, 62, 64, 67, 84–87, 89, 93, 93, 282, 328, 329, 351, 368, 411–413]. Gravitational waves (GWs) from BH mergers observed by Advanced LIGO [4], Advanced Virgo [5], and KAGRA [6] are one of the few ways to measure the spins of stellar-mass BHs directly.

Spin precession is, however, notoriously difficult to measure from GW signals [98, 143, 314, 408]. Informative spin measurements are rare to begin with: most observations are consistent with negligible spin magnitudes, and contain no information about spin precession [285–287]. Moreover, among observed BBHs with interesting spin configurations, a few are contaminated by non-Gaussian noise transients ("glitches" [144–146]). The anti-aligned spins of GW191109_010717 [158] and the large spin precession of GW200129_065458 [159], for example, are both dependent on the assumed glitch model. Measuring precessing spins is further complicated by waveform systematics: different models for the waveform can yield different in-plane spin measurements for the same signals [48]. Therefore, obtaining robust spin measurements can be even more challenging than other BBH properties like mass, not only because the imprint is weaker but also because it is more subtle to model.

We submit that a detailed understanding of *signal morphology* (i.e., strain over time) can lead to more robust spin inference and interpretation. Connecting astrophysical parameters with waveform phenomenology—or how these

parameters imprint cycle-by-cycle onto observable GW signals—bolsters the measurements' robustness. For instance, if spin precession consistently imprints a distinct waveform feature, identifying that feature in a new GW event would strengthen confidence in the measurement and clarify potential systematics. Additionally, consider the highly probable situation in which we detect a GW that overlaps with a glitch or its residual [146]: with *a priori* knowledge of how and where precession imprints on GWs, we can better identify whether or not that glitch can mimic precession. This knowledge is especially powerful in the case where the imprint of precession is subtle and can be easily mimicked by non-Gaussianities in data. In this way, signal morphology is a tool to help safeguard against spurious spin measurements caused by data quality issues or waveform systematics. A similar idea of dividing data into time or frequency chunks to compare how data and simulations accumulate signal-to-noise ratio (SNR) is used in GW searches [414–416] and glitch mitigation [158]. Working in the time domain allows us to draw a connection between the waveform morphology and the dynamics of the source, leading to better intuitions.

The link between waveform morphology and system parameters is especially challenging for high-mass ("heavy") BBHs with redshifted (i.e., detector-frame) component masses $\gtrsim 50 + 50\,M_\odot$. These systems merge at the low-frequency edge of the sensitivity band for current ground-based interferometers [141], resulting in a short-duration detectable GW signal dominated by the merger phase of coalescence. In such cases, up to seventeen BBH parameters[1] must be inferred from just a few observed waveform cycles. If a cycle overlaps with a glitch, a large portion of the signal is compromised. Moreover, the observational signature of BBH parameters on the merger phase remains poorly understood. Unlike in the inspiral which is described by, e.g., the post-Newtonian formalism [148, 149], spin is not analytically tractable in the merger [150]. Precession is particularly difficult to measure, as the precession timescale is comparable to the duration of the signal itself [78]. A sizable fraction of the most recent catalog of GW transients [287] is consistent with being high-mass, and therefore plagued by these issues. Although the typical BBH detected with GWs is consistent with having detector-frame masses of less than $40 + 40\,M_\odot$, generating an observable inspiral, about *one third*[2] of the BBHs in the third

[1]Fifteen parameters describe a non-eccentric BBH, and an additional two are necessary if the orbit is eccentric.

[2]Nuance is necessary when interpreting the significance of high-mass BBH merger de-

GW transient catalog [287] have a median inferred detector-frame total mass $M \gtrsim 100\,M_\odot$, which corresponds to a merger at $\sim$150 Hz and only $\sim$5 observable cycles in-band.

Investigations with numerical relativity (NR) and other simulated data (e.g., with NR surrogates and ringdown models) suggest that, despite observational challenges, information about spin precession *does* exist in heavy BBH signals [103, 104, 151–154, 156, 157]. However, studies of simulated high-mass GW signals, such as Ref. [103] and Ref. [104], find that the observability of spin precession in these systems is contingent on many factors: the specific spin configuration of the system, the angle at which the BBH source is oriented with respect to the Earth, and the Gaussian noise realization. In some cases, even changing spin tilt angles by as little as order one-tenth of a radian can affect whether or not spin precession is observed, cf. Fig. 8 of Ref. [103]. Generally, it is not always understood why precession inference is informative or not in any specific signal. The relationship between the input (parameters) and output (strain) of waveform models and the *measurability* of these parameters via said strain remains murky.

In Ref. [101] we used *time-domain inference* [367, 395] to understand the morphological imprint of spin precession in one particular signal: GW190521 [20, 107], the most massive and highly precessing BBH confidently observed to date.[3] We dissected GW190521's data in the time domain and performed parameter estimation independently on different temporal sub sets of data, i.e., before/after different "cutoff times." With this method, we localized the segment of data driving the previously elusive measurement of spin precession, under the assumption of a quasicircular orbit: *the constraint hinges on a short and quiet portion of the pre-merger data that is dampened out with respect to the signal's loud merger*. Looking into the system's source dynamics, the relative suppression of this pre-merger cycle corresponds to the tilt of the binary's direction of maximal GW emission away from the Earth as the orbital plane precesses. If the system were *not* precessing (and we assume the orbit is non-eccentric), such a change in emission direction would not occur, leaving the final pre-merger amplitude louder than it was for GW190521.

tections. This back-of-the-envelope estimate uses the standard astrophysical probability significance threshold [287] and not the more stringent thresholds used when searching for intermediate mass BHs [417].

[3] GW190521's status as most massive has since been eclipsed by GW231123_135430 [21].

Here, we expand upon Ref. [101] and use *simulated* data to further investigate the signal morphology conditions that underpin measurements of spin precession in high-mass systems. Like past studies, our motivating question is: should it be surprising that out of $\mathcal{O}(100)$ detections the most massive BBH is also the most highly precessing, i.e., does it hinge on a fine-tuned configuration? Furthermore, when precession *is* constrained, does our *interpretation* of the morphological origin of the GW190521 measurement hold generically true, i.e., do the relative strengths of the pre-merger and merger data always drive the measurement? This paper is the first in a series of two throughout which we address these questions. In this first paper, we start from the "best fit" (maximum likelihood; max. $\mathcal{L}$) parameters of GW190521 [101] and consider a variety of heavy, highly-precessing signals. In the Ch. 9, we investigate effects that could *mimic* precession in high-mass signals, namely eccentricity [290] and intermediate mass ratio [292].

By tweaking GW190521's morphology along five different axes in parameter space, we examine how features in the signal impact the confidence of the spin precession measurement for both the full and restricted subsets of data, i.e., the inspiral vs. ringdown. Throughout this work, we define the "ringdown" as the data after the signal's peak, and the "inspiral" as the data before the signal's peak, where the peak refers to the waveform amplitude integrated over the celestial sphere [418]. We keep the six spin degrees of freedom and mass ratio fixed, meaning all of our simulated signals have identical spin evolution. The effect of spin magnitudes and angles on inference in high-mass signals has been explored in other works [103, 104]. We vary the system's SNR, adjust merger frequency by changing the total mass at a fixed SNR, and explore extrinsic angular parameters (inclination, polarization,and phase). Each of these variations affects the *observed* waveform morphology and precession measurability in different ways. We find that:

- Spin precession is measurable later into and past the merger regime of the signal as the SNR increases. At a sufficiently high SNR, precession can be constrained using ringdown data alone, with informative measurements persisting for start times up to $20\,M$ after the signal's peak.

- Decreasing the total mass, and thereby increasing the duration of the detectable inspiral, makes the signal more informative about precession. At

sufficiently low masses ($M \lesssim 100\,M_\odot$), precession can be constrained using the inspiral data alone at GW190521's SNR.

- A fine-tuned extrinsic angular configuration is not necessary to measure spin precession, meaning that precession is constrained away from the prior for many different combinations of the inclination, polarization, and phase angles. The inclination and phase angles do, however, drastically affect measurability, ranging from signals which have no information to being more informative than the real GW190521 data. The polarization angle does not affect precession inference for the configurations we study. These trends hold true for the full and restricted durations of data.

- The relationship between loud merger and quiet final pre-merger does not always drive spin precession constraints, as it did for GW190521 [101]. In some signals, inference does hinge on the relative amplitude of the pre-merger cycle, but generically, the informativeness and the ratio of the signal's pre-merger strain peak to overall maximal amplitude are not negatively correlated.

The remainder of the paper is organized as follows. We briefly describe our methodology in Sec. 8.2. In Sec. 8.3 we present inference on the max. $\mathcal{L}$ waveform for GW190521 from Ref. [101]. Section 8.4 presents results from changing different parameters of the max. $\mathcal{L}$ waveform, with Secs. 8.4, 8.4, and 8.4 corresponding to the first three bullets above. Finally, we conclude in Sec. 8.5.

8.2 Methods

In this section, we establish spin notation (Sec. 8.2), and then describe our analysis approach, which can be broadly outlined in three steps. First, we inject GW190521-like signals and conduct parameter estimation in the time domain before and after various cutoff times in the data. The methods for generating and performing inference on simulated signals in the time domain are described in Sec. 8.2. Then, we assess the informativeness of posteriors across different time segments by computing the Jensen-Shannon divergence between each posterior and its prior, as described in Sec. 8.2. Finally, we trace the informativeness of the posteriors as a function of cutoff time to localize what features in the waveform may be driving the measurements of spin precession and other astrophysical phenomena. Throughout, we look at

the whitened time-domain strain and SNR, obtained through the procedures described in Sec. 8.2.

Spin parametrization

At leading order, BH spin imprints on GWs from merging BBHs in two ways. The strongest effect—modulation of the signal duration—comes from the spin components *parallel* to the binary's orbital angular momentum ("aligned spin"). These are parametrized via the effective spin $\chi_{\rm eff}$ which is approximately conserved in the inspiral [193, 289],

$$\chi_{\rm eff} = \frac{\chi_1 \cos\theta_1 + q\,\chi_2 \cos\theta_2}{1+q} \in [-1, 1]\,, \tag{8.1}$$

where χ_i are spin magnitudes, $\cos\theta_i$ are the polar angles between the spins and the orbital angular momentum, and $q \equiv m_2/m_1$ is the binary's mass ratio. Subscript 1 corresponds to the primary (more massive) BH, while 2 corresponds to the secondary (less massive).

Spin components *misaligned* with the orbital angular momentum cause the binary's orbital plane to precess, which induces amplitude and phase modulation as the direction of maximum GW emission sweeps toward and away from the Earth [199, 312]. Although alternatives exist [204, 205], we report spin precession constraints using the canonical "effective precessing spin" parameter $\chi_{\rm p}$ [77, 200, 203],

$$\chi_{\rm p} = \max\left[\chi_1 \sin\theta_1, \left(\frac{3+4q}{4+3q}\right) q\,\chi_2 \sin\theta_2\right] \in [0, 1]\ , \tag{8.2}$$

for ease of comparison to past work. In Ref. [101], we showed that other precession parameters behave analogously to $\chi_{\rm p}$ for GW190521. Spin inference is correlated with other parameters [98, 103, 104, 187, 298, 300, 301]. Thus, while we focus on $\chi_{\rm p}$, we also examine $\chi_{\rm eff}$, detector-frame total mass M, and mass ratio q.

Given that the spin vectors evolve over time due to precessional dynamics [199], they must be quoted at a specific reference time or frequency. We define all spin quantities at $t = -100\,M$ [16, 48]. For reference, in geometric units, $1\,M_\odot = 4.93 \times 10^{-6}\,$s and hence the max. $\mathcal{L}$ point for GW190521 has $M = 258.7\,M_\odot = 1.27 \times 10^{-3}\,$s. Using units of mass allows us to compare systems with different total masses by referring to the analogous points in the orbital evolution (see Sec. 8.4) and facilitates comparison with ringdown analyses [214], which report results in units of detector-frame remnant mass.

Analyzing signals in the time domain

Inferring the parameters of GW sources is typically conducted in the frequency domain. Because detector noise is stationary, its covariance matrix becomes diagonal in the Fourier domain when periodic boundary conditions are imposed, enabling efficient likelihood calculations [284, 391]. However, if we want to isolate a specific segment of the signal in time, it may not be possible to enforce periodicity, meaning either non-trivial likelihood modifications must be performed in the Fourier domain [e.g., 215, 419], or the data must be analyzed directly in the time domain [367]. We use time-domain inference [101, 367], based on methods originally derived for studying BH ringdowns [222, 395, 396]. The code base used for this analysis has been released as `tdinf` [161].

We use simulated GW signals with a zero noise realization, assuming the sensitivity of the LIGO and Virgo detectors at the time of GW190521 [407]. We opt for zero noise because we are interested in studying the behavior of simulated signals on average, and zero noise is equivalent to averaging over Gaussian realizations [420]. The specific Gaussian noise realization will impact the posteriors [103, 104] in the standard way, i.e., by shifting the best-fit parameters [421]. We confirm that we can reproduce the GW190521 inference with the max. $\mathcal{L}$ waveforms in zero noise (see Fig. 8.1), unlike Ref. [104].

We use the inspiral-merger-ringdown (IMR) numerical relativity surrogate waveform NRSur7dq4 [16] for both injection and recovery.[4] For each signal, we select a series of cutoff times, truncate the data at each time, and analyze the data before ("pre-cutoff") and after ("post-cutoff") independently, as well as the full signal. Consistent with NRSur7dq4, we define the time of coalescence ($t = 0$) as the time of peak total waveform amplitude integrated over the celestial sphere [418]; see Eq. (5) of Ref. [16].

During inference, we draw samples from the posterior of the binary total mass M, mass ratio q, spin magnitude χ_i and tilt angles θ_i, azimuthal inter-spin angle ϕ_{12}, azimuthal precession cone angle ϕ_{JL}, luminosity distance d_L, and phase of coalescence φ. Geocenter time of coalescence t_0, right ascension α, declination δ, and polarization angle ψ are fixed to their true values; this fixes the time of arrival in each detector. Ref. [101] showed that fixing these parameters did not affect conclusions for GW190521. We confirm that the

[4] To generate NRSur7dq4 with parameters defined at $t = -100\,M$, see Sec. 8.2, we use a custom version of LALSuite with the flag `fref=-1` [16, 48, 319].

same is true for simulated data by repeating select analyses while varying over all parameters. Priors are the same as Table I in Ref. [101], except that we use a wider uniform total mass prior to encompass each simulated system's true total mass value. We simulate data from the three-detector network of LIGO Livingston (LLO), LIGO Hanford (LHO), and Virgo with a sampling rate of 2048 Hz, assuming the GW190521 event power spectral densities publicly available on the Gravitational-wave Open Science Center [405, 406]. We analyze the frequency range of 11–1024 Hz. Our time domain inference code [161] utilizes a Markov Chain Monte Carlo sampler implemented in Python with `emcee` [252]. Each analysis runs with 512 chains for 50,000 steps.

At face value, there are parallels between our time-domain inference framework and traditional IMR consistency tests (IMRCTs) [e.g., 394, 422, 423] used to test General Relativity [424]. However, while both involve conducting parameter estimation on different portions of a GW signal, the two have distinct methods and goals. IMRCTs operate in the frequency domain—frequency-domain truncation is not equivalent to time-domain truncation beyond the stationary phase approximation, e.g., in the merger or when higher-order modes are present. One could theoretically slice GW data in any way in the time-frequency plane and require consistency within the different segments, but not all bases provide equal insight. Segmenting in the time domain allows us to draw connections to a BBH's source dynamics in a more direct way than other bases. Moreover, IMRCTs are, as the name suggests, consistency tests: they compare the posterior from different parts of the signal to see if the data are consistent with each other. In this work, we are not concerned with the question of consistency, but rather in understanding how information is accumulated throughout a signal.

Quantifying a parameter's "measurability"

We use the Jensen-Shannon divergence (JSD) [425–427] to quantify how different two posterior probability distributions are. The JSD is a statistical quantity for the similarity between two probability distributions, P and Q. It is the symmetrized and smoothed version of the Kullback-Lieber (KL) divergence D [343, 428], such that

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\left[D\left(P\middle|\middle|\frac{P+Q}{2}\right) + D\left(Q\middle|\middle|\frac{P+Q}{2}\right)\right], \tag{8.3}$$

where the KL divergence is defined as follows:

$$D(P||Q) = \int_{-\infty}^{\infty} p(x) \log_2 \frac{p(x)}{q(x)} dx \,, \tag{8.4}$$

where p and q are the probability densities of P and Q. The JSD is normalized with a minimum of 0 nats, corresponding to the two distributions P and Q being identical, and a maximum dependent on the base of the logarithm used to calculate $D(P||Q)$. We use a base 2, corresponding to a maximum of $\log(2) = 0.69$ nats. Lower JSDs indicate distributions that are more similar.

To define how "measurable" a parameter is, we follow Refs. [104, 254] and calculate the JSD between its one-dimensional marginal posterior and prior. Correlated priors between two parameters can lead to spuriously strong one-dimensional JSDs, i.e, confidently measuring one may induce a prior-driven measurement in the other [50]; we further discuss this in Appendix 8.A. The KL divergence is estimated using a kernel density estimate (KDE) over a finite number of samples; it is thus subject to Monte Carlo error. As an order of magnitude approximation, when generating sets of $N = 10,000$ random draws from the $\chi_{\rm p}$ prior distribution, we find JSDs ranging $\mathcal{O}(10^{-4} - 10^{-6})$. When using a smaller number of samples, $N = 1,000$, the JSDs increase up to $\mathcal{O}(10^{-3})$. Figure 8.9 in the Appendix shows examples of such JSD null distributions for $N = 3,000$ samples. As such, to be conservative when dealing with varying numbers of samples and differently shaped distributions (KDEs are sensitive to the tails of distributions), we define JSDs $\gtrsim 10^{-2}$ as certainly indicating a difference between two distributions, and JSDs $\lesssim 10^{-4}$ as indicating identical distributions.

Whitening a signal and calculating its signal-to-noise ratio

We visualize GW signals with *whitened* waveforms, a helpful representation of GW data which brings the noise to a white-Gaussian baseline so we can better judge what goes into the likelihood by eye. In the Fourier domain, whitened waveforms $\hat{h}(f)$ are obtained by dividing the original waveform $h(f)$ by the square root of the noise power spectral density, $S_n(f)$:

$$\hat{h}(f) = \sqrt{\frac{f_s}{2\, S_n(f)}}\, h(f) \,, \tag{8.5}$$

where f_s is the sampling rate. We derive the corresponding whitened time series $\hat{h}(t)$ by inverse Fourier transforming. The scaling of Eq. (8.5) ensures

Parameter	Symbol	max. $\mathcal{L}$ value ($f_{\rm ref} = 11\,{\rm Hz}$)	max. $\mathcal{L}$ value ($t_{\rm ref} = -100\,M$)
Detector-frame total mass	M	$258.74\,M_\odot$	$277.15\,M_\odot$
Mass ratio	q	0.97	0.86
Primary spin	$\vec{\chi}_1$	(0.67, 0.04, 0.29)	(−0.87, 0.18, −0.39)
Secondary spin	$\vec{\chi}_2$	(−0.05, −0.76, −0.58)	(−0.17, −0.75, 0.59)
Effective spin	$\chi_{\rm eff}$	−0.13	0.06
Effective precessing spin	$\chi_{\rm p}$	0.74	0.89
Inclination angle	ι	1.16 rad	1.11 rad
Polarization angle	ψ	2.30 rad	2.30 rad
Phase	φ	1.31 rad	−3.12 rad
Luminosity distance	d_L	3253.11 MPc	3190.22 MPc
Right ascension	α	6.26 rad	6.26 rad
Declination	δ	−1.15 rad	−1.15 rad
Coalescence GPS time	t_0	1242442967.407715 s	1242442967.4064941 s

Table 8.1: The maximum likelihood (max. $\mathcal{L}$) parameters from two time-domain inference analyses conducted on GW190521 by Ref. [101]: one where all parameters are defined at $f_{\rm ref} = 11\,{\rm Hz}$ and another at $t_{\rm ref} = -100\,M$. The $f_{\rm ref} = 11\,{\rm Hz}$ parameters are used as the baseline for the results presented in Sec. 8.3, Sec. 8.4, and Sec. 8.4. The $t_{\rm ref} = -100\,M$ parameters serve as the baseline for results in Sec. 8.4 for consistency when looking at sources from different total masses. Since they are obtained from a sampling algorithm and not a dedicated "peak finding" algorithm both sets of parameters should be interpreted as "the maximum likelihood sample," rather than the absolute maximum of the likelihood surface.

that the units of $\hat{h}(t)$ are standard deviations of the noise, a dimensionless quantity denoted by σ.

When dissecting signals in the time domain, we calculate the *optimal network SNR* in different time windows, denoted ρ. For a signal x, the SNR in a single detector $\mathbb{D}$ is calculated in the time domain over a region of time T by taking the inner product of the signal with itself:

$$\rho_{\mathbb{D}} = \langle x|x\rangle_T^{1/2}\,, \tag{8.6}$$

where the time-domain inner product is defined as

$$\langle x|y\rangle_T = \sum_{i,j=0}^{N} x_i C_{ij}^{-1} y_j\,, \tag{8.7}$$

where x and y are time series with N time steps over the duration T, with index 0 corresponding to the first point in the region of T. The covariance matrix

C_{ij} is derived from a given power spectral density following the procedure presented in Sec. III of Ref. [367]. The SNR for a network of detectors is then obtained by summing the single-detector SNRs in quadrature,

$$\rho = \left(\sum_{\mathbb{D}} {\rho_{\mathbb{D}}}^2 \right)^{1/2} .$$

Here, $\mathbb{D} \in \{\text{LHO}, \text{LLO}, \text{Virgo}\}$. Using this method we calculate the SNR over different time slices of the data, e.g., the SNR in a signal's inspiral versus ringdown.

8.3 Analyzing the GW190521 maximum likelihood waveform

The signals we consider are all based on the maximum likelihood waveform for GW190521 found in Ref. [101], with parameters listed in Table 8.1. We herein call this the "max. $\mathcal{L}$ waveform." Before examining the impact of various system parameters by varying them one at a time in Sec. 8.4, we here confirm that the max. $\mathcal{L}$ waveform behaves similarly to the real event, both when analyzed in full and under time cutoffs.

Results from inference on the max. $\mathcal{L}$ waveform compared to the real data are shown in Fig. 8.1. We begin with the *full-signal analysis* (top row), comprising the entire data segment. In the first column, we plot the whitened GW190521 data (gray) superimposed with the max. $\mathcal{L}$ waveform (black) in LLO.[5] The remainder of the first row shows the posteriors from the real GW190521 data (gray filled histograms) and the max. $\mathcal{L}$ waveform (black empty histograms) for the total detector-frame mass M, mass ratio q, and effective precessing spin $\chi_{\rm p}$.[6] The true max. $\mathcal{L}$ values are indicated by vertical red lines, with which all full posteriors are consistent. Posteriors for M and $\chi_{\rm p}$ are in high agreement with those from the real data, yielding JSDs of $\mathcal{O}(10^{-3})$ between them.[7] The q posterior for the max. $\mathcal{L}$ waveform is slightly more informative than that from

[5] We plot waveforms in LLO because it is the detector in which GW190521 is the loudest: the max. $\mathcal{L}$ optimal SNR is 10.8 in LLO, 8.6 in LHO, and 2.5 in Virgo.

[6] We exclude the effective spin $\chi_{\rm eff}$ from this figure because it is consistent with the prior for the full data and all time slices. We present the evolution of $\chi_{\rm eff}$ and other spin parameters in any accompanying data release. All spin parameters lose/gain information at the same cutoff times.

[7] The latter is notable: $\chi_{\rm p}$ inference from the max. $\mathcal{L}$ waveform is more consistent with the real data than previous studies, e.g. Fig. 3 of Ref. [103] or Fig. 2 of Ref. [156]. We have verified that this difference comes down to the choice of "max. $\mathcal{L}$" waveform: while we use the parameters from Ref. [101], Ref. [103] and Ref. [156] use parameters from Ref. [404], given in Table II of Ref. [103]. When we simulate data based on that maximum likelihood waveform and perform inference in the time domain, we recover posteriors consistent with

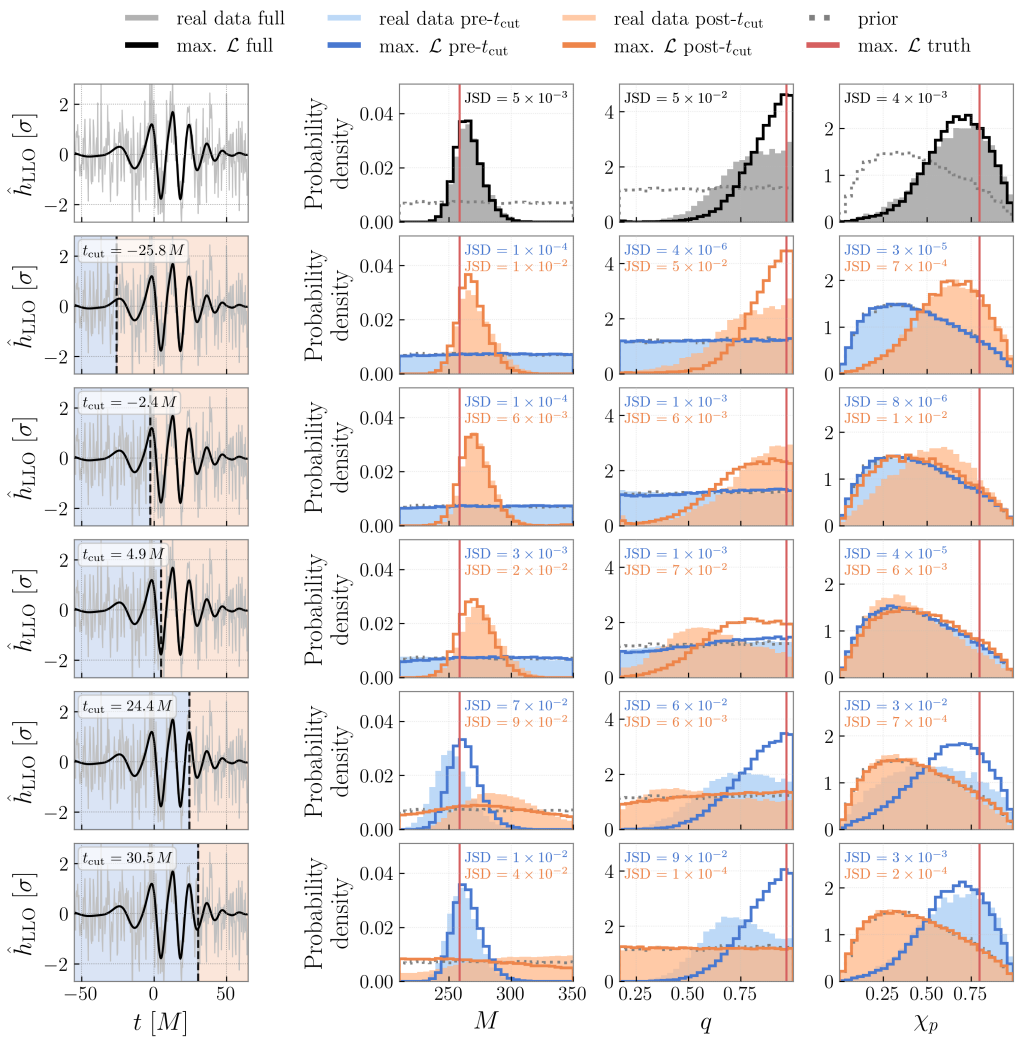


Figure 8.1: Evolution of posteriors for representative cutoff times for GW190521 and its max. $\mathcal{L}$ waveform, exemplifying that the max. $\mathcal{L}$ waveform behaves similarly to the real data, both for the full signals and as a function of cutoff time. Each row shows different segments of data analyzed: the top row shows the full segment of real data (gray) and the full max. $\mathcal{L}$ waveform (black), while subsequent rows show results from data before (blue) and after (orange) different cutoff times: $t_{\rm cut} = \{-25.6, -2.4, 4.9, 24.4, 30.5\}\,M$ from top to bottom. The lighter shaded histograms correspond to the real GW190521 data, while the darker empty histograms are for the max. $\mathcal{L}$ waveform. JSDs between the two posteriors are given in inset text on each plot. *First column:* The whitened max. $\mathcal{L}$ waveform from the full analysis (black) along with the whitened LLO data (gray). The blue/orange shaded regions highlight the data informing the same-color posteriors in the remaining columns, and the black-dashed vertical line is the cutoff time. *Second through fourth columns:* Marginalized posteriors on the detector-frame total mass M, mass ratio q, and effective precessing spin $\chi_{\rm p}$ inferred from each segment of data. Priors on each parameter are represented by the gray-dotted histograms and vertical red lines mark the true values for the max. $\mathcal{L}$ waveform (Table 8.1).

the real data, although they both peak at equal masses ($q = 1$). Here, the JSD between the max. $\mathcal{L}$ and real data posteriors is an order of magnitude larger at 5×10^{-2}, just above our threshold for difference.

We now turn to how inference progresses over time. The second through final rows of Fig. 8.1 show the results from inference conducted on data before (blue) and after (orange) five representative cutoff times. The dashed vertical line in the first column marks each cutoff, with blue and orange shaded regions representing pre- and post-cutoff data, respectively. The filled (empty) histograms are posteriors from the analysis conducted on real data (max. $\mathcal{L}$ waveform). The two analyses are again largely in agreement: JSDs between the analogous posteriors are $\lesssim \mathcal{O}(10^{-2})$. Most relevantly, ***inference of*** $\chi_{\rm p}$ ***from the*** max. $\mathcal{L}$ ***waveform hinges on the relationship between the final pre-merger cycle and the post-peak data, just like for the real GW190521 event***. For both cases, the data loses the bulk of its precession information between approximately $-25\,M$ (second row) and $-2\,M$ (third row) corresponding to the exclusion of final pre-merger cycle.

While our method is intrinsically data-driven, as it connects the measurability of parameters to observed waveform morphology, it can be useful to think about the underlying physical mechanisms that may generate the waveform morphology that drives our measurements. In Ref. [101], we found that the suppressed final pre-merger cycle of GW190521 was driven by the orbital plane precessing away from our line of sight. The same interpretation holds for the max. $\mathcal{L}$ signal, as analogous correlations between the inclination and $\chi_{\rm p}$ posteriors in the full versus excluding-pre-merger data can be seen in our accompanying data release.

Though the behaviors of GW190521 and it's max. $\mathcal{L}$ waveform are qualitatively similar as a function of cutoff time, small differences remain. The max. $\mathcal{L}$ waveform loses information about $\chi_{\rm p}$ slightly earlier than the real data in the post-cutoff analyses, and gains it back earlier in the pre-cutoff analyses. As can be seen in orange histograms in the third row of Fig. 8.1, the max. $\mathcal{L}$ posterior for the post-$t = -2\,M$ analysis is near identical to the prior, while that from the real data is shifted to slightly higher values. In the pre-cutoff

theirs (peaking at $\chi_{\rm p} \sim 0.6$). Thus the time-domain analysis of GW190521 happened to find a max. $\mathcal{L}$ point that yields a more similar posterior to the real data. We attribute this to little more than a lucky chance in stochastic sampling.

analyses, when the data before $t = 24\,M$ are included, the max. $\mathcal{L}$ signal becomes informative again (fifth row; blue). For the real data, this does not happen until $30\,M$ (sixth row). Additionally, the M and q posteriors behave marginally differently around the peak amplitude of the data: the post-cutoff q posterior for the real data shifts to lower values earlier than the max. $\mathcal{L}$ waveform (fourth row), while the M posterior for the real data retains post-cutoff information longer (fifth row). Generally, the max. $\mathcal{L}$ waveform is more informative on q than the real data.

Perfect agreement between the max. $\mathcal{L}$ waveform and the real GW190521 signal is not expected, because they differ in noise realization and true signal parameters.[8] The two are, however, guaranteed to at least be relatively consistent within the uncertainty implied by Gaussian noise as encoded in the posterior of the original analysis [142, 168]. The same is not necessarily true when analyzing *subsets* of the data, especially under mismodeling of the signal and/or noise. This is because the waveform that is the best fit (e.g., minimizes the residual) for a *portion* of the data will in general not be the same waveform that minimizes the residual over a bigger window—a template may match well for part of the signal but not the rest. For example, as shown in Fig. 2 of Ref. [101], the max. $\mathcal{L}$ waveform found for GW190521's full data is not the same as that when just its merger-ringdown data are analyzed. That the max. $\mathcal{L}$ signal behaves similarly to the real GW190521 makes it a worthy baseline to use in the remainder of this study.

We interpret the time-cut tests performed here and in the subsequent sections as nontrivial consistency checks of the interpretation of the real GW190521 data, in a similar vein to the χ^2 tests performed in GW searches [414–416] and glitch mitigation [158]. Those consistency checks test how SNR is accumulated. A waveform is split into frequency (or time) bins of equal optimal SNR. If given data are equivalent to the sum of that waveform and Gaussian noise, the matched-filter SNR in these bins should then follow a χ^2 distribution. If an unexpected amount of matched-filter SNR is accumulated in a bin, this is attributed to, e.g, a glitch. Our cutoff-time analyses are similar in spirit. Here, we use the accumulation of *information* about a parameter (in the form

[8] In the high SNR regime where the likelihood becomes approximately Gaussian, full agreement in the posteriors from the data and its max. $\mathcal{L}$ simulation is expected [168, 187, 421]. This is not the case at GW190521's SNR.

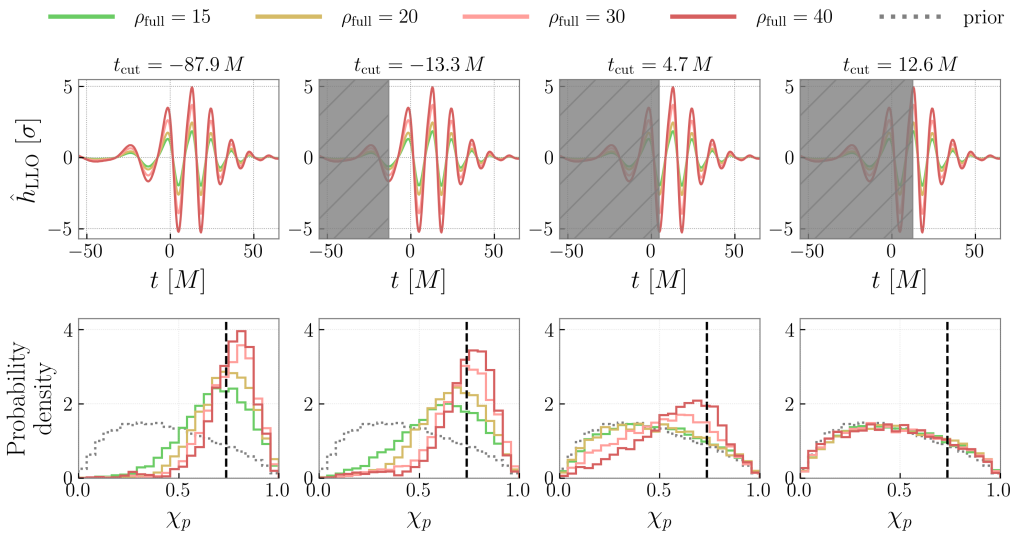


Figure 8.2: Evolution of the $\chi_{\rm p}$ posterior for the GW190521 max. $\mathcal{L}$ waveform at different full-signal optimal network SNRs: $\rho_{\rm full} = 15$ (green), 20 (yellow), 30 (pink), and 40 (red). *Top row:* Whitened strain $\hat{h}$ in LLO. The gray hatched shading hides data excluded from a given analysis. *Bottom row:* Posteriors for $\chi_{\rm p}$ inferred from the non-shaded data in the corresponding column in the top row. The prior is shown as gray dotted, while the black dashed vertical line is the true value. Louder signals have more sharply peaked $\chi_{\rm p}$ posteriors at all times. The $\rho_{\rm full} = 30$ and 40 waveforms retain some information about $\chi_{\rm p}$ in the post-peak data alone (third column).

of the JSD), rather than SNR, to see if a signal (in this case, the max. $\mathcal{L}$) is consistent with data (GW190521).

8.4 Exploring Spin Measurability

Having verified its similarity over cutoff time to the true event, we now use GW190521's max. $\mathcal{L}$ waveform as a starting point to explore the impact of various parameters on waveform morphology and parameter measurability. We simulate a suite of signals from BBHs with the same six intrinsic spin degrees of freedom and mass ratio. This means each system has the *same intrinsic spin evolution*. We also fix the time of coalescence and sky position (right ascension and declination). One parameter at a time, we then vary the SNR (Sec. 8.4), total mass (Sec. 8.4), and three extrinsic angles (phase, inclination, and polarization; Sec. 8.4), each of which changes the *observed* waveform projected onto detectors in different ways. The parameters we target roughly correspond to the three degrees of freedom of a simple sine wave: SNR controls the signal amplitude, the total mass determines the frequency and

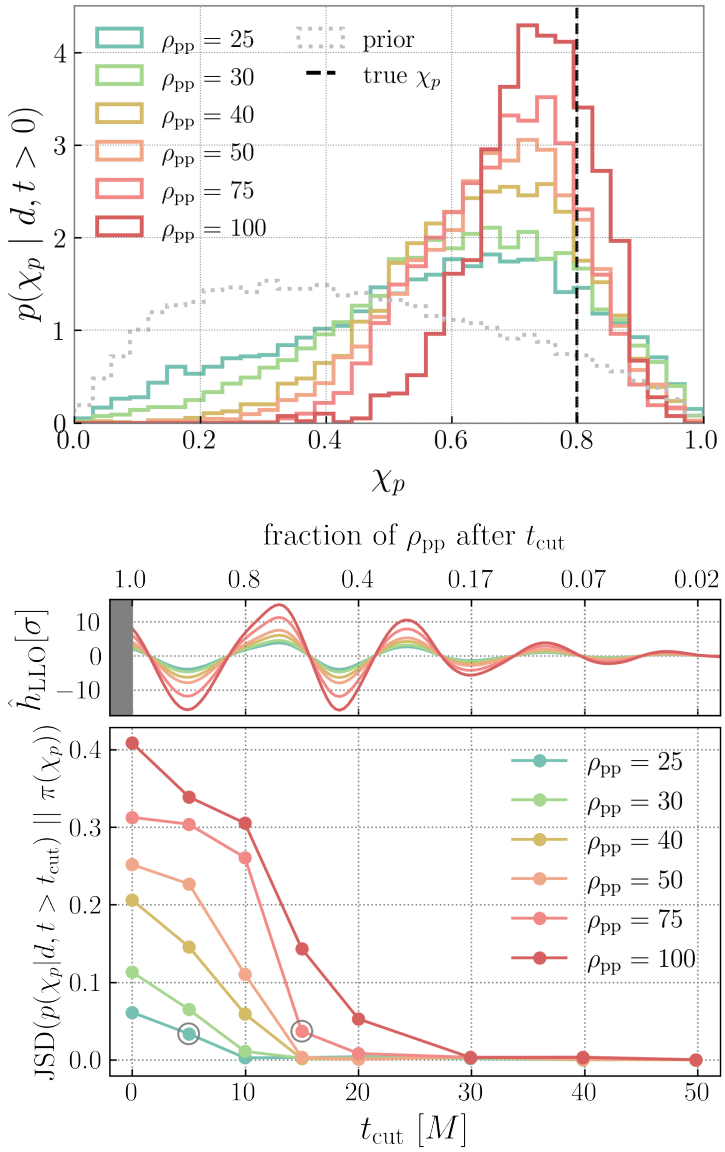


Figure 8.3: Information about $\chi_{\rm p}$ in the ringdown of the max. $\mathcal{L}$ waveform at different SNRs. These signals have $\rho_{\rm pp} = 25, 30, 40, 50, 75, 100$, which correspond to full-signal SNRs of $\rho = 30.5, 36.5, 48.7, 60.9, 91.4, 121.8$. For reference, the max. $\mathcal{L}$ waveform has $\rho_{\rm pp} = 11.4$. *Top panel:* Posteriors on $\chi_{\rm p}$ from the post-peak data ($t > 0\,M$) for different post-peak SNRs (different colors). The ringdown alone has measurable information about $\chi_{\rm p}$, as the posteriors are not consistent with the prior (gray dotted). *Bottom panel:* JSD between the $\chi_{\rm p}$ posterior and prior from data after different cutoff times (horizontal axis) for different SNRs (different colors). The whitened strain of each signal in LLO is plotted above, with the upper axis labeling the fraction of $\rho_{\rm pp}$ remaining after $t_{\rm cut}$. The two points circled in gray are an example of posteriors with comparable JSDs originating from signals with different SNRs. For loud enough signals, $\chi_{\rm p}$ can be measured up to late times of $20\,M$ into the ringdown.

observable length, and the extrinsic angles (while fixing the SNR) determine the phase. Past work [99, 101] has shown that the number of observable cycles and their strain amplitude strongly matter for the determination of intrinsic BBH properties like precession. Thus, we investigate how *the way that we observe the same BBH system* impacts our ability to infer such properties. In doing so, we hope to illuminate whether GW190521's parameters were fine-tuned to measure precession, and if the morphological origin of its precession information is universal.

Imprints of spin precession in the ringdown

In Fig. 8.1 we showed that, similar to the real data, the last inspiral cycle is necessary to constrain $\chi_{\rm p}$ in GW190521's max. $\mathcal{L}$ waveform. Now, we explore whether this is a general characteristic of the signal—meaning the ringdown alone carries no information about precession—or if the ringdown does encode precession information, but it can only be extracted at a higher SNR, as would be expected from Refs. [157, 214]. To do this, we scale the luminosity distance of the max. $\mathcal{L}$ parameters (while fixing the detector-frame total mass) to give a desired network optimal SNR, leveraging the two's inverse proportionality [183]. We then repeat the analysis of Sec. 8.3. In this section, we denote the full-signal SNR as as $\rho_{\rm full}$ and the post-peak (ringdown) SNR as $\rho_{\rm pp}$.

Results for various full-signal SNRs and time cuts are shown in Fig. 8.2. As expected, louder signals yield $\chi_{\rm p}$ posteriors that are increasingly constrained away from the prior when the full signal is analyzed—seen, by proxy, in the first column. Up to $\rho_{\rm full} = 20$, the post-cutoff results behave analogously to GW190521, which has $\rho_{\rm full} = 14$: seeing the final pre-merger cycle in comparison to the loud merger is crucial for constraining $\chi_{\rm p}$ away from the prior, as seen in the transition from the second to the third columns for $\rho_{\rm full} = 15$ (green) and 20 (yellow). For the $\rho_{\rm full} = 30$ and 40 cases, however, the post-$t = 0$ posteriors for $\chi_{\rm p}$ are *not* identical to the prior (third column), meaning that if a GW190521-like signal is loud enough, there is some evidence for precession in the post-peak signal alone.

We next investigate how far into the ringdown $\chi_{\rm p}$ can be measured for increasingly loud GW190521-like signals. We generate waveforms with increasing *post-peak* SNRs ($\rho_{\rm pp}$), i.e., ρ over the region where $t > 0$, with results shown in

Fig. 8.3. The top panel of Fig. 8.3 shows the posteriors for $\chi_{\rm p}$ when analyzing just the post-peak data (i.e., post-$t_{\rm cut} = 0$). We again use the NRSur7dq4 waveform, *not* ringdown-specific models like a sum of quasi-normal modes (QNMs), e.g., [15, 153, 155, 157, 209–215]. For all cases (various colors), $\chi_{\rm p}$ is constrained well away from the prior (gray dotted). ***Precession can be measured from the ringdown alone using an IMR waveform model, given sufficient SNR.*** This is not surprising, as the inspiral dynamics are imprinted onto QNM complex amplitudes [153, 155–157, 215, 223–225, 225], and these should be informative at sufficiently high SNR, as we discuss more in the conclusions. While our results use NRSur7dq4, other parametrized waveform models, such as IMRPhenomXO4a [246] or SEOBNRv5PHM$_{\rm w/asym}$ [244], may yield similar results.

We next cut off increasingly more data and repeat the analysis. To visualize the signal being excluded with each cut, the middle panel of Fig. 8.3 shows the whitened waveforms in LLO. The fraction of $\rho_{\rm pp}$ remaining in the signal after each $t_{\rm cut}$ is labeled on the upper axis. For example, the data after $t_{\rm cut} = 20\,M$ has 40% of the SNR of the data after $t_{\rm cut} = 0\,M$. This is independent of the total SNR.

In the bottom panel, we plot the JSD between the $\chi_{\rm p}$ posterior $p(\chi_{\rm p}|t > t_{\rm cut})$ and prior $\pi(\chi_{\rm p})$ as a function of cutoff time for each signal (different colors). Across all SNRs, as more data are excluded, i.e., $t_{\rm cut}$ increases, the JSDs decrease, indicating that the $\chi_{\rm p}$ posterior becomes increasingly consistent with the prior. For $\rho_{\rm pp} \leq 50$, the JSD between the $\chi_{\rm p}$ posterior and prior is $\gtrsim 10^{-2}$ for up to $10{-}15\,M$ into the ringdown, indicating informative $\chi_{\rm p}$ posteriors. For the even higher SNR signals of $\rho_{\rm pp} = \{75, 100\}$, there is information up to $t_{\rm cut} \sim 20\,M$. ***Pre-merger spin precession dynamics are imprinted very late into the ringdown.*** A log-scaled version of this plot and for other parameters can be found in Fig. 8.9 in Appendix 8.A. The total mass, for example, can be measured up to $t_{\rm cut} = 40{-}50\,M$.

In the bottom panel of Fig. 8.3, the two points circled in gray are an example of posteriors with comparable JSDs originating from signal regions with different SNRs. The $t_{\rm cut} = 5\,M$, $\rho_{\rm pp} = 25$ case (teal) yields a JSD of 0.033 over a region of data with SNR $0.89 \cdot 25 = 22.2$. The $t_{\rm cut} = 15\,M$, $\rho_{\rm pp} = 75$ analysis (coral), on the other hand, has a comparable JSD of 0.037 but an SNR that is a factor of two larger: $\rho = 0.62 \cdot 75 = 46.5$. ***It is thus not just the total SNR of the***

signal in a region that matters, but the actual morphology contained in that region.

Observing more inspiral bolsters spin precession inference

Having explored the measurability of spin precession in the post-peak (ringdown) data, we now turn to the pre-peak (inspiral) data. Observing more (less) of the inspiral maps to increasing (decreasing) the binary's merger frequency, which is achieved by decreasing (increasing) the system's detector-frame total mass [168], due to the detector's low-frequency sensitivity cutoff. For GW190521 and its max. $\mathcal{L}$ waveform, spin precession was not measurable in the inspiral alone—only one or two cycles are present for $t < 0\,\mathrm{M}$, as can be seen in the third row of Fig. 8.1.

Figure 8.4 shows results of measuring spin precession in signals with various total masses, ranging from $M = 80\,M_\odot$ to $500\,M_\odot$. Corresponding results for the measured total mass and mass ratio for these signals are provided in Appendix 8.C. Since all scales of the signal depend on the total mass, we here define waveform parameters for the simulated signals and during inference at $t = -100\,M$, ensuring that spin parameters are specified at the same point in the orbital evolution for all masses.[9] Defining parameters in this way guarantees that only M (and thus the amount of visible inspiral) varies, while the intrinsic spin configuration remains unchanged [16, 103]. We fix the total SNR of the system to that of the max. $\mathcal{L}$ signal ($\rho = 14$). The corresponding merger frequencies are listed in the second column of Table 8.2. The top row of Fig. 8.4 shows the whitened strain in LLO with time in units of seconds. Each signal has a different merger frequency (in Hz), and correspondingly a different fraction of SNR in the inspiral versus ringdown. For example, the $M = 80\,M_\odot$ signal (magenta) has the longest visible inspiral, while the $500\,M_\odot$ signal (light blue) has the longest visible ringdown. In the second row we repeat this plot but with time in units of M. Again we see that different signals have different amounts of observable inspiral, and additionally, all of the signals' peaks/troughs line up cycle-by-cycle when plotted in these time units.

We again truncate the data at different times and independently analyze before and after each cutoff. We use cutoff times of fixed M that correspond to

[9] This is different from other sections, where we map the spins to $t = -100\,M$ in post-processing. The max. $\mathcal{L}$ parameters for GW190521 analyzed at $t_{\mathrm{ref}} = -100\,M$ are given in the rightmost column of Table 8.1.

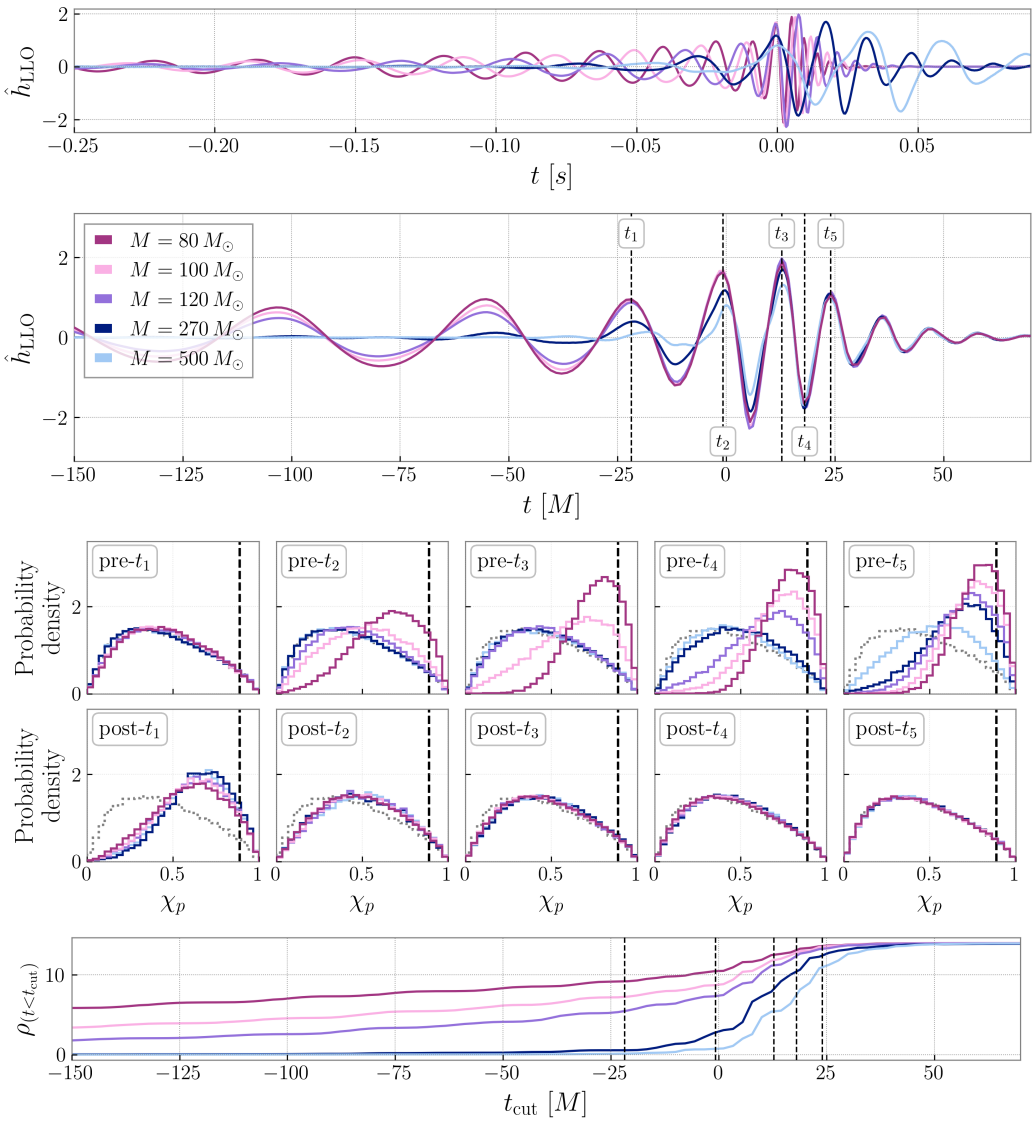


Figure 8.4: Waveforms and inference results for signals with different detector-frame total masses but the same SNR: $M = 80\,M_\odot$ (magenta), $100\,M_\odot$ (pink), $120\,M_\odot$ (purple), $270\,M_\odot$ (dark blue), and $500\,M_\odot$ (light blue). *First row:* Whitened waveforms in LLO, with the time axis plotted in units of seconds. *Second row:* Whitened waveforms in LLO, with the time axis plotted in units of detector-frame total mass. Since all other parameters are the same, the waveform cycles line up when plotted in units of M. Cutoff times $t_i \in \{-21.8, -0.8, 12.8, 18.0, 24.1\}\,M$, corresponding to specific peaks/troughs of the signals, are labeled with vertical black dashed lines. *Third and fourth rows:* $\chi_{\rm p}$ posteriors for the pre- and post-cutoff analyses for each signal (corresponding colors to top two rows). We exclude the full-signal analysis results, as they are indistinguishable from the pre-t_5 results. *Fifth row:* SNR accumulation over time for the different signals.

peaks/troughs of the signals (and thus to different numbers of seconds per signal), indicated by the vertical dashed lines in the second row of Fig. 8.4 and labeled as t_i. The posteriors for $\chi_{\rm p}$ from the pre-cutoff analyses are shown in the third row of Fig. 8.4. For reference, pre-t_2 corresponds most closely to just the inspiral data being analyzed. We exclude the full-signal analysis results from the figure, as they are indistinguishable from the pre-t_5 results. For all cutoff times, the lower the mass, the more informative the pre-cutoff signal is about spin precession. More visible inspiral for a system with the same intrinsic spin configuration yields a stronger $\chi_{\rm p}$ constraint, as ***precession has a clear imprint on the inspiral*** [e.g., 188, 203, 429]. For the lowest-mass signal we analyze ($M = 80\,M_\odot$; magenta), the $\chi_{\rm p}$ posterior is constrained away from the prior when *just* the inspiral is analyzed, with hints also present in the $M = 100\,M_\odot$ system (pink).

The post-cutoff posteriors are shown in the fourth row of Fig. 8.4. Although louder pre-peak data yield more informative $\chi_{\rm p}$ constraints, the same is not true about the post-peak data at these masses and SNR: here, in contrast to Sec. 8.4, increasing the total mass does *not* make the ringdown notably more informative about spin precession for these test cases, as can be seen in the post-t_2 and onward $\chi_{\rm p}$ posteriors. The signal with the loudest post-peak SNR is $500\,M_\odot$ with $\rho_{\rm pp} = 12$, which is consistent with being uninformative with respect to the varying-SNR analyses presented in Sec. 8.4. There, the $\chi_{\rm p}$ information in the ringdown shows up somewhere between a $\rho_{\rm pp}$ of 17 and 25. It is likely that for our test signals, the ringdown SNR is simply not large enough to measure precession.

Contrasting the pre- and post-cutoff results, the data between t_1 and t_2, i.e., the last pre-merger cycle, are necessary for measuring $\chi_{\rm p}$ in these different-mass signals. This is true regardless of how many cycles are observable before t_1. For example, we consider the $M = 80\,M_\odot$ case where precession is the most strongly constrained. The pre-t_1 and post-t_2 analyses are both uninformative about $\chi_{\rm p}$, while the pre-t_2 and post-t_1 analyses *are* informative. The final pre-merger cycle again seems to inform $\chi_{\rm p}$ measurements, no matter the total mass and amount of observable inspiral. In Sec. 8.4, we explore whether the cycle's importance holds true under different extrinsic angular configurations. While we directly observe that the orbital angular momentum tilts away from the line of sight in GW190521—regardless of total mass—we do not necessarily

expect this behavior to be universal. Still, for GW190521, the importance of the final pre-merger cycle persists at different total masses.

Even though all signals in Fig. 8.4 have the same total SNR, this SNR *accumulates* differently over time. This is shown in the bottom row, where we plot $\rho_{(t<t_{\rm cut})}$: the SNR of the signal before the corresponding time $t_{\rm cut}$ on the horizontal axis. The pre- and post-peak SNRs and their ratios are also given in Table 8.2. The relationship between total mass, pre-peak SNR, and $\chi_{\rm p}$ measurement strength is strictly monotonic: the larger the mass, the lower the pre-peak SNR, the less informative the signal is about $\chi_{\rm p}$. To aid in understanding the nature of this correlation, Fig. 8.5 shows the JSD between each pre-cutoff $\chi_{\rm p}$ posterior and prior versus the pre-cutoff SNR for many cutoff times. These two quantities are linearly correlated,[10] with a global line of best fit shown with a gray dashed line.

In other words, the SNR over a region of inspiral data has the *leading-order* effect on the measurability of $\chi_{\rm p}$ in that region, following common intuition. But as we found in Sec. 8.4, the SNR *alone* is not the full picture. The morphology contained within a given region bolsters or weakens precession measurability on top of the dominant effect of SNR, as we highlight in the inset axis of Fig. 8.5. In the regime where $\chi_{\rm p}$ is measurable (above the black dashed horizontal line), the $M = 80\,M_\odot$ signal is the most informative about precession (highest JSDs) at a fixed pre-cutoff SNR, and the JSDs decrease as total mass increases. The fact that, across the board, spin precession constraints weaken with total mass makes the fact that we mostly measure precession in high-mass events in real data [287] all the more intriguing.

The extrinsic angles of GW190521 were not finely tuned to measure spin precession

The measurability of spin precession in GW190521-like signals has been shown to depend sensitively on the exact intrinsic spin configuration of the system [103], making GW190521's informative $\chi_{\rm p}$ measurement seemingly fine tuned. We now explore the impact of the *extrinsic* projection and the orientation of the binary, specifically looking at three angular parameters: *inclination* ι, *polarization* ψ, and *phase* φ. We keep sky position (right ascension

[10] The Pearson correlation coefficient [430] between the pre-cutoff SNR and the $\log_{10}$ JSD is 0.95 with a p-value of 10^{-29}, strongly indicating a linear trend.

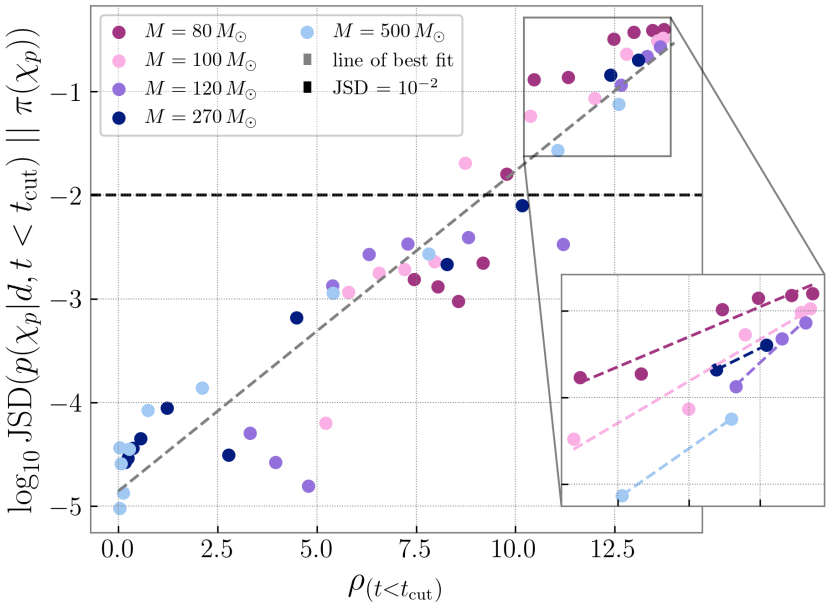


Figure 8.5: JSD between the pre-cutoff χ_{p} posterior and the prior versus the pre-cutoff SNR for different total masses (colors) and a range of cutoff times. The gray dashed line is the global line of best fit across all masses. The horizontal black dashed line marks the threshold we define for an informative posterior, JSD $\gtrsim 10^{-2}$. The inset zooms in on the informative region, with lines of best fit for each mass individually (colored dashed lines).

M	Merger f	Pre-peak ρ	Post-peak ρ	ρ ratio
80 $M_\odot$	205 Hz	10.5	8.3	1.3
100 $M_\odot$	164 Hz	8.7	9.5	0.9
120 $M_\odot$	137 Hz	7.4	10.4	0.7
270 $M_\odot$	59 Hz	2.9	11.6	0.3
500 $M_\odot$	31 Hz	0.8	12.2	0.1

Table 8.2: Merger frequency and SNR information of the signals with different total masses shown in Fig. 8.4. For comparison, the max. $\mathcal{L}$ signal has a merger frequency of 56 Hz. The method for extracting the merger frequencies from the data is given in Appendix 8.B; see Fig 8.11. The third and fourth columns list the optimal network SNRs for the pre-peak data (inspiral) and post-peak data (ringdown). Their ratio is given in the final column. The SNR is monotonically increasing/decreasing for the ringdown/inspiral as total mass increases. All systems have the same total SNR.

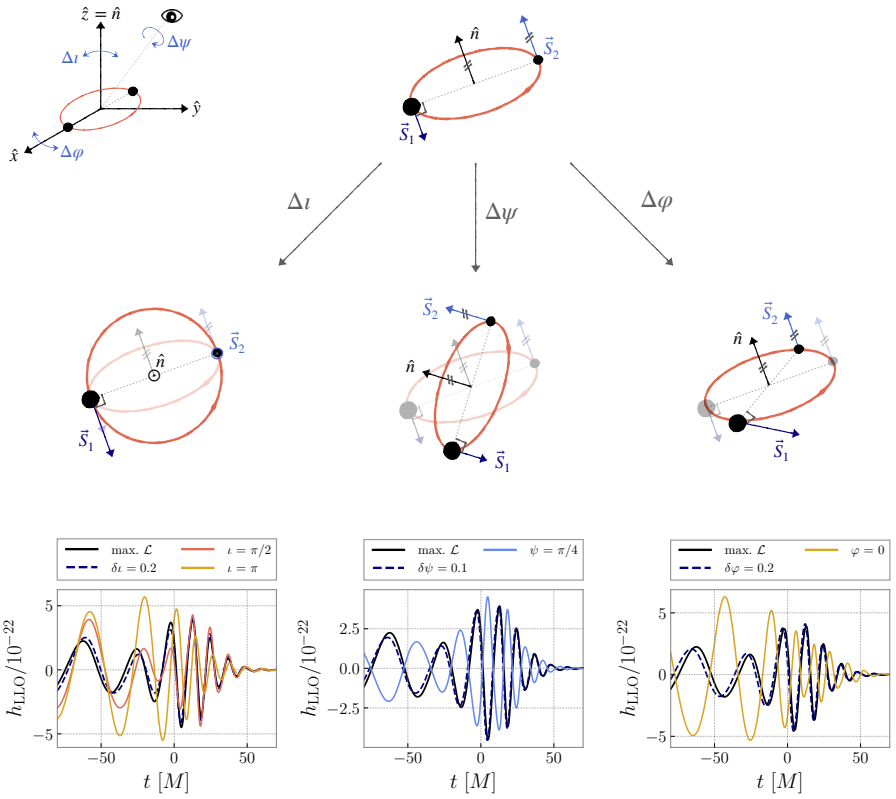


Figure 8.6: Schematic diagram explaining the effect of changing three extrinsic angle parameters on a binary's configuration and the resultant gravitational waveform: inclination ι (left), polarization ψ (middle), and phase φ (right). Changing each angle is equivalent to a rigid rotation of one of the Euler angles, as illustrated in the smaller diagram in the upper left, showing the binary's frame with respect to a stationary observer (eyeball). For the remaining diagrams, the binary is drawn on the plane of the sky, i.e., as viewed from an observer on Earth. The quasicircular orbit of the BBH is drawn in red, the spins $\vec{S}_i$ of the primary and secondary BHs are in blue, and the normal vector to the orbital plane $\hat{n}$ is in black. For demonstrative purposes, $\vec{S}_1$ lies in the orbital plane, perpendicular to the line connecting the BHs, while $\vec{S}_2$ is parallel to $\hat{n}$. Changing the inclination angle changes the degree to which the binary points toward/away from the viewer. This changes the projected ellipticity of the orbit onto the plane of the sky (the true binary is circular). Varying the polarization angle rotates $\hat{n}$ around the line of sight, leaving the shape of the ellipse projected onto the sky unchanged but altering its orientation. Finally, changing the phase angle keeps $\hat{n}$ fixed, but rotates the orbit around $\hat{n}$, keeping the spin components rigidly fixed in the binary's source frame. Under each diagram, we also show several examples of waveforms where the corresponding angle differs from GW190521's max. $\mathcal{L}$ (black) while the rest are kept constant. We show the effect of a small perturbation of each angle (dark blue dashed), and some extreme cases (yellow, red, and light blue). All signals have the same SNR; examples where just the angles are varied, leading to different SNRs, are shown in Fig. 8.13 in Appendix 8.D.

and declination) fixed for all simulated signals and in parameter estimation. The extrinsic angular parameters do not control a binary's intrinsic dynamics but do impact the *observed* GW signal—we provide equations for quantifying these dependencies in Appendix 8.D.

The phase and polarization are not astrophysically significant, meaning that they are expected to be randomly distributed across the binaries in the universe: there is no known astrophysical process or selection effect [340, 341, 431] that could tune the polarization or phase angle of a binary to specific values with respect to the Earth. Inclination too is expected to be approximately isotropically distributed in the astrophysical population.[11] However, due to selection effects, the inclination is not isotropic in the *observed* population. At a fixed distance/SNR, we are most sensitive to inclinations of $\iota = 0$ (face on) or π (face off); see Appendix 8.D. The degeneracy between inclination and luminosity distance [168, 301, 434–436] further impacts detectability. Thus, if the extrinsic angles of GW190521 are finely tuned (to non-face on/off configurations)—meaning that even a slight variation would render precession unobservable—that could be an indication that the waveform model is overfitting. If the true signal lies outside the waveform model's manifold, inference may yield sharply peaked posteriors in a narrow region of parameter space that best approximates the data [437], indicating model inefficiency. Alternatively, we may have simply been lucky: perhaps only heavy BBHs with a finely tuned configuration lead to an observable precession cycle, and GW190521 happened to be one such case.

Figure 8.6 provides a schematic illustration of the effect of ι, ψ, and φ on the observed configuration of a BBH's (non-eccentric) orbit and resultant GW emission. The top row depicts an example BBH as it would appear on the plane of the sky to an observer on Earth, with the orbit drawn in red and the individual black hole spins, $\vec{S}_i$, in blue. The normal vector $\hat{n}$ to the binary's orbital plane is shown in black, pointing at an angle out of the page. In the second row, we sketch how varying ι, ψ, and φ (from left to right) alters the

[11] The only BBH formation environment predicted to yield specific values of ι is the disks of active galactic nuclei (AGN), where BBHs' ι are preferentially aligned or anti-aligned with the orbital angular momentum of the disk [e.g., 120, 432]. This produces an astrophysical ι distribution with two sub-populations: one that is isotropic, and another clustered at specific inclinations of BBH-forming AGN-disks. Since AGN disk orientations are themselves isotropic with respect to Earth, the positions of these over-densities are also randomly distributed, and have not yet been observed with GW data [433].

observed BBH's geometry. In the third row, we show the impact of changing the three angles away from GW190521's max. $\mathcal{L}$ configuration (black) on the emitted GW signal projected onto LLO, showing both small perturbations (dark blue dashed) and more extreme deviations (yellow, red, and light blue solid) for each parameter. Within each column, all waveform parameters but two are kept constant: the varied angle, and luminosity distance (which is scaled such that all signals have the same SNR). The specific combination of ι, ψ, and φ plays a key role in determining how much of a precession cycle is visible in a waveform.

In the following bullet points, we detail how each extrinsic angle affects the BBH orbit and waveform. All angles are defined at the same time in the signal ($f = 11\,\mathrm{Hz}$, corresponding to $t = -212\,M$, all systems have the same total mass).

- The *inclination angle* ι determines the tilt of the binary's orbit relative to the observer's line of sight, i.e., the dot product between $\hat{n}$ and the line of sight. While the binary's true orbit is circular, changing ι affects the ellipticity of the orbit projected onto the plane of the sky. If viewed face-on/off ($\iota = 0$ or π), the projected orbit is a circle (left sketch in the second row of Fig. 8.6); if viewed edge-on ($\iota = \pi/2$) it is a line. The inclination angle can significantly alter the observed signal morphology, as it changes what combination of GW radiation modes are observed; see Appendix 8.D. For instance, at $\iota = \pi$ (yellow waveform in lower-left subplot of Fig. 8.6) the final pre-merger cycle has a large amplitude compared to the merger, whereas $\iota = \pi/2$ (red waveform) yields an extreme suppression of this same cycle.

- Changes in the *polarization angle* ψ rotate the projected orbit on the plane of the sky without altering its shape/ellipticity, shifting the waveform's polarization state [181]. This is shown in the center diagram of the second row of Fig. 8.6. Crucially, this rotation does not make the $\hat{n}$ point more toward or away from the observer, but rather changes its azimuthal angle around the observer's line of sight. In other words, changes in ψ keep the dot product between $\hat{n}$ and the line of sight constant. Additionally, varying ψ fully preserves the complex amplitude envelope of the waveform (Fig. 8.13), but shifts the timing of the peaks and troughs of the strain, making it the most intuitive form of a "phase shift" for precessing BBHs.

- Finally, variations in the so-called *phase angle* φ rotate the binary's orbit *around* $\hat{n}$—equivalently, rotate the BHs along their fixed orbit—as shown in the right-hand diagram of the second row of Fig. 8.6. Importantly, this rotation keeps the spin components rigidly fixed in the binary's source frame, i.e., the Cartesian components of the spins are the same in the frame of the source, where the x-axis is defined as the line from m_2 to m_1 at the reference time, and the z-axis is along $\hat{n}$ [16, 48]. In precessing BBHs—where higher-order modes are significant—φ plays a more complicated role than the simple phase shift from which it derives its name, which only applies in the case of non-precessing, equal-mass systems; see Appendix 8.D. Like the inclination angle, φ can significantly alter the signal morphology for precessing systems, as exemplified by the yellow waveform in the lower right-hand plot of Fig. 8.6.

As mentioned, Appendix 8.D provides further details about the role of inclination, polarization, and phase in precessing systems by giving equations. There we also contrast the impact of the phase angle on precessing systems versus with equal-mass non-spinning systems.

Figure 8.7 shows results from full signals with a range of inclinations (left column), polarizations (middle column), and phases (right column). All signals are scaled to have the same full-signal SNR as GW190521's max. $\mathcal{L}$, and—aside from the one differing angle—share its other parameters. In inference, only the sky position, time of coalescence, and polarization are fixed to their injected values, as was true in the analyses of the previous sections. The top row shows $\chi_{\rm p}$ posteriors for each signal (various colors), compared to GW190521's max. $\mathcal{L}$ posterior (black) and the prior (gray dotted). The second row plots the JSD between each posterior and the prior as a function of the varied angle. With all other parameters fixed to GW190521's max. $\mathcal{L}$ configuration at $\rho = 14$, we are most sensitive to precession at $\iota \sim 1.5 - 2$ radians, $\psi \sim 0.8$ or 2.2 radians, and $\varphi \sim 0.8$ radians. ***GW190521's*** max. $\mathcal{L}$ ***reconstruction is not the most informative about spin precession over all the extrinsic angle configurations***. GW190521's max. $\mathcal{L}$ signal is tied for most informative over the polarization angle (although the JSD variation between the different values of ψ is minute), and falls squarely in the middle over inclination and phase angle. ***This configuration—which is very similar to the real GW190521 data—is not exceptionally fine-tuned to measure spin precession***.

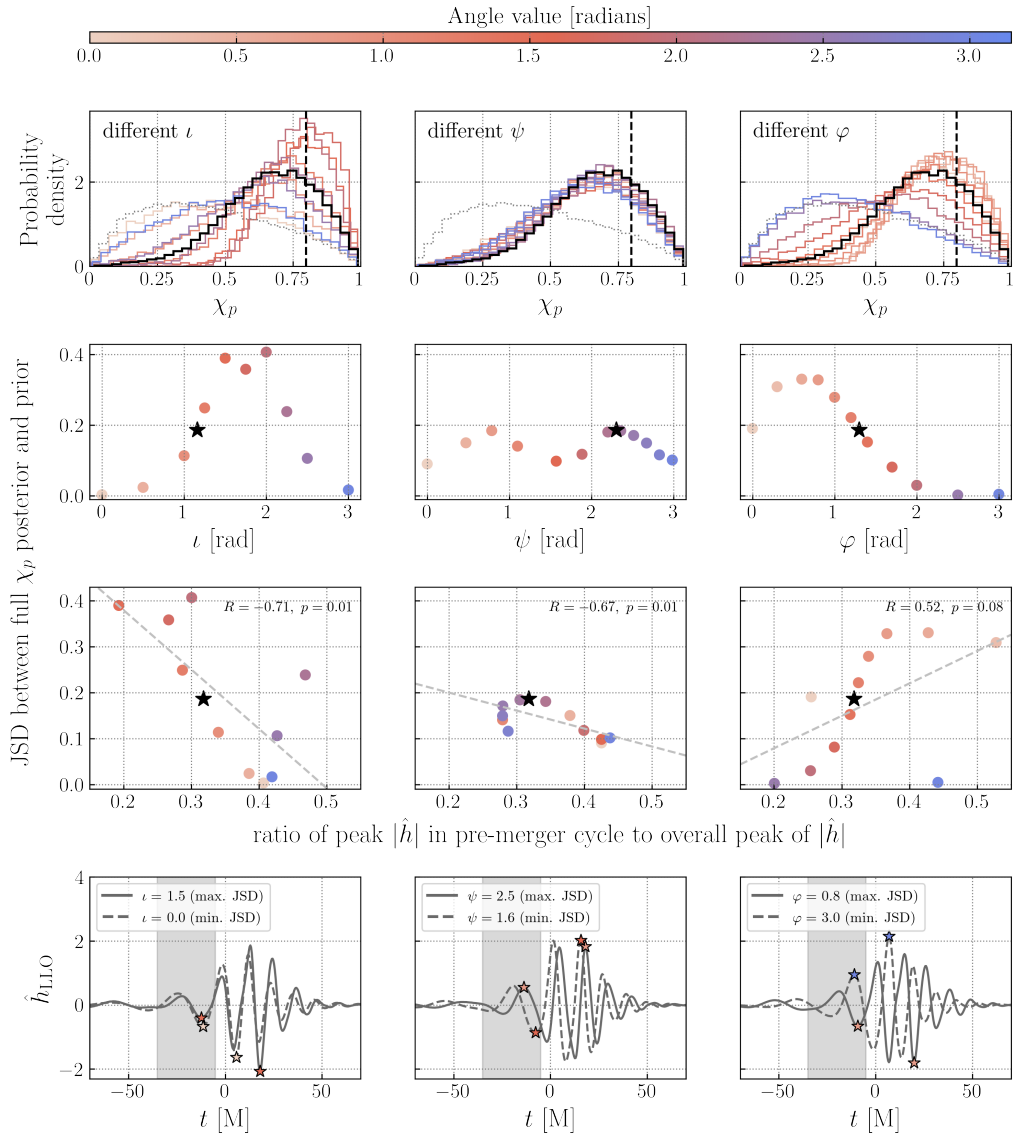


Figure 8.7: Results from full-duration signal analyses where one extrinsic angle at a time is changed from GW190521's max. $\mathcal{L}$ configuration: inclination angle ι (left column), polarization angle ψ (middle column), and phase angle φ (right column). (*First row*) χ_{p} posteriors for all signals with varied angles (colorbar) compared to the max. $\mathcal{L}$ posterior (black) and prior (gray dotted). (*Second row*) JSD between the marginalized χ_{p} posteriors and the prior as a function of each angle. GW190521's max. $\mathcal{L}$ signal is indicated with a black star. (*Third row*) JSD versus the ratio of the peak whitened strain in the pre-merger cycle to the overall signal strain peak. In gray-dashed are lines of best fit, intended to visualize the net trend strength and direction. Pearson R correlation coefficients [430] and associated p-values are in each upper right corner. (*Fourth row*) Comparisons of the whitened strain in LLO of signals which are the most informative (maximum JSD, solid) versus least informative (minimum JSD, dashed) about χ_{p}. Stars show the values extracted to calculate the x-axis quantity in the third row.

Each angle affects the measurability of $\chi_{\rm p}$ in different ways. In our test cases, the polarization angle ψ does not significantly impact the $\chi_{\rm p}$ constraints for the full signal,[12] whereas ι and φ each have much more dramatic effect, yielding $\chi_{\rm p}$ posteriors that range from uninformative (JSD $\sim 10^{-3}$) to being very informative (JSD ~ 0.4), even though all signals have the same SNR. The strong influence of ι is expected [103, 104]. Sensitivity to φ is also reasonable, given its impact on signal morphology (Fig. 8.6), stemming from its effect on the superposition of higher-order modes [184, 185, 438] (Appendix 8.D). The polarization angle has been suggested to affect the measurability of precession for certain inclinations, based on its impact on the precessing SNR [104, 201, 376]. Ref. [104] find that the precessing SNR is sensitive to ψ (cf. their Fig. 8), although that study was based on a different GW190521-like signal with a different value of φ and different intrinsic spin configuration. The individual effects of extrinsic angles are difficult to disentangle; varying them independently while holding others fixed, such as when calculating the precessing SNR, fails to fully capture their joint influence across the full parameter space. As a result, polarization may impact $\chi_{\rm p}$ measurability in some inclination-phase configurations (as in Ref. [104]), but not in others (as seen in our results). Ultimately, it is the position within the higher-dimensional space of extrinsic (and intrinsic) parameters that governs the influence on precession measurability [103].

Next, we evaluate how precession measurability varies with signal strength at different times. We focus on the relationship between the pre-merger versus merger strain amplitude. Ref. [101] found an anti-correlation between the pre-merger strain amplitude and $\chi_{\rm p}$ measurability, which we now test across the various extrinsic angle configurations. The third row of Fig. 8.7 shows the ratio of the signal's peak whitened strain amplitude in the final pre-merger cycle to that in the merger, plotted against the JSD between the full $\chi_{\rm p}$ posterior and prior. The peak merger amplitude (denominator of ratio) is the maximal value of absolute value of the whitened strain $|\hat{h}|$ across all times. Defining the final pre-merger is not precise nor universal, as it involves deciding the time at which the inspiral ends and merger begins. In Fig. 8.7, the peak final pre-merger strain amplitude (numerator of ratio) is calculated as the first local

[12] While our inference was conducted with ψ fixed, we have verified that the results remain robust when varying it: we perform inference for a selection of test cases while varying ψ and find results unchanged.

maxima of $|\hat{h}|$ under the conditions $t < -5\,M$ (shaded gray band) and $|\hat{h}| < 1$. We choose $-5M$ as a heuristic time for defining the end of the final pre-merger cycle because it is before peak emission but not so far into the inspiral that the signal becomes unobservable for GW190521's mass. The $|\hat{h}|$ cap of 1 removes ambiguity for several cases (e.g., $\varphi = 0$) whose final-pre merger cycle seems, by eye, to be pushed even earlier.

The relationships between $\chi_{\rm p}$ informativeness and the relative strength of the pre-merger cycle's amplitude are not linear.[13] The best-fit lines (gray dashed) are intended to illustrate first-order trends, but clearly do not capture the underlying correlational structure. If the findings for GW190521 are generic, we would expect a consistent negative correlation, meaning smaller strain amplitude ratios correlate with more informative $\chi_{\rm p}$ posteriors. However, ***the relationship between $\chi_{\rm p}$ and the ratio of pre-merger to overall strain is not generic***. The trends are in general weak, especially when varying ψ, and are not consistently negative. The overall trend found when varying φ is positive (although not well fit by a line), the opposite to what one might have expected. The bottom row of Fig. 8.7 highlights the signals that are the most (solid lines) and least (dashed lines) informative about $\chi_{\rm p}$. The stars mark the pre-merger and overall strain values whose ratio is plotted in the third row. Interestingly, in these extreme cases, the trend from GW190521 *does* hold: the most informative signals exhibit a more suppressed pre-merger cycle (gray shaded region) than the least. This is particularly curious for φ where the overall trend is not negative—the $\varphi = 3$ signal is an outlier in its uninformative $\chi_{\rm p}$ measurement which we believe relates the fact that it has a cycle at $t \sim 5\,M$ with a peak amplitude in LLO right below our threshold of $|\hat{h}| < 1$.

Having studied how the extrinsic phases matter for spin precession measurability, we now look further into what part of the signal impacts these measurements by turning to the pre-/post-cutoff-time $\chi_{\rm p}$ results. Figure 8.8 shows post-cutoff results for $t_{\rm cut} \in \{-30, -20, -10, 0\}\,M$ (left to right columns). The pre-cutoff results at these times for all angular configurations are consistent

[13] The strain ratio and JSD have Pearson R correlation coefficients [430] of -0.71, -0.67, and 0.52 when varying inclination, polarization, and phase respectively. The first two relationships have p-values of 0.01, indicating that the linear fit is good. For phase, the p-value is 0.08, indicating the linear relationship is insufficient to describe the data. This can all be confirmed by eye, looking at the third row of Fig. 8.7.

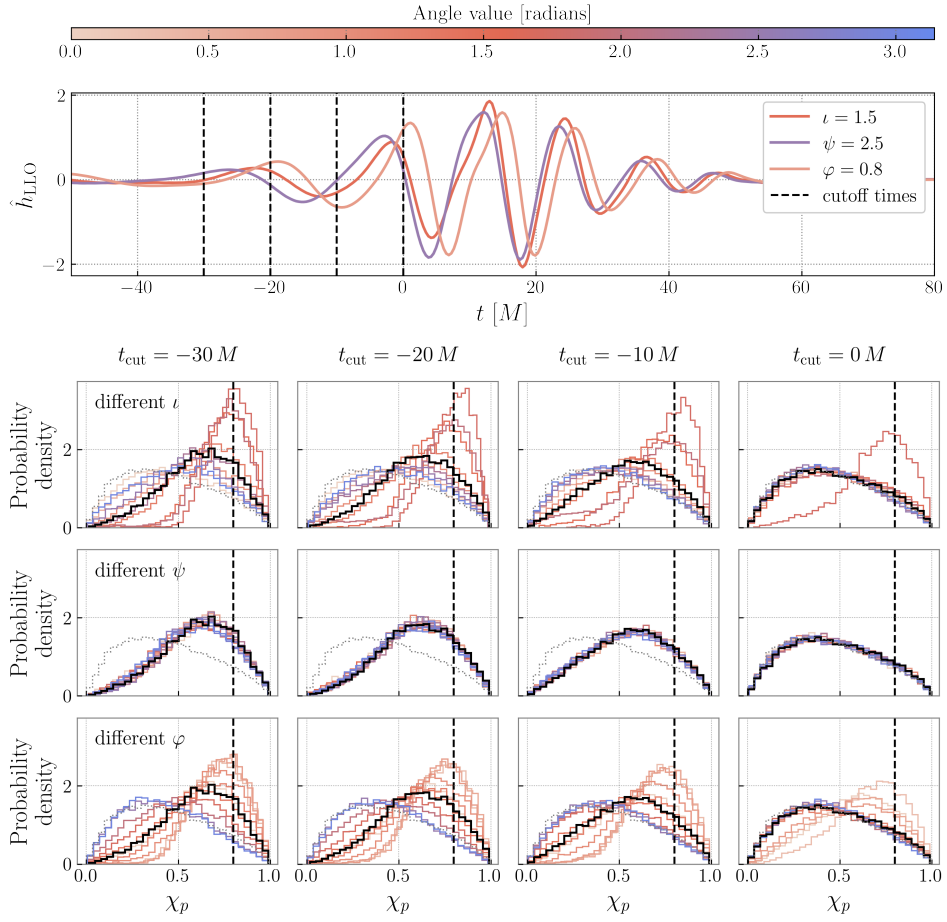


Figure 8.8: Results from post-cutoff analyses for signals with various extrinsic angles. (*First row*) Waveforms from the most informative variation of each angle, as identified in Fig. 8.7, with cutoff times shown for reference (black dashed lines). All signals have the same total mass. (*Second through fourth rows*) $\chi_{\rm p}$ posteriors for the post-cutoff analyses where $t_{\rm cut} = -30\,M$ (first column), $-20\,M$ (second column), $-10\,M$ (third column), and $0\,M$ (fourth column). Each row shows posteriors from simulated signals where one extrinsic angle is changed from the max. $\mathcal{L}$ injection (black): inclination angle (top row), polarization angle (middle row), and phase angle (bottom row). The color bar indicates the value of the angle that is varied from the max. $\mathcal{L}$ configuration. The $\chi_{\rm p}$ prior is shown in gray dotted lines. We here only show post-cutoff analysis results, as the pre-cutoff $\chi_{\rm p}$ posteriors for all signals at the selected $t_{\rm cut}$ values are consistent with the prior.

with the prior and are, therefore, omitted. The whitened waveforms in LLO for the most informative variation of each angle from the full-signal analyses are shown in the top row for reference, with the cutoff times marked with black dashed lines. Results with varied ι, ψ, and φ are given in each of rows two through four.

Just like for the full signal, varying ψ does not affect the post-cutoff posteriors (third row of Fig. 8.8). Against our intuition, all posteriors evolve identically despite differences in the timing of peaks and troughs of the strain. Here we are cutting at fixed *times* in the strain, which do not correspond to a fixed number of cycles from merger for different angular configurations. The polarization angle does not change the signal's amplitude envelope (Appendix 8.D). Thus, we posit that the *amplitude envelope may matter more for constraining precession than the precise timing or values of strain maxima and zero-crossings*. Furthermore, the polarization angle does not affect the evolution of a binary's inclination over time, meaning it does not change what part of a precessional cycle is visible in the data, complicating our interpretation from GW190521 [101] of what underlying physics leads to precession being measurable.

On the other hand, ι and φ matter substantially for $\chi_{\rm p}$ evolution(second and fourth rows of Fig. 8.8), and the evolution of inclination over time. ***A subset of signals with specific values of ι and φ have ringdowns*** (post-peak data, fourth column of Fig. 8.8) ***that are informative about $\chi_{\rm p}$ alone, even at the same SNR as GW190521.*** At GW190521's SNR and intrinsic spin configuration, observable amplitude modulations in the inspiral due to changing the direction of GW emission are not prerequisite to measuring precession—it can instead be observed by its imprint via mode mixing on the ringdown alone [157].

8.5 Conclusions

Ever since the detection of GW190521, there has been a need to understand how spin precession can be imprinted in the signals from heavy compact binaries, which are only observable in current ground-based gravitational-wave detectors for a few cycles. Ref. [101] used time-domain inference to localize the measurement of precession in GW190521: it originates from a subtle interaction between the pre- and post-peak signal, rather than from any iso-

lated segment of the data, e.g., inspiral or ringdown alone. In this paper, we built on Ref. [101] and explored the morphological imprints of spin precession in simulated GW190521-like (high-mass, highly-precessing) GW signals using time-domain inference. The goal was to understand whether the features observed in GW190521 can be replicated in simulated signals and whether its configuration was special in any regard, building on past studies [103, 104]. We focused on the effects of SNR, total mass, and extrinsic angular parameters on the measurability of spin precession over a range of time segments, with an emphasis on pre- and post-peak data.

We first showed that GW190521's behavior is indeed mimicked by its max. $\mathcal{L}$ injection, for both full and truncated data (Fig. 8.1), and that as we increase the SNR of this signal precession becomes, as expected, better constrained (Fig. 8.2). When varying the total mass at the same SNR, we found that signals from lighter systems yield more informative $\chi_{\rm p}$ posteriors. Precession is known to be measurable in the inspiral [e.g., 188, 203, 429], so this trend is expected, as lower-mass systems have a longer visible inspiral. For a low enough total mass ($M \lesssim 100\,M_\odot$), precession is measurable when *only* the inspiral data are visible (Fig. 8.4). For higher-mass events, precession can become more or less apparent for certain angular configurations. We find that the ability to measure spin precession in GW190521-like signals is sensitive (although *not fine-tuned*) to the extrinsic angular configuration. Moreover, the specific waveform cycles that inform the $\chi_{\rm p}$ measurement differ as we change the system's angular configuration. While there are more than a few combinations of inclination, polarization, and phase that lead to $\chi_{\rm p}$ being constrained away from the prior, specific values of the inclination and phase angles do render precession unmeasurable (Fig. 8.7). At GW190521's max. $\mathcal{L}$ inclination and phase, the polarization angle does not strongly affect $\chi_{\rm p}$ constraints for either the full or any pre-/post-cutoff data, which may be explained by their shared amplitude envelope.

At high enough SNRs, we can measure precession in the ringdown alone using an IMR waveform model (Fig. 8.2). When the post-peak SNR is sufficiently large ($\rho_{\rm pp} \sim 100$), precession information remains in the signal up to $\sim 20\,M$ after the peak (Fig. 8.3). Past work has explored this same idea with QNM ringdown models, for both for GW190521 itself [107, 214, 215, 398] and various precessing numerical relativity simulations [151, 153, 155–157] rather than

an IMR waveform model like NRSur7dq4. The fact that the ringdown has imprints of precession is not unexpected for sufficiently high SNRs. Ref. [157] found that, for near-equal-mass binaries, the ratio between the amplitudes of the 200:220 and 210:220 modes are correlated with the remnant spin misalignment angle—a proxy for spin-orbit misalignment and, therefore, precession. In this vein, Ref. [214] found that the GW190521 ringdown data are consistent with a significant 210 QNM amplitude, which would be unexpected for a non-precessing BBH.

Even at GW90521's mass and SNR of ~ 14, certain combinations of inclination and phase render precession measurable in the ringdown (Fig. 8.8). This finding is in agreement to the complementary QNM studies discussed above [e.g., 157, 214]: the information about spins before the merger is expected to be encoded in not just the observed QNM complex frequency spectrum—which suffers from degeneracies inherent to the map from inspiral to remnant properties—but also these *modes' relative amplitudes and phases*. A direct consequence of the no-hair conjecture [e.g., 439] in General Relativity is that complex ringdown frequencies are uniquely determined by mass and spin; multiple sets of progenitor properties can yield the same BH remnant, and thus the same ringdown frequencies. Even though these degeneracies may explain why, for example, we are able to infer precession from an SNR ~ 10 inspiral (Fig. 8.4) but not an SNR ~ 10 ringdown (Fig. 8.3) generated by a BBH with the same spin and angular configuration, they do not entirely govern the circumstances that render precession's imprint informative. Which ringdown modes manifest and with what complex amplitudes are determined by the masses and spin magnitudes *plus* the BBH's full spin/angular configuration [440]. Modeling the imprint of spin precession and other inspiral dynamics on the ringdown is an area of active research [153, 155, 156, 215, 223–225, 225, 440], and a closer comparison with NR surrogate findings is warranted in a future study.

Ref. [103] and Ref. [104] have both conducted complementary studies on measuring spin precession in high-mass BBHs, both using frequency-domain inference. Ref. [103] explored various total mass and spin configurations at selected mass ratios, inclinations, and SNRs. They found that precession measurability is sensitive to the *intrinsic* spin dynamic, such as spin tilt angles. Furthermore, they concluded that the uncertainty in spin measurements increases with the total mass, as do we. They simulated signals and measured spins at a fixed

frequency in Hz rather than a fixed time in units of M, meaning their different mass signals *also* had different intrinsic spin configurations. When comparing results with spins fixed at the same reference time, they found a stronger trend (cf. their Fig. 11), which is even more in agreement with ours. Formulating our results in terms of the symmetric 90% credible interval width (CI_{90}) used in Ref. [103] to assess a posterior's precision yields analogous results: between a total mass of 80 and 250 $M_\odot$, the CI_{90} for full-signal χ_p posteriors increases by ~ 0.2. In terms of the extrinsic angles, they also observed improved χ_p measurement at higher inclinations (1.5 radians vs. 0.5 radians), consistent with our findings.

Ref. [104] examined the role of detector noise in GW190521-like signals, concluding that it impacts spin precession measurability in the expected way. They also analyzed variations in intrinsic spin configuration, SNR, and inclination/polarization. They suggested that spin precession is not measurable in the full signal until SNR 45, whereas we found that at this SNR spin precession is measurable in the ringdown alone, and measurable in the full signal at SNR 14. In their test cases, they found that inclination matters less and polarization matters more than we do. Together, these studies and ours highlight the dependence of spin precession measurability on multiple interdependent parameters.

More broadly, we found that not just the SNR in a region of a signal matters for constraining precession, but also the specific signal morphology contained within that region (Figs. 8.3 and 8.5). Furthermore, the morphological features needed to measure spin precession are not the same for all high-mass signals. We do not find that a loud merger and quiet pre-merger cycle consistently drive spin precession constraints in high-mass signals, as was the case for GW190521. While precession inference sometimes hinges on this morphology, the informativeness of a χ_p posterior is not universally anti-correlated with the ratio of the signal's pre-merger to overall amplitude peak (Fig. 8.7). As we vary SNR, total mass, and the extrinsic angles, different signals lose information about χ_p at different times in their evolution.

Sometimes, specific waveform features directly correlate with physical properties, e.g., the merger frequency and total mass, or the ratio of the 221:220 ringdown mode amplitudes and the angle between the remnant BH's spin and the progenitor's angular momentum vector [157]. We tested whether an

analogous universal feature links $\chi_{\rm p}$ and the late inspiral and merger strain amplitudes, but did not identify any such relationship.

Our results show that, while not fine-tuned, our ability to infer precession is *sensitive* to a BBH's total mass and angular parameters at current detector sensitivity. While this is not a selection effect in the traditional sense—as it does not impact the *detectability* of events—it does analogously influence the *measurability* of the associated astrophysical phenomena. This sensitivity necessitates hand selecting a small number of highly precessing events on which to perform follow-up studies such as this. We may be mis-identifying other events as non-precessing due to imprecise measurements. Perhaps there is an observed GW whose spin precession *would* have confidently been observed if the system had a different extrinsic angular configuration. Looking to the future, ground-based GW detectors will become increasingly sensitive, alleviating these concerns.

In LIGO-Virgo-KAGRA's upcoming fifth observing run ("A_+ sensitivity") and future "$A_\sharp$ sensitivity,"GW190521's SNR would be $\sim$ 40 and $\sim$ 80 respectively [441, 442]. In Cosmic Explorer, it would increase to $\sim$ 450 [443]. This means that, had this event been observed at even A_+ sensitivity, precession would have been constrained in GW190521's post-peak data alone. Looking ahead, if precessing BBHs are abundant in the cosmos—as theoretical models suggest—we will be able to identify that they are indeed precessing more robustly, opening the door to a stronger understanding of how and where BBHs form.

Ch. 8 Data and Code Availability

The code `tdinf` [161] was written by the same authors as this manuscript and was used to perform all of the time-domain analyses presented in this work. Notebooks to generate all the figures appearing in the text are available on Github at `https://github.com/simonajmiller/heavy-precessing-bbh-time-domain` The repository additionally contains supplemental figures at and animations showing results from more cutoff times and simulated signals:

- `gif_figure_01`: Animation of M, q, and $\chi_{\rm p}$ posteriors for GW190521 and its max. $\mathcal{L}$ injection, similar to Fig. 8.1.

- gif_figure_02: Animation of $\chi_{\rm p}$ posteriors for different SNR signals, similar to Fig. 8.2.
- gif_figure_04: Animation of $\chi_{\rm p}$ posteriors for different total mass signals, similar to Fig. 8.4.

Data necessary to reproduce figures and animations are available on Zenodo at https://zenodo.org/record/15474960. The data release includes posteriors for M, q, $\chi_{\rm eff}$, and $\chi_{\rm p}$ for each analysis. Additional data are available upon request.

Ch. 8 Acknowledgments

We thank Harrison Siegel, Rhiannon Udall, Eliot Finch, and Will Farr for helpful discussions about ringdown analyses and parameter estimation, as well as Barry McKernan and K. E. Saavik Ford for insights into the expected astrophysical BBH inclination angle distribution. SJM and KC were supported by NSF Grants PHY-2308770 and PHY-2409001. The Flatiron Institute is a division of the Simons Foundation. VV acknowledges support from NSF Grant No. PHY-2309301 and UMass Dartmouth's Marine and Undersea Technology (MUST) Research Program funded by the Office of Naval Research (ONR) under Grant No. N00014-23-1-2141. SH was supported by the National Science Foundation Graduate Research Fellowship under Grant DGE-1745301. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. The authors are thankful for LIGO Laboratory computing resources, funded by the National Science Foundation Grants PHY-0757058 and PHY-0823459, as well as the Hawk computing cluster provided by Cardiff University and supported by STFC Grants ST/I006285/1 and ST/V005618/1. This research has made use of data or software obtained from the Gravitational Wave Open Science Center (gwosc.org) [406, 407], a service of the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation, as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Coun-

cil. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. KAGRA is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica (AS) and National Science and Technology Council (NSTC) in Taiwan.

Software: `tdinf` [161], EMCEE [252], LALSUITE [319], PYCBC [444], GWPY [445], NUMPY [323], SCIPY [353], H5PY [402], H5IFY [446], MATPLOTLIB [320], SEABORN [354], RINGDOWN [222, 367], GWTOOLS [403], PANDAS [324, 325].

Ch. 8 Appendices

8.A Measuring more mass and spin parameters in the post-peak data and exploring conditional priors

Section 8.4 explores how late into the post-peak signal $\chi_{\rm p}$ can be measured at different SNRs by computing the JSD between the post-cutoff posterior and prior as the cutoff time $t_{\rm cut}$ increases. Figure 8.9 presents an analogous version to the bottom panel of Fig. 8.3, here showing the $\log_{10}$ JSD evolution of more parameters: $\chi_{\rm eff}$, M, and q, along with $\chi_{\rm p}$ for comparison. The signals lose information about each parameter at different times: first $\chi_{\rm p}$, then $\chi_{\rm eff}$, then q, and finally M. The effective spin $\chi_{\rm eff}$ remains the least informative throughout the whole signal, never yielding a JSD greater than 0.01, even in the highest-SNR case. The total mass M has the most information at late times in the ringdown, yielding JSDs $\gtrsim 10^{-2}$ (our approximate threshold for measurability) up to $t_{\rm cut} = 40{-}50\,M$ after the signal's peak. This is expected, since M is directly proportional to the frequency of the ringdown modes, which is a primary observable [15, 222].

After the point where posteriors become uninformative, their actual $\log_{10}$ JSD values become less meaningful, due to the increase in fractional uncertainty intrinsic to finite random sampling. Posteriors for $\chi_{\rm p}$ with, e.g., JSD $= 10^{-3}$ and JSD $= 10^{-4}$, are indistinguishable by eye. For reference, in Fig. 8.9 we plot JSDs between distributions approximated by sets of $N = 3000$ random samples from each marginal prior (gray histograms), which is effectively a null distri-

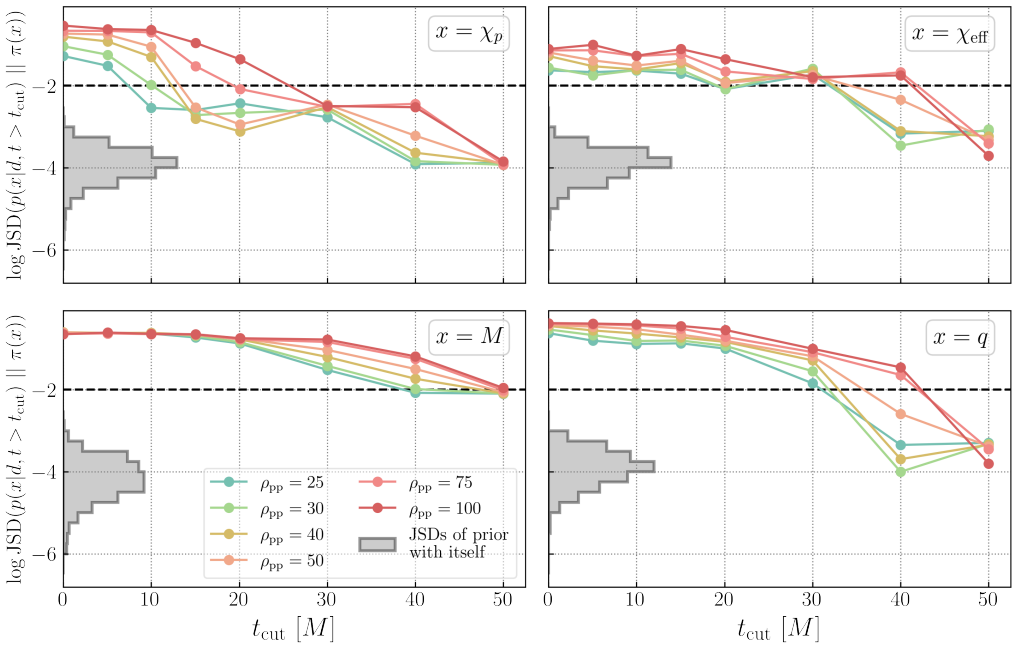


Figure 8.9: Jensen Shannon Divergence ($\log_{10}$ JSD) between the posterior $p(x|d, t > t_{\rm cut})$ and prior $\pi(x)$ for $x \in \{\chi_{\rm p}, \chi_{\rm eff}, M, q\}$ from data after different cutoff times $t_{\rm cut}$ (horizontal axis) for different signals with different post-peak SNRs ($\rho_{\rm pp}$; different colors). Distributions of the JSD between two sets of $N = 3000$ samples from the marginal prior for each parameter are shown in gray histograms; our conservative benchmark for measurability of $\log_{10}$ JSD $= -2$ is indicated by the black dashed line. See Fig. 8.3 and the corresponding text for more details. The top left panel is the same as the bottom panel of Fig. 8.3 except in log-space.

bution for what the JSD can look like when a parameter is fully uninformed by the data. Our conservative threshold for measurability, $\log_{10}$ JSD $= -2$, is shown with a black line, slightly above these histograms. At $t = 30\,M$, both the $\chi_{\rm eff}$ and $\chi_{\rm p}$ JSD curves cluster at a value just above or below this threshold. There is no apparent unusual feature in the waveforms or posteriors at this time. The clustering likely reflects random variation as posteriors become poorly constrained. From this point onward, the $\chi_{\rm p}$ posteriors at all SNRs are uninformative. The correlation between $\chi_{\rm p}$ and the (uninformative) $\chi_{\rm eff}$ explains its clustering at the same cutoff time.

Correlated priors between two (or more) parameters can lead to spuriously strong one-dimensional JSDs, as discussed in Ref. [50]. We use the data in Fig. 8.9 to explore the degree to which conditional priors are affecting our results. As we are primarily concerned with measuring spin precession, we

investigate correlations between $\chi_{\rm p}$ and the other parameters. The inference prior is uniform in M, q, and spin magnitudes, and is isotropic in spin angles. The prior between M and $\chi_{\rm p}$ is uncorrelated, which we validate by calculating a negligible Pearson correlation coefficient [430] between samples drawn from the prior. The test yields a p-value of 0.81, indicating non-correlation. On the other hand, the priors on q, $\chi_{\rm eff}$, and $\chi_{\rm p}$ are all correlated [238, 239].

To assess how conditional priors from measuring $\chi_{\rm eff}$ and q influence the $\chi_{\rm p}$ results, we compute the JSD between the full $\chi_{\rm p}$ prior and the $\chi_{\rm p}$ prior conditioned on $\chi_{\rm eff}$ and/or q being close to the true value. We focus on conditioning using the $\rho_{\rm pp} = 100$ post-$t_{\rm cut} = 0$ posterior because it provides the most informative measurements, making potential effects the most pronounced. We define "close" for $\chi_{\rm eff}$ as within one standard deviation of the posterior relative to the injected value. For q, since all posteriors peak near the true value at $q = 1$, we define "close" as greater than the $5^{\rm th}$ percentile of the posterior. Conditioning on $\chi_{\rm eff}$ adds a negligible 2×10^{-3} nats of information to $\chi_{\rm p}$, while conditioning on q or both parameters increases it by 2×10^{-2} nats. This suggests that $\chi_{\rm eff}$ has little impact on $\chi_{\rm p}$ constraints, which aligns with Fig. 8.9 where $\chi_{\rm eff}$ is shown to contribute less information than the mass parameters. However, we must evaluate whether and how q influences our $\chi_{\rm p}$ measurements.

Instead of defining the measurability of precession by comparing a given $\chi_{\rm p}$ posterior to its *full* prior, we instead calculate the JSD between the posterior and the *conditional* prior [50]. Figure 8.10 shows how the JSD is affected by this correction as a function of cutoff time, for the highest (red) and lowest (teal) SNR case of Fig. 8.9. For each cutoff time, we compute the conditional $\chi_{\rm p}$ prior by selecting only $\chi_{\rm p}$ prior samples where the corresponding q exceeds the $5^{\rm th}$ percentile of the post-cutoff posterior. The JSDs from the unconditioned $\chi_{\rm p}$ priors (solid lines) are always larger than those from the priors conditioned on q (dashed lines), although the effect is small. Crucially, the conditioning does not alter the cutoff time at which the $\chi_{\rm p}$ posteriors become uninformative (i.e., JSD $\lesssim 10^{-2}$), nor the general trend. In the $\rho_{\rm pp} = 25$ case, this transition remains between $5{-}10\,M$ after the signal peak, while for $\rho_{\rm pp} = 100$, it remains at $20{-}30\,M$.

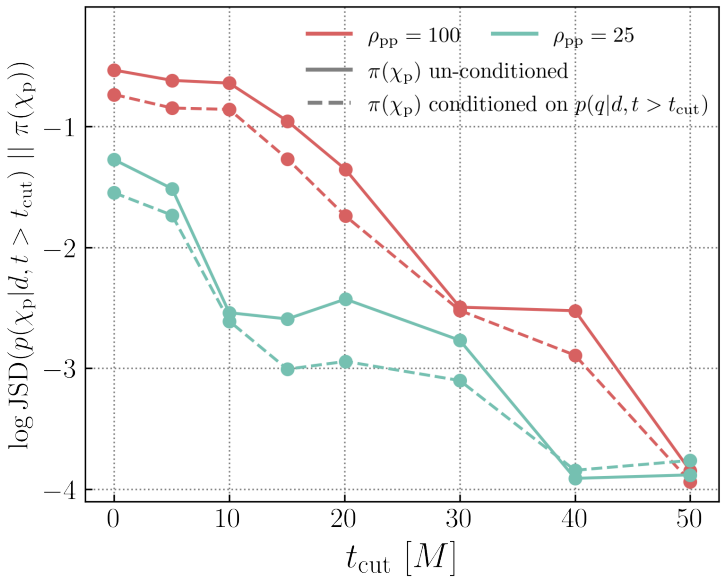


Figure 8.10: $\log_{10}$ JSD between the $\chi_{\rm p}$ posterior and full prior $\pi(\chi_{\rm p})$ (solid lines) compared to that computed with $\pi(\chi_{\rm p})$ conditioned on the q posterior (dashed lines), as a function of cutoff time $t_{\rm cut}$ for post-peak SNRs of $\rho_{\rm pp} = 100$ (red) and 25 (teal).

8.B Extracting the merger frequency

We define the merger frequency of a GW signal as the frequency at which the signal reaches its maximum amplitude. We extract it in the frequency domain; cf. Fig. 8.11 for the five different total mass cases presented in Sec. 8.4. To average out fluctuations in the signal (lighter lines), we apply a smoothing Savitzky-Golay filter [447] (yields the darker lines). We then find the region in frequency space over which the smoothed signal has a net-increasing amplitude spectral density (ASD), and calculate the frequency at which the ASD is the largest over that region. We identify these values as the merger frequencies and mark them with dashed vertical lines in Fig. 8.11. The figure only shows data in the LIGO Livingston detector; we extract consistent merger frequencies from LIGO Hanford and Virgo. While we opt for a data-driven approach, merger frequencies can also be calculated using the post-Newtonian formalism [148, 149] or numerical relativity surrogate dynamics [16].

8.C Measuring the mass parameters for systems with different M

Section 8.4 looks at signals with different total masses and thus different amounts of visible inspiral. In the main text we presented the evolution of the $\chi_{\rm p}$ posteriors as various data are excluded; see Fig. 8.4. Analogous plots

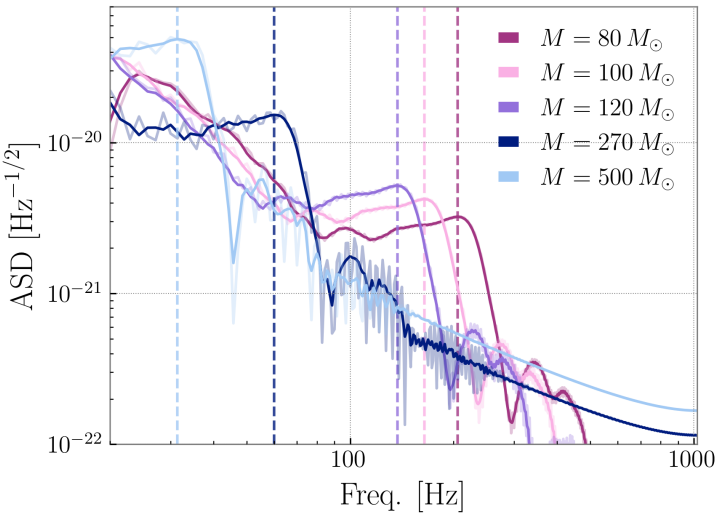


Figure 8.11: The ASD in LLO versus frequency for five signals with different total masses (different colors). The lighter shade is the true signal, while the overlaid darker shade is the signal with a smoothing filter applied. The dashed vertical lines indicate the merger frequencies, i.e., the frequencies at which each smoothed ASD has a local maximum corresponding to merger. These frequencies are listed in Table 8.2.

for the M and q posteriors are provided in Fig. 8.12. For the three lowest mass signals, M and q can be measured in the inspiral alone (pre-t_2 and earlier). For all masses, M and q can be measured in the ringdown alone (post-t_2), albeit weakly for the highest mass case. The total mass is always measured better for the lower-mass systems. On the other hand, the mass ratio is better measured in post-cutoff data for the higher-mass systems. Only the $M \leq 120\,M_\odot$ signals contain information about q before t_1 and t_2, while all signals have similar q information before t_4 and t_5. These mass parameters behave differently than $\chi_{\rm p}$, which hinges specifically on including the data between t_1 and t_2 (final pre-merger cycle).

8.D Impact of extrinsic angles: Equations and waveforms

Section 8.4 explores the impact of three extrinsic angles—inclination ι, polarization ψ, and phase φ—on GW morphology and resultant measurability of spin precession. Figure 8.6 illustrates how variations in each angle influence both the observed binary configuration on the plane of the sky and the corresponding GW signal. Here, we additionally provide analytic expressions that capture how each angle affects a GW signal. Changing the polarization angle by $\Delta\psi$ causes the two GW polarizations h_+ and $h_\times$ to mix, cf. Eq. (36) of

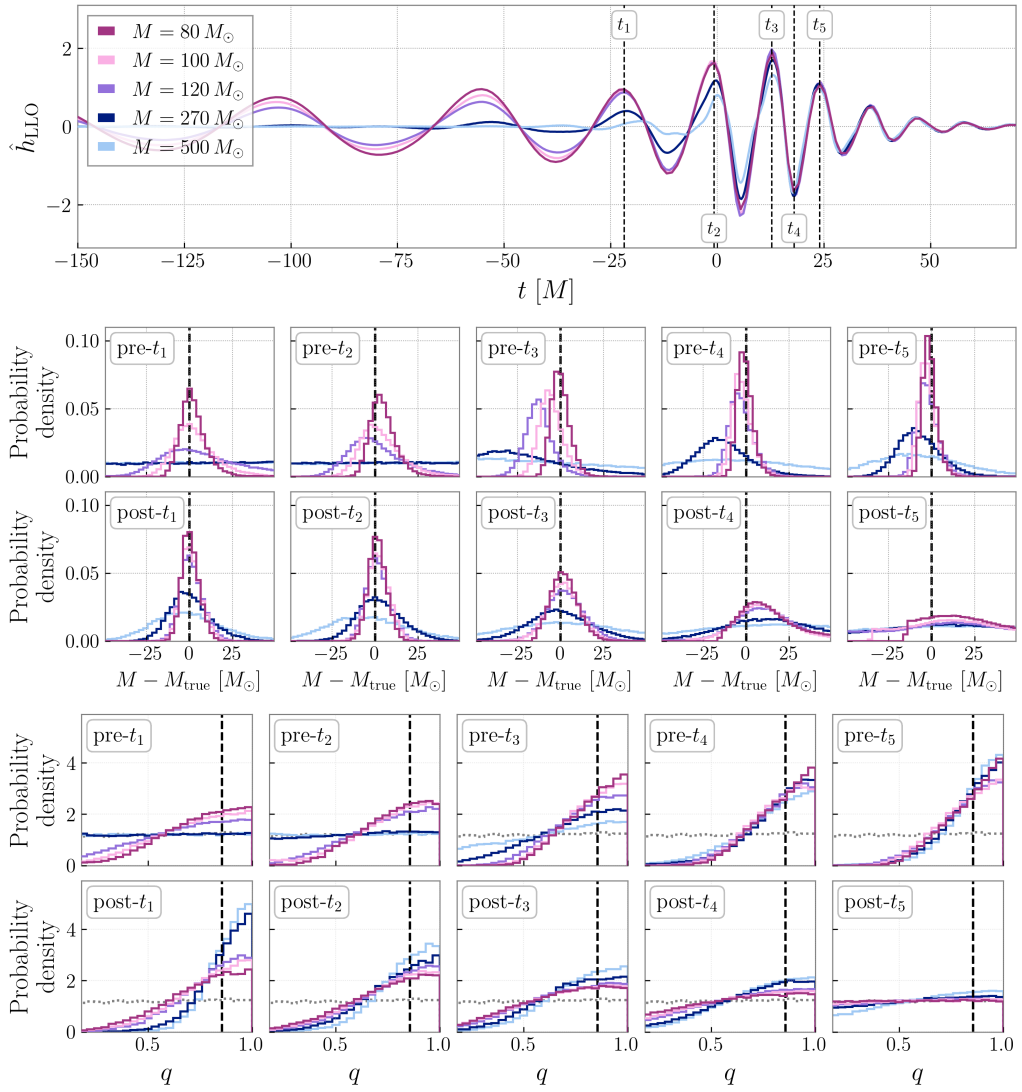


Figure 8.12: Evolution of the total mass M and mass ratio q as a function of cutoff time for signals with different total masses, presented in Sec. 8.4. Since all signals shown here have different total masses, we here plot posteriors on the inferred total mass minus the true total mass for a more direct comparison. The top panel is copied from Fig. 8.4 in the main text for reference. See the caption of Fig. 8.4 for more details.

Ref. [181],

$$\begin{aligned} h_+ \to h'_+ &= h_+ \cos 2\Delta\psi - h_\times \sin 2\Delta\psi \,, \\ h_\times \to h'_\times &= h_+ \sin 2\Delta\psi + h_\times \cos 2\Delta\psi \,. \end{aligned} \tag{8.8}$$

With some simple trigonometry, we can see that this transformation does not change the strain amplitude $h = \sqrt{h_+^2 + h_\times^2}$. In the top left panel of Fig. 8.13, we confirm this visually: the amplitude envelope for GW190521's max. $\mathcal{L}$ waveform (red) is conserved as ψ is varied. The amplitude envelope traces for all the other signals shown fall directly under that from the max. $\mathcal{L}$. For reference, these signals are plotted in LLO, with the amplitude envelopes scaled by the antenna pattern amplitude. This is not the case when ι or φ are varied, as shown via the amplitude envelope traces (pink) in the top right and bottom left panels of Fig. 8.13.

In the frame of the source [448], the two polarizations can be represented by a complex time series decomposed into the sum of spin -2 weighted spherical harmonics, dependent on the inclination angle ι and phase φ [148, 184–186]:

$$h_+ - ih_\times = \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} Y_{\ell m}^{(-2)}(\iota, \varphi) h_{\ell m}(\lambda, t) \,, \tag{8.9}$$

assuming we are infinitely away from the source. The modes $h_{\ell m}$ are each a function of time t and intrinsic parameters λ (masses and spins), and can be expressed in terms of their amplitude $A_{\ell m}$ and phase $\Phi_{\ell m}$ evolution

$$h_{\ell m}(\lambda, t) = A_{\ell m}(\lambda, t) e^{-i\Phi_{\ell m}(\lambda, t)} \,, \tag{8.10}$$

which are in turn approximated by post-Newtonian theory, cf. Eqs. (218, 327-329) of Ref. [148] or Eq. (4.17) of Ref. [188], only for the inspiral phase. The spherical harmonics are separable for ι and φ:

$$Y_{\ell m}^{(-2)}(\iota, \varphi) = \sqrt{\frac{2\ell + 1}{4\pi}} d_{(-2)}^{\ell m}(\iota) e^{im\varphi} \,, \tag{8.11}$$

where $d_{(-2)}^{\ell m}(\iota)$ are the spin -2 Wigner d functions [449]. The dominant $\ell = 2$, $m = \pm 2$ modes are [186]

$$\begin{aligned} Y_{2,2}^{(-2)}(\iota, \varphi) &= \sqrt{\frac{5}{4\pi}} \cos^4\left(\frac{\iota}{2}\right) e^{2i\varphi} \,, \\ Y_{2,-2}^{(-2)}(\iota, \varphi) &= \sqrt{\frac{5}{4\pi}} \sin^4\left(\frac{\iota}{2}\right) e^{-2i\varphi} \,, \end{aligned} \tag{8.12}$$

from which it can be shown that [181]

$$
\begin{aligned}
h_+ &= \sqrt{\frac{5}{4\pi}}\left(\frac{1+\cos^2\iota}{2}\right) A_{22}\,\cos\left(\Phi_{22}-2\varphi\right), \\
h_\times &= \sqrt{\frac{5}{4\pi}}\cos\iota\, A_{22}\,\sin\left(\Phi_{22}-2\varphi\right),
\end{aligned}
\tag{8.13}
$$

assuming planar symmetry (e.g. non-precessing systems). Written in this way, the ι and φ dependencies of the 22 mode are apparent [168]:

$$
h_+ \propto \frac{1+\cos^2\iota}{2}\,, \quad h_\times \propto \cos\iota\,, \tag{8.14}
$$

while φ is a simple phase shift. For non-precessing, equal-mass binaries, the 22 mode closely approximates the signal [185, 450]. The bottom right panel of Fig. 8.13 shows waveforms from such a system over a range of φ values, illustrating that φ just adds a phase shift to the signal while leaving the strain amplitude evolution unchanged. The non-spinning, equal-mass case underlies the general intuition of φ being the signal's "phase." However, in systems that are high-mass, precessing, and/or have unequal masses, the 22 mode alone is an insufficient approximation, and excluding higher-order modes leads to biased parameter estimates [314]. In these cases, interference between multiple ℓm modes makes the role of φ more complicated than a trivial phase shift, as shown in the bottom left panel of Fig. 8.13.

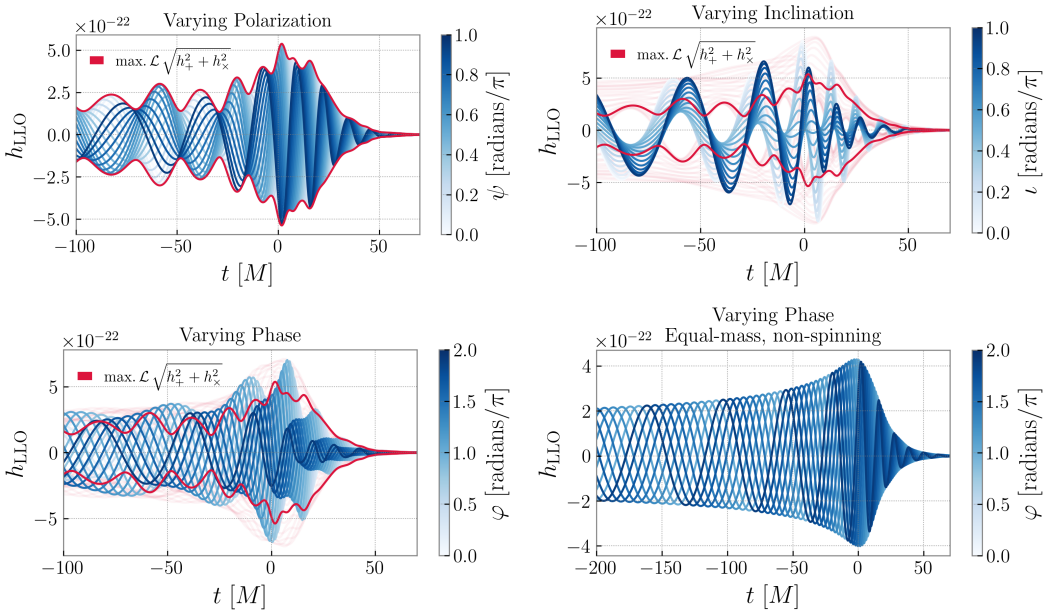


Figure 8.13: Effect of varying the polarization (top left), inclination (top right), and phase angle (bottom left) over a range of angles (blues, colorbar) from the max. $\mathcal{L}$ configuration on a projected waveform in LLO, at a fixed SNR. The max. $\mathcal{L}$ waveform amplitude envelope (red) is invariant under changes in polarization, but not inclination or phase angle. The pink traces show the amplitude envelope for each signal; for ψ they all fall directly under the red. For comparison, varying the phase angle of an equal-mass, non-spinning BBH is shown in the bottom right panel.

Chapter 9

EXPLORING ECCENTRIC AND INTERMEDIATE MASS RATIO INTERPRETATIONS OF GW190521 WITH TIME-DOMAIN INFERENCE

This chapter contains analyses completed in 2023-2024 that have not been otherwise published. The text was first written in 2024 and has been mildly adapted for this dissertation.

Abstract

The GW190521 signal was inferred to originate from a precessing equal-mass BBH; however, alternative explanations suggested it could be instead an unequal-mass or eccentric system. Here we explore the potential degeneracy between precession and eccentricity in LVK signals from heavy BBHs. We do so by applying time domain inference techniques to high-mass and eccentric waveforms consistent with GW190521. The time domain analysis allows us to better quantify the extent of this degeneracy and trace back to the signal morphology: precession, unequal masses, and eccentricity can all suppress the last cycle before merger. However, precession can achieve this to a greater degree than the alternative options leading to a better fit. We establish that the GW190521 data evolve over time in a manner most consistent with the precession hypothesis.

9.1 Introduction

The high-mass event GW190521 has been found to be consistent with competing source interpretations [e.g., 290, 292], implying that different physical effects can yield similar observational imprints when only a few cycles are observable. In this work, we ask: do these different source dynamics additionally imprint differently on each waveform *cycle*?

Our analysis centers the idea that even if multiple templates with different source dynamics are reasonable fits to the entirety of some data, they may not behave the same as a function of truncation time. We may gain information by comparing their respective evolutions to those of the real data—akin to a χ^2 signal consistency penalty used for data-quality assessment [e.g., 158, 451].

In Ref. [102] (Ch. 8) we explored simulated signals with the same *intrinsic* source properties as GW190521. We now use the same time-domain inference methods (see Ch. 6) but pivot to looking at injections with totally *different* sets of intrinsic parameters, corresponding to alternate interpretations of GW190521. We focus on comparing two alternative source scenarios discussed in the literature: spin precession vs. eccentricity (Section 9.2) and equal vs. unequal masses (Section 9.3). We find that the near-equal-mass, precessing injection behaves more similarly to the real GW190521 data than an unequal-mass precessing injection or an equal-mass eccentric injection.

9.2 Eccentricity e

A prominent, but contested, interpretation of GW190521 is that it is an eccentric merger: various analyses find evidence for [104, 290, 291] and against [385, 386] GW190521 having a non-circular orbit. It remains challenging to conclusively say one way or the other, because currently there exist no inspiral-merger-ringdown waveform models that include both spin precession *and* eccentricity. Without waveform models including both, it will be hard to fully distinguish between these effects, especially on short-duration merger-dominated signals [99] like GW190521. At the very least, precessing eccentric models will be able to confirm if both are comparably valid solutions for a signal, or if both are required in tandem.

Spin precession and eccentricity, while distinct physical effects, can both suggest similar BBH formation scenarios: they are primarily associated with BBH dynamical formation [1, 2, 452]. Both phenomena also modulate the signal amplitude on timescales between the orbital period and the GW inspiral, albeit with different morphologies and on different timescales [78, 363]. Spin precession modulates the amplitude of a BBH waveform as the BBH's orbital plane—and therefore direction of maximal GW emission—moves towards and away from the observer. Eccentricity, on the other hand, introduces modulation as the BHs accelerate more vs. less over the course of their orbit. In eccentric systems, the amplitude has burst-like features when the BBH passes through pericenter, i.e. when the BHs are closest and therefore accelerating most. While long inspiral observations would clearly distinguish these effects, they can become degenerate when only a few waveform cycles near merger are visible [99]. Both could, theoretically, cause the final pre-merger cycle to be dampened out with respect to what it would be for a circular non-precessing

source in a way that is difficult to distinguish without seeing a full precession or eccentric cycle.

GW190521's status as eccentric merger has both support and opposition in the literature. Ref. [290] and Ref. [291] find that GW190521's data prefers eccentricities at or near the upper boundary of the explored range. Ref. [290], however, uses the now-outdated waveform model SEOBNRE [453, 454], which is only for aligned spins and eccentricities less than $e = 0.2$ at 10 Hz. They find an eccentricity of $e \geq 0.11$ at 90% credibility; however, their posterior rails against the upper bound of the prior, suggesting that the true eccentricity may be higher. Ref. [291] instead use a coarse grid of numerical relativity simulations to estimate the eccentricity of GW190521. Among the ∼600 simulations studied, ∼100 of which include both spin precession *and* eccentricity, they find that GW190521 is most consistent with high eccentricity ($e = 0.69$) and spin precession ($\chi_{\rm p} = 0.66$). Ref. [293] analyze GW190521 under the assumption that it is generated by the merger of non-spinning black holes on *hyperbolic* orbits. They find that the signal is most consistent (more so than a precessing quasi-circular scenario) with two black holes undergoing two encounters and then merging. They use the TEOBRESUMS waveform model [455, 456] with only the $l = m = 2$ multipole, and assume the black holes are non-spinning.

Ref. [386] and Ref. [385], on the other hand, find that GW190521 is *not* consistent with being an eccentric signal. These two studies use the waveform models SEOBNRV4EHM [457, 458] and TEOBRESUMS [455, 456] respectively, both of which are more recent than SEOBNRE. To further complicate things for merger-dominated signals, all of the aforementioned EOB waveform models only model eccentricity in the *inspiral*; they stitch eccentric inspirals together with standard quasi-circular merger-ringdowns, as eccentricity is expected to largely decay away by merger for most astrophysically relevant configurations [363]. Clearly, the inference of eccentricity on GW190521 is *highly* dependent on the waveform model used, and therefore requires further investigation to mitigate waveform systematics and identify which features are actually driving these measurements.

In this work, we explore what happens when we recover an eccentric non-precessing injection with a precessing non-eccentric waveform model. Past work finds that eccentricity *can* be mistaken for spin precession [99, 104]; we wish to see if this claim holds true for a GW190521-like signal—not just

for the full data, but also as a function of time—and dig further into the data's morphology to see what waveform features are driving these claims. We utilize the in-development waveform model TEOBESUMSDALI [459], as it is valid for the largest range of parameters to date: for $q > 1/6$, aligned spins in the range $(-0.99, +0.6)$ and eccentricities up to $\sim$0.7. Specifically, we look at an injection generated with the maximum likelihood parameters found for GW190521 when it is analyzed using our TD inference code with TEOBESUMSDALI. This waveform has a detector-frame total mass of $M = 218\,M_\odot$, a mass ratio of $q = 0.4$, an effective spin of $\chi_{\rm eff} = -0.09$, and an eccentricity of $e = 0.47$ at 1.9 Hz. We again conduct parameter estimation with NRSUR7DQ4.

We use our pre-/post-cutoff analyses to track the behavior of the posteriors on the canonical max. $\mathcal{L}$ (labeled "precessing injection" for clarity) and eccentric injections and compare this to the real GW190521 data, with results shown in Figure 9.1. Looking at the waveforms in the top row, it is clear that the eccentric injection has a larger amplitude in the final pre-merger cycle (between t_3 and t_5). The posteriors for M, q, $\chi_{\rm eff}$, and $\chi_{\rm p}$ from analyzing the full precessing injection, eccentric injection, and real GW190521 data are shown in the top row of the grid of histograms. Posteriors for M, q, $\chi_{\rm eff}$, and $\chi_{\rm p}$ for five representative post-cutoff time analyses, labeled t_1 through t_5 for convenience, are given in each following row.

We find little confusion between eccentricity and precession. For the full-signal analysis, the recovered $\chi_{\rm p}$ for the eccentric injection is largely consistent with the prior, which is shown in gray dotted lines for comparison. Interestingly, the total mass and mass ratio of the eccentric injection are not found accurately by inference with the NR surrogate, but *are* both consistent with the real GW190521 data. As we are injecting and recovering with different waveforms that incorporate different physical effects, bias is not surprising but can tell us which parameters may be degenerate with eccentricity. The eccentric and precessing waveforms we examine are intentionally similar, in the sense that they are both tuned to fit the same data. Thus, if spin precession and eccentricity *are* degenerate in this region of parameter space (i.e. heavy BBHs), eccentricity *should* be mistaken for precession in this injection. The fact that they are not provides compelling evidence that spin precession and eccentricity can be distinguishable, even for merger-dominated signals.

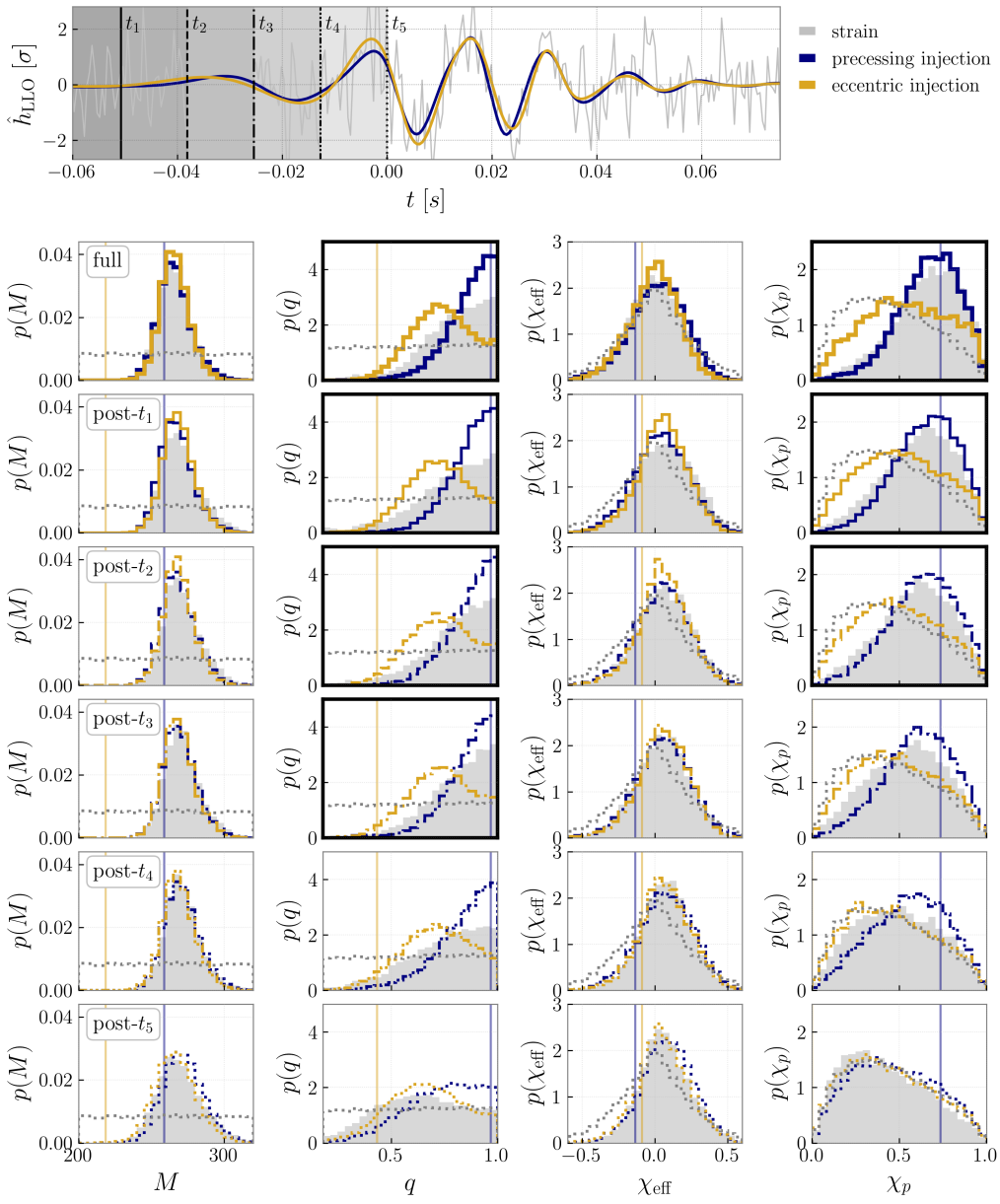


Figure 9.1: *Top panel:* Two signals that are consistent with GW190521 (whitened LLO data in gray): the precessing, quasi-circular injection (navy) and the eccentric, non-precessing injection (gold), each of which is the max. $\mathcal{L}$ parameters corresponding to a different waveform. *Bottom panel:* Comparison of the time evolution of the posteriors for total mass M (first column), mass ratio q (second column), effective spin χ_{eff} (third column), and effective precessing spin χ_{p} (fourth column) for injections corresponding to two interpretations of GW190521 (navy and gold, same as top panel) versus the real GW190521 data itself (gray). Each row corresponds to one cutoff time from the top panel; indicated with labels t_i and matching line-styles. Injected values are shown with vertical lines. The bold sets of axes highlight where the GW190521 posteriors are most consistent with the precessing injection, rather than the eccentric injection.

The behaviors of the eccentric and precessing injections also differ over cutoff time. As the M and $\chi_{\rm eff}$ posteriors for the two cases are both consistent with the real data at all cutoff times shown, we here focus on differences between q and $\chi_{\rm p}$. The eccentric injection has a smaller recovered mass ratio, and is typically less consistent with the real data; although, we note that in the post-t_5 analysis (i.e. analyzing just the merger-ringdown of the signals), the real data is actually more consistent with the eccentric injection than the precessing injection for q, albeit very mildly.

In Ref. [101], we linked the measurement of spin precession in GW190521 to the dampening out of the final pre-merger cycle with respect to the loud merger. We compared the inferred strain amplitudes at the peak of the final pre-merger cycle when that cycle is (precession inferred) versus is *not* (precession agnostic) included in parameter estimation; see Figure 2 therein. We found that the whitened strain amplitude in LLO, $\hat{h}$, at that point was anti-correlated with $\chi_{\rm p}$: larger inferred strain corresponds to lower $\chi_{\rm p}$. We repeat that analysis here for the precessing and eccentric injections, with results shown in the left-hand column of Figure 9.3. In the top left subplot, we plot the strain versus time in LLO for the full (darker, solid lines) and post-$t = 0$ (lighter, dashed lines) analyses conducted on the eccentric (gold) and precessing (blue) injections. In the bottom left, we plot the 50% and 90% credible intervals of the posterior on the strain amplitude at $t = -5$ ms versus $\chi_{\rm p}$; injected values are indicated, for comparison, with horizontal and vertical lines. The time $t = -5$ ms is a representative time near the peak of the final pre-merger cycle.

Zooming in, we observe that for the precessing injection, including (dark) versus excluding (light) the final pre-merger cycle in the analysis results in inferred strain amplitudes that are inconsistent at the 50% credible level. This inconsistency is not present for the eccentric injection. When the pre-$t = 0$ data is excluded, the $\chi_{\rm p}$ posteriors for both injections are consistent with the prior. However, when the full data *is* included, the analysis on the eccentric injection can detect that the strain amplitude of the final pre-merger cycle is larger than expected for high $\chi_{\rm p}$. The $\hat{h}_{\rm LLO}(-5\,{\rm ms})$-$\chi_{\rm p}$ posteriors for all these analyses align along the same correlation axis, but their exact positions *along* this axis depend on how much of the data is visible and the corresponding strain amplitude it reveals. Figure 9.3 illuminates that features in the later portion of the signal are different between the two injections; in one case, they

lead you to expect a higher amplitude at $t = -5$ ms than in the other. The degree of tension in the pre- vs. post-peak signal may be less in the eccentric case.

9.3 Mass ratio q

While all analyses that use NRSur7dq4 agree that GW190521 is most likely a precessing, near-equal-mass binary [20, 101, 107, 460], Ref. [292] find support for a precessing BBH with highly *unequal* masses ($q < 1/10$) using the waveform model IMRPhenomXPHM [245] and a prior uniform in source-frame total mass and inverse mass ratio. Although NRSur7dq4 is understood to be better tuned to numerical relativity at high masses and spins and reproduces this region of parameter space more faithfully [48], IMRPhenomXPHM is able to model more extreme mass ratios than NRSur7dq4. Ref. [292] show that a more extreme mass ratio ($q \sim 1/10$ to $q \sim 1/20$) yields a waveform with a final pre-merger cycle that is of lower amplitude than the more similar mass scenario; see their Figure 4. In Ref. [101], we also find that the data indeed suggest the presence of a suppressed premerger cycle, but the waveform achieves this by increasing precession, not unequal masses.

Both IMRPhenomXPHM *and* NRSur7dq4 find two solutions for GW190521: one with near-equal masses and one with highly unequal masses. The relative weight between these two modes in the posterior depends on the waveform and the prior. Other analyses [20, 101] also find the unequal-mass mode in the GW190521 data, although the mode is barely visible due to the mass ratio prior.

To explore the scenario of unequal masses, we take the full-data posterior from Ref. [101] and find the highest likelihood point with unequal masses ($q \approx 0.23$) and large in-plane spins ($\chi_{\rm p} \approx 0.85$). The injection with these parameters is shown in red in the top panel of Figure 9.2, labeled "unequal-mass injection." The max. $\mathcal{L}$ injection used throughout this work, here labeled "equal-mass injection," is shown in navy blue. The GW190521 data are shown in gray. All three signals are whitened with the reference power spectral densities and plotted in LIGO Livingston. Both simulated waveforms appear to be in phase with the data, but there are visible differences in their instantaneous amplitudes.

Analogously to Figure 9.1, we compare the behavior of the equal and unequal-mass injections to the real GW190521, with results shown in Figure 9.2. In

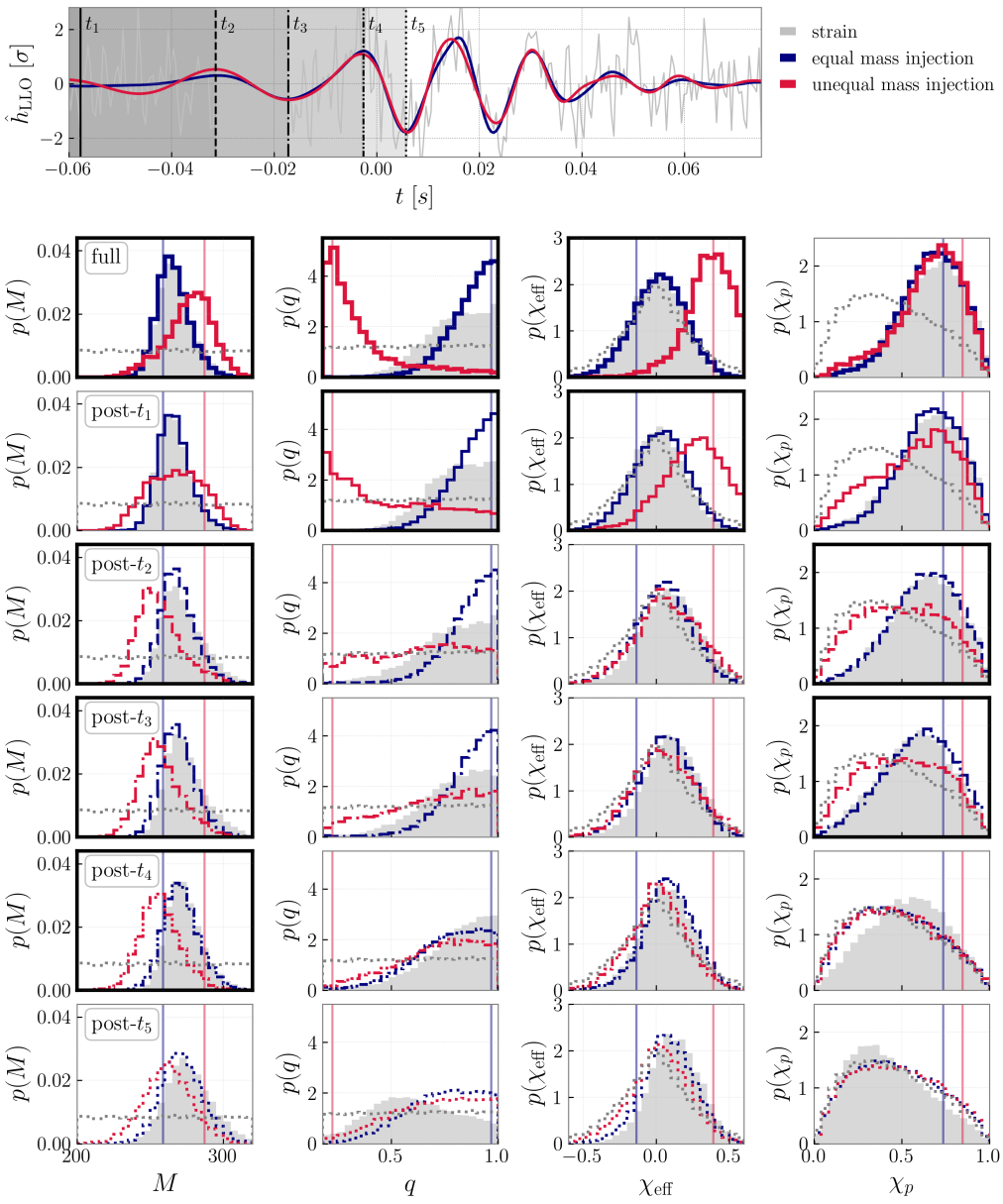


Figure 9.2: Same as Figure 9.1 but instead of comparing the max. $\mathcal{L}$ equal-mass injection (navy) to an eccentric injection, we here compare to an injection with unequal masses, $q \approx 0.23$ (red). The full posteriors from GW190521 (gray) and their time evolution are more consistent with those inferred from the equal-mass injection (navy) than the unequal-mass injection (red). This is most notable for the bold sets of axes. Injected values for the two injections are shown with vertical lines. Line styles for the histograms correspond to those of the vertical line denoting cutoff times in the top panel.

the top sub-plot, the unequal-mass injection is shown in red; our canonical max. $\mathcal{L}$ injection (here called "equal-mass injection" for clarity) is again shown in navy blue. Each subsequent row corresponds to one cutoff time, which map to peaks/troughs of the max. $\mathcal{L}$ waveform.

When looking at the full posteriors, the two injected signals have different sets of parameters. The posteriors for both signals peak at their injected values, shown by the vertical lines. The posteriors for GW190521 (gray histograms) much more closely resemble the equal-mass injection (navy histograms) than the unequal-mass injection (red histograms). Most striking are the differences in q and $\chi_{\rm eff}$. The unequal-mass injection peaks at small q while the equal-mass injection and the real data peak at $q \approx 1$. We recover a substantially more positive $\chi_{\rm eff}$ posterior with the unequal-mass injection than the equal-mass injection and the real data.

The inconsistency between the real data and the unequal-mass injection persists for segments before and after the various cutoff times. Subplots where the discrepancy is particularly prominent are highlighted by a thick black outline. When the majority of the visible inspiral is analyzed (top row; post-t_1), both injections and the real data are consistent with their respective full analyses. The early inspiral contributes little information, as this region of the data is barely visible in the detectors due to their insensitivity at low frequencies.

Continuing down the rows, we can see that the real data behaves more like the equal-mass injection than the unequal-mass injection as a function of cutoff time. The bold sets of axes in Figure 9.2 highlight where the discrepancies between the real data and the unequal-mass injection are most notable. When the early inspiral is excluded from the analysis (third row; post-t_2), substantial information about q and $\chi_{\rm p}$ is lost for the unequal-mass injection; both posteriors shift toward consistency with the prior. The M posterior for the unequal-mass injection also shifts to lower values. The equal-mass injection retains information in the post-t_2 analysis for all of the plotted parameters, as does the real data. As more and more inspiral is cut out, the post-cutoff time posteriors for both injections begin to converge, as information is lost. We verify that the higher level of similarity between GW190521 and the equal-mass injection also holds for the pre-cutoff time analyses.

We next examine the signal morphology for the waveforms recovered by NRSur7dq4 for the equal vs. unequal-mass injections. In the top right subplot

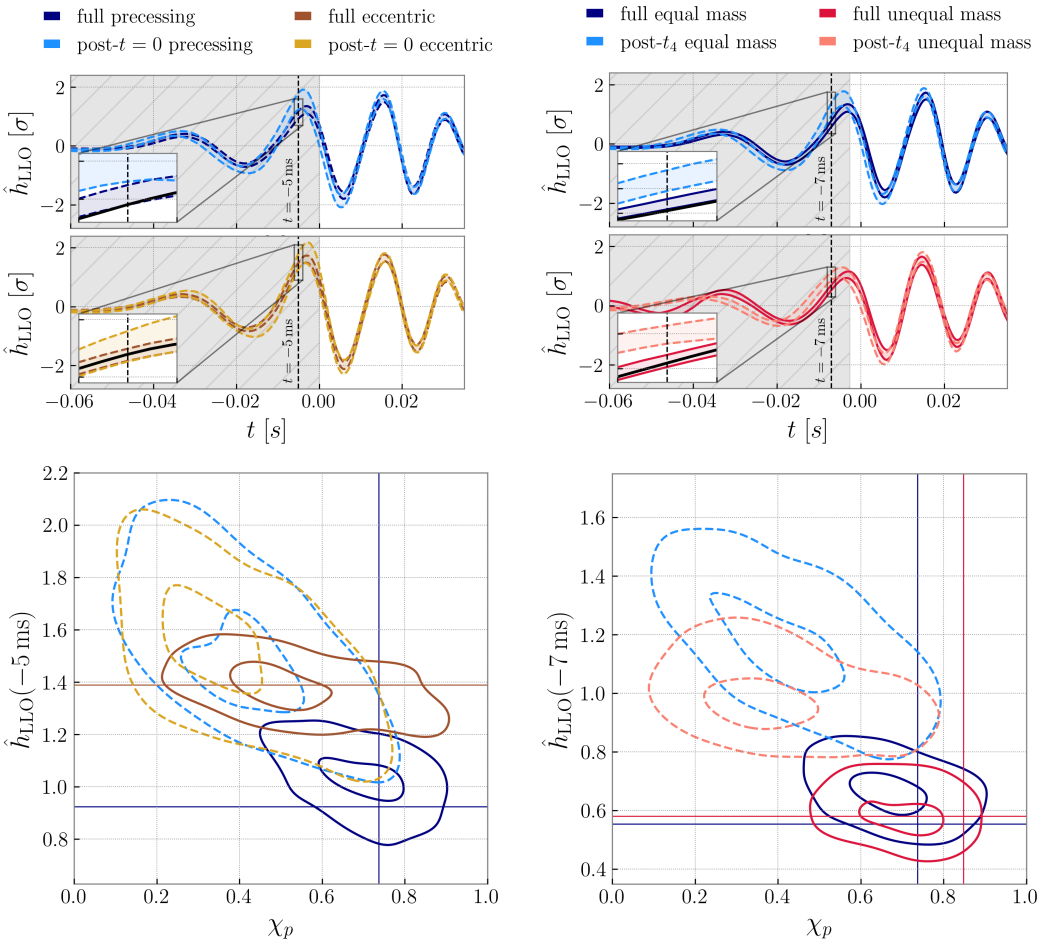


Figure 9.3: *(Top panels)* 50% credible intervals for the reconstructed whitened strain in LLO for various runs. The right column shows the injections with different mass ratios; the left column shows precessing vs. eccentric. Results from the equal-mass precessing injections are shown in blue; from the unequal-mass precessing injections in red; and the equal-mass eccentric non-precessing injections in yellow. Darker, solid lines correspond to the full signals, lighter, dashed lines show post-cutoff analyses. Data excluded in the post-cutoff analyses are shaded; the insets show that the reconstructed strain is inconsistent in the final pre-merger cycle when that data is versus is not included in inference. The black solid line in the inset is the injected waveform, for comparison. *(Bottom panels)* 50% and 90% credible intervals of the two-dimensional posteriors for $\chi_{\rm p}$ and whitened LLO strain $\hat{h}$ at $t = -7$ ms (right) and $t = -5$ ms (left), which are representative times at the peak of the final pre-merger cycle of the injected signals that are indicated with a vertical dashed line in the top panels. The horizontal and vertical lines indicate the injected values.

of Figure 9.3, we plot the strain versus time in LLO for the full (darker, solid lines) and post-$t = 0$ (lighter, dashed lines) analyses conducted on the unequal-mass (red) and equal-mass (blue) injections. The bottom right panel shows 50% and 90% credible intervals for the two-dimensional posterior on $\chi_{\rm p}$ and $\hat{h}$ in the final pre-merger cycle. Here, we measure $\hat{h}$ in LLO at a representative time of $t = -7$ ms, as indicated by the inset of waveform reconstruction plots. As indicated by the vertical dashed line in the top subplot, this time corresponds to near the peak of the final pre-merger cycle, and is excluded in the post-t_4 analysis (shaded region). We again find the anti-correlation between $\chi_{\rm p}$ and $\hat{h}_{\rm LLO}(-7\ {\rm ms})$, both between the two runs and *within* each posterior.

Interestingly, the $\hat{h}_{\rm LLO}$ values inferred for the unequal-mass injection (both full and post-t_4 runs) are slightly *smaller* than for the equal-mass injection, even though the *injected* strains are nearly equal at $t = -7$ ms, as indicated by the horizontal lines in the bottom right subplot of Figure 9.3. Furthermore, there is also dephasing between the recovered waveforms in the full versus post-t_4 analyses for the unequal-mass injection that is not present for the equal-mass injection, which we believe is driven by total mass M. Accordingly, we turn to other parameters to see what may be driving these differences.

In Figure 9.4, we plot the 50% and 90% credible intervals on the posteriors for the whitened strain amplitude in LLO at $t = -7$ ms, this time investigating correlations with the total mass M (top left panel) and mass ratio q (bottom right panel); the $M - q$ posteriors are in the bottom left panel. Correlations between $\hat{h}$ and the mass parameters are much more apparent for the unequal-mass injection (red) than the equal-mass injection (blue). Including versus excluding the final pre-merger cycle does not drastically alter the mass inference for the equal-mass case. For the unequal-mass case, on the other hand, excluding the final pre-merger cycle causes the mass posteriors to jump from low-q higher-M to the high-q lower-M region of parameter space occupied by the equal-mass injection.

Synthesizing these correlations, we conclude that, for the equal-mass injection, our interpretation of GW190521 holds: an observed suppression of the final pre-merger cycle with respect to the merger drives inference for $\chi_{\rm p}$ (and less-so M or q). Meanwhile, for the unequal-mass injection, the inference of M, q, and $\chi_{\rm p}$ are more intertwined in terms of what specifically is being informed by this feature in the data. In this case, the dampening is attributed to a confluence of

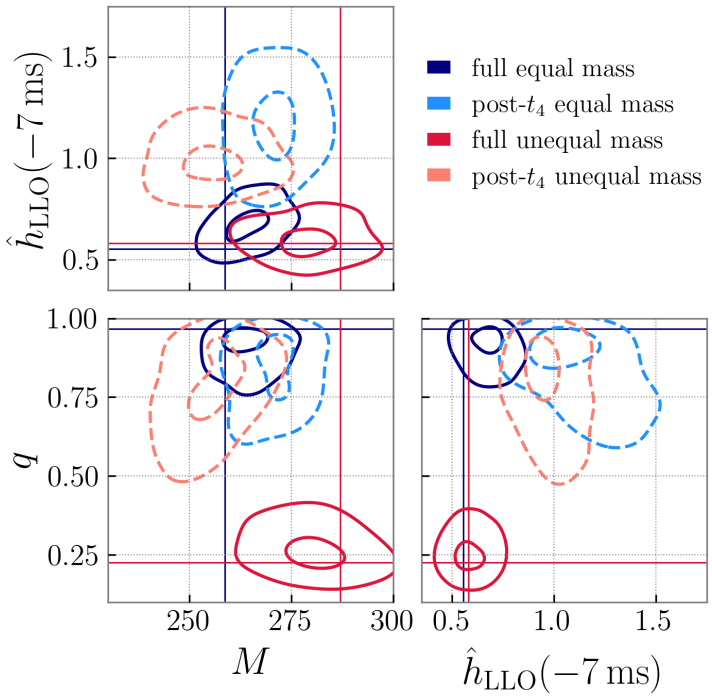


Figure 9.4: 50% and 90% credible intervals for the two-dimensional posteriors for total mass M, mass ratio q, and whitened LLO strain at $t = -7$ ms (see Figure 9.3) from the equal-mass (blue) and unequal-mass (red) injections. Results from analyzing the full signal are shown in the darker, solid line for each color; results from analyzing just after t_4, as marked in Figure 9.2, are in the lighter, dashed lines. Injected parameters are indicated by the horizontal and vertical lines.

larger total mass and more unequal masses, along with the system's large χ_{p}. Our interpretation is in agreement with Ref. [292]'s explanation that unequal masses can also cause a dampening of the final pre-merger cycle of GW190521.

9.4 Conclusion

We find that the behavior of GW190521 agrees most with the precessing, equal-mass (i.e., the event's max. $\mathcal{L}$) injection, rather than the unequal-mass precessing injection or the eccentric injection—all three of which are chosen to be highly consistent with the real event data. Notably, this consistency not only holds true for the full signal, but *also as a function of truncation time*, highlighting the critical advantage of using time-domain analysis: it enhances robustness against potential systematics by capturing how the signal's characteristics evolve cycle-by-cycle, rather than relying solely on its total information content. The three interpretations are not necessarily all mutu-

ally exclusive: all three could be (and in general, probably are) present at some level in the data; but at present we do not have an eccentric precessing waveform model to investigate that. Nonetheless, the fact that information is distributed throughout GW190521's waveform cycles in the same way as a precessing, quasi-circular injection bolsters the claim that GW190521 is more consistent with a near-equal-mass quasi-circular source than an eccentric or unequal-mass one.

Chapter 10

USING TIME-DOMAIN INFERENCE TO LINK THE ANTI-ALIGNED SPIN OF GW191109 WITH A GLITCH

An extended version of this chapter is published as Ref. [158]:

Rhiannon Udall, Sophie Hourihane, **Simona J. Miller**, Derek Davis, Katerina Chatziioannou, Maximiliano Isi, and Howard Deshong. Antialigned spin of GW191109: Glitch mitigation and its implications. *Phys. Rev. D*, 111(2):024046, 2 2025. ***SJM*** *conducted the time-domain inference analysis presented in Section 3b and wrote the associated text.*

The sections of Ref. [158] excluded from this thesis are those that do not reflect ***SJM****'s notable contribution to the work.*

10.1 Introduction

Reported in the third gravitational wave (GW) transient catalog (GWTC-3) [287], GW191109_010717 (more concisely GW191109) stands out among existing binary black hole (BBH) signals. With source-frame primary and secondary masses of $m_1 = 65^{+11}_{-11}\,M_\odot$ and $m_2 = 47^{+15}_{-13}\,M_\odot$ (90% symmetric credible intervals), it is among the most massive events. Furthermore, there is significant support for black hole (BH) spins anti-aligned with the orbital angular momentum: the mass-weighted effective spin [193, 289, 461] is $\chi_{\rm eff} = -0.29^{+0.42}_{-0.31}$. For these reasons, as well as support for unequal masses, $q = m_2/m_1 = 0.73^{+0.21}_{-0.24}$, spin-precession, and hints of eccentricity [390, 462], the binary is potentially of dynamical and/or hierarchical origin [62, 463] and impacts population inference [334, 464].

Multiple GW191109 properties hint toward a dynamical origin. High masses, above the pair-instability supernova (PISN) limit of $45 - 70\,M_\odot$ (depending on modeling assumptions) [131, 371], may require a hierarchical mechanism in order to form and merge. Asymmetric masses, in particular, might imply the merger of a second- and a first-generation BH [463]. Furthermore, population synthesis simulations of isolated formation scenarios typically find little support for spins anti-aligned with the orbital angular momentum, unless supernova kicks are exceptionally high [62, 84, 85]. Finally, eccentricity would

also be challenging to explain except by dynamical processes [389, 390, 465], due to the rapid orbit circularization by GW emission [363].

Given their astrophysical implications, the inferred properties of GW191109 are worth scrutinizing. The first potential source of systematics is the waveform used to model the signal. GWTC-3 employed the IMRPhenomXPHM [245] and SEOBNRv4PHM approximants [466], with inference performed by Bilby [243, 254] and RIFT [257] respectively. Both models include the physical effects of higher-order modes and spin-precession, and headline results (as quoted above) are their average. However, GW191109 is flagged for systematic differences between approximants [287], especially for the binary inclination (edge-on versus face-on/off respectively) and the longer $\chi_{\rm eff} > 0$ tail with IMRPhenomXPHM. A third waveform, NRSur7dq4 [16], was employed in Ref. [460]. A direct surrogate of numerical relativity simulations, NRSur7dq4 is expected to be the most accurate available model for systems with high masses and spins [16, 294, 460]. These results bolster the evidence for dynamical origin, with a more negative spin, $\chi_{\rm eff} = -0.38^{+0.21}_{-0.20}$, asymmetric masses, $q = 0.65^{+0.20}_{-0.19}$, and a precessing spin parameter [203] of $\chi_p = 0.59^{+0.26}_{-0.27}$. While waveform systematics remain relevant, the broad agreement between three waveforms (including a direct surrogate to numerical relativity) that $\chi_{\rm eff} \lesssim 0$ to varying credibility, suggests that subsequent interpretations of its formation history remain valid.

A second potential source of systematics concerns modeling the detector noise. Around GW191109's arrival, both LIGO [4] detectors experienced a terrestrial noise transient known as a scattered light glitch [146, 287, 467]. The Virgo detector [5] was offline at this time, and so only the LIGO detectors contributed to the observation. In LIGO Hanford (LHO), the glitch power was at a nadir while the event was in the detection band, making its impact on the inferred parameters negligible, see App. A of Ref. [158]. As such, we ignore the LHO glitch going forward. By contrast, glitch power in the Livingston detector (LLO) was directly coincident in time and frequency with the signal, a circumstance which could bias astrophysical inference [159, 468–471]. Specifically, glitch power extends up to $\sim 40\,\mathrm{Hz}$, coincident with the signal, see Fig. 10.1. Spin parameters might be particularly susceptible to such data quality issues due to the relatively smaller imprint they leave on signals compared to, e.g., the BH masses. For example, GW200129 shows evidence of

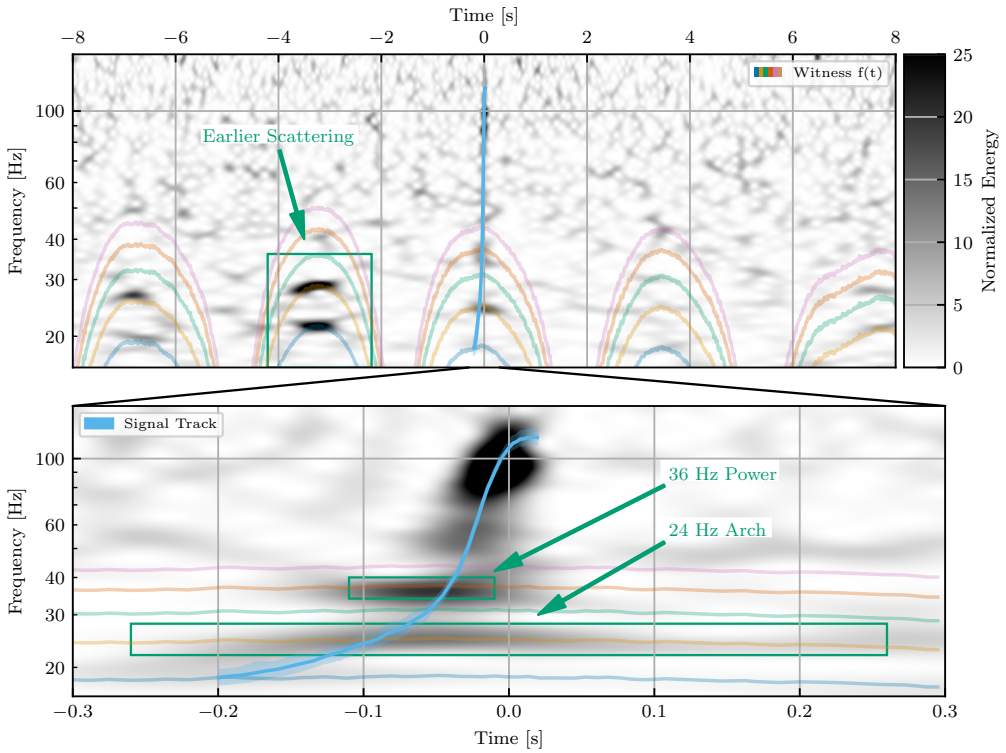


Figure 10.1: Spectrograms of the original (before glitch mitigation) data in LLO centered around the time of GW191109. The top panel shows $\pm 8\,\mathrm{s}$ of data, while the bottom panel zooms in around the event. Onto this, we plot the time-frequency tracks of the scattered light glitch, as predicted by the motion observed in the witness channel `L1:SUS-ETMX_L2_WIT_L_DQ`. This is the witness to the penultimate stage of the reaction chain pendulum for the X-arm end test mass. The scattering surface is the final stage of the reaction chain, and so this witness does not perfectly capture the motion of the scattering surface; to compensate, we apply a static coefficient of 1.38 to the predicted frequency, such that it is calibrated to the prominent scattering arches $\sim 3\,\mathrm{s}$ before the event. We also plot the inferred signal from a NRSur7dq4 analysis of full-bandwidth data after glitch subtraction. We annotate three regions of interest: the prominent scattering before the event (top panel), the long-duration excess power at 24 Hz (bottom panel), and the short-duration excess power at 36 Hz (bottom panel). Both the 24 Hz and the 36 Hz excess power coincide with expected glitch arches, however only the former has an arch-like shape.

spin-precession [287, 294], but its significance depends on how the glitch that overlapped that signal is modeled [159, 295].

The headline GWTC-3 results were obtained after an estimate for the glitch had been subtracted from the data. The two-step process involved first mod-

eling the signal and the glitch with a flexible sum of coherent and incoherent wavelets respectively with BayesWave [472–474]. Second, a fair draw from the glitch posterior was subtracted and the system parameters were inferred as quoted above. This procedure has been shown to generally lead to unbiased mass and (aligned) spin inference [468, 470]. However, uncertainties remain related to BayesWave's glitch model and in the fair draw chosen to be subtracted. These effects were investigated in Ref. [146], albeit with a simpler waveform model with single-spin precession and no higher-order modes, IMRPhenomPv2 [316]. Glitch mitigation was found to affect the $\chi_{\rm eff}$ inference by a similar amount as waveform systematics. Completely removing the glitch-affected data, i.e. all LLO data below 40 Hz, instead resulted in a dramatic shift of $\chi_{\rm eff}$ to positive values $\chi_{\rm eff} = 0.27^{+0.24}_{-0.48}$.

The stark impact of glitch-affected data on astrophysically-impactful spin inference motivates our study. We extend Ref. [146] to explore the manner in which the data inform the system parameters. Using NRSur7dq4 and a frequency-domain analysis, we find that the LLO data between 30 and 40 Hz are crucial for spin inference: excluding $30 - 40$ Hz data shifts the probability of $\chi_{\rm eff} < 0$ from 99.4% to 32.2%, effectively wiping out any preference for for anti-aligned spins. A similar time-domain analysis [101] highlights the role of the data $0.1 - 0.04$ s prior to merger. These data, which inform the $\chi_{\rm eff} < 0$ measurement, coincide in time and frequency with excess power in LLO, see Fig. 10.2 and in particular the excess power at ~ 36 Hz. To check whether such dramatic shifts in support for $\chi_{\rm eff} < 0$ are possible from Gaussian noise alone, we analyze 100 simulated signals consistent with GW191109. We find that shifts of this magnitude are unlikely but not impossible as 6% of the simulations experience a larger shift than GW191109.

10.2 Modeling the Compact Binary Signal

We use both time- and frequency-domain techniques to model the signal with either waveform approximants for compact binary signals or, more generically, with sine-Gaussian wavelets. All analyses consider data surrounding the nominal trigger time of GW191109, GPS time 1257296855.22, and employ a sampling rate of 1024 Hz, with the maximum analysis frequency set to 7/8 of the Nyquist frequency. Unless otherwise noted, analyses that model only the compact binary (and not the glitch) use a minimum frequency of 20 Hz in both detectors. We use standard compact-binary priors [254], notably uniform in

detector-frame component masses and spin magnitude and orientation.

Frequency-domain inference

Frequency domain analyses with waveform approximants are based on BILBY [243, 254] with its implementation of the DYNESTY sampler [253] and BAYESWAVE [469], both analyzing 4 s of data. The former models the signal with NRSUR7DQ4 [16] and the latter with IMRPHENOMXPHM [245] (though for consistency we also perform checks with the former using IMRPHENOMXPHM). NRSUR7DQ4 supports a minimum mass ratio of 0.25 and minimum detector-frame chirp mass of $35\,M_\odot$; neither restriction affects the analysis. We extend into the extrapolation region in spins, setting a maximum spin magnitude of 0.99. For comparison, we also perform analyses with BAYESWAVE where the signal is modeled as a flexible sum of coherent sine-Gaussian wavelets [472, 474].

Time-domain inference

While GW inference is typically conducted in the frequency domain for computational efficiency, it can equivalently be conducted in the time domain [222, 367, 395, 396]. Frequency domain analyses are non-local in time; to avoid non-trivial likelihood modifications [215], time-domain inference is necessary in order to isolate purely temporal features of the data. Below, we truncate the GW191109 data at different times around the 36 Hz excess power, and independently conduct inference on the pre- or post-cutoff-time data. For this, we use the time-domain inference code employed in Ref. [101] to study the GW190521 properties and which was based on time-domain implementations targeting post-merger data [222, 395]. All time-domain results are based on regions of 1 s of data around GW191109's trigger time and employ NRSUR7DQ4 [16]. The PSDs used in the time domain analyses are the same as those in the frequency domain, i.e., from the GWTC-3 data release [406].

10.3 Understanding the GW191109 Inference

In this section, we explore the relation between the GW191109 inference, especially the $\chi_{\rm eff} < 0$ measurement, and the glitch-affected data. In Fig. 10.1 we show spectrograms of the original data (without any glitch mitigation) in LLO at the time of the event. Arch-like traces (multiple colors) show the glitch time-frequency tracks as predicted by a witness channel. The light blue track corresponds to GW191109 as inferred with NRSUR7DQ4 from data after

the glitch was subtracted. The upper panel presents 16 s of data; scattering arches are visible leading up to the event. In the bottom panel, we focus on the vicinity of the signal and highlight the intersection of the signal track with visible excess power along the projected scattering tracks. The first is at $\sim$24 Hz and has the expected duration and morphology of a scattering arch. The second is at $\sim$36 Hz and while it coincides with the glitch track predicted by the witness, the excess power duration is short and does not match the expected behavior of slow scattering. As noted in Ref. [146], this 36 Hz excess power is not included in the original BayesWave glitch reconstruction and thus not subtracted in the GWTC-3 data.

We begin by confirming and extending the results of Ref. [146] with NRSur7dq4. Analyzing data from each detector separately we confirm that the measurement is driven solely by LLO, which prefers $\chi_{\rm eff} < 0$ at 99.6%, compared to 20.0% in LHO. Coherent analysis of both detectors tends to the LLO conclusion due to LLO's higher sensitivity in the relevant frequency range, shown below to be 20$-$40 Hz. Indeed, the maximum likelihood waveform from the coherent analysis accumulates 20% (8%) of its SNR squared in LLO (LHO) for frequencies below 40 Hz. This estimate further suggests that LHO data cannot aid in determining whether the critical $\sim$36 Hz excess power is part of the signal or the glitch.

Similar differences in parameter inference per detector are present for other parameters as well, notably the detector-frame total mass M and luminosity distance D_L; see footnote 1 for a discussion of the correlation between $\chi_{\rm eff}$ and D_L. For example, in individual detector analyses the detector-frame total mass is $M = 133^{+14}_{-14}\,{\rm M}_\odot$ ($M = 162^{+21}_{-20}\,{\rm M}_\odot$) in LLO (LHO), while the luminosity distance is $D_L = 1630^{+1360}_{-850}$ Mpc ($D_L = 2760^{+2300}_{-1570}$ Mpc) in LLO (LHO). The corresponding source-frame total mass remains the same as the increases in detector-frame mass and distance effectively "cancel out." Though different, these estimates are still consistent with each other within statistical uncertainties so there is no indication of a discrepancy across detectors as was the case for GW200129 [159]. Moreover, these differences do not lead to diverging astrophysical interpretations like the $\chi_{\rm eff}$ inference; we therefore focus on the latter in what follows.

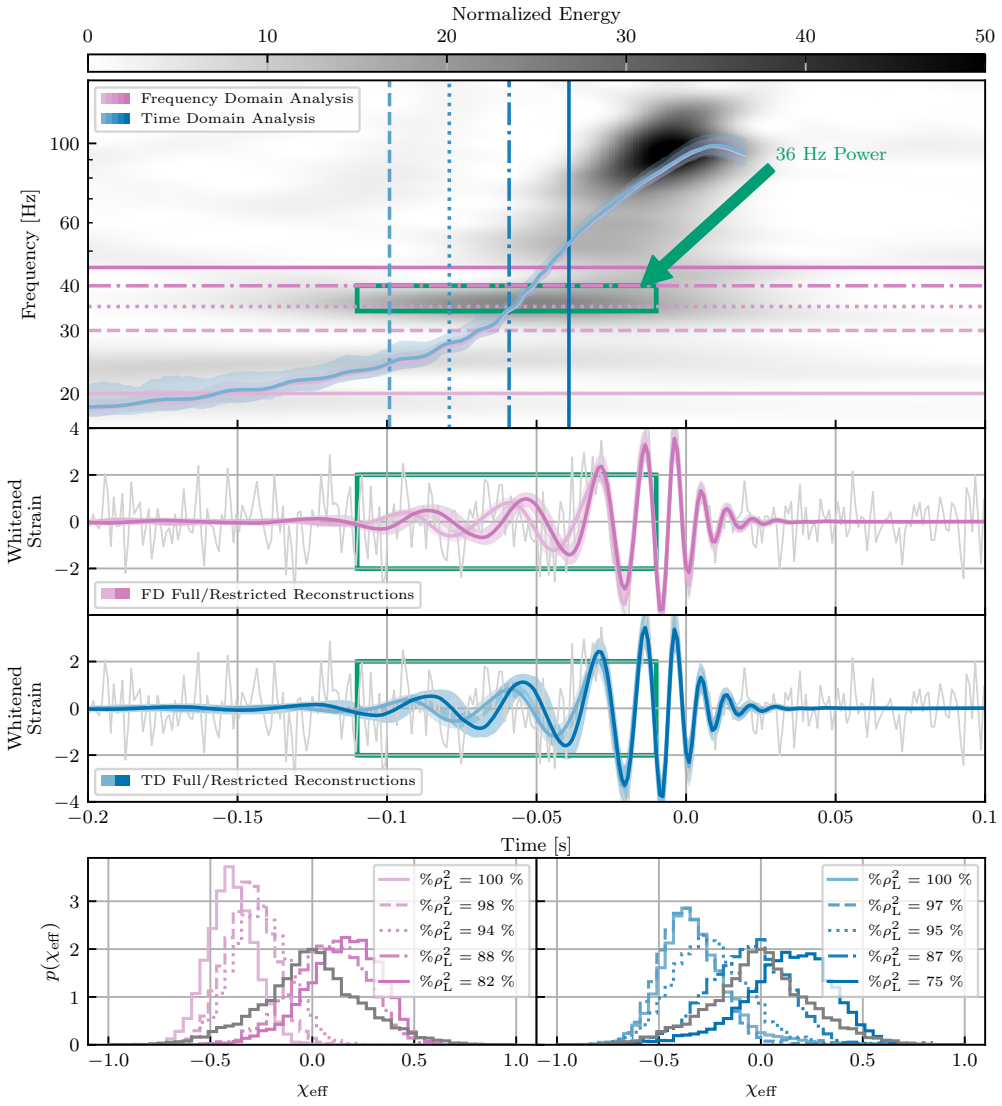


Figure 10.2: Tracing the $\chi_{\rm eff}$ inference across frequencies and times. The top panel shows the spectrogram of the glitch-subtracted data around GW191109, with residual excess power at 36 Hz highlighted along with the signal track. We progressively remove data in the frequency domain (pink) and the time domain (blue) and reanalyze the restricted data. Vertical and horizontal lines in the top panel denote the time and frequency cuts, respectively; only data to the left or above these lines are analyzed. The two middle panels show the whitened time-domain data (grey) and signal reconstruction (pink and blue). Lighter colors correspond to the analyses of the full data, while darker colors correspond to the most restricted data (frequencies above 40 Hz and times from −0.04 s before merger onwards). The bottom row shows the $\chi_{\rm eff}$ prior (gray) and $\chi_{\rm eff}$ marginal posteriors from analyses with varying levels of data restriction, each corresponding to the lines on the top panel. The legend notes the SNR squared ρ^2 fraction in Livingston that remains in the analysis band after each data restriction.

Tracing inference across frequencies

To more precisely track the origin of the $\chi_{\rm eff} < 0$ measurement across LLO data, we perform a series of coherent 2-detector analyses where we successively restrict the LLO frequencies, incrementing the minimum frequency f_L by 5 Hz from 20 to 45 Hz. We use the glitch-subtracted data where the 24 Hz arch from Fig. 10.1 has been subtracted, but the 36 Hz excess power has not [287]. A subset of these results are shown in Fig. 10.2 (pink shading). The top panel shows a spectrogram of the glitch-subtracted data; compared to Fig. 10.1, there is no excess noise at ∼24 Hz.

Marginalized posteriors for $\chi_{\rm eff}$ are shown in the bottom panel. The legend denotes the percentage of the total SNR squared ρ^2 (computed based on the maximum-likelihood full-band signal) that remains in the analysis window after each restriction. Removing data between 20−30 Hz (solid vs dashed horizontal lines in the top panel and histograms in the bottom panel) or 30−35 Hz (dashed vs dotted) removes 6% of ρ^2 but does not dramatically alter inference: $\chi_{\rm eff} < 0$ is still preferred at 96.3% for $f_L = 35$ Hz. Such small shifts are likely consistent with the SNR reduction and regression to the prior (gray). Removing data $35 - 40$ Hz (dotted vs dot-dashed) removes an additional 6% of ρ^2 and instead results in an abrupt shift in $\chi_{\rm eff}$, with $\chi_{\rm eff} < 0$ now only at 32.2%, a moderate preference for positive values.[1] Further bandwidth reduction does not modify the $\chi_{\rm eff}$ posterior substantially (solid dark pink). These results indicate that it is the LLO data between 35 and 40 Hz that are crucial for measuring $\chi_{\rm eff}$, coinciding with the 36 Hz excess power visible in both the original data, Fig. 10.1, and the glitch subtracted data, Fig. 10.2.

The second panel from the top of Fig. 10.2 shows the whitened time-domain reconstructions. We compare signal reconstructions from two analyses with dramatically different $\chi_{\rm eff}$ posteriors: the full bandwidth analysis that prefers $\chi_{\rm eff} < 0$ against the $f_L = 40$ Hz analysis with a mildly positive $\chi_{\rm eff}$. While the two analyses are conducted on different data subsets, we can still evaluate the waveforms across the same times and plot them together. The reconstructions are consistent during the merger (corresponding to high frequencies included

[1] Since the $\chi_{\rm eff}$ prior is centered at zero, this shift to mildly positive values goes beyond regression to the prior. We attribute this to a mild $\chi_{\rm eff} - D_L$ degeneracy that arises for merger-only signals. The uniform-in-volume prior favors larger D_L and results in larger $\chi_{\rm eff}$ to compensate for the amplitude reduction. This degeneracy is less pronounced when the signal inspiral is visible, as then $\chi_{\rm eff}$ is constrained by the inspiral phase evolution beyond just the merger amplitude.

in both analyses), but start diverging $2-3$ cycles before merger. By eye, the full-band reconstruction better matches the data for $t \approx -0.6\,\mathrm{s}$, corresponding to the 36 Hz excess power. When that power is included in the analysis, the signal model absorbs it by setting $\chi_{\mathrm{eff}} < 0$ and pushing the GW cycle to earlier times. If that power is not part of the analysis, χ_{eff} is no longer required to be negative and the 36 Hz excess power is left unaccounted for.

This conclusion raises the question of whether the 36 Hz excess power is part of the signal or part of a glitch that remained unsubtracted. Though the shift in the χ_{eff} posterior is suggestive of anomalous noise, it is possible that it is at least partly due to loss of information as 6% of ρ^2 in LLO is contained in the $35-40\,\mathrm{Hz}$ frequency band. In Sec. 10.3 we contextualize this χ_{eff} shift with simulated signals.

Tracing inference across times

Having identified the crucial frequencies for χ_{eff} inference, here we do the same across time with the time-domain analysis described in Sec. 10.2. When used on the full dataset, frequency- and time-domain analyses should yield equivalent results. Indeed, we find consistent posteriors for χ_{eff} when analyzing GW191109 in the frequency and time domains, as seen by the solid histograms in Fig. 10.2.[2]

However, as Fig. 10.1 shows, there is no 1-to-1 mapping between time and frequency for the glitch. Though not as apparent, the same is true for the signal beyond the inspiral regime or due to spin-precession and higher-order modes. Truncating the data in the time domain is, therefore, not equivalent to truncating in the frequency domain, as the former allows us to probe the effect of individual cycles (or parts of cycles) of the signal or the glitch.

Results from progressively excluding the earlier portion of the signal in the time-domain are shown in Fig. 10.2 (blue shading). We find broadly similar

[2]The time- and frequency-domain analyses employ different priors on masses, luminosity distance, and time. The time-domain inference uses priors which are uniform in detector-frame total mass, mass ratio, and luminosity distance; and are normally distributed in geocenter time, centered at 1257296855.2114642 with a width of 0.005 seconds. We confirm that the differences in time and the mass priors effect the posteriors minimally. Reweighting between the two luminosity distance priors proves difficult due to finite sampling and upweighting portions of parameter space with no support in the posterior. However, the luminosity distance posteriors from the time- and frequency-domain analyses are in high agreement when the full data is analyzed, despite using different priors.

results as the frequency-domain analysis: the full data yield preference for $\chi_{\rm eff} < 0$. As the segment that contains the 36 Hz excess power is progressively removed (blue vertical lines in the top panel), the $\chi_{\rm eff}$ posterior shifts to being principally positive (equivalent blue histograms in the bottom panel). Overall, the data 0.1−0.04 s before merger are crucial for $\chi_{\rm eff} < 0$ inference. Compared to the frequency-domain results, the shift in the $\chi_{\rm eff}$ posterior is more gradual, likely due to the fact that the 36 Hz power is more concentrated in frequency, hence no time "cut" abruptly completely excludes it. Waveform reconstructions from the time-domain analysis (third panel from the top in Fig. 10.2) yield consistent conclusions.

Simulated signals

We investigate the degree to which the abrupt shift in the $\chi_{\rm eff}$ posterior in Fig. 10.2 is consistent with SNR loss from removing data with simulated signals. We simulate 100 signals drawn from the GW191109 full-band posterior, add them to Gaussian noise drawn from the GW191109 PSDs in LLO and LHO, and analyze the full data versus the > 40 Hz data in LLO independently. Signals have true values $\chi_{\rm eff} < 0$ but as data and signal SNR are removed when $f_L = 40$ Hz, we expect the posterior to become more prior-like and shift toward $\chi_{\rm eff} = 0$. For each simulated signal, Fig. 10.3 shows the probability of $\chi_{\rm eff} \leq 0$ from the full-data, $f_L = 20$ Hz, and the restricted-data, $f_L = 40$ Hz, analysis.

For almost all signals removing low-frequency LLO data results in a $\chi_{\rm eff}$ posterior that shifts closer to the prior and positive values (lying below the diagonal) as expected for signals with true values of $\chi_{\rm eff} < 0$. In most cases, this shift is marginal, and posteriors stay majority-negative, as evidenced by the high density (64% of all signals) in the top right quadrant (green axes). The next most likely outcome is the bottom right quadrant (orange axes), which contains 34% of the signals, including GW191109: here the $\chi_{\rm eff}$ posterior shifts from favoring negative to positive values. Among these, GW191109 is one of the more extreme cases, exhibiting a shift more significant than 94% of the simulations. Therefore, we conclude that the $\chi_{\rm eff}$ shift presented in Fig. 10.2 is *unlikely*, but not impossible, to be explained by a random Gaussian noise instantiation, i.e. without needing to invoke residual glitch power. In App. E of Ref. [158], we present further results based on a χ^2 test used in search algorithms that tracks how SNR is accumulated along the signal [414–416]. Consistent with

Fig. 10.3, the test is inconclusive: the full-band analysis has behavior more extreme than most simulations, but it is not strongly inconsistent with them.

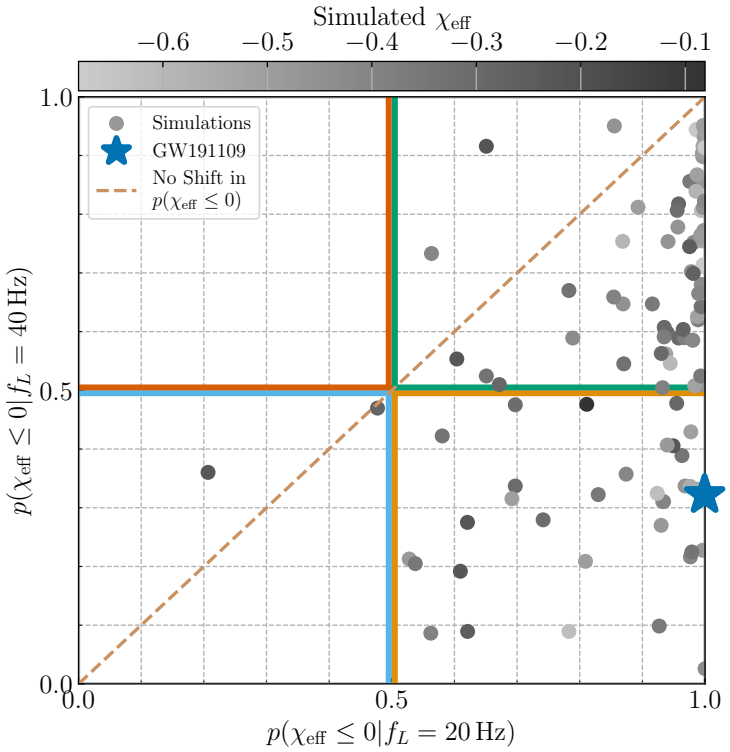


Figure 10.3: Shifts in the probability of $\chi_{\rm eff} \leq 0$ for 100 simulated signals consistent with GW191109 in Gaussian noise (dots) and the real signal (cross). The x-axis corresponds to a full-band analysis, while the y-axis corresponds to a restricted-band analysis with $f_L = 40\,{\rm Hz}$. Going clockwise, the top left quadrant (red-orange axes) would contain cases where the posterior shifted from majority positive to negative (of which there were none), the top right quadrant (green axes) contains cases which were majority negative in both full- and restricted-band analyses, the bottom right quadrant (orange axes) contains cases which started majority negative and became majority positive (including GW191109), and the bottom left quadrant (blue axes) contains cases which were consistently majority positive. The $x = y$ line (dashed brown) corresponds to no shift in the probability for $\chi_{\rm eff} \leq 0$.

Chapter 11

TESTING HAWKING'S AREA LAW WITH GW250114

*This chapter contains **SJM**'s contributions to Ref. [18]:*

> LIGO-Virgo-KAGRA Scientific Collaboration incl. **Simona J. Miller**. GW250114: Testing Hawking's Area Law and the Kerr Nature of Black Holes. *Phys. Rev. Lett.*, 135(11):111403, 11 2025. ***SJM** was one of two key analysts who generated the data leading to the first-ever high-confidence observational confirmation of Hawking's Area Law.*

Some text in this chapter is written novelly for this thesis, while some is adapted from or published verbatim in Refs. [18] or [475].

Summary of Contribution

My time-domain inference studies, described in detail in the preceding chapters of Part II, gave me the unique ability to perform a high-profile test of fundamental physics: in a recent LVK Collaboration paper about the loudest binary black hole to date, GW250114 [18], I **was a key analyst in the first-ever observational confirmation of Hawking's Area Law** which states that the black hole horizon area—a simple function of mass and spin—cannot decrease over time [476]. The open-source computational tools I developed (`tdinf` [161]) allow gravitational-wave data to probe not just their own astrophysical sources, but also to test the fundamental laws that govern our physical universe.

In this chapter, I describe the Area Law [Sec. 11.1] and how it can be tested with gravitational waves [Sec. 11.2]. I then introduce the gravitational-wave event GW250114 [Sec. 11.3] and present our results from using it to probe for Area Law non-violation [Sec. 11.4], including a discussion of how the significance of such non-violation is quantified. Technical analysis details are provided in Appendix 11.A.

11.1 The Area Law

Also known as the second law of black hole mechanics, Hawking's Area Law [476] states that **the black hole horizon area $\mathcal{A}$ cannot decrease over time**.[1] The law relies on three prerequisites: first, that the null-energy condition is satisfied;[2] second, that weak cosmic censorship holds;[3] and third, that general relativity is true. When dealing with the merging black holes observed by LIGO [4], Virgo [5], and KAGRA [6], we can reasonably assume that all three conditions are met [e.g., 19, 24].

In our early days of education, we (hopefully) learn that the surface area A of a sphere with radius R is $A = 4\pi R^2$. While a black hole does not have a radius in the classical sense, it does have an *event horizon* defining a surface at some coordinate-dependent "distance" from the singularity at its center. In Kerr spacetime [46], a black hole with mass m and spin $\chi \in [0, 1]$ has a "radius" $\mathcal{R}$ governed by the relationship

$$\mathcal{R}^2 = 2m^2 \left(1 + \sqrt{1 - \chi^2}\right) , \tag{11.1}$$

in units where $G = c = 1$. The so-called *horizon area* $\mathcal{A}$ of a Kerr black hole is then [485]

$$\mathcal{A}(m, \chi) \equiv 4\pi\, \mathcal{R}^2 = 8\pi\, m^2 \left(1 + \sqrt{1 - \chi^2}\right) . \tag{11.2}$$

As shown in Fig. 11.1, increasing χ from its minimal to maximal value (at a fixed m) decreases a black hole's "radius" by $\sim 29\%$, corresponding to a 50% decrease in the horizon area.

11.2 Testing the Area Law with a Merging Binary Black Hole

When applied to a binary black hole system, the Area Law states that the sum of the areas of the two initial, progenitor black holes (subscript i) must

[1] The law was proven by Hawking [476], but previously explored as the so-called "irreducible mass" by Christodoulou [477] and Christodoulou & Ruffini [478], as well as stated by Penrose and Floyd [479]—all in the years 1970-1971.

[2] Simply put, an "energy condition" is a generalization of the notion that "energy should be positive" to classical field theories of gravity like general relativity [480]. The null-energy condition, in particular, demands that the contraction of the Ricci tensor with any null vector x^μ is non-negative: $R_{\mu\nu} x^\mu x^\nu \geq 0$ [480]. The null-energy condition is actually violated by Hawking radiation, which extracts energy from a black hole due to the spontaneous creation of particle-antiparticle pairs slightly outside of the event horizon [481]. In this case, the Area Law is superseded by a generalized law which considers both the entropy of the black hole and the Hawking radiation [482, 483].

[3] Weak cosmic censorship means that all singularities are to be surrounded by an event horizon [484], rendering them "invisible" to observers outside of the event horizon, i.e., there are no naked singularities.

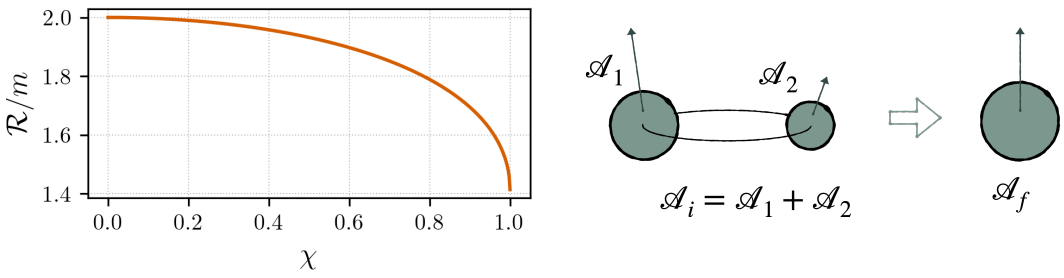


Figure 11.1: *Left:* Black hole event horizon "radius" $\mathcal{R}$ (in units of mass m) as a function of spin magnitude $\chi \in [0, 1]$. *Right:* Schematic visualization of Hawking's Area Law for a binary black hole merger.

be smaller than the area of the final, remnant black hole (subscript f):

$$\mathcal{A}_f - \mathcal{A}_i \geq 0, \quad \text{where } \mathcal{A}_i = \mathcal{A}_1 + \mathcal{A}_2 \ , \tag{11.3}$$

as visualized in Fig. 11.1. To work with unitless quantities, we use the *fractional* change in area:

$$\Delta\mathcal{A} \equiv \frac{\mathcal{A}_f - \mathcal{A}_i}{\mathcal{A}_i} \geq 0 \ . \tag{11.4}$$

Thus, confirming the Area Law using a gravitational-wave observation from a merging black hole binary requires, first, *independently measuring the initial and final black hole areas* and, second, *quantifying the credibility at which the fractional difference in these areas* $\Delta\mathcal{A}$ *is greater than 0* [395].

By using waveform models calibrated to general relativity, typical full-signal gravitational-wave inference implicitly assumes that the Area Law holds true. Thus, to achieve *independent* initial and final area measurements, we must *separately* extract the properties of the component black holes from the pre-merger ("inspiral") portion of the signal, and the remnant black hole from the post-merger ("ringdown") signal. By discarding data between the inspiral and ringdown (i.e., the merger), such a test becomes sensitive to any physics that violates the conditions laid out in Sec. 11.1 and leads the inspiral and ringdown regimes to be better described independently, rather than coherently. Conducting parameter estimation in the time domain—rather than the typical frequency domain—allows gravitational-wave strain data to be cleanly split into the inspiral and ringdown regimes [Ch. 6].

11.3 GW250114

Observed nearly an exact decade later, GW250114 [18] is a near carbon copy of the first-ever gravitational wave detection, GW150914 [17]—in all but one

facet: it is by far the loudest signal thus far observed, with a whopping signal-to-noise ratio (SNR) of 80 [18].

> *By "loudest," I mean that the amplitude of the signal is much larger than the background noise in our detectors. Think of it like being at a dinner party where everyone is quietly chatting. Previous gravitational waves detected by LIGO-Virgo-KAGRA are somewhere between a loud laugh or a shout over the din of the party. GW250114 is akin to a fire alarm going off.*
>
> — ***SJM*** in IFL Science [475]

Figure 11.2 shows time and time-frequency representations of GW250114 in the LIGO Livingston and LIGO Hanford detectors, where it can be seen that the signal's strain amplitude reaches $> 10\sigma$ above the noise. GW250114's unprecedentedly high SNR offers highly precise constraints on the component and remnant masses and spins—and thus areas [Eq. 11.2]—of its binary black hole source, making it a prime candidate for observationally testing Hawking's Area Law [395].

11.4 Results

In the remainder of this chapter, I describe how we independently measure the initial and final areas of GW250114's binary black hole source and use these constraints to estimate a significance for the non-detection of a violation of Hawking's Area Law.

(a) Initial horizon area

The initial black holes sourcing GW250114 are considered at wide separations, where they obey the Kerr metric and the total area is the sum of the individual areas. Assuming that the binary orbit is quasi-circular [100] and well-described by general relativity [24, 424, 487–489], we model the signal with the numerical relativity surrogate waveform model NRSur7dq4 [16] up to some time $t_<$. Using the time-domain inference package `tdinf` [161], we then infer (among other parameters) the masses and spins of the initial black holes, informed *only* by data before $t_<$, and from them generate a posterior probability distribution on the initial area $\mathcal{A}_i = \mathcal{A}_1 + \mathcal{A}_2$. Analysis settings are given in Appendix 11.A.

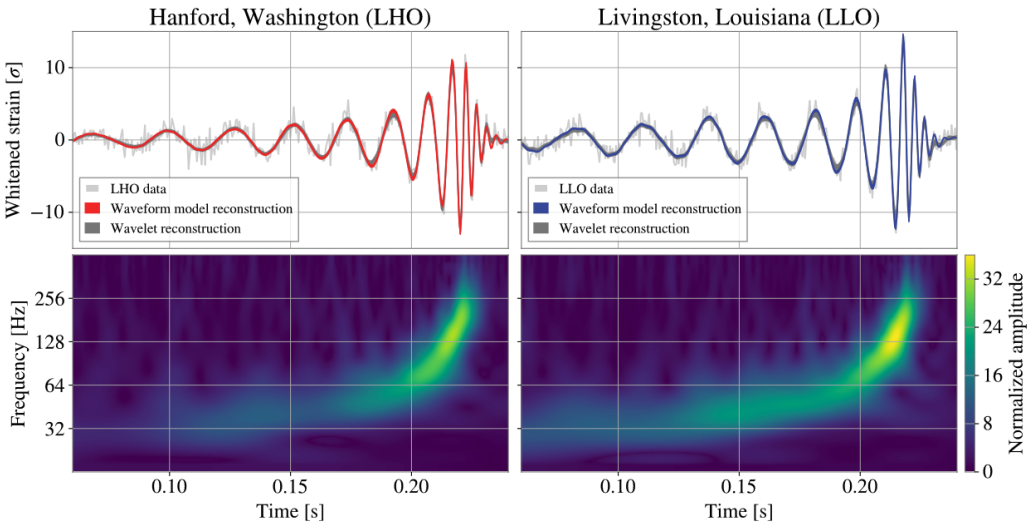


Figure 11.2: Data from LIGO Hanford (left) and LIGO Livingston (right) and the GW250114 signal reconstruction. Times are relative to January 14th, 2025, 08:22:03 UTC. The top panels show whitened data versus time and signal reconstructions (90% credible regions), either with a waveform model for black hole binaries in general relativity [16] or via a model-agnostic wavelet-based approach [474, 486]. The bottom panels show a time–frequency spectrogram of the data.

In general relativity, there is not a cleanly defined value of t where the inspiral ends and the merger begins. Regardless, we must decide upon some $t_<$ to use for our analysis. The choice involves a tradeoff between SNR and potential contamination from the merger: if $t_<$ is too conservatively chosen (i.e., too early in the signal), we wager unnecessary loss of SNR that may render the Area Law test uninformative. If $t_<$ is too late, we may gain SNR at the risk of potentially describing the beginning of the non-linear merger regime of the signal, rendering the test effectively useless.

We turn to a physically-motivated model: to interpret the pre-merger truncation times, we use the gravitational-wave luminosity as a proxy for how "relativistic" the system is [148]. Based on the full-signal analysis, the luminosity is sharply peaked at merger and drops to 10% of its maximum at $-36\,t_M$[4] before $t_{\rm peak}$.[5] Rounding to the nearest $5\,t_M$, we conservatively select $t_< = -40\,t_M$ for the flagship Area Law test, a choice that excludes the two

[4] Inspiral truncation times are quoted in units of the total mass from a corresponding full-signal analysis, $t_M \equiv (1+z)\,G\,M/c^3 = 0.354$ ms.

[5] A proxy for the time of merger, $t_{\rm peak}$ is the peak of the strain over the celestial sphere, given by Eq. (3) in Ref. [18]. See Table 11.1.

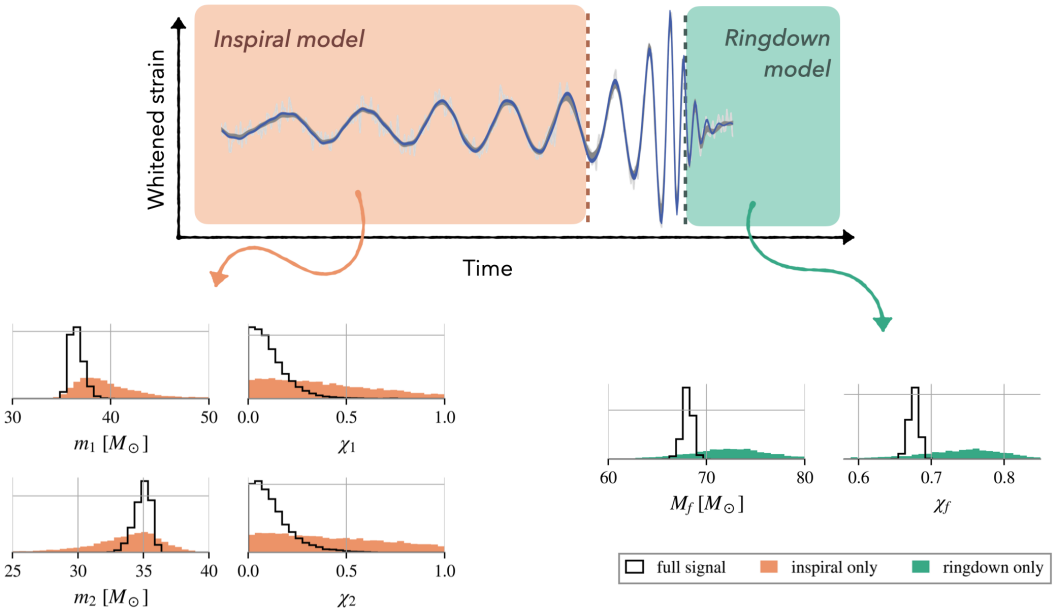


Figure 11.3: Posterior distributions for the component (subscripts 1,2) and remnant (subscript f) masses and spins for GW250114 for inspiral-only (orange), ringdown-only (green) and full-signal (black) analyses, along with a schematic depiction of the method. Inspiral measurements are from data before $t_< = -40\,t_M$, using the NRSur7dq4 waveform model. Ringdown measurements are from data after $t_> = 10.5\,t_{M_f}$ using the 220 QNM.

loudest signal cycles. For reference, the total signal SNR before $t_< = -40\,t_M$ is 55; even truncated, GW250114 has a higher SNR than any other signal to date [9]. The component mass and spin constraints from the $t_< = -40\,t_M$ analysis are shown in orange in Fig. 11.3. Additional choices for $t_<$ are also explored in Figs. 11.4 & 11.5.

(b) Final horizon area

The mass and spin of GW250114's remnant black hole are inferred through the post-merger signal using a superposition of quasi-normal modes (QNMs) [15, 207–209, 490]; see section titled "The Ringing of the Remnant Black Hole" in Ref. [18] for details on the QNM analysis. Similarly to the inspiral calculation, picking a post-merger analysis start time requires balancing: *(i)* independence from the full-signal analysis that assumes the Area Law, *(ii)* maximizing the amount of data and thus the SNR, and *(iii)* using a model within its regime of validity.

We adopt the most basic QNM model, i.e., just the single $\ell = |m| = 2, n = 0$

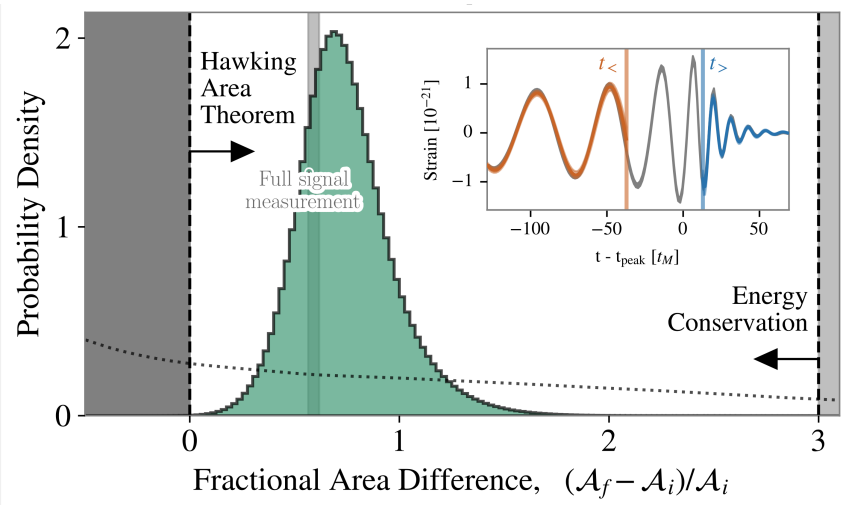


Figure 11.4: Fractional difference between the area of the final black hole, $\mathcal{A}_f$, and the initial black holes, $\mathcal{A}_i$. Posterior (histogram) and prior (dotted line) for $t_< = -40\,t_M$. The histogram is produced from all pairwise combinations of `ringdown` and `pyRing` post-merger samples with `tdinf` pre-merger samples. The shading on the right highlights configurations that would violate energy conservation $M_f \leq m_1 + m_2$, a bound saturated for maximally-spinning initial black holes that merge into a nonspinning remnant [491–493]. The shading on the left highlights configurations that would violate the Area Law, $(\mathcal{A}_f - \mathcal{A}_i)/\mathcal{A}_i < 0$. The vertical gray band is the 90% credible interval from the full-signal analysis, i.e., the general relativity expectation for GW250114. The inset illustrates the analysis via the signal reconstruction in LIGO Livingston from the full-signal (gray), and when split into the pre-merger (orange) and post-merger (blue) stages; the merger phase is included in neither analysis. We find that $\mathcal{A}_f > \mathcal{A}_i$ to high credibility, indicating that GW250114 obeys Hawking's Area Law.

(or "220") mode, at its earliest time of applicability, $t_> = 10.5\,t_{M_f}$.[6] Details about the selection of this time are given in Ref. [18]. We use the 220 mode's inferred frequency and damping rate to calculate the mass and spin of the final Kerr black hole—shown in green in Fig. 11.3—and hence its area, $\mathcal{A}_f$. Additional cutoff times and modes are explored in Fig. 11.5.

(c) Change in area

Figure 11.4 shows the fractional difference between the final and initial black hole horizon areas, $\Delta\mathcal{A}$ [Eq. 11.4], for GW250114 with $t_< = -40\,t_M$ and $t_> = 10.5\,t_{M_f}$. We find that $\Delta\mathcal{A} > 0$ at the 4.4σ level. The method used to estimate

[6] Ringdown start times are quoted in units of the final mass, $t_{M_f} \equiv (1+z)\,G\,M_f/c^3 = 0.337$ ms. See Table 11.1.

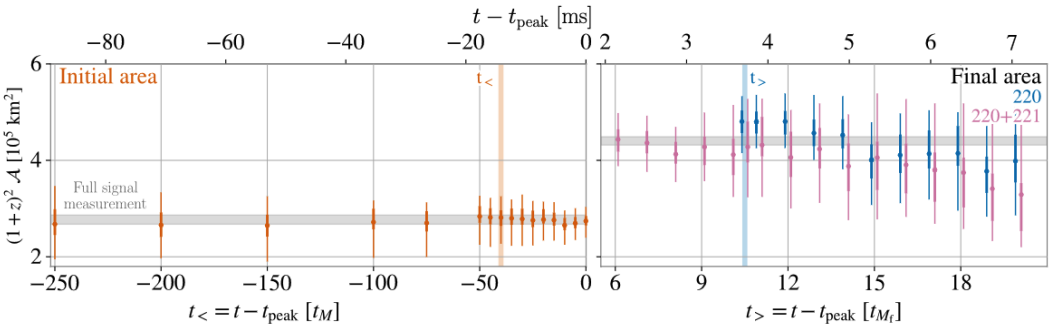


Figure 11.5: Measurements for the initial and final black hole areas as a function of fit time. *Left:* Initial black hole area inferred by analyzing the pre-merger signal with `tdinf`, truncated at the time indicated on the x-axis. *Right:* Final black hole areas inferred by analyzing the post-merger signal with `ringdown`, with the analysis start time indicated on the x-axis. Results are shown for the two-mode analysis starting at $t_> = 6\,t_{M_f}$, while the single-mode results are only presented from $t_> = 10.5\,t_{M_f}$ and onwards. Vertical bands highlight the reference areas quoted in Fig. 11.4: at $-40\,t_M$ before the peak strain for the inspiral analysis and $10.5\,t_{M_f}$ after the peak strain for the single-mode ringdown analysis. Horizontal grey bands indicate the 90% credible intervals for the initial and final areas as inferred from the full inspiral-merger-ringdown analysis, where the Area Law is assumed to hold *a priori*.

significance is described in Sec. 11.4d. The strength of GW250114 enables this first-ever high-confidence observational support of Hawking's Area Law, even while excluding the loudest portions of the signal and without any modeling of the nonlinear merger dynamics through numerical relativity, cf. inset of Fig. 11.4. Similar analyses of GW150914 [395, 419] yielded $\sim 2\sigma$ results, albeit with less conservative assumptions: a pre-merger analysis that extended to the waveform peak and a quasinormal mode model that presumed circular polarization.

Figure 11.5 presents initial and final black hole area posteriors for even more values of $t_>$ and $t_<$, as well as those from an additional two-mode ringdown QNM model. We find $\Delta\mathcal{A} > 0$ to at least 3.4σ for all times $t_< > -250\,t_M$, exceeding 5σ for $t_< \geq -10\,t_M$. The two-mode QNM model at its earliest time of applicability, $t_> = 6\,t_{M_f}$, also agrees with the Area Law at 3.6σ (Fig. 11.6). As expected, the uncertainty in the remnant area grows as the start time is pushed to later parts of the data where the signal is weaker; for the initial area, the opposite is true and the uncertainty grows as the analysis end time is pushed earlier. All results are further consistent with the general relativity

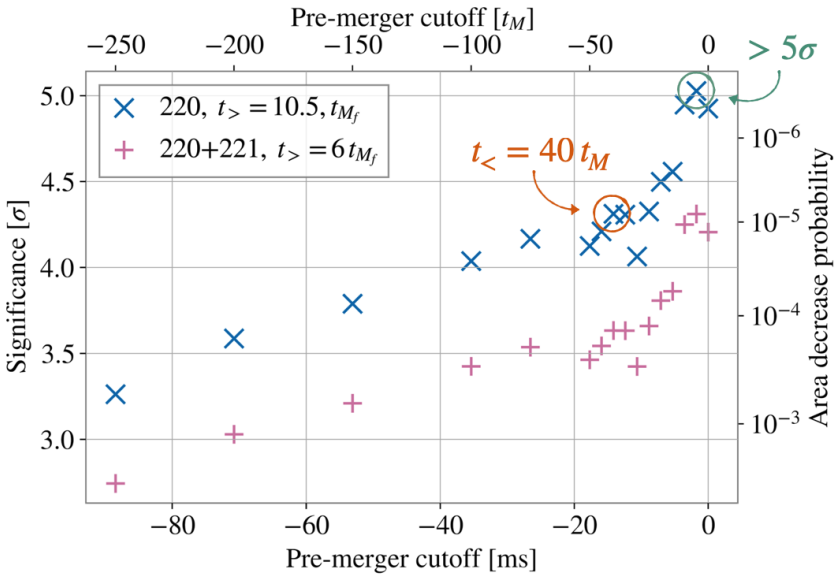


Figure 11.6: The significance of a non-detection of a violation of Hawking's Area Law (Eq. 11.5) as a function of inspiral end-time $t_<$. Two cases of ringdown modes and start times are shown: the 220 mode with $t_> = 10.5\,t_{M_f}$ in blue and the 220+221 modes with $t_> = 6\,t_{M_f}$ in pink, both generated using `ringdown` [367]. As $t_<$ increases, i.e., more inspiral data are analyzed and the SNR correspondingly increases, the significance that the data obey Hawking's Area Law also increases, reaching $> 5\sigma$ when $t_< = -2.5t_M$.

expectation for GW250114, obtained from the full-signal analysis with NRSur7dq4, which considers the entire signal coherently and obeys the Area Law *a priori*.

(d) Assessing significance

The significance of a non-detection of a violation of Hawking's Area Law is related to the separation between the posterior distributions on the initial and final areas. We estimate this significance as $X\sigma$ with

$$X = \frac{\mu_f - \mu_i}{\sqrt{\sigma_f^2 + \sigma_i^2}} \tag{11.5}$$

where μ, σ are the means and standard deviations of the initial (i) and final (f) area posteriors. Following literature in cosmology about resolving the Hubble tension [494, 495], we opt to use the first two cumulants to estimate significance, rather than the tails of the distribution. The latter is highly sensitive to sampling error and is unreliable past the $\sim 4\sigma$ level.

Figure 11.6 plots the significance X given in Eq. (11.5) as a function of inspiral

end-time ($t_<$) with two different combinations of ringdown modes and start times ($t_>$). The case highlighted in Fig. 11.4 is circled in orange, while that with $X > 5$ is circled in green. The final areas used in Fig. 11.6 were generated using `ringdown` [367]. If we instead use the final area measurements generated with `pyRing` [396], the significance that $\Delta\mathcal{A} > 0$ increases by $\sim 0.5\sigma$.

Ch. 11 Data and Code Availability

Strain data from the LIGO detectors for GW250114 are available from the Gravitational Wave Open Science Center at https://doi.org/10.7935/1g4j-2028. All the material required for reproducing the figures in Ref. [18], including scripts and posterior distributions from the analyses, is available in the data release on Zenodo at https://doi.org/10.5281/zenodo.16877102.

Ch. 11 Appendices

11.A Inspiral Analysis Details

To enable the sharp truncation of the data and gravitational-wave model at a time before the binary merger [393, 395, 419], we carry out the analysis in the time-domain [Ch. 6], guided by the reference parameters in Table 11.1. We use the `tdinf` inference package [101, 102, 161], which samples the 15-dimensional parameter space of a quasicircular black-hole binary using the `emcee` sampler [252].

Parameter	Value	Description
$t_{\rm peak}$ (geocenter)	1420878141.235932 s	Reference time at geocenter
α	2.333 rad	Right ascension of source
δ	0.190 rad	Declination of source
ψ	1.329 rad	Polarization angle
$t_{\rm peak}^{\rm LHO}$	1420878141.2190118 s	Reference time at LIGO Hanford
$t_{\rm peak}^{\rm LLO}$	1420878141.2165165 s	Reference time at LIGO Livingston
$(1+z)\,M_f$	$68.409\,M_\odot$	Reference final black hole mass
t_{M_f}	0.337 ms	Reference $(1+z)M_f$ in units of time
$(1+z)\,M$	$71.849\,M_\odot$	Reference total mass
t_M	0.354 ms	Reference $(1+z)M$ in units of time

Table 11.1: Reference parameters for the GW250114 inspiral and ringdown analyses.

We adopt the same settings as the full-signal frequency-domain analyses with `bilby` presented in Ref. [18] in terms of the data, trigger time, power spectral density, and bandwidth. Differences between the time and frequency domain analyses are: (1) the amount of data considered, (2) certain priors, and (3) the treatment of calibration. The time-domain analysis of the full signal is based on 1.4 s of data, which are appropriately truncated. The priors are the same as the `bilby` analysis other than the masses (uniform in total mass and mass ratio), distance (uniform in luminosity distance), and time (Gaussian centered around the geocenter trigger time with a standard deviation of 0.01 s). The distance prior has a minimal effect on the inferred area, which is based on redshifted masses that are minimally correlated with the distance. The composite Area Law prior is broad and relatively flat, as seen in Fig. 11.4. Lastly, the time-domain analysis neglects the uncertainty over the detector calibration, which `bilby` marginalizes over. We verified that calibration has minimal impact by repeating the `bilby` analysis neglecting calibration uncertainty.

To obtain properly-converged independent posterior samples from `tdinf`, we use the autocorrelation length (ACL) from the full ensemble of walkers to determine the burn-in period and thinning of the chains. Depending on the truncation time, the lengths of these chains range between 100,000–500,000 steps, each with 512 walkers. We use a burn-in of at least five times the maximum ACL across sampled parameters, corresponding to $> 40\%$ the chain length, and thin by half of the minimum ACL. For some of the truncation times, a handful of chains are not converged to the bulk of the posterior at the end of the burn-in, instead representing a secondary mode in likelihood and sky location. For these chains, we manually extend the burn-in period until they converge to the bulk of the posterior. We have confirmed that the posterior is not restricted by any of the prior edges.

We have also verified that the `tdinf` and `bilby` analyses of the full GW250114 signal yield statistically identical posteriors when reweighted to the same prior for both NRSur7dq4 and IMRPhenomXPHM—shown in Fig. 11.7 for component masses and spins. Finally, using GR and non-GR injections, we confirm that `tdinf` and `ringdown` are able to identify when a signal does *not* obey the Area Law, as shown in Fig. 11.8.

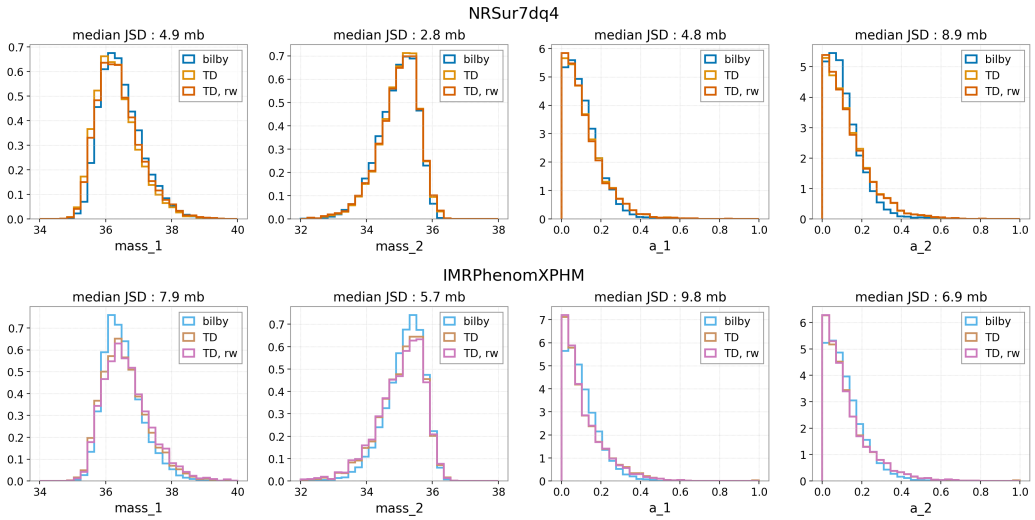


Figure 11.7: Comparison of m_1, m_2, χ_1, χ_2 posteriors (from left to right) obtained with time- and frequency-domain analyses of the full GW250114 signal using NRSur7dq4 [16] (top row) and IMRPhenomXPHM [245] (bottom row). Frequency-domain results obtained with `bilby` are shown in blue, time-domain results with `tdinf` are in orange, and time-domain results reweighted to the `bilby` priors are in red/pink. Each panel is labeled with the Jensen-Shannon Divergence (JSD) [425] in milli-bits (mb) between the `bilby` and reweighted `tdinf` posteriors. To account for finite sampling in JSD calculation, we quote the median JSD over 100 random draws of 5,000 samples per-posterior. All JSDs are below 10 mb, indicating consistency [Sec. 8.2].

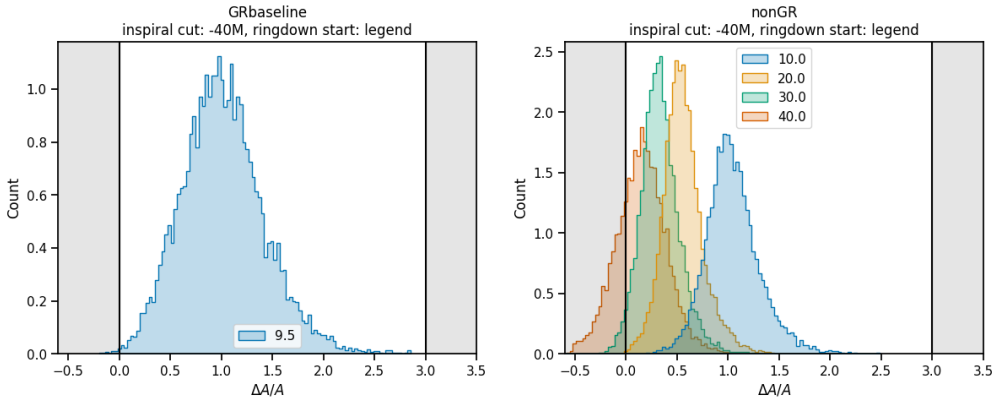


Figure 11.8: Confirmation that the `tdinf` + `ringdown` method is able to identify when a signal does *not* obey the Area Law. *Left:* $\Delta\mathcal{A}$ posterior from a simulated binary black hole signal that is consistent with GR for $t_< = -40\,t_M$ and $t_> = 9.5\,t_{M_f}$. The posterior is confidently constrained above $\Delta\mathcal{A} = 0$. *Right:* $\Delta\mathcal{A}$ posteriors from a simulated binary black hole signal that is *not* consistent with GR $t_< = -40t_M$, and various $t_>$ (colors). Depending on the specific choice of $t_>$ and $t_<$, the initial and final area posteriors move around drastically, as does $\Delta\mathcal{A}$. For the 220 ringdown analysis, we attribute this to the beating between the two strongly non-Kerr modes.

Part III

What do we know about the spins of the binary black hole population?

Chapter 12

NO EVIDENCE THAT THE MAJORITY OF BLACK HOLES IN BINARIES HAVE ZERO SPIN

This chapter is published as Ref. [164]:

Thomas A. Callister, **Simona J. Miller**, Katerina Chatziioannou, and Will M. Farr. No Evidence that the Majority of Black Holes in Binaries Have Zero Spin. *Astrophys. J. Lett.*, 937(1):L13, 9 2022. ***SJM*** *conceptualized the project jointly with co-authors, ran all spin magnitude and tilt population analyses, wrote around half of the paper's text, and made Figures 5-9, 12, 14, and 15.*

Abstract

The spin properties of merging black holes observed with gravitational waves can offer novel information about the origin of these systems. The magnitude and orientations of black hole spins offer a record of binaries' evolutionary history, encoding information about massive stellar evolution and the astrophysical environments in which binary black holes are assembled. Recent analyses of the binary black hole population have yielded conflicting portraits of the black hole spin distribution. Some work suggests that black hole spins are small but non-zero and exhibit a wide range of misalignment angles relative to binaries' orbital angular momenta. Other work concludes that the majority of black holes are non-spinning while the remainder are rapidly rotating and primarily aligned with their orbits. We revisit these conflicting conclusions, employing a variety of complementary methods to measure the distribution of spin magnitudes and orientations among binary black hole mergers. We find that the existence of a sub-population of black hole with vanishing spins is not required by current data. Should such a sub-population exist, we conclude that it must contain $\lesssim 60\%$ of binaries. Additionally, we find evidence for significant spin-orbit misalignment among the binary black hole population, with some systems exhibiting misalignment angles greater than 90°, and see no evidence for an approximately spin-aligned sub-population.

12.1 Introduction

The spins of black holes in merging binaries detected with gravitational waves promise to illuminate open questions in massive stellar evolution and compact binary formation. The orientations of component black hole spins may differentiate between binaries formed via isolated stellar evolution and those formed dynamically in clusters or the disks of active galactic nuclei, and additionally offer a means of measuring natal kicks that black holes receive upon their formation [67, 85–87, 89, 93, 120, 282, 329, 412, 496–498]. Spin magnitudes, meanwhile, are determined by poorly-understood angular momentum processes operating in stellar cores, and may be further affected by binary processes such as tidal torques or mass transfer [58, 59, 81, 83, 86]. Black holes with large spin magnitudes might also point to hierarchical assembly in dense environments, involving component black holes that are themselves the products of previous mergers [43, 61, 63, 67, 368, 499].

Despite the large astrophysical interest, spin measurements are hampered by the fact that spin dynamics have a relatively weak imprint on the gravitational-wave signal. The main effect of spins (anti)parallel to the Newtonian orbital angular momentum is to (speed-up) slow-down the binary inspiral and merger. Spins in the plane of the orbit, on the other hand, give rise to precession that modulates the amplitude and phase of emitted gravitational waves. Even with informative measurements of these effects, however, it is not straightforward to constrain all six degrees of freedom independently.

Recent work [165, 282, 332] has yielded conflicting conclusions regarding the distribution of spins among the binary black hole population witnessed by Advanced LIGO [4] and Virgo [5]. Specifically:

- Do binary black holes have small but non-zero spins that may be misaligned significantly with the Newtonian orbital angular momentum, i.e., with spin-orbit misalignments $> 90°$?
- Or, do a majority of binaries have spins that are identically zero or, if nonzero, preferentially aligned with the Newtonian orbital angular momentum, i.e., with misalignments $< 90°$?

These questions highlight the subtleties and difficulties inherent in statistical analysis of weakly informative measurements. This debate also hinges on

a variety of *technical difficulties* related to hierarchical inference of narrow population features using discretely sampled data.

The two possibilities listed above carry considerably different astrophysical implications for the assembly and evolution of binary black hole mergers. Systems arising from isolated binary evolution are traditionally expected to have spins preferentially parallel to their orbital angular momenta [58, 80, 83, 500]. Significant spin-orbit misalignment, in contrast, is considered difficult to achieve under canonical isolated binary evolution and so would indicate either the presence of alternative formation channels or a change in the paradigm of isolated binary evolution [86, 89, 90, 93, 329]. The degree to which black holes are observed to be primarily spinning or non-spinning, meanwhile, would support or refute theories positing highly efficient angular momentum transport in stellar interiors [51, 76, 501].

Our goal in this paper is to revisit these incompatible conclusions. In Sec. 12.2 we review recent literature and describe in more detail what is known, unknown, and still debated about binary black hole spins. We then employ a string of increasingly sophisticated analyses to study the population of binary black hole spins and determine what we can and cannot robustly conclude about the magnitudes and orientations of black hole spins. We begin in Sec. 12.3 with a simple counting argument, which demonstrates that the fraction of black holes that are non-spinning is consistent with zero. We then turn to full hierarchical analyses of the binary black hole population, studying the distributions of effective inspiral spins (Sec. 12.4) and component spin magnitudes and orientations (Sec. 12.5). In all cases, we find that the fraction of non-spinning black holes can comprise up to $60-70\%$ of the total population, but that this fraction cannot be confidently bounded away from zero. Overall, the inferred spin magnitude distribution is consistent with a single population extending smoothly from zero up to magnitudes of approximately 0.4. Additionally, we find a preference for considerable spin-orbit misalignments among the binary black hole population, with some spins inclined by more than 90° relative to their orbits.

12.2 The spins of black holes in binaries

Each component black hole in a binary has dimensionless spin vectors χ_1 and χ_2. Six parameters are needed to fully specify these two spins, with each spin

vector characterized by a magnitude $0 \leq \chi_i \leq 1$, tilt angle θ_i, and azimuthal angle ϕ_i, where $i \in \{1, 2\}$ in a coordinate system where the z-axis is aligned with the Newtonian orbital angular momentum.

Not all spin degrees of freedom are as dynamically important in a gravitational-wave signal, however. While the two spin magnitudes are conserved throughout the binary evolution [modulo horizon absorption effects; 336, 502], the four spin angles vary due to spin-precession. A combination known as the *effective inspiral spin* is conserved under spin-precession to at least the 2PN order[1] [193]

$$\chi_{\rm eff} = \frac{\chi_1 \cos\theta_1 + q\chi_2 \cos\theta_2}{1+q} \in [-1, 1]\,, \tag{12.1}$$

where $q = m_2/m_1 < 1$ is the mass ratio. Though not conserved under spin-precession, the *effective precessing spin*

$$\chi_{\rm p} = \max\left[\chi_1 \sin\theta_1, \left(\frac{3+4q}{4+3q}\right) q\,\chi_2 \sin\theta_2\right] \in [0, 1]\,, \tag{12.2}$$

reflects the degree of in-plane spin and characterizes spin-precession dynamics [203]. [2] Although modern waveform models make use of the full 6-dimensional spin parameter space [16, 245, 466, 503], earlier versions were constructed in terms of $\chi_{\rm eff}$ and $\chi_{\rm p}$, leveraging their relevance in binary dynamics.

At current signal strengths it is not possible to meaningfully constrain all 6 spin parameters. When exploring the population of compact binary spins, we therefore generally work in one of two lower-dimensional spaces.

- The most straightforward approach is to constrain the distribution of "effective" spin parameters. Though the $\chi_{\rm eff}$ and $\chi_{\rm p}$ distributions do not unambiguously reveal information about individual component spins, they do enable *categorical* conclusions to be made regarding compact binary spins. A non-vanishing $\chi_{\rm p}$ distribution, for example, indicates that spins are not perfectly aligned with binary orbits, while the identification of negative $\chi_{\rm eff}$ requires at least some component spins to be inclined by more that $\theta_i = 90°$. A common approach, and the baseline model that we will extend below, is to treat the marginal distribution of $\chi_{\rm eff}$ as a truncated Gaussian [162, 504]. We refer to this as the `Gaussian` model; see Appendix 12.A.

[1] An NPN order is defined as being proportional to $(u/c)^{2N}$ compared to its leading order term, where u is a characteristic velocity of the system and c is the speed of light.

[2] Recent work has also explored alternative parameters which better capture the imprint of spin-precession in gravitational-wave signals [204, 205].

- Going one step beyond the effective spin parameters, we can directly model the distribution of component spin magnitudes and tilt angles. A popular choice is to treat the component spin magnitude distribution as a Beta distribution, while spin tilts are drawn from a mixture between two components, an isotropic component and a preferentially aligned component [351, 505]. The azimuthal spin angles are ignored and presumed to be uniformly distributed as $\phi_1, \phi_2 \sim U[0, 2\pi]$ [this assumption was relaxed in 506]. We assume that component spin magnitudes and tilts are independently and identically distributed within this model and refer to it as the `Beta+Mixture` model,[3] discussed further in Appendix 12.A.

What we know about black hole spins

Before proceeding to explore the prevalence of zero-spin events and extreme spin-orbit misalignment (i.e., spin tilts larger than 90°), it is useful to first examine the conclusions that are broadly and robustly recovered by different analyses and authors.

i. Black hole spins are not all maximal. Following the first four binary black hole detections by Advanced LIGO, Refs. [93] and [32] determined that if all component spins are aligned ($\cos\theta_i = 1$), then their magnitudes must be small, $\chi_i \lesssim 0.3$. If spins were assumed to be isotropically-oriented, though, near-extremal spins were still allowed. Ref. [507] conducted a similar analysis using an expanded catalog, and also found large spins to be disfavored, regardless of their orientation. The degeneracy between spin magnitude and orientation was later broken by efforts to simultaneously measure these two properties. Using the `Default` model, [505] and Ref. [327] found that typical spin magnitudes are small, with 50% of black holes having $\chi \lesssim 0.3$. Ref. [327] furthermore revisited the analysis of Ref. [32], now finding that large spin magnitudes are moderately disfavored *even when* assuming isotropic orientations. Ref. [504], meanwhile, studied the $\chi_{\rm eff}$ distribution, leveraging the `Gaussian` model to conclude that effective spins are concentrated about zero. They argued that, if component spins are aligned, then the measured $\chi_{\rm eff}$ distribution implies that component spin magnitudes are $\chi \lesssim 0.1$. Ref. [504] additionally measured the fraction of binaries whose secondaries have maximal spins due to tidal spin-up;

[3] A closely related model in which the spin tilts are not independently and identically distributed but instead *both* originate from either the isotropic or the aligned component is called the `Default` spin model in Ref. [282] and Ref. [283].

they found the fraction to be consistent with zero and bounded to < 0.3. To date, no confident detection has exhibited unambiguously large $\chi \gtrsim 0.5$.

ii. Black hole spins are not all zero. Despite evidence pointing towards preferentially small spin magnitudes, not *all* black holes can be non-spinning. Using the `Gaussian` effective spin model, Refs. [504] and [162] concluded that the $\chi_{\rm eff}$ distribution is *inconsistent* with a delta-function at zero; hence the $\chi_{\rm eff}$ distribution possessed either a non-zero mean or non-zero width. This conclusion was bolstered by Ref. [282], who found that the $\chi_{\rm eff}$ distribution is centered at ~ 0.05 with a non-zero width ≥ 0.08, and that the component spin magnitude distribution peaks at small values but also with a non-vanishing width ~ 0.15. Several individual events such as GW151226 [296] and GW190517 [286] are also confidently known to possess spin, although their component spins are individually poorly measured (see e.g. Fig. 12.11).

iii. Black holes exhibit a range of spin-orbit misalignment angles. Since spin-precession is a subtle effect, the spin tilts of individual binaries are highly uncertain. Analyses of the population, however, indicate that spins are not purely aligned but instead exhibit a range of misalignment angles. In their analysis of the $\chi_{\rm eff}$ distribution, Ref. [507] reported evidence against pure spin-orbit alignment. Ref. [282] later employed the `Default` model to directly measure the distribution of misalignment angles, recovering a possible preference for alignment but ruling out perfect alignment at high credibility. Ref. [282] furthermore extended the `Gaussian` model to jointly measure the mean and variance of both $\chi_{\rm eff}$ and $\chi_{\rm p}$. They found a delta-function at $\chi_{\rm p} = 0$ to be disfavored, indicating the presence of spin-orbit misalignment in the population. Ref. [165], meanwhile, used an extended version of the `Default` model to measure the *maximum* spin-orbit misalignment angle among the binary black hole population, finding that the observed population requires some misalignment angles exceeding $\theta \gtrsim 65^\circ$ ($\cos\theta \leq 0.43$) at 99% credibility. Evidence for spin precession identified in individual events is also under active investigation [294, 297, 399, 499, 508–511].

What is under debate about black hole spins

i. Do most black holes have zero spin? Although it is agreed that *not all* black holes can possess zero spin, a debated question is whether *most* do. Using both the `Default` and `Gaussian` models, Ref. [282] found no indication of an

excess of zero-spin systems; predictive checks designed to test the goodness-of-fit of these models found these continuous unimodal distributions to be good descriptors of the observed population. In the context of hierarchical black hole formation, Ref. [368] and Ref. [63] directly measured the fraction of "first-generation" black holes with zero spin, finding this fraction to be consistent with zero. A different conclusion was drawn in Ref. [332], who modeled the $\chi_{\rm eff}$ distribution not as a single Gaussian but via a mixture of three components,

$$\begin{aligned} p(\chi_{\rm eff}|\zeta_{\rm pos},\zeta_0,\sigma) = \zeta_{\rm pos}\,&\mathcal{N}_{[0,1]}(\chi_{\rm eff}|0,\sigma) \\ &+ \zeta_0\,\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|0,0.04) \\ &+ (1-\zeta_{\rm pos}-\zeta_0)\,\mathcal{N}_{[-1,0]}(\chi_{\rm eff}|0,\sigma)\,, \end{aligned} \tag{12.3}$$

corresponding to a half-Gaussian encompassing $\chi_{\rm eff} > 0$, a half-Gaussian encompassing $\chi_{\rm eff} < 0$, and a Gaussian centered at zero. This third component was intended to capture systems with $\chi_{\rm eff} = 0$; its finite standard deviation (fixed to 0.04) was chosen to mitigate sampling effects, a technical issue we discuss further below. Ref. [332] argued that a significant fraction of observed binaries are possibly associated with this zero-spin sub-population. They reported a maximum likelihood value of $\zeta_0 \approx 0.5$, although ζ_0 remained consistent with zero. A similar but stronger conclusion was forwarded by Ref. [165]. Working in the component spin domain, they extended the `Default` model to include an additional sub-population whose spin magnitudes are identically zero for both binary components. Ref. [165] concluded that $54^{+21}_{-25}\%$ of binaries are members of the zero-spin sub-population at 90% credibility.[4]

ii. Do some black holes have large spin magnitudes? Ref. [282] performed a series of predictive checks, testing the goodness of fit of their models against observation. They concluded that black hole spins are well described by a single unimodal distribution concentrated at small-but-nonzero values. In contrast, Ref. [165] argued that, although they infer most binary black holes to be non-spinning, the remaining binaries are members of a distinct rapidly-spinning sub-population. This secondary population is claimed to exhibit a broad range of spin magnitudes, centered at $\chi \approx 0.45$ but extending to maximal spins. Ref. [410] noted that some individual events exhibit more confidently positive spin than others, speculating that they comprise a secondary

[4] During the late stages of preparation of this study, the exact numerical results of Ref. [165] were updated to account for an analysis bug, though their main conclusions remain unchanged. Numbers in this manuscript correspond to their updated results.

population of more rapidly spinning events. However, their results are based on inspection of individual posteriors and not hierarchical inference of the underlying population, and so it is unclear how their conclusions compare to those of Ref. [165].

iii. Do extreme spin-orbit misalignments exist? Although all analyses agree that at least a moderate degree of spin-orbit misalignment exists, the question of *extreme* misalignments, i.e., $\theta > 90^\circ$, remains. Using the `Gaussian` model, Ref. [282] inferred that at least 12% of binaries have negative $\chi_{\rm eff}$, possible only if one or both component spins have $\theta > 90^\circ$. To determine whether this conclusion was a proper *measurement* or simply an extrapolation of their model, [282] introduced a variable lower bound $\chi_{\rm eff,min}$ on the Gaussian $\chi_{\rm eff}$ distribution, finding the data to require $\chi_{\rm eff,min} < 0$ at 99% credibility. Ref. [332], however, found that this support for negative effective spins is diminished when allowing for the possibility of a zero-spin population as in Eq. (12.3). Ref. [165], meanwhile, explored another variant of the `Default` component spin model, now introducing a variable truncation bound on the spin-tilt distribution. They found this minimum $\cos\theta$ to be consistent with zero. Motivated by these analyses, Ref. [283] further extended the `Gaussian` model to include both a variable truncation bound and a possible zero-spin sub-population. They recovered diminished evidence for negative effective spins, but still recovered a preference for $\chi_{\rm eff,min} < 0$, now at 90% credibility. To date, no individual events discovered by the LIGO-Virgo-KAGRA Collaboration have confidently negative $\chi_{\rm eff}$ or component spins unambiguously inclined by more than 90° [287]. Independent reanalyses of LIGO/Virgo data have identified several candidates with confidently negative $\chi_{\rm eff}$ [496, 512], although most of these candidates do not pass the significance threshold adopted in Ref. [332].

12.3 A counting experiment

The central question of whether the majority of detected black hole binaries have vanishing spins admits a quick back-of-the-envelope estimate. Fully marginalized likelihoods[5] have been obtained for every event in GWTC-2 [63, 286] under two different prior hypotheses: (i) both binary components are non-spinning (NS), with spin magnitudes fixed to $\chi_{1,2} = 0$, and (ii) the black holes

[5]Sometimes known as "evidence," though in this paper we reserve this term for non-technical use.

are spinning (S), with spin magnitudes and cosine tilts distributed uniformly across the intervals $0 \leq \chi_{1,2} \leq 1$ and $-1 \leq \cos\theta_{1,2} \leq 1$. The ratio of the fully marginalized likelihoods gives the Bayes factor $\mathcal{B}^{\rm NS}_{\rm S}$ between non-spinning and spinning hypotheses. Such Bayes factors serve as a primary input in the analysis of Ref. [165], which makes use of parameter estimation samples obtained under both the non-spinning and spinning priors; the Bayes factors between hypotheses is critical in determining how to properly combine these samples.

We start by considering a simple one-parameter model for the fraction of non-spinning binaries, ζ. Given a catalog of $N_{\rm obs}$ observations and data $\{d\}$, the likelihood of $\{d\}$ is

$$\begin{aligned} p(\{d\}|\zeta) &= \prod_{i=1}^{N_{\rm obs}} p(d_i|\zeta) \\ &= \prod_{i=1}^{N_{\rm obs}} \left[p(d_i|{\rm NS})\, p({\rm NS}|\zeta) + p(d_i|{\rm S})\, p({\rm S}|\zeta)\right], \end{aligned} \tag{12.4}$$

where in the second line we have written the likelihood for each individual event as the sum of two terms corresponding to the non-spinning (NS) and the spinning (S) hypothesis. By definition, $p({\rm NS}|\zeta) = \zeta$ and $p({\rm S}|\zeta) = 1-\zeta$. Substituting these expressions into Eq. (12.4) gives

$$\begin{aligned} p(\{d\}|\zeta) &= \prod_{i=1}^{N_{\rm obs}} \left[p(d_i|{\rm NS})\, \zeta + p(d_i|{\rm S})\, (1-\zeta)\right] \\ &= \prod_{i=1}^{N_{\rm obs}} p(d_i|{\rm S}) \left[\frac{p(d_i|{\rm NS})}{p(d_i|{\rm S})}\, \zeta + (1-\zeta)\right] \\ &\propto \prod_{i=1}^{N_{\rm obs}} \left[\mathcal{B}^{\rm NS}_{{\rm S},i}\, \zeta + (1-\zeta)\right], \end{aligned} \tag{12.5}$$

where the likelihood ratio $p(d_i|{\rm NS})/p(d_i|{\rm S})$ is the non-spinning vs. spinning Bayes factor $\mathcal{B}^{\rm NS}_{{\rm S},i}$.

The Bayes factors computed in Ref. [286] and used by Ref. [165] were obtained via nested sampling [253, 513, 514]. To evaluate Eq. (12.5), we instead use posterior samples under the spinning hypothesis to calculate $\mathcal{B}^{\rm NS}_{\rm S}$ via a Savage-Dickey density ratio. The Bayes factors we compute generally agree with those used as inputs in Ref. [165], although with some notable exceptions that may contribute to the discrepancies between their results and our own; see Appendix 12.G.

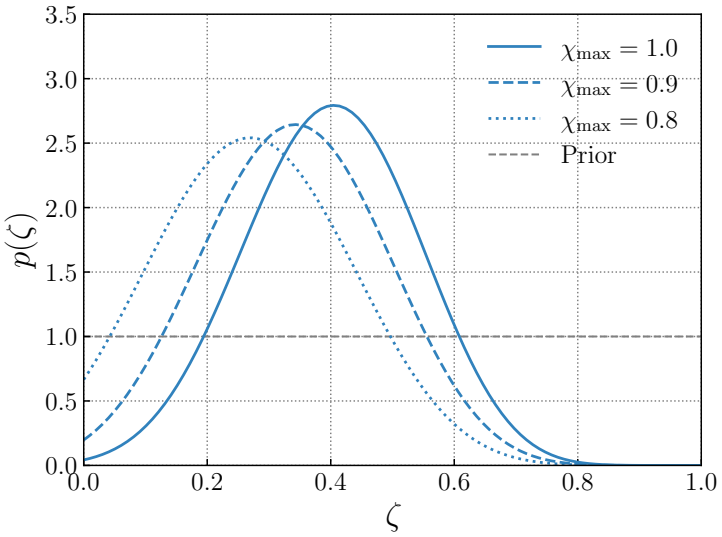


Figure 12.1: Posterior on the fraction ζ of binary black holes in which both components have zero spin, as obtained in the simple counting experiment of Sec. 12.3. This counting experiment uses only events detected through GWTC-2 [286], and furthermore only relies on the fully-marginalized likelihoods for each binary under spinning ($0 \leq \chi_1, \chi_2 \leq \chi_{\rm max}$) and non-spinning ($\chi_1 = \chi_2 = 0$) priors. Values of $\zeta \gtrsim 0.8$ are definitively ruled out. Whether or not ζ is consistent with zero, however, depends more sensitively on assumptions regarding the distribution of black hole spin magnitudes. If we assume that black hole spins range uniformly up to $\chi_{\rm max} = 1.0$, $\zeta = 0$ is disfavored. At the same time, relaxing $\chi_{\rm max}$ to slightly smaller values yields posteriors increasingly consistent with zero, which would indicate no distinct sub-population of non-spinning systems.

The solid curve in Fig. 12.1 shows our resulting posterior on ζ using only GWTC-2 events for which such Bayes factors are available. We find that zero-spin fractions $\zeta \gtrsim 0.8$ are excluded at high credibility. It also appears that $\zeta = 0$ is disfavored, which would imply the presence of at least a few zero-spin systems. However, note that the spinning hypothesis (S) *requires* that spin magnitudes be distributed uniformly up to $\chi_{\rm max} = 1$, a possibility that is heavily disfavored as discussed in Sec. 12.2. What happens if we adopt a more plausible prior distribution for the spinning hypothesis? To answer this question, the dashed and dotted curves in Fig. 12.1 show the ζ posterior given by Eq. (12.5) if we recompute $\mathcal{B}^{\rm NS}_{\rm S}$ but now assume maximum spin magnitudes of $\chi_{\rm max} = 0.9$ and 0.8, respectively, among the "spinning" population. We see that even these small adjustments to $\chi_{\rm max}$ further rule out large ζ and increasingly support $\zeta = 0$.

If, rather than our Savage-Dickey estimates, we instead use the same nested sampling Bayes factors adopted by Ref. [165], we now much more strongly rule out $\zeta = 0$. In Appendix 12.G we track the origin of this behavior to one event, GW190408_181802, whose nested sampling Bayes factor appears significantly overestimated. This system is reported to have a large Bayes factor $\mathcal{B}^{\rm NS}_{\rm S} \sim 130$ in favor of the non-spinning hypothesis, but this conclusion is not supported by the posterior on this system's spins (and thus the Savage-Dickey density ratio); see Fig. 12.14. When we exclude GW190408_181892 from our sample, we obtain consistent $p(\zeta)$ posteriors from both sets of Bayes factors. When including this event, however, the nested sampling Bayes factors cause $p(\zeta = 0)$ to be underestimated by a factor of $\sim 10^2$ relative to the result obtained with Savage-Dickey ratios (again see Fig. 12.14). This could account for the nonzero fraction of non-spinning events found in Ref. [165].

The initial check presented in Fig. 12.1 suggests that the conclusion that most binary black holes comprise a distinct non-spinning sub-population is inconsistent with the parameter estimates for individual binary black systems. At the same time, however, we saw that exact quantitative conclusions depend sensitively on assumptions regarding other aspects of the binary black hole population. We therefore need to undertake a more complete hierarchical analysis, measuring the zero-spin fraction ζ while simultaneously fitting the distribution of black hole spin magnitudes and orientations.

12.4 No sharp features in the effective spin distribution

Going beyond our simple counting experiment, we next consider hierarchical inference of the *effective* spin distribution. An excess of events with vanishing spins would have stark implications for the distribution of effective spin parameters. In Fig. 12.2 we compare the $\chi_{\rm eff}$ distribution implied by the results of Ref. [165] and Ref. [283]. If most black holes are indeed non-spinning, we should correspondingly see a prominent spike at $\chi_{\rm eff} = 0$, and if such a spike exists it should be robustly measurable using an appropriate model.

There are two benefits to searching for this excess of zero-spin systems in the $\chi_{\rm eff}$ domain, before proceeding to even more carefully explore the distribution of component spins. First, inference of the $\chi_{\rm eff}$ distribution offers an independent and complementary check on the existence of a prominent zero-spin population: such a feature should be detectable either in the space of

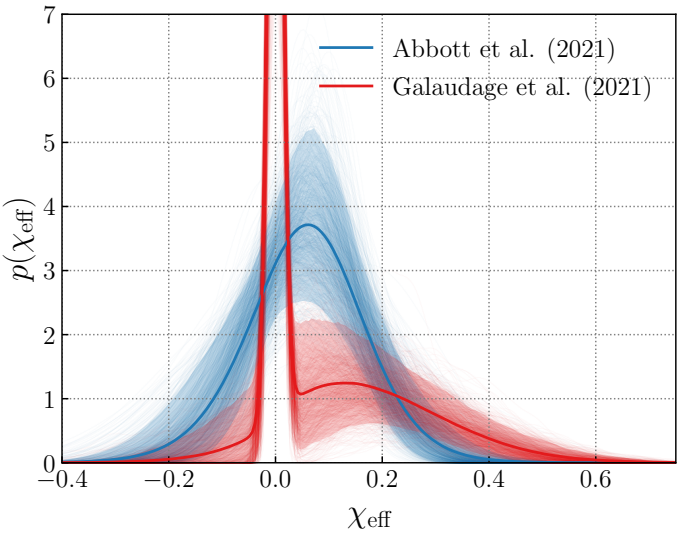


Figure 12.2: Marginal posterior distribution for $\chi_{\rm eff}$ using results from Ref. [283] that have no excess of zero-spin events (blue) and results from Ref. [165] that imply that $\sim 50\%$ of binaries have zero spins (red). The solid lines trace the mean distribution inferred by each model while shaded regions denote the 90% credible regions; the light traces trace individual draws from the population posterior under each model. We see that, if a significant fraction of binary black holes are indeed non-spinning, this should be clearly identifiable through their $\chi_{\rm eff}$ distribution. This figure is modeled after Fig. 5 of Ref. [165].

component spins or effective spins. Second, by analyzing $\chi_{\rm eff}$ and not the higher-dimensional space of spin magnitudes and tilts, we can more easily avoid systematic uncertainties due to finite sampling effects. As detailed in Appendix 12.B, the core ingredients in any hierarchical analysis are the posteriors $p(\lambda_i|d_i)$ on the properties λ_i of our individual binaries (labeled by $i \in [1, N_{\rm obs}]$). In general, however, we do not have direct access to these posteriors, but instead have discrete *samples* $\{\lambda_{i,j}\}$ drawn from the posteriors, where $j \in [1, N_i]$ enumerates the samples from posterior i and N_i is the total number of samples for this event. We therefore ordinarily replace integrals over $p(\lambda_i|d_i)$ with Monte Carlo averages over these discrete samples. The fundamental assumption underlying this approach is that the posterior samples are sufficiently dense relative to the population features of interest. In this paper, however, we are concerned with very narrow features in the binary black hole spin distribution, see Fig. 12.2. In this case, we cannot automatically assume that Monte Carlo averages over posterior samples will yield accurate results.

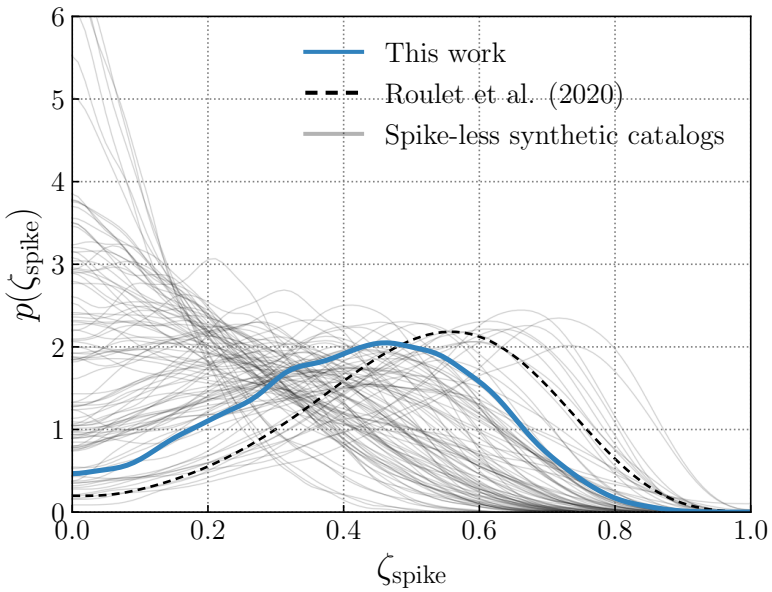


Figure 12.3: Marginalized posterior on the fraction $\zeta_{\rm spike}$ of binary black holes comprising a distinct sub-population with $\chi_{\rm eff} = 0$. The blue curve shows results obtained through analysis of the binary black holes in GWTC-3 [287]. We find $\zeta_{\rm spike}$ to be consistent with zero, indicating no evidence for an excess of zero-spin systems. Nevertheless, $p(\zeta_{\rm spike})$ peaks suggestively at $\zeta_{\rm spike} \approx 0.5$. We demonstrate that this is not unexpected by repeatedly generating and analyzing mock catalogs of $\chi_{\rm eff}$ measurements, drawn from an intrinsically spike-less population described by a simple Gaussian. These catalogs yield posteriors similar to our own, often (and incorrectly) disfavoring $\zeta_{\rm spike} = 0$. As discussed in the main text, this behavior is due to a degeneracy between $\zeta_{\rm spike}$ and the inferred mean of the spinning "bulk" population. For reference, the black dashed line shows the posterior on the zero-spin fraction ζ_0 inferred in Ref. [332], which is qualitatively consistent with both our GWTC-3 measurement and the simulated spike-less measurements.

Analysis of the $\chi_{\rm eff}$ distribution allows us to circumvent these sampling issues by alternatively representing each event's posterior as a Gaussian kernel density estimate (KDE) over its posterior samples. This approach effectively imparts a finite "resolution" to each posterior sample, and allows us to assess the likelihood of arbitrarily narrow population features that would otherwise cause the typical Monte Carlo procedure to break down. Further details about the KDE representation of posteriors are provided in Appendix 12.E.

Motivated by Fig. 12.2, we attempt to measure the presence of any zero-spin sub-population by modeling the $\chi_{\rm eff}$ distribution as a mixture between a broad "bulk" population, centered at $\mu_{\rm eff}$ with width $\sigma_{\rm eff}$, and a narrow "spike"

centered at zero

$$p(\chi_{\rm eff}|\zeta_{\rm spike},\epsilon,\mu_{\rm eff},\sigma_{\rm eff}) = \zeta_{\rm spike}\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|0,\epsilon) + (1-\zeta_{\rm spike})\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|\mu_{\rm eff},\sigma_{\rm eff})\,. \tag{12.6}$$

We call this the `GaussianSpike` model; see App. 12.A for further details. Leveraging the KDE posterior representation introduced above, we take $\epsilon = 0$, such that the spike is infinitely narrow at $\chi_{\rm eff} = 0$. We hierarchically infer the parameters of this model using every binary black hole detection in GWTC-3 [287] with a false alarm rate below $1\,{\rm yr}^{-1}$; see Appendix 12.B for further details on the data we use.

The blue curve in Fig. 12.3 shows our resulting marginal posterior on the fraction $\zeta_{\rm spike}$ of binary black holes comprising a zero-spin spike. We find that $\zeta_{\rm spike}$ remains consistent with zero, indicating no evidence for an over-density of events at $\chi_{\rm eff} = 0$. For reference, the "zero-spin" fraction measured by Ref. [332] [ζ_0 in Eq. (12.3)] is shown as a dashed black curve. The results from Ref. [332] are qualitatively consistent with our own; exact agreement isn't expected due to the different models and different data employed in each analysis.

Although $\zeta_{\rm spike}$ is not bounded away from zero, a non-zero value is nevertheless preferred, with a maximum posterior value of $\zeta_{\rm spike} \approx 0.5$. *We demonstrate that this behavior is not unexpected from intrinsically spike-less populations.* We perform a series of null tests, repeatedly simulating and analyzing catalogs of events drawn from a spike-less population and with uncertainties typical of current detections; see Appendix 12.F for further details on the simulations. The grey curves in Fig. 12.3 show the posteriors on $\zeta_{\rm spike}$ given by these synthetic catalogs. Despite being drawn from a spike-less population, the simulated catalogs generically yield posteriors morphologically similar to the those obtained using actual observations, with some posteriors that even more extremely (and incorrectly) disfavor $\zeta_{\rm spike} = 0$.

The cause of this behavior is a degeneracy between $\mu_{\rm eff}$ and $\zeta_{\rm spike}$. The mock catalogs that strongly disfavor $\zeta = 0$ are typically those that have events with moderately high, well-measured effective spins. The presence of such events increases the inferred mean $\mu_{\rm eff}$ of the "bulk" population, pulling the bulk away from those events near $\chi_{\rm eff} \approx 0$ and leaving them to be absorbed into a zero-spin sub-population. We have verified that removing the most rapidly-spinning

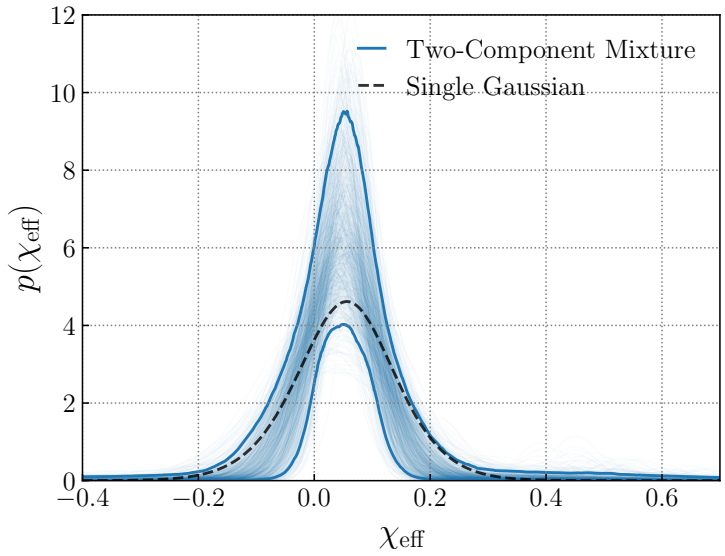


Figure 12.4: The $\chi_{\rm eff}$ distribution as inferred by a `BimodalGaussian` effective spin model, defined as the sum of two Gaussians (see Appendix 12.A). Solid blue lines denote the 90% credible intervals, while blue light curves are select individual draws from the posterior. For reference, the dashed black curve shows the mean $\chi_{\rm eff}$ distribution as inferred using a *single* Gaussian. Both results are consistent with one another, indicating no evidence for bimodal features, narrow or otherwise, in the binary black hole $\chi_{\rm eff}$ distribution.

events from these mock catalogs indeed acts to resolve the apparent tension in Fig. 12.3; see Appendix 12.F. This demonstration further emphasizes the need for additional caution when drawing strong astrophysical conclusions based on narrow population features, particularly in the regime when uncertainties on individual events are much larger than the features of interest.

Going beyond the question of a narrow spike at zero, we now more generally ask if there is evidence for *any* bimodality in the $\chi_{\rm eff}$ distribution of binary black holes. We explore this question using the `BimodalGaussian` model (see Appendix 12.A) in which the "spike" in Eq. (12.6) is replaced with a second, independent Gaussian with a variable mean and width. The $\chi_{\rm eff}$ distribution inferred under this model is shown in Fig. 12.4, where light traces show individual draws from our population posterior and the slide lines mark 90% credible bounds. We find no evidence that the effective spin distribution deviates from a simple unimodal shape. For reference, the dashed black line marks the mean distribution inferred using a simple `Gaussian` model. Inference using the more complex `BimodalGaussian` remains extremely consistent with this simple result, with both models yielding consistent means ($0.05^{+0.03}_{-0.03}$

and $0.06^{+0.04}_{-0.03}$ under the `Gaussian` and `BimodalGaussian` models, respectively) and standard deviations ($0.09^{+0.03}_{-0.04}$ and $0.12^{+0.04}_{-0.08}$).

An alternative way to test for the presence of additional structure in the $\chi_{\rm eff}$ distribution is to ask about the *predictive power* of our models: are there systematic residuals between the $\chi_{\rm eff}$ values we observe and those predicted by each model? In Appendix 12.C we subject the `Gaussian`, `GaussianSpike`, and `BimodalGaussian` models to predictive tests, finding that all three models, once fitted, successfully predict the observed $\chi_{\rm eff}$ values among GWTC-3. The fact that the simple `Gaussian` models passes this check once more points to the lack of observational evidence for additional structure in the binary black hole $\chi_{\rm eff}$ distribution, whether a spike, a secondary mode, or still other feature.

12.5 The population of spin magnitudes and tilts

Preliminary results so far, based on the Bayes factor counting experiment in Sec. 12.3 and hierarchical $\chi_{\rm eff}$ analyses in Sec. 12.4, do not point to an excess of zero-spin events. Here we confirm these conclusions via a more complete inference of the component spin magnitude and tilt distributions, under a series of increasingly complex models.

As a baseline model [282, 283, 351, 505], we take each component spin magnitude to be distributed following a Beta distribution,

$$p(\chi_i|\alpha,\beta) = \frac{\chi_i^{1-\alpha}\,(1-\chi_i)^{1-\beta}}{c(\alpha,\beta)}\,, \tag{12.7}$$

where $c(\alpha,\beta)$ is a normalization constant. Every spin tilt, meanwhile, is distributed as a mixture between an isotropic component and a preferentially-aligned component, modeled as a half-Gaussian centered at $\cos\theta = 1$:

$$p(\cos\theta_i|f_{\rm iso},\sigma_t) = \frac{f_{\rm iso}}{2} + (1-f_{\rm iso})\mathcal{N}_{[-1,1]}(\cos\theta_i|1,\sigma_t)\,. \tag{12.8}$$

We refer to Eqs. (12.7) and (12.8) as the `Beta+Mixture` model. A related version of this model in which both component spin tilts are together drawn from either the isotropic or the aligned component is also called the `Default` spin model in Ref. [282] and Ref. [283].

Ref. [165] explored an extension to the `Default` model that allowed for an excess of systems with zero spin and a variable bound on the spin tilt distribution; it was termed the `Extended` model in that study. Motivated by

these extensions, we introduce the possibility of a zero-spin "spike" in the spin magnitude distribution, modeled as a half-Gaussian with finite width $\epsilon_{\rm spike}$ centered at zero:

$$p(\chi_i|\alpha,\beta,f_{\rm spike},\epsilon_{\rm spike}) \\ = f_{\rm spike}\,\mathcal{N}_{[0,1]}(\chi_i|0,\epsilon_{\rm spike}) + (1-f_{\rm spike})\frac{\chi_i^{1-\alpha}\,(1-\chi_i)^{1-\beta}}{c(\alpha,\beta)}\,. \quad (12.9)$$

Following Ref. [165] we also introduce a variable truncation bound $z_{\rm min}$ on the spin-tilt distribution:

$$p(\cos\theta_i|f_{\rm iso},\sigma_t,z_{\rm min}) = \frac{f_{\rm iso}}{1-z_{\rm min}} + (1-f_{\rm iso})\mathcal{N}_{[z_{\rm min},1]}(\cos\theta_i|1,\sigma_t)\,, \quad (12.10)$$

for $z_{\rm min} \leq \cos\theta_i \leq 1$. We refer to Eqs. (12.9) and (12.10) together as the `BetaSpike+TruncatedMixture` model. To better understand how conclusions regarding a zero-spin excess and the prevalence of spin-orbit misalignment are related, we also consider variants of this model in which only one extended feature is present: a zero-spin spike but no $\cos\theta$ truncation (`BetaSpike+Mixture`) or a $\cos\theta$ truncation but no spike (`Beta+TruncatedMixture`). See Appendix 12.A for further details.

Our full `BetaSpike+TruncatedMixture` model differs from the `Extended` model of Ref. [165] in two ways. First, we do not fix the width $\epsilon_{\rm spike}$ of the vanishing spin subpopulation, but instead treat it as a free parameter to be inferred from the data. This allows us to *test* whether a narrow sub-population is actually preferred by the data, similar to our investigation with the `BimodalGaussian` model in Sec. 12.4 above. We adopt a prior requiring $\epsilon_{\rm spike} \geq 0.03$. This lower limit is motivated by tracking our number of effective posterior samples per event (see further discussion below), and by the effective population resolution of our catalog, which we estimate as

$$\frac{1}{\sigma^2_{\rm obs}} = \sum_{i=1}^{N_{\rm obs}}\left(\frac{1}{\sigma^2_{\chi_1,i}} + \frac{1}{\sigma^2_{\chi_2,i}}\right)\,, \quad (12.11)$$

where $\sigma^2_{\chi_1,i}$ and $\sigma^2_{\chi_2,i}$ are the variances of the spin magnitude posteriors for each event i. We find $\sigma_{\rm obs} = 0.02$, approximately equal to our lower bound on $\epsilon_{\rm spike}$.

Second, the `Extended` model does not allow for independently and identically distributed spin magnitudes and orientations. Instead, component spin magnitudes are either *both* vanishing or *both* non-vanishing in a given binary. This

Model	$f_{\rm spike}$	$z_{\rm min}$	$z_{1\%}$	$\chi_{1\%}$	$\chi_{99\%}$	$\chi_{\rm eff,1\%}$	$\chi_{\rm eff,99\%}$
`Beta+Mixture`	–	–	$-0.96^{+0.07}_{-0.02}$	$0.02^{+0.05}_{-0.02}$	$0.58^{+0.18}_{-0.16}$	$-0.20^{+0.10}_{-0.13}$	$0.27^{+0.12}_{-0.09}$
`BetaSpike+Mixture`	$0.22^{+0.31}_{-0.22}$	–	$-0.96^{+0.08}_{-0.02}$	$0.004^{+0.010}_{-0.003}$	$0.64^{+0.20}_{-0.19}$	$-0.21^{+0.10}_{-0.13}$	$0.30^{+0.13}_{-0.10}$
`Beta+TruncatedMixture`	–	$-0.55^{+0.19}_{-0.22}$	$-0.53^{+0.18}_{-0.21}$	$0.03^{+0.05}_{-0.03}$	$0.57^{+0.19}_{-0.14}$	$-0.10^{+0.06}_{-0.10}$	$0.27^{+0.12}_{-0.09}$
`BetaSpike+TruncatedMixture`	$0.33^{+0.27}_{-0.33}$	$-0.49^{+0.26}_{-0.26}$	$-0.47^{+0.27}_{-0.25}$	$0.003^{+0.008}_{-0.002}$	$0.67^{+0.20}_{-0.21}$	$-0.10^{+0.08}_{-0.11}$	$0.30^{+0.15}_{-0.10}$

Table 12.1: Median and 90% credible intervals on various physical quantities of interest with the component spin models. From left to right we present: the fraction of black holes that belong in the (finite-width) zero-spin spike ($f_{\rm spike}$; see Fig. 12.6 for posterior), the minimum value of the spin tilt ($z_{\rm min}$; Fig. 12.6 for posterior), 1% lower value of the spin tilt posterior distribution ($z_{1\%}$), the 1%/99% lower/upper values of the spin magnitude posterior distribution ($\chi_{1\%}/\chi_{99\%}$), and the 1%/99% lower/upper values of the effective spin distribution ($\chi_{\rm eff,1\%}/\chi_{\rm eff,99\%}$). We choose to report 1% and 99% values since $1/N_{\rm obs} = 0.014 \simeq 1\%$ for our $N_{\rm obs} = 69$ events.

choice precludes astrophysical scenarios such as tidal spin-up, which is expected to affect only one component spin in a given binary. Similarly, within the `Extended` model the spin tilts are both either members of the isotropic component or the preferentially aligned component in Eq. (12.10). Here, we instead assume that all component spin magnitudes and tilts are independently drawn from Eqs. (12.9) and (12.10).

When hierarchically analyzing the $\chi_{\rm eff}$ distribution above, we relied on a KDE representation of individual event likelihoods $p(d_i|\lambda_i)$ to mitigate sampling uncertainties and evaluate population models with narrow features. This trick cannot be straightforwardly applied to inference of the component spin distribution, due to both the increased dimensionality and the fact that the sharp feature of interest (a spike at $\chi_1 = \chi_2 = 0$) lies on the boundary of parameter space, rather than the center. We will therefore return to standard Monte Carlo averaging over posterior samples when hierarchically inferring the component spin distribution. To diagnose possible breakdowns in our inference due to finite sampling effects, we monitor the effective number of posterior samples, $N_{\rm eff}$, informing our Monte Carlo estimates for each event's likelihood. As discussed in Appendix 12.D, we explicitly track $\min\left[N_{{\rm eff},i}(\Lambda)\right]$, the minimum effective sample count across all events for a proposed population Λ, and use this quantity to define which regions of parameter space we can and cannot make claims about. In particular, we find that we expect reliable population inference when $\epsilon_{\rm spike} > 0.03$.

Figure 12.5 shows the measured spin magnitude and tilt distributions under our four component spin models, and we show the posteriors obtained on the

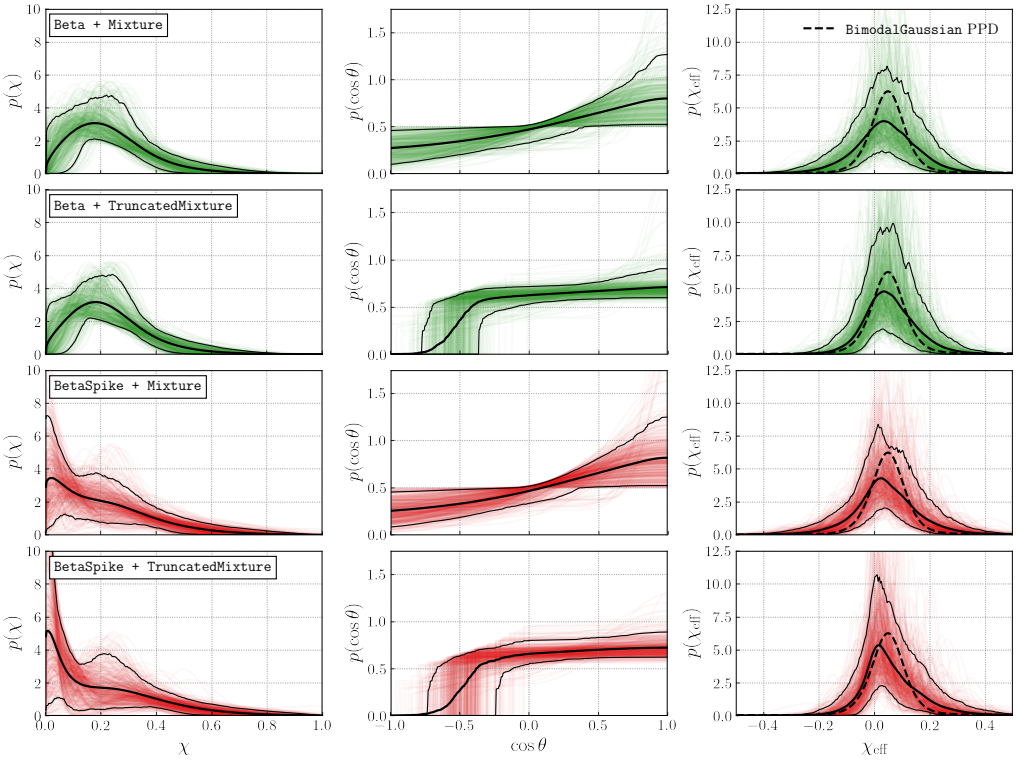


Figure 12.5: Inferred distributions of component spin magnitudes (left), spin-orbit tilts (middle), and effective inspiral spins (right) of binary black holes under the various component spin models considered in this paper. Solid black lines denote the median and the 90% credible intervals, while red/green light curves correspond to individual draws from the posterior for each model. Among models that allow for a distinct sub-population of non-spinning systems (bottom two rows), we infer that the fraction of binaries in such systems is consistent with zero. Meanwhile, we find that the $\cos\theta$ distribution confidently extends to negative values; models that allow for a sharp truncation in the $\cos\theta$ distribution require this truncation to occur at $z_{\rm min} < 0$. Despite the varying component spin magnitude and tilt distributions recovered by these different models, all yield similar $\chi_{\rm eff}$ distributions. For reference, the dashed black curves in the right-hand column show the mean $\chi_{\rm eff}$ distribution obtained by direct inference with the `BimodalGaussian` model of Sec. 12.4; all four component spin models give $\chi_{\rm eff}$ distributions that are consistent with this result.

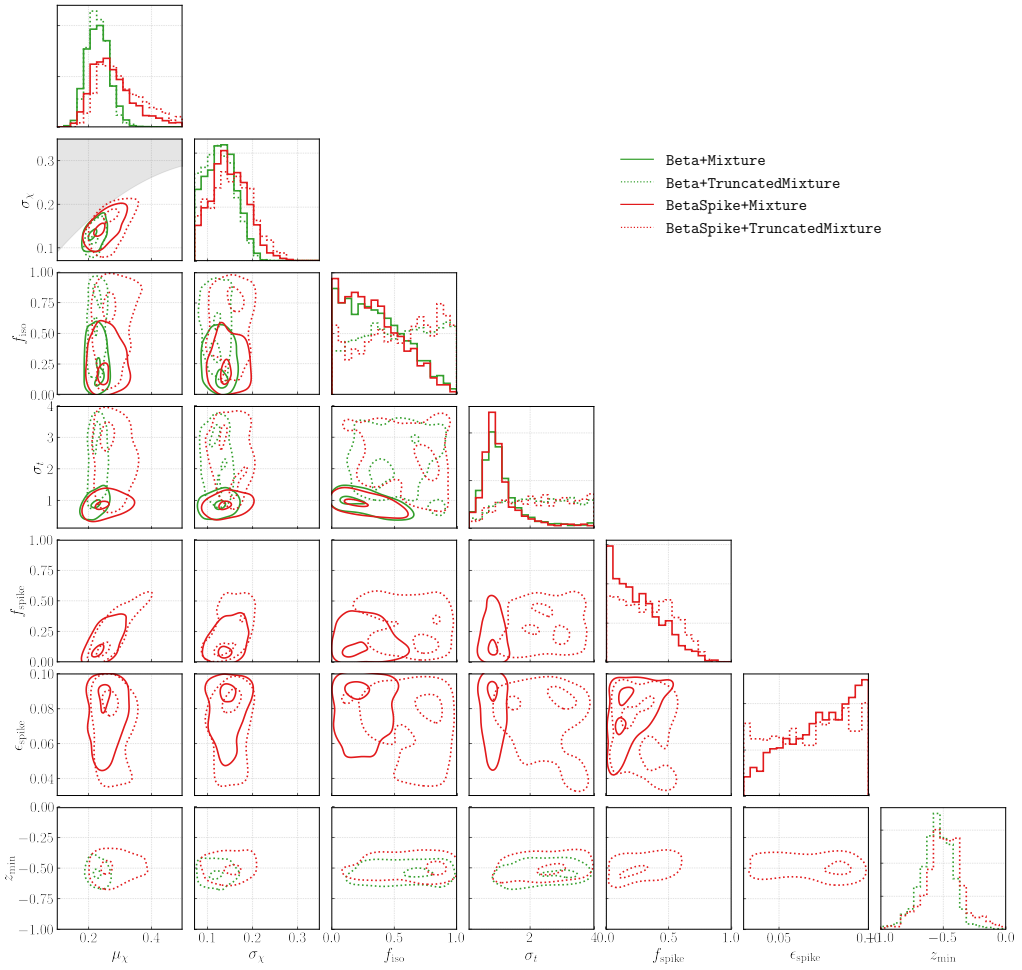


Figure 12.6: One- and two-dimensional marginalized posteriors for the parameters of each component spin model we consider (see Appendix 12.A). These posteriors correspond to the measured distributions of component spin magnitudes and tilts shown above in Fig. 12.5. Some parameters are defined only for a subset of component spin models. The shaded region in the joint μ_χ–σ_χ posterior is the region excluded by the prior cut on the shape parameters of the spin magnitude Beta distribution, see Appendix 12.A. We find that the fraction $f_{\rm spike}$ of black holes comprising a zero-spin sub-population is consistent with zero. The data also require that at least some spins are misaligned by more than 90° relative to their Newtonian orbital angular momentum; among models that include a variable truncation bound $z_{\rm min}$ on the $\cos\theta$ distribution, this bound is inferred to be confidently ≤ 0, regardless of assumptions about a possible zero-spin sub-population. Allowing such a truncation bound, meanwhile, significantly impacts constraints on $f_{\rm iso}$, the fraction of binaries with isotropically-oriented spins, as well as σ_t, the width of a preferentially spin-aligned mixture component (see Eq. (12.8)). As seen in Fig. 12.5, introduction of a truncation bound yields a significantly flatter $p(\cos\theta)$ distribution, corresponding here to larger values of $f_{\rm iso}$ and σ_t.

hyperparameters of each model in Fig. 12.6. To hierarchically measure the component spin distribution, we again use all binary black holes in GWTC-3 with false alarm rates below $1\,\mathrm{yr}^{-1}$; see Appendix 12.B for details.

For those models allowing a distinct zero-spin spike, the left panels of Fig. 12.5 indicate that such a feature is not required to exist. We instead see inferred spin magnitude distributions consistent with a single, smooth function that remains finite at $\chi = 0$ and most events have spin magnitudes in the $\chi \in (0, 0.3)$ range. Correspondingly, the posteriors on f_{spike} in Fig. 12.6 are consistent with zero. Compared to the fraction ζ_{spike} with $\chi_{\mathrm{eff}} = 0$ (see Fig. 12.3), we find that f_{spike} is more consistent with zero. We interpret this variation as reflective of the systematic uncertainty in exactly how the question "*Does there exist an excess of zero-spin systems?*" is posed: whether in the component spin or effective spin domains, with a delta-function at zero or a finite-width spike, etc. Despite these differences, all results indicate that the presence of a zero-spin population is not required by the current data. A zero-spin population is not *precluded*, though: both sets of results bound any zero-spin fraction to $\lesssim 60\%$, suggesting that a distinct zero-spin population could yet emerge in future data.

The fraction f_{spike} is primarily correlated with μ_χ, the mean of the "bulk" spin magnitude population. Larger μ_χ values reduce the capability of this "bulk" Beta distribution to accommodate events with small spin; these events will therefore necessarily be assigned to the "spike" and thus increase the value of f_{spike}. This is similar to the phenomenon identified in Sec. 12.4 above. Additionally, all four component spin models yield similar spin magnitude distributions above $\chi \gtrsim 0.4$, suggesting that the data are robustly consistent with the absence of large spin magnitudes. Table 12.1 lists the 1st and 99th spin magnitude percentiles ($\chi_{1\%}$ and $\chi_{99\%}$ respectively) under each model; our most conservative estimate gives $\chi_{99\%} = 0.67^{+0.20}_{-0.21}$.

In contrast to Ref. [165], we left the width $\epsilon_{\mathrm{spike}}$ of our "spike" population as a free parameter in order to test whether the data indeed require a narrow feature at $\chi = 0$. Although we recover largely uninformative constraints on $\epsilon_{\mathrm{spike}}$, Fig. 12.6 in fact shows a slight preference for *large* $\epsilon_{\mathrm{spike}}$, further indicating that no narrow features are confidently present in the spin magnitude distribution. Our conclusions regarding small $\epsilon_{\mathrm{spike}}$, however, are limited by the finite sampling effects discussed above. In Appendix 12.D we study the

number $N_{\rm eff}$ of effective posterior samples as a function of $\epsilon_{\rm spike}$. We find that $N_{\rm eff}$ depends sensitively on $\epsilon_{\rm spike}$, and we caution that values of $\epsilon_{\rm spike} \lesssim 0.04$ are possibly subject to significant Monte Carlo uncertainty. Even with this restriction, however, we find no preference for a small $\epsilon_{\rm spike}$ in the posterior. This conclusion is further bolstered by the analysis of the effective spin distribution in Sec. 12.4 above, which was *not* subject to finite sampling effects and which similarly found no evidence for an excess of zero-spin systems.

Irrespective of modeling assumptions regarding a zero-spin sub-population, all four component spin models exhibit significant support at negative $\cos\theta$ in Fig. 12.5. Models that allow for a truncation in the spin tilt distribution consistently infer this truncation to be at negative values; in Fig. 12.6 we correspondingly see that $z_{\rm min} \leq 0$ at high significance. This result signals the presence of events with spins misaligned by more than 90° from their Newtonian orbital angular momentum. Table 12.1 gives the 1st percentile ($z_{1\%}$) on $\cos\theta$ inferred by each model, with $z_{1\%} = -0.47^{+0.27}_{-0.25}$ in the most conservative case. Notably, the addition of a truncation in the spin tilt distribution significantly "flattens" the recovered $\cos\theta$ distribution, no longer requiring any peak at $\cos\theta = 1$. This behavior is due to the fact that the data disfavor an equal density of events at $\cos\theta = +1$ and $\cos\theta = -1$. Since an isotropic distribution is necessarily symmetric at $\cos\theta = \pm 1$, models that allow for a mixture of isotropic and aligned spin tilts must compensate by adding more weight to the "primarily-aligned" Gaussian component in Eq. (12.8). No such compensation is necessary if the spin tilt distribution is allowed to truncate. In models including a $z_{\rm min}$ truncation, all events are either assigned to the "isotropic" (but now truncated) component, or the Gaussian component itself is stretched to near-isotropy; in Fig. 12.6 we see that $f_{\rm iso}$ becomes consistent with 1 and σ_t shifts to prefer large values when a truncation is included in the model.

This result signals the presence of events with tilt angles greater than 90° and hence the possibility of binaries with effective spins $\chi_{\rm eff} < 0$. The third column in Fig. 12.5 shows the $\chi_{\rm eff}$ distributions implied by each component model. Despite the wide range of possible features possessed by each model, they all yield similar $\chi_{\rm eff}$ distributions that exhibit asymmetry about zero but that extend to negative values. As a consistency check, we also compare these implied $\chi_{\rm eff}$ distributions to our result obtained through direct inference

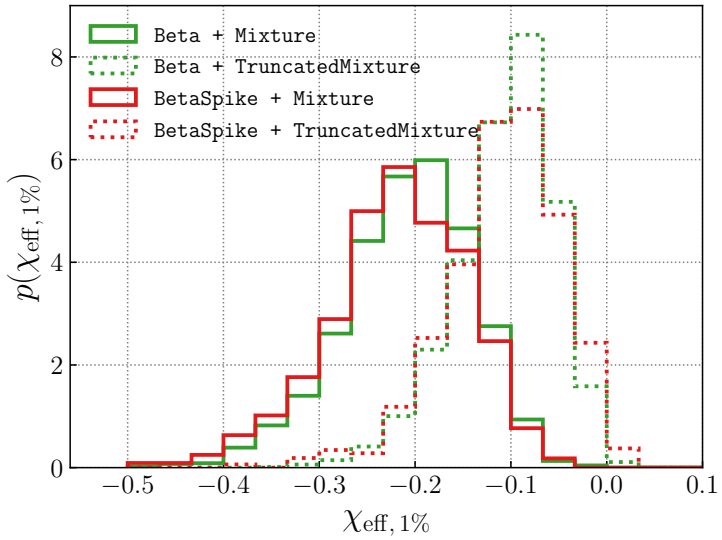


Figure 12.7: Posterior for the 1st percentile ($\chi_{\mathrm{eff},1\%}$) of the χ_{eff} distributions implied by each of our component spin models, as shown in the right-hand column of Fig. 12.5. The median values and 90% credible intervals for $\chi_{\mathrm{eff},1\%}$ are also reported in Table 12.1. Each of our component spin models require the binary black hole population to contain spins misaligned by more than 90° relative to their Newtonian orbital angular momentum. Correspondingly, $\chi_{\mathrm{eff},1\%}$ is inferred to likely (although not *necessarily*) be negative. The $\chi_{\mathrm{eff},1\%}$ values recovered here are consistent with the minimum χ_{eff} values measured by Ref. [282, 283].

on χ_{eff}. The dashed curve in each panel shows the mean obtained under the `BimodalGaussian` effective spin model. Each of the four component spin models yields consistent χ_{eff} distributions, although they are suggestive of possible additional skewness [282, 283]. As an additional consistency check and safeguard against erroneous conclusions due to poorly-fitting models, we subject each of the component spin models to predictive checks, as introduced in Sect. 12.4 above. The results, shown in Appendix 12.C, indicate that all four models provide a good fit to observation.

As a proxy for the minimum χ_{eff} values implied by our component spin results, Fig. 12.7 shows posteriors on the 1st percentile ($\chi_{\mathrm{eff},1\%}$) of the effective spin distribution corresponding to each model. Under the `BetaSpike+Truncated Mixture` model, for example, we find $\chi_{\mathrm{eff},1\%} = -0.10^{+0.08}_{-0.11}$, and bound $\chi_{\mathrm{eff},1\%} < 0$ at 98.8% credibility. This finding is consistent with the conclusions drawn by Ref. [282] and Ref. [283], who added a lower truncation in the χ_{eff} distribution and inferred $\chi_{\mathrm{eff,min}} < 0$ at $\sim 90\%$ credibility. Our conclusions are

also consistent with Ref. [332], who modeled the $\chi_{\rm eff}$ distribution as the sum of a positive component, a negative, and a "near-zero" component whose 2σ width spanned $-0.08 < \chi_{\rm eff} < 0.08$; see Eq. (12.3). This near-zero component is broad enough to likely encompass the small but negative $\chi_{\rm eff,1\%}$ values we report here.

Finally, Figs. 12.5 and 12.6 reveal how questions regarding zero-spin sub-populations and spin-orbit misalignment are potentially related. The inclusion or exclusion of a zero-spin sub-population (`Beta+TruncatedMixture` vs. `BetaSpike+TruncatedMixture`) has a negligible effect on our conclusions regarding the location of $z_{\rm min}$. On the other hand, introduction of a spin tilt truncation (`BetaSpike+Mixture` vs. `BetaSpike+TruncatedMixture`) more noticeably impacts the inferred $f_{\rm spike}$ posterior, resulting in less support for zero. As discussed above, though, the spin tilt truncation has a much larger effect on $f_{\rm iso}$, which becomes consistent with 1 (i.e., no excess of spin-aligned events) when a spin tilt truncation is included in the model. This same effect can be seen in Fig. 4 of [165].

12.6 Discussion and Conclusions

In this paper we have sought to explore the three open questions posed in Sec. 12.2 via three complementary routes: a counting experiment using only the fully-marginalized likelihoods for each event, hierarchical analysis of the binary black hole effective spin distribution, and hierarchical analysis of the binaries' component spin distribution. Each of these three routes yielded consistent conclusions, which we summarize below.

i. ***We find no evidence for an excess of zero-spin events.*** Using each of our three methods we find that the fraction of black holes in a distinct zero-spin sub-population is consistent with zero. We furthermore verified that this behavior is common among analyses of synthetic populations lacking a zero-spin excess. We therefore conclude that the observational data do not presently support the notion that the majority of black holes in merging binaries are non-spinning. Given current observations, a non-spinning population can comprise at most $\lesssim 60-70\%$ of merging binary black holes.

ii. ***The inferred population is consistent with less than 1% of spin magnitudes above*** $\chi \sim 0.6$**.** In each of the component spin models explored in Sec. 12.5, the inferred population is nearly identical for spin magnitudes

$\chi \gtrsim 0.4$, falling rapidly, with negligible support for $\chi \gtrsim 0.6$. This conclusion is robust against modeling systematics, including a possible truncation in the spin tilt distribution or a zero-spin sub-population.

iii. ***Binary black holes exhibit a broad range of spin-orbit misalignment angles, with some angles greater than*** $\mathbf{90^\circ}$***.*** In all models, we find a preference for a flat and broad distribution of spin-orbit misalignment angles. When we introduce a lower truncation bound on the $\cos\theta$ distribution, we confidently infer that this truncation bound must be negative, thus requiring the presence of systems with anti-aligned spins among the binary black hole population. Additionally, the minimum $\chi_{\rm eff}$ among the binary black hole population is inferred to be confidently negative under each component spin model.

Our conclusions are consistent with the findings of Ref. [282] and Ref. [283], who reported evidence for spins misaligned by more than 90° and no modeling tension that would indicate the existence of a large zero-spin sub-population. Our conclusions are moreover in broad agreement with Ref. [332]. Figure 12.3 shows the inferred fraction of events with $\chi_{\rm eff} = 0$ from this work and the inferred fraction of events in the "zero-spin" sub-population [ζ_0 in Eq. (12.3)] from Ref. [332]. Both results are in qualitative agreement, consistent with a zero-spin fraction of zero but peaking at ~ 0.5. As demonstrated in Sec. 12.4, posteriors of this form arise generically when analyzing mock catalogs of events drawn from a population lacking a zero-spin sub-population.

Ref. [332] find that the fraction of events in their "negative-spin" sub-population is consistent with zero; this too is consistent with our result. As discussed in Sec. 12.5, we infer $\chi_{\rm eff,1\%}$ to be negative but small in magnitude. Because Ref. [332]'s "zero-spin" sub-population has a broad standard deviation of 0.04, events with negative but small-in-magnitude $\chi_{\rm eff}$ are counted as members of this zero-spin sub-population, rather than associated with the "negative-spin" sub-population. Indeed, Ref. [332] conclude that the *sum* of their "zero-spin" and "negative-spin" sub-populations is non-zero. This is evident from Fig. 3 of Ref. [332] where $\zeta_{\rm pos}$ is inconsistent with 1.

Shortly after the initial circulation of our study, an independent and complementary study of binary black hole spins was presented by Ref. [515]. They sought to explore evidence for mass ratio reversal in compact binary formation, employing models in which component spins are *not* independently and

identically distributed. As in our study, though, they additionally considered the possibility of zero-spin spikes appearing in the spin magnitude distribution. Their findings corroborate our own, indicating that $< 46\%$ of black hole primaries and $< 36\%$ of secondaries have vanishing spins.

At the same time, our conclusions are generally inconsistent with those reported by Ref. [165]. Although our posterior distributions do have overlap with those of Ref. [165] to within statistical uncertainty, differences between them result in *qualitatively different conclusions regarding the nature of binary black hole spins*. Below, we detail several differences between our analyses and those of Ref. [165] and comment on whether each could be contributing to the discrepancy.

First, an important categorical difference between our analyses and those conducted in Ref. [165] is the sample of gravitational-wave events analyzed. While our counting experiment in Sec. 12.3 uses only GWTC-2 [286] events, the hierarchical analyses in Sec. 12.4 and Sec. 12.5 make use of the full GWTC-3 catalog (see Appendix 12.B for the exact event selection criteria). Ref. [165], meanwhile, analyzed only GWTC-2 events, as GWTC-3 had not yet been released at the time of their study. To gauge whether our differing data sets contribute to the disagreement between our own conclusions and those of Ref. [165], in Appendix 12.H we show results obtained using only GWTC-2 events. Our results are qualitatively unchanged, with GWTC-2 yielding no evidence for a distinct sub-population of non-spinning black holes. Additionally, GWTC-2 contains spin-orbit misalignment greater than 90°, though at a slightly reduced credibility compared to GWTC-3 (97.1% vs. 99.7% quantiles).

Unlike our analyses, which use a single set of posterior samples for each event, Ref. [165] make use of two distinct sets of samples for each binary, obtained under spinning and non-spinning parameter estimation priors. In order to perform inference with both sets of samples, it is necessary to quantify the Bayes factors between these priors for each event. As we mentioned in Sec. 12.3 and illustrate further in Appendix 12.G, there is at least one event whose fully-marginalized likelihood under the non-spinning hypothesis is significantly overestimated in the data set used by Ref. [165]. This could cause Ref. [165] to spuriously confirm the existence of a distinct non-spinning sub-population.

Another factor may be the demand by Ref. [165] that the zero-spin subpopulation have vanishing width. In Sec. 12.5, we did not fix the width $\epsilon_{\rm spike}$

of our approximate zero-spin spike, but let it vary as another free parameter to be inferred from the data. In addition to concluding that the fraction $f_{\rm spike}$ of events occupying this spike is consistent with zero, we also found no preference for the spike to be narrow, even if it *were* to exist. If we nevertheless *require* $\epsilon_{\rm spike}$ to be small, however, we do see an increased preference for larger $f_{\rm spike}$. It is therefore possible that the strict delta-function model adopted by Ref. [165] is systematically affecting their results. We note, however, that our posteriors in the region $\epsilon_{\rm spike} \leq 0.04$ may be subject to elevated Monte Carlo variance (see Appendix 12.D) and so we cannot confidently conclude that the demand of $\epsilon_{\rm spike} = 0$ is responsible for the discrepant conclusions.

Although we may be able to reproduce Ref. [165]'s measurement of $f_{\rm spike}$ via artificially demanding small $\epsilon_{\rm spike}$, we are unable to reproduce their conclusions regarding $z_{\rm min}$, the lower truncation bound on the component spin $\cos\theta$ distribution. While we infer $z_{\rm min}$ to be confidently negative, indicating spins misaligned by more than 90°, Ref. [165] infer this parameter to be consistent with zero or positive values. Qualitatively, binaries with large in-plane spins are difficult to distinguish from binaries with very small or vanishing aligned spins – both give the same $\chi_{\rm eff}$ values. Therefore, if Ref. [165] are overestimating the fraction $f_{\rm spike}$ of non-spinning systems, they may correspondingly be *underestimating* the prevalence of systems with significant spin-orbit misalignments. This said, neither Ref. [165]'s results nor our own exhibit any significant correlation between $z_{\rm min}$ and $f_{\rm spike}$.

Finally, although our spin models were heavily inspired by those of Ref. [165], ours *do not* reduce to theirs in the limit $\epsilon_{\rm spike} = 0$. Our component spin models assume that individual spins are independently and identically distributed between spike and bulk sub-populations, whereas Ref. [165] assume that *both* spins in a given binary together lie in the spike or the bulk of the component spin distribution. Similarly, Ref. [165] require that both spins are together members of the isotropic or preferentially-aligned components of the spin-tilt distribution. We have verified that this modeling choice is not responsible for the different conclusions regarding $f_{\rm spike}$, as can be seen in Fig. 12.17 in Appendix 12.H. In fact, we find that $f_{\rm spike}$ is constrained to even *smaller* values if we alternatively adopt Ref. [165]'s modeling convention. This is expected: under the convention of Ref. [165], a component spin can only contribute to $f_{\rm spike}$ if its companion is *also* consistent with zero spin. Under our assump-

tion of independent and identically distributed spins, in contrast, a component spin can be identified as a member of the zero-spin spike regardless of its companion's spin measurement. Such differing conventions and how they impact population measurements are further elaborated upon in Appendix 12.H.

Overall, we find a component spin distribution that is consistent with a single, smooth function. Our results suggest that the spin distribution is possibly non-zero at $\chi = 0$, but without a requirement for a distinct and discontinuous sub-population of non-spinning black holes (see Fig. 12.5). Such a behavior cannot be easily captured by the common modeling choice of a non-singular Beta distribution, which is constrained to vanish at $\chi = 0$. This tension is illustrated in the joint μ_χ-σ_χ posterior of Fig. 12.6, which shows the inferred mean and variance of the spin magnitude distribution railing against the prior cut that ensures non-singular behavior. When the spin magnitudes are modeled with a single Beta distribution (green posteriors in Fig. 12.5), μ_χ must be small in order to accommodate events with small but finite spins. Events with larger spins, meanwhile, can nominally be captured with a large σ_χ. However, σ_χ *cannot* be large when μ_χ is small, as seen by the prior cut in Fig. 12.6. The introduction of a zero-spin "spike" relieves this tension (even if no such spike is actually present in the data). Small-spin events can now be captured by the spike, allowing the Beta distribution to shift to higher μ_χ and σ_χ to accommodate higher-spin events. This phenomenon can also be seen in the correlation, noted above, between $f_{\rm spike}$ and μ_χ: the spin magnitude distribution is best described either as a single Beta distribution that peaks at low values (small $f_{\rm spike}$ and μ_χ), or a "spike" combined with a Beta distribution that peaks at higher values (large $f_{\rm spike}$ and μ_χ). The combination of these two effects yields a smooth spin magnitude distribution that does not exhibit distinct sub-populations. This discussion suggests that a Beta distribution might be a sub-optimal model for spin magnitudes going forward.

The observed presence or absence of a distinct zero-spin sub-population, rapidly spinning black holes, and/or significantly misaligned black hole spins would all carry considerable theoretical implications for the formation channels and astrophysical processes driving compact binary evolution. At the same time, hierarchical inference of the compact binary spin distribution is technically difficult, relying on a large number of highly uncertain measurements that are themselves finitely sampled. When drawing observational conclusions about

the nature of compact binary spins, ensuring that results are reproducible under different complementary approaches can help mitigate these concerns and determine which conclusions are driven by priors, which by modeling systematics, and which by the data itself.

It certainly remains *possible* that the binary black hole mergers witnessed by Advanced LIGO and Virgo contain sub-populations of non-spinning and/or nearly-aligned binaries, but this scenario is not presently required by GW observations. If such sub-populations were to appear in an analysis of future data, they could be taken as evidence towards the hypothesis that merging black holes are born in the stellar field, with spins acquired primarily through tidal spin-up [1, 88, 165, 516–518]. Such astrophysical conclusions are currently unsupported by the data, however. Further detections made in the upcoming O4 observation run and beyond will shine further light on the nature of compact binary spins and, consequently, the evolutionary origin of the black hole mergers we see today.

Ch. 12 Data and Code Availability

Code used to produce the results of this study is available via Github at `github.com/tcallister/gwtc3-spin-studies`. Our data, meanwhile, can be obtained on Zenodo at `doi.org/10.5281/zenodo.6555145`.

Ch. 12 Acknowledgments

We thank our anonymous referee, the AAS Data Editors, and Ilya Mandel for their valuable comments on our manuscript, as well as Eric Thrane, Colm Talbot, and Shanika Galaudage for numerous discussions about these results. We also thank Javier Roulet for sharing data from Ref. [332] with us and for insightful feedback on this study, and Christopher Berry, Charlie Hoy, Vaibhav Tiwari, and Mike Zevin for their thoughtful comments.

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 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. This material is based upon work supported by NSF's LIGO Laboratory which

is a major facility fully funded by the National Science Foundation. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by NSF Grants PHY-0757058 and PHY-0823459.

Software: `astropy` [355, 356], `emcee` [252], `h5py` [402], `jax` [258], `matplotlib` [320], `numpy` [323], `numpyro` [259, 260], `scipy` [353].

Ch. 12 Appendices

12.A Spin population models

We employ two broad categories of parametrized models for the black hole spins: models on the effective spin parameters and models on the spin components. A summary of all models, their corresponding parameters, and their priors is presented in Tables 12.2 and 12.3. All component spin models ignore the azimuthal spin angles, assuming they are distributed according to the uniform prior used during the original parameter estimation [286]. As measurements of compact binary spins and mass ratios are generally correlated, we hierarchically measure the distribution of binary mass ratios alongside spins, assuming that the secondary mass distribution follows

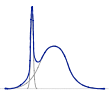
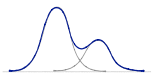
$$p(m_2|m_1) \propto m_2^{\beta_q} \quad (m_{\rm min} \leq m_2 \leq m_1) \ , \tag{12.12}$$

and inferring the power-law index β_q. We adopt a Gaussian prior $N(0,3)$ on β_q. The distributions of primary masses and redshifts, in contrast, have a negligible impact on conclusions regarding spin. We assume that primary masses follow the `PowerLaw+Peak` model [357], with parameters fixed to their (one-dimensional) median values as inferred in Ref. [283]. Following the notation of Ref. [283], we take $\alpha = 3.5$, $m_{\rm min} = 5.0$, $m_{\rm max} = 88.2$, $\lambda_{\rm peak} = 0.03$, $\mu_m = 33.6$, $\sigma_m = 4.7$, and $\delta_m = 4.9$. Meanwhile, we assume that the source-frame binary black hole merger density rate evolves as

$$R(z) \propto \frac{dV_c}{dz} (1+z)^{2.7} \ , \tag{12.13}$$

where $\frac{dV_c}{dz}$ is the differential comoving volume per unit redshift.

Table 12.2: Summary of effective spin models we employ, including their names, an example $p(\chi_{\rm eff})$ plot, parameters, priors, and comments. $\mathrm{U}(x_{\rm min}, x_{\rm max})$ denotes a uniform prior between $x_{\rm min}$ and $x_{\rm max}$, $\mathrm{LU}(x_{\rm min}, x_{\rm max})$ signifies a log-uniform prior across the same bounds, and $\delta(0)$ a delta-function prior at zero.

Model Name	$p(\chi_{\rm eff})$	Parameter	Prior	Comments
Gaussian		$\mu_{\rm eff}$ $\sigma_{\rm eff}$	U(-1,1) LU(0.01,1)	Simplest Gaussian model
GaussianSpike		$\mu_{\rm eff}$ $\sigma_{\rm eff}$ $\zeta_{\rm spike}$ ϵ	U(-1,1) LU(0.01,1) U(0,1) $\delta(0)$	Mixture of a Gaussian "bulk" with a distinct "spike" sub-population of non-spinning systems
BimodalGaussian		$\mu_{{\rm eff},a}$ $\mu_{{\rm eff},b}$ $\sigma_{{\rm eff},a}$ $\sigma_{{\rm eff},b}$ ζ_a	U(-1,1) $\mathrm{U}(-1,1)$ LU(0.01,1) LU(0.01,1) U(0.5,1)	Mixture of two Gaussians with variable means and widths

Effective spin models

In this subsection, we list the various models considered when exploring the distribution of effective inspiral spins among binary black hole mergers. See Table 12.2 for illustrations of each model, as well as the priors adopted for each parameter.

`Gaussian`. Our simplest model assumes that effective inspiral spins as Gaussian-distributed with mean $\mu_{\rm eff}$ and variance $\sigma^2_{\rm eff}$,

$$p(\chi_{\rm eff}|\mu_{\rm eff}, \sigma_{\rm eff}) = \mathcal{N}_{[-1,1]}(\chi_{\rm eff}|\mu_{\rm eff}, \sigma_{\rm eff})\,, \tag{12.14}$$

where $\mathcal{N}_{[a,b]}$ denotes a Gaussian distribution truncated within $[a, b]$. Our priors on $\mu_{\rm eff}$ and $\sigma_{\rm eff}$ are listed in Table 12.2. This model was proposed in Refs. [504] and [162] and has been employed and extended in Refs. [282, 283, 365, 519, 520].

`GaussianSpike`. To initially assess the prevalence of identically non-spinning black holes, we extend the `Gaussian` model to include a narrow "spike" centered at $\chi_{\rm eff} = 0$ with width $\epsilon \ll 1$,

$$\begin{aligned} &p(\chi_{\rm eff}|\zeta_{\rm spike}, \epsilon, \mu_{\rm eff}, \sigma_{\rm eff}) \\ &\qquad = \zeta_{\rm spike}\,\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|0, \epsilon) + (1-\zeta_{\rm spike})\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|\mu_{\rm eff}, \sigma_{\rm eff})\,. \end{aligned} \tag{12.15}$$

The fraction of binary black holes comprising the zero-spin spike is given by $\zeta_{\rm spike}$. A similar model was studied by Ref. [283], who imparted a finite width

to the zero-spin spike population. In this work, we fix $\epsilon = 0$ following the modeling choice of Ref. [165], leveraging the kernel density estimation trick discussed in Appendix 12.E to accurately and stably evaluate the likelihood for this delta-function spike.

`BimodalGaussian`. Given the inferred absence of a zero-width spike with `GaussianSpike`, we introduce this model to more generally assess *any* potential multimodality. We model the distribution of $\chi_{\rm eff}$ as a mixture of two Gaussians with arbitrary means and standard deviations

$$p(\chi_{\rm eff}|\zeta_a, \mu_{{\rm eff},a}, \sigma_{{\rm eff},a}, \mu_{{\rm eff},b}, \sigma_{{\rm eff},b}) = \\ \zeta_a \mathcal{N}_{[-1,1]}(\chi_{\rm eff}|\mu_{{\rm eff},a}, \sigma_{{\rm eff},a}) + (1-\zeta_a)\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|\mu_{{\rm eff},b}, \sigma_{{\rm eff},b})\,. \quad (12.16)$$

The mixing fraction ζ_a is constrained to be ≥ 0.5, solving the "label-switching" problem by demanding that $\mu_{{\rm eff},a}$ and $\sigma_{{\rm eff},a}$ are the mean and standard deviation of the dominant sub-component. No further prior constraints are placed on $\mu_{{\rm eff},a}$ and $\sigma_{{\rm eff},a}$ or $\mu_{{\rm eff},b}$ and $\sigma_{{\rm eff},b}$.

Component spin models

We next list the various models used to study the distribution of component spin magnitudes and orientations among binary black holes. See Table 12.3 for illustrations and priors.

`Beta+Mixture`. In our simplest component spin model, we assume that spin magnitudes are independently draw from identical Beta distributions,

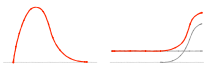
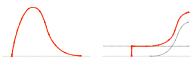
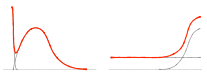
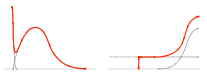
$$p(\chi_i|\alpha,\beta) = \frac{\chi_i^{\alpha-1}(1-\chi_i)^{\beta-1}}{c(\alpha,\beta)}, \quad (12.17)$$

where $c(\alpha,\beta)$ normalizes the distribution to unity. The two shape parameters α and β are constrained such that $\alpha,\beta > 1$ to ensure that the distribution is bounded. This choice of parameters *requires* that $p(\chi_i = 0) = p(\chi_i = 1) = 0$, thus *a priori* assuming that there are no systems with exactly vanishing spin. We sample not in α and β directly but rather in the mean μ_χ and standard deviation σ_χ of the Beta distribution. The shape parameters α and β are calculated from μ_χ and σ_χ through

$$\alpha = \mu_\chi \nu\,, \quad \beta = (1-\mu_\chi)\nu, \quad \nu = \frac{\mu_\chi(1-\mu_\chi)}{\sigma_\chi^2} - 1\,. \quad (12.18)$$

The region of the μ_χ and σ_χ parameter space excluded by restriction $\alpha,\beta > 1$ can be seen in Figure 12.6. The spin tilt distribution is a mixture between

Table 12.3: Summary of component spin models we employ, including their names, an example χ and $\cos\theta$ plot, parameters, priors, and comments. $\mathrm{U}(x_{\rm min}, x_{\rm max})$ denotes a uniform prior between $x_{\rm min}$ and $x_{\rm max}$, while $\mathrm{LU}(x_{\rm min}, x_{\rm max})$ signifies a log-uniform prior.

Model Name	$p(\chi)$	$p(\cos\theta)$	Parameter	Prior	Comments
Beta+Mixture			μ_χ σ_χ σ_t $f_{\rm iso}$	U(0,1) U(0.07,0.5) U(0.1,4) U(0,1)	Simplest component spin model; additional prior requirement that $\alpha, \beta > 1$, see Eq. (12.18)
Beta+TruncatedMixture			μ_χ σ_χ σ_t $f_{\rm iso}$ $z_{\rm min}$	U(0,1) U(0.07,0.5) U(0.1, 4) U(0,1) U(-1,1)	Introduces a truncation in the $\cos\theta$ distribution at some minimum value (maximum misalignment angle)
BetaSpike+Mixture			μ_χ σ_χ $f_{\rm spike}$ $\epsilon_{\rm spike}$ σ_t $f_{\rm iso}$	U(0,1) U(0.07,0.5) U(0,1) U(0.03, 0.1) U(0.1,4) U(0,1)	Spin magnitude distribution modified to include Beta-distribution "bulk" and a finite-width "spike" at approximately zero spin
BetaSpike+TruncatedMixture			μ_χ σ_χ $f_{\rm spike}$ $\epsilon_{\rm spike}$ σ_t $f_{\rm iso}$ $z_{\rm min}$	U(0,1) U(0.07,0.5) U(0,1) U(0.03, 0.1) U(0.1, 4) U(0,1) U(-1,1)	Most complex component spin model; includes both the multiple two spin magnitude sub-populations and the truncation in the spin tilt distribution

two components: a uniform isotropic component and a preferentially-aligned component

$$p(\cos\theta_i|f_{\rm iso},\sigma_t) = \frac{f_{\rm iso}}{2} + (1-f_{\rm iso})\mathcal{N}_{[-1,1]}(\cos\theta_i|1,\sigma_t)\,. \tag{12.19}$$

Here, $f_{\rm iso}$ is the fraction of events comprising the isotropic subpopulation, while the second term is a half-Gaussian peaking at $\cos\theta = 1$. Each component spin tilt is independently drawn from Eq. (12.19). A model of this form was introduced in Refs. [351] and [505], with the restriction that both spin tilts together lie in either the isotropic or the preferentially-aligned component, and is referred to as the `Default` model in Refs. [165, 282, 283].

`BetaSpike+Mixture`. In order to assess the presence of non-spinning black hole binaries, we emulate Ref. [165] and extend the `Beta+Mixture` model to include a half-Gaussian "spike" that peaks at $\chi_i = 0$ but has a finite width $\epsilon_{\rm spike}$:

$$p(\chi_i|\alpha,\beta,f_{\rm spike},\epsilon_{\rm spike}) = \\ f_{\rm spike}\mathcal{N}_{[0,1]}(\chi_i|0,\epsilon_{\rm spike}) + (1-f_{\rm spike})\frac{\chi_i^{1-\alpha}(1-\chi_i)^{1-\beta}}{c(\alpha,\beta)}. \tag{12.20}$$

A fraction $f_{\rm spike}$ of the population falls in this (approximately) zero-spin spike while $1 - f_{\rm spike}$ lives in the bulk, parametrized again by a Beta distribution. Unlike Ref. [165], we do not require both black holes in a given binary to occupy the same sub-population, but instead assume that component spins are independently and identically drawn from Eq. (12.20). The spin tilt distribution is the same as that in the `Beta+Mixture` model, given by Eq. (12.19).

`Beta+TruncatedMixture`. Further motivated by Ref. [165], we alternately extend the `Beta+Mixture` model by instead modifying the $\cos\theta$ distribution, augmenting it with a tunable lower truncation bound $z_{\rm min}$:

$$p(\cos\theta_i|f_{\rm iso},\sigma_t) = \frac{f_{\rm iso}}{1-z_{\rm min}} + (1-f_{\rm iso})\mathcal{N}_{[z_{\rm min},1]}(\cos\theta_i|1,\sigma_t)\,, \tag{12.21}$$

with $p(\cos\theta_i) = 0$ for $\cos\theta_i \leq z_{\rm min}$. The spin magnitude distribution is the same as that in the `Beta+Mixture` model, given in Eq. (12.17).

`BetaSpike+TruncatedMixture`. Finally, we consider both extensions together, incorporating both the (finite-width) zero-spin sub-population in Eq. (12.20) and the truncated spin tilt distribution in Eq. (12.21).

12.B Hierarchical inference and data analyzed

In this appendix we briefly describe our hierarchical inference framework as well as the exact data we use in this study. Given a catalog of gravitational-wave events with data $\{d_i\}_{i=1}^{N_{\rm obs}}$, the likelihood that such events arise from a population with parameters Λ is [340–342, 521]

$$p(\{d\}|\Lambda) \propto \xi(\Lambda)^{-N_{\rm obs}} \prod_i \int d\lambda_i\, p(d_i|\lambda_i) p(\lambda_i|\Lambda)\,, \tag{12.22}$$

where λ_i denotes the parameters (masses, spins, etc.) of each individual binary, and we have marginalized over the overall rate of binary mergers assuming a logarithmically-uniform prior. The detection efficiency $\xi(\Lambda)$ is the fraction of events that we expect to successfully detect given the proposed population Λ. If $P_{\rm detection}(\lambda)$ is the probability that an event with parameters λ is successfully recovered by detection pipelines, then

$$\xi(\Lambda) = \int d\lambda\, p(\lambda|\Lambda) P_{\rm detection}(\lambda)\,. \tag{12.23}$$

In order to evaluate Eq. (12.22), we require the likelihood $p(d_i|\lambda_i)$ for each detection. In most cases, however, we do not have access to these likelihoods,

but rather the *posterior* $p(\lambda_i|d_i)$ obtained using some default prior $p_{\rm pe}(\lambda)$. In this case we rewrite Eq. (12.22) as

$$p(\{d\}|\Lambda) \propto \xi(\Lambda)^{-N_{\rm obs}} \prod_i \int d\lambda_i \, \frac{p(\lambda_i|d_i)}{p_{\rm pe}(\lambda_i)} p(\lambda_i|\Lambda) \,. \tag{12.24}$$

Furthermore, we generally do not know $p(\lambda_i|d_i)$ itself, but instead have a discrete set of *independent samples* $\{\lambda_{i,j}\}_{j=1}^{N_i}$ drawn from $p(\lambda_i|d_i)$, where N_i is the number of posterior samples used for event i. The standard course of action is to approximate the integral within Eq. (12.24) via a Monte Carlo average over these posterior samples,

$$p(\{d\}|\Lambda) \propto \xi(\Lambda)^{-N_{\rm obs}} \prod_i \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{p(\lambda_{i,j}|\Lambda)}{p_{\rm pe}(\lambda_{i,j})} \,. \tag{12.25}$$

We can similarly recast the detection efficiency [Eq. (12.23)] as a Monte Carlo average. Given a number $N_{\rm inj}$ of injected signals drawn from some reference distribution $p_{\rm inj}(\lambda)$, the detection efficiency may be approximated via

$$\xi(\Lambda) = \frac{1}{N_{\rm inj}} \sum_{i=1}^{N_{\rm found}} \frac{p(\lambda_i|\Lambda)}{p_{\rm inj}(\lambda_i)} \,, \tag{12.26}$$

summing over the $N_{\rm found}$ injections that pass our detection criteria.

The fundamental assumption made in Eqs. (12.25) and (12.26) is that posterior samples (and recovered injections) are sufficiently dense relative to the population features of interest. In this paper, however, we are concerned with very narrow features in the binary black hole spin distribution: Gaussians of width $\epsilon \ll 1$ or even true delta functions. We cannot therefore be automatically assured that Eqs. (12.25) and (12.26) will accurately represent our likelihood and detection efficiency. In Appendix 12.D we will further assess the accuracy of these Monte Carlo averages, and in Appendix 12.E describe how to bypass them completely.

For this study, we consider the subset of black holes among GWTC-3 with false alarm rates below $1\,{\rm yr}^{-1}$ [287]. We do not include GW190814 [522] or GW190917 [285], both of which have secondary masses $m_2 < 3\,M_\odot$ and are outliers with respect to the binary black hole population [282, 283]. This leaves us with a total of 69 binary black holes in our sample. We use parameter estimation samples made publicly available through the Gravitational-Wave

Open Science Center[6] [405, 523]. For events first published in GWTC-1[7] [26] we use the "`Overall_posterior`" parameter estimation samples. We adopt the "`PrecessingSpinIMRHM`" samples for events first published in GWTC-2[8] [286] and GWTC-2.1[9][285], and for new events in GWTC-3 [287] use the "`C01:Mixed`" samples[10]. These choices correspond to a union of samples obtained with different waveform families. All samples include spin precession effects, while the `PrecessingSpinIMRHM` and `C01:Mixed` samples from GWTC-2, GWTC-2.1, and GWTC-3 additionally include the effects of higher order modes (parameter estimation accounting for higher order modes was not available in GWTC-1). We evaluate the detection efficiency using the set of successfully recovered binary black hole injections, provided by the LIGO-Virgo-KAGRA collaborations, spanning their first three observing runs[11].

12.C Model Checking

In order to trust physical conclusions drawn from phenomenological models, we must be sure that the models themselves provide a reasonable fit to observation. A particularly powerful means of model-checking is to perform predictive tests: comparing the *predictions* made by a fitted model to our actual observed data. In this section, we subject each of the effective and component spin models used in this paper to such predictive tests.

Figure 12.8 compares predicted and observed $\chi_{\rm eff}$ values under the `Gaussian`, `GaussianSpike`, and `BimodalGaussian` models explored in Sect. 12.4. The ensemble of traces in each panel reflects the uncertainties in the hyperparameters of each model, as well as our uncertainties in the observed properties of each event. Specifically, these figures are generated via the following algorithm:

0. Perform hierarchical inference using the model in question, as described in Appendix 12.B, to obtain posteriors on the hyperparameters Λ of the model.

1. From the posterior on Λ, draw a single hyperparameter sample Λ_i.

[6] https://www.gw-openscience.org/
[7] GWTC-1 samples available at https://dcc.ligo.org/LIGO-P1800370/public
[8] GWTC-2 samples available at https://dcc.ligo.org/LIGO-P2000223/public
[9] GWTC-2.1 samples available at https://zenodo.org/record/5117703
[10] GWTC-3 samples available at https://zenodo.org/record/5546663
[11] https://zenodo.org/record/5636816

2. Use this Λ_i to define a new prior $p(\lambda|\Lambda_i)$ on the properties of individual events. Reweight the posterior of each observed binary to this new prior, as in e.g. Ref. [162], and randomly select a single sample λ_j from each event. The resulting set $\{\lambda\}_{\rm obs}$ constitutes one realization of "observed" values.[12]

3. Similarly, reweight the set of successfully found injections described above to the proposed prior $p(\lambda|\Lambda_i)$. Randomly select $N_{\rm obs}$ values from these injections; the resulting set $\{\lambda\}_{\rm inj}$ is one realization of "predicted" values. Drawing predicted values from found injections, rather than directly from the proposed population Λ_i, serves to accurately capture relevant selection effects.

4. Independently sort $\{\lambda\}_{\rm obs}$ and $\{\lambda\}_{\rm inj}$ and plot them against one another, yielding a single "Observed vs. Predicted" trace as in Fig. 12.8.

5. Repeat Steps 1-4.

A model that accurately captures features in our observed data will yield an ensemble of traces centered on the Predicted $\chi_{\rm eff}$ = Observed $\chi_{\rm eff}$ diagonal, shown as a dashed black line. Systematic departures away from the diagonal, on the other hand, would indicate a failure of the fitted model to reflect true features in the data. All three effective spin models show good predictive power in Fig. 12.8, yielding traces distributed symmetrically about the diagonal. Note that the large variance in the `BimodalGaussian` results reflects our correspondingly large uncertainty about the $\chi_{\rm eff}$ distribution at very negative and very positive values. The elevated variance is symmetric about the diagonal, though, and so is not a sign of model failure.

Figure 12.9 shows analogous predictive checks using the four component spin models explored in Sect. 12.5. We investigate each model's ability to predict observed spin magnitudes (first and second columns), spin tilts (third and fourth columns), and effective spins (final column). Each model shows good predictive power, with no significant departures from the diagonal that would indicate model failure. The `Beta+Mixture` and `BetaSpike+Mixture` models, though, may show slight signs of tension in their ability to predict $\cos\theta_1$ and $\chi_{\rm eff}$. For each model, 70% of traces lie above the diagonal at Predicted $\cos\theta_1$ =

[12] Note that reweighting single-event posteriors following Steps 1 and 2 avoids any "double-counting" of information [350].

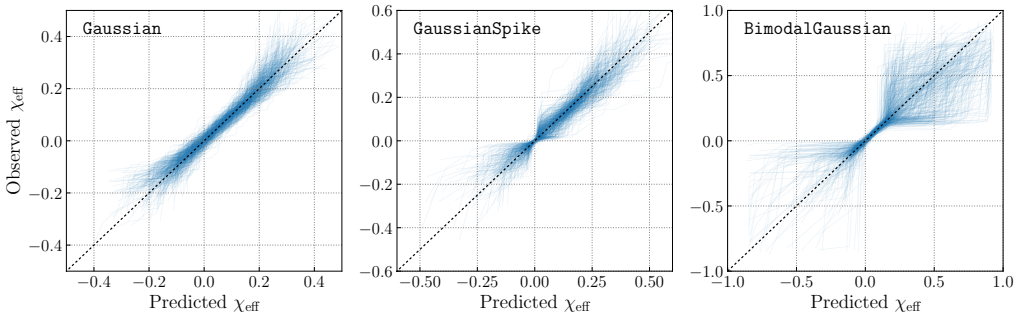


Figure 12.8: Predictive checks of the three effective spin models explored in Sec. 12.4. Under each model, we repeatedly generate sets of "Observed" $\chi_{\rm eff}$ values drawn from our binary black hole observations, and sets of "Predicted" values according to the fitted models; see Appendix 12.C for further details. Each trace among the three subplots represents one such realization, with the spread among traces reflecting both our uncertainty in the properties of each observed binary as well as uncertainty in model hyperparameters. A model that provides a good fit to observation will yield an ensemble of traces centered symmetrically around the diagonal, whereas any systematic departure from the diagonal would indicate model failure. No such departures are seen, and so all three effective spin models are able to successfully reproduce observation.

-0.5, possibly indicating a tendency to over-predict the occurrence of large and negative $\cos\theta$ values. Accordingly, both models slightly over-predict the prevalence of large negative $\chi_{\rm eff}$, with $\sim 75\%$ of traces rising above the diagonal at Predicted $\chi_{\rm eff} = -0.2$. This behaviour is consistent with the preference for a truncation at $z_{\rm min} \approx -0.5$ found using the `Beta+TruncatedMixture` and `BetaSpike+TruncatedMixture` models. These tendencies are slight, however, and so cannot yet be taken as an indicator of model failure; more data will be required to better resolve the underlying spin distributions and determine which (if any) models can no longer provide a good descriptor of observation.

12.D Effective samples

In order to identify and mitigate possible issues due to such finite sampling effects when estimating Eq. (12.25), we track the number of "effective samples" informing the Monte Carlo estimates of the likelihood for every event. Given a set of N_i posterior samples $\{\lambda_{i,j}\}_{j=1}^{N_i}$ for each event i, the number of "effective samples" under a proposed population Λ is

$$N_{{\rm eff},i}(\Lambda) \equiv \frac{\left[\sum_{j=1}^{N_i} w_{i,j}(\Lambda)\right]^2}{\sum_{j=1}^{N_i}\left[w_{i,j}(\Lambda)\right]^2}\,, \tag{12.27}$$

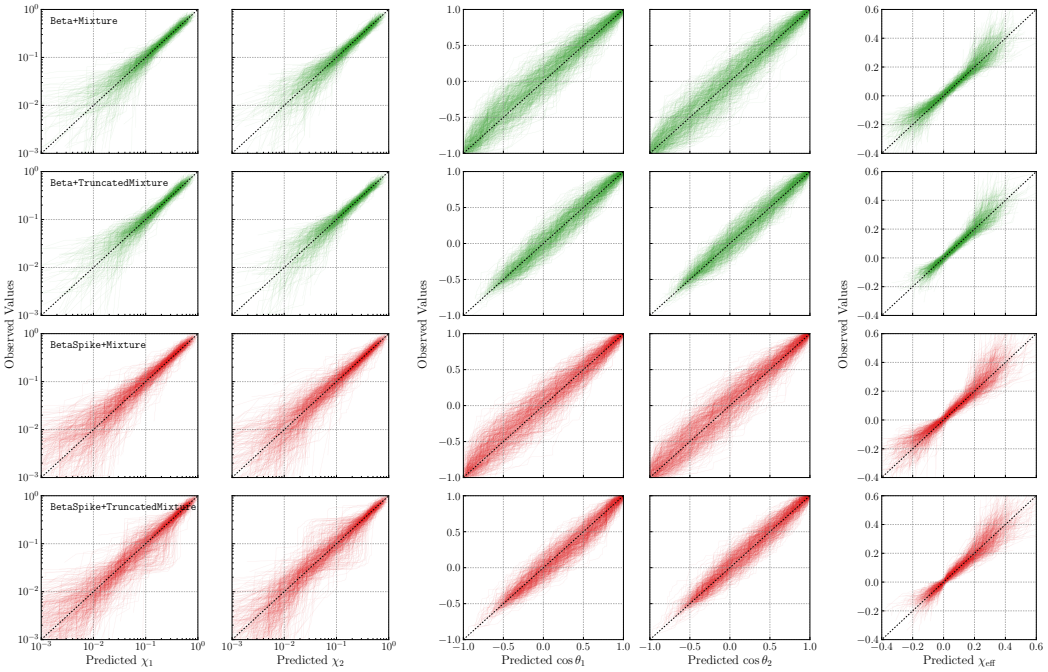


Figure 12.9: As in Fig. 12.8 above, but for the component spin models studied in Sec. 12.5. We explore each model's ability to predict the observed spin magnitudes (first and second columns), spin tilts (third and fourth columns), and effective spins (fifth column). Again, we see no clear departures from the diagonal, such that none of the four models can be rejected as a poor descriptor of our data. The `Beta+Mixture` and `BetaSpike+Mixture` models may exhibit slight tension with observation, possibly over-predicting the occurrence of very negative $\cos\theta_1$ and $\chi_{\rm eff}$, but additional data are required to confirm the significance of this tendency.

where $w_{i,j}(\Lambda) = p(\lambda_{i,j}|\Lambda)/p_{\rm pe}(\lambda_{i,j})$. Small $N_{{\rm eff},i}(\Lambda)$ indicates that the given event is sparsely sampled in the region where the population $p(\lambda|\Lambda)$ is concentrated, and hence the likelihood may be dominated by sampling variance. In these regions, we should not necessarily trust the results of Monte-Carlo-based hierarchical inference. To avoid such regions, Ref. [283] impose a data-dependent *prior cut*, rejecting those Λ for which $\min\left[N_{{\rm eff},i}(\Lambda)\right]$ falls below some threshold value. We avoid imposing such a cut, which amounts to an implicit (and often stringent) prior on Λ. Instead, we compute and track $\min\left[N_{{\rm eff},i}(\Lambda)\right]$ for each component spin population model we consider. Additionally, for models with distinct "spike" and "bulk" spin sub-populations (i.e. the half-Gaussian at zero and the Beta distribution, respectively), we track the minimum effective sample count within each of these sub-populations.

Our primary concern in tracking $N_{\rm eff}$ is to calibrate the allowed width $\epsilon_{\rm spike}$

of a possible zero-spin spike in the component spin distribution. Due to finite sampling effects, we cannot let this spike width be arbitrarily small, but must bound $\epsilon_{\rm spike}$ to sufficiently large values to ensure reasonable $N_{\rm eff}$. Ref. [524] argue that $N_{\rm eff} \sim 10$ per event is sufficiently high to ensure accurate marginalization over each event's parameters. Figure 12.10 illustrates how $\min\left[N_{{\rm eff},i}\right]$ varies as a function of $\epsilon_{\rm spike}$; each point represents a posterior sample drawn from the `BetaSpike+TruncatedMixture` results in Fig. 12.6. The left panel shows $N_{\rm eff}$ as computed across the full spin distribution, while the center panel shows the effective sample count taken just over the "bulk" Beta distribution (via fixing $f_{\rm spike} = 0$ when computing Eq. (12.27)). The effective sample counts across the total population and in the bulk are largely uncorrelated with $\epsilon_{\rm spike} > 0.03$ and are everywhere large, with $\min N_{\rm eff} \gtrsim 10$.

The right panel of Fig. 12.10, meanwhile, shows the number of effective samples contained in the zero-spin spike (obtained via fixing $f_{\rm spike} = 1$ in Eq. (12.27)). The number of effective samples in the spike is extremely correlated with $\epsilon_{\rm spike}$. This is expected, as a wider spike will encompass more posterior samples and thus have a larger $\min\left[N_{{\rm eff},i}\right]$. The blue points show $\min\left[N_{{\rm eff},i}\right]$ taken across all events. This minimum sample count is unacceptably low [524], with every population sample yielding at least one binary with only ~ 1 effective sample. However, we argue that this is not in fact concerning.[13] The events with a very low number of effective samples in the spike are the same events that are unambiguously spinning, and hence confidently belong in the bulk and not the spike. Their extremely low number of effective samples in the spike, therefore, does not affect inference. The left-hand panel in Fig. 12.11 illustrates the component spin magnitude posterior for one such "confidently spinning" event that unambiguously rules out $\chi_1 = \chi_2 = 0$. The red dots in Fig. 12.10 show the minimum effective sample count if we now exclude these confidently-spinning events (GW190517, GW190412, GW151226, and GW191204, specifically). This minimum count still contains events that are likely (but not necessarily unambiguously) non-spinning, such as the event shown in the center panel of Fig. 12.11. The most critical measure of stability is whether $N_{\rm eff}$ is large for events such as GW150914 (right panel of Fig. 12.11)

[13] If one wishes to avoid this argument altogether, the alternative method discussed in Appendix 12.E and used to infer the binary $\chi_{\rm eff}$ distribution is accurate in the presence of narrow population features, and produces results that agree with those discussed in this section.

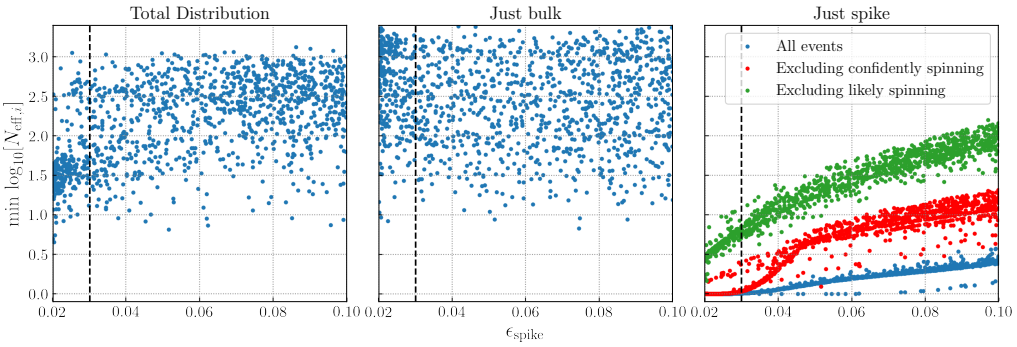


Figure 12.10: Scatter plot of the minimum (across events) number of effective samples $\min[N_{\mathrm{eff},i}]$ and the width of the zero-spin spike $\epsilon_{\mathrm{spike}}$ for the `BetaSpike+TruncatedMixture` model. Each point represents a single draw from the posterior shown in Fig. 12.6. Left panel: The minimum number of effective samples as computed across the entire population model. Middle panel: The minimum number of effective samples in the "bulk" (Beta distribution) spin magnitude sub-population. Right panel: The minimum number of effective samples in the "spike" (zero-spin half-Gaussian) spin magnitude sub-population. Dashed vertical lines denote $\epsilon_{\mathrm{spike}} = 0.03$, the lower limit we impose on this parameter in our prior. Although the minimum sample count in the spike appears unacceptably low, the minimum is driven by events that are confidently spinning and which therefore do not impact our inference regarding the zero-spin spike (see the left panel of Fig. 12.11). Red points illustrate the minimum effective samples when excluding events these events that are confidently spinning. Green points show the minimum effective sample count if we further exclude events that are "likely" spinning (e.g. middle panel of Fig. 12.11), focusing only on those events that are well-described by zero-spin. To ensure a reasonable number of effective samples among these remaining events, we bound $\epsilon_{\mathrm{spike}} > 0.03$, and note that the region $\epsilon_{\mathrm{spike}} \lesssim 0.04$ may be subject to elevated Monte Carlo variance.

whose posteriors are finite up to and including $\chi_1 = \chi_2 = 0$. The green points in Fig. 12.10, therefore, show the minimum sample count when we further exclude "likely nonspinning" events, defined as any event with less than $1/200$ of its samples in the region $\chi_1, \chi_2 < 0.1$. The minimum number of effective samples across remaining events now rises to ~ 5 at $\epsilon_{\mathrm{spike}} = 0.03$ and $\gtrsim 10$ at $\epsilon_{\mathrm{spike}} = 0.04$. Given this, we choose to limit $\epsilon_{\mathrm{spike}} \geq 0.03$, and caution that the region $\epsilon_{\mathrm{spike}} < 0.04$ may be subject to increased Monte Carlo averaging error.

12.E Avoiding samples with likelihood KDEs

Our ordinary population likelihood, Eq. (12.25), relies on Monte Carlo averages over posterior samples, an approach that can behave poorly when evaluating

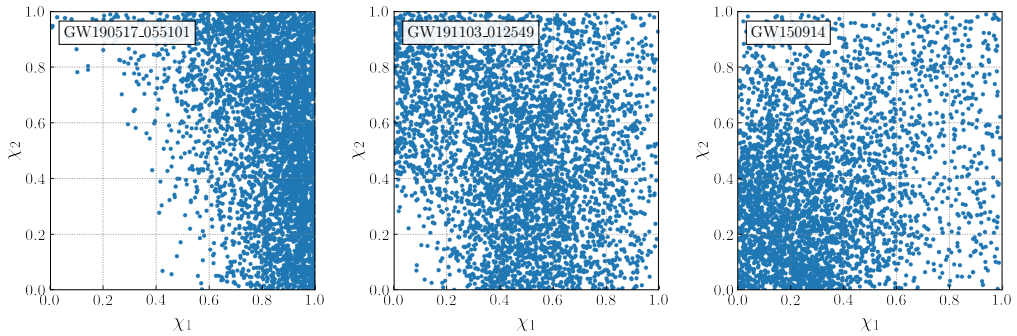


Figure 12.11: Spin magnitude posterior samples demonstrating the three classes of event discussed in Fig. 12.10 above. The left-hand panel illustrates a confidently spinning binary which is clearly identified as a member of the spinning "bulk" population, and whose low effective sample count in the spike is therefore unimportant. The middle panel illustrates a "likely" spinning event, defined by having less than a fraction 1/200 of its posterior samples in the region $\chi_1, \chi_2 < 0.1$. The right panel, finally, shows an event belonging to neither category; these remaining events are used to generate the green points in Fig. 12.10 above.

narrow population features as described above. Within the `GaussianSpike` effective spin model, for example, in the limit that $\epsilon \to 0$ there will be precisely *no* posterior samples falling in the zero-spin spike, and hence we will be unable to precisely evaluate the model's likelihood. This limitation is not physical, but purely due to our representation of each event's posterior as a set of discrete samples. As an alternative approach, we can abandon samples altogether and instead represent the likelihoods of individual events as Gaussian kernel density estimates (KDEs). This alternative representation remains well-behaved even when the population features of interest are narrow, provided that there are enough samples in the neighborhood of the narrow feature to estimate the KDE.

For convenience, let $\tilde{\lambda}$ represent all individual-event parameters *except* $\chi_{\rm eff}$. Under the discrete sample representation, the ordinary Monte Carlo average in Eq. (12.25) is obtained by identifying individual events' likelihoods $p(d_i|\tilde{\lambda}_i, \chi_{{\rm eff},i})$ as a sum of delta-functions located at each posterior sample j:

$$p(d_i|\tilde{\lambda}_i, \chi_{{\rm eff},i}) \propto \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\delta(\tilde{\lambda}_i - \tilde{\lambda}_{i,j})\delta(\chi_{{\rm eff},i} - \chi_{{\rm eff},i,j})}{p_{\rm pe}(\tilde{\lambda}_{i,j}, \chi_{{\rm eff},i,j})} \,. \tag{12.28}$$

In the KDE approach, we impart a finite "resolution" to each posterior sample,

replacing the delta-functions with Gaussians of width $\sigma_{\rm kde}$:

$$p(d_i|\tilde{\lambda}_i,\chi_{{\rm eff},i}) \propto \frac{1}{N_i}\sum_{j=1}^{N_i}\frac{\delta(\tilde{\lambda}_i-\tilde{\lambda}_{i,j})N(\chi_{{\rm eff},i}|\chi_{{\rm eff},i,j},\sigma_{\rm kde})}{p_{\rm pe}(\tilde{\lambda}_{i,j},\chi_{{\rm eff},i,j})}\,. \tag{12.29}$$

Equation (12.29) is now a continuous function of $\chi_{{\rm eff},i}$, and hence allows us to evaluate the likelihood of a population with arbitrarily narrow features. This comes at a cost. As shown in Eq. (12.22), evaluation of a population likelihood requires evaluation of the integral $\int d\tilde{\lambda}\,d\chi_{\rm eff}\,p(d_i|\tilde{\lambda},\chi_{\rm eff})p(\tilde{\lambda},\chi_{\rm eff}|\Lambda)$. This integration is trivial when the likelihoods are composed purely of delta functions, as in Eq. (12.28), but not necessarily straightforward using Eq. (12.29).

Fortunately, both the individual-event likelihoods $p(d_i|\tilde{\lambda}_i,\chi_{{\rm eff},i})$ and our population models $p(\tilde{\lambda},\chi_{\rm eff}|\Lambda)$ are mixtures of Gaussians in $\chi_{\rm eff}$, and so we can analytically carry out the required integration over $\chi_{\rm eff}$. For the `GaussianSpike` model, for example, the result is

$$\begin{aligned}
p(\{d\}|\Lambda) &\propto \xi(\Lambda)^{-N_{\rm obs}}\prod_i\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{p(\tilde{\lambda}_{i,j}|\Lambda)}{p_{\rm pe}(\tilde{\lambda}_{i,j},\chi_{{\rm eff},i,j})}\\
&\qquad\times\int d\chi_{{\rm eff},i}\,N(\chi_{{\rm eff},i}|\chi_{{\rm eff},i,j},\sigma_{\rm kde})\Big[\zeta_{\rm spike}N_{[-1,1]}(\chi_{{\rm eff},i}|0,\epsilon)\\
&\qquad\qquad+(1-\zeta_{\rm spike})N_{[-1,1]}(\chi_{{\rm eff},i}|\mu_{\rm eff},\sigma_{\rm eff})\Big]\\
&\propto \xi(\Lambda)^{-N_{\rm obs}}\prod_i\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{p(\tilde{\lambda}_{i,j}|\Lambda)}{p_{\rm pe}(\tilde{\lambda}_{i,j},\chi_{{\rm eff},i,j})}\\
&\qquad\times\left[\mathcal{A}_{i,j}\,\zeta_{\rm spike}\exp\left(-\frac{(\chi_{{\rm eff},i,j})^2}{2(\epsilon^2+\sigma_{\rm kde}^2)}\right)\right.\\
&\qquad\qquad\left.+\mathcal{B}_{i,j}\,(1-\zeta_{\rm spike})\exp\left(-\frac{(\chi_{{\rm eff},i,j}-\mu_{\rm eff})^2}{2(\sigma_{\rm eff}^2+\sigma_{\rm kde}^2)}\right)\right],
\end{aligned} \tag{12.30}$$

with

$$\mathcal{A}_{i,j} = \frac{{\rm Erf}\left(\frac{(1-\chi_{{\rm eff},i,j})\epsilon^2+\sigma_{\rm kde}^2}{\sqrt{2\epsilon^2\sigma_{\rm kde}^2(\epsilon^2+\sigma_{\rm kde}^2)}}\right)+{\rm Erf}\left(\frac{(1+\chi_{{\rm eff},i,j})\epsilon^2+\sigma_{\rm kde}^2}{\sqrt{2\epsilon^2\sigma_{\rm kde}^2(\epsilon^2+\sigma_{\rm kde}^2)}}\right)}{2\sqrt{2\pi(\epsilon^2+\sigma_{\rm kde}^2)}{\rm Erf}\left(\frac{1}{\sqrt{2\epsilon^2}}\right)}\,, \tag{12.31}$$

and

$$\mathcal{B}_{i,j} = \frac{{\rm Erf}\left(\frac{(1-\chi_{{\rm eff},i,j})\sigma_{\rm eff}^2+(1-\mu_{\rm eff})\sigma_{\rm kde}^2}{\sqrt{2\sigma_{\rm eff}^2\sigma_{\rm kde}^2(\sigma_{\rm eff}^2+\sigma_{\rm kde}^2)}}\right)+{\rm Erf}\left(\frac{(1+\chi_{{\rm eff},i,j})\sigma_{\rm eff}^2+(1+\mu_{\rm eff})\sigma_{\rm kde}^2}{\sqrt{2\sigma_{\rm eff}^2\sigma_{\rm kde}^2(\sigma_{\rm eff}^2+\sigma_{\rm kde}^2)}}\right)}{\sqrt{2\pi(\sigma_{\rm eff}^2+\sigma_{\rm kde}^2)}\left[{\rm Erf}\left(\frac{1-\mu_{\rm eff}}{\sqrt{2\sigma_{\rm eff}^2}}\right)+{\rm Erf}\left(\frac{1+\mu_{\rm eff}}{\sqrt{2\sigma_{\rm eff}^2}}\right)\right]}\,. \tag{12.32}$$

In an exactly analogous fashion, when computing the detection efficiency $\xi(\Lambda)$ we can replace a Monte Carlo average over successfully found injections [as in Eq. (12.26)] with a KDE over injections:

$$\xi(\Lambda) \propto \sum_{i=1}^{N_{\rm found}} \frac{p(\tilde{\lambda}_i|\Lambda)}{p_{\rm inj}(\tilde{\lambda}_i, \chi_{{\rm eff},i})} \int d\chi_{\rm eff}\, N(\chi_{\rm eff}|\chi_{{\rm eff},i}, \sigma_{\rm kde})\, p(\chi_{\rm eff}|\Lambda)\,. \tag{12.33}$$

As above, the integration can be carried out analytically when (as in our case) the population model $p(\chi_{\rm eff}|\Lambda)$ is a mixture of Gaussians.

As a demonstration and validation of this approach, we revisit inference of the GWTC-3 effective spin distribution via the `GaussianSpike` model, but now fix the "spike" at $\chi_{\rm eff} = 0$ to have a finite width $\epsilon > 0$. We repeat the analysis in two ways: (i) using the ordinary Monte Carlo representation of Eqs. (12.25) and (12.26), and (ii) using the KDE representation of Eqs. (12.30) and (12.33). Figure 12.12 shows the marginal posterior for $\zeta_{\rm spike}$ for different choices of ϵ. When ϵ is large (i.e., the "spike" is broad compared to the typical inter-sample spacing) the two methods give identical results. As we let ϵ approach zero, however, the Monte Carlo average produces divergent results, while the KDE likelihood representation yields stable and converging posteriors.

12.F Mock injection campaign

In Fig. 12.3 in the main text, we showed posteriors on $\zeta_{\rm spike}$ obtained by analyzing mock catalogs of binary black holes drawn from an intrinsically spike-less population. In this appendix we describe our procedure for generating and analyzing these mock catalogs.

To generate mock $\chi_{\rm eff}$ measurements, we assume an underlying astrophysical distribution following the `Gaussian` effective spin model, with mean $\mu_{\rm eff} = 0.06$ and standard deviation $\sigma_{\rm eff} = 0.09$. These values are chosen to match the median measurements of $\mu_{\rm eff}$ and $\sigma_{\rm eff}$ obtained when hierarchically analyzing GWTC-3 with the `Gaussian` effective spin model. For each mock catalog, we draw 69 "true" spin values from this Gaussian population:

$$\chi_{\rm eff,true} \sim N(0.06, 0.09). \tag{12.34}$$

We then add a random measurement uncertainty to each mock event. Assuming Gaussian likelihoods for each $\chi_{\rm eff}$ measurement, "observed" maximum-likelihood values are generated via

$$\chi_{\rm eff,obs} \sim N(\chi_{\rm eff,true}, \sigma_{\rm obs}), \tag{12.35}$$

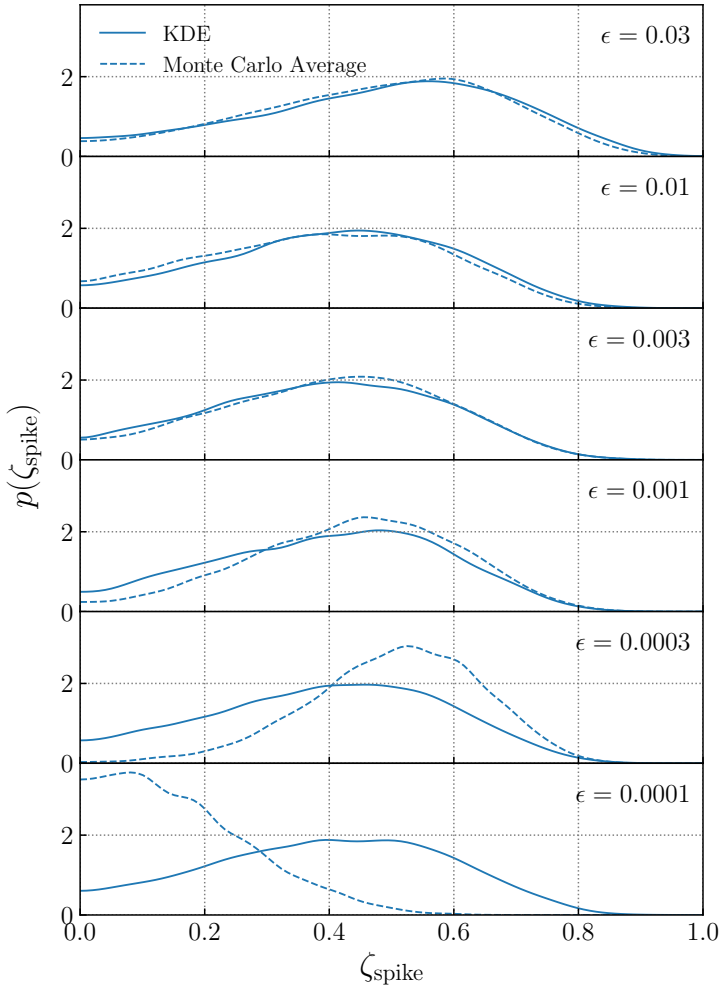


Figure 12.12: Validation of the KDE likelihood approach discussed in Appendix 12.E. Each subplot shows the marginalized posterior for the fraction of events contained within a narrow spike at $\chi_{\rm eff} = 0$. Solid lines show posteriors computed using the KDE method, while dashed lines show posteriors obtained using ordinary Monte Carlo averaging (Appendix 12.B). Results are shown for a variety of spike widths, from broad spikes with standard deviation $\epsilon = 0.03$ to narrow spikes with $\epsilon = 10^{-4}$. We see that the KDE method yields consistent and convergent results as ϵ approaches zero, while the Monte Carlo averaging gives increasingly divergent results for small spike widths due to growing Monte Carlo errors.

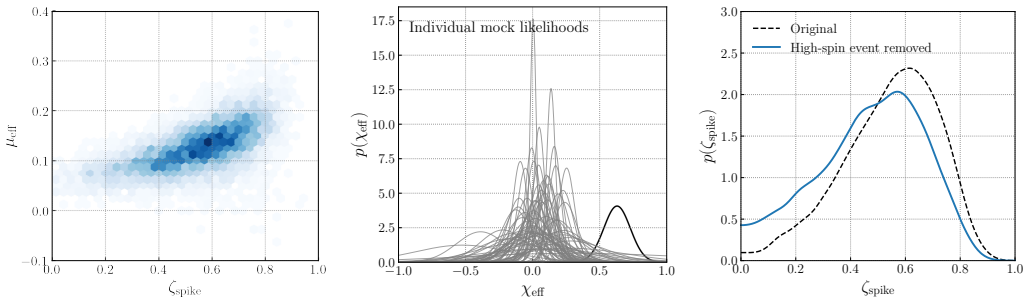


Figure 12.13: Follow-up exploration of the injected mock catalog that most strongly (and incorrectly) excludes $\zeta_{\rm spike} = 0$ in Fig. 12.3. Left: The joint posterior on $\zeta_{\rm spike}$ and $\mu_{\rm eff}$ from the selected catalog. In general, we find significant correlations between $\zeta_{\rm spike}$ and the mean $\mu_{\rm eff}$ of the spinning "bulk" population. Mock catalogs that most strongly disfavor $\zeta_{\rm spike} = 0$ are generally those that have the largest inferred $\mu_{\rm eff}$. Middle: The individual events comprising the selected catalog. The event with the largest measured $\chi_{\rm eff}$ (and hence the event that most strongly influences $\mu_{\rm eff}$) is highlighted. Right: The solid blue curve shows the updated $\zeta_{\rm spike}$ posterior after removing the highlighted high-spin event. Compared to the original $\zeta_{\rm spike}$ measurement (dashed black curve), the updated posterior is now more (correctly) consistent with $\zeta_{\rm spike} = 0$.

where $\sigma_{\rm obs}$ is the standard deviation of each event's likelihood. These uncertainties are themselves randomly distributed to match the range of uncertainties exhibited by real measurements. The $\chi_{\rm eff}$ likelihoods among binary black holes in GWTC-3 have log standard deviations that are approximately normally distributed, with a mean $\log_{10}\sigma_{\rm obs}$ of -0.9 and a standard deviation of 0.3. When drawing mock observed values in Eq. (12.35), we therefore assign each event a random measurement uncertainty according to

$$\log_{10}\sigma_{\rm obs} \sim N(-0.9, 0.3). \tag{12.36}$$

Since both our population model and our mock likelihoods are Gaussian, the hierarchical likelihood for $\mu_{\rm eff}$ and $\sigma_{\rm eff}$ can be calculated analytically, exactly as was done in Appendix 12.E above. Specifically, the likelihood $p(\{d\}|\mu_{\rm eff}, \sigma_{\rm eff})$ of our mock catalog is of the same form as Eq. (12.30), with the observed $\chi_{\rm eff,obs}$ values in place of $\chi_{{\rm eff},i,j}$, the measurement uncertainties $\sigma_{\rm obs}$ in place of ϵ, and neglecting the summation over j.

As highlighted in Sec. 12.4, many of our mock catalogs exhibit $\zeta_{\rm spike}$ posteriors qualitatively similar to the posterior obtained using real data, encompassing zero but peaking at $\zeta_{\rm spike} \approx 0.5$. Some mock catalogs even more confidently (and incorrectly) appear to rule out the injected value of $\zeta_{\rm spike} = 0$. We find

that this elevated "false-alarm probability" can be explained by a degeneracy between $\zeta_{\rm spike}$ and $\mu_{\rm eff}$ that occurs when analyzing relatively small number of individually uncertain measurements. To illustrate this behavior, we choose the mock catalog whose $\zeta_{\rm spike}$ posterior *most strongly* rules out zero in Fig. 12.3. The left-hand panel in Fig. 12.13 shows the joint posterior on $\zeta_{\rm spike}$ and $\mu_{\rm eff}$ given by this catalog. These parameters are correlated: if the Gaussian "bulk" is shifted to a larger mean, events located at negative or near-zero $\chi_{\rm eff}$ values are left behind and must be absorbed into the zero-spin spike. To test this intuition, we show in the middle panel of Fig. 12.13 the individual likelihoods that comprise this mock catalog. The event with the largest $\chi_{\rm eff}$ (and hence the event pulling $\mu_{\rm eff}$ to large values) is highlighted. If we remove this high-spin event and redo the hierarchical inference on this catalog, we obtain the blue $\zeta_{\rm spike}$ posterior shown in the right-hand panel of Fig. 12.13. After removing the high spin event, we now find $\zeta_{\rm spike}$ to be considerably more consistent with zero.

12.G Bayes factors through the Savage Dickey density ratio

Our initial counting experiment in Sec. 12.3 uses only the Bayes factors for each event in GWTC-2 between non-spinning ($\chi_1 = \chi_2 = 0$) and spinning ($\chi_1, \chi_2 \geq 0$) hypotheses. Such Bayes factors also act as inputs in the analysis of Ref. [165]; they form Bayes factors via the ratios of fully-marginalized likelihoods computed for each event via nested sampling under non-spinning and spinning priors [63, 286]. Instead, we compute Bayes factors directly from the posterior samples for each event. For nested models where the simpler model (e.g. the non-spinning model) is contained within a more complex model (the spinning model with $\chi_1 = \chi_2 = 0$), the Bayes factor between models is given through the Savage-Dickey density ratio [525]

$$\mathcal{B}^{\rm NS}_{\rm S} = \frac{p(\chi_1 = 0, \chi_2 = 0|d)}{p(\chi_1 = 0, \chi_2 = 0)}\,, \tag{12.37}$$

where $p(\chi_1 = 0, \chi_2 = 0|d)$ and $p(\chi_1 = 0, \chi_2 = 0)$ are the posterior and prior densities at $\chi_1 = \chi_2 = 0$, respectively. See Appendix B of Ref. [526] for a proof of this equation when the more complex model has additional parameters (here, the spin angles) that are absent from the simpler model. For every binary black hole in GWTC-2, the left panel of Fig. 12.14 shows the

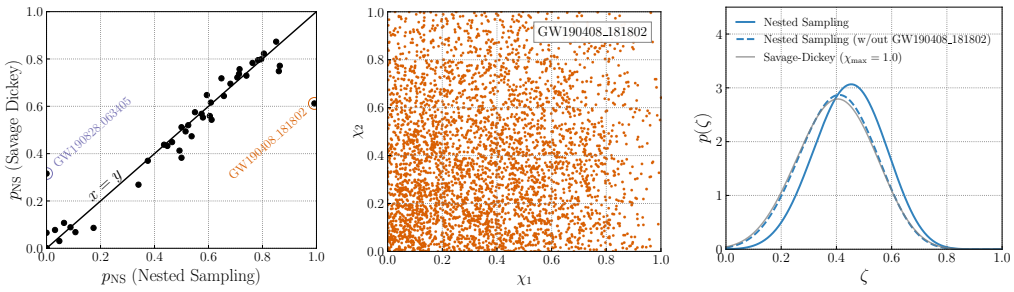


Figure 12.14: Left panel: The probabilities $p_{\rm NS}$ for each binary black hole in GWTC-2 to be non-spinning, as computed in two different ways: through a Savage Dickey density ratio and using fully-marginalized likelihoods computed via nested sampling and reported in Ref. [286]. The circles denote GW190408_181802 and GW190828_063405, whose nested sampling results appear to be outliers. Center panel: The spin magnitude posterior for GW190408_181802. Nested sampling yields a Bayes factor of $\mathcal{B}^{\rm NS}_{\rm S} \sim 130$ in favor of the hypothesis that this event is non-spinning. This would require the spin magnitude posterior to be ~ 130 times larger than the (flat) prior at $\chi_1 = \chi_2 = 0$. This is inconsistent, though, with the true posterior on the spin magnitudes of GW190408_181802, suggesting that the nested sampling Bayes factor in favor of the non-spinning hypothesis is significantly overestimated. Right panel: As a consistency check, we repeat the counting experiment from Sec. 12.3 using the nested sampling Bayes factors serving as input to Ref. [165], rather than our Savage-Dickey estimates. We show posteriors obtained using the entire GWTC-2 catalog (solid blue curve) and results after excluding GW190408_181802 from our sample (dashed blue). As mentioned in Sec. 12.3, we obtain much closer agreement with the Savage-Dickey result (grey curve) after excluding GW190408_181802.

probability that the event is non-spinning,

$$p_{\rm NS} = \frac{p_{\rm NS}}{p_{\rm S} + p_{\rm NS}} = \frac{\mathcal{B}^{\rm NS}_{\rm S}}{1 + \mathcal{B}^{\rm NS}_{\rm S}}, \tag{12.38}$$

as implied by both sets of Bayes factors. We find general agreement, within numerical uncertainties, between our Bayes factors computed via Savage Dickey density ratios and those computed via nested sampling [286]. We note that Savage Dickey density ratios, which rely on KDEs over posterior samples, may be less reliable for events with very little support at $(\chi_1 = 0, \chi_2 = 0)$. In these cases, however, a Savage Dickey density ratio would still conclude that $p_{\rm NS} \to 0$ consistently.

There are two significant outliers, however. Nested sampling returns $\mathcal{B}^{\rm NS}_{\rm S} \sim 130$ in favor of the non-spinning hypothesis for the event GW190408_181802. If this Bayes factor were correct, then GW190408_181802 is almost certainly

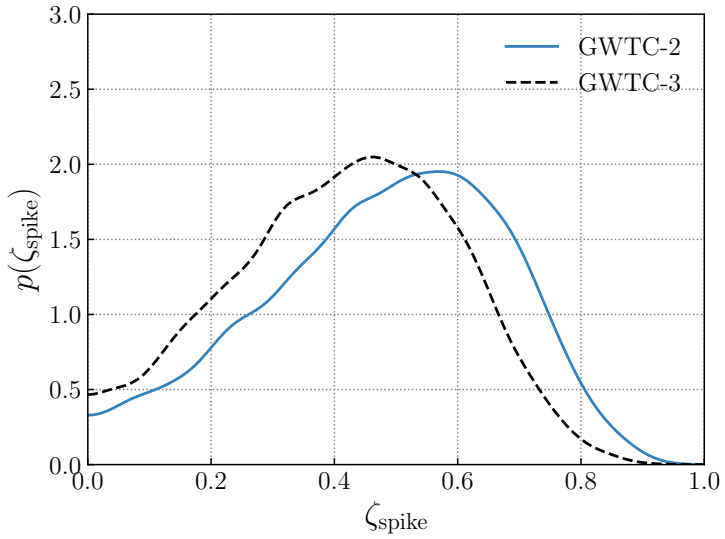


Figure 12.15: The fraction of binary black holes among GWTC-2 [solid blue line; 286] with $\chi_{\rm eff} = 0$, inferred using the `GaussianSpike` effective spin model. For comparison, the dashed black curve shows the result obtained using GWTC-3 [287]; this is the same result shown in Fig. 12.3 in the main text. Both results are consistent with one another, showing no evidence for a distinct sub-population of events with $\chi_{\rm eff} = 0$, although the GWTC-3 result more strongly favors $\zeta_{\rm spike} = 0$.

non-spinning, guaranteeing that a non-zero number of events will be placed in a non-spinning sub-population. This is indeed the conclusion drawn by Ref. [165] with the nested sampling Bayes factors. Such a large preference for vanishing spins is not supported by this event's spin magnitude posteriors (right panel); it would require that the posterior is ~ 130 times larger than the prior at $(\chi_1 = \chi_2 = 0)$. A Savage Dickey density ratio instead gives an agnostic $\mathcal{B}^{\rm NS}_{\rm S} \approx 1.6$ for GW190408_181802. The opposite situation is encountered for GW190828_063405, which is reported to have $\mathcal{B}^{\rm NS}_{\rm S} \sim 1.7 \times 10^{-5}$ from nested sampling. Inspection of the posterior again shows that the posterior and the Bayes factor are inconsistent, as the former remains finite at $(\chi_1 = \chi_2 = 0)$. Although these two outliers were identified by comparison of nested sampling and Savage Dickey density ratio results, the conclusion that the nested sampling Bayes factors are misestimated can be drawn through visual inspection of the posterior samples alone.

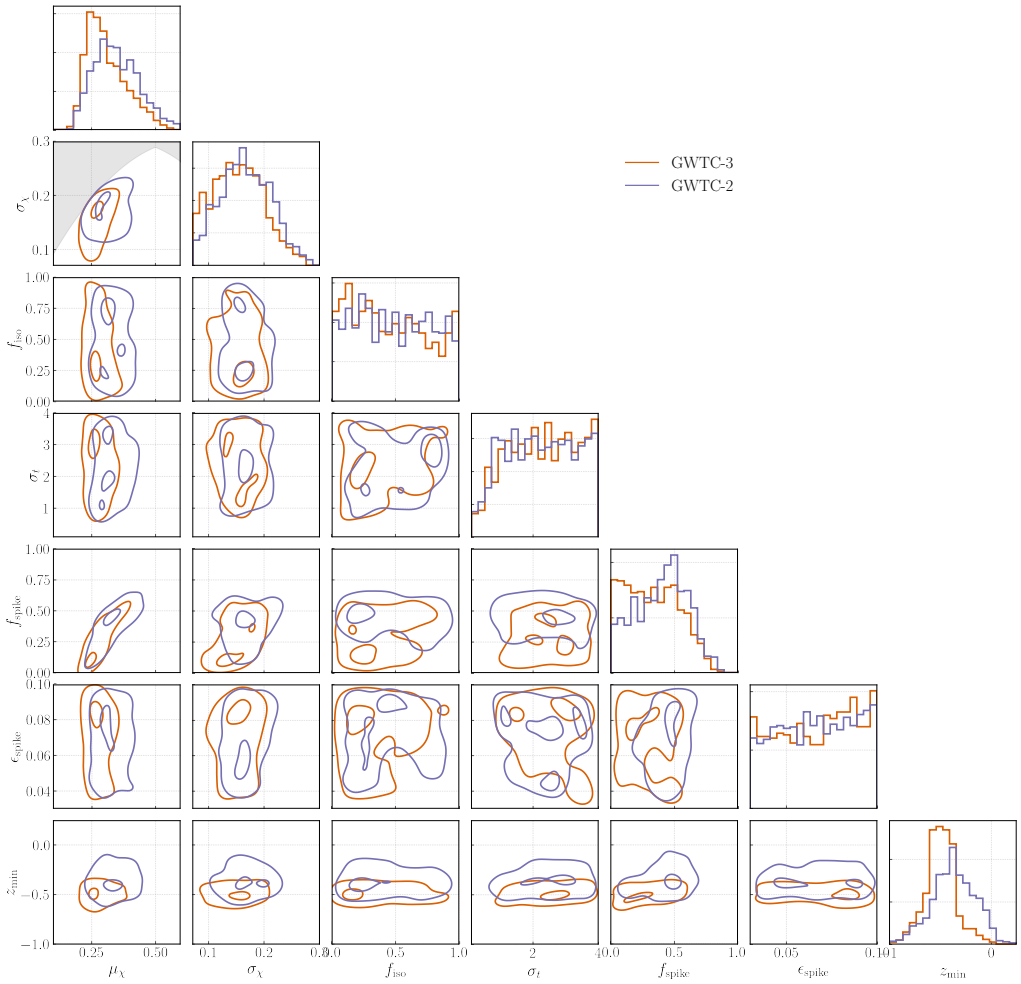


Figure 12.16: Posterior on the parameters of the `BetaSpike+TruncatedMicture` component spin model, now using only GWTC-2 events (purple). The shaded region in the μ_χ–σ_χ two-dimensional posterior is the region excluded by the prior cut on the shape parameters of the beta distribution $\alpha, \beta > 1$. For reference, results using the full GWTC-3 event list are shown in orange; the GWTC-3 result is identical to the one presented in Fig. 12.6. We find broadly consistent results between the two catalogs, including no requirement for a zero-spin sub-population as well as a preference for spins misaligned by more than 90° relative to binaries' Newtonian orbital angular momentum.

12.H Different Catalogs and Model Variations

All results in Secs. 12.4 and 12.5 of the main text rely on the latest GWTC-3 catalog of gravitational-wave detections [287], as detailed in Appendix 12.B. However, the results of Ref. [165] (to which we frequently compare) instead made use of GWTC-2 [286], since GWTC-3 results were published only after their study. Moreover, our model differs slightly from ours in how compo-

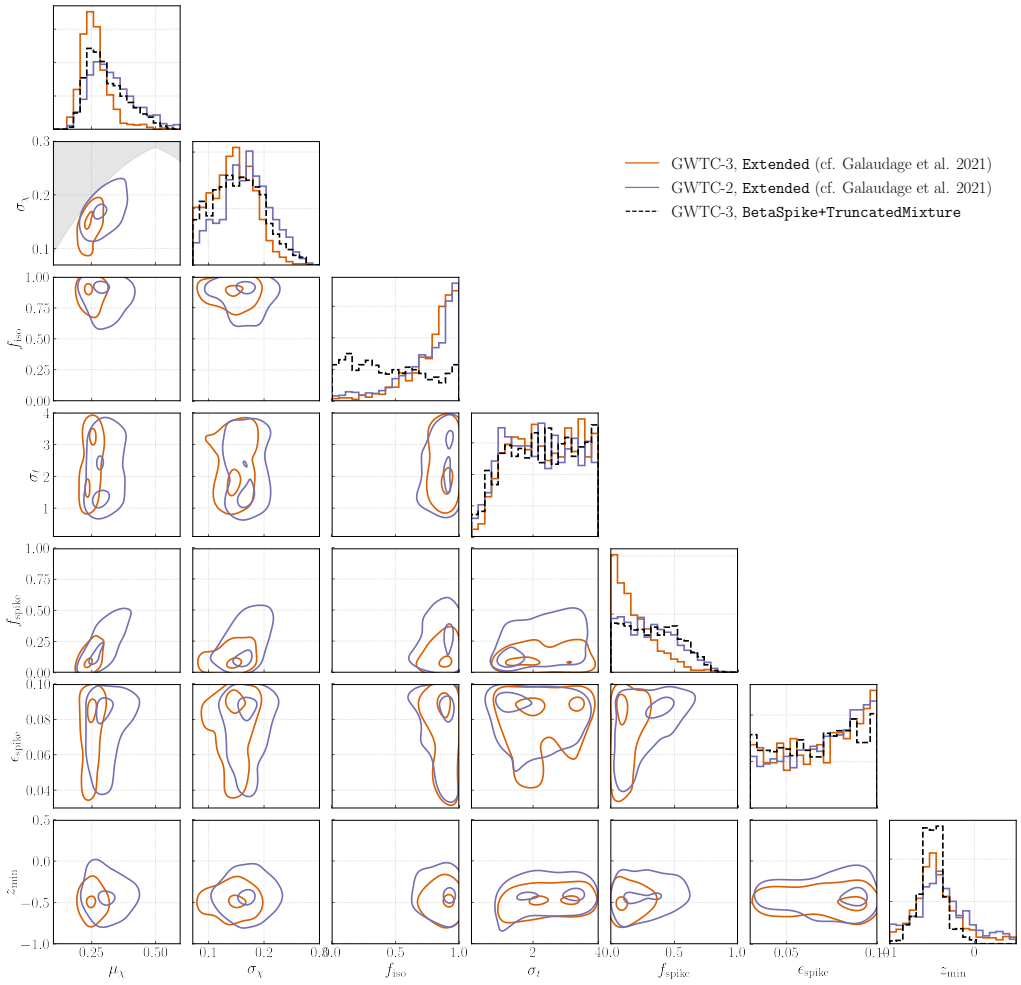


Figure 12.17: Posterior on the parameters of the `Extended` component spin model (Eqs. 12.39, 12.40) using GWTC-3 data (orange) and just GWTC-2 data (purple). The one dimensional marginalized posteriors for the `BetaSpike+TruncatedMicture` model for GWTC-3 are shown in black dashed lines for comparison; this result is identical to that shown in Figs. 12.6 and 12.16. The `Extended` model infers a population with smaller $f_{\rm spike}$ and larger $f_{\rm iso}$ than `BetaSpike+TruncatedMicture`; it also has more support for $z_{\rm min} > 0$ than `BetaSpike+TruncatedMicture`, although the $z_{\rm min}$ posterior peaks at a negative value for both cases. The posteriors for the remaining population parameters are qualitatively similar between the two models.

nent spin magnitudes and tilt angle are paired. We assume that spin magnitudes χ_1 and χ_2 are independently and identically distributed (IID), as are tilt angles $\cos\theta_1$ and $\cos\theta_2$. In our `BetaSpike+TruncatedMixture` model, for example, one component spin could lie in the zero-spin spike while its companion spin could lie in the Beta distribution "bulk." Ref. [165], on the other hand, adopt preferential pairing and require that both component spins in a given binary occupy the same mixture component (bulk or spike; aligned or isotropic). In order to enable a more direct comparison between our results, in this appendix we show results obtained using only binary black holes among GWTC-2, as well as results generated with an alternative pairing model that directly matches that of Ref. [165].

Figure 12.15 illustrates the fraction of binary black holes inferred to be non-spinning, using the `GaussianSpike` effective spin model (see Secs. 12.4 and 12.A) and analyzing data from GWTC-2. Like our results with GWTC-3, results using only GWTC-2 are consistent with $\zeta_{\rm spike} = 0$, indicating no evidence for a distinct sub-population of non-spinning events. Similarly, Fig. 12.16 shows results when analyzing GWTC-2 binary black holes with the `BetaSpike+TruncatedMixture` component spin model. We again find consistent results between GWTC-2 and GWTC-3 events. When analyzing only events in GWTC-2, we infer the fraction $f_{\rm spike}$ of approximately non-spinning systems to be consistent with zero, and find no requirement for the width of this possible sub-population to be narrow (small $\epsilon_{\rm spike}$). Using GWTC-2, we also conclude that the distribution of spin-orbit misalignment angles extends beyond 90°, with the truncation $z_{\rm min}$ on the $\cos\theta$ distribution inferred to be negative (97.1% credibility), although with slightly decreased certainty compared to GWTC-3 (99.7% credibility).

Finally, Fig. 12.17 shows results obtained under an alternative version of our `BetaSpike+TruncatedMixture` model that directly matches the pairing function used by Ref. [165] in their `Extended` model. Explicitly, spin magnitudes are jointly distributed as

$$\begin{aligned} p(\chi_1,\chi_2|\alpha,\beta,f_{\rm spike},\epsilon_{\rm spike}) = \\ f_{\rm spike}\,\mathcal{N}_{[0,1]}(\chi_1|0,\epsilon_{\rm spike})\,\mathcal{N}_{[0,1]}(\chi_2|0,\epsilon_{\rm spike}) \\ + (1-f_{\rm spike})\frac{\chi_1^{1-\alpha}\,(1-\chi_1)^{1-\beta}}{c(\alpha,\beta)}\frac{\chi_2^{1-\alpha}\,(1-\chi_2)^{1-\beta}}{c(\alpha,\beta)}, \end{aligned} \quad (12.39)$$

and the joint tilt angle distribution is

$$p(\cos\theta_1, \cos\theta_2 | f_{\rm iso}, \sigma_t) = \frac{f_{\rm iso}}{(1-z_{\rm min})^2} + (1-f_{\rm iso})\mathcal{N}_{[z_{\rm min},1]}(\cos\theta_1|1,\sigma_t)\mathcal{N}_{[z_{\rm min},1]}(\cos\theta_2|1,\sigma_t)\,. \quad (12.40)$$

Figure 12.17 shows results from this alternative model using both GWTC-2 and GWTC-3. For reference, the dashed black distributions show results from our usual `BetaSpike+TruncatedMixture` model. Compared to `BetaSpike+TruncatedMixture`, the `Extended` model has *more* support at $f_{\rm spike} = 0$, strengthening the conclusion that the fraction of black holes with negligible spin is consistent with zero. For the tilt angle distribution, we find that data still has a strong preference for negative $z_{\rm min}$ under the `Extended` model, but with decreased confidence (84.6% quantile for GWTC-2 and 90.4% quantile for GWTC-3) than with `BetaSpike+TruncatedMixture`. Perhaps the starkest difference between results generated with the `Extended` and `BetaSpike+TruncatedMixture` models is that under the `Extended`, the data even more strongly prefer isotropically distributed tilt angles over aligned, as seen by the $f_{\rm iso}$ peaking strongly at unity. This again indicates no strong preference, given current data, that black hole spins are collectively nearly aligned with their orbital angular momenta.

Chapter 13

THE ASTROPHYSICAL DISTRIBUTION OF BINARY BLACK HOLE SPINS IN GWTC-4.0

This chapter is published as Section 6.3 and Appendices B.5, D.2, D.3, and D.4 of Ref. [23]:

LIGO-Virgo-KAGRA Scientific Collaboration incl. **Simona J. Miller**. GWTC-4.0: Population Properties of Merging Compact Binaries. 8 2025. *As an invited member of the 8-person paper writing team,* ***SJM*** *wrote and made figures for everything spin-related (Section 6.3; Appendices B.5, D.2, and D.3; Figures 7-9, 17-19, and 21) and provided detailed feedback and edits on the full paper throughout the writing process. Additionally,* ***SJM*** *was a key analyst, producing effective spin measurements shown in Section 6.3.2 and Appendix D.3, and was the central coordinator of all production, result review, and synthesis spin analyses in the paper.*

The sections of Ref. [23] excluded from this thesis are those that do not reflect ***SJM****'s notable contribution to the work.*

$\sim\sim\sim$

We next present the spin distribution of the BBH population through O4a. We model spins using two different parameterizations: the magnitudes χ_i and tilt angles θ_i (Section 13.1), and the effective inspiral spin $\chi_{\rm eff}$ and effective precessing spin $\chi_{\rm p}$ (Section 13.2). The effective inspiral spin $\chi_{\rm eff}$ is the mass-weighted average of the component spins *aligned* with the binary's orbital angular momentum [193, 289]. The effective precessing spin $\chi_{\rm p}$ characterizes the degree of relativistic precession caused by spin–orbit misalignment, capturing the effect of the *in-plane* spin components [77, 200, 203, 205]; these are defined in Equations 15 and 16 of Ref. [7]. More details about spin parameterization are given in Appendix 13.A.

Previous analyses of the BBH observed through O3 found that BH spin vectors tend to be small in magnitude ($\chi \lesssim 0.3$) and preferentially—but not exclusively—lie above the orbital plane, $\cos\theta > 0$[282, 283, 327]. The new observations in GWTC-4.0 further support these conclusions and additionally suggest more structure in the component and effective spin distributions, enabled by changes to the models used in previous population analyses: *(i)* the

spin magnitude distribution has support at $\chi \approx 0$, *(ii)* the spin tilt distribution may peak away from perfect alignment with the orbital angular momentum, and *(iii)* the $\chi_{\rm eff}$ distribution is asymmetric about its peak. We elaborate upon these new features in the following subsections.

13.1 Spin Magnitudes and Tilts

Spin magnitudes and tilt angles provide insight about BBH formation and evolution [e.g., 1, 2]; we begin with spin magnitudes. If angular momentum transport in stars is efficient, stellar cores rotate slowly, resulting in small spin magnitudes for isolated BH [31, 51, 527]. However, binary interactions can significantly influence BH spins through mechanisms such as tides [53, 54, 58, 80, 83, 528] and accretion [53, 54, 86, 529–531]. If a BBH forms from a pre-existing isolated stellar binary, tidal interactions can spin up the progenitor of the second-born BH [57, 83], yielding spin magnitudes of $\chi \approx 0.2{-}0.4$ [60], although these can be reduced by stellar winds [55]. Chemically homogeneous evolution—involving tidally locked, high-mass, low-metallicity, close binaries—may produce even larger spins [38, 56]. Alternatively, large spin magnitudes may point toward a hierarchical merger origin, where one or both BH are remnants of previous BBH mergers [43, 61, 62, 64–71, 368], which are predicted to have $\chi \sim 0.7$ [73].

With this astrophysical context, we present our measurement of the spin magnitude distribution through O4a. The left column of Figure 13.1 shows the marginal distributions of χ using the strongly modeled approach (Gaussian Component Spins; blue) and weakly modeled approach (B-Spline; green). The Gaussian Component Spins model (Equation 13.1) describes the χ population as a truncated Gaussian distribution. This is a departure from the Default spin model of GWTC-3.0. There, a non-singular Beta distribution was used [283], which forces $p(\chi) = 0$ at $\chi = 0, 1$ and thus cannot measure contributions to the population at near-minimal or near-maximal spins. Allowing for more model flexibility at the χ distribution's boundaries is crucial, as there exists an ongoing discussion in the literature about whether or not there is an over-density of BBH with $\chi \lesssim 0.01$ [164, 165, 273, 334, 368, 515]. We find that the Gaussian Component Spins model is preferred over the Default spin model of GWTC-3.0 by $\log_{10} \mathcal{B} = 0.66$; additional parametric spin magnitude (and tilt) models are discussed in Appendix 13.B. The widening of the 90% credible regions for the χ and $\cos\theta$ distributions at their boundaries under the

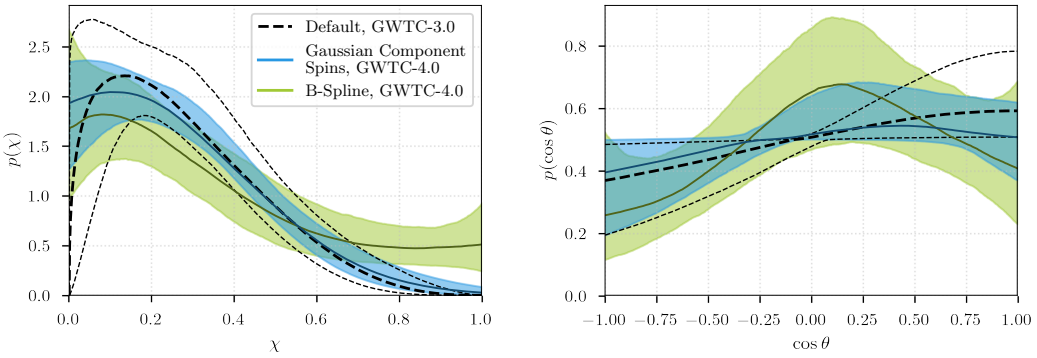


Figure 13.1: Marginal spin magnitude χ (left) and cosine tilt angle $\cos\theta$ (right) distributions under the Gaussian Component Spins (blue) and B-Spline (green) models. The results from the GWTC-3.0 Default are shown for comparison (black dashed); note that this model is different from what is used for GWTC-4.0. The GWTC-3.0 analysis employed a Beta distribution for spin magnitudes rather than a truncated normal, and fixed the location of the Gaussian component of the spin tilt angle distribution to $\cos\theta = 1$ rather than letting it vary freely in inference. The solid lines show the median of the inferred distributions from GWTC-4.0, and the shaded regions show their 90% credible regions. The thick dashed lines show the median from GWTC-3.0, while the thin dashed lines show its 90% credible region.

B-Spline model seen in Figure 13.1 is a prior-driven effect common in spline modeling [e.g., 266, 267].

In GWTC-4.0, we constrain $\mathbf{p(\chi \approx 0) > 0}$ under both the strongly and weakly modeled approaches. At 90% credibility, our recovered spin magnitude distribution peaks between $\chi = 0.01$-0.23, as measured by the μ_χ location parameter. Broadly, BH spins are inferred to be predominantly non-extremal, with the Gaussian Component Spins model finding that 90% of BH having $\chi < 0.57$. The comparative dearth of observed BBH with large spins disfavors a population dominated by second-generation BH. However, the precise fraction of systems with large spin magnitudes is model dependent: the Gaussian Component Spins and B-Spline models only disagree at 90% credibility for $\chi > 0.83$. At $\chi = 0.8$, 0.9, and 1.0, their $p(\chi)$ distributions differ at the 85%, 96%, and 98% levels, respectively. While the Gaussian Component Spins model approaches $p(\chi) \sim 0$ at $\chi = 1$, the B-Spline model infers a larger, nearly flat distribution from $\chi = 0.6$ to 1. Similar behavior was found in GWTC-3.0 [272], and attributed to a subpopulation with near-uniform spin magnitudes. The discrepancy between our two models may arise

from limited flexibility in the strongly modeled approach; a single truncated Gaussian cannot increase the probability at $\chi \gtrsim 0.8$ without also increasing it at $\chi \sim 0.2{-}0.5$. In GWTC-4.0, a handful of events do have preferentially large spin magnitudes, including GW231123 which has primary spin $\chi_1 > 0.5$ with high confidence [21].

Under both the Gaussian Component Spins and B-Spline models, the results presented in Figure 13.1 assume that the spin magnitudes are IID, meaning that the function describing the population distribution is factorizable in terms of χ_1 and χ_2, and the two are described by the same set of hyperparameters. The data prefer identically distributed spin magnitudes over those which are non-identically distributed, cf. Table 13.4. However, mathematically, the primary and secondary spins cannot be IID if the purported correlation between q and $\chi_{\rm eff}$ is true [532]. We probe this correlation in Section 6.5.1 of Ref. [23], and find support for its existence, albeit with evidence that has diminished since GWTC-3.0. Thus, we interpret the Bayes factor in favor of IID spins as a statement that, under the Gaussian Component Spins model, we cannot yet say with statistical certainty that the spins are not identically distributed. This statement is likely driven by the fact that individual-event spin magnitude posterior distributions are typically wide, making χ_1 and χ_2 hard to distinguish. In general, χ_1 and χ_2 are not expected to be identically distributed in nature; if q and $\chi_{\rm eff}$ are indeed correlated, more informative χ_i measurements may eventually reveal their non-identical nature. Figure 13.4 in Appendix 13.B presents posteriors on the Gaussian Component Spins hyperparameters under the assumption that spin magnitudes (and tilts) are identically versus non-identically distributed.

Next, Figure 13.2 shows the inferred spin magnitude distributions if, rather than sorting by the more versus less massive BH, we instead sort by the BH with the larger (subscript A) and smaller (subscript B) spin magnitude [533]. The magnitudes χ_A and χ_B are derived quantities and are not fit independently: to generate their distributions, we take results from the Gaussian Component Spins model and impose order statistics, assuming that χ_A is the larger of two draws from the $p(\chi)$ distribution in the left panel of Figure 13.1 and χ_B is the smaller [283, 533]. This analysis does not assume any sort of pairing function between spins. Spin sorting offers an alternative way to visualize and interpret the spin information from the IID model. The more

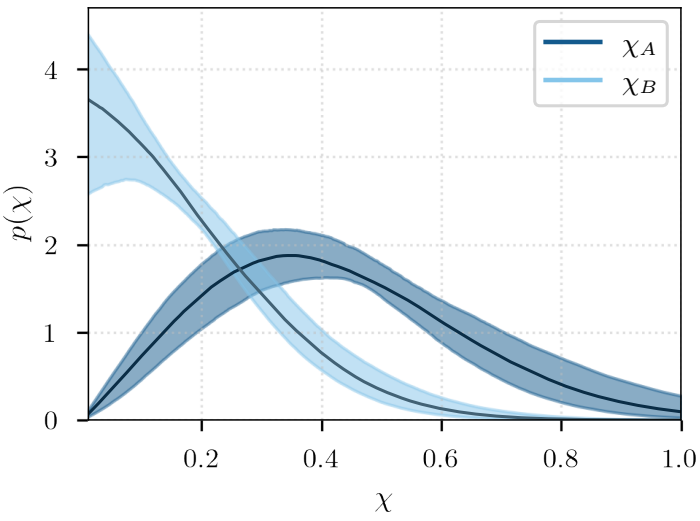


Figure 13.2: Larger (χ_A, darker blue) and smaller (χ_B, lighter blue) spin magnitude distributions, derived from imposing order statistics on the GAUSSIAN COMPONENT SPINS χ distribution shown in Figure 13.1. The solid lines show the median of each inferred distribution, and the shaded regions show the 90% credible intervals.

rapidly spinning component has a wide spin magnitude distribution, with a PPD peaking at $\chi_A = 0.35$, and support up to $\chi_{A,99\%} = 0.92^{+0.04}_{-0.08}$ (value of χ_A at which each $p(\chi_A)$ trace reaches its 99th percentile, serving as a proxy for the distribution's maximum). The vanishing probability at $\chi_A = 0$ is a Jacobian effect of the order statistics. However, if both BH had $\chi \approx 0$, the χ_A distribution would be much more strongly peaked at small values [534], meaning that spin-sorting results on GWTC-4.0 indicate that **at least one BH per binary has $\chi \gtrsim 0$**. The more slowly spinning component, on the other hand, is consistent with a narrower distribution peaking at $\chi_B = 0$, and only has support up to $\chi_{B,99\%} = 0.6^{+0.09}_{-0.07}$. The observation that spin χ_B magnitudes are preferentially small could indicate small natal spins for at least one of the two BH, while the fact that the population is consistent with only one BH per binary having large spin supports the existence of some variety of spin-up mechanism in BBH evolution.

We next turn to the spin tilt distribution. Many authors argue that if BBH are formed primarily in the isolated binary scenario, large tilt angles are hard to explain without invoking large supernovae kicks and inefficient tides [84–89], although others claim that, depending on the specifics of poorly understood supernovae physics, even small kicks can misalign a binary [52, 90]. Complete

anti-alignment ($\cos\theta = -1$) is a possible result of mass transfer in the isolated channel [531]. On the other hand, BBH formed dynamically in stellar clusters are predicted to have isotropically distributed spin orientations, as there is no *a priori* preferential spin direction in these environments [39, 91–93]. However, recent studies indicate that mechanisms could possibly exist for dynamically formed BBH to have slight preference for $\cos\theta > 0$ [110, 112, 113], especially for BBH formed dynamically in the disks of active galactic nuclei [96, 97].

The right-hand column of Figure 13.1 shows the marginal distribution of the cosine of the tilt angle, $\cos\theta$. The $\cos\theta$ population under the Gaussian Component Spins model is a mixture between isotropic and truncated Gaussian sub-populations (Equation 13.2). Following GWTC-3.0 and earlier population analyses, we assume that spin tilt angles are NID under the Gaussian Component Spins model, meaning that while $\cos\theta_1$ and $\cos\theta_2$ are described by the same hyperparameters, the population distribution is *not* separable in terms of the two: we require that both BH in a binary are drawn from the same sub-population (either the isotropic or the truncated Gaussian). NID tilts are favored by the data over non-identical distribution, cf. Table 13.4. The B-Spline model naturally assumes spin tilt angles are IID, as it does not probe separate tilt sub-populations.

The Default model of GWTC-3.0 fixed the location of the Gaussian sub-population at exact spin–orbit alignment [283, 328, 351], making it a half-Gaussian peaking at $\cos\theta = 1$. In GWTC-4.0, the strongly and weakly modeled approaches both find that **the spin tilt distribution may peak away from exact spin–orbit alignment**. The Gaussian Component Spins model finds that the $\cos\theta$ distribution reaches its maximum between -0.36 and 0.94 at 90% credibility; consistently, the B-Spline model finds the peak to lie between -0.23 and 0.96. The possibility that the tilt angle distribution peaks away from alignment is also found in GWTC-3.0 under various models which permit such a feature [266, 267, 318]. It is not impossible, however, for the peak of a $\cos\theta$ distribution to be inferred away from alignment even when the true underlying population does peak at $\cos\theta = 1$ [535]. The inferred peak of the GWTC-4.0 $\cos\theta$ distribution is more strongly constrained away from unity than nearly all of those spuriously found with simulated catalogs of the same size [cf. Figure 13.4 with Figure 4 of Ref. 535].

Under both the strongly and weakly modeled approaches, the $\cos\theta$ distribu-

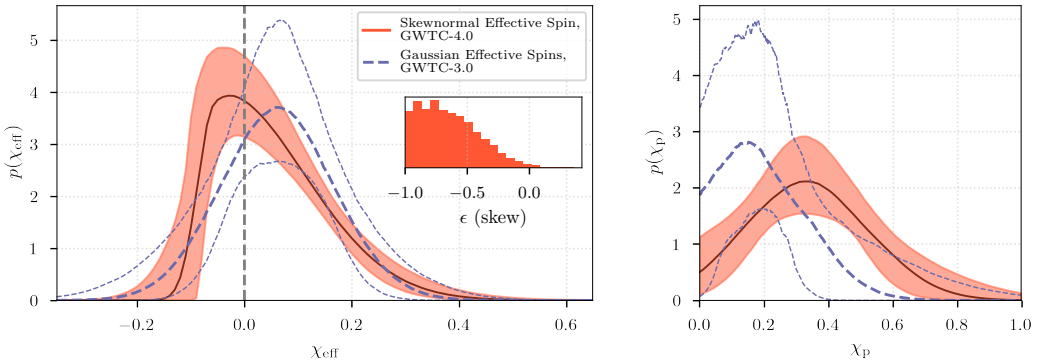


Figure 13.3: Marginal $\chi_{\rm eff}$ (left panel) and $\chi_{\rm p}$ (right panel) distributions. The solid lines show the median of each inferred distribution, and the shaded regions show the 90% credible intervals. The GWTC-4.0 results under the Skew-normal Effective Spin model are shown in red; the GWTC-3.0 results under the Gaussian Effective Spins model are shown in purple dashed. The histogram inset in the left panel shows the skew parameter ϵ for the Skew-normal Effective Spin model. There is significant preference for a skewed $\chi_{\rm eff}$ distribution, indicated by $\epsilon < 0$ with 99.3% credibility.

tion shown in Figure 13.1 has support across a wide range of spin tilts, with a slightly larger fraction of positive $\cos\theta$ compared to negative. Given this broad support, we next ask the question: what is the lowest spin tilt angle that is absolutely required to fit GWTC-4.0 reasonably? To probe this, we use a model which is similar to the Gaussian Component Spins model but with $p(\cos\theta < t_{\rm min}) = 0$ for an inferred value $t_{\rm min}$, where $t \equiv \cos\theta$ [164, 334]; see Equation (13.3). We find $t_{\rm min} = 0.71^{+0.15}_{-0.19}$ at 90% credibility. That the posterior for the minimum required $\cos\theta$ is inconsistent with -1 but remains confidently negative is a somewhat unexpected astrophysical result. If most BBH form in isolation, a minimum tilt cutoff is plausible but would be expected to be closer to zero, potentially even positive. Conversely, dynamically formed BBH should exhibit isotropic tilts extending down to $\cos\theta = -1$. Our observed tilt distribution therefore does not preclude either broad scenario of isolated or dynamical formation—or a mixture of both. However, the fact that the inferred minimum required tilt is significantly negative suggests the contribution of a dynamical formation channel to the BBH population. We further discuss the astrophysical interpretation of negative spin tilts and potential sub-populations in Section 13.2 in the context of effective spins.

13.2 Effective Spins

We next turn to the inferred distributions of the effective spins $\chi_{\rm eff}$ and $\chi_{\rm p}$. Figure 13.3 shows the marginal distributions of $\chi_{\rm eff}$ (left) and $\chi_{\rm p}$ (right) under the Skew-normal Effective Spin model (red solid; Equation 13.12), compared to the Gaussian Effective Spins model result from GWTC-3.0 (purple dashed; Equation 13.11). The Skew-normal Effective Spin model [536] differs from the previously-used Gaussian Effective Spins in two ways. First, it allows for asymmetry in the $\chi_{\rm eff}$ marginal distribution. Second, the marginal $\chi_{\rm p}$ distribution is a truncated Gaussian which is not correlated with $\chi_{\rm eff}$. On GWTC-4.0 data, the Gaussian Effective Spins model finds that $\chi_{\rm eff}$ and $\chi_{\rm p}$ are preferentially uncorrelated, although this conclusion depends on analysis settings; see Figure 13.6 in Appendix 13.C and the discussion therein.

We find that the $\chi_{\rm eff}$ distribution is skewed and asymmetric about zero with more support for positive values. Asymmetry can be directly probed with the ϵ skewness parameter of the Skew-normal Effective Spin model, defined in Equation (13.12). We find that the skewness ϵ is less than 0 (symmetry) with 99.3% credibility, as can be seen in the inset of the left panel of Figure 13.3. This corresponds to a wider $\chi_{\rm eff}$ distribution to the right of the peak, and a narrower distribution to the left, and is known as *positive skew*. As discussed in Section 6.5.1 of Ref. [23], asymmetry is also found when allowing for a linear or spline correlation between $\chi_{\rm eff}$ and mass ratio. Results for the GWTC-4.0 $\chi_{\rm eff}$ distribution measured with more models, including the Gaussian Effective Spins model and with the weakly modeled approach are presented in Figures 13.6 and 13.7 in the Appendix.

We now discuss BBH with spin tilts lying below the orbital plane. These interesting probes of dynamical BBH formation have $\cos\theta < 0$ and thus negative $\chi_{\rm eff}$. In Table 13.1, we present 90% bounds for three quantities which probe negative $\chi_{\rm eff}$ in the population, using both the Skew-normal Effective Spin and Gaussian Effective Spins models for a straightforward comparison to GWTC-3.0. First, we report the first percentile of $\chi_{\rm eff}$ distribution, which serves as a proxy for the distribution's minimum without necessitating the inclusion of sharp features in the distribution itself [164], constraining it to fall between -0.27 and -0.08. The Gaussian Effective Spins model finds more extremal minimum spins than the Skew-normal

Table 13.1: Summary of probes of spin misalignment, as measured by $\chi_{\rm eff}$.

Model	$\chi_{\rm eff,1\%}$	Fraction $\chi_{\rm eff} < 0$	HM Fraction
Gaussian Effective Spins (GWTC-3.0)	$-0.18^{+0.09}_{-0.12}$	$0.28^{+0.12}_{-0.13}$	$< 3.1 \times 10^{-2}$
Gaussian Effective Spins (GWTC-4.0)	$-0.2^{+0.06}_{-0.07}$	$0.34^{+0.09}_{-0.1}$	$< 1.9 \times 10^{-2}$
Skew-Normal Effective Spin (GWTC-4.0)	$-0.11^{+0.04}_{-0.08}$	$0.34^{+0.06}_{-0.08}$	$< 1.3 \times 10^{-4}$

Note— Summary of probes of spin misalignment, as measured by $\chi_{\rm eff}$. We give 90% credible intervals for the $\chi_{\rm eff}$ value of the first percentile of the distribution ($\chi_{\rm eff,1\%}$), serving as a proxy for the minimum $\chi_{\rm eff}$, and the fraction of $\chi_{\rm eff} < 0$. The HM fraction provides an upper limit to the fraction of BBH of hierarchical merger origin, equal to 1/0.16 times fraction of systems with $\chi_{\rm eff} < -0.3$ [537, 538]; we provide its 90% upper limit. Posteriors for these quantities for these and other models are plotted in Figure 13.7.

Effective Spin model, likely due to its required symmetry about its peak; fitting the positive side of the distribution forces a longer tail into the negative region. Second, we report the fraction of the population with negative $\chi_{\rm eff}$ to be 0.24–0.42, with the two models finding nearly identical results. Assuming isotropy, an upper bound on the fraction of BBH formed dynamically in gas-free environments can be placed by doubling the fraction of negative $\chi_{\rm eff}$ [Equation 7 of Ref. 282]; we thus find that at most 84% of BBH form dynamically. Third, we give the 90% upper limit on the fraction of BBH coming from the hierarchical merger (HM) scenario. The HM fraction is bounded by the consideration that $\sim 16\%$ of BBH coming from the HM formation channel will have $\chi_{\rm eff} < -0.3$ [537, 538]. With the Skew-normal Effective Spin and Gaussian Effective Spins models, we limit this fraction to $\lesssim 3\%$. Figure 13.7 in Appendix 13.D shows posterior distributions on the quantities in Table 13.1 for the models presented here and those using the weakly modeled approach; with more flexible models, we more generously limit the HM fraction to $\lesssim 74\%$.

It is possible that the support for $\chi_{\rm eff} < 0$ may be a byproduct of the models fitting for a peak at $\chi_{\rm eff} \sim 0$ without having the flexibility for a sharp decline beyond $\chi_{\rm eff} < 0$. This type of sharp feature could occur if component spin magnitudes cluster around $\chi \sim 0$ but their tilts seldom reach below $\cos\theta = 0$. In Section 13.1, we address this by directly fitting for the minimum required $\cos\theta$ which is found to be confidently negative, and corresponds to a fraction 0.41 ± 0.05 of systems with spins laying below the orbital plane. These probes of negative spin indicate that **between ∼20–40% the BBH population has**

spins which are more than 90 degrees misaligned with the orbital angular momentum.

It is unlikely that the entire observed BBH population originates from a single formation channel [35, 45, 276, 539, 540]. Features in our spin distributions indeed suggest the presence of multiple sub-populations. A purely random spin channel would produce a symmetric distribution around $\chi_{\rm eff} = 0$ [91, 93, 329]. However, for all models, the observed $\chi_{\rm eff}$ distribution is asymmetric about zero, c.f., Figure 13.7. This asymmetry suggests contribution from a preferentially aligned subpopulation [85, 536, 541]. The skew observed in the SKEW-NORMAL EFFECTIVE SPIN, as well as the $(q, \chi_{\rm eff})$ LINEAR and $(q, \chi_{\rm eff})$ SPLINE models (see Section 6.5.1 of Ref. [23] and Figure 13.7) further supports the presence of an aligned component, either as a sub-dominant sub-population or a dominant sub-population with small spin magnitudes. We find a preference for small spin magnitude (Figure 13.1), perhaps favoring the latter.

Finally, we discuss the effective precessing spin $\chi_{\rm p}$, which can provide additional insight about spin precession in the population. The right panel of Figure 13.3 shows the marginal inferred $\chi_{\rm p}$ distribution (red). We find that precession exists on a population level, as our models do not support $(\mu_p, \sigma_p) = (0, 0)$ at high credibility. The GWTC-4.0 results support larger $\chi_{\rm p}$ than GWTC-3.0, shown for comparison in Figure 13.3 (purple dashed). However, different likelihood convergence criteria (Appendix A.1 in Ref. [23]) were used between GWTC-4.0 and GWTC-3.0, meaning these results are not directly comparable. Under the less-stringent GWTC-3.0 convergence criterion, the $\chi_{\rm p}$ distribution inferred from GWTC-4.0 data is more consistent with that found in GWTC-3.0, albeit still with less support for low $\mu_{\rm p}$ and low $\sigma_{\rm p}$. The inferred $\chi_{\rm p}$ distribution is more dependent on analysis settings, specifically the threshold on the variance of the log-likelihood, than other parameters. This is largely because the individual-event prior has no support at $\chi_{\rm p} = 0$ making it technically difficult to reweight individual-event posteriors to the population. We discuss the sensitivity of the $\chi_{\rm p}$ distribution to analysis settings in detail in Appendix 13.C; see Figures 13.5 and 13.6.

Ch. 13 Appendices

13.A Spin Models

BBH spins can be parameterized in several ways which are useful for GW data analysis. Here, we use dimensionless spin magnitudes for the BH χ_1 and χ_2, and cosine tilt angles $\cos\theta_1$ and $\cos\theta_2$, as well as the effective spin parameters $\chi_{\rm eff}$ [193, 194, 289] and $\chi_{\rm p}$ [77, 200, 203]. The effective inspiral spin $\chi_{\rm eff}$ used because it is typically the most precisely measured BBH spin parameter, due to its lower-order appearance post-Newtonian expansions [188]. Other parametrizations of spin precession exist [e.g., 201, 204, 205] but for reasons of convention, we only work with the typical parameterization given in Equation 16 of Ref. [7]. When modeling the effective spins, we use the analytic per-event parameter-estimation prior on $\chi_{\rm eff}$, $\chi_{\rm p}$, and q [239] in the calculation of the hierarchical likelihood. For all models presented in this work, we assume that azimuthal angles ϕ_i are distributed uniformly between 0 and 2π—the same as their parameter estimation prior—due to their typically uninformative individual-event posteriors. See Table 3 of Ref. [7] for more information about spin parameters.

We report results with spin vectors defined at the reference frequencies used for inference on each signal [8]. The effective spin $\chi_{\rm eff}$ is approximately conserved throughout the inspiral [78, 193], so its population distribution should be unaffected by the choice of reference frequency. While the effective precessing spin $\chi_{\rm p}$ and the spin angles are dependent on reference frequency, they are comparatively weakly constrained by the data. Additionally, their parameter estimation priors are invariant under time evolution, meaning measurements are robust against different reference frequencies. At the current number of events and because of model dependence while measuring tilt distributions, we do not expect significant differences between the inferred tilt or $\chi_{\rm p}$ distributions at different reference frequencies [47]. Our approach is consistent with past LVK analyses, where evolved spins have not been used [282, 283, 327].

Component Spin Models

Gaussian Component Spins: We model the spin magnitudes (χ_i) as a truncated Gaussian distribution between 0 and 1, assuming they are identically and independently distributed:

$$\pi(\chi_i|\mu_\chi,\sigma_\chi) = \mathcal{N}_{[0,1]}(\chi_1|\mu_\chi,\sigma_\chi)\mathcal{N}_{[0,1]}(\chi_2|\mu_\chi,\sigma_\chi)\,. \tag{13.1}$$

Table 13.2: Summary of GAUSSIAN COMPONENT SPINS model parameters for spin magnitudes (Equation 13.1) and tilt angles (Equation 13.2, 13.3).

Parameter	Description	Prior
μ_χ	Location of the χ distribution	U(0, 1)
σ_χ	Width of the χ distribution	U(0.005, 1)
μ_t	Location of the Gaussian component of the $\cos\theta$ distribution	U(−1, 1)
σ_t	Width of the Gaussian component of the $\cos\theta$ distribution	U(0.01, 4)
ζ	Fraction in the Gaussian component of the $\cos\theta$ distribution	U(0, 1)
$t_{\rm min}$	Minimum of the $\cos\theta$ distribution	U(−1, 1)

NOTE— U stands for a uniform prior.

This choice attempts to rectify shortcomings of the non-singular Beta distribution used to model the spin magnitudes in previous work [282, 283, 327]; see Equation (13.4) below. The non-singular Beta distribution is forced to go to $\pi(\chi) = 0$ at $\chi = 0, 1$, which crucially does not allow for measurements of contributions to the population at $\chi = 0$ or $\chi = 1$, even though non-trivial contributions may exist [164, 165, 273]. Allowing the Beta distribution to be singular would add limited additional model flexibility, only adding the option for $\pi(\chi) = \infty$ at the $\chi = 0, 1$, but not anywhere in between 0 and ∞. Thus, we opt for the truncated Gaussian model, which can take on a continuous range of values at χ's boundaries.

We model the distribution of the cosine spin tilt angle ($\cos\theta_i$) as a mixture between a Gaussian distribution truncated on −1 to 1 and an isotropic distribution, assuming they are identically but *not* independently distributed:

$$\pi(\cos\theta_i|\mu_t, \sigma_t, \zeta) = \zeta\,\mathcal{N}_{[-1,1]}(\cos\theta_1|\mu_t, \sigma_t)\mathcal{N}_{[-1,1]}(\cos\theta_2|\mu_t, \sigma_t) + \frac{1-\zeta}{4}\,. \quad (13.2)$$

We here allow for the location of the Gaussian sub-population to vary, following [318], rather than fixing it at $\mu_t = 1$ as was done in previous work [282, 283, 327]. Priors on the GAUSSIAN COMPONENT SPINS hyperparameters are given in Table 13.2.

MINIMUM TILT MODEL: For the spin tilts, we also use a model in which $t_{\rm min}$ (where $t \equiv \cos\theta$), a hard lower truncation of $\pi(\cos\theta)$, is informed by the data.

In this case the population distribution becomes

$$\pi(\cos\theta_i|\mu_t,\sigma_t,\zeta,t_{\rm min}) = \zeta\,\mathcal{N}_{[t_{\rm min},1]}(\cos\theta_1|\mu_t,\sigma_t)\mathcal{N}_{[t_{\rm min},1]}(\cos\theta_2|\mu_t,\sigma_t) + \frac{1-\zeta}{(1-t_{\rm min})^2} \quad (13.3)$$

if $\cos\theta_i > t_{\rm min}$ and 0 if $\cos\theta_i \leq t_{\rm min}$. Equation (13.2) is the same as Equation (13.3) with $t_{\rm min} = -1$. The prior on $t_{\rm min}$ and other hyperparameters are given in Table 13.2.

Beta Distribution Spin Magnitude: In Appendix 13.B, we compare the Gaussian Component Spins model to alternatives. For the spin magnitudes, these are the Constrained and Unconstrained Beta Distributions, which both follow the form:

$$\pi(\chi_i|\alpha,\beta) \propto {\chi_i}^{\alpha-1}(1-\chi_i)^{\beta-1}\ . \quad (13.4)$$

The Constrained Beta was the Default model for GWTC-3.0 [283] and requires the shape parameters $\alpha,\beta > 1$, which forces $\pi(\chi_i = 0,1) = 0$ [282, 505]. The Unconstrained Beta relaxes this constraint on the the shape parameters. If $0 < \alpha,\beta < 1$, the Beta distribution becomes singular, meaning $\pi(\chi_i = 0,1) = \infty$. Following [283], we sample the mean and standard deviation of the Beta distribution, rather than α and β, and use the same priors for μ_χ and σ_χ as given in Table 13.2 for the Gaussian Component Spins model. For the Constrained Beta case, we additionally impose the cut $\alpha,\beta > 1$. The relationship between $\{\mu_\chi,\ \sigma_\chi\}$ and $\{\alpha,\ \beta\}$ is given in Equation 5 of [327].

Identical versus Non-identically Distributed Spin Magnitudes & Tilts

In Appendix 13.B, we investigate whether or not the primary and secondary spins are identically distributed, assuming the Gaussian Component Spins model. *Identical* distribution means that the two distributions share the same set of hyperparameters, while *non-identical* means that each is described by different hyperparameters. Spin magnitudes are also *independently* distributed, meaning that $\pi(\chi_1,\chi_2)$ is separable in terms of χ_1 and χ_2. The acronym IID means they are independently and identically distributed, while IND means independently and non-identically distributed:

$$\chi_i \text{ IID} \implies \pi(\chi_{1,2}|\mu_\chi,\sigma_\chi) = \mathcal{N}_{[0,1]}(\chi_1|\mu_\chi,\sigma_\chi)\,\mathcal{N}_{[0,1]}(\chi_2|\mu_\chi,\sigma_\chi)\ , \quad (13.5)$$

$$\chi_i \text{ IND} \implies \pi(\chi_{1,2}|\mu_{\chi,1},\sigma_{\chi,1},\mu_{\chi,2},\sigma_{\chi,2}) = \mathcal{N}_{[0,1]}(\chi_1|\mu_{\chi,1},\sigma_{\chi,1})\,\mathcal{N}_{[0,1]}(\chi_2|\mu_{\chi,2},\sigma_{\chi,2})\ . \quad (13.6)$$

Table 13.3: Summary of GAUSSIAN EFFECTIVE SPINS (Equation 13.11) and SKEW-NORMAL EFFECTIVE SPIN (Equation 13.12) spin parameters.

Parameter	Description	Prior **G**	Prior **SN**
$\mu_{\rm eff}$	Location of the $\chi_{\rm eff}$ distribution	U(−1, 1)	U(−1, 1)
$\sigma_{\rm eff}$	Width of the $\chi_{\rm eff}$ distribution	U(0.05, 1)	LU(0.01, 4)
μ_p	Location of the χ_p distribution	U(0.05, 1)	U(0.01, 1)
σ_p	Width of the χ_p distribution	U(0.07, 1)	LU(0.01, 1)
ρ	Degree of correlation between $\chi_{\rm eff}$ and χ_p	U(−0.75, 0.75)	N/A
ϵ	Skew of the $\chi_{\rm eff}$ distribution	N/A	U(−1, 1)

NOTE— The first column of priors (**G**) gives those used for the GAUSSIAN EFFECTIVE SPINS results, the second column (**SN**) is SKEW-NORMAL EFFECTIVE SPIN. U stands for a uniform prior; LU for log-uniform.

The tilt angles, on the other hand, are *non-independently* distributed, meaning that $\pi(\cos\theta_1, \cos\theta_2)$ is not separable. The acronynm NID means they are non-independently but identically distributed, while NND means non-independently and non-identically distributed:

$$\cos\theta_i \text{ NID} \implies \pi(\cos\theta_{1,2}|\zeta, \mu_t, \sigma_t) \tag{13.7}$$
$$= \zeta\,\mathcal{N}_{[-1,1]}(\cos\theta_1|\mu_t, \sigma_t)\,\mathcal{N}_{[-1,1]}(\cos\theta_2|\mu_t, \sigma_t) + \frac{1-\zeta}{4}\ ,$$
$$\cos\theta_i \text{ NND} \implies \pi(\cos\theta_{1,2}|\zeta, \mu_{t,1}, \sigma_{t,1}, \mu_{t,2}, \sigma_{t,2}) \tag{13.8}$$
$$= \zeta\,\mathcal{N}_{[-1,1]}(\cos\theta_1|\mu_{t,1}, \sigma_{t,1})\,\mathcal{N}_{[-1,1]}(\cos\theta_2|\mu_{t,2}, \sigma_{t,2}) + \frac{1-\zeta}{4}\ .$$

For the non-identically distributed cases, each component's hyperparameters have the same priors as those listed in Table 13.2 for the identically distributed case. Table 13.4 gives Bayes factors between different combinations of IID/IND spin magnitudes and NID/NND spin tilts; see associated dicsussion in Appendix 13.B.

Effective Spin Models

GAUSSIAN EFFECTIVE SPINS: We here assume that the distribution of $\chi_{\rm eff}$ and $\chi_{\rm p}$ across the BBH population is a bivariate truncated Gaussian [162, 504] characterized by the location and width of the $\chi_{\rm eff}$ and $\chi_{\rm p}$ distributions, and the covariance between them:

$$\pi(\chi_{\rm eff}, \chi_{\rm p}|\mu, \mathbf{\Sigma}) \propto \mathcal{N}(\chi_{\rm eff}, \chi_{\rm p}|\mu, \mathbf{\Sigma})\ , \tag{13.9}$$

where $\mu = (\mu_{\rm eff}, \mu_{\rm p})$ and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{\rm eff}^2 & \rho\,\sigma_{\rm eff}\,\sigma_{\rm p} \\ \rho\,\sigma_{\rm eff}\,\sigma_{\rm p} & \sigma_{\rm p}^2 \end{pmatrix} \ . \tag{13.10}$$

This expands to:

$$\pi(\chi_{\rm eff}, \chi_{\rm p}|\mu, \mathbf{\Sigma}) \propto \\ \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(\chi_{\rm eff}-\mu_{\rm eff})^2}{\sigma_{\rm eff}^2} - \frac{2\rho(\chi_{\rm eff}-\mu_{\rm eff})(\chi_{\rm p}-\mu_{\rm p})}{\sigma_{\rm eff}\sigma_{\rm p}} + \frac{(\chi_{\rm p}-\mu_{\rm p})^2}{\sigma_{\rm p}^2}\right)\right] \ . \tag{13.11}$$

The bivariate Gaussian is truncated over the range $\chi_{\rm eff} \in [-1, 1]$, $\chi_{\rm p} \in [0, 1]$ and is normalized numerically. Priors on the hyperparameters are given in Table 13.3.

Skew-normal Effective Spin: To account for observational evidence that the $\chi_{\rm eff}$ distribution is not symmetric [365, 536, 542], we additionally model $\chi_{\rm eff}$ as a skewed, truncated Gaussian distribution:

$$\pi(\chi_{\rm eff}|\mu_{\rm eff}, \sigma_{\rm eff}, \epsilon) \propto \begin{cases} (1+\epsilon)\,\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|\mu_{\rm eff}, \sigma_{\rm eff}(1+\epsilon)) & \chi_{\rm eff} \leq \mu_{\rm eff} \ , \\ (1-\epsilon)\,\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|\mu_{\rm eff}, \sigma_{\rm eff}(1-\epsilon)) & \chi_{\rm eff} \geq \mu_{\rm eff} \ . \end{cases} \tag{13.12}$$

This model includes a parameter ϵ which describes the skew of the distribution, such that $\epsilon > 0$ characterizes a distribution with more support for $\chi_{\rm eff} < \mu_{\rm eff}$, while $\epsilon < 0$ has more support when $\chi_{\rm eff} > \mu_{\rm eff}$. The Skew-normal Effective Spin distribution reduces to a standard, symmetric Gaussian in the case that $\epsilon = 0$. We here model $\chi_{\rm p}$ as a truncated normal distribution and infer its location and width. Priors on the hyperparameters are given in Table 13.3.

13.B Model Comparison Study: Spin Magnitudes and Tilt Angles

We consider two additional spin magnitude models beyond the Gaussian Component Spins model from Section 13.1: the Constrained Beta and Unconstrained Beta models, both defined by Equation (13.4). We also consider one additional spin tilt model, the Isotropic + Aligned Gaussian, which fixes $\mu_t = 1$ in the Gaussian Component Spins model and served as the Default in GWTC-3.0 [283, 351]. The first five rows of Table 13.4 give the log Bayes factors between these models. Within the Gaussian Component Spins model, we also test whether the primary and secondary spins are identically and independently distributed (see Appendix 13.A), with log Bayes factors given in the bottom three rows of Table 13.4. The Gaussian Component Spins model with IID spin magnitudes and NID spin tilts performs the best, and is adopted as the default component spin model in the main text.

Table 13.4: Comparison of different parametric models for spin magnitudes χ_i and tilt angles $\cos\theta_i$.

χ_i model		$\cos\theta_i$ model		$\log_{10}\mathcal{B}$
Truncated Gaussian ★	IID	Isotropic + Truncated Gaussian	NID	0.0
Constrained Beta	IID	Isotropic + Aligned Gaussian	NID	−0.66
Constrained Beta	IID	Isotropic + Truncated Gaussian	NID	−0.64
Unconstrained Beta	IID	Isotropic + Truncated Gaussian	NID	−0.98
Truncated Gaussian	IID	Isotropic + Aligned Gaussian	NID	−0.08
Truncated Gaussian	IND	Isotropic + Truncated Gaussian	NID	−1.09
Truncated Gaussian	IID	Isotropic + Truncated Gaussian	NND	−0.17
Truncated Gaussian	IND	Isotropic + Truncated Gaussian	NND	−0.91

NOTE— The first row is the Default model for GWTC-4.0 (GAUSSIAN COMPONENT SPINS), while the second is the Default that was used for GWTC-3.0 and earlier [282, 283, 327, 351]. The third through final are other models explored. All log Bayes factors ($\log_{10}\mathcal{B}$) are with respect to the first row.

Figure 13.4 shows posteriors for the magnitude and tilt hyperparameters assuming that χ_i and $\cos\theta_i$ are both identically distributed (gray) versus both non-identically distributed (blue and red). The primary and secondary spins have consistent hyperparameter distributions. Assuming identical distribution makes the hyperparameters more-precisely constrained—yielding tighter 90% bands on the resultant population distributions—but does not affect the overall shape of the distributions. The only difference of note is that χ_2 has less support for a small μ_χ than χ_1, potentially hinting at more highly spinning secondary BHs on a population level.

13.C Dependence of the Effective Spin Distribution on the Likelihood Variance Cut

We next investigate the effect of the likelihood variance threshold (see Appendix A.1 in Ref. [23]) on our strongly modeled effective spin results. To ensure that posteriors generated using Monte Carlo estimation are trustworthy, i.e., the likelihood estimate is converged, we enforce that all hyperparameter samples yield a log-likelihood variance ${\sigma^2}_{\ln\hat{\mathcal{L}}} < 1$ for all analyses presented in the main text [281]. In GWTC-3.0, however, a different quantity was used to assess the Monte Carlo uncertainty: the effective number of independent

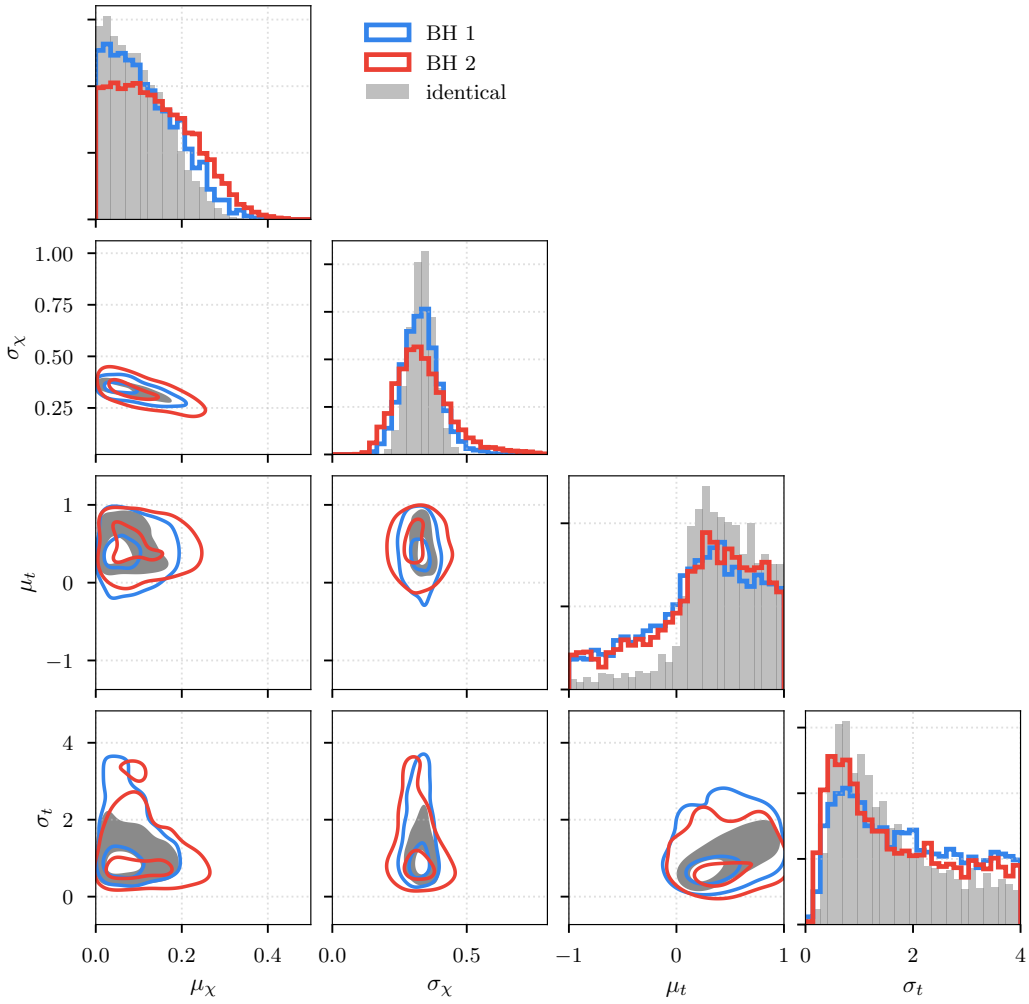


Figure 13.4: Distribution of Gaussian Component Spins hyperparameters for the primary BH (blue) and secondary BH (red) when neither magnitudes nor tilts are assumed identical (last row of Table 13.4), compared to when they are both assumed identical (gray, first row of Table 13.4). The contours of the two dimensional distributions mark the 50th and 90th percentiles.

samples, $N_{\rm eff}$ [280]. It was there imposed that $N_{\rm eff}$ be greater than $4N_{\rm events}$ for the sensitivity injections used to calculate the detection efficiency and greater than 10 for every event in the catalog of CBCs. This $N_{\rm eff}$ cut is generally less stringent than the approach taken for GWTC-4.0. The $\chi_{\rm p}$ distribution is sensitive to which method is used to cut out samples.

For the most direct comparison to GWTC-3.0, we here use the Gaussian Effective Spins model. Figure 13.5 shows the posteriors for the Gaussian Effective Spins parameters where the ${\sigma^2}_{\ln\hat{\mathcal{L}}}$ (purple) versus $N_{\rm eff}$ (green) cuts are done on the GWTC-4.0 results. The LVK GWTC-3.0 results (with the $N_{\rm eff}$ cut) are shown in comparison (black). Figure 13.6 shows the resultant marginal and joint $\chi_{\rm eff}$–$\chi_{\rm p}$ population distributions. There are three main differences between the results with the two cuts on GWTC-4.0 data:

1. The ${\sigma^2}_{\ln\hat{\mathcal{L}}}$ cut yields a $\mu_{\rm p}$ posterior which is constrained away from zero, while the the $N_{\rm eff}$ cut does not.

2. The $N_{\rm eff}$ cut allows for wider $\chi_{\rm p}$ distributions (larger $\sigma_{\rm p}$) than the ${\sigma^2}_{\ln\hat{\mathcal{L}}}$ cut.

3. Under the $N_{\rm eff}$ cut, $\chi_{\rm eff}$ and $\chi_{\rm p}$ are preferentially positively correlated, while under the ${\sigma^2}_{\ln\hat{\mathcal{L}}}$ cut we remain agnostic, with preference for small-to-zero correlation.

Thus, the claims that the $\chi_{\rm p}$ distribution does not peak at zero, and that the data prefer $\chi_{\rm eff}$ and $\chi_{\rm p}$ being uncorrelated at the population level, are not necessarily astrophysical in origin. Rather, they are driven by regions of the parameter-space that our events and sensitivity injections allow us reliably probe with Monte Carlo likelihood estimators.

The particular sensitivity of the $\chi_{\rm p}$ distribution to these cuts can be at-least partly attributed to the use of a uniform and isotropic spin prior in GWTC-4.0 parameter estimation, which induces a $\chi_{\rm p}$ prior that goes to 0 at $\chi_{\rm p} = 0, 1$. Thus, the hierarchical likelihood is prone to yield a small $N_{\rm eff}$ and/or large ${\sigma^2}_{\ln\hat{\mathcal{L}}}$ when $\chi_{\rm p}$ is near-minimal or near-maximal. Using alternative spin priors in parameter estimation could aid in getting more effective samples at small $\chi_{\rm p}$, improving our ability to probe the presence or absence of negligibly precessing BH at the population level. The Monte Carlo uncertainty from working with a finite number of samples to estimate the likelihood will only grow more

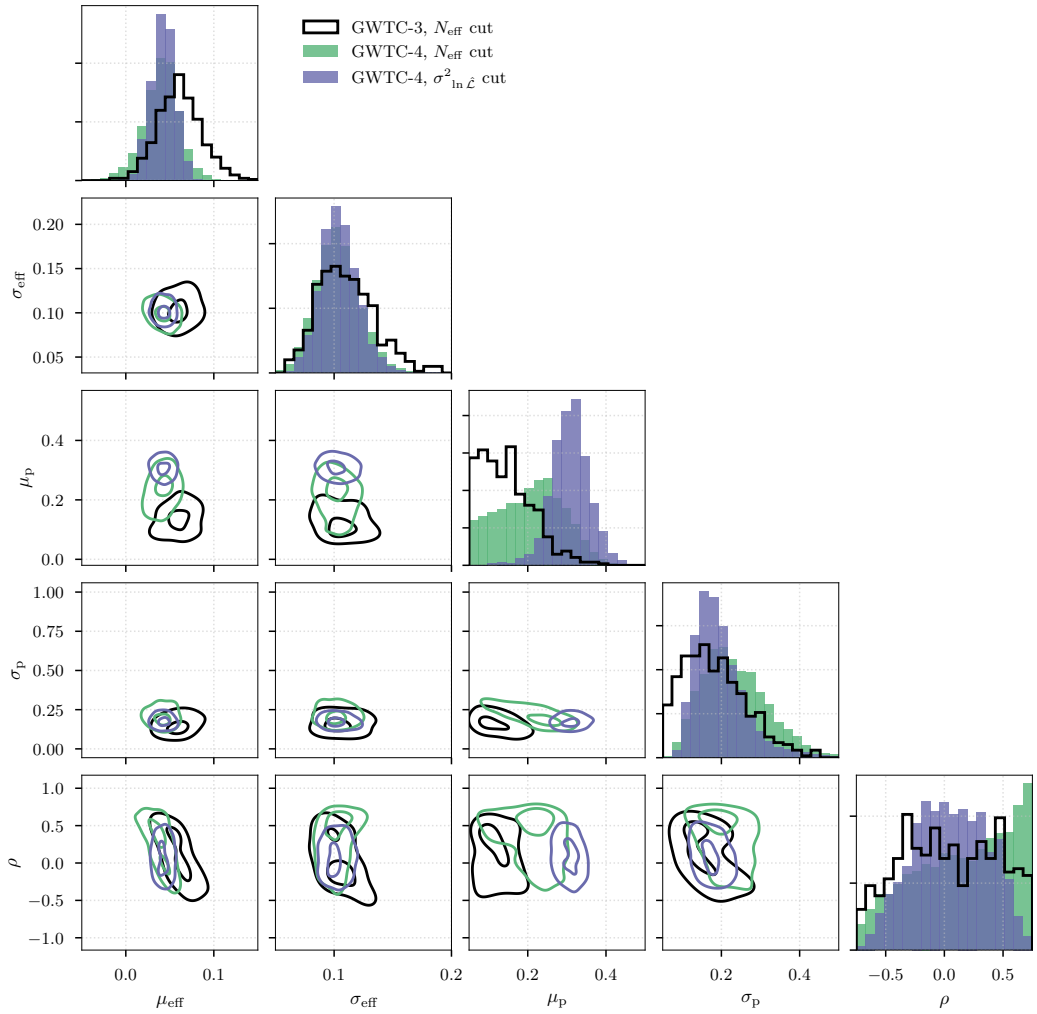


Figure 13.5: Posterior for the Gaussian Effective Spins model hyperparameters, as defined in Table 13.3, for GWTC-4.0 under two methods of cutting out samples that may lead to an unconverged likelihood. Posteriors excluding samples with a substantially large log-likelihood variance ($\sigma^2_{\ln\hat{\mathcal{L}}}$ cut) are shown in purple; those excluding samples with a substantially small number of effective samples ($N_{\rm eff}$ cut) are in green. In black are the GWTC-3.0 results with the $N_{\rm eff}$ cut for comparison. The $\chi_{\rm eff}$ hyperparameters are not affected by the cuts, while the $\chi_{\rm p}$ and joint-distribution hyperparameters are. The contours of the two dimensional distributions mark the 50th and 90th percentiles.

problematic as the number of observed CBC increases [281]; development of mitigation techniques is an area of active research [e.g., 164, 273, 274, 277, 278, 281, 543–546].

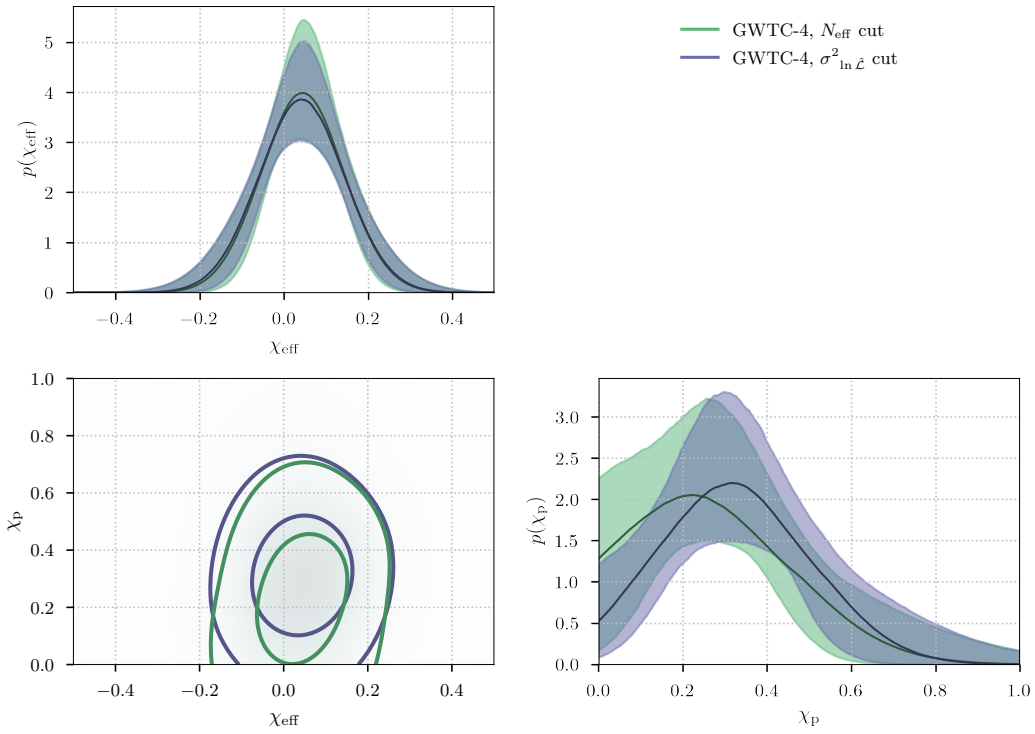


Figure 13.6: Marginal and joint $\chi_{\rm eff}$ and $\chi_{\rm p}$ distributions under the Gaussian Effective Spins model using two methods of cutting out hyperparameter samples that may lead to an unconverged likelihood (see Figure 13.5). The marginal distributions show the median (solid line) and 90% credible interval on the probability density for each population distribution (shaded). The joint distribution is the PPD with contours marking the 50th and 90th quantiles.

13.D Comparison of Weakly Modeled Approach Results

In addition to the Skew-normal Effective Spin model and various correlation models presented for $\chi_{\rm eff}$ in the main text, we use the weakly modeled approach for a consistency check and find the distributions shown in the top panel of Figure 13.7. We plot the $\chi_{\rm eff}$ distribution directly inferred with a binned Gaussian process (BGP; blue) as well as that reconstructed from the spin magnitude and tilt B-Spline model results (green), and compare them to the Skew-normal Effective Spin (red-orange), Gaussian Effective Spins (purple), and $(q, \chi_{\rm eff})$ Spline (maroon) results from the main text. The two approaches paint the same qualitative picture: the $\chi_{\rm eff}$ distribution peaks at small values. However, there is variation between the weakly modeled approach and strongly modeled approach; the weakly modeled distributions are wider than the strongly modeled distributions. In Figure 13.7, we additionally plot posteriors distributions on statistics derived from each $\chi_{\rm eff}$ distribution

(c.f., Table 13.1), with results summarized as follows:

- *First percentile of the $\chi_{\rm eff}$ distribution* ($\chi_{\rm eff,1\%}$): The three strongly modeled distributions find consistent first-percentiles, around ~ 0.2, while the weakly modeled find that the $\chi_{\rm eff}$ distribution extends to lower values. In the case of the BGP, the first-percentile distribution rails against the minimum allowed value of -0.7.

- *Fraction of BBH with negative $\chi_{\rm eff}$*: All models yield consistent posteriors with one another and find that the fraction of BBH with negative $\chi_{\rm eff}$ is greater than zero and less than ~ 0.8. The BGP posterior on this fraction is the widest, and the B-Spline posterior peaks at a higher values than the near-identical strongly modeled posteriors.

- *HM fraction*: The HM fraction is a heuristic for the upper limit of the fraction of BBH coming from hierarchical mergers; it is calculated as 1/0.16 times the fraction of BBHs with $\chi_{\rm eff} < -0.3$, based on the fact that $\sim 16\%$ of hierarchical mergers are theorized to have $\chi_{\rm eff} < -0.3$ [537, 538]. The different models yield a range of 90% upper limits for the HM fraction. With the three parametric models, we consistently find the HM fraction to be small at $\lesssim 8\%$. The two more flexible non-parametric models, on the other hand, find support up to larger values of 27% (B-Spline) or 74% (BGP).

- *Skew of the $\chi_{\rm eff}$ distribution about its peak*: Each model finds fairly different values of the skew, which is here defined as the difference between the fraction of $\chi_{\rm eff} > \chi_0$ and $\chi_{\rm eff} < \chi_0$, where χ_0 is the value of $\chi_{\rm eff}$ at which the distribution peaks [536, Equation 5]. By design, the Gaussian Effective Spins model will always have a skew of ~ 0: truncated normal distributions are approximately symmetric about their peak if the truncation occurs substantially far from the distribution's bulk, which is here the case. All other models find a preference for positive skew, meaning that the distribution has more support for $\chi_{\rm eff}$ above its peak than below. This is the most pronounced for the Skew-normal Effective Spin, which is entirely inconsistent with a skew of 0.

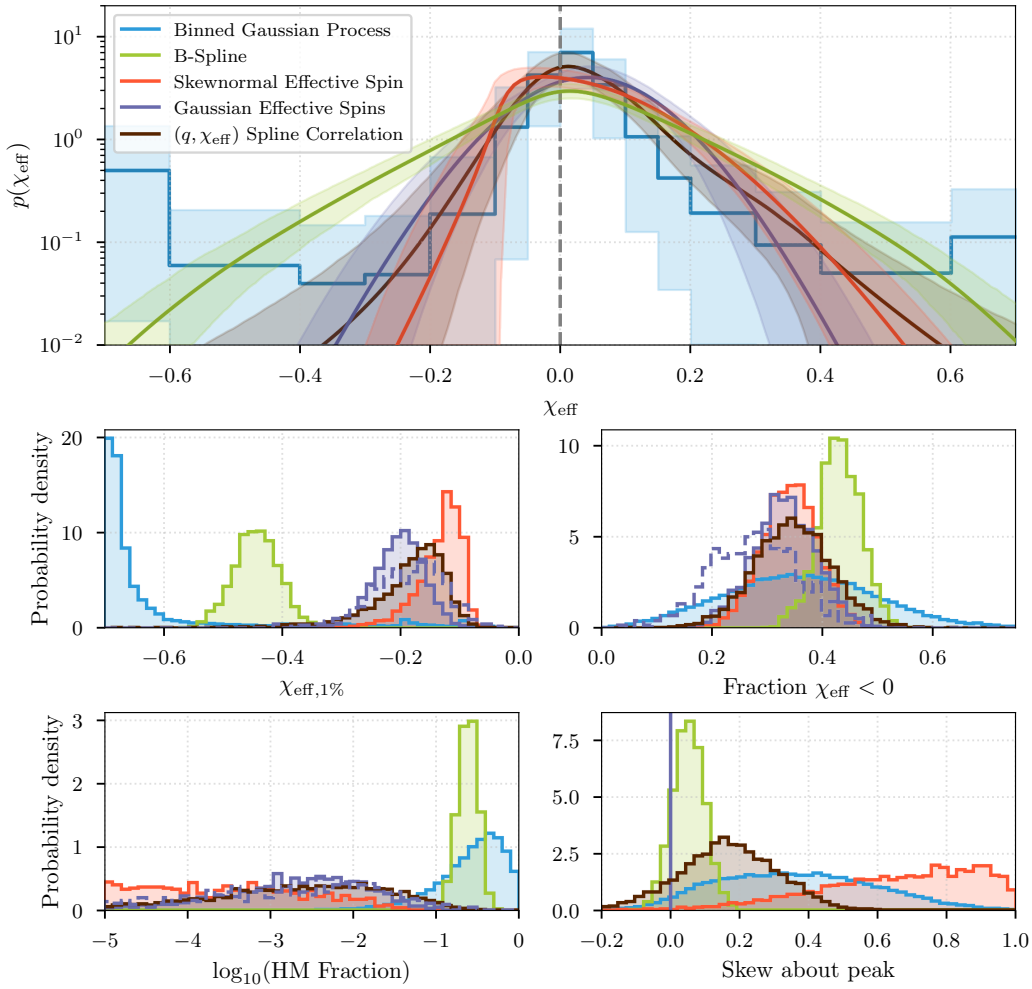


Figure 13.7: *First row:* Marginal $\chi_{\rm eff}$ distributions with BGP (blue) and B-Spline (green) models, as compared to the Skew-normal Effective Spin (red-orange), Gaussian Effective Spins (purple), and $(q, \chi_{\rm eff})$ Spline (maroon) models. The PPD (average) is shown with a dark line, and the 90% confidence intervals are shaded. *Second and third rows:* Posteriors on the first percentile of the $\chi_{\rm eff}$ distribution (upper left), the fraction of $\chi_{\rm eff} < 0$ (upper right), HM fraction (lower left), and the skew (lower right). See Section 13.2 for a discussion of these quantities. Colors correspond to the top panel; purple dashed is GWTC-3.0 with the Gaussian Effective Spins model for comparison.

Chapter 14

LIMITATIONS AND IMPROVEMENTS FOR POSTERIOR PREDICTIVE CHECKS

This chapter is published as Ref. [166]:

Simona J. Miller, Sophia Winney, Katerina Chatziioannou, and Patrick M. Meyers. Posterior Predictive Checks for Gravitational-wave Populations: Limitations and Improvements. 4 2026. ***SJM*** *conceptualized the project, mentored S. Winney (undergraduate) and collaboratively conducted analysis together, wrote the majority of the text, and made Figures 1-3, 5, 6, and 10-18.*

Abstract

When selecting a model to characterize an astrophysical population, it is crucial to assess whether that model fits the data and, if not, how it can be improved. To this end, posterior predictive checks (PPCs) are a widely-used statistical test of model fit when inferring gravitational-wave source populations. However, PPCs exhibit limitations when assessing single-event parameters with large measurement uncertainty, like the spin tilt angles of the binary black holes (BBHs) observable with the LIGO-Virgo-KAGRA (LVK) detectors. When single-event inference is prior-dominated, traditional PPCs fail to flag even very poor model fits. In this work, we assess the efficacy of various alternative PPCs on poorly-constrained parameters. We compare PPCs conducted on event- vs. data-level parameters (e.g. posterior samples vs. maximum likelihood points), and explore two additional event-level PPCs: partial predictive checks and split predictive checks. Independent of measurement uncertainty, we find that PPCs on maximum likelihood parameters are always more discerning of model misspecification than any event-level PPC. However, when investigating simulated GWTC-3.0-like catalogs, none of the alternative PPCs show significant improvement over those traditionally used, indicating that at that sensitivity, any limited information in the data about spin tilts is insufficient to diagnose model misspecification. Finally, we apply our suite of PPCs to the spin magnitude and tilt distributions inferred in the most recent LVK catalog, GWTC-4.0. We conclude that the Gaussian Component

Spins model used therein under-predicts BBHs with large spin magnitudes and over-predicts those with perfectly anti-aligned tilts.

14.1 Introduction

The hundreds of gravitational-wave (GW) signals from compact binary mergers [9, 285–287] observed over the past decade by the LIGO [4], Virgo [5], and KAGRA [6] (LVK) detectors have yielded constraints on the population properties of coalescing binary black holes (BBHs) with masses $3-300\,M_\odot$ [23, 282, 283, 327]. The distributions of mass, spin (i.e., intrinsic angular momentum), and redshift across this BBH population are inferred by, first, selecting some model for the underlying population and, second, constraining that model's parameters using hierarchical Bayesian methods [cf. Sec. 2 of Ref. 23]. Since this process relies on a (sometimes phenomenological) population model, it is important to assess whether that model is indeed a good fit to the observed data, and to understand when conclusions are driven by the data versus the model.

Several model-checking procedures are regularly used in GW population analyses [547]. For instance, one can infer the distribution of mass, spin, and redshift across the same population using different models and compare these models' (dis)agreement [e.g., 23, 267, 283, 318, 548]. Of particular utility is contrasting the behavior of strongly-parametrized models (like Gaussians or power laws) to more flexible, weakly-parametrized models which make minimal assumptions about the shape of the underlying population (like splines, autoregressive processes, or binned-Gaussian processes) [e.g., 266–272]. However, while one can make sure that each is consistent with the maximum likelihood population [549, 550], whether or not two inferred distributions are "similar enough" remains difficult to quantify, especially given the high dimensionality of BBH parameter-space. Another commonly utilized tool for comparing population models is the Bayes factor—a ratio between the marginal likelihoods of two models. However, Bayes factors tend to be highly prior-dependent, rely on obscure criteria for quantifying significance, and provide only an aggregate test of model performance without demonstrating which aspects of the data a non-preferred model fails to capture. As a result, Bayes factors are most helpful in ranking models, not improving them [242].

In this work, we instead focus on posterior predictive checks (PPCs) [349, 551–

553]. Compared to other model-checking procedures, PPCs have the advantage of both providing interpretable, quantitative results about model-fit *and* suggesting sites of model improvement. Broadly, PPCs evaluate the performance of models by checking the consistency between data predicted by the model and current observations. They assess whether statistically significant differences exist between the observed GW catalog and catalogs of simulated GW events drawn from the inferred population distribution. In GW population analyses, PPCs are traditionally conducted on "event-level" parameters: the true underlying astrophysical BH masses, spins, etc. [e.g., 237, 283, 338, 547, 554]. Alternatively, PPCs can be conducted on "population-level" parameters (hyper-parameters), i.e., features in the population distribution [e.g., 327, 339, 365, 535, 548, 555], or "data-level" parameters like maximum likelihood points or search pipeline results [338, 555]. In this work, we *(i)* contrast event and data-level PPCs and *(ii)* implement alternative predictive checks on event-level parameters.

Our study is motivated by previous work in Ref. [160], which demonstrated that the commonly-used event-level PPCs exhibit significant limitations when looking at *poorly-constrained* BBH parameters. When individual event data are weakly informative about a particular parameter, event-level PPCs nearly always indicate that a population model is a "good fit" to the data, even if we know *a priori* it is not. In this regime, the behavior of event-level PPCs becomes dominated by the population model, rather than the limited information contained in the data—the exact problem we wish to diagnose with such tests. Specifically, event-level PPCs incorrectly indicate that a poor model is a good fit because they test a *combination of the data and prior* (i.e., population model) against that prior. In contrast, the data-level PPCs we explore in this work test *just the data* against the prior. We conclude that data-level PPCs provide a more robust assessment of *(i)* to what extent the data are actually informative and *(ii)* if the data are indeed informative, whether they disagree with the choice of population model.

Developing trustworthy model-checking algorithms for poorly constrained parameters is crucial, as some of the most astrophysically interesting BBH properties—spin magnitudes and directions—are indeed only weakly informative. A robust measurement of the BBH spin distribution is central when untangling which BBH formation and evolutionary mechanisms [e.g., 1, 2] contribute to

the observed population. Spins offer unique insight into angular momentum transport in stars [31, 51], supernovae kicks [52, 84–90], and two-body interactions hypothesized to impact BBH formation and evolution, such as tides, accretion, and stellar winds [e.g., 53–60]. Spin-orbit misalignment can also be used to identify dynamically formed BBHs, e.g., in globular clusters [39, 91–93], the disks of active galactic nuclei [e.g., 96, 97, 120, 122, 123], and/or via hierarchical mergers [e.g., 43, 61–71].

Previous studies have found that the BBH population has preferentially small but non-zero spin magnitudes, and a wide range of angles ("tilts") between the spin and orbital angular momenta [23]. While these broad conclusions remain qualitatively unimpeachable, the exact quantitative results are model dependent, due to the aforementioned large measurement uncertainty [160, 318, 535]. Model-dependence is especially prominent when trying to probe narrow population features, such as the exact fraction of non-spinning BBHs, or features informed by the tails of the population, like the fraction of BBHs with spin vectors pointing below the orbital plane [164, 165, 334, 556, 557]. Thus, in this and previous [160] work, we choose to focus on population-level measurements of BBH spin tilts as they provide an astrophysically-motivated testbed for the behavior of PPCs when measurement uncertainty is large.

In this paper, we *explore a wide range of predictive checks on GW population data*. In Sec. 14.2, we summarize our findings and provide a road map for which sections are likely relevant for various audiences. In subsequent sections, we revisit PPCs in general, explain their mathematical formulation in detail, describe posterior predictive p-values, and explore various choices one can make when selecting precisely what predicted vs. observed data are compared. We apply three alternative PPCs to spin tilts, for both simulated populations (where we know the true underlying distribution and intentionally measure it with an insufficiently complex model) under several individual-event likelihood models, as well as the LVK's most recent catalog of GW events, GWTC-4.0 [7, 9].

14.2 Summary of Main Results

We explore the below predictive checks on GW population data, and conclude the following:

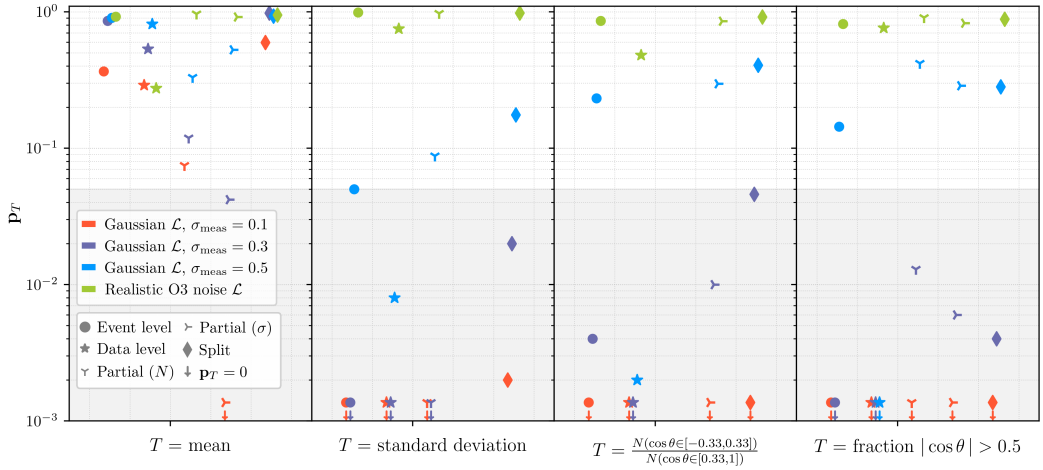


Figure 14.1: Summary of the posterior predictive p-values ($\mathbf{p}_T$) for all PPCs (shapes), test statistics (T; subplots), and single-event likelihoods (different colors) that we explore in this work. Here, we are using an intentionally misspecified model to measure the BBH tilt ($\cos\theta$) distribution: the true distribution is bimodal, while our population model is a single Gaussian (Fig. 14.4). The goal is to successfully identify this model misspecification with low p-values ($\mathbf{p}_T < 0.05$, gray shaded region). Each $\mathbf{p}_T$ is generated from 1,000 PPC traces. Smaller values of $\mathbf{p}_T$ indicate a more robust test of model fit. Points in the bottom row atop downward arrows have $\mathbf{p}_T < 10^{-3} \approx 0$, the smallest value we can probe with 1,000 PPC traces. The unit-less horizontal axis groups the p-values by T (each subplot). From left to right, T is the mean, standard deviation, ratio of the number of events with $\cos\theta \in [-0.33, 0.33]$ to those with $\cos\theta \in [0.33, 1]$, and the fraction of events with $|\cos\theta| > 0.5$. The horizontal placement within each subplot is arbitrary and selected to aid readability. We show results for different PPC (between each gridline per-subplot) and single-event likelihood ($\mathcal{L}$, slightly offset from each other for readability). We investigate both Gaussian single-event likelihood models with increasing measurement uncertainty (red, purple, and blue), as well as realistic O3 noise model (green), as discussed in Sec. 14.3. For the definition of $\mathbf{p}_T$, see Secs. 14.3 and 14.4. For descriptions of the various PPCs, see Secs. 14.4 (event vs. data level), 14.5 (partial), and 14.6 (split). In the legend, "Partial (N)" refers to the partial PPC where $T = N(\cos\theta \in [-0.33, 0.33])/N(\cos\theta \in [0.33, 1])$ is fixed between predicted and observed catalogs, while "Partial (σ)" refers to that where the standard deviation is fixed between. These cases are thus respectively excluded from the third and second columns. Throughout this work, we refer back to the above figure to explain results in greater detail.

1. **Event vs. Data level PPCs**: Event-level PPCs are conducted on true single-event parameters. They involve taking single-event posterior draws, which by definition depend on the choice of prior. We also conduct PPCs on the data itself. Here we use the maximum likelihood parameters, which do not depend on the prior.

   Averaged over many realizations, *data-level PPCs are always equally or more discerning of model misspecification than event-level PPCs*, as shown in Figs. 14.4 and 14.6. This behavior stems from the event-level PPCs' dependence on the prior, and is thus most stark for poorly-constrained parameters like spins.

   To further elucidate event and data level PPCs, we have included a toy-model implementation of both in our data release and Appendix 14.D.

2. **Partial PPCs** target a feature in the population distribution (like the mean, standard deviation, or fraction above some cutoff) and isolate information orthogonal to it.

   The efficacy of the partial PPC depends on which feature we probe. We find that it is *more discerning when the targeted feature is well-predicted by the model*, Fig. 14.8, rather than for features that the model cannot capture, Fig. 14.7.

3. **Split PPCs** divide the observed data into two sub-sets: one subset is used to infer the population distribution and the other is used to generate predictive catalogs. Split PPCs are consistently the least informative PPC that we test, Fig. 14.9.

We conduct both partial and split PPCs on only event-level parameters.

We first investigate a simulated population where we intentionally mismodel the distribution of the tilt angle, $\cos\theta$, and assess each PPC's ability to identify mismatch between model and data. To quantify each PPC's ability to identify this model misspecification, we select features in the distribution to probe (called test statistics T) and from these calculate posterior predictive p-values ($\mathbf{p}_T$), as explained in Sec. 14.3. Figure 14.1 provides a summary of $\mathbf{p}_T$ for each PPC and T across a range of single-event measurement uncertainties and thus provides a summary of our main results. Specifically, we present results where T is:

1. The mean of $\cos\theta$ across a catalog.
2. The standard deviation of $\cos\theta$ across a catalog.
3. The ratio of number of events in a catalog with $\cos\theta \in [-0.33, 0.33]$ to those with $\cos\theta \in [0.33, 1]$.
4. The fraction of events in a catalog with $|\cos\theta| > 0.5$.

If $\mathbf{p}_T < 0.05$ (gray shaded region), that combination of PPC and T is able to identify that the model is indeed a poor fit to the data. Figure 14.1 summarizes our main conclusions: the data level PPCs are always equally or more discerning of model misspecification (i.e., lower $\mathbf{p}_T$) than the traditional event-level PPCs, which are themselves always more discerning than the split PPCs, while the performance of the partial PPC depends on which degree of freedom is targeted. Throughout this work, we will refer back to Fig. 14.1 to explain results in greater detail.

Figure 14.1 also showcases that when calculating posterior predictive p-values, the choice of test statistic is highly important for the test's efficacy. The choice of T is naturally driven by what we want to probe (i.e., is our population model predicting an over-abundance of aligned spins?), but it also depends on the model itself. For instance, a Gaussian model should robustly infer the population's mean regardless of how well it is able to fit the full underlying shape, leading to the mean being a poor choice of test statistic.

Investigations on simulated GWTC-3.0-like catalogs (green in Fig. 14.1) indicate, however, that none of the PPCs are able to robustly identify model misspecification for any T that we probe. Thus, we conclude that at that sensitivity, any limited information in the data about spin tilts is likely insufficient to diagnose model misspecification. However, when applying our suite of PPCs to GWTC-4.0 (Fig. 14.10), we find evidence for model misspecification in the spin magnitude and tilt distributions inferred in Ref. [23] with the Gaussian Component Spins model: the model under-predicts BBHs with large spin magnitudes and over-predicts those with perfectly anti-aligned tilts.

The remainder of the paper is organized as follows. For an overview of the formalism of PPCs and posterior predictive p-values, we direct readers to Secs. 14.3 and 14.3. To learn about conducting PPCs on different levels of parameters (event vs. data level), see Sec. 14.4. To learn about partial and split

predictive checks, two alternative PPCs that change the distributions from which events are drawn, see Secs. 14.5 and 14.6 respectively. In Sec. 14.7, we apply these various checks to GWTC-4.0 data. We present recommendations and future work in Sec. 14.8.

14.3 Methods

To determine various posterior-predictive checking methods' ability to identify discrepancies between population models and observed data, we apply an intentionally misspecified model to a simulated population where the true, underlying parameter distributions are known. In Sec. 14.3, we describe the simulated population used in this work. More details about hierarchical inference and our methods of population simulation and recovery are given in Appendix 14.A. We then present a high-level overview of PPCs (14.3) and associated p-values (14.3), with more technical details given in Appendix 14.B.

Simulated population and its inference

The simulated BBH population distribution in this work is taken from Ref. [160], which explored three simulated populations with the same effective aligned spin ($\chi_{\rm eff}$) distributions but different underlying spin magnitude and tilt distributions. The dimensionless quantity $\chi \in [0, 1]$ captures the magnitude of the spin angular momentum of the BH, where $\chi = 1$ is the maximal spin a BH can support given its mass. The tilt parameter $\theta \in [0, \pi]$ is the angle between the spin vector of the BH and the Newtonian orbital angular momentum vector of the BBH system. We parametrize the tilt through its cosine, $\cos\theta \in [-1, 1]$, where 1 and -1 indicate perfect alignment and anti-alignment with the binary's orbital angular momentum, and 0 indicates fully in-plane spin. GW signals are primarily influenced not by spin magnitudes and tilts [e.g., 143, 306], but instead by two "effective" spin parameters [e.g., 191, 289]: the effective aligned spin ($\chi_{\rm eff}$) containing the spin components that are aligned with the orbital angular momentum [193], and the effective precessing spin ($\chi_{\rm p}$) containing the misaligned components [77, 200, 203–205] which induce spin-orbit precession.

In this work, we focus on the LowSpinAligned population of Ref. [160], shown in their Fig. 1 and replicated in our Fig. 14.11 in Appendix 14.A. This population has a spin magnitude distribution peaking at $\chi = 0.1$, with nearly all $\chi < 0.5$. Its tilt distribution is *bimodal* with peaks at $\cos\theta = 1$ and

$\cos\theta = -1$, or perfect alignment and anti-alignment. These magnitude and tilt distributions produce a $\chi_{\rm eff}$ distribution consistent with that observed in real GW data to date, i.e., through the first part of the LVK's fourth observing run (O4a) [23, 162, 282, 283]: narrow and centered slightly above $\chi_{\rm eff} = 0$. The simulated mass and redshift populations follow the median distributions found analyzing GWTC-3.0, containing the data through LVK's third observing run (O3) [283].

We consider catalogs of 70 BBH events (the approximate number of events in GWTC-3.0) and use the same individual-event likelihood models investigated in Ref. [160], corresponding to different noise scenarios:

1. **Realistic O3 noise likelihood**: Individual-event posteriors were sampled stochastically via nested sampling [253] with the parameter-estimation code Bilby [243], using realistic O3 sensitivity [141] and the waveform model IMRPhenomXPHM [245].

2. **Gaussian likelihood**: Individual-event spin magnitude and cosine-tilt posteriors were constructed using simulated Gaussian likelihoods with standard deviations $\sigma_{\rm meas} = 0.1, 0.3, 0.5$. Samples from the Gaussian spin posteriors were randomly paired with those from the full stochastically sampled Bilby posteriors for all other parameters, i.e., we assumed spins were uncorrelated with mass and redshift.

For reference, the realistic $\cos\theta$ uncertainty at O3 sensitivity averages $\sigma_{\rm meas} \approx 0.5$, which corresponds to posterior distributions often spanning the full allowed parameter range of $\cos\theta \in [-1, 1]$. However, the realistic noise $\cos\theta$ posteriors are non-Gaussian and are correlated with other parameters.

To measure the astrophysical population, we used a *truncated Gaussian distribution* in $\cos\theta$ and inferred its mean and width. This model was deliberately *not* able to capture the full structure of the true, underlying bimodal population. In the remainder of this work, we use the individual-event and hyper-posterior samples obtained in Ref. [160] to test whether various PPC methods can identify the discrepancy between the true and inferred $\cos\theta$ distributions. We refer readers to Ref. [160] (especially Appendices A, B, and C) as well as our Appendix 14.A for more details related to our simulated populations and inference.

Posterior Predictive Checks

In this section, we describe the concept and associated mathematics of PPCs. In full generality, PPCs are a means to assess the accuracy of a model fit to observed data $y^{\rm obs}$ based on its ability to predict future data $y^{\rm pred}$ (also known as *replications* of data) [551]. The mathematical underpinning of PPCs is that observed data should be plausible under an inferred model's *posterior predictive distribution* (PPD). Assuming that the data can be described by a model with parameters Θ, the PPD is

$$\mathrm{PPD}(y^{\rm pred}|y^{\rm obs}) \equiv p(y^{\rm pred}|y^{\rm obs}) = \int p(y^{\rm pred}|\Theta)\, p(\Theta|y^{\rm obs})\, d\Theta\,,$$

where $p(\Theta|y^{\rm obs})$ is what the observed data tell you about the assumed model's parameters and $p(y^{\rm pred}|\Theta)$ is how the model generates data given parameters. Conceptually, the PPD is the distribution of data one would expect to observe in the future ($y^{\rm pred}$) given the data we have already observed ($y^{\rm obs}$) and assuming those data can be described by a model with parameters Θ.

In the case of GW populations, the data y in question are a *catalog* of $N_{\rm obs}$ independent BBH detections: $y = \vec{d} \equiv \{d_i\}_{N_{\rm obs}}$. Throughout this work, we use vectors to describe catalogs, i.e., collections of single-event data. We assume that each BBH in the catalog can be described by individual-event parameters λ (masses, spins, redshift, etc.) via some waveform model for its GW signal. Accordingly, the posterior distribution $p(\lambda|d)$ is the probability that strain data d consist of a superposition of Gaussian noise and a waveform with parameters λ. Going up one level in the hierarchy, the distribution of λ across the astrophysical BBH population $p(\lambda|\Lambda)$ is then described by a set of *hyperparameters* Λ; see Eq. (14.21).

When dealing with a hierarchical model structure, the PPD is a joint distribution on data $d^{\rm pred}$ and parameters $\lambda^{\rm pred}$:

$$\begin{aligned}\mathrm{PPD}(d^{\rm pred}, \lambda^{\rm pred}|\vec{d}^{\rm obs}) &\equiv p(d^{\rm pred}, \lambda^{\rm pred}|\vec{d}^{\rm obs}) \\ &= \int p(d^{\rm pred}, \lambda^{\rm pred}|\Lambda)\, p(\Lambda|\vec{d}^{\rm obs})\, d\Lambda \\ &= \int \mathcal{L}(d^{\rm pred}|\lambda^{\rm pred})\, \pi_{\rm pop}(\lambda^{\rm pred}|\Lambda)\ p(\Lambda|\vec{d}^{\rm obs})\, d\Lambda\,,\end{aligned} \tag{14.1}$$

where $\mathcal{L}(d^{\rm pred}|\lambda^{\rm pred})$ is the single-event likelihood, $\pi_{\rm pop}(\lambda^{\rm pred}|\Lambda)$ is the population model, and $p(\Lambda|\vec{d}^{\rm obs})$ is the hyper-posterior. Most commonly in GW

population analyses, Eq. (14.1) is integrated over $d^{\rm pred}$, yielding a PPD on only parameters $\lambda^{\rm pred}$:

$$\mathrm{PPD}(\lambda^{\rm pred}|\vec{d}^{\rm obs}) \equiv p(\lambda^{\rm pred}|\vec{d}^{\rm obs}) = \int \pi_{\rm pop}(\lambda^{\rm pred}|\Lambda)\; p(\Lambda|\vec{d}^{\rm obs})\, d\Lambda \;. \tag{14.2}$$

We can also express the PPD for a full catalog with data $\vec{d}^{\rm pred}$ and parameters $\vec{\lambda}^{\rm pred}$ (such that $\vec{\lambda} = \{\lambda_i\}_{N_{\rm obs}}$) rather than a single event:

$$\mathrm{PPD}(\vec{d}^{\rm pred}, \vec{\lambda}^{\rm pred}|\vec{d}^{\rm obs}) = \int \mathcal{L}(\vec{d}^{\rm pred}|\vec{\lambda}^{\rm pred})\, \pi_{\rm pop}(\vec{\lambda}^{\rm pred}|\Lambda)\; p(\Lambda|\vec{d}^{\rm obs})\, d\Lambda \,, \tag{14.3}$$

where the likelihood in the integrand is a product over single-event likelihoods

$$\mathcal{L}(\vec{d}^{\rm pred}|\vec{\lambda}^{\rm pred}) = \prod_i^{N_{\rm obs}} \mathcal{L}(d_i^{\rm pred}|\lambda_i^{\rm pred}) \;, \tag{14.4}$$

as is the population distribution,

$$\pi_{\rm pop}(\vec{\lambda}^{\rm pred}|\Lambda) = \prod_i^{N_{\rm obs}} \pi_{\rm pop}(\lambda_i^{\rm pred}|\Lambda) \;. \tag{14.5}$$

Equations (14.3), (14.4), and (14.5) are useful when calculating posterior predictive p-values, as discussed in the following section. We return to discussing PPDs in Sec. 14.4.

In practice, PPCs are conducted numerically by comparing $\vec{d}^{\rm obs}$ and/or $\vec{\lambda}^{\rm obs}$ to a series of simulated $\vec{d}^{\rm pred}$ and/or $\vec{\lambda}^{\rm pred}$ drawn from Eq. (14.3). All PPCs explored in this work can be broken down into the following general algorithmic steps, also visualized in Fig. 14.2:

1. Choose a parameter x on which to conduct the check. In this work we either look at true underlying values of $x \in \lambda$ (denoted $x_{\rm true}$; an example of an *event-level parameter*) or associated maximum likelihood values of x (denoted $x_{\mathrm{max.}\,\mathcal{L}}$; an example of a *data-level parameter*). These quantities are explained in detail in Sec. 14.4.

2. Draw $N_{\rm obs}$ values of x that represent the observed data, which we will call $\vec{x}^{\rm obs}$. This constitutes one "observed catalog" with values of x consistent with $\vec{d}^{\rm obs}$. The exact distribution x is drawn from depends on the specific PPC in question, as explained in Secs. 14.4, 14.5, and 14.6.

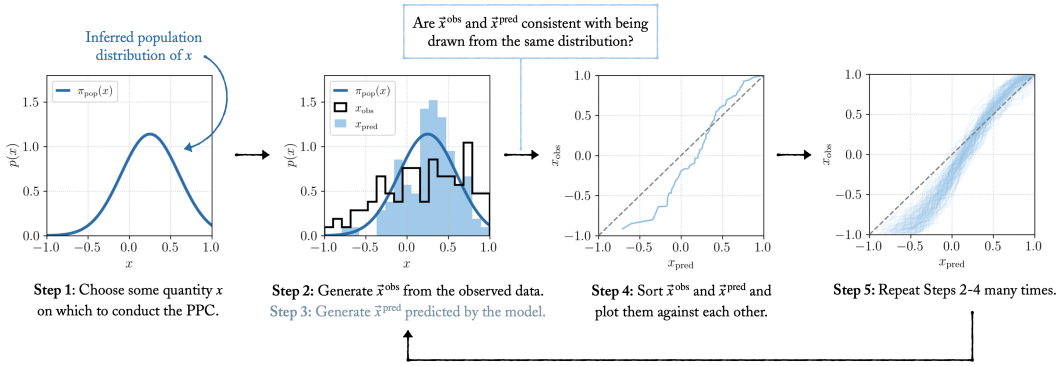


Figure 14.2: Schematic showing the steps for conducting a PPC on an arbitrary parameter x with an inferred distribution $\pi_{\rm pop}(x)$. In this toy example, we pretend that we have perfectly measured $\pi_{\rm pop}(x)$ with some parameterized model (dark blue line in panels 1 and 2). We can tell that the observed data—whose true underlying x distribution we are trying to ascertain—are *inconsistent* with the inferred population distribution because catalogs with x-values generated from each (e.g., black vs. blue histograms in panel 2) produce *non-diagonal PPC traces* (panels 3 and 4). Thus, $\pi_{\rm pop}$ is insufficient to capture the shape of the true underlying distribution of x. The different types of PPCs explored in this work correspond to exactly what happens between the first two panels i.e., how predicted and observed catalogs are generated.

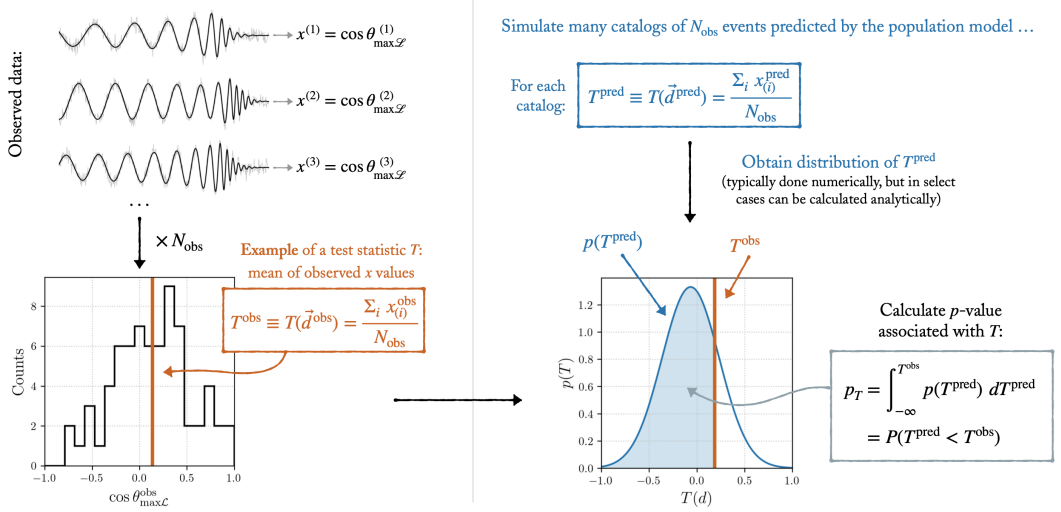


Figure 14.3: Schematic demonstrating how to calculate p_T (Eq. 14.7). In this example, we look at the data-level parameter $x = \cos\theta_{\rm max.\,\mathcal{L}}$ and use the mean of x as our test statistic T. Refer to Fig. 14.2 for how observed and predicted data (catalogs) are generated.

3. Draw $N_{\rm obs}$ values of x predicted by the model, which we will call $\vec{x}^{\,\rm pred}$. This constitutes a "predicted catalog" with values of x consistent with one possible instantiation of $\vec{d}^{\,\rm pred}$. These are drawn from a distribution analogous to that in Step 2.

4. Sort the elements of $\vec{x}^{\,\rm pred}$ and $\vec{x}^{\,\rm obs}$ each from smallest to largest, and plot them against each other. This generates one "trace" of a PPC.

5. Repeat steps 2 through 4 many times to generate a collection of PPC traces.

6. Choose some test statistic T to assess the discrepancy between predicted and observed data. Calculate an associated p-value. This is discussed in detail in Sec. 14.3 and visualized in Fig. 14.3.

In the limit of a perfectly measured population distribution with $N_{\rm obs} = \infty$, each of the $\vec{x}^{\,\rm pred}$ vs. $\vec{x}^{\,\rm obs}$ traces will follow an exact diagonal line. However, in the realistic scenario where we have finite $N_{\rm obs}$, the *average* of the traces should be consistent with the diagonal if the model is a good fit to the data [e.g., 162, 338, 348–350]. Thus, any systematic differences between $\vec{d}^{\,\rm pred}$ and $\vec{d}^{\,\rm obs}$—indicating the population model's inability to correctly infer $\vec{d}^{\,\rm obs}$—will manifest as non-diagonal $\vec{x}^{\,\rm pred}$ vs. $\vec{x}^{\,\rm obs}$ traces. For the above to be a valid technique, $\vec{d}^{\,\rm pred}$ must be generated under the exact same set of assumptions as $\vec{d}^{\,\rm obs}$, such as detection threshold, single-event likelihood model, etc.

Posterior Predictive p-values

A classical p-value is the probability that the value of a test statistic T evaluated on replications of data is *at least as extreme* as the one observed, assuming some model for the distribution of T. A p-value closer to 0 indicates that the data are less likely to have occurred under said model. This idea can be extended to Bayesian statistics, where we move from a single test statistic that summarizes the data, to a distribution of test statistics that summarizes both the data and our uncertainty on parameters of the model. In this case, the test statistic depends not just on the data but also the *parameters* describing the data. The PPD given in Eq. (14.1) yields a *posterior predictive p-value* [551],

$$p_T(y^{\rm obs}) = \iint I_{[T(y^{\rm pred},\Theta)\geq T(y^{\rm obs},\Theta)]}\; p(y^{\rm pred}|\Theta)\; p(\Theta|y^{\rm obs})\; dy^{\rm pred} d\Theta\,, \qquad (14.6)$$

where I is an indicator function, meaning $I_X = 1$ if the conditional X is true and 0 if X is false. The value of p_T thus quantifies the probability that

replicated data $y^{\rm pred}$ drawn from a model $p(y^{\rm pred}|\Theta)$, where Θ is informed by observed data $y^{\rm obs}$ via $p(\Theta|y^{\rm obs})$, could be *more extreme* than the observed data $\vec{d}^{\rm obs}$ according to some test statistic T.[1] The posterior predictive p-value has previously been used in searching for GWs with Pulsar Timing Arrays, where the GW detection summary statistic depends upon a posterior distribution for individual pulsar noise models [558–561].

Calculating p_T requires the choice of the test statistic T, a function from data and/or parameter space to the real numbers, which is compared between the observed catalog of GW events and replications of catalogs predicted by the model [551]. Figure 14.3 provides a schematic for the process for calculating p_T for an example T. Examples of T include the mean, standard deviation, minimum, or maximum of some quantity. As shown in Fig. 14.1, throughout this work, we use four choices of T: *(i)* the mean of $\cos\theta$ draws, *(ii)* the standard deviation of $\cos\theta$ draws, *(iii)* the ratio of the draws where $\cos\theta \in [-0.33, 0.33]$ to those where $\cos\theta \in [0.33, 1]$, and *(iv)* the fraction of draws with $|\cos\theta| > 0.5$.

Choosing T is context dependent, and a smart choice is necessarily tailored to whichever aspect of the model one is trying to probe. For instance, if one wanted to test whether a model correctly predicts aligned tilts, a strong T would be, e.g., the fraction of events with $\cos\theta > 0.5$ or the ratio of $\cos\theta \sim 1$ to $\cos\theta \sim 0$. Additionally, the efficacy of a given T depends on the population model itself. For example, if one used a Gaussian distribution as the population model, it will by definition correctly infer the population's mean (assuming the true mean is within the prior), making this likely a poor choice of T. In Sec. 14.4 and Appendix 14.D, we show how some choices of T can be much more informative than others.

In the case of hierarchical inference, the posterior predictive p-value is calculated from a test statistic over the joint posterior predictive distribution of data *and* parameters, thus requiring a more complicated integral than Eq. (14.6). Here, we are looking at the probability that $T(\vec{d}^{\rm pred}, \vec{\lambda}^{\rm pred}) \geq T(\vec{d}^{\rm obs}, \vec{\lambda}^{\rm obs})$ where $\vec{\lambda}^{\rm pred}$ are the parameters describing $\vec{d}^{\rm pred}$ and $\vec{\lambda}^{\rm obs}$ are the parameters describing $\vec{d}^{\rm obs}$. To generate a single value of p_T to assess model fit, we perform

[1] In the definition in Eq. (14.6), "more extreme" means too large: $T(y^{\rm pred}, \Theta) > T(y^{\rm obs}, \Theta)$. We generalize the p-value definition in Eq. (14.8) to expand "extreme" to mean too large *or* too small.

the following integral:[2]

$$p_T(\vec{d}^{\,\text{obs}}) = \iiiint I_{[T(\vec{d}^{\,\text{pred}},\vec{\lambda}^{\text{pred}})\geq T(\vec{d}^{\,\text{obs}},\vec{\lambda}^{\text{obs}})]} \times p(\vec{d}^{\,\text{pred}},\vec{\lambda}^{\text{pred}}\,|\Lambda)\; p(\vec{\lambda}^{\text{obs}},\Lambda|\vec{d}^{\,\text{obs}})\, d\Lambda\, d\vec{\lambda}^{\text{pred}}\, d\vec{\lambda}^{\text{obs}}\, d\vec{d}^{\,\text{pred}}\,. \quad (14.7)$$

The two probability distributions in the integrand of Eq. (14.7) can be understood as follows. The observed data give us constraints on our models' single-event parameters and hyper-parameters: $p(\vec{\lambda}^{\text{obs}},\Lambda|\vec{d}^{\,\text{obs}})$. The population model then generates single-event parameters, which in turn are used to produce predicted data: $p(\vec{d}^{\,\text{pred}},\vec{\lambda}^{\text{pred}}\,|\Lambda)$. Equations (14.12) and (14.13) in Sec. 14.4 present specific cases where Eq. (14.7) can be simplified: T that depend solely on $\vec{d}$ (data-level) versus solely on $\vec{\lambda}$ (event-level).

Using the above definitions in Eqs. (14.6) and (14.7), p_T is defined on the range $[0,1]$, with values closer to 0.5 indicating a good fit. For ease of comparison with more widely-used frequentist (rather than Bayesian) p-values, we decide to scale p_T to obtain a new quantity,

$$\mathbf{p}_T = 1 - 2\times|\,p_T - 0.5\,|\,, \quad (14.8)$$

such that $\mathbf{p}_T$ closer to 1 indicates a better fit, and $\mathbf{p}_T$ closer to 0 is obtained when T^{obs} is either greater *or* less than T^{pred}.

Unlike frequentist p-values, posterior predictive p-values are *not uniform under the null hypothesis* (model matches data), complicating their interpretation. This is a known feature of Bayesian model checking [349, 551, 562–565]. For traditional PPCs, posterior predictive p-values tend to concentrate around $p_T \sim 0.5$ (equivalently, $\mathbf{p}_T \sim 1$) under the null hypothesis because the same data are used to both estimate parameters and evaluate the test statistic [349]. In this way, the posterior predictive p-value tends to be conservative, with Ref. [563] showing that $P(p_T \leq \alpha) \leq 2\alpha$ when the model is a good fit to the data. Our Eq. (14.8) implies then that $P(\mathbf{p}_T \leq \alpha) \leq \alpha$ under the null hypothesis. Therefore, in this casting, we use the canonical threshold of $\mathbf{p}_T \lesssim 0.05$ as a *conservative* indication of discrepancy between model and data. In the majority of this work, we intentionally use a model that poorly fits the data and thus desire small values of $\mathbf{p}_T$: the smaller the $\mathbf{p}_T$, the more robustly

[2] One could also calculate a *distribution* of p-values over any of Λ, λ, or d^{pred}, instead of marginalizing over all of these parameters as is done in Eq. (14.7).

that test probes model misspecification. In the remainder of this work, when we mention a posterior-predictive p-value, we are referring to Eq. (14.8) unless otherwise specified.

14.4 Event vs. Data Level PPCs

In hierarchical inference, tests for goodness-of-fit can be conducted on different "levels" of parameters in the hierarchy: population-level, event-level, or data-level parameters [264, 338]. In this work, we avoid the population-level case. These are carried out, for example, by performing leave-one-out analyses with events in the observed catalog and comparing the resultant Λ posteriors. Alternatively, they can assess the robustness of a particular inferred *feature* in population (e.g., an overdensity at some value of mass or spin): by simulating populations *without* the specific feature of interest, one can quantify how often we spuriously infer the presence of that feature due to Poisson noise [339], a form of the p-value discussed in the preceding section. Here, we instead focus on comparing event vs. data-level PPCs.

In GW data analysis, PPCs are most commonly conducted on **event-level** parameters [with exceptions, e.g., 338, 555]. Specifically, we traditionally look at "true" parameters: the mass, spin, etc. that describe the actual underlying astrophysical system. The probability of drawing parameters $\lambda^{\rm pred}_{\rm true}$ given our observed data $\vec{d}^{\rm obs}$ and population model $\pi_{\rm pop}$ is given in Eq. (14.2), where we average the population distribution $\pi_{\rm pop}(\lambda^{\rm pred}_{\rm true}|\Lambda)$ over the hyper-posterior $p(\Lambda|\vec{d}^{\rm obs})$.

In reality, the observed data have a further restriction: detectability. To faithfully compare predicted and observed catalogs, we therefore must incorporate a detection threshold on the predicted data. In other words, we care about a model's predictive power to generate true parameters that lead to a *detectable* GW signal, and so must incorporate selection effects into Eq. (14.2):

$$p(\lambda^{\rm pred}_{\rm true}\,|\,\vec{d}^{\rm obs}, {\rm det}) = {\rm PPD}(\lambda^{\rm pred}_{\rm true}|\vec{d}^{\rm obs}) P_{\rm det}(\lambda^{\rm pred}_{\rm true})$$

$$= {\rm PPD}(\lambda^{\rm pred}_{\rm true}|\vec{d}^{\rm obs}) \int I_{[D(d)>D_{\rm thr}]}\, \mathcal{L}(d|\lambda^{\rm pred}_{\rm true})\, dd\ . \quad (14.9)$$

where $P_{\rm det}$ is the detection probability, equivalent to the indicator function I, equaling 0 or 1 depending on whether our *detection statistic* D evaluated on the data passes some known threshold $D_{\rm thr}$, and the PPD is given in Eq. (14.2). In

the integral in Eq. (14.9), we average $\mathcal{L}(d|\lambda^{\rm pred}_{\rm true})$ over all possible replications of data d containing a signal produced by a BBH with true parameters $\lambda^{\rm pred}_{\rm true}$ (i.e., all possible noise instantiations) where that signal is detectable (expressed via $I_{[D(d)>D_{\rm thr}]}$). Equation (14.9) is used as the basis for the event-level PPCs.[3]

PPCs conducted on the **data-level** are instead concerned with quantities that characterize the actual data recorded in the detectors, such as the source parameters' maximum likelihood (max. $\mathcal{L}$) values

$$\lambda_{\max.\,\mathcal{L}} \equiv \{\lambda \mid \mathcal{L}(d|\lambda) = \max[\mathcal{L}(d|\lambda)]\}\,. \tag{14.10}$$

In the cases we consider, event-level parameters are determined *probabilistically* from the data: any posterior draw has, by definition, an equal probability of being the true underlying parameters of the BBH. In contrast, data-level parameters stem *deterministically* from the data: each posterior distribution only has one max. $\mathcal{L}$ point.

We can write a probability distribution on the *data* ($d^{\rm pred}$) of a future detection, rather than the true BBH source parameters that (combined with random noise) generated said data [338]:

$$p(d^{\rm pred} \mid \vec{d}^{\rm obs}, \text{det}) = \text{PPD}(d^{\rm pred}|\vec{d}^{\rm obs}) P_{\rm det}(d^{\rm pred})$$

$$= \int I_{[D(d^{\rm pred})>D_{\rm thr}]}\; \mathcal{L}(d^{\rm pred}|\lambda)\; \text{PPD}(\lambda|\vec{d}^{\rm obs})\, d\lambda\,, \tag{14.11}$$

where the PPD in the integrand is again calculated via the integral in Eq. (14.2). Throughout this work, we use the same symbol $P_{\rm det}$ and change the argument between d and λ depending on which we are referring to, such that $P_{\rm det}(\lambda) = \int P_{\rm det}(d)\mathcal{L}(d|\lambda)dd$. Equations (14.9) and (14.11) are similar, except the former integrates over noise instantiations (dd) while the latter integrates over masses, spins, etc. ($d\lambda$). Equation (14.11) is used as the basis for the data-level PPCs: to generate $\lambda^{\rm pred}_{\max.\,\mathcal{L}}$, we draw $d^{\rm pred}$ from Eq. (14.11) and then calculate $\lambda^{\rm pred}_{\max.\,\mathcal{L}}$ from $d^{\rm pred}$ under the same likelihood model used for original parameter estimation on $d^{\rm obs}$.

In the following subsections, we describe *algorithmically* how to conduct PPCs using Eqs. (14.9) and (14.11), which we hope are more illuminating than simply

[3] Technically, Eqs. (14.9) and (14.11) should include a factor of $1/p(\text{det}|d^{\rm obs})$. However, $p(\text{det}|d^{\rm obs}) = 1$ because we have, by definition, detected our observed events.

looking at the integrals. We present results for PPC traces (14.4) and p-values (14.4) for event-level and data-level parameters, and discuss pros and cons of each. Additionally, in Appendix 14.D we repeat these exercises with a simple, one-dimensional, analytic toy model; we direct readers there to gain intuition.

Event vs. data level PPC traces

Here, we add specificity to the general PPC algorithm presented in Sec. 14.3 for both our event- and data-level cases. Algorithmically, one trace of a traditional event-level PPC is generated with the following steps [350]:

1. Draw one sample of Λ from the hyperposterior $p(\Lambda|\vec{d}^{\rm obs})$.
2. *Observed values* ($\vec{\lambda}^{\rm obs}_{\rm true}$): Draw one sample from each of the $N_{\rm obs}$ individual-event posteriors in the observed catalog, *reweighted* [350] from its parameter-estimation prior to $\pi_{\rm pop}(\lambda_{\rm true}|\Lambda)$.
3. *Predicted values* ($\vec{\lambda}^{\rm pred}_{\rm true}$): Draw $N_{\rm obs}$ *detectable*[4] values of $\lambda_{\rm true}$ from the distribution $\pi_{\rm pop}(\lambda_{\rm true}|\Lambda)$, following Eq. (14.9).

In this work, we repeat the above steps, stochastically sampling over Λ to obtain 1,000 realizations of predicted and observed catalogs.

For the data-level PPC, there is only one set of observed values, consisting solely of the maximum likelihood parameters from each observed individual-event posterior ($\lambda^{\rm obs}_{\rm max.\,\mathcal{L}}$). To create each set of predicted values, we:

1. Draw one sample of Λ from the hyperposterior $p(\Lambda|\vec{d}^{\rm obs})$.
2. Draw $N_{\rm obs}$ *detectable* values of $\lambda_{\rm true}$ from the distribution $\pi_{\rm pop}(\lambda_{\rm true}|\Lambda)$ (Eq. (14.9)).
3. Generate a GW signal for each $\lambda_{\rm true}$ and inject it into random Gaussian noise drawn from our power spectral density.[5]

[4] Specifically, we draw $N_{\rm obs}$ values from the found injections used to determine detection efficiency (Eq. (14.23)) reweighted from the injected distribution to $\pi_{\rm pop}$.

[5] We self-consistently used a threshold on optimal signal-to-noise ratio (SNR) to quantify detectability when analyzing the simulated populations in Ref. [160] and thus do the same in this work, as discussed in Appendix 14.A. The use of optimal SNR means that detectability can be assessed in Step 3, rather than Step 4 (as would be necessitated for matched-filter SNR).

4. For each simulated signal, find the maximum likelihood value of the parameters ($\lambda^{\rm pred}_{\rm max.\,\mathcal{L}}$) using the same method as was used for $\lambda^{\rm obs}_{\rm max.\,\mathcal{L}}$.

As with the event-level PPCs, we stochastically sample over the Λ hyperposterior and generate 1,000 traces.

For both the predicted and observed catalogs in the Gaussian single-event likelihood case, we assume that the max. $\mathcal{L}$ spin magnitudes and tilts are drawn from a truncated Gaussian distribution with width $\sigma_{\rm meas}$ centered at $\lambda_{\rm true}$, i.e., for each event $\cos\theta_{\rm max.\,\mathcal{L}} \sim \mathcal{N}_{[-1,1]}(\cos\theta_{\rm true}, \sigma_{\rm meas})$, with a corresponding single-event posterior $p(\cos\theta_{\rm true}|d) = \mathcal{N}_{[-1,1]}(\cos\theta_{\rm max.\,\mathcal{L}}, \sigma_{\rm meas})$.

For the realistic noise case, we find the max. $\mathcal{L}$ parameters through a least-squares minimization algorithm implemented in the parameter-estimation code `cogwheel` [256, 298, 566]. Likelihood *optimization*, rather than taking the max. $\mathcal{L}$ *sample* from a full posterior is necessary: performing full parameter estimation to generate posterior samples is computationally infeasible when dealing with many thousands of simulated events. Moreover, we find that the max. $\mathcal{L}$ posterior sample is not a particularly robust data-level parameter for spin tilts, as parameter estimation is not designed to optimize the likelihood but rather to well-approximate the full posterior distribution. For leading-orders parameters like the chirp mass, the max. $\mathcal{L}$ sample and true max. $\mathcal{L}$ value are close. However, for weakly-informative parameters like $\cos\theta$, the two can differ significantly. We discuss our approach in more detail and provide a comparison with different methods in Appendix 14.C.

Figure 14.4 shows the results of the event-level (left column within each color) versus data-level (right column within each color) PPC traces for $\cos\theta$ of our simulated population with $N_{\rm obs} = 70$ events. The event-level PPC results are a replication of those shown in Figs. 3 and 4 of Ref. [160] (although using a different method to calculate error bars in each bottom row). Four different sets of individual-event posteriors are tested: realistic posteriors for O3 sensitivity sampled with Bilby (green), and synthetic Gaussian posteriors with measurement uncertainties $\sigma_{\rm meas} = 0.1$ (red), 0.3 (purple), and 0.5 (blue). For comparison, the $\cos\theta$ posteriors from actual GW detections from O3 have average measurement uncertainties $\sigma_{\rm meas} \gtrsim 0.5$ and are non-Gaussian [287]. We show results for a case where the population *is* well-described by the model in Appendix 14.D.

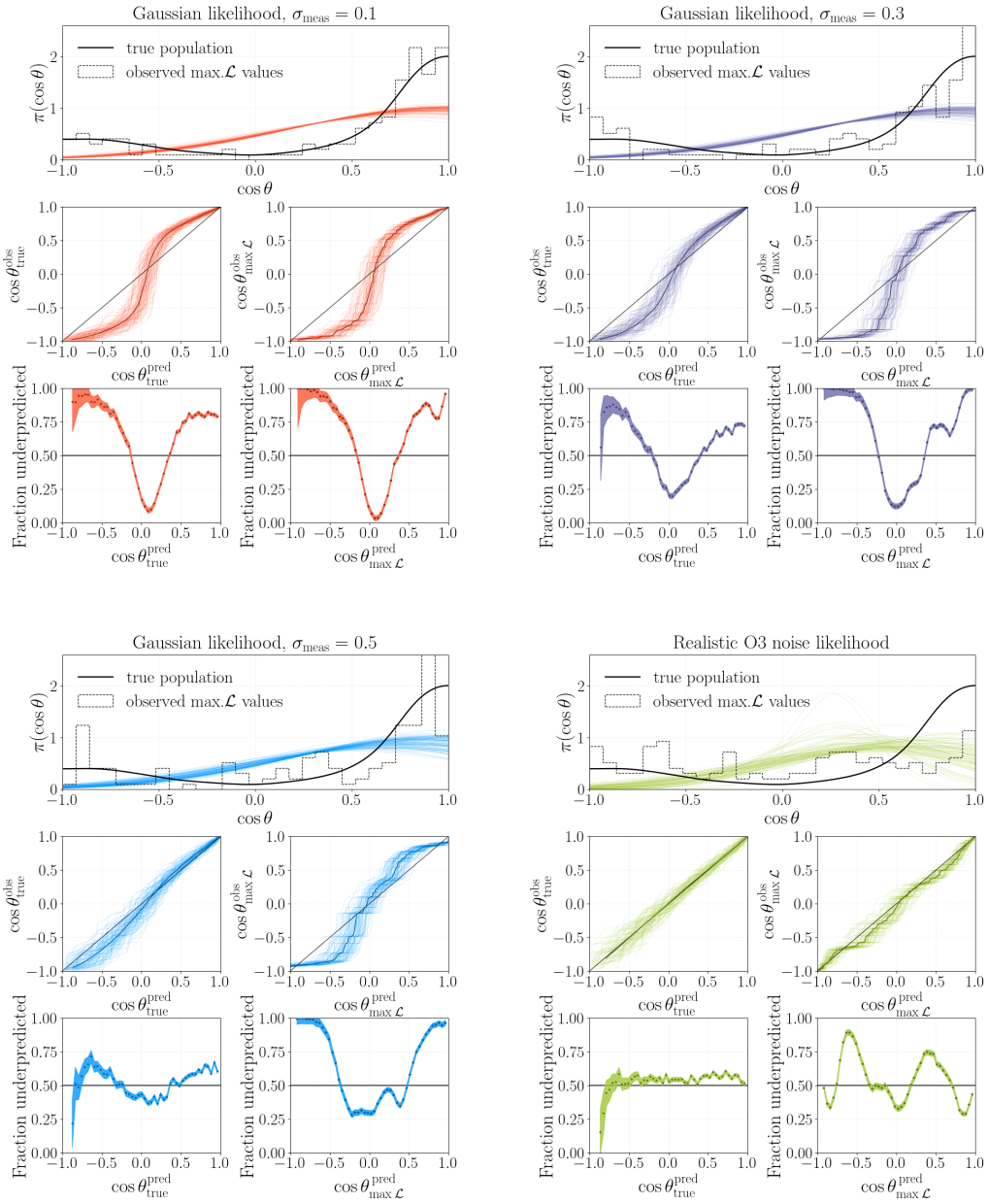


Figure 14.4: Results of population inference and event-level versus data-level PPCs for $\cos\theta$ using a Gaussian population model on simulated catalogs containing 70 events which are from a bimodal population. Each quadrant/color shows a different individual-event likelihood: Gaussian likelihoods with measurement uncertainty $\sigma_{\rm meas} = 0.1$ (red), 0.3 (purple), and 0.5 (blue) versus the realistic O3 noise likelihood (green). The following descriptions correspond to the subplots within each quadrant. *(First row)*: The true (black, solid) population distribution for $\cos\theta$ compared to traces obtained from hyper-posterior draws (colors). Maximum likelihood values of $\cos\theta$ from the observed histogram are shown in the black-dashed histograms. *(ctd. on following page)*

Figure 14.4 *(previous page)*: *(Second row)*: Traditional event-level (left) versus data-level (right) PPC traces from 100 catalogs. The bold line is the average of the traces across the catalog instantiations. If the population model is a good fit to the data, these traces should follow the diagonal (black) on average. *(Third row)*: The fraction of catalogs within a given bin of $\cos\theta$ for which the traces in the middle row have slope < 1, indicating the model underpredicts the true population. Points mark the average over ten 100-trace PPCs; the shaded region indicates $\pm N^{-1/2}$ error where N is the total number of draws in that bin across all 1,000 traces.

As demonstrated in the top row of Fig. 14.4, a Gaussian population model yields a poor fit (various colors of traces) to the true underlying bimodal $\cos\theta$ distribution (single black trace) of our simulated population. This stark discrepancy makes this particular simulated population a useful probe into the efficacy of different PPCs. We again emphasize that we are deliberately using a population model that cannot fit all the features of the true underlying $\cos\theta$ distribution. Our goal is to identify this mismatch without prior knowledge about the true population: when analyzing real data, all we have access to are inferred populations, not the truth.

With low individual-event measurement uncertainties of $\sigma_{\rm meas} = 0.1, 0.3$ (red, purple), the discrepancy between the observed and predicted draws is visually clear in both the event-level and data-level PPCs: their traces are highly non-diagonal. However, when $\sigma_{\rm meas}$ increases to 0.5 (blue) or we use realistic O3 noise (green), individual event likelihoods become poorly-constrained; their posteriors become more prior-dominated, and the event-level PPC traces begin to approach the diagonal. In these cases, the event-level PPC indicates that the model is a good fit to the data even though *a priori* we know it is not.

Thus, Fig. 14.4 shows that **if individual-event posterior distributions are prior-dominated, event-level PPCs have difficulty diagnosing an inaccurate model.** This concerning trend stems from the fact that when we re-weight our observed individual-event posteriors to the population (i.e., a new prior) in Step 2 of the event-level PPC algorithm, each becomes near-identical to the population distribution, causing nearly identical observed and predicted event-level PPC traces. When individual-event measurement uncertainty is high, the information contained in the population distribution dominates, **regardless of whether that information comes from our**

(perhaps arbitrary) assumptions about the shape of the distribution or the actual limited information in the observed events themselves.

Data-level PPCs avoid reweighting individual-event posteriors to the population, making them immune to the prior-related shortcomings of event-level PPCs when dealing with large individual-event measurement uncertainty. For reference, we show the observed max. $\mathcal{L}$ values of $\cos\theta$ in black-dashed histograms within the top row of each quadrant of Fig. 14.4. For the $\sigma_{\rm meas} = 0.5$ single-event likelihood especially, the data-level PPC shows significantly more non-diagonality, properly identifying the poor fit between the model and the data. Deviations from the diagonal are also more apparent in the data-level than the event-level PPC for the realistic noise case (Fig. 14.4, green), but the deviations follow a pattern that differs from the similar shapes of the Gaussian posteriors. Such behavior is not unexpected: Gaussian intuition only goes so far, and realistic GW likelihoods have a much more rich structure. **Data-level PPCs are a better tool than traditionally-used, event-level PPCs for assessing population model misspecification in the case of poorly-constrained parameters, like spin tilts**. In Appendix 14.D, we find analogous behavior with a toy model.

In addition to just telling us that a model is a poor fit to the data, PPCs tell us *where* that model failed so that we can improve it. To visualize this, we calculate the fraction of events over/under-predicted by our Gaussian population model as a function of $\cos\theta$. This is done by looking at the *slopes* of the traces. If, at a given value of $\cos\theta$, the slope of a PPC trace is steeper than the diagonal, then the model is *over-predicting* events with that $\cos\theta$; if instead the slope is shallower than the diagonal, it is *under-predicting*. If the model is a good fit to the data, the fraction of events over/under predicted should be consistent—within error bars—with 0.5 for all values of $\cos\theta$. If not, the fraction of events over/under predicted can tell us where new features should be added to the population model to provide a better fit—going beyond the utility of Bayes factors.

The third row within each quadrant of Fig. 14.4 shows the fraction of traces that under-predict the number of events in a binned range of $\cos\theta$, corresponding to the fraction of traces in that bin with a slope < 1. The slope of each trace is determined via linear regression over the $\cos\theta$ bin. We repeat this experiment with ten 100-trace PPCs; the dots are the average fraction

under-predicted over these ten PPCs and the shaded region encloses $\pm N^{-1/2}$ error, where N is the total number of draws in the $\cos\theta$ bin across all 1,000 traces. Near $\cos\theta \sim -1$, there are less data in both the predicted and observed catalogs, leading to consistently larger uncertainty in those regions regardless of single-event likelihood model.

In the $\sigma_{\rm meas} = 0.1, 0.3$ cases, both event- and data-level PPCs accurately identify the inability of the Gaussian model to capture the bimodal nature of the true population distribution, with a fraction under-predicted clearly inconsistent with 0.5 across $\cos\theta$. For the $\sigma_{\rm meas} = 0.5$ case, the event-level PPC shows little scatter around a fraction under-predicted of 0.5, while the data-level PPC shows large deviation. The *shape* of the fraction under-predicted curve mirrors the shape of the true population relative to the population model: for $\cos\theta \lesssim -0.5$ and $\gtrsim 0.5$, the model under-predicts the truth while for $-0.5 \lesssim \cos\theta \lesssim 0.5$, the model over-predicts. If *a priori* we did not know the shape of the true $\cos\theta$ distribution, seeing a PPC result like this would tell us that there is unresolved bimodality.

The fraction under-predicted curve for the realistic-noise case (green) has a similar shape to the Gaussian cases between $\cos\theta \in [-0.5, 0.5]$, but differs as $\cos\theta \to \pm 1$. We consider several plausible explanations. First, data-level PPCs are sensitive to randomness from small-number statistics in the observed catalog. We compare each predicted trace to the *same* observed trace, so a random "bump" in the observed max. $\mathcal{L}$ values from simple Poisson uncertainty can have an outsized effect. From looking at just the fraction-under predicted plots, we suspect this is happening, e.g., at $\cos\theta \sim 0.7$ in the $\sigma_{\rm meas} = 0.3$ case and $\cos\theta \sim 0.3$ in the $\sigma_{\rm meas} = 0.5$ case, where there are deviations from the U-shaped fraction under-predicted curve. The observed max. $\mathcal{L}$ histograms in Fig. 14.4 confirm this suspicion. The behavior of the realistic noise case could be explained by similar (albeit larger) deviations as $\cos\theta \to \pm 1$. Second, it could instead be the case that the likelihood contains so little information about $\cos\theta$ that *all* we are seeing is the impact of small-number statistics. This explanation is supported by the fact that very different true/injected $\cos\theta$ distributions all produce corresponding max. $\mathcal{L}$ $\cos\theta$ distributions that look near-identical, as seen in Fig. 14.12 in Appendix 14.C. Third, the impact of random Gaussian noise instantiations on the full 15-dimensional BBH likelihood might differ from the intuition we gained with the Gaussian likelihoods—especially

near the boundaries of $\cos\theta = \pm 1$.

The observed max.$\mathcal{L}$ histogram for the realistic noise makes it difficult to distinguish between which of the preceding three scenarios is most impacting our results. Probably, it is some combination of all three: the $\cos\theta_{\text{max.}\mathcal{L}}$ distribution for the realistic noise case is flatter than those for the Gaussian likelihoods, therefore retaining less information about the true underlying population. Thus, it is indeed true that the realistic noise case exhibits different behavior from the Gaussian case as $\cos\theta_{\text{max.}\mathcal{L}} \to \pm 1$. Additionally, the $\cos\theta^{\text{obs}}_{\text{max.}\mathcal{L}}$ distribution shows seemingly random over-densities at $\cos\theta \sim -0.7$ and 0.4, corresponding to the turning points in the fraction under-predicted plot.

Event vs. data level p-values

PPC traces like those shown in the middle row of each quadrant of Fig. 14.4 allow for a qualitative assessment of a model's goodness of fit to the true, underlying population. While the fraction under-predicted across parameter space (bottom row within each quadrant of Fig. 14.4) is one useful metric to move towards a quantitative result, we can also look at posterior predictive p-values ($\mathbf{p}_T$), as described in Sec. 14.3. Equations to calculate p_T for a test statistic T using event and data-level parameters are given in Eqs. (14.12) and (14.13), respectively:

$$\text{Event level}: \quad p_T(\vec{d}^{\,\text{obs}}) = \iiint I_{[T(\vec{\lambda}^{\,\text{pred}}) \geq T(\vec{\lambda}^{\,\text{obs}})]}\; P_{\text{det}}(\vec{\lambda}^{\,\text{pred}}, \vec{\lambda}^{\,\text{obs}})\; \pi_{\text{pop}}(\vec{\lambda}^{\,\text{pred}}|\Lambda) \times\; p(\vec{\lambda}^{\,\text{obs}}|\vec{d}^{\,\text{obs}}, \Lambda)\; p(\Lambda|\vec{d}^{\,\text{obs}})\; d\Lambda\; d\vec{\lambda}^{\,\text{obs}}\; d\vec{\lambda}^{\,\text{pred}}\,, \tag{14.12}$$

$$\text{Data level}: \quad p_T(\vec{d}^{\,\text{obs}}) = \iiint I_{[T(\vec{d}^{\,\text{pred}}) \geq T(\vec{d}^{\,\text{obs}})]}\; P_{\text{det}}(\vec{d}^{\,\text{pred}})\; \mathcal{L}(\vec{d}^{\,\text{pred}}|\vec{\lambda}^{\,\text{pred}}) \times\; \pi_{\text{pop}}(\vec{\lambda}^{\,\text{pred}}|\Lambda)\; p(\Lambda|\vec{d}^{\,\text{obs}})\; d\Lambda\; d\vec{\lambda}^{\,\text{pred}}\; d\vec{d}^{\,\text{pred}}\,. \tag{14.13}$$

These are derived from Eq. (14.7) in Appendix 14.B. To get $\mathbf{p}_T$, we then plug Eqs. (14.12) and (14.13) into Eq. (14.8). Except for in extremely rare cases, these p-values cannot be calculated analytically.

In both Eqs. (14.12) and (14.13), $p(\Lambda|\vec{d}^{\,\text{obs}})$ is the hyperposterior on Λ given the

observed data $\vec{d}^{\rm obs}$ (Eq. (14.21)) and $\pi_{\rm pop}(\vec{\lambda}^{\rm pred}|\Lambda)$ is the corresponding population distribution (Eq. (14.5)). Algorithmically, we draw Λ from $p(\Lambda|\vec{d}^{\rm obs})$ and then $\vec{\lambda}^{\rm pred}$ from $\pi_{\rm pop}(\vec{\lambda}^{\rm pred}|\Lambda)$ for each event in the catalog. In the case of the event-level p_T of Eq. (14.12), this Λ is also used to calculate $p(\vec{\lambda}^{\rm obs}|\vec{d}^{\rm obs},\Lambda)$, the population-reweighted individual-event posteriors, from which we draw $\vec{\lambda}^{\rm obs}$. Then, the test statistic T is calculated on $\vec{\lambda}^{\rm obs}$ and $\vec{\lambda}^{\rm pred}$ and the indicator function I is evaluated. Selection effects on the predicted parameters (Eq. (14.9)) are incorporated via the $P_{\rm det}(\vec{\lambda}^{\rm pred},\vec{\lambda}^{\rm obs})$ term.

Alternatively, in the case of the data-level p_T (Eq. (14.13)), there is no individual-event posterior, just the likelihood $\mathcal{L}(\vec{d}^{\rm pred}|\vec{\lambda}^{\rm pred})$ (Eq. (14.4)) which, crucially, does not depend on Λ. Algorithmically, the population model generates single-event parameters $\lambda^{\rm pred}$, which in turn are used to produce predicted data $d^{\rm pred}$. The predicted data are produced by evaluating a waveform model at $\lambda^{\rm pred}$ and injecting it onto Gaussian noise, using the same waveform model and noise power-spectral density as the observed data. To obtain a catalog $\vec{d}^{\rm pred}$ this is repeated $N_{\rm obs}$ times. Then, the test statistic T is calculated on $\vec{d}^{\rm obs}$ and $\vec{d}^{\rm pred}$ and the indicator function I is evaluated. Here, we think of the calculation of T as *including* finding max. $\mathcal{L}$ parameters from each $\vec{d}$. Selection effects on the predicted data are applied via $P_{\rm det}(\vec{d}^{\rm pred})$, cf. Eq. (14.11). Since by-definition we have already detected $\vec{d}^{\rm obs}$, no corresponding selection term needs to be applied.

Figure 14.5 shows distributions for a number of test statistics T and gives corresponding values of $\mathbf{p}_T$ for each. These $\mathbf{p}_T$ were previously shown in Fig. 14.1. As mentioned in preceding sections, we consider four choices of T: (from left to right) the mean, the standard deviation, the ratio of number of events with $\cos\theta \in [-0.33, 0.33]$ to $\cos\theta \in [0.33, 1]$, and the fraction with $|\cos\theta| > 0.5$. The final two choices of T are selected specifically to probe the bimodality in our population of interest. We again investigate results from the Gaussian (red, purple, blue) and realistic-noise (green) individual-event likelihood models.

The top row of Fig. 14.5 compares T computed on the event-level $\cos\theta_{\rm true}$: the horizontal axis shows T from an predicted catalog generated from one hyper-parameter draw Λ while the vertical axis T from a observed catalog with the same Λ. In other words, one scatter point corresponds to T calculated on one event-level PPC trace in Fig. 14.4. Each event-level p_T is calculated from

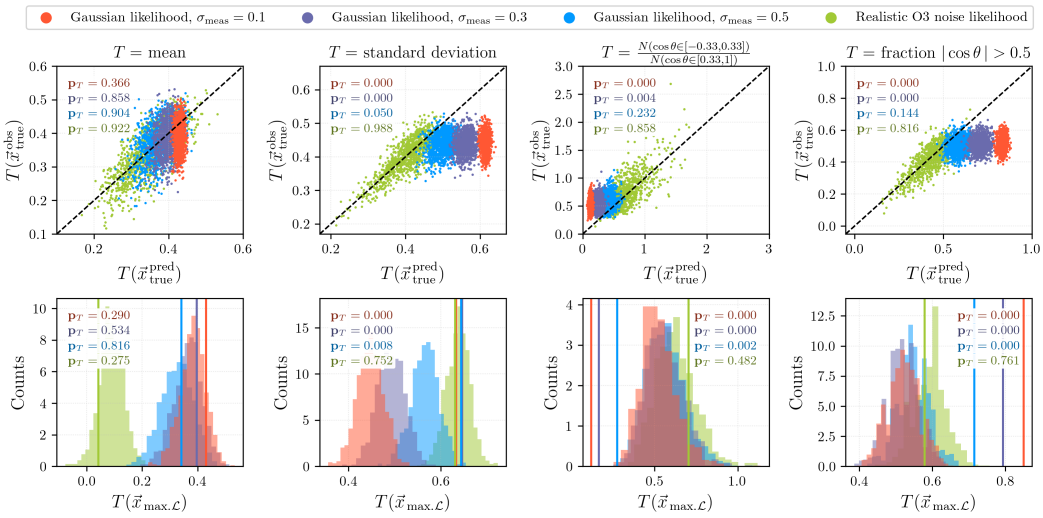


Figure 14.5: Distributions and associated posterior predictive p-values ($\mathbf{p}_T$) for four test statistics T calculated on $x = \cos\theta$ for 500 catalogs, each with 70 events, using the four likelihood variants: Gaussian likelihoods with $\sigma_{\rm meas} = 0.1$ (red), 0.3 (purple), and 0.5 (blue), compared to the realistic O3 noise likelihood (green). From left to right, T is the mean of $\cos\theta$ for each catalog, its standard deviation, the ratio of number of events with $\cos\theta \in [-0.33, 0.33]$ to $\cos\theta \in [0.33, 1]$, and the fraction of events with $|\cos\theta| > 0.5$. The top row shows T on event-level parameters ($\cos\theta_{\rm true}$), with values calculated from $\vec{d}^{\rm pred}$ on the horizontal axis and $\vec{d}^{\rm obs}$ on the vertical axis. The bottom row shows T on data-level parameters ($\cos\theta_{\rm max.\,\mathcal{L}}$), with the histogram showing those from $\vec{d}^{\rm pred}$ and the vertical line showing that from $\vec{d}^{\rm obs}$.

the fraction of scatter-points below the diagonal, which is then scaled to $\mathbf{p}_T$ (Eq. (14.8)). The bottom row of Fig. 14.5 shows each T on the data-level $\cos\theta_{\rm max.\,\mathcal{L}}$. The histogram represents values of T from predicted catalogs; the vertical line is T from the (single) observed catalog. Each data-level p_T is calculated from the fraction of points to the right of the vertical line, then again scaled to $\mathbf{p}_T$. Model misspecification is successfully identified when $\mathbf{p}_T \lesssim 0.05$.

For all four likelihoods, the data-level $\mathbf{p}_T$ are more discerning than the event-level. Most notable is the Gaussian likelihood with $\sigma_{\rm meas} = 0.5$ (blue) when T is the ratio of number of systems with $\cos\theta \in [-0.33, 0.33]$ to $\cos\theta \in [0.33, 1]$ (third column) or fraction of systems with $|\cos\theta| > 0.5$ (fourth column). When T is the former, the event-level is $\mathbf{p}_T = 0.232$, indicating a good model fit, while the data-level is $\mathbf{p}_T = 0.002$, well below the threshold for significance. The contrast is even more stark for the latter T: the event level $\mathbf{p}_T = 0.144 > 0.05$

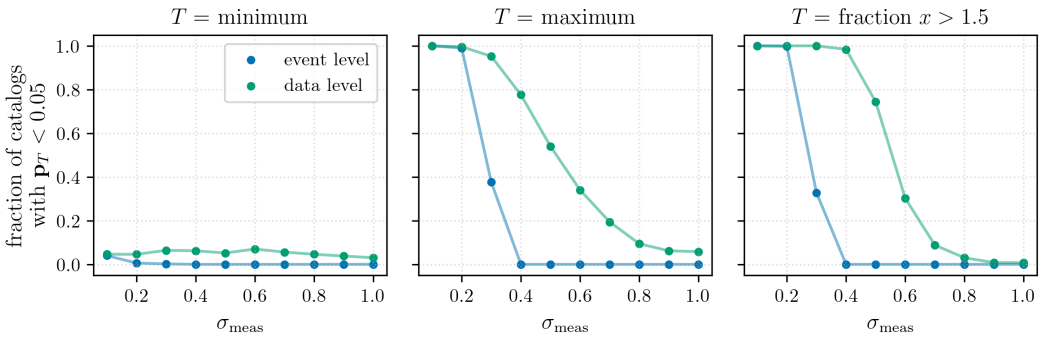


Figure 14.6: Single-event measurement uncertainty σ_{meas} versus the fraction of 500 simulated universes where $\mathbf{p}_T < 0.05$, using the analytic toy model described in Appendix 14.D and three representative test statistics T. (Note that although the hyper-posterior can be can be calculated analytically for the toy model, the distribution of p_T cannot for these choices of test statistic, hence sampling over 500 universes.) The employed population model is intentionally unable to fully characterize the true underlying distribution; see Fig. 14.17. A fraction of 1 means that the test can always tell that there is model mismatch; a fraction of 0 means it can never tell that there is model mismatch. The data-level $\mathbf{p}_T$ (green) using max. $\mathcal{L}$ parameters are always more discerning of model misspecification than the event-level $\mathbf{p}_T$ (blue) using true parameters, although both become uninformative when single-event measurement uncertainty is sufficiently large. The over-all degree of informativeness depends on which T is used. Here, e.g., T = minimum (first panel) is substantially less informative than T = maximum (second panel) or T = fraction of $x > 1.5$ (third panel).

and the data-level $\mathbf{p}_T = 0$.

For all T explored, the realistic O3 noise likelihood case (green) yields an event-level $\mathbf{p}_T > 0.8$, giving quantitative support to our claim that the event-level PPC cannot identify model misspecification. The data-level $\mathbf{p}_T$ fares marginally better, but is still always > 0.05. At least for this particular observed catalog instantiation, the data-level PPC is not considerably more discerning than the event-level PPC when detector noise is realistic. Of particular concern is the case where T is the fraction of events with $|\cos\theta| > 0.5$. Comparing the true underlying $\cos\theta$ distribution to the inferred population, the two are clearly very different: $\sim 85\%$ of the injected events have $|\cos\theta| > 0.5$, while the model predicts $\sim 50\%$ do. However, the corresponding observed max. $\mathcal{L}$ values have only $\sim 50\%$ with $|\cos\theta| > 0.5$, as seen in Fig. 14.4. The likelihood is sufficiently uninformative about $\cos\theta$ that we are losing nearly all information about the corresponding true distribution. The lack of informa-

tion about $\cos\theta$ in the realistic-noise likelihood is further highlighted in the figures in Appendix 14.C.

To move beyond a single test case, we compare event vs. data-level $\mathbf{p}_T$ over *many instantiations* of the observed catalog (i.e., many simulated universes). Such a test is only computationally feasible using a toy model and analytic likelihood, the details of which are given in Appendix 14.D. Once again measuring an underlying bimodal distribution with a Gaussian population model, we look at single-event likelihoods with ten values of σ_{meas} linearly spaced between 0.1 and 1. For each σ_{meas}, we generate 500 instantiations of observed catalogs, and generate 1,000 PPC traces for each. As in Fig. 14.5, we select several representative T. For each T and σ_{meas}, we calculate the fraction of these 500 catalogs for which $\mathbf{p}_T < 0.05$, indicating strong evidence for model mismatch (which we know exists), and plot this fraction vs. σ_{meas} in Fig. 14.6. A fraction of 1 means that the test can always tell that there is model mismatch; a fraction of 0 means it can never tell that there is model mismatch. Figure 14.6 thus shows that for the range of cases we consider **the data-level $\mathbf{p}_T$ is *always* equally or more discerning of model misspecification than the event-level $\mathbf{p}_T$.** The supremacy of the data-level $\mathbf{p}_T$ persists regardless of the measurement uncertainties and which T we consider—although both data- and event-level PPCs become less discerning as σ_{meas} increases and eventually reach a point where neither is informative. We may be hitting this limit in the realistic O3-noise likelihood explored in this work. Figure 14.6 also reiterates that **the choice of T that are most discerning depends on the specific population features of interest and the model used to measure it.**

While we can only confirm that data-level PPCs are more discerning than event-level PPCs in the specific cases explored in this work, we expect their superiority to translate across applications, as event-level PPCs will always include the impact of the prior. However, there remains a reason one might prefer to conduct an event-level PPC: using max. $\mathcal{L}$ values has a significantly larger computational cost. Obtaining max. $\mathcal{L}$ parameters for an arbitrarily large number of synthetic GW events is computationally intensive due to the high-dimensional parameter space and non-analytic likelihood model. Moreover, *consistently* computing max. $\mathcal{L}$ values for predicted vs. observed catalogs becomes finicky in the case of testing real observed data (using, by necessity,

simulated predicted data). The role of waveform models, power spectral densities, selection effects, etc. must be consistent between the (real) observed and (synthetic) predicted data, which becomes difficult when real GW catalogs consist of a mixture of each of these attributes, some of which are selected via difficult-to-replicate human decisions. Developing faster, more robust methods for determining max. $\mathcal{L}$ parameters (or other data-level quantities) across observed and predicted GW catalogs is left for future work.

14.5 Partial PPCs

One shortcoming of the traditional event-level PPCs of Sec. 14.4 is that they use information present in $\vec{d}^{\rm obs}$ to both estimate parameters *and* assess model incompatibility using $\mathbf{p}_T$ [562]. Partial PPCs (pPPCs) address this issue by avoiding the double-use of data by fixing a degree of freedom between observed and predicted catalogs [349, 564]. The distribution from which the predicted catalogs are drawn when conducting a pPPC changes from the traditional PPCs explored in Sec. 14.4. In the traditional PPC, the catalogs are drawn from the full posterior predictive distribution, which is conditioned on the observed data; in the pPPC, the distribution from which catalogs are drawn is *additionally* constrained to match the observed value of a chosen test statistic T_0. While the pPPC method we use still occurs at the event level and therefore uses reweighted posteriors, it only uses information not present in $T_0(\vec{d}^{\rm obs})$ when drawing from the posterior predictive distribution:

$$\mathrm{PPD}^{\rm pPPC}(\lambda^{\rm pred}_{\rm true}|\vec{d}^{\rm obs}, T_0) = \int \pi_{\rm pop}(\lambda^{\rm pred}_{\rm true}|\Lambda, T_0(\vec{d}^{\rm obs}))\, p(\Lambda|\vec{d}^{\rm obs})\, d\Lambda\,, \quad (14.14)$$

resulting in the p-value given in Eq. (14.15):

$$\text{Partial}: \quad p_T^{\rm pPPC}(\vec{d}^{\rm obs}, T_0) = \iiint I_{[T(\vec{\lambda}^{\rm pred})\geq T(\vec{\lambda}^{\rm obs})]}\; P_{\rm det}(\vec{\lambda}^{\rm pred}, \vec{\lambda}^{\rm obs}) \quad (14.15)$$
$$\times\; \pi_{\rm pop}(\vec{\lambda}^{\rm pred}|\Lambda, T_0(\vec{d}^{\rm obs}))\; p(\vec{\lambda}^{\rm obs}|\vec{d}^{\rm obs}, \Lambda)\; p(\Lambda|\vec{d}^{\rm obs})\; d\Lambda\; d\vec{\lambda}^{\rm obs}\; d\vec{\lambda}^{\rm pred}\,.$$

By fixing one degree of freedom between the predicted and observed catalogs, we can assess whether the remaining degrees of freedom become more informative.

To implement the pPPC, we use the following steps:

1. Choose a test statistic T_0 to keep constant between the predicted and observed catalogs.

2. Draw one sample of Λ from the hyperposterior $p(\Lambda|\vec{d}^{\rm obs})$.

3. *Observed values* ($\vec{\lambda}^{\rm obs}_{\rm true}$): Draw one sample from each individual-event posterior in the observed catalog, *reweighted* from its parameter-estimation prior to $\pi_{\rm pop}(\lambda_{\rm true}|\Lambda)$. Calculate $T_0(\vec{d}^{\rm obs})$.

4. *Predicted values* ($\vec{\lambda}^{\rm pred}_{\rm true}$): Draw $N_{\rm obs}$ *detectable* values of $\lambda_{\rm true}$ from the distribution $\pi_{\rm pop}(\lambda_{\rm true}|\Lambda)$. Calculate $T_0(\vec{d}^{\rm pred})$. If

$$T_0(\vec{d}^{\rm obs}) - \delta T \leq T_0(\vec{d}^{\rm pred}) \leq T_0(\vec{d}^{\rm obs}) + \delta T_0,$$

   continue to the next step. Otherwise, redraw $N_{\rm obs}$ events until the condition is satisfied, resulting in a predictive distribution following Eq. (14.14). We choose $\delta T_0 = (2N_{\rm events})^{-1/2}$.[6]

5. Repeat steps 2-4 for each Λ.

Ideally, using the pPPC to fix one statistic between the observed and predicted catalogs would lead to a more discerning p-value for other test statistics.

We focus on two cases of T_0. First, we choose a T_0 which is highly mismatched between the inferred vs. true underlying population distributions: the ratio of tilts with $\cos\theta \in [-0.33, 0.33]$ to tilts with $\cos\theta \in [0.33, 1]$. Results are shown in Fig. 14.7. Second, we define T_0 as the standard deviation of $\cos\theta$ draws per-catalog—a quantity which is theoretically well-measured by the Gaussian population model— with results shown in Fig. 14.8.

For both choices of T_0, conducting pPPCs on the results where the single-event likelihood is Gaussian with $\sigma_{\rm meas} = 0.1$ (red) and 0.3 (purple) yields predicted vs. observed traces that drastically differ in shape from those of the traditional event-level PPCs in Fig. 14.4. In these cases, fixing one degree of freedom does indeed yield more informative predictive checks and corresponding $\mathbf{p}_T$ for

[6] We cannot set $\delta T_0 = 0$ because we are dealing with catalogs of finite size. Instead, we want δT_0 to be as close to zero as possible. However, in reality disagreement between the predicted and observed populations can so extreme that it is practically impossible to satisfy

$$T_0(\vec{d}^{\rm obs}) - \delta T \leq T_0(\vec{d}^{\rm pred}) \leq T_0(\vec{d}^{\rm obs}) + \delta T_0$$

via random population draws if δT_0 is sufficiently small. We consider that standard error when dealing with finite events is $\propto (N_{\rm events})^{-1/2}$. The factor of 2 is included for additional constraining power; if increased we encounter sampling insufficiency. Figures 14.7 and 14.8 show the minor spread of $T_0(\vec{d}^{\rm pred})$ vs. $T_0(\vec{d}^{\rm obs})$ about the diagonal due to our choice of δT_0.

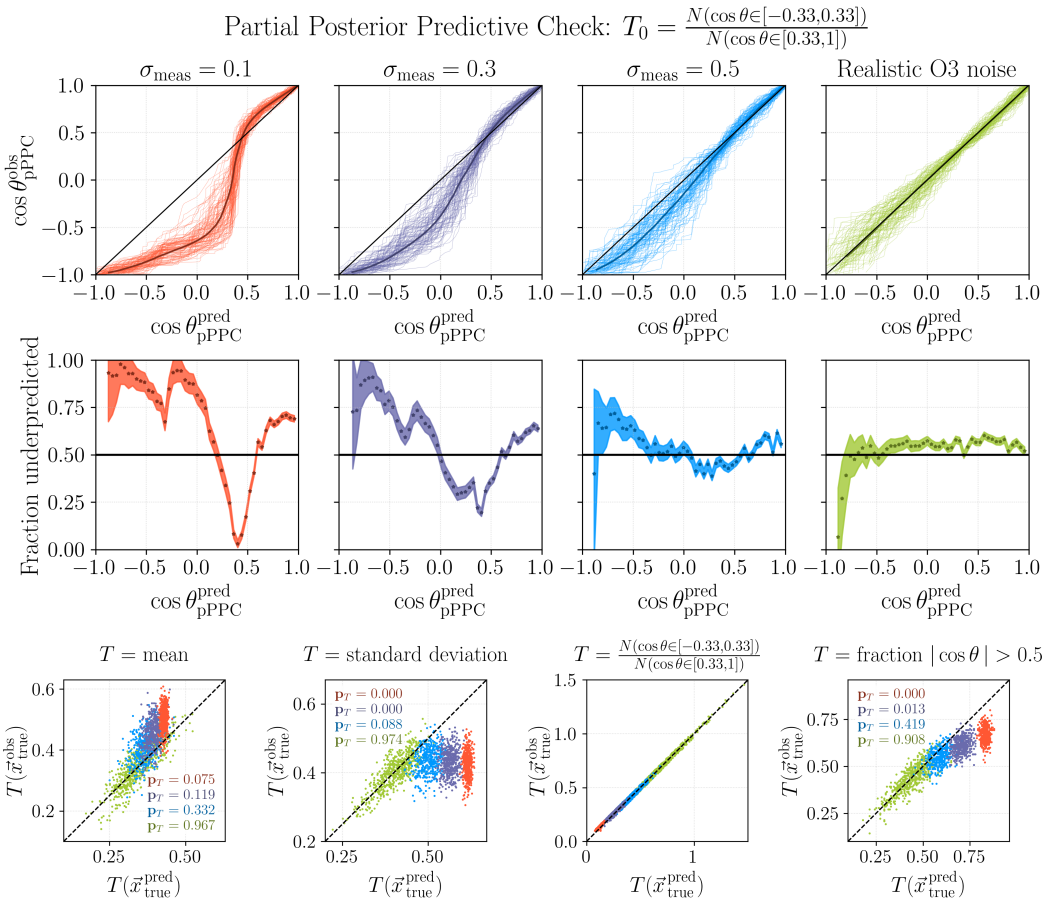


Figure 14.7: PPC traces (top row), fraction under-predicted (middle row), and $\mathbf{p}_T$ (bottom row) using a pPPC where the ratio of tilts with to tilts with $\cos\theta \in [0.33, 1]$ is the fixed degree of freedom, T_0. The third column of the bottom row shows predicted vs. observed T_0, which naturally follows the diagonal. Colors and labels retain meaning from Figs. 14.4 & 14.5.

(some of) the non-fixed test statistics T. The most pronounced improvement is when T = mean and T_0 = standard deviation, as reflected in the first column of Fig. 14.1. Here, the pPPC with T_0 as the standard deviation is actually *more* informative than even the data-level PPC. However, for other test statistics, fixing T_0 actually makes $\mathbf{p}_T$ less discerning (higher $\mathbf{p}_T$) than the traditional event-level PPC; e.g., when T = fraction $|\cos\theta| > 0.5$.

Looking towards the cases with higher single-event measurement uncertainty (Gaussian $\sigma_{\text{meas}} = 0.5$ and the realistic O3 noise), neither pPPC shows improvement over the traditional, event-level PPC. The only potentially significant feature is that the pPPC in Fig. 14.7 hints at the model under-predicting $\cos\theta \sim -0.5$. However, for most cases with large measurement uncertainty,

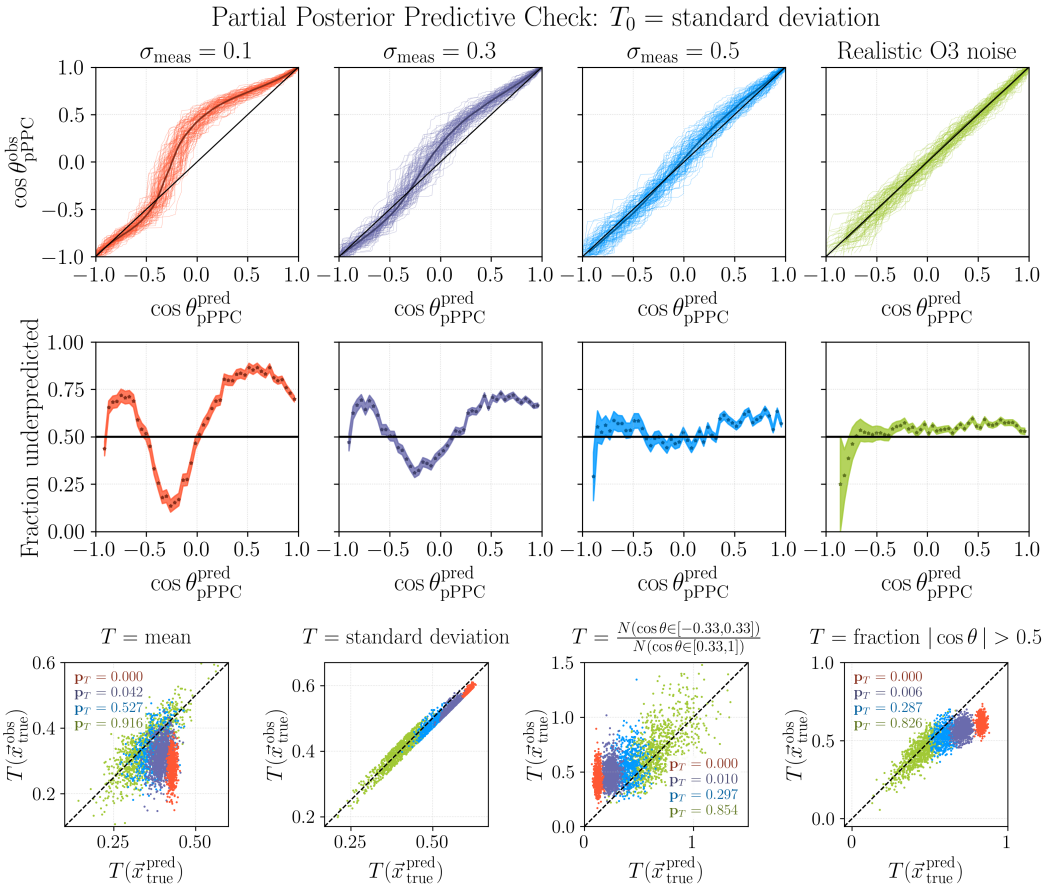


Figure 14.8: Same as Fig. 14.7 but where the fixed degree of freedom T_0 is the standard deviation of a catalog's $\cos\theta$ draws.

the reweighting process already results in predicted values that overly reflect the model, to the extent that fixing $T_0(\vec{d}^{\rm obs}) = T_0(\vec{d}^{\rm pred})$ has a sub-dominant effect and the data-level PPCs still reign supreme.

In conclusion, the efficacy of the pPPC depends on single-event measurement uncertainty and which specific population features we decide to probe. We find that, generally, **pPPCs are more discerning when the targeted feature is well-predicted by the model** (Fig. 14.8), rather than one that the model cannot capture (Fig. 14.7). However, when single-event uncertainty is large, reweighting has already created biased individual event posteriors, and there is no discernible advantage in using the pPPC to diagnose model misspecification over the traditional event-level PPC.

14.6 Split Predictive Checks

Lastly, we implement the split predictive check (SPC) [567] (also known as holdout predictive checks [568]) as an alternative way to perform event-level PPCs. To conduct an SPC, we divide the observed data into two sub-sets: one to infer the population distribution and another to generate predictive catalogs. The process is analogous to a leave-one-out analysis, in which population inference is performed on all but one event to see how the resultant population prior impacts the posterior of the left-out event [e.g., 275, 282, 327, 338]. The SPC again changes the distribution from which predicted catalogs are drawn:

$$p^{\rm SPC}_{\rm pop}(\lambda^{\rm pred}_{\rm true}|\{\vec{d}^{\rm obs}_1,\vec{d}^{\rm obs}_2\}) \int \pi_{\rm pop}(\lambda^{\rm pred}_{\rm true}|\Lambda,\vec{d}^{\rm obs}_1)\,p(\Lambda|\vec{d}^{\rm obs}_2)\,d\Lambda\,, \tag{14.16}$$

where here, the full set of observed data are $\vec{d}^{\rm obs} = \{\vec{d}^{\rm obs}_1,\vec{d}^{\rm obs}_2\}$. This PPD results in the $\mathbf{p}_T$ in given in Eq. (14.17), where we can see that the held out data $\vec{d}^{\rm obs}_2$ are used to generate the hyperposterior $p(\Lambda|\vec{d}^{\rm obs}_2)$, which is then used to draw $\lambda^{\rm pred}$ and reweight $p(\lambda^{\rm obs}|\vec{d}^{\rm obs}_1)$, from which we draw $\lambda^{\rm obs}$:

$$\begin{aligned}&\text{Split}:\\ p^{\rm SPC}_T(\vec{d}^{\rm obs}_1,\vec{d}^{\rm obs}_2) = &\iiint I_{[T(\vec{\lambda}^{\rm pred})\geq T(\vec{\lambda}^{\rm obs})]}\;P_{\rm det}(\vec{\lambda}^{\rm pred},\vec{\lambda}^{\rm obs})\\ &\times\;\pi_{\rm pop}(\vec{\lambda}^{\rm pred}|\Lambda)\;p(\vec{\lambda}^{\rm obs}|\vec{d}^{\rm obs}_1,\Lambda)\;p(\Lambda|\vec{d}^{\rm obs}_2)\;d\Lambda\;d\vec{\lambda}^{\rm obs}\;d\vec{\lambda}^{\rm pred}\,.\end{aligned} \tag{14.17}$$

To implement the SPC, we need to (in theory) rerun hierarchical inference on many combinations of events in the observed catalog. However, to avoid actually repeatedly re-running hierarchical inference, we use a hyperposterior reweighting scheme. The hierarchical likelihood of a catalog $\vec{d}^{\rm obs}$ (Eq. (14.22)) is the product of the hierarchical likelihoods of each event in that catalog. Therefore, we can write

$$\mathcal{L}(\vec{d}^{\rm obs}|\Lambda) = \mathcal{L}(\vec{d}^{\rm obs}_1|\Lambda)\,\mathcal{L}(\vec{d}^{\rm obs}_2|\Lambda)\;. \tag{14.18}$$

To generate Λ draws from $p(\Lambda|\vec{d}^{\rm obs}_2)$ in Eq. (14.17), we can then perform *weighted* draws from the existing $p(\Lambda|\vec{d}^{\rm obs})$ with weights w_Λ equivalent to

$$w_\Lambda = \frac{\mathcal{L}(\vec{d}^{\rm obs}_2|\Lambda)}{\mathcal{L}(\vec{d}^{\rm obs}|\Lambda)} = \frac{1}{\mathcal{L}(\vec{d}^{\rm obs}_1|\Lambda)}\,, \tag{14.19}$$

where

$$\mathcal{L}(\vec{d}^{\rm obs}_1|\Lambda) = \prod_{\rm i=0}^{n_1}\mathcal{L}(d^{\rm obs,i}|\Lambda) \tag{14.20}$$

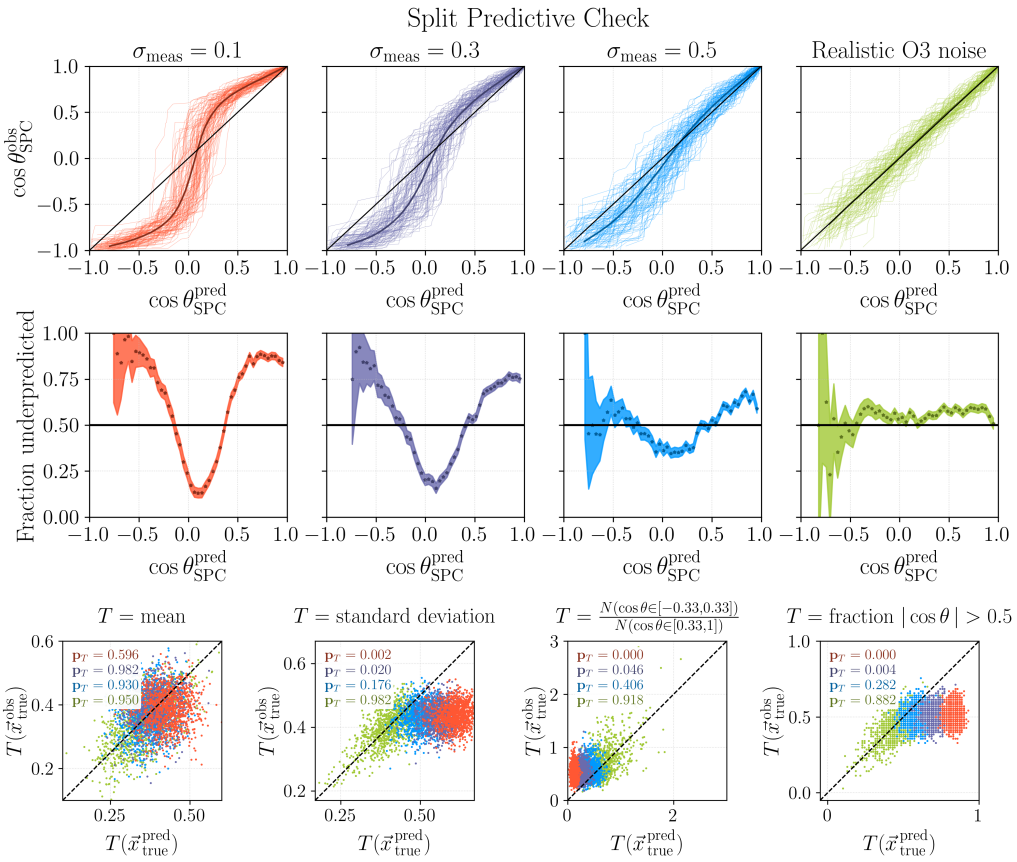


Figure 14.9: Same as Fig. 14.7 but for the SPC, rather than pPPC.

and n_1 is the number of events in the subset of events $\vec{d}^{\rm obs}_1$. These weights can be entirely calculated in post-processing, meaning we do not need to re-run hierarchical inference on any subsets of events.

Algorithmically, to conduct an SPC, we perform the following steps:

1. Randomly partition the full observed catalog into disjoint subsets $\vec{d}^{\rm obs}_1$ and $\vec{d}^{\rm obs}_2$, each containing half of the events, such that $n_1 = n_2 = N_{\rm obs}/2$.

2. Draw one sample of Λ from the hyperposterior $p(\Lambda|\vec{d}^{\rm obs}_2)$. In practice, to do this we perform a *weighted* draw from the already available $p(\Lambda|\vec{d}^{\rm obs})$ with the weights given in Eq. (14.19).

3. *Observed values* ($\vec{\lambda}^{\rm obs}_{\rm true}$): Draw one sample from each of the $N/2$ individual-event posteriors in $\vec{d}^{\rm obs}_1$, reweighted from its parameter-estimation prior to $\pi_{\rm pop}(\lambda_{\rm true}|\Lambda)$ [350].

4. *Predicted values* ($\vec{\lambda}^{\rm pred}_{\rm true}$): Draw $N/2$ detectable values of $\lambda_{\rm true}$ from the distribution $\pi_{\rm pop}(\lambda_{\rm true}|\Lambda)$.

5. Repeat steps 1-4 many times.

The results of the SPC are shown in Fig. 14.9. We find that the SPC shows no advantage over the traditional event-level PPC, and in fact is the least discerning of all the PPCs we test; see Fig. 14.1 for a comparison. Because we test SPCs on event-level parameters, they still involve the reweighting of individual events to the population model, causing the method to fail in identifying model mismatch. We attribute the SPC's heightened failure to the fact that it utilizes smaller predicted vs. observed catalogs (here, by a factor of two), leading to a wider spread in the traces and p-values in Fig. 14.9, effectively "blurring out" some of the information from Fig. 14.4. We leave exploring the efficacy of SPCs with an increased catalog size, as well as different ratios of $n_1 : n_2$, to future work.

14.7 PPCs applied to GWTC-4.0

We have shown that among our tested suite of PPCs, data-level PPCs are holistically the most capable of identifying model mismatch. Thus, we apply and compare the traditional event-level and data-level PPCs of Sec. 14.4 to the BBH mergers in the LVK's Fourth Gravitational-Wave Transient Catalog (GWTC-4.0) [7, 9], using the "Gaussian Component Spins" model of Ref. [23]. Under this model, the spin magnitude (χ) model is described by a Gaussian truncated on $[0, 1]$, and $\cos\theta$ is a mixture of a Gaussian distribution truncated on $[-1, 1]$ and an isotropic distribution. These population models are defined in Eqs. (B26) and (B27) of Ref. [23].

For the event-level PPCs, we use all 153 BBHs from GWTC-4.0 with a false-alarm rate of less than one per year [9]; for the data-level case, we only use the 84 new BBHs from O4a when generating PPC traces. The latter restriction allows the use a single injection set—that for O4a sensitivity estimates [265, 570][7]—and a single representative power spectral density [7, 34, 574, 575] for the predicted catalogs. To find max. $\mathcal{L}$ values for the data-level PPCs, we use

[7] The sensitivity estimates from the LVK's first two observing runs are semi-analytic [362] and do not contain all extrinsic parameters necessary to re-produce the GW signals [571, 572]. Sensitivity estimates from the LVK's third observing run also exclude the merger phase [573].

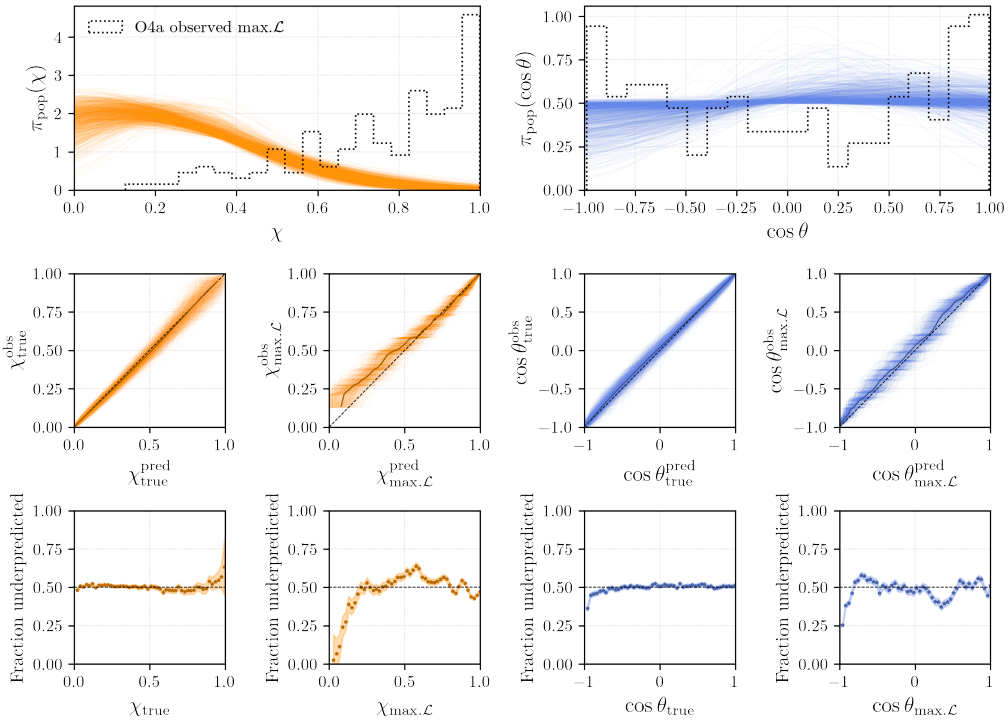


Figure 14.10: Results of standard event-level (left within each color) and data-level (right within each color) PPCs for GWTC-4.0 spin magnitudes (orange) and tilt (blue). *(First row)*: Recovered population distribution for χ and $\cos\theta$ (traces) given in Ref. [23], reproduced from their Fig. 7 [569] and compared to max. $\mathcal{L}$ values from the observed O4a population (histogram). Note that direct comparison of the traces versus histograms does not indicate model misspecification: the former showcases true parameters (used for event-level PPCs) while the latter shows max. $\mathcal{L}$ parameters (used for data-level PPCs); see also Fig. 14.14 in Appendix 14.C. *(Second row)*: PPC traces from 100 catalogs. *(Third row)*: Fraction of traces in each binned χ or $\cos\theta$ with a slope < 1, corresponding to the fraction of events in that parameter-range under-predicted by population model. The points and shaded region retain their meaning from Fig. 14.4.

the parameter-estimation code `cogwheel` [256, 298, 566] on both the predicted and observed catalogs and follow the procedure described in Appendix 14.C.

Figure 14.10 shows results for the event-level and data-level PPCs on GWTC-4.0 spin magnitudes (orange) and tilt angles (blue). Focusing first on the event-level PPCs, our results indicate that the **GWTC-4.0 population model is under-predicting systems with large spin magnitudes and over-predicting systems with perfectly anti-aligned tilts**. Similar conclusions were offered in the LVK analysis [23] using a weakly-parameterized B-Spline model [266, 272] for χ and $\cos\theta$, as well as a strongly-parameterized tilt model which enforces a hard cutoff minimum in $\cos\theta$, inferred to be between -0.90 and 0.56 at 90% credibility. Comparing our results to Fig. 7 in Ref. [23] indicates that the B-Spline model might be a better fit to the data than the Gaussian Components Spins model. The requirement of high spin magnitudes and a dearth of perfectly anti-aligned tilts are similarly found by Ref. [550] in the maximum likelihood population framework [549]. Aside from these discrepancies at large χ and negative $\cos\theta$, the event-level PPCs do not indicate any significant mismatch between the population model and the GWTC-4.0 data. Broadly, their traces tend to follow the 1:1 line between observed and predicted catalogs, indicating a good fit.

Turning to the data-level PPC results, we see substantially more deviation from the diagonal, potentially indicative of model misspecification. The fact that the observed max. $\mathcal{L}$ parameters for O4a (histograms in the top row of Fig. 14.10) do not agree with the population traces is not a direct indicator of model misspecification, as the former is a data-level quantity while the latter is an event-level quantity. Rather, we use the event-level traces to then generate predicted max. $\mathcal{L}$ values, as described in Appendix 14.C and shown in Fig. 14.14. As with the simulated populations (Fig. 14.4), the traces for the max. $\mathcal{L}$ PPCs are highly sensitive to the local fluctuations resulting from the particular noise realizations of each observed event. As discussed in Sec. 14.4, these fluctuations make it difficult to ascertain which features are "real" versus spurious. For the spin magnitude case, the minimum observed $\chi_{\mathrm{max.}\,\mathcal{L}} \approx 0.1$. If this feature is indeed representative of the underlying population, then the Gaussian χ population model is over-predicting spin magnitudes $\lesssim 0.1$. The data-level $\cos\theta$ PPC finds the same behavior as the event-level PPC as $\cos\theta \to -1$, but with more confidence. The remaining fluctuations in the χ and $\cos\theta$

max. $\mathcal{L}$ traces are consistent with the random Poisson variation we observe even with the lowest single-event uncertainty in Fig. 14.4.

We turn to p-values to further assess the significance to the potential model misspecification near the tails of the GWTC-4.0 χ and $\cos\theta$ distributions. Using the minimum and maximum of each quantity as a test statistic, we find that only the minimum of the $\cos\theta$ distribution yields a p-value below our threshold of significance. Here, $\mathbf{p}_T = 0.005$ for the data-level case, indicating model misspecification. The fact that the minimum of the χ distribution returns a data-level $\mathbf{p}_T = 0.37$ indicates that the observed minimum max. $\mathcal{L}$ of $\chi \sim 0.1$ is consistent with random noise fluctuations.

14.8 Recommendations and Future Work

To conclude, we remind readers that we summarize our main results at the beginning of this work in Sec. 14.2, and to refer back to Fig. 14.1 to view the p-values for all PPCs and test statistics explored in this work. For readers looking to use PPCs to probe model misspecification in GW population analyses, we make the following recommendations:

1. For moderate- to poorly-constrained parameters, we recommend the use of data-level PPCs alongside event-level PPCs. While any model misspecification uncovered by event-level PPCs is robust, we **caution against over-interpreting event-level PPC results that indicate a *good* model fit**.

2. Additionally, we advise against assigning significance to every fluctuation in data-level PPC traces that might stem from Poisson noise rather than a feature in the underlying astrophysical population. Small-number effects naturally cause fluctuations in fraction-underpredicted. Assessing the consistency of cumulative density functions [555] or comparing $\mathbf{p}_T$-values are complementary diagnostics to ensure a mismodeled feature is robustly identified.

3. For well-constrained parameters, we recommend the use of the traditional, event-level PPCs: these are more computationally efficient, and in the low measurement-uncertainty limit, perform equally as well as data-level PPCs. If negligible information exists about a parameter in the single-event likelihood, neither class of PPC will be informative.

4. When calculating posterior predictive p-values, we recommend always using a variety of test statistics, rather than just one.

5. If trying to probe a specific feature in the data, partial PPCs can be a useful tool, but we only recommend their use for moderately- to well-constrained parameters.

6. At current GW catalog size, we discourage the use of split-PPCs.

Moving forwards, we wish to explore PPCs on alternative data-level quantities beyond just the maximum likelihood parameters, such as search statistics. Likelihood optimization for poorly constrained parameters is difficult (Appendix 14.C), warranting both turning to other data-level properties as well as developing faster, more accurate methods for GW likelihood optimization. Additionally, given that decreased measurement uncertainty drastically improves PPC performance, we hope to investigate whether only conducting PPCs on subsets of events that pass an SNR threshold [576] yields any improvement. Finally, all of the methods explored in this work can also be applied to non-parametric models; we leave comparing PPCs on parametric to non-parametric approaches to future work.

Robust model assessment for GW source populations is crucial for downstream astrophysical implications. There is only so much information our data can provide us with, and knowing those limits is preferable to drawing false conclusions. Model checking will continue to be important as the catalog of observed GW signals grows in size, and our conclusions become increasingly influential.

Ch. 14 Data and Code Availability

Notebooks to generate all the figure (excluding Figs. 14.2 and 14.3) are available on Github at `https://github.com/simonajmiller/GW-PPCs`, as is a toy model elucidating the difference between event and data-level PPCs. The individual-event posterior samples and hyper-posterior samples used as the basis of this work are from Ref. [160]. All PPC traces generated from those data and GWTC-4.0 are included in the Github repository linked above.

Ch. 14 Acknowledgments

We thank Maya Fishbach, Thomas Callister, and Amanda Farah for the initial discussions that inspired this work in 2023; Javier Roulet and Colm Talbot

for assistance with likelihood optimization; and Noah Wolfe, Matthew Mould, Sofia Alvarez-Lopez, Jack Heinzel, and Salvatore Vitale for many insightful discussions about PPCs throughout this project. In particular, we extend immense gratitude to Sofia Alvarez-Lopez for serving as our internal LVK reviewer, and thank Derek Davis and Eric Thrane for helpful comments during review.

S.J.M. and K.C. were supported by NSF Grants PHY-2308770 and PHY-2409001. P.M.M. acknowledges support through the NSF Physics Frontiers Center award 1430284 and 2020265, and through the Gravitational Physics Professorship at ETH Zurich. This work was also supported by NSF Grant PHY-2150027 as part of the LIGO Caltech REU Program, which funded S.W.

The authors are thankful for LIGO Laboratory computing resources, funded by the National Science Foundation Grants PHY-0757058 and PHY-0823459. This research has made use of data or software obtained from the Gravitational Wave Open Science Center (gwosc.org), a service of the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation, as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. KAGRA is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica (AS) and National Science and Technology Council (NSTC) in Taiwan.

Software: EMCEE [252], BILBY [243], DYNESTY [253], COGWHEEL [256, 298, 566], LALSUITE [319], NUMPY [323], SCIPY [353], H5PY [402], H5IFY [446], MATPLOTLIB [320], SEABORN [354], PANDAS [324, 325], ASTROPY [355, 356].

Ch. 14 Appendices

14.A Simulated Population and Inference

We use the simulated LowSpinAligned population from Ref. [160] as the basis for probing the efficacy of various posterior predictive checking methods throughout this work. The population's bimodal $\cos\theta$ distribution allows for probing model mismatch with a (unimodal) Gaussian tilt model. For reference, we reproduce the LowSpinAligned population's underlying and detectable parameter distributions in Fig. 14.11. A detailed description of the simulated populations and their individual-event and population inference can be found in Appendices A, B, and C of Ref. [160]. Essential details are summarized as follows. For each underlying population we found the corresponding detectable population by applying a network optimal signal-to-noise ratio (SNR) cut of 10 in the LIGO Livingston, LIGO Hanford, and Virgo detector network using O3 power spectral densities [141]. GW data were simulated using IMRPhenomX-PHM waveform model [245] with Gaussian noise. Samples were drawn from the 15-dimensional posterior of binary parameters using the IMRPhenomX-PHM waveform model, with standard priors for all parameters. Posteriors were sampled stochastically using the implementation of the nested sampler Dynesty [253] in the parameter-estimation code Bilby [243]. A network optimal SNR cut of 10 was performed on the samples to ensure consistency with the selection process [264].

We establish the following notation for hierarchical/population inference. The posterior on population-level parameters Λ given a GW catalog $\vec{d}$ (also called the "hyper-posterior") is

$$p(\Lambda|\vec{d}) = \frac{\mathcal{L}(\vec{d}|\Lambda)\pi(\Lambda)}{p(\vec{d})}\,, \tag{14.21}$$

where the data $\vec{d} = \{d_i\}$ are from $N_{\rm obs}$ independent events, and the population likelihood (or "hierarchical likelihood") is

$$\begin{aligned}\mathcal{L}(\vec{d}|\Lambda) &\propto \prod_i^{N_{\rm obs}} \frac{\int \mathcal{L}(d_i|\lambda)\,\pi_{\rm pop}(\lambda|\Lambda)\,d\lambda}{\xi(\Lambda)} \\ &= \frac{1}{\xi(\Lambda)^{N_{\rm obs}}} \prod_i^{N_{\rm obs}} \int \mathcal{L}(d_i|\lambda)\,\pi_{\rm pop}(\lambda|\Lambda)\,d\lambda\,,\end{aligned} \tag{14.22}$$

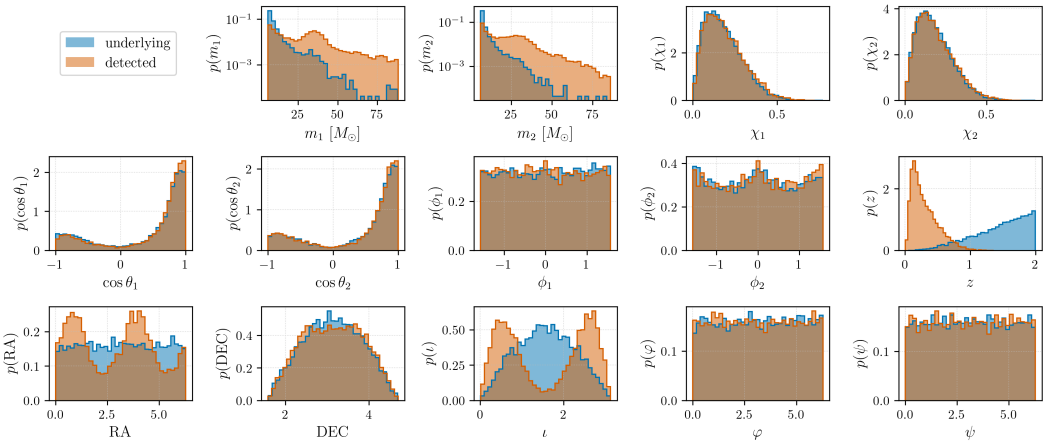


Figure 14.11: Underlying (blue) and detected (red-orange) distributions of all 15 BBH parameters for the LowSpinAligned population from Ref. [160], reproduced here for reference. In order, these are: masses m_i, spin magnitudes χ_i, spin tilt angles $\cos\theta_i$, spin azimuthal angles ϕ_i, redshift z, sky position (RA, DEC), inclination ι, phase φ, and polarization angle ψ.

with individual-event likelihoods $\mathcal{L}(d_i|\lambda)$ and a population distribution $\pi_{\rm pop}(\lambda|\Lambda)$. The detection efficiency $\xi(\Lambda)$—equivalent to the expected fraction of the population described by Λ—is calculated by

$$\xi(\Lambda) = \int \pi_{\rm pop}(\lambda|\Lambda)\, P_{\rm det}(\lambda)\, d\lambda\,, \tag{14.23}$$

where $P_{\rm det}(\lambda)$ is the probability that an individual event with parameters λ is detected. In Ref. [160], we conducted hierarchical inference using Markov Chain Monte Carlo (MCMC) sampling with the code emcee [252]. We used a Beta distribution for $\pi_{\rm pop}(\chi_i)$, Gaussian distribution for $\pi_{\rm pop}(\cos\theta_i)$, and mass/redshift model from Ref. [283]. Monte Carlo uncertainty was accounted for by excluding regions not meeting the effective sample criterion proposed by Ref. [280]. To estimate $P_{\rm det}$, we simulated many signals with λ drawn from a reference distribution and stored which are detectable [264, 280, 341, 524, 577]. We direct readers to Section 2 of Ref. [23] or Appendix C of Ref. [160] for more details about hierarchical Bayesian inference.

14.B Derivation of posterior predictive p-values

In this Appendix, we present the derivation of the expressions for the event- and data-level posterior predictive p-values given in Eqs. (14.12) and (14.13). We start from Eq. (14.7). For notational ease, define $y = \vec{d}^{\rm obs}$, $x = \vec{d}^{\rm pred}$, and

correspondingly $\lambda_y = \vec{\lambda}^{\rm obs}$, $\lambda_x = \vec{\lambda}^{\rm pred}$. In this notation, Eq. (14.7) becomes

$$p_T(y) = \iiiint I_{[T(x,\lambda_x)\geq T(y,\lambda_y)]}\, p(x,\lambda_x|\Lambda)\, p(\lambda_y,\Lambda|y)\; d\Lambda\, d\lambda_x\, d\lambda_y\, dx\,, \quad (14.24)$$

where $p(\lambda_y,\Lambda|y)$ indicates how observed data y give us constraints on $\{\lambda_y,\Lambda\}$, and $p(x,\lambda_x|\Lambda)$ is how the model with hyper-parameters Λ generates $\{x,\lambda_x\}$. We can expand each of these terms using conditional probabilities:

$$p(\lambda_y,\Lambda|y) = p(\lambda_y|y,\Lambda)\, p(\Lambda|y)\,, \quad (14.25)$$

and

$$p(x,\lambda_x|\Lambda) = p(x|\lambda_x,\Lambda)\, p(\lambda_x|\Lambda) = \mathcal{L}(x|\lambda_x)\, \pi_{\rm pop}(\lambda_x|\Lambda)\;. \quad (14.26)$$

In the above two expansions, $p(\lambda_y|y,\Lambda)$ is the population-informed posterior, $p(\Lambda|y)$ is the hyperposterior, $\mathcal{L}(x|\lambda_x)$ is the (normalized) product of single-event likelihoods over each event (Eq. 14.4), and $\pi_{\rm pop}(\lambda_x|\Lambda)$ is the population distribution (Eq. 14.5). Substituting into Eq. (14.24) yields

$$p_T(y) = \iiiint I_{[T(x,\lambda_x)\geq T(y,\lambda_y)]}\, \mathcal{L}(x|\lambda_x)\, \pi_{\rm pop}(\lambda_x|\Lambda)\, p(\lambda_y|y,\Lambda)\, p(\Lambda|y)\; d\Lambda\, d\lambda_x\, d\lambda_y\, dx\;. \quad (14.27)$$

Going from Eq. (14.27) to Eqs. (14.12) and (14.13) comes down to how we define our choice of test statistic T.

Derivation of the event-level p-value (Eq. 14.12): For event-level PPCs, we look at T that only depend on parameters $\{\lambda_x,\lambda_y\}$, not the data itself. Thus, we can integrate over x in Eq. (14.27) because the indicator function I no longer depends on x:

$$p_T^{\rm event}(y) = \int \mathcal{L}(x|\lambda_x)\, dx \iiint I_{[T(\lambda_x)\geq T(\lambda_y)]}\, \pi_{\rm pop}(\lambda_x|\Lambda)\, p(\lambda_y|y,\Lambda)\, p(\Lambda|y)\; d\Lambda\, d\lambda_x\, d\lambda_y$$
$$\implies p_T^{\rm event}(y) = \iiint I_{[T(\lambda_x)\geq T(\lambda_y)]}\, \pi_{\rm pop}(\lambda_x|\Lambda)\, p(\lambda_y|y,\Lambda)\, p(\Lambda|y)\; d\Lambda\, d\lambda_x\, d\lambda_y\,,$$

because the likelihood $\mathcal{L}(x|\lambda_x)$ is normalized in x for any λ_x. Finally, we must incorporate selection effects (cf. Eq. 14.9 in the main text):

$$p_T^{\rm event}(y) = \iiint I_{[T(\lambda_x)\geq T(\lambda_y)]}\, P_{\rm det}(\lambda_x,\lambda_y)\, \pi_{\rm pop}(\lambda_x|\Lambda)\, p(\lambda_y|y,\Lambda)\, p(\Lambda|y)\; d\Lambda\, d\lambda_x\, d\lambda_y\;. \quad (14.28)$$

Equation (14.28) is equivalent to Eq. (14.12).

Derivation of the data-level p-value (Eq. 14.13): For data-level PPCs, we look at T that only depend on the data. We are comparing the observed data

y (in the absence of any assumptions about Λ) to the data generated by the model described by Λ. Thus, we can integrate over λ_y in Eq. (14.27) because the indicator function I no longer depends on λ_y:

$$p_T^{\rm data}(y) = \int p(\lambda_y|y,\Lambda)\,d\lambda_y \iiint I_{[T(x)\geq T(y)]}\,\mathcal{L}(x|\lambda_x)\,\pi_{\rm pop}(\lambda_x|\Lambda)\,p(\Lambda|y)\,d\Lambda\,d\lambda_x\,dx$$
$$\implies p_T^{\rm data}(y) = \iiint I_{[T(x)\geq T(y)]}\,\mathcal{L}(x|\lambda_x)\,\pi_{\rm pop}(\lambda_x|\Lambda)\,p(\Lambda|y)\,d\Lambda\,d\lambda_x\,dx\,,$$

because the posterior $p(\lambda_y|y,\Lambda)$ is normalized in λ_y for any y,Λ. Again, we must incorporate selection effects, yielding

$$p_T^{\rm data}(y) = \iiint I_{[T(x)\geq T(y)]}\,P_{\rm det}(x)\,\mathcal{L}(x|\lambda_x)\,\pi_{\rm pop}(\lambda_x|\Lambda)\,p(\Lambda|y)\,d\Lambda\,d\lambda_x\,dx\,. \quad (14.29)$$

Equation (14.29) is equivalent to Eq. (14.13).

14.C Optimizing the Full 15-dimensional BBH GW likelihood

Many computational tools exist to obtain robust estimates of the full 15-dimensional[8] BBH GW posterior distribution [e.g., 243, 256, 257, 261]. However, perhaps counterintuitively, calculating the maximum likelihood parameters for such a distribution remains more difficult. Naively, one could simply approximate the posterior distribution via stochastic sampling, and then take the sample that happens to have the highest likelihood value. For leading-order parameters (like chirp mass), this approximate max. $\mathcal{L}$ will be fairly close to the true max. $\mathcal{L}$. For hard-to-constrain parameters like spin tilts that only weakly influence the likelihood, the same is not guaranteed, as shown in the left-hand panel of Fig. 14.12

Both the max. $\mathcal{L}$ sample and the true max. $\mathcal{L}$ point are valid data-level parameters, as long as they are compared in an apples-to-apples way between predicted and observed data. However, running full parameter estimation on tens of thousands of simulated GW signals is computationally infeasible. Thus, while max. $\mathcal{L}$ samples are readily available for the observed catalog, they remain difficult to obtain for predicted catalogs. As such, we turn to likelihood optimization for both the predicted and observed catalogs to obtain true max. $\mathcal{L}$ values.

To optimize the GW likelihood, we use the parameter estimation code `cogwheel` [298, 566, 578] for both the simulated GWTC-3.0-like case (Sec. 14.4) and the real

[8] (17-dimensional if including orbital eccentricity)

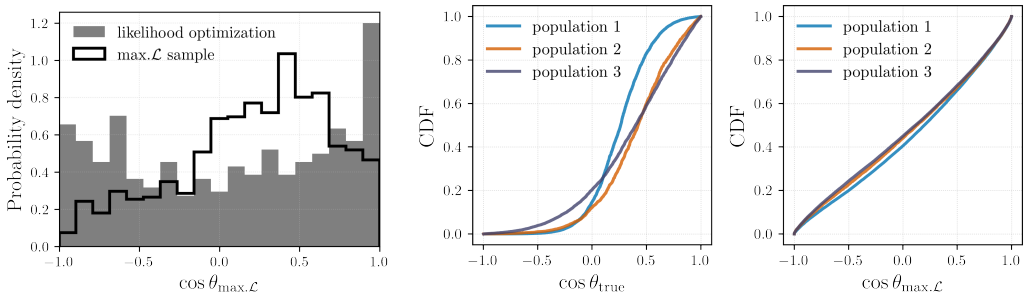


Figure 14.12: (*Left*): Histogram of max. $\mathcal{L}$ samples from stochastically sampled posteriors (empty) versus max. $\mathcal{L}$ values obtained via full likelihood optimization (filled) for the same underlying population. (*Middle, right*): The cumulative density functions (CDFs) for three example populations' true and corresponding max. $\mathcal{L}$ $\cos\theta$ distributions, for the case of realistic O3 noise. While the three true $\cos\theta$ distributions are very different, the max. $\mathcal{L}$ distributions are nearly indistinguishable.

GWTC-4.0 case (Sec. 14.7). The built-in likelihood optimizer in `cogwheel` uses the differential evolution algorithm encoded in `scipy` [353]. A comparison of injected spin tilts vs. their recovered max. $\mathcal{L}$ values for three example populations are shown in the middle and right-hand panels of Fig. 14.12. These three underlying populations are distinct in tilt (middle panel). However, their max. $\mathcal{L}$ distributions are nearly identical (right panel). If we trust that the likelihood optimizer is indeed finding the true max. $\mathcal{L}$ tilts, then Fig. 14.12 indicates that the likelihood (at O3 sensitivity) contains very little information about $\cos\theta$. This is further reflected in Fig. 14.13, where we compare the predicted vs. observed $\cos\theta$ distributions for the four single-event likelihoods explored in this work. With Gaussian likelihoods, the predicted vs. observed max. $\mathcal{L}$ distributions are visibly distinct, indicating that these likelihoods indeed contain information about $\cos\theta$. For the realistic O3-noise likelihood, the predicted vs. observed max. $\mathcal{L}$ distributions become visibly similar, echoing what is showcased in Figs. 14.4 and 14.12. Additionally, as single-event measurement uncertainty increases (left to right), the max. $\mathcal{L}$ distributions look increasingly dissimilar to the injected, true distribution (black-dashed), further exemplifying the likelihood's loss of information as σ_{meas} increases.

As hinted at in Fig. 14.13, the distribution of true vs. max. $\mathcal{L}$ parameters will generally not be identical. We show this explicitly for the case of GWTC-4.0 results in Fig. 14.14 for spin magnitudes χ (orange) and tilts $\cos\theta$ (blue), cf. Sec. 14.7 in the main text. Even if the left (max. $\mathcal{L}$ parameters) vs. right

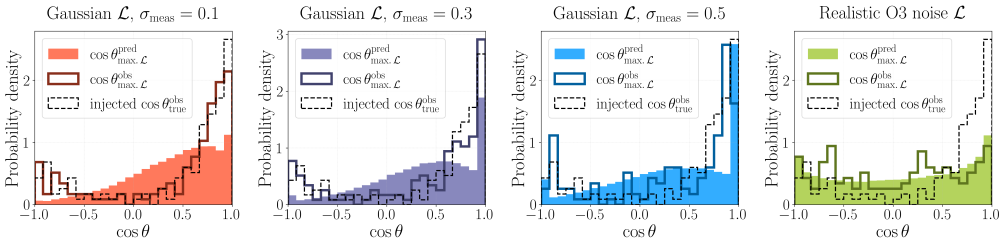


Figure 14.13: Predicted (filled) vs. observed (open) max. $\mathcal{L}$ $\cos\theta$ distributions for Gaussian likelihoods with $\sigma_{\rm meas} = 0.1$ (red), 0.3 (purple), and 0.5 (blue), and the realistic O3 noise likelihood (green). The histograms show the concatenation of all PPC traces in Fig. 14.4. In black-dashed, we additionally plot the injected $\cos\theta$ distribution, which illuminates that as $\sigma_{\rm meas}$ increases (left to right), the max. $\mathcal{L}$ distribution retains very little information from the true underlying distribution.

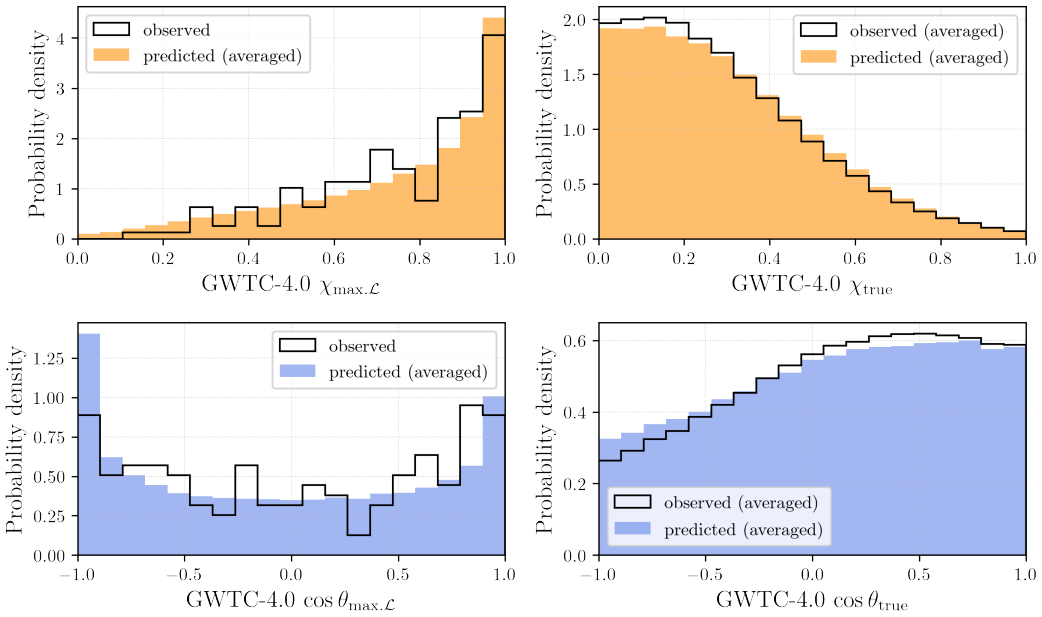


Figure 14.14: Observed (empty) versus predicted (filled) parameter distributions for max. $\mathcal{L}$ (left column) and true (right column) values of χ (top row) and $\cos\theta$ (bottom row) for GWTC-4.0, averaged over all of the traces in Fig. 14.10.

(true parameters) subplots house visibly different distributions, the observed (empty) vs. predicted (filled) histograms within each subplot look similar, corresponding to PPCs which indicate good model fit. Figure 14.14 also helps visualize our conclusion that the GWTC-4.0 Gaussian Component Spins model [23] over-estimates BBHs with $\cos\theta \sim -1$.

14.D Simple Toy Model

To help gain intuition, we present a simple toy model where both the individual-event likelihoods and population distribution are Gaussian. In this specific case, the hierarchical likelihood of Eqs. (14.21) and (14.22) can be calculated analytically. We look at a single individual-event level variable called x, which can take on any real value between $-\infty$ and ∞, and assume no selection effects, i.e., every value of x is equally likely to be detected. The Gaussian population distribution for $\pi_{\rm pop}(x)$ is described by two parameters: the mean μ and standard deviation σ. We implement uniform priors on both. Thus, in this toy model, Eq. (14.21) (with $\Lambda = \{\mu, \sigma\}$ becomes

$$p(\mu, \sigma|\vec{d}) \propto \mathcal{L}(\vec{d}|\mu, \sigma) \propto \prod_i^N \int_{-\infty}^{\infty} \mathcal{L}(d_i|x)\, \pi_{\rm pop}(x|\mu, \sigma)\, dx\,, \tag{14.30}$$

where the population distribution model is

$$\pi_{\rm pop}(x|\mu, \sigma) = \mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{Exp}\Big[-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big]\,. \tag{14.31}$$

and the individual-event likelihood is

$$\mathcal{L}(d_i|x) = \mathcal{N}(x|x_i^{\rm max.\mathcal{L}}, \sigma_{\rm meas}) = \frac{1}{\sqrt{2\pi\sigma_{\rm meas}^2}} \mathrm{Exp}\Big[-\frac{1}{2}\Big(\frac{x - x_i^{\rm max.\mathcal{L}}}{\sigma_{\rm meas}}\Big)^2\Big]\,, \tag{14.32}$$

where $\sigma_{\rm meas}$ is the measurement uncertainty and $x_i^{\rm max.\mathcal{L}}$ is the maximum likelihood point for d_i. Under this Gaussian individual-event likelihood,

$$x_i^{\rm max.\mathcal{L}} \sim \mathcal{N}(x_i^{\rm true}, \sigma_{\rm meas})\,. \tag{14.33}$$

Solving Eq. (14.30)—the integral of the product of two Gaussian distributions—yields

$$p(\mu, \sigma|\vec{d}) \propto \prod_i^N \mathcal{N}\big(x_i^{\rm max.\mathcal{L}}|\mu, \sqrt{\sigma^2 + \sigma_{\rm meas}^2}\big)\,, \tag{14.34}$$

or equivalently,

$$p(\mu, \sigma|\vec{d}) \propto [2\pi(\sigma^2 + \sigma_{\rm meas}^2)]^{-N/2} \times \mathrm{Exp}\Big[-\frac{1}{2(\sigma^2 + \sigma_{\rm meas}^2)} \sum_i^N (x_i^{\rm max.\mathcal{L}} - \mu)^2\Big]\,. \tag{14.35}$$

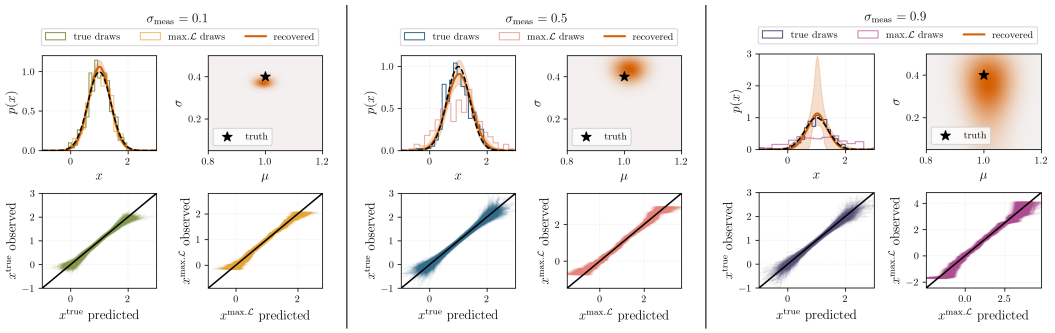


Figure 14.15: Results from hierarchical inference performed analytically with a Gaussian population model and Gaussian individual-event likelihoods. From the left to right sub-figures divided by vertical lines, the per-event measurement uncertainty increases from $\sigma_{\rm meas} = 0.1, 0.5, 0.9$. Within each sub-figure: *(Top left)* The true underlying population distribution (black dashed) compared to a 300 event catalog drawn from that distribution, with true values of x and corresponding max.$\mathcal{L}$ values of x. The median and 90% credible region of the reconstructed distribution (from μ, σ in the top right panel) are shown in orange. *(Top right)* Hierarchical likelihood (Eq. (14.35)) on μ, σ given the set $\vec{x}_{\rm max.\mathcal{L}}$ in the histogram. Darker orange indicates a higher likelihood, and the true underlying value is marked with a black star. *(Bottom left)* Event-level PPC traces, consisting of $x^{\rm true}$ for 1,000 catalogs each with 300 events. *(Bottom right)* Data-level PPC traces, consisting of $x_{\rm max.\mathcal{L}}$ for the same number of catalogs and events.

The likelihood in Eq. (14.35) is analytically tractable for a given set of $\{x_i^{\rm max.\mathcal{L}}\}$. In this toy model, the maximum likelihood values of x are randomly drawn from a set of $\{x_i^{\rm true}\}$ using Eq. (14.33). The true values of x can come from any underlying population distribution.

We first present an example in Appendix 14.D where the true underlying population is also Gaussian, and thus can be accurately recovered by the population model. Then, in Appendix 14.D, we look at a case where the true underlying distribution is highly non-Gaussian, and therefore cannot be well-fit by the population model. In each, we perform event and data-level PPCs and calculate posterior predictive p-values for a series of test statistics.

In which the Gaussian population model is a good fit to the data

We begin by examining hierarchical inference, PPCs, and posterior predictive p-values for a case where the Gaussian population model is a good fit to the observed data, with results shown in Figs. 14.15 and 14.16. For the true underlying population, we use a Gaussian distribution with $\mu^{\rm true} = 0.4$ and $\sigma^{\rm true} = 0.4$,

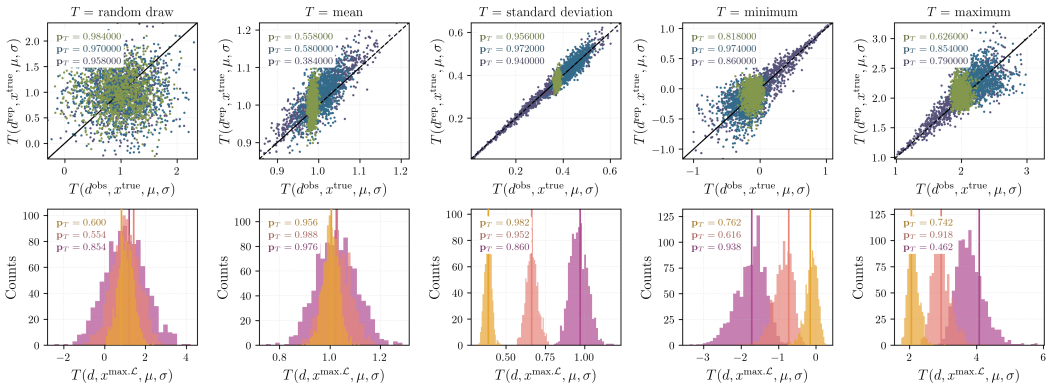


Figure 14.16: Distributions and associated posterior predictive p-values ($\mathbf{p}_T$) for five test statistics T from the PPCs shown in Fig. 14.15, with per-event measurement uncertainties $\sigma_{\rm meas} = 0.1$ (lightest shade of each color), 0.5 (middle shade), and 0.9 (darkest shade). From left to right, T is a random draw of x from the catalog, the catalog's mean, its standard deviation, minimum, and maximum. The top row show T on event-level parameters ($x^{\rm true}$), with values calculated from $\vec{d}^{\rm obs}$ on the horizontal axis and $\vec{d}^{\rm pred}$ on the vertical axis. The bottom row shows T on data-level parameters ($x_{\rm max.\,\mathcal{L}}$), with the histogram showing those from $\vec{d}^{\rm pred}$ and the vertical line showing that from $\vec{d}^{\rm obs}$.

shown in the upper left panel of each sub-figure of Fig. 14.15 by a black-dashed line. We simulate 300 events from this population, whose true values $\{x_i^{\rm true}\}$ and corresponding $\{x_i^{\rm max.\mathcal{L}}\}$ are shown in each histogram, where we look at three possible measurement uncertainties: $\sigma_{\rm meas} = 0.1, 0.5, 0.9$, from left to right sub-figures. The upper right subplots show the posterior probability (orange gradient) on μ, σ—as calculated analytically with Eq. (14.35)—compared to the true value (black star); deeper orange indicates a higher likelihood. From the posterior on μ, σ we can reconstruct the population distribution over x: the orange lines and bands in each upper left subplot shows the median and 90% credible interval of the reconstructed distribution. As expected, the Gaussian population model accurately recovers the true distribution, with uncertainty increasing with $\sigma_{\rm meas}$.

The second row of Fig. 14.15 shows event-level (lower left, cool-toned colors) and data-level (lower-right, warm-toned colors) PPC traces. Each of the 1,000 traces corresponds to a 300-event catalog. These are calculated using the procedures outlined in Sec. 14.4 of the main text. As in the main text, the event-level PPCs are calculated on the values of $x^{\rm true}$, while the data-level are calculated on values of $x_{\rm max.\,\mathcal{L}}$. For all $\sigma_{\rm meas}$, both the event- and data-level PPCs are, by eye, consistent with the diagonal, indicating good model fit.

We next calculate posterior-predictive p-values ($\mathbf{p}_T$) using five example test statistics, T, with results shown in Fig. 14.16. From left to right, T is defined as a random draw from a catalog, the catalog's mean, its standard deviation, its minimum, and its maximum—where one catalog corresponds to one trace from the second row of Fig. 14.15. The first row (scatter-plots) shows test statistics and associated p-values from the event-level PPCs, while the second (histograms) shows data-level results. Colors correspond to Fig. 14.15. All p-values, for both the event- and data-level tests, are well above the acceptable threshold of $\mathbf{p}_T = 0.05$, quantifying what we already know: the model is a good fit to the data.

In which the Gaussian population model is a poor fit to the data

We repeat the above exercise with a simulated population that *cannot* be accurately recovered by a Gaussian model, with results shown in Figs. 14.17 and 14.18. Here, the true underlying population is a bimodal distribution—the mixture of two Gaussian distributions, with the sub-dominant component having $\mu_1, \sigma_1 = (0.5, 0.4)$, the dominant component having $\mu_2, \sigma_2 = (2, 0.2)$, and a mixing fraction of 0.2 between them. Within each sub-figure of Fig. 14.17, the true population distribution is shown in the upper left panel with a black-dashed line, again compared to our simulated observed catalog of 300 values of $x_{\rm true}$ and corresponding $x_{\rm max.\,\mathcal{L}}$. The analytically-calculated hierarchical likelihood on μ, σ is shown in each upper right panel, with darker orange indicating a higher likelihood.

Clearly, the Gaussian population model does not find the true underlying population. Its 90% recovered credible interval (orange, upper left subplot of Fig. 14.17) does not encompass the truth, for any of the three values of $\sigma_{\rm meas}$. The true values of the mean and standard deviation of either of the true distribution's subpopulations are far away from the most probable μ, σ for the single Gaussian. This makes sense: the Gaussian distribution is indeed finding the mean and width of the underlying population, these two degrees of freedom just do not fully characterize the true, more complicated distribution.

For the $\sigma_{\rm meas} = 0.1$ case, both the event and data-level PPCs shown deviations from the diagonal, while neither do when $\sigma_{\rm meas} = 0.9$. We thus focus on the middle case of $\sigma_{\rm meas} = 0.5$, which is coincidentally the closest analog to BBH spins. The discrepancy between model and truth can be clearly seen

by eye in the data-level PPC traces, shown in pink in the bottom right panel of Fig. 14.17's middle sub-figure—but *not* so much in the event-level traces (lower left panel, blue).

The strength of the data-level PPC is also reflected in the posterior predictive p-values in Fig. 14.18. For $\sigma_{\text{meas}} = 0.5$, the event-level p-values are all well above the threshold of 0.05, even though we know *a priori* that the model is not a good fit to the data. For some choices of T, the data-level p-values are much more informative. Although the random draw (first column), mean (second column), and standard deviation (third column) test statistics would indicate that our model is a very good fit (which makes sense, given that we are using a Gaussian model which *does* accurately constrain the mean and standard deviation degrees of freedom), the others are either more inconclusive or return a negative result. Most notable is when T is the distribution's maximum (fourth column): In the event-level case, $\mathbf{p}_T = 0.34$, well above the threshold. In the data-level case, however, $\mathbf{p}_T = 0.004$, an order of magnitude below the threshold. The histogram of T further tells us that the model nearly always over-predicts the maximum $x_{\text{max.}\mathcal{L}}$ of a catalog. In the upper-left panel of Fig. 14.17, we see that the distributions' upper tails are where the true and inferred populations are most discrepant, consistent with what is reported by $\mathbf{p}_T$. The fraction of events with $x > 1.5$ (the inflection point in the true underlying bimodal distribution) is also flagged by the data-level PPC, with $\mathbf{p}_T = 0.06$ in the data level case—right on the threshold for passing—as compared to $\mathbf{p}_T = 0.7$ in the event-level PPC. Even the $\sigma_{\text{meas}} = 0.9$ case has a relatively low $\mathbf{p}_T = 0.108$ for this final choice of test statistic.

The Gaussian toy model has trivial computational cost, making it easy to repeat population-inference and calculations of $\mathbf{p}_T$ for many instantiations of $\vec{d}^{\text{obs}}$, a process which remains computationally infeasible for the results presented in the main text. We repeat the results shown in Fig. 14.18 over ten values of σ_{meas} linearly spaced between 0.1 and 1. For each σ_{meas} we generate 500 instantiations of observed catalogs catalogs (thus repeating the exercise in Fig. 14.18 5000 times total). For each T and σ_{meas}, we calculate the fraction of these 500 catalogs for which $\mathbf{p}_T < 0.05$, indicating strong evidence for model mismatch (which we know exists). A fraction of 1 means that the test can always tell that there is model mismatch; a fraction of 0 means it can never tell that there is model mismatch. Results are presented in Fig. 14.6

in the main text, and show that the data-level $\mathbf{p}_T$ are always more discerning of model misspecification than the event-level $\mathbf{p}_T$, although both become uninformative when single-event measurement uncertainty is sufficiently large. For comparison, when repeating this analysis with the good-fit population of Appendix 14.D, the fraction of $\mathbf{p}_T < 0.05$ is clustered at 0 for all test statistics at both the event and data level parameters.

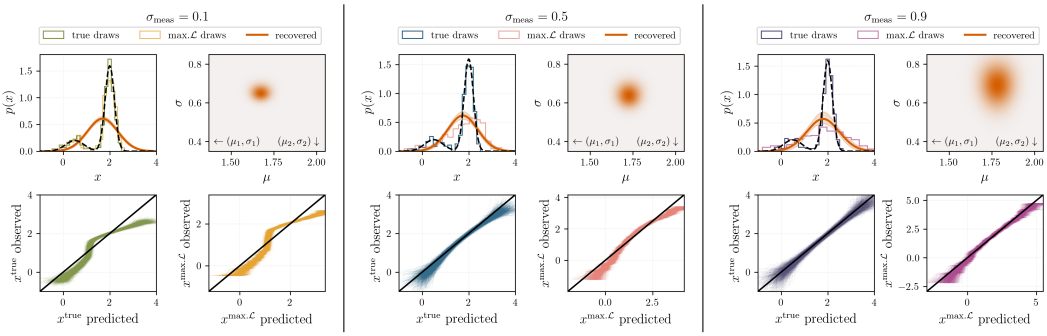


Figure 14.17: Same as Fig. 14.15 but with a true underlying distribution that is bimodal, and therefore cannot be accurately recovered by the Gaussian population model. For the $\sigma_{\rm meas} = 0.1$ case, both the event and data-level PPCs shown deviations from the diagonal. For the $\sigma_{\rm meas} = 0.5$ case, only the data-level PPC is informative.

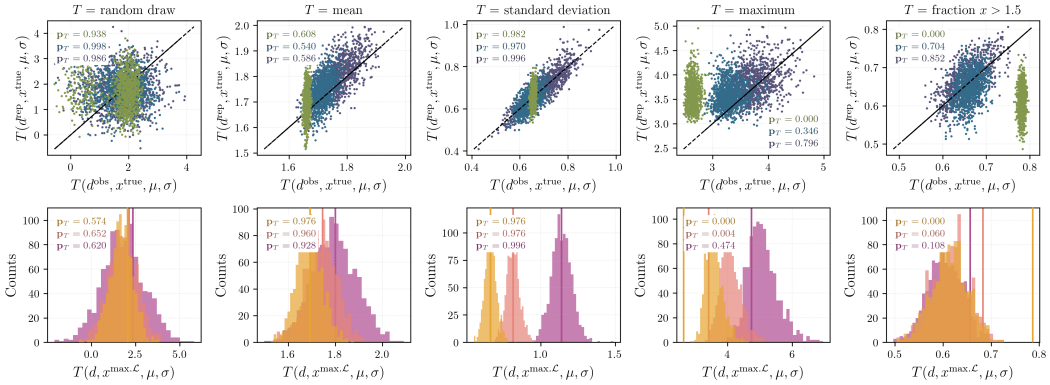


Figure 14.18: Similar to Fig. 14.16 but for the bimodal underlying population shown in Fig. 14.17, and (from left to right) test statistics T defined as a random draw from a catalog, the catalog's mean, its standard deviation, its maximum, and the fraction greater than 1.5 (the inflection point of the bimodal distribution). Colors correspond to Fig. 14.17.

Chapter 15

CONCLUSION

15.1 Summary of dissertation

In this dissertation, I have presented single-event and population-level measurements of the spins of merging binary black holes (BBHs) via gravitational waves (GWs), along with the tools I have developed to ensure those measurements are robust. In summary, I found that, for individual events, we must **connect measured parameters with their phenomenology** in GW signals. To make spin measurements for the BBH population robust, we need to thoroughly **test model behavior** against simulated populations, understand the role of **Monte Carlo uncertainty**, and whenever feasible, use **data-level quantities to probe model misspecification**, rather than event-level.

In more detail, Part I centered the question *what spin parameters are measurable?* I began with an overview of GW signals and how these encode the spins of a BBH merger source in the inspiral (pre-merger) and ringdown (post-merger) phases of coalescence (Ch. 2), followed by an introduction to the Bayesian framework used to extract spins and other source properties from observed GW data for single events and their astrophysical population (Ch. 3). I then presented a data-driven method for recovering the imprint of spin in individual GW signals, which revealed that the leading-order effective spin parameters described in Ch. 2 are insufficient to fully describe the data as less and less inspiral is visible, i.e., the BBH total mass increases (Ch. 4). Part I concluded with an exploration of spin measurability at the population-level, where I found that we can distinguish between BBH populations with different spin magnitude and tilt distributions, even if they have the same leading-order effective-spin distribution (Ch. 5).

In Part II, I dissected GW data from several high-mass BBHs in the time domain to answer *where in GW signals do measurements of BBH spin originate?* I began with a description of how I used time-domain inference to track changes in parameter estimates as various time segments of the data are included/excluded (Ch. 6). Using the method, I showed that GW190521's previously-elusive spin precession measurement stems from the suppression of

the signal's final pre-merger cycle with respect to the loud merger, and linked that morphology to the precession-induced motion of the binary's orbital plane (Ch. 7). For massive and highly precessing signals, I found that GW190521's morphology is not generic: whether or not precession is actually measurable depends on the specific (and astrophysically random) angular orientation of the BBH, with some configurations yielding informative precession constraints in the ringdown data alone (Ch. 8). I also applied time-domain inference to localize the anti-aligned spin measurement of GW191109 to a stretch of the inspiral concurrent with a detector glitch (Ch. 10), as well as to find significant observational support for Hawking's Area Law with GW250114 (Ch. 11). My work in Part II highlights that each GW detection deserves careful scrutiny to validate claims of unique spin configurations and safeguard against false astrophysical conclusions.

Finally, in Part III, I moved beyond constraints from individual GW events and asked *what do we know about the spins of the BBH population?* I first showed that the majority of BHs in binaries in GWTC-3.0 have non-negligible spin (Ch. 12). Then, I presented population measurements of BBH spin in GWTC-4.0, where we found that the distribution of effective inspiral spins is asymmetric about zero with skew toward positive values, that $\sim 90\%$ of BHs have spin magnitudes $\chi < 0.57$ (with a model-dependent hint of a high-spin feature), and that a wide range of spin-orbit misalignment exists in the population. The GWTC-4.0 findings are consistent with an observed BBH population stemming from multiple formation channels (Ch. 13). I ended by describing the importance of model checking in GW population analyses, and presented posterior predictive checks that are better-equipped to identify model misspecification than those most commonly-used in the field (Ch. 14).

15.2 Future plans

A decade after the dawn of GW observation, GW astronomy remains a young and expanding field. There are many exciting avenues of continued research that have sprung from my dissertation work. In this final section, I describe some immediate and longer-term aspirational endeavors.

Closed-form expressions for spins in the merger

As discussed throughout this dissertation, spin magnitudes and tilts are often poorly constrained, and canonical effective spin parameters [e.g., 77, 193,

200, 203, 289] are derived from early inspiral dynamics [e.g., 149, 313, 315] (see Ch. 2). Therefore, existing effective spin parameters are almost certainly not the best descriptors of the late-inspiral, merger, and ringdown data more readily visible in high-mass BBHs [150, 223, 313].

Chapter 4 laid the groundwork for unearthing more-precise effective spin parameters to characterize BBHs of different masses. I plan to build off of our previous work and investigate the possibility of *deriving closed-form expressions* analogous to $\chi_{\rm p}$ that best capture the first-order effects of precessing spin in the merger and ringdown. To gain intuition, I will utilize data-driven methods like those presented in Ch. 4, as well as a Fisher Information approach [e.g., 421].

More and faster time-domain inference

The time-domain inference framework presented in Ch. 6 opens the door for a rich and continued understanding of the link between BBH source morphology and parameter measurability, especially for high-mass systems. In this vein, I have already started to analyze GW231123 [21], the most massive BBH thus far detected (at the time of writing), with my TD-inference approach. Preliminary results show that the imprint of spin precession on GW231123 is morphologically distinct from GW190521, as shown in the Pocket Material at the end of this dissertation. In the immediate future, I plan to dive deeper into GW231123's morphology and compare my findings with ringdown-only measurements [232]. When more high-mass BBHs are inevitably detected in the LVK's next observing run (O5) and beyond, I will similarly infer their spin-induced morphology.

Additionally, TD inference has extraordinary potential to help understand the supposed degeneracy between precessing and eccentric high-mass BBHs [99, 100, 248, 579]. However, waveform models with eccentricity remain more computationally intensive than those that assume quasi-circular orbits, and waveforms with both eccentricity *and* precession remain relatively nascent. The Gohberg-Semencul representation of the inverse of the Toeplitz covariance matrix [580] can be used to accelerate matrix-vector products and thus further speed up time-domain likelihood calculations. Following Ref. [581], I plan to incorporate the Gohberg-Semencul inverse Toeplitz representation into the `tdinf` code [161] and, building on Ch. 9, continue to hunt for degeneracies

between precession and eccentricity.

Ringdown-based measurements for precessing BBHs

Ringdown analysis is a natural next direction given my existing skills and interests: ringdown data are also analyzed in the time domain, but with the aforementioned flexible yet simplistic damped-sinusoid (QNM) model (Sec. 2.5) rather than templated inspiral-merger-ringdown (IMR) waveforms like NRSur7dq4 [16] or IMRPhenomXPHM [245]. Even state-of-the-art IMR models are potentially error-prone at the transition from inspiral to merger. Extracting high-fidelity measurements of the progenitor masses and spins from the ringdown *without relying on potentially systematic-ridden IMR models* will either strengthen confidence in existing results or bring tensions between models to light. For example, different IMR waveform models yield drastically different estimates for some of GW231123's spin parameters [21, 247, 248]; ringdown-only measurements could help discriminate between IMR waveform models, eliminate systematics, and narrow in on the truth.

Currently, there exist "ringdown-surrogate" models which use simple machine learning techniques like Gaussian processes to calculate QNMs from BBH source parameters. At present, these only exist for aligned-spin systems [223, 400, 582, 583]. Incorporating our most up-to-date understanding of precessional dynamics is crucial, since our observed high-mass BBHs tend towards misaligned spin. Luckily, ringdown surrogates that include spin precession are under development. To use such ringdown surrogate models to infer source properties, we must be able to map *from* QNMs *back to* progenitor masses and spins. I am interested in "inverting" precessing ringdown surrogates to re-analyze the BBHs in the most recent LVK gravitational-wave transient catalog.

Population-level tests of Hawking's Area Law

Next, I propose to extend the Area Law analysis from GW250114 (Ch. 11) to the population level. Just as we can measure the astrophysical distribution of BBH spins, so too can we constrain the change in area from component to remnant BHs at the population level with hierarchical inference (Ch. 3). GW250114 already yielded $\gtrsim 4\sigma$ agreement with the Area Law; adding ensemble data from other mergers may further increase this significance.

Inferring the astrophysical spin distribution in the presence of quiet glitches

Disconcertingly, Ref. [105] (for which I am a co-author) showed that even SNR < 5 glitches (which are essentially undetectable) can drastically bias BBH spin measurements. Given uncertainty in the rate of quiet glitches [e.g., 584, 585], it remains difficult to quantify their impact at the population level on real data. Through an injection-recovery study of a simulated glitch-contaminated population, I hope to investigate a worst-case scenario for if and how single-event bias incurred by quiet glitches can propagate to the population.

High-mass BBHs in next-generation GW detectors

The past decade of GW observation has been exceptionally fruitful, and yet we've only just scratched the surface of all that GWs can reveal about the cosmos. The next generation of GW detectors are currently under development, with Cosmic Explorer (CE) [27], the Einstein Telescope (ET) [28], and the Laser Interferometer Space Antenna (LISA) [29] all targeted to be built in the 2030s. This set of detectors will enable the observation of more, louder BBH signals over a wider range of frequencies [e.g., 586–588]. The ground-based CE and ET are sensitive over a similar frequency band as LVK detectors, i.e., $\sim 10 - 1000\,\mathrm{Hz}$. However, while LVK observatories can only observe redshifts $z \leq \mathcal{O}(1)$, CE/ET will look out to $z \sim 100$ [443]—meaning they will see all stellar mass BBH mergers throughout the full history of the universe.

In contrast, the space-based LISA will probe a much lower GW frequency range, i.e., milli- to deci-Hertz [589], from different compact binary populations like double white dwarf and supermassive BH mergers. The large number and long duration of observable astrophysical GW sources in LISA leads to unavoidable signal overlap in time-frequency, necessitating a *global fit* rather than separate inference of individual source properties [e.g., 590, 591]. Current solutions for extracting BBH signals from LISA data and classifying their source properties, e.g. with ringdown measurements [213, 592, 593], remain in their infancy.

While all three upcoming missions are expected to detect GW signals from the $\mathcal{O}(100\,M_\odot)$ BBHs discussed in this dissertation, the way these signals will manifest differs per detector and each will come with unique challenges that must be diagnosed, *especially for inferring spins* [e.g., 594–597]. In the future,

I hope to help set constraints on the measurability of source parameters for high-mass, highly-precessing BBH in the CE/ET detector network and LISA's global fit. Involvement with CE, ET, and LISA will open the door for continued research avenues of the utmost astrophysical importance throughout my full academic tenure.

Bibliography

[1] Ilya Mandel and Alison Farmer. Merging stellar-mass binary black holes. *Phys. Rept.*, 955:1–24, 2022.

[2] Michela Mapelli. Astrophysics of stellar black holes. *Proc. Int. Sch. Phys. Fermi*, 200:87–121, 2020.

[3] LIGO Scientific Collaboration and Virgo Collaboration. Detection science summary. Technical report, LIGO, California Institute of Technology, 2016.

[4] J. Aasi, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, et al. Advanced LIGO. *Classical and Quantum Gravity*, 32(7):074001, Apr 2015.

[5] F. Acernese, M. Agathos, K. Agatsuma, D. Aisa, N. Allemandou, et al. Advanced Virgo: a second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity*, 32(2):024001, Jan 2015.

[6] T. Akutsu et al. Overview of KAGRA: Detector design and construction history. *PTEP*, 2021(5):05A101, 2021.

[7] The LIGO Scientific, Virgo, and KAGRA Collaborations. GWTC-4.0: An Introduction to Version 4.0 of the Gravitational-Wave Transient Catalog. 8 2025.

[8] A. G. Abac et al. GWTC-4.0: Methods for Identifying and Characterizing Gravitational-wave Transients. 8 2025.

[9] A. G. Abac et al. GWTC-4.0: Updating the Gravitational-Wave Transient Catalog with Observations from the First Part of the Fourth LIGO-Virgo-KAGRA Observing Run. 8 2025.

[10] B. P. Abbott, R. Abbott, T. D. Abbott, et al. Gw170817: Observation of gravitational waves from a binary neutron star inspiral. *Phys. Rev. Lett.*, 119:161101, Oct 2017.

[11] R. Abbott et al. Observation of Gravitational Waves from Two Neutron Star–Black Hole Coalescences. *Astrophys. J. Lett.*, 915(1):L5, 2021.

[12] A. G. Abac et al. Observation of Gravitational Waves from the Coalescence of a 2.5–4.5 $M_\odot$ Compact Object and a Neutron Star. *Astrophys. J. Lett.*, 970(2):L34, 2024.

[13] Eric Poisson and Clifford M. Will. Gravitational waves from inspiraling compact binaries: Parameter estimation using second postNewtonian wave forms. *Phys. Rev. D*, 52:848–855, 1995.

[14] Alessandra Buonanno, Bala Iyer, Evan Ochsner, Yi Pan, and B. S. Sathyaprakash. Comparison of post-Newtonian templates for compact binary inspiral signals in gravitational-wave detectors. *Phys. Rev. D*, 80:084043, 2009.

[15] Saul A. Teukolsky. Perturbations of a Rotating Black Hole. I. Fundamental Equations for Gravitational, Electromagnetic, and Neutrino-Field Perturbations. *The Astrophysical Journal*, 185:635–648, October 1973.

[16] Vijay Varma, Scott E. Field, Mark A. Scheel, Jonathan Blackman, Davide Gerosa, Leo C. Stein, Lawrence E. Kidder, and Harald P. Pfeiffer. Surrogate models for precessing binary black hole simulations with unequal masses. *Phys. Rev. Research.*, 1:033015, 2019.

[17] The LIGO Scientific and Virgo Collaborations. GW150914: First results from the search for binary black hole coalescence with Advanced LIGO. *Phys. Rev. D*, 93(12):122003, 2016.

[18] A. G. Abac et al. GW250114: Testing Hawking's Area Law and the Kerr Nature of Black Holes. *Phys. Rev. Lett.*, 135(11):111403, 2025.

[19] A. G. Abac et al. Black Hole Spectroscopy and Tests of General Relativity with GW250114. *Phys. Rev. Lett.*, 136(4):041403, 2026.

[20] R. Abbott et al. GW190521: A Binary Black Hole Merger with a Total Mass of $150M_{\odot}$. *Phys. Rev. Lett.*, 125(10):101102, 2020.

[21] A. G. Abac et al. GW231123: A Binary Black Hole Merger with Total Mass 190–265 $M_{\odot}$. *Astrophys. J. Lett.*, 993(1):L25, 2025.

[22] A. G. Abac et al. GW241011 and GW241110: Exploring Binary Formation and Fundamental Physics with Asymmetric, High-spin Black Hole Coalescences. *Astrophys. J. Lett.*, 993(1):L21, 2025.

[23] A. G. Abac et al. GWTC-4.0: Population Properties of Merging Compact Binaries. 8 2025.

[24] A. G. Abac et al. GWTC-4.0: Tests of General Relativity. I. Overview and General Tests. 3 2026.

[25] The LIGO Scientific, Virgo, and KAGRA Collaborations. GWTC-4.0: Constraints on the Cosmic Expansion Rate and Modified Gravitational-wave Propagation. 9 2025.

[26] 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. *Phys. Rev. X*, 9(3):031040, 2019.

[27] David Reitze et al. Cosmic Explorer: The U.S. Contribution to Gravitational-Wave Astronomy beyond LIGO. *Bull. Am. Astron. Soc.*, 51(7):035, 2019.

[28] Marica Branchesi et al. Science with the Einstein Telescope: a comparison of different designs. *JCAP*, 07:068, 2023.

[29] Pau Amaro-Seoane, Heather Audley, Stanislav Babak, John Baker, Enrico Barausse, Peter Bender, Emanuele Berti, Pierre Binetruy, Michael Born, Daniele Bortoluzzi, Jordan Camp, Chiara Caprini, Vitor Cardoso, Monica Colpi, John Conklin, Neil Cornish, Curt Cutler, Karsten Danzmann, Rita Dolesi, Luigi Ferraioli, Valerio Ferroni, Ewan Fitzsimons, Jonathan Gair, Lluis Gesa Bote, Domenico Giardini, Ferran Gibert, Catia Grimani, Hubert Halloin, Gerhard Heinzel, Thomas Hertog, Martin Hewitson, Kelly Holley-Bockelmann, Daniel Hollington, Mauro Hueller, Henri Inchauspe, Philippe Jetzer, Nikos Karnesis, Christian Killow, Antoine Klein, Bill Klipstein, Natalia Korsakova, Shane L Larson, Jeffrey Livas, Ivan Lloro, Nary Man, Davor Mance, Joseph Martino, Ignacio Mateos, Kirk McKenzie, Sean T McWilliams, Cole Miller, Guido Mueller, Germano Nardini, Gijs Nelemans, Miquel Nofrarias, Antoine Petiteau, Paolo Pivato, Eric Plagnol, Ed Porter, Jens Reiche, David Robertson, Norna Robertson, Elena Rossi, Giuliana Russano, Bernard Schutz, Alberto Sesana, David Shoemaker, Jacob Slutsky, Carlos F. Sopuerta, Tim Sumner, Nicola Tamanini, Ira Thorpe, Michael Troebs, Michele Vallisneri, Alberto Vecchio, Daniele Vetrugno, Stefano Vitale, Marta Volonteri, Gudrun Wanner, Harry Ward, Peter Wass, William Weber, John Ziemer, and Peter Zweifel. Laser Interferometer Space Antenna. February 2017.

[30] Christopher S. Reynolds. Observational Constraints on Black Hole Spin. *Ann. Rev. Astron. Astrophys.*, 59:117–154, 2021.

[31] Jim Fuller, Anthony L. Piro, and Adam S. Jermyn. Slowing the spins of stellar cores. *Mon. Not. Roy. Astron. Soc.*, 485(3):3661–3680, 2019.

[32] Ben Farr, Daniel E. Holz, and Will M. Farr. Using Spin to Understand the Formation of LIGO and Virgo's Black Holes. *Astrophys. J. Lett.*, 854(1):L9, 2018.

[33] Hui Tong, Maya Fishbach, and Eric Thrane. Spinning Spectral Sirens: Robust Cosmological Measurement Using Mass–Spin Correlations in the Binary Black Hole Population. *Astrophys. J.*, 985(2):220, 2025.

[34] E. Capote et al. Advanced LIGO detector performance in the fourth observing run. *Phys. Rev. D*, 111(6):062002, 2025.

[35] Ilya Mandel and Floor S. Broekgaarden. Rates of compact object coalescences. *Living Rev. Rel.*, 25(1):1, 2022.

[36] Gijs Nelemans and C. A. Tout. Reconstructing the evolution of white dwarf binaries: Further evidence for an alternative algorithm for the outcome of the common-envelope phase in close binaries. *Mon. Not. Roy. Astron. Soc.*, 356:753–764, 2005.

[37] L. A. C. van Son, S. E. de Mink, M. Renzo, S. Justham, E. Zapartas, K. Breivik, T. Callister, W. M. Farr, and C. Conroy. No Peaks without Valleys: The Stable Mass Transfer Channel for Gravitational-wave Sources in Light of the Neutron Star–Black Hole Mass Gap. *Astrophys. J.*, 940(2):184, 2022.

[38] S. E. de Mink and I. Mandel. The chemically homogeneous evolutionary channel for binary black hole mergers: rates and properties of gravitational-wave events detectable by advanced LIGO. *Mon. Not. Roy. Astron. Soc.*, 460(4):3545–3553, 2016.

[39] Carl L. Rodriguez, Meagan Morscher, Bharath Pattabiraman, Sourav Chatterjee, Carl-Johan Haster, and Frederic A. Rasio. Binary Black Hole Mergers from Globular Clusters: Implications for Advanced LIGO. *Phys. Rev. Lett.*, 115(5):051101, 2015. [Erratum: Phys.Rev.Lett. 116, 029901 (2016)].

[40] Giacomo Fragione and Joseph Silk. Repeated mergers and ejection of black holes within nuclear star clusters. *Mon. Not. Roy. Astron. Soc.*, 498(4):4591–4604, 2020.

[41] Sambaran Banerjee. Stellar-mass black holes in young massive and open stellar clusters – V. comparisons with LIGO-Virgo merger rate densities. *Mon. Not. Roy. Astron. Soc.*, 503(3):3371–3385, 2021.

[42] B. Mckernan et al. Constraining Stellar-mass Black Hole Mergers in AGN Disks Detectable with LIGO. *Astrophys. J.*, 866(1):66, 2018.

[43] Zoheyr Doctor, Daniel Wysocki, Richard O'Shaughnessy, Daniel E. Holz, and Ben Farr. Black Hole Coagulation: Modeling Hierarchical Mergers in Black Hole Populations. 11 2019.

[44] Ken K. Y. Ng et al. Measuring properties of primordial black hole mergers at cosmological distances: Effect of higher order modes in gravitational waves. *Phys. Rev. D*, 107(2):024041, 2023.

[45] Michael Zevin, Simone S. Bavera, Christopher P. L. Berry, Vicky Kalogera, Tassos Fragos, Pablo Marchant, Carl L. Rodriguez, Fabio Antonini, Daniel E. Holz, and Chris Pankow. One Channel to Rule Them All? Constraining the Origins of Binary Black Holes Using Multiple Formation Pathways. *Astrophys. J.*, 910(2):152, 2021.

[46] Roy P. Kerr. Gravitational field of a spinning mass as an example of algebraically special metrics. *Phys. Rev. Lett.*, 11:237–238, Sep 1963.

[47] Matthew Mould and Davide Gerosa. Gravitational-wave population inference at past time infinity. *Phys. Rev. D*, 105(2):024076, 2022.

[48] Vijay Varma, Maximiliano Isi, Sylvia Biscoveanu, Will M. Farr, and Salvatore Vitale. Measuring binary black hole orbital-plane spin orientations. *Phys. Rev. D*, 105(2):024045, 2022.

[49] Sylvia Biscoveanu. Probing Spin–Orbit Resonances with the Binary Black Hole Population. *Astrophys. J.*, 996(1):95, 2026.

[50] Daria Gangardt, Davide Gerosa, Michael Kesden, Viola De Renzis, and Nathan Steinle. Constraining black-hole binary spin precession and nutation with sequential prior conditioning. *Phys. Rev. D*, 106(2):024019, 2022. [Erratum: Phys.Rev.D 107, 109901 (2023)].

[51] Jim Fuller and Linhao Ma. Most Black Holes are Born Very Slowly Rotating, 2019.

[52] Vishal Baibhav and Vicky Kalogera. Revising the Spin and Kick Connection in Isolated Binary Black Holes. 12 2024.

[53] P. Hut. Tidal evolution in close binary systems. *Astronomy and Astrophysics*, 99:126–140, June 1981.

[54] W. Packet. On the spin-up of the mass accreting component in a close binary system. *Astronomy and Astrophysics*, 102(1):17–19, September 1981.

[55] C. A. Tout and J. E. Pringle. Spin-down of rapidly rotating, convective stars. *Monthly Notices of the Royal Astronomical Society*, 256:269–276, May 1992.

[56] Ilya Mandel and Selma E. de Mink. Merging binary black holes formed through chemically homogeneous evolution in short-period stellar binaries. *Mon. Not. Roy. Astron. Soc.*, 458(3):2634–2647, 2016.

[57] Ying Qin, Pablo Marchant, Tassos Fragos, Georges Meynet, and Vicky Kalogera. On the Origin of Black-Hole Spin in High-Mass X-ray Binaries. *Astrophys. J. Lett.*, 870(2):L18, 2019.

[58] 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. *Astron. Astrophys.*, 616:A28, 2018.

[59] Simone S. Bavera et al. The impact of mass-transfer physics on the observable properties of field binary black hole populations. *Astron. Astrophys.*, 647:A153, 2021.

[60] Linhao Ma and Jim Fuller. Tidal Spin-up of Black Hole Progenitor Stars. *Astrophys. J.*, 952(1):53, 2023. [Erratum: Astrophys.J. 965, (2024)].

[61] Carl L. Rodriguez, Michael Zevin, Pau Amaro-Seoane, Sourav Chatterjee, Kyle Kremer, Frederic A. Rasio, and Claire S. Ye. Black holes: The next generation—repeated mergers in dense star clusters and their gravitational-wave properties. *Phys. Rev. D*, 100(4):043027, 2019.

[62] Rachel C. Zhang, Giacomo Fragione, Chase Kimball, and Vicky Kalogera. On the Likely Dynamical Origin of GW191109 and Binary Black Hole Mergers with Negative Effective Spin. *Astrophys. J.*, 954(1):23, 2023.

[63] Chase Kimball et al. Evidence for Hierarchical Black Hole Mergers in the Second LIGO–Virgo Gravitational Wave Catalog. *Astrophys. J. Lett.*, 915(2):L35, 2021.

[64] Ethan Payne, Kyle Kremer, and Michael Zevin. Spin Doctors: How to Diagnose a Hierarchical Merger Origin. *Astrophys. J. Lett.*, 966(1):L16, 2024.

[65] Davide Gerosa and Maya Fishbach. Hierarchical mergers of stellar-mass black holes and their gravitational-wave signatures. *Nature Astron.*, 5(8):749–760, 2021.

[66] Maya Fishbach, Daniel E. Holz, and Ben Farr. Are LIGO's Black Holes Made From Smaller Black Holes? *Astrophys. J. Lett.*, 840(2):L24, 2017.

[67] Davide Gerosa and Emanuele Berti. Are merging black holes born from stellar collapse or previous mergers? *Phys. Rev. D*, 95(12):124046, 2017.

[68] Matthew Mould, Davide Gerosa, and Stephen R. Taylor. Deep learning and Bayesian inference of gravitational-wave populations: Hierarchical black-hole mergers. *Phys. Rev. D*, 106(10):103013, 2022.

[69] Parthapratim Mahapatra, Anuradha Gupta, Marc Favata, K. G. Arun, and B. S. Sathyaprakash. Remnant Black Hole Kicks and Implications for Hierarchical Mergers. *Astrophys. J. Lett.*, 918(2):L31, 2021.

[70] Parthapratim Mahapatra, Debatri Chattopadhyay, Anuradha Gupta, Marc Favata, B. S. Sathyaprakash, and K. G. Arun. Predictions of a simple parametric model of hierarchical black hole mergers. *Phys. Rev. D*, 111(2):023013, 2025.

[71] Parthapratim Mahapatra, Debatri Chattopadhyay, Anuradha Gupta, Fabio Antonini, Marc Favata, B. S. Sathyaprakash, and K. G. Arun. Reconstructing the Genealogy of LIGO-Virgo Black Holes. *Astrophys. J.*, 975(1):117, 2024.

[72] Cailin Plunkett, Thomas Callister, Michael Zevin, and Salvatore Vitale. Signatures of a subpopulation of hierarchical mergers in the GWTC-4 gravitational-wave dataset. 1 2026.

[73] Carlos O. Lousto, Manuela Campanelli, Yosef Zlochower, and Hiroyuki Nakano. Remnant Masses, Spins and Recoils from the Merger of Generic Black-Hole Binaries. *Class. Quant. Grav.*, 27:114006, 2010.

[74] P. Marchant, N. Langer, P. Podsiadlowski, T. M. Tauris, and T. J. Moriya. A new route towards merging massive black holes. *Astronomy & Astrophysics*, 588:A50, 2016.

[75] Jeff Riley, Ilya Mandel, Pablo Marchant, Ellen Butler, Kaila Nathaniel, Coenraad Neijssel, Spencer Shortt, and Alejandro Vigna-Gomez. Chemically homogeneous evolution: a rapid population synthesis approach. *Mon. Not. Roy. Astron. Soc.*, 505(1):663–676, 2021.

[76] H. C. Spruit. Differential rotation and magnetic fields in stellar interiors. *Astron. Astrophys.*, 349:189, 1999.

[77] Patricia Schmidt, Mark Hannam, Sascha Husa, and P. Ajith. Tracking the precession of compact binaries from their gravitational-wave signal. *Phys. Rev. D*, 84:024046, 2011.

[78] Davide Gerosa, Michael Kesden, Ulrich Sperhake, Emanuele Berti, and Richard O'Shaughnessy. Multi-timescale analysis of phase transitions in precessing black-hole binaries. *Phys. Rev. D*, 92:064016, 2015.

[79] Simon Stevenson and Teagan A. Clarke. Constraints on the contributions to the observed binary black hole population from individual evolutionary pathways in isolated binary evolution. *Mon. Not. Roy. Astron. Soc.*, 517(3):4034–4053, 2022.

[80] Matias Zaldarriaga, Doron Kushnir, and Juna A. Kollmeier. The expected spins of gravitational wave sources with isolated field binary progenitors. *Mon. Not. Roy. Astron. Soc.*, 473(3):4174–4178, 2018.

[81] Michael Zevin and Simone S. Bavera. Suspicious Siblings: The Distribution of Mass and Spin across Component Black Holes in Isolated Binary Evolution. *Astrophys. J.*, 933(1):86, 2022.

[82] Nathan Steinle and Michael Kesden. Signatures of spin precession and nutation in isolated black-hole binaries. *Phys. Rev. D*, 106(6):063028, 2022.

[83] Simone S. Bavera, Tassos Fragos, Ying Qin, Emmanouil Zapartas, Coenraad J. Neijssel, Ilya Mandel, Aldo Batta, Sebastian M. Gaebel, Chase Kimball, and Simon Stevenson. The origin of spin in binary black holes: Predicting the distributions of the main observables of Advanced LIGO. *Astron. Astrophys.*, 635:A97, 2020.

[84] Vassiliki Kalogera. Spin orbit misalignment in close binaries with two compact objects. *Astrophys. J.*, 541:319–328, 2000.

[85] Davide Gerosa, Emanuele Berti, Richard O'Shaughnessy, Krzysztof Belczynski, Michael Kesden, Daniel Wysocki, and Wojciech Gladysz. Spin orientations of merging black holes formed from the evolution of stellar binaries. *Phys. Rev. D*, 98(8):084036, 2018.

[86] Nathan Steinle and Michael Kesden. Pathways for producing binary black holes with large misaligned spins in the isolated formation channel. *Phys. Rev. D*, 103(6):063032, 2021.

[87] Daniel Wysocki, Davide Gerosa, Richard O'Shaughnessy, Krzysztof Belczynski, Wojciech Gladysz, Emanuele Berti, Michael Kesden, and Daniel E. Holz. Explaining LIGO's observations via isolated binary evolution with natal kicks. *Phys. Rev. D*, 97(4):043014, 2018.

[88] Simon Stevenson. Biases in Estimates of Black Hole Kicks from the Spin Distribution of Binary Black Holes. *Astrophys. J. Lett.*, 926(2):L32, 2022.

[89] Thomas A. Callister, Will M. Farr, and Mathieu Renzo. State of the Field: Binary Black Hole Natal Kicks and Prospects for Isolated Field Formation after GWTC-2. *Astrophys. J.*, 920(2):157, 2021.

[90] Thomas M. Tauris. Tossing Black Hole Spin Axes. *Astrophys. J.*, 938:66, 2022.

[91] Carl L. Rodriguez, Sourav Chatterjee, and Frederic A. Rasio. Binary Black Hole Mergers from Globular Clusters: Masses, Merger Rates, and the Impact of Stellar Evolution. *Phys. Rev. D*, 93(8):084029, 2016.

[92] Carl L. Rodriguez, Pau Amaro-Seoane, Sourav Chatterjee, and Frederic A. Rasio. Post-Newtonian Dynamics in Dense Star Clusters: Highly-Eccentric, Highly-Spinning, and Repeated Binary Black Hole Mergers. *Phys. Rev. Lett.*, 120(15):151101, 2018.

[93] Will M. Farr, Simon Stevenson, M. Coleman Miller, Ilya Mandel, Ben Farr, and Alberto Vecchio. Distinguishing Spin-Aligned and Isotropic Black Hole Populations With Gravitational Waves. *Nature*, 548:426, 2017.

[94] Kyle Kremer, Claire S. Ye, Nicholas Z. Rui, Newlin C. Weatherford, Sourav Chatterjee, Giacomo Fragione, Carl L. Rodriguez, Mario Spera, and Frederic A. Rasio. Modeling Dense Star Clusters in the Milky Way and Beyond with the CMC Cluster Catalog. *Astrophys. J. Suppl.*, 247(2):48, 2020.

[95] Carl L. Rodriguez, Newlin C. Weatherford, Scott C. Coughlin, Pau Amaro-Seoane, Katelyn Breivik, Sourav Chatterjee, Giacomo Fragione, Fulya Kıroğlu, Kyle Kremer, Nicholas Z. Rui, Claire S. Ye, Michael Zevin, and Frederic A. Rasio. Modeling Dense Star Clusters in the Milky

Way and beyond with the Cluster Monte Carlo Code. *The Astrophysical Journal Supplement*, 258(2):22, February 2022.

[96] Yi-Han Wang, Barry McKernan, Saavik Ford, Rosalba Perna, Nathan W. C. Leigh, and Mordecai-Mark Mac Low. Symmetry Breaking in Dynamical Encounters in the Disks of Active Galactic Nuclei. *Astrophys. J. Lett.*, 923(2):L23, 2021.

[97] 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. *Mon. Not. Roy. Astron. Soc.*, 514(3):3886–3893, 2022.

[98] Salvatore Vitale, Ryan Lynch, John Veitch, Vivien Raymond, and Riccardo Sturani. Measuring the spin of black holes in binary systems using gravitational waves. *Phys. Rev. Lett.*, 112(25):251101, 2014.

[99] Isobel M. Romero-Shaw, Davide Gerosa, and Nicholas Loutrel. Eccentricity or spin precession? Distinguishing subdominant effects in gravitational-wave data. *Mon. Not. Roy. Astron. Soc.*, 519(4):5352–5357, 2023.

[100] Divyajyoti, Sumit Kumar, Snehal Tibrewal, Isobel M. Romero-Shaw, and Chandra Kant Mishra. Blind spots and biases: The dangers of ignoring eccentricity in gravitational-wave signals from binary black holes. *Phys. Rev. D*, 109(4):043037, 2024.

[101] Simona J. Miller, Maximiliano Isi, Katerina Chatziioannou, Vijay Varma, and Ilya Mandel. GW190521: Tracing imprints of spin-precession on the most massive black hole binary. *Phys. Rev. D*, 109(2):024024, 2024.

[102] Simona J. Miller, Maximiliano Isi, Katerina Chatziioannou, Vijay Varma, and Sophie Hourihane. Measuring spin precession from massive black hole binaries with gravitational waves: Insights from time-domain signal morphology. *Phys. Rev. D*, 112(10):104046, 2025.

[103] Sylvia Biscoveanu, Maximiliano Isi, Vijay Varma, and Salvatore Vitale. Measuring the spins of heavy binary black holes. *Phys. Rev. D*, 104(10):103018, 2021.

[104] Yumeng Xu and Eleanor Hamilton. Measurability of precession and eccentricity for heavy binary-black-hole mergers. *Phys. Rev. D*, 107(10):103049, 2023.

[105] Rhiannon Udall, Sophie Bini, Katerina Chatziioannou, Derek Davis, Sophie Hourihane, Yannick Lecoeuche, Jess McIver, and Simona Miller. Inferring the spins of merging black holes in the presence of data-quality issues. *Phys. Rev. D*, 113(4):042005, 2026.

[106] Sophie Hourihane and Katerina Chatziioannou. Glitches far from transient gravitational-wave events do not bias inference. 6 2025.

[107] R. Abbott et al. Properties and Astrophysical Implications of the 150 $M_\odot$ Binary Black Hole Merger GW190521. *Astrophys. J. Lett.*, 900(1):L13, 2020.

[108] Shrinivas F. Kulkarni, Steve McMillan, and Piet Hut. Stellar black holes in globular clusters. *Nature*, 364:421, 1993.

[109] Isobel M. Romero-Shaw, Kyle Kremer, Paul D. Lasky, Eric Thrane, and Johan Samsing. Gravitational waves as a probe of globular cluster formation and evolution. *Mon. Not. Roy. Astron. Soc.*, 506(2):2362–2372, 2021.

[110] A. A. Trani, A. Tanikawa, M. S. Fujii, N. W. C. Leigh, and J. Kumamoto. Spin misalignment of black hole binaries from young star clusters: implications for the origin of gravitational waves events. *Mon. Not. Roy. Astron. Soc.*, 504(1):910–919, 2021.

[111] Sambaran Banerjee, Holger Baumgardt, and Pavel Kroupa. Stellar-mass black holes in star clusters: implications for gravitational wave radiation. *Mon. Not. Roy. Astron. Soc.*, 402:371, 2010.

[112] Sambaran Banerjee, Aleksandra Olejak, and Krzysztof Belczynski. Symmetry Breaking in Merging Binary Black Holes from Young Massive Clusters and Isolated Binaries. *Astrophys. J.*, 953(1):80, 2023.

[113] Fulya Kıroğlu, James C. Lombardi, Kyle Kremer, Hans D. Vanderzyden, and Frederic A. Rasio. Spin–Orbit Alignment in Merging Binary Black Holes Following Collisions with Massive Stars. *Astrophys. J. Lett.*, 983(1):L9, 2025.

[114] Jordan Barber, Debatri Chattopadhyay, and Fabio Antonini. Black hole binary mergers in dense star clusters: the importance of primordial binaries. *Mon. Not. Roy. Astron. Soc.*, 527(3):7363–7381, 2023.

[115] Jordan Barber and Fabio Antonini. Formation and evolution of binary black holes in N-body simulations of star clusters with up to two million stars. *Mon. Not. Roy. Astron. Soc.*, 538(2):639–658, 2025.

[116] Ugo N. Di Carlo, Nicola Giacobbo, Michela Mapelli, Mario Pasquato, Mario Spera, Long Wang, and Francesco Haardt. Merging black holes in young star clusters. *Mon. Not. Roy. Astron. Soc.*, 487(2):2947–2960, 2019.

[117] Maya Fishbach and Giacomo Fragione. Globular cluster formation histories, masses, and radii inferred from gravitational waves. *Mon. Not. Roy. Astron. Soc.*, 522(4):5546–5557, 2023.

[118] Giacomo Fragione and Bence Kocsis. Black hole mergers from an evolving population of globular clusters. *Phys. Rev. Lett.*, 121(16):161103, 2018.

[119] Stefano Torniamenti, Michela Mapelli, Carole Périgois, Manuel Arca Sedda, Maria Celeste Artale, Marco Dall'Amico, and Maria Paola Vaccaro. Hierarchical binary black hole mergers in globular clusters: Mass function and evolution with redshift. *Astron. Astrophys.*, 688:A148, 2024.

[120] B. McKernan, K. E. S. Ford, R. O'Shaughnessy, and D. Wysocki. Monte Carlo simulations of black hole mergers in AGN discs: Low $\chi_{\rm eff}$ mergers and predictions for LIGO. *Mon. Not. Roy. Astron. Soc.*, 494(1):1203–1216, 2020.

[121] B. McKernan, K. E. S. Ford, W. Lyra, and H. B. Perets. Intermediate mass black holes in AGN disks: I. Production & Growth. *Mon. Not. Roy. Astron. Soc.*, 425:460, 2012.

[122] B. McKernan and K. E. S. Ford. Constraining the LVK AGN channel with black hole spins. *Mon. Not. Roy. Astron. Soc.*, 531(3):3479–3485, 2024.

[123] Y. Yang, I. Bartos, Z. Haiman, B. Kocsis, Z. Marka, N. C. Stone, and S. Marka. AGN Disks Harden the Mass Distribution of Stellar-mass Binary Black Hole Mergers. *Astrophys. J.*, 876(2):122, 2019.

[124] Kyle Kremer, Mario Spera, Devin Becker, Sourav Chatterjee, Ugo N. Di Carlo, Giacomo Fragione, Carl L. Rodriguez, Claire S. Ye, and Frederic A. Rasio. Populating the upper black hole mass gap through stellar collisions in young star clusters. *Astrophys. J.*, 903(1):45, 2020.

[125] Simon Stevenson, Matthew Sampson, Jade Powell, Alejandro Vigna-Gómez, Coenraad J. Neijssel, Dorottya Szécsi, and Ilya Mandel. The Impact of Pair-instability Mass Loss on the Binary Black Hole Mass Distribution. *The Astrophysical Journal*, 882(2):121, September 2019.

[126] D. D. Hendriks, L. A. C. van Son, M. Renzo, R. G. Izzard, and R. Farmer. Pulsational pair-instability supernovae in gravitational-wave and electromagnetic transients. *Mon. Not. Roy. Astron. Soc.*, 526(3):4130–4147, 2023.

[127] Pablo Marchant, Mathieu Renzo, Robert Farmer, Kaliroe M. W. Pappas, Ronald E. Taam, Selma de Mink, and Vassiliki Kalogera. Pulsational pair-instability supernovae in very close binaries. 10 2018.

[128] Fabio Antonini, Thomas Callister, Fani Dosopoulou, Isobel M. Romero-Shaw, and Debatri Chattopadhyay. Inferring the pair-instability mass gap from gravitational wave data. *Phys. Rev. D*, 112(6):063040, 2025.

[129] Hui Tong et al. Evidence of the pair instability gap in the distribution of black hole masses. *Nature*, 4 2026.

[130] C. L. Fryer, S. E. Woosley, and A. Heger. Pair instability supernovae, gravity waves, and gamma-ray transients. *Astrophys. J.*, 550:372–382, 2001.

[131] 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. *The Astrophysical Journal*, 887(1):53, December 2019.

[132] Pablo Marchant and Takashi Moriya. The impact of stellar rotation on the black hole mass-gap from pair-instability supernovae. *Astron. Astrophys.*, 640:L18, 2020.

[133] M. Renzo, R. Farmer, S. Justham, Y. Götberg, S. E. de Mink, E. Zapartas, P. Marchant, and N. Smith. Predictions for the hydrogen-free ejecta of pulsational pair-instability supernovae. *Astron. Astrophys.*, 640:A56, 2020.

[134] Mario Spera and Michela Mapelli. Very massive stars, pair-instability supernovae and intermediate-mass black holes with the SEVN code. *Mon. Not. Roy. Astron. Soc.*, 470(4):4739–4749, 2017.

[135] Aryanna Schiebelbein-Zwack and Maya Fishbach. The Mass Density of Merging Binary Black Holes over Cosmic Time. *Astrophys. J.*, 970(2):128, 2024.

[136] Amanda M. Farah, Thomas A. Callister, Jose María Ezquiaga, Michael Zevin, and Daniel E. Holz. No Need to Know: Toward Astrophysics-free Gravitational-wave Cosmology. *Astrophys. J.*, 978(2):153, 2025.

[137] Grégoire Pierra, Alberto Colombo, and Simone Mastrogiovanni. Non-Parametric Reconstruction of the Hubble Parameter from the Fourth Gravitational Wave Transient Catalog and DESI Baryonic Acoustic Oscillations. 11 2025.

[138] Hiromichi Tagawa, Masayuki Umemura, and Naoteru Gouda. Mergers of accreting stellar-mass black holes. *Mon. Not. Roy. Astron. Soc.*, 462(4):3812–3822, 2016.

[139] Kohei Inayoshi, Eli Visbal, and Zoltán Haiman. The Assembly of the First Massive Black Holes. *Ann. Rev. Astron. Astrophys.*, 58:27–97, 2020.

[140] Alex Sicilia, Andrea Lapi, Lumen Boco, Mario Spera, Ugo N. Di Carlo, Michela Mapelli, Francesco Shankar, David M. Alexander, Alessandro

Bressan, and Luigi Danese. The Black Hole Mass Function Across Cosmic Times. I. Stellar Black Holes and Light Seed Distribution. *Astrophys. J.*, 924(2):56, 2022.

[141] A. Buikema et al. Sensitivity and performance of the advanced ligo detectors in the third observing run. *Phys. Rev. D*, 102:062003, Sep 2020.

[142] J. Veitch and A. Vecchio. Bayesian approach to the follow-up of candidate gravitational wave signals. *Phys. Rev. D*, 78:022001, Jul 2008.

[143] Salvatore Vitale, Ryan Lynch, Vivien Raymond, Riccardo Sturani, John Veitch, and Philp Graff. Parameter estimation for heavy binary-black holes with networks of second-generation gravitational-wave detectors. *Phys. Rev. D*, 95(6):064053, 2017.

[144] B. P. Abbott et al. Characterization of transient noise in Advanced LIGO relevant to gravitational wave signal GW150914. *Class. Quant. Grav.*, 33(13):134001, 2016.

[145] Derek Davis et al. LIGO detector characterization in the second and third observing runs. *Class. Quant. Grav.*, 38(13):135014, 2021.

[146] 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. *Class. Quant. Grav.*, 39(24):245013, 2022.

[147] Vijay Varma, Scott E. Field, Mark A. Scheel, Jonathan Blackman, Lawrence E. Kidder, and Harald P. Pfeiffer. Surrogate model of hybridized numerical relativity binary black hole waveforms. *Phys. Rev. D*, 99(6):064045, 2019.

[148] Luc Blanchet. Gravitational Radiation from Post-Newtonian Sources and Inspiralling Compact Binaries. *Living Rev. Rel.*, 17:2, 2014.

[149] Alessandro Nagar et al. Time-domain effective-one-body gravitational waveforms for coalescing compact binaries with nonprecessing spins, tides and self-spin effects. *Phys. Rev. D*, 98(10):104052, 2018.

[150] Karen Kang, Simona J. Miller, Katerina Chatziioannou, and Deborah Ferguson. Mapping parameter correlations in spinning binary black hole mergers. *Phys. Rev. D*, 112(6):064020, 2025.

[151] 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. *Phys. Rev. D*, 87(4):044038, 2013.

[152] Scott A. Hughes, Anuj Apte, Gaurav Khanna, and Halston Lim. Learning about black hole binaries from their ringdown spectra. *Phys. Rev. Lett.*, 123(16):161101, 2019.

[153] Eliot Finch and Christopher J. Moore. Modeling the ringdown from precessing black hole binaries. *Phys. Rev. D*, 103(8):084048, 2021.

[154] Ioannis Kamaretsos, Mark Hannam, and B. Sathyaprakash. Is black-hole ringdown a memory of its progenitor? *Phys. Rev. Lett.*, 109:141102, 2012.

[155] Eleanor Hamilton, Lionel London, Jonathan E. Thompson, Edward Fauchon-Jones, Mark Hannam, Chinmay Kalaghatgi, Sebastian Khan, Francesco Pannarale, and Alex Vano-Vinuales. Model of gravitational waves from precessing black-hole binaries through merger and ringdown. *Phys. Rev. D*, 104(12):124027, 2021.

[156] Eleanor Hamilton, Lionel London, and Mark Hannam. Ringdown frequencies in black holes formed from precessing black-hole binaries. *Phys. Rev. D*, 107(10):104035, 2023.

[157] Hengrui Zhu et al. Black hole spectroscopy for precessing binary black hole coalescences. *Phys. Rev. D*, 111(6):064052, 2025.

[158] Rhiannon Udall, Sophie Hourihane, Simona Miller, Derek Davis, Katerina Chatziioannou, Max Isi, and Howard Deshong. Antialigned spin of GW191109: Glitch mitigation and its implications. *Phys. Rev. D*, 111(2):024046, 2025.

[159] Ethan Payne, Sophie Hourihane, Jacob Golomb, Rhiannon Udall, Richard Udall, Derek Davis, and Katerina Chatziioannou. Curious case of GW200129: Interplay between spin-precession inference and data-quality issues. *Phys. Rev. D*, 106(10):104017, 2022.

[160] Simona J. Miller, Zoe Ko, Tom Callister, and Katerina Chatziioannou. Gravitational waves carry information beyond effective spin parameters but it is hard to extract. *Phys. Rev. D*, 109(10):104036, 2024.

[161] Simona J. Miller, Sophie Hourihane, Maximiliano Isi, Rhiannon Udall, and Katerina Chatziioannou. `tdinf`: time domain parameter estimation for gravitational-wave signals, August 2025.

[162] Simona Miller, Thomas A. Callister, and Will Farr. The Low Effective Spin of Binary Black Holes and Implications for Individual Gravitational-Wave Events. *Astrophys. J.*, 895(2):128, 2020.

[163] M. Coleman Miller and Jon M. Miller. The Masses and Spins of Neutron Stars and Stellar-Mass Black Holes. *Phys. Rept.*, 548:1–34, 2014.

[164] Thomas A. Callister, Simona J. Miller, Katerina Chatziioannou, and Will M. Farr. No Evidence that the Majority of Black Holes in Binaries Have Zero Spin. *Astrophys. J. Lett.*, 937(1):L13, 2022.

[165] Shanika Galaudage, Colm Talbot, Tushar Nagar, Deepnika Jain, Eric Thrane, and Ilya Mandel. Building Better Spin Models for Merging Binary Black Holes: Evidence for Nonspinning and Rapidly Spinning Nearly Aligned Subpopulations. *Astrophys. J. Lett.*, 921(1):L15, November 2021.

[166] Simona J. Miller, Sophia Winney, Katerina Chatziioannou, and Patrick M. Meyers. Posterior Predictive Checks for Gravitational-wave Populations: Limitations and Improvements. 4 2026.

[167] Simona J. Miller. Using gravitational-wave signals to model the distribution of spin across the binary black hole population, 2020. Includes bibliographical references (pages 2, 103–110).

[168] Michele Maggiore. *Gravitational Waves. Vol. 1: Theory and Experiments*. Oxford University Press, 2007.

[169] Michele Maggiore. *Gravitational Waves: Volume 2: Astrophysics and Cosmology*. Oxford University Press, 2018.

[170] J. B. Hartle. *Gravity: An Introduction To Einstein's General Relativity*. Pearson, 2014.

[171] Thomas A. Moore. *A General Relativity Workbook*. University Science Books, 2013.

[172] Sean M Carroll. *Spacetime and Geometry: An Introduction to General Relativity*. Addison-Wesley, San Francisco, 2004.

[173] Kip S Thorne and Roger D Blandford. *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*. Princeton University Press, Princeton, NJ, 2017.

[174] Charles W. Misner, Kip S. Thorne, and John Archibald Wheeler. *Gravitation*. W. H. Freeman, San Francisco, 1973.

[175] Edwin Hubble. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173, 1929.

[176] Adam G Riess, Alexei V Filippenko, Peter Challis, Alejandro Clocchiatti, Alan Diercks, Peter M Garnavich, Ronald L Gilliland, Craig J Hogan, Saurabh Jha, Robert P Kirshner, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal*, 116(3):1009, 1998.

[177] Saul Perlmutter, G Aldering, G Goldhaber, RA Knop, P Nugent, PG Castro, S Deustua, S Fabbro, A Goobar, DE Groom, et al. Measurements of ω and λ from 42 high-redshift supernovae. *The Astrophysical Journal*, 517(2):565, 1999.

[178] P. A. R. Ade et al. Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.*, 594:A13, 2016.

[179] Wendy L. Freedman. Measurements of the Hubble Constant: Tensions in Perspective. *Astrophys. J.*, 919(1):16, 2021.

[180] K. S. Thorne. THE THEORY OF GRAVITATIONAL RADIATION: AN INTRODUCTORY REVIEW. In *Les Houches Summer School on Gravitational Radiation*, 1982.

[181] Maximiliano Isi. Parametrizing gravitational-wave polarizations. *Class. Quant. Grav.*, 40(20):203001, 2023.

[182] Alexandre Le Tiec and Jérôme Novak. Theory of Gravitational Waves. In Gerard Auger and Eric Plagnol, editors, *An Overview of Gravitational Waves: Theory, Sources and Detection*, pages 1–41. 2017.

[183] David W. Hogg. Distance measures in cosmology. 5 1999.

[184] Vijay Varma and Parameswaran Ajith. Effects of nonquadrupole modes in the detection and parameter estimation of black hole binaries with nonprecessing spins. *Phys. Rev. D*, 96(12):124024, 2017.

[185] Juan Calderón Bustillo, Sascha Husa, Alicia M. Sintes, and Michael Pürrer. Impact of gravitational radiation higher order modes on single aligned-spin gravitational wave searches for binary black holes. *Phys. Rev. D*, 93(8):084019, 2016.

[186] Cameron Mills and Stephen Fairhurst. Measuring gravitational-wave higher-order multipoles. *Phys. Rev. D*, 103(2):024042, 2021.

[187] Curt Cutler and Eanna E. Flanagan. Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral wave form? *Phys. Rev. D*, 49:2658–2697, 1994.

[188] K. G. Arun, Alessandra Buonanno, Guillaume Faye, and Evan Ochsner. Higher-order spin effects in the amplitude and phase of gravitational waveforms emitted by inspiraling compact binaries: Ready-to-use gravitational waveforms. *Phys. Rev. D*, 79:104023, May 2009.

[189] Poulami Dutta Roy, Sayantani Datta, and K. G. Arun. Tests of general relativity at the fourth post-Newtonian order. *Phys. Rev. D*, 110(6):064002, 2024.

[190] Gihyuk Cho, Rafael A. Porto, and Zixin Yang. Gravitational radiation from inspiralling compact objects: Spin effects to the fourth post-Newtonian order. *Phys. Rev. D*, 106(10):L101501, 2022.

[191] Ken K. Y. Ng, Salvatore Vitale, Aaron Zimmerman, Katerina Chatziioannou, Davide Gerosa, and Carl-Johan Haster. Gravitational-wave astrophysics with effective-spin measurements: asymmetries and selection biases. *Phys. Rev. D*, 98(8):083007, 2018.

[192] Michael Pürrer, Mark Hannam, P. Ajith, and Sascha Husa. Testing the validity of the single-spin approximation in inspiral-merger-ringdown waveforms. *Phys. Rev. D*, 88:064007, 2013.

[193] Etienne Racine. Analysis of spin precession in binary black hole systems including quadrupole-monopole interaction. *Phys. Rev. D*, 78:044021, 2008.

[194] Thibault Damour. Coalescence of two spinning black holes: an effective one-body approach. *Phys. Rev. D*, 64:124013, 2001.

[195] Héctor Estellés, Sascha Husa, Marta Colleoni, David Keitel, Maite Mateu-Lucena, Cecilio García-Quirós, Antoni Ramos-Buades, and Angela Borchers. Time-domain phenomenological model of gravitational-wave subdominant harmonics for quasicircular nonprecessing binary black hole coalescences. *Phys. Rev. D*, 105(8):084039, 2022.

[196] Manuela Campanelli, C. O. Lousto, and Y. Zlochower. Spinning-black-hole binaries: The orbital hang up. *Phys. Rev. D*, 74:041501, 2006.

[197] James Healy and Carlos O. Lousto. Hangup effect in unequal mass binary black hole mergers and further studies of their gravitational radiation and remnant properties. *Phys. Rev. D*, 97(8):084002, 2018.

[198] Curt Cutler et al. The Last three minutes: issues in gravitational wave measurements of coalescing compact binaries. *Phys. Rev. Lett.*, 70:2984–2987, 1993.

[199] Theocharis A. Apostolatos, Curt Cutler, Gerald J. Sussman, and Kip S. Thorne. Spin-induced orbital precession and its modulation of the gravitational waveforms from merging binaries. *Phys. Rev. D*, 49:6274–6297, Jun 1994.

[200] Patricia Schmidt, Mark Hannam, and Sascha Husa. Towards models of gravitational waveforms from generic binaries: A simple approximate mapping between precessing and non-precessing inspiral signals. *Phys. Rev. D*, 86:104063, 2012.

[201] Stephen Fairhurst, Rhys Green, Charlie Hoy, Mark Hannam, and Alistair Muir. Two-harmonic approximation for gravitational waveforms from precessing binaries. *Phys. Rev. D*, 102(2):024055, 2020.

[202] Héctor Estellés, Antoni Ramos-Buades, Sascha Husa, Cecilio García-Quirós, Marta Colleoni, Leïla Haegel, and Rafel Jaume. Phenomenological time domain model for dominant quadrupole gravitational wave signal of coalescing binary black holes. *Phys. Rev. D*, 103(12):124060, 2021.

[203] Patricia Schmidt, Frank Ohme, and Mark Hannam. Towards models of gravitational waveforms from generic binaries II: Modelling precession effects with a single effective precession parameter. *Phys. Rev. D*, 91(2):024043, 2015.

[204] Lucy M. Thomas, Patricia Schmidt, and Geraint Pratten. New effective precession spin for modeling multimodal gravitational waveforms in the strong-field regime. *Phys. Rev. D*, 103(8):083022, 2021.

[205] Davide Gerosa, Matthew Mould, Daria Gangardt, Patricia Schmidt, Geraint Pratten, and Lucy M. Thomas. A generalized precession parameter $\chi_{\rm p}$ to interpret gravitational-wave data. *Phys. Rev. D*, 103(6):064067, 2021.

[206] Michael Boyle et al. The SXS Collaboration catalog of binary black hole simulations. *Class. Quant. Grav.*, 36(19):195006, 2019.

[207] C. V. Vishveshwara. Stability of the schwarzschild metric. *Phys. Rev. D*, 1:2870–2879, May 1970.

[208] William H. Press. Long Wave Trains of Gravitational Waves from a Vibrating Black Hole. *The Astrophysical Journal Letters*, 170:L105, December 1971.

[209] Subrahmanyan Chandrasekhar and S. Detweiler. The quasi-normal modes of the schwarzschild black hole. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 344(1639):441–452, 08 1975.

[210] Kostas D. Kokkotas and Bernd G. Schmidt. Quasinormal modes of stars and black holes. *Living Rev. Rel.*, 2:2, 1999.

[211] S. Detweiler. Black holes and gravitational waves. III - The resonant frequencies of rotating holes. *The Astrophysical Journal*, 239:292–295, July 1980.

[212] Olaf Dreyer, Bernard J. Kelly, Badri Krishnan, Lee Samuel Finn, David Garrison, and Ramon Lopez-Aleman. Black hole spectroscopy: Testing

general relativity through gravitational wave observations. *Class. Quant. Grav.*, 21:787–804, 2004.

[213] Emanuele Berti, Vitor Cardoso, and Clifford M. Will. On gravitational-wave spectroscopy of massive black holes with the space interferometer LISA. *Phys. Rev. D*, 73:064030, 2006.

[214] Harrison Siegel, Maximiliano Isi, and Will M. Farr. Ringdown of GW190521: Hints of multiple quasinormal modes with a precessional interpretation. *Phys. Rev. D*, 108(6):064008, 2023.

[215] Collin D. Capano, Miriam Cabero, Julian Westerweck, Jahed Abedi, Shilpa Kastha, Alexander H. Nitz, Yi-Fan Wang, Alex B. Nielsen, and Badri Krishnan. Multimode Quasinormal Spectrum from a Perturbed Black Hole. *Phys. Rev. Lett.*, 131(22):221402, 2023.

[216] Teagan A. Clarke et al. Toward a self-consistent framework for measuring black hole ringdowns. *Phys. Rev. D*, 109(12):124030, 2024.

[217] Emanuele Berti et al. Black hole spectroscopy: from theory to experiment. 5 2025.

[218] Gregorio Carullo. Black hole spectroscopy: status report. *General Relativity and Gravitation*, 57(5):76, 2025.

[219] Erin Coleman and Eliot Finch. High-overtone ringdown fits: start time, no-hair tests, and correlations. 12 2025.

[220] Xisco Jiménez Forteza and Pierre Mourier. High-overtone fits to numerical relativity ringdowns: Beyond the dismissed n=8 special tone. *Phys. Rev. D*, 104(12):124072, 2021.

[221] Naritaka Oshita and Vitor Cardoso. Reconstruction of ringdown with excitation factors. *Phys. Rev. D*, 111(10):104043, 2025.

[222] Maximiliano Isi, Matthew Giesler, Will M. Farr, Mark A. Scheel, and Saul A. Teukolsky. Testing the no-hair theorem with GW150914. *Phys. Rev. Lett.*, 123(11):111102, 2019.

[223] Mark Ho-Yeuk Cheung, Emanuele Berti, Vishal Baibhav, and Roberto Cotesta. Extracting linear and nonlinear quasinormal modes from black hole merger simulations. *Phys. Rev. D*, 109(4):044069, 2024. [Erratum: Phys.Rev.D 110, 049902 (2024)].

[224] Costantino Pacilio, Swetha Bhagwat, Francesco Nobili, and Davide Gerosa. Flexible mapping of ringdown amplitudes for nonprecessing binary black holes. *Phys. Rev. D*, 110(10):103037, 2024.

[225] Lorena Magaña Zertuche et al. High-precision ringdown surrogate model for nonprecessing binary black holes. *Phys. Rev. D*, 112(2):024077, 2025.

[226] E. W. Leaver. An analytic representation for the quasi-normal modes of Kerr black holes. *Proc. Roy. Soc. Lond. A*, 402:285–298, 1985.

[227] Xisco Jiménez-Forteza, David Keitel, Sascha Husa, Mark Hannam, Sebastian Khan, and Michael Pürrer. Hierarchical data-driven approach to fitting numerical relativity data for nonprecessing binary black holes with an application to final spin and radiated energy. *Phys. Rev. D*, 95(6):064024, 2017.

[228] Lorena Magaña Zertuche, Leda Gao, Eliot Finch, and Gregory B. Cook. Multimode ringdown modelling with `qnmfits` and `KerrRingdown`. 2 2025.

[229] Vijay Varma, Davide Gerosa, Leo C. Stein, François Hébert, and Hao Zhang. High-accuracy mass, spin, and recoil predictions of generic black-hole merger remnants. *Phys. Rev. Lett.*, 122(1):011101, 2019.

[230] Matteo Boschini et al. Extending black-hole remnant surrogate models to extreme mass ratios. *Phys. Rev. D*, 108(8):084015, 2023.

[231] Guido Da Re et al. Modeling the BMS transformation induced by a binary black hole merger. *Phys. Rev. D*, 111(12):124019, 2025.

[232] Harrison Siegel, Nicole M. Khusid, Maximiliano Isi, and Will M. Farr. GW231123 ringdown: interpretation as multimodal Kerr signal. 11 2025.

[233] Leo C. Stein. qnm: A Python package for calculating Kerr quasinormal modes, separation constants, and spherical-spheroidal mixing coefficients. *J. Open Source Softw.*, 4(42):1683, 2019.

[234] Alan H. Guth and Mohammad Hossein Namjoo. Bayesian questions with frequentist answers. 8 2023.

[235] Thomas Bayes and Richard Price. An essay towards solving a problem in the doctrine of chances. communicated by mr. price, in a letter to john canton, m. a. and f. r. s. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

[236] Matthew Mould, Rodrigo Tenorio, and Davide Gerosa. Gravitational-wave astronomy requires population-informed parameter estimation. 4 2026.

[237] Eric Thrane and Colm Talbot. An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models. *Publications of the Astronomical Society of Australia*, 36, March 2019.

[238] T. Callister. A Thesaurus for Common Priors in Gravitational-Wave Astronomy. 4 2021.

[239] Masaki Iwaya, Kazuya Kobayashi, Soichiro Morisaki, Kenta Hotokezaka, and Tomoya Kinugawa. Analytic joint priors of effective spin parameters and inference of the spin distribution of binary black holes. *Phys. Rev. D*, 111(10):103046, 2025.

[240] Kazuya Kobayashi, Masaki Iwaya, Soichiro Morisaki, Kenta Hotokezaka, and Tomoya Kinugawa. Impact of Spin Priors on the Population Inference of Merging Binary Black Holes. *J. Phys. Conf. Ser.*, 3177(1):012078, 2026.

[241] Vaishak Prasad and B. S. Sathyaprakash. Are black hole spins truly near-zero? 2 2026.

[242] Maximiliano Isi, Will M. Farr, and Katerina Chatziioannou. Comparing Bayes factors and hierarchical inference for testing general relativity with gravitational waves. *Phys. Rev. D*, 106(2):024048, 2022.

[243] Gregory Ashton, Moritz Hübner, Paul D. Lasky, Colm Talbot, Kendall Ackley, Sylvia Biscoveanu, Qi Chu, Atul Divakarla, Paul J. Easter, Boris Goncharov, Francisco Hernandez Vivanco, Jan Harms, Marcus E. Lower, Grant D. Meadors, Denyz Melchor, Ethan Payne, Matthew D. Pitkin, Jade Powell, Nikhil Sarin, Rory J. E. Smith, and Eric Thrane. Bilby: A user-friendly bayesian inference library for gravitational-wave astronomy. *The Astrophysical Journal Supplement Series*, 241(2):27, April 2019.

[244] Héctor Estellés, Alessandra Buonanno, Raffi Enficiaud, Cheng Foo, and Lorenzo Pompili. Adding equatorial-asymmetric effects for spin-precessing binaries into the SEOBNRv5PHM waveform model. 6 2025.

[245] Geraint Pratten et al. Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes. *Phys. Rev. D*, 103(10):104056, 2021.

[246] Jonathan E. Thompson, Eleanor Hamilton, Lionel London, Shrobana Ghosh, Panagiota Kolitsidou, Charlie Hoy, and Mark Hannam. PhenomXO4a: a phenomenological gravitational-wave model for precessing black-hole binaries with higher multipoles and asymmetries. *Phys. Rev. D*, 109(6):063012, 2024.

[247] Sophie Bini, Krzysztof Król, Katerina Chatziioannou, and Maximiliano Isi. Impact of waveform systematics and Gaussian noise on the interpretation of GW231123. *Phys. Rev. D*, 113(8):083036, 2026.

[248] Aasim Jan, Sophia Nicolella, Deirdre Shoemaker, and Richard O'Shaughnessy. Measuring eccentricity and addressing waveform systematics in GW231123. *Phys. Rev. D*, 113(8):084052, 2026.

[249] Arnab Dhani, Sebastian H. Völkel, Alessandra Buonanno, Hector Estelles, Jonathan Gair, Harald P. Pfeiffer, Lorenzo Pompili, and Alexandre Toubiana. Systematic Biases in Estimating the Properties of Black Holes Due to Inaccurate Gravitational-Wave Models. *Phys. Rev. X*, 15(3):031036, 2025.

[250] Valerio De Luca, Loris Del Grosso, Francesco Iacovelli, Andrea Maselli, and Emanuele Berti. Systematic biases from ignoring environmental tidal effects in gravitational wave observations. *Phys. Rev. D*, 111(12):124046, 2025.

[251] Veome Kapil, Luca Reali, Roberto Cotesta, and Emanuele Berti. Systematic bias from waveform modeling for binary black hole populations in next-generation gravitational wave detectors. *Phys. Rev. D*, 109(10):104043, 2024.

[252] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306, March 2013.

[253] Joshua S. Speagle. DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *Mon. Not. Roy. Astron. Soc.*, 493(3):3132–3158, April 2020.

[254] 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. *Mon. Not. Roy. Astron. Soc.*, 499(3):3295–3319, 2020.

[255] Gregory Ashton and Colm Talbot. B ilby-MCMC: an MCMC sampler for gravitational-wave inference. *Mon. Not. Roy. Astron. Soc.*, 507(2):2037–2051, 2021.

[256] Javier Roulet, Jonathan Mushkin, Digvijay Wadekar, Tejaswi Venumadhav, Barak Zackay, and Matias Zaldarriaga. Fast marginalization algorithm for optimizing gravitational wave detection, parameter estimation, and sky localization. *Phys. Rev. D*, 110(4):044010, 2024.

[257] Jacob Lange, Richard O'Shaughnessy, and Monica Rizzo. Rapid and accurate parameter inference for coalescing, precessing compact binaries. 5 2018.

[258] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[259] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. 2019.

[260] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[261] Maximilian Dax, Stephen R. Green, Jonathan Gair, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-Time Gravitational Wave Science with Neural Posterior Estimation. *Phys. Rev. Lett.*, 127(24):241103, 2021.

[262] Javier Roulet, Marco Crisostomi, Lucy M. Thomas, and Katerina Chatziioannou. labrador: A domain-optimized machine-learning tool for gravitational wave inference. 4 2026.

[263] T. A. Callister. Observed Gravitational-Wave Populations. 10 2024.

[264] Reed Essick and Maya Fishbach. Ensuring Consistency between Noise and Detection in Hierarchical Bayesian Inference. *Astrophys. J.*, 962(2):169, 2024.

[265] Reed Essick et al. Compact binary coalescence sensitivity estimates with injection campaigns during the LIGO-Virgo-KAGRA Collaborations' fourth observing run. *Phys. Rev. D*, 112(10):102001, 2025.

[266] Bruce Edelman, Ben Farr, and Zoheyr Doctor. Cover Your Basis: Comprehensive Data-driven Characterization of the Binary Black Hole Population. *Astrophys. J.*, 946(1):16, 2023.

[267] Jacob Golomb and Colm Talbot. Searching for structure in the binary black hole spin distribution. *Phys. Rev. D*, 108(10):103009, 2023.

[268] Thomas A. Callister and Will M. Farr. Parameter-Free Tour of the Binary Black Hole Population. *Phys. Rev. X*, 14(2):021005, 2024.

[269] Anarya Ray, Ignacio Magaña Hernandez, Katelyn Breivik, and Jolien Creighton. Searching for Binary Black Hole Subpopulations in Gravitational-wave Data Using Binned Gaussian Processes. *Astrophys. J.*, 991(1):17, 2025.

[270] Jack Heinzel, Matthew Mould, Sofía Álvarez-López, and Salvatore Vitale. High resolution nonparametric inference of gravitational-wave populations in multiple dimensions. *Phys. Rev. D*, 111(6):063043, 2025.

[271] Sofia Alvarez-Lopez, Jack Heinzel, Matthew Mould, and Salvatore Vitale. Nowhere left to hide: revealing realistic gravitational-wave populations in high dimensions and high resolution with PixelPop. 6 2025.

[272] Jaxen Godfrey, Bruce Edelman, and Ben Farr. Cosmic Cousins: Identification of a Subpopulation of Binary Black Holes Consistent with Isolated Binary Evolution. 4 2023.

[273] Asad Hussain, Maximiliano Isi, and Aaron Zimmerman. Hints of Spin-magnitude Correlations and a Rapidly Spinning Subpopulation of Binary Black Holes. *Astrophys. J.*, 996(1):71, 2026.

[274] Michele Mancarella and Davide Gerosa. Sampling the full hierarchical population posterior distribution in gravitational-wave astronomy. *Phys. Rev. D*, 111(10):103012, 2025.

[275] Reed Essick, Amanda Farah, Shanika Galaudage, Colm Talbot, Maya Fishbach, Eric Thrane, and Daniel E. Holz. Probing Extremal Gravitational-wave Events with Coarse-grained Likelihoods. *Astrophys. J.*, 926(1):34, 2022.

[276] Storm Colloms, Christopher P. L. Berry, John Veitch, and Michael Zevin. Exploring the Evolution of Gravitational-wave Emitters with Efficient Emulation: Constraining the Origins of Binary Black Holes Using Normalizing Flows. *Astrophys. J.*, 988(2):189, 2025.

[277] Ana Lorenzo-Medina and Thomas Dent. A physically modelled selection function for compact binary mergers in the LIGO-Virgo O3 run and beyond. *Class. Quant. Grav.*, 42(4):045008, 2025.

[278] Thomas A. Callister, Reed Essick, and Daniel E. Holz. Neural network emulator of the Advanced LIGO and Advanced Virgo selection function. *Phys. Rev. D*, 110(12):123041, 2024.

[279] Jack Heinzel and Salvatore Vitale. When (not) to trust Monte Carlo approximations for hierarchical Bayesian inference. 9 2025.

[280] Will M. Farr. Accuracy Requirements for Empirically-Measured Selection Functions. *Research Notes of the AAS*, 3(5):66, 2019.

[281] Colm Talbot and Jacob Golomb. Growing pains: understanding the impact of likelihood uncertainty on hierarchical Bayesian inference for gravitational-wave astronomy. *Mon. Not. Roy. Astron. Soc.*, 526(3):3495–3503, 2023.

[282] R. Abbott et al. Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog. *Astrophys. J. Lett.*, 913(1):L7, 2021.

[283] R. Abbott et al. Population of Merging Compact Binaries Inferred Using Gravitational Waves through GWTC-3. *Phys. Rev. X*, 13(1):011048, 2023.

[284] Benjamin P Abbott et al. A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals. *Class. Quant. Grav.*, 37(5):055002, 2020.

[285] 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. *Phys. Rev. D*, 109(2):022001, 2024.

[286] R. Abbott et al. GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. *Phys. Rev. X*, 11:021053, 2021.

[287] R. Abbott et al. GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run. *Phys. Rev. X*, 13:041039, 2023.

[288] Antoni Ramos-Buades, Alessandra Buonanno, Héctor Estellés, Mohammed Khalil, Deyan P. Mihaylov, Serguei Ossokine, Lorenzo Pompili, and Mahlet Shiferaw. Next generation of accurate and efficient multipolar precessing-spin effective-one-body waveforms for binary black holes. *Phys. Rev. D*, 108(12):124037, 2023.

[289] P. Ajith et al. Inspiral-merger-ringdown waveforms for black-hole binaries with non-precessing spins. *Phys. Rev. Lett.*, 106:241101, 2011.

[290] Isobel M. Romero-Shaw, Paul D. Lasky, Eric Thrane, and Juan Calderon Bustillo. GW190521: orbital eccentricity and signatures of dynamical formation in a binary black hole merger signal. *Astrophys. J. Lett.*, 903(1):L5, 2020.

[291] V. Gayathri, J. Healy, J. Lange, B. O'Brien, M. Szczepanczyk, Imre Bartos, M. Campanelli, S. Klimenko, C. O. Lousto, and R. O'Shaughnessy. Eccentricity estimate for black hole mergers with numerical relativity simulations. *Nature Astron.*, 6(3):344–349, 2022.

[292] Alexander H. Nitz and Collin D. Capano. GW190521 may be an intermediate mass ratio inspiral. *Astrophys. J. Lett.*, 907(1):L9, 2021.

[293] Rossella Gamba, Matteo Breschi, Gregorio Carullo, Simone Albanesi, Piero Rettegno, Sebastiano Bernuzzi, and Alessandro Nagar. GW190521 as a dynamical capture of two nonspinning black holes. *Nature Astron.*, 7(1):11–17, 2023.

[294] Mark Hannam et al. General-relativistic precession in a black-hole binary. *Nature*, 610(7933):652–655, 2022.

[295] Ronaldas Macas, Andrew Lundgren, and Gregory Ashton. Revisiting the evidence for precession in GW200129 with machine learning noise mitigation. *Phys. Rev. D*, 109(6):062006, 2024.

[296] B. P. Abbott et al. GW151226: Observation of Gravitational Waves from a 22-Solar-Mass Binary Black Hole Coalescence. *Phys. Rev. Lett.*, 116(24):241103, 2016.

[297] R. Abbott et al. GW190412: Observation of a binary-black-hole coalescence with asymmetric masses. *Phys. Rev. D*, 102(4):043015, August 2020.

[298] Javier Roulet, Seth Olsen, Jonathan Mushkin, Tousif Islam, Tejaswi Venumadhav, Barak Zackay, and Matias Zaldarriaga. Removing degeneracy and multimodality in gravitational wave source parameters. *Phys. Rev. D*, 106(12):123015, 2022.

[299] Deborah Ferguson. Optimizing the placement of numerical relativity simulations using a mismatch predicting neural network. *Phys. Rev. D*, 107(2):024034, 2023.

[300] Emily Baird, Stephen Fairhurst, Mark Hannam, and Patricia Murphy. Degeneracy between mass and spin in black-hole-binary waveforms. *Phys. Rev. D*, 87(2):024035, 2013.

[301] Stephen Fairhurst, Charlie Hoy, Rhys Green, Cameron Mills, and Samantha A. Usman. Simple parameter estimation using observable features of gravitational-wave signals. *Phys. Rev. D*, 108(8):082006, 2023.

[302] Benjamin J. Owen. Search templates for gravitational waves from inspiraling binaries: Choice of template spacing. *Phys. Rev. D*, 53:6749–6761, 1996.

[303] Mark Hannam, Duncan A. Brown, Stephen Fairhurst, Chris L. Fryer, and Ian W. Harry. When can gravitational-wave observations distinguish between black holes and neutron stars? *Astrophys. J. Lett.*, 766:L14, 2013.

[304] Juan Calderón Bustillo, Pablo Laguna, and Deirdre Shoemaker. Detectability of gravitational waves from binary black holes: Impact of precession and higher modes. *Phys. Rev. D*, 95(10):104038, 2017.

[305] S. D. Mohanty. Hierarchical search strategy for the detection of gravitational waves from coalescing binaries: Extension to postNewtonian wave forms. *Phys. Rev. D*, 57:630–658, 1998.

[306] Michael Pürrer, Mark Hannam, and Frank Ohme. Can we measure individual black-hole spins from gravitational-wave observations? *Phys. Rev. D*, 93(8):084042, 2016.

[307] Katerina Chatziioannou, Geoffrey Lovelace, Michael Boyle, Matthew Giesler, Daniel A. Hemberger, Reza Katebi, Lawrence E. Kidder, Harald P. Pfeiffer, Mark A. Scheel, and Béla Szilágyi. Measuring the properties of nearly extremal black holes with gravitational waves. *Phys. Rev. D*, 98(4):044028, 2018.

[308] Chinmay Kalaghatgi and Mark Hannam. Investigating the effect of in-plane spin directions for precessing binary black hole systems. *Phys. Rev. D*, 103(2):024024, 2021.

[309] M. V. van der Sluys, C. Röver, A. Stroeer, V. Raymond, I. Mandel, N. Christensen, V. Kalogera, R. Meyer, and A. Vecchio. Gravitational-Wave Astronomy with Inspiral Signals of Spinning Compact-Object Binaries. *Astrophys. J. Lett.*, 688:L61, 2008.

[310] Marc van der Sluys, Vivien Raymond, Ilya Mandel, Christian Rover, Nelson Christensen, Vicky Kalogera, Renate Meyer, and Alberto Vecchio. Parameter estimation of spinning binary inspirals using Markov-chain Monte Carlo. *Class. Quant. Grav.*, 25:184011, 2008.

[311] J. Veitch et al. Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library. *Phys. Rev. D*, 91(4):042003, 2015.

[312] Lawrence E. Kidder, Clifford M. Will, and Alan G. Wiseman. Spin effects in the inspiral of coalescing compact binaries. *Phys. Rev. D*, 47:R4183–R4187, May 1993.

[313] Geraint Pratten, Sascha Husa, Cecilio Garcia-Quiros, Marta Colleoni, Antoni Ramos-Buades, Hector Estelles, and Rafel Jaume. Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for nonprecessing quasicircular black holes. *Phys. Rev. D*, 102(6):064001, 2020.

[314] Feroz H. Shaik, Jacob Lange, Scott E. Field, Richard O'Shaughnessy, Vijay Varma, Lawrence E. Kidder, Harald P. Pfeiffer, and Daniel Wysocki. Impact of subdominant modes on the interpretation of gravitational-wave signals from heavy binary black hole systems. *Phys. Rev. D*, 101(12):124054, 2020.

[315] Serguei Ossokine, Tim Dietrich, Evan Foley, Reza Katebi, and Geoffrey Lovelace. Assessing the Energetics of Spinning Binary Black Hole Systems. *Phys. Rev. D*, 98(10):104057, 2018.

[316] Mark Hannam, Patricia Schmidt, Alejandro Bohé, Leïla Haegel, Sascha Husa, Frank Ohme, Geraint Pratten, and Michael Pürrer. Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms. *Phys. Rev. Lett.*, 113(15):151101, 2014.

[317] Gregory Ashton, Sarah Thiele, Yannick Lecoeuche, Jess McIver, and Laura K. Nuttall. Parameterised population models of transient non-Gaussian noise in the LIGO gravitational-wave detectors. *Class. Quant. Grav.*, 39(17):175004, 2022.

[318] Salvatore Vitale, Sylvia Biscoveanu, and Colm Talbot. Spin it as you like: The (lack of a) measurement of the spin tilt distribution with LIGO-Virgo-KAGRA binary black holes. *Astron. Astrophys.*, 668:L2, 2022.

[319] LIGO Scientific Collaboration. LIGO Algorithm Library - LALSuite. Free software (GPL), 2019.

[320] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[321] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[322] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[323] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[324] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[325] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[326] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam

Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017.

[327] B. P. Abbott, R. Abbott, T. D. Abbott, et al. Binary Black Hole Population Properties Inferred from the First and Second Observing Runs of Advanced LIGO and Advanced Virgo. *Astrophys. J. Lett.*, 882(2):L24, Sep 2019.

[328] Salvatore Vitale, Ryan Lynch, Riccardo Sturani, and Philip Graff. Use of gravitational waves to probe the formation channels of compact binaries. *Class. Quant. Grav.*, 34(3):03LT01, 2017.

[329] Carl L. Rodriguez, Michael Zevin, Chris Pankow, Vasilliki Kalogera, and Frederic A. Rasio. Illuminating Black Hole Binary Formation Channels with Spins in Advanced LIGO. *Astrophys. J. Lett.*, 832(1):L2, 2016.

[330] Sebastian Khan, Frank Ohme, Katerina Chatziioannou, and Mark Hannam. Including higher order multipoles in gravitational-wave models for precessing binary black holes. *Phys. Rev. D*, 101(2):024056, 2020.

[331] Serguei Ossokine et al. Multipolar Effective-One-Body Waveforms for Precessing Binary Black Holes: Construction and Validation. *Phys. Rev. D*, 102(4):044055, 2020.

[332] Javier Roulet, Horng Sheng Chia, Seth Olsen, Liang Dai, Tejaswi Venumadhav, Barak Zackay, and Matias Zaldarriaga. Distribution of effective spins and masses of binary black holes from the LIGO and Virgo O1–O3a observing runs. *Phys. Rev. D*, 104(8):083010, 2021.

[333] Katerina Chatziioannou and Will M. Farr. Inferring the maximum and minimum mass of merging neutron stars with gravitational waves. *Phys. Rev. D*, 102(6):064063, 2020.

[334] Hui Tong, Shanika Galaudage, and Eric Thrane. Population properties of spinning black holes using the gravitational-wave transient catalog 3. *Phys. Rev. D*, 106(10):103019, 2022.

[335] Kashif Alvi. Energy and angular momentum flow into a black hole in a binary. *Phys. Rev. D*, 64:104020, 2001.

[336] Eric Poisson. Absorption of mass and angular momentum by a black hole: Time-domain formalisms for gravitational perturbations, and the small-hole / slow-motion approximation. *Phys. Rev. D*, 70:084044, 2004.

[337] LIGO Scientific Collaboration and Virgo Collaboration. Noise curves used for simulations in the update of the observing scenarios paper, 2020.

[338] Maya Fishbach, Will M. Farr, and Daniel E. Holz. The Most Massive Binary Black Hole Detections and the Identification of Population Outliers. *Astrophys. J. Lett.*, 891(2):L31, 2020.

[339] Amanda M. Farah, Bruce Edelman, Michael Zevin, Maya Fishbach, Jose María Ezquiaga, Ben Farr, and Daniel E. Holz. Things That Might Go Bump in the Night: Assessing Structure in the Binary Black Hole Mass Spectrum. *Astrophys. J.*, 955(2):107, 2023.

[340] Thomas J. Loredo. Accounting for source uncertainties in analyses of astronomical survey data. *AIP Conf. Proc.*, 735(1):195–206, 2004.

[341] Ilya Mandel, Will M. Farr, and Jonathan R. Gair. Extracting distribution parameters from multiple uncertain observations with selection biases. *Mon. Not. Roy. Astron. Soc.*, 486(1):1086–1093, 2019.

[342] Salvatore Vitale, Davide Gerosa, Will M. Farr, and Stephen R. Taylor. Inferring the properties of a population of compact binaries in presence of selection effects. In *Handbook of Gravitational Wave Astronomy*. Springer Singapore, 2020.

[343] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[344] Andrew Gelman Samantha R Cook and Donald B Rubin. Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.

[345] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv e-prints*, page arXiv:1804.06788, April 2018.

[346] Sylvia Biscoveanu, Colm Talbot, and Salvatore Vitale. The effect of spin mismodelling on gravitational-wave measurements of the binary neutron star mass distribution. *Mon. Not. Roy. Astron. Soc.*, 511(3):4350–4359, 2022.

[347] Maya Fishbach and Daniel E. Holz. Minding the gap: GW190521 as a straddling binary. *Astrophys. J. Lett.*, 904(2):L26, 2020.

[348] Sandip Sinharay and Hal S. Stern. Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111(1-2):209–221, 2003.

[349] M. J. Bayarri and M. E. Castellanos. Bayesian Checking of the Second Levels of Hierarchical Models. *arXiv e-prints*, page arXiv:0802.0743, February 2008.

[350] Thomas Callister. Reweighting single event posteriors with hyperparameter marginalization, July 2021.

[351] Colm Talbot and Eric Thrane. Determining the population properties of spinning black holes. *Phys. Rev. D*, 96(2):023012, 2017.

[352] Rory J. E. Smith, Colm Talbot, Francisco Hernandez Vivanco, and Eric Thrane. Inferring the population properties of binary black holes from unresolved gravitational waves. *Mon. Not. Roy. Astron. Soc.*, 496(3):3281–3290, 2020.

[353] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[354] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[355] Astropy Collaboration. Astropy: A community Python package for astronomy. *Astronomy & Astrophysics*, 558:A33, October 2013.

[356] Astropy Collaboration. The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *Astronomical J.*, 156(3):123, September 2018.

[357] Colm Talbot and Eric Thrane. Measuring the binary black hole mass spectrum with an astrophysically motivated parameterization. *Astrophys. J.*, 856(2):173, 2018.

[358] P. A. R. Ade et al. Planck 2013 results. XVI. Cosmological parameters. *Astron. Astrophys.*, 571:A16, 2014.

[359] K. S. Thorne. Gravitational radiation. In *Three Hundred Years of Gravitation*, pages 330–458. 1987.

[360] Lee Samuel Finn and David F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. *Phys. Rev. D*, 47:2198–2219, 1993.

[361] Bruce Allen, Warren G. Anderson, Patrick R. Brady, Duncan A. Brown, and Jolien D. E. Creighton. FINDCHIRP: An Algorithm for detection of gravitational waves from inspiraling compact binaries. *Phys. Rev. D*, 85:122006, 2012.

[362] Reed Essick. Semianalytic sensitivity estimates for catalogs of gravitational-wave transients. *Phys. Rev. D*, 108(4):043011, 2023.

[363] P. C. Peters. Gravitational radiation and the motion of two point masses. *Phys. Rev.*, 136:B1224–B1232, Nov 1964.

[364] Luc Blanchet, Thibault Damour, Bala R. Iyer, Clifford M. Will, and Alan G. Wiseman. Gravitational-radiation damping of compact binary systems to second post-newtonian order. *Phys. Rev. Lett.*, 74:3515–3518, May 1995.

[365] Thomas A. Callister, Carl-Johan Haster, Ken K. Y. Ng, Salvatore Vitale, and Will M. Farr. Who Ordered That? Unequal-mass Binary Black Hole Mergers Have Larger Effective Spins. *Astrophys. J. Lett.*, 922(1):L5, 2021.

[366] Colm Talbot et al. Inference with finite time series: II. The window strikes back. *Class. Quant. Grav.*, 42(23):235023, 2025.

[367] Maximiliano Isi and Will M. Farr. Analyzing black-hole ringdowns. 7 2021.

[368] Chase Kimball, Colm Talbot, Christopher P. L. Berry, Matthew Carney, Michael Zevin, Eric Thrane, and Vicky Kalogera. Black Hole Genealogy: Identifying Hierarchical Mergers with Gravitational Waves. *Astrophys. J.*, 900(2):177, 2020.

[369] Giacomo Fragione, Bence Kocsis, Frederic A. Rasio, and Joseph Silk. Repeated Mergers, Mass-gap Black Holes, and Formation of Intermediate-mass Black Holes in Dense Massive Star Clusters. *Astrophys. J.*, 927(2):231, 2022.

[370] S. E. Woosley. Pulsational Pair-Instability Supernovae. *Astrophys. J.*, 836(2):244, 2017.

[371] S. E. Woosley and Alexander Heger. The Pair-Instability Mass Gap for Black Holes. *Astrophys. J. Lett.*, 912(2):L31, 2021.

[372] K. Belczynski et al. The Effect of Pair-Instability Mass Loss on Black Hole Mergers. *Astron. Astrophys.*, 594:A97, 2016.

[373] L. A. C. van Son, S. E. de Mink, F. S. Broekgaarden, M. Renzo, S. Justham, E. Laplace, J. Moran-Fraile, D. D. Hendriks, and R. Farmer. Polluting the pair-instability mass gap for binary black holes through super-Eddington accretion in isolated binaries. *Astrophys. J.*, 897(1):100, 2020.

[374] M. Renzo, M. Cantiello, B. D. Metzger, and Y. F. Jiang. The Stellar Merger Scenario for Black Holes in the Pair-instability Gap. *Astrophys. J. Lett.*, 904(2):L13, 2020.

[375] Juan Calderón Bustillo, Nicolas Sanchis-Gual, Alejandro Torres-Forné, and José A. Font. Confusing Head-On Collisions with Precessing Intermediate-Mass Binary Black Hole Mergers. *Phys. Rev. Lett.*, 126(20):201101, 2021.

[376] Rhys Green, Charlie Hoy, Stephen Fairhurst, Mark Hannam, Francesco Pannarale, and Cory Thomas. Identifying when Precession can be Measured in Gravitational Waveforms. *Phys. Rev. D*, 103(12):124023, 2021.

[377] Seth Olsen, Javier Roulet, Horng Sheng Chia, Liang Dai, Tejaswi Venumadhav, Barak Zackay, and Matias Zaldarriaga. Mapping the likelihood of GW190521 with diverse mass and spin priors. *Phys. Rev. D*, 104(8):083036, 2021.

[378] M. Coleman Miller and Douglas P. Hamilton. Production of intermediate-mass black holes in globular clusters. *Mon. Not. Roy. Astron. Soc.*, 330:232, 2002.

[379] Davide Gerosa and Emanuele Berti. Escape speed of stellar clusters from multiple-generation black-hole mergers in the upper mass gap. *Phys. Rev. D*, 100(4):041301, 2019.

[380] Giacomo Fragione, Abraham Loeb, and Frederic A. Rasio. On the Origin of GW190521-like events from repeated black hole mergers in star clusters. *Astrophys. J. Lett.*, 902(1):L26, 2020.

[381] B. McKernan, K. E. S. Ford, B. Kocsis, W. Lyra, and L. M. Winter. Intermediate-mass black holes in AGN discs – II. Model predictions and observational constraints. *Mon. Not. Roy. Astron. Soc.*, 441(1):900–909, 2014.

[382] Krzysztof Belczynski. The most ordinary formation of the most unusual double black hole merger. *Astrophys. J. Lett.*, 905(2):L15, 2020.

[383] László Gondán and Bence Kocsis. High eccentricities and high masses characterize gravitational-wave captures in galactic nuclei as seen by Earth-based detectors. *Mon. Not. Roy. Astron. Soc.*, 506(2):1665–1696, 2021.

[384] V. De Luca, V. Desjacques, G. Franciolini, P. Pani, and A. Riotto. GW190521 Mass Gap Event and the Primordial Black Hole Scenario. *Phys. Rev. Lett.*, 126(5):051101, 2021.

[385] H. L. Iglesias et al. Eccentricity Estimation for Five Binary Black Hole Mergers with Higher-order Gravitational-wave Modes. *Astrophys. J.*, 972(1):65, 2024.

[386] Antoni Ramos-Buades, Alessandra Buonanno, and Jonathan Gair. Bayesian inference of binary black holes with inspiral-merger-ringdown waveforms using two eccentric parameters. *Phys. Rev. D*, 108(12):124063, 2023.

[387] Juan Calderón Bustillo, Nicolas Sanchis-Gual, Alejandro Torres-Forné, José A. Font, Avi Vajpeyi, Rory Smith, Carlos Herdeiro, Eugen Radu, and Samson H. W. Leong. GW190521 as a Merger of Proca Stars: A Potential New Vector Boson of 8.7×10^{-13} eV. *Phys. Rev. Lett.*, 126(8):081101, 2021.

[388] Josu C. Aurrekoetxea, Thomas Helfer, and Eugene A. Lim. Coherent Gravitational Waveforms and Memory from Cosmic String Loops. *Class. Quant. Grav.*, 37(20):204001, 2020.

[389] Michael Zevin, Isobel M. Romero-Shaw, Kyle Kremer, Eric Thrane, and Paul D. Lasky. Implications of Eccentric Observations on Binary Black Hole Formation Channels. *Astrophys. J. Lett.*, 921(2):L43, 2021.

[390] Isobel M. Romero-Shaw, Paul D. Lasky, and Eric Thrane. Four Eccentric Mergers Increase the Evidence that LIGO–Virgo–KAGRA's Binary Black Holes Form Dynamically. *Astrophys. J.*, 940(2):171, 2022.

[391] Michael Unser. On the approximation of the discrete karhunen-loeve transform for stationary processes. *Signal Processing*, 7(3):231–249, 1984.

[392] Barak Zackay, Tejaswi Venumadhav, Javier Roulet, Liang Dai, and Matias Zaldarriaga. Detecting gravitational waves in data with non-stationary and non-Gaussian noise. *Phys. Rev. D*, 104(6):063034, 2021.

[393] Miriam Cabero, Collin D. Capano, Ofek Fischer-Birnholtz, Badri Krishnan, Alex B. Nielsen, Alexander H. Nitz, and Christopher M. Biwer. Observational tests of the black hole area increase law. *Phys. Rev. D*, 97(12):124069, 2018.

[394] Abhirup Ghosh et al. Testing general relativity using golden black-hole binaries. *Phys. Rev. D*, 94(2):021101, 2016.

[395] Maximiliano Isi, Will M. Farr, Matthew Giesler, Mark A. Scheel, and Saul A. Teukolsky. Testing the Black-Hole Area Law with GW150914. *Phys. Rev. Lett.*, 127(1):011103, 2021.

[396] Gregorio Carullo, Walter Del Pozzo, and John Veitch. Observational Black Hole Spectroscopy: A time-domain multimode analysis of GW150914. *Phys. Rev. D*, 99(12):123029, 2019. [Erratum: Phys.Rev.D 100, 089903 (2019)].

[397] Michael Boyle, Robert Owen, and Harald P. Pfeiffer. A geometric approach to the precession of compact binaries. *Phys. Rev. D*, 84:124011, 2011.

[398] Collin D. Capano, Jahed Abedi, Shilpa Kastha, Alexander H. Nitz, Julian Westerweck, Yi-Fan Wang, Miriam Cabero, Alex B. Nielsen, and Badri Krishnan. Statistical validation of the detection of a sub-dominant quasi-normal mode in GW190521. 9 2022.

[399] Héctor Estellés et al. A Detailed Analysis of GW190521 with Phenomenological Waveform Models. *Astrophys. J.*, 924(2):79, 2022.

[400] Ssohrab Borhanian, K. G. Arun, Harald P. Pfeiffer, and B. S. Sathyaprakash. Comparison of post-Newtonian mode amplitudes with numerical relativity simulations of binary black holes. *Class. Quant. Grav.*, 37(6):065006, 2020.

[401] Jahed Abedi, Collin D. Capano, Shilpa Kastha, Alexander H. Nitz, Yi-Fan Wang, Julian Westerweck, Alex B. Nielsen, and Badri Krishnan. Spectroscopy for asymmetric binary black hole mergers. *Phys. Rev. D*, 108(10):104009, 2023.

[402] Andrew Collette, Thomas Kluyver, Thomas A Caswell, James Tocknell, Jerome Kieffer, et al. h5py: 3.2.1, March 2021.

[403] Chad Galley. `GWTools`: A collection of useful gravitational wave tools, 2022.

[404] R. Abbott et al. GW190521 posterior samples, November 2021.

[405] Rich Abbott et al. Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo. *SoftwareX*, 13:100658, 2021.

[406] R. Abbott et al. Open Data from the Third Observing Run of LIGO, Virgo, KAGRA, and GEO. *Astrophys. J. Suppl.*, 267(2):29, 2023.

[407] R. Abbott et al. GW190521, August 2023.

[408] Stephen Fairhurst, Rhys Green, Mark Hannam, and Charlie Hoy. When will we observe binary black holes precessing? *Phys. Rev. D*, 102(4):041302, 2020.

[409] Charlie Hoy and Vivien Raymond. PESummary: the code agnostic Parameter Estimation Summary page builder. *SoftwareX*, 15:100765, 2021.

[410] Charlie Hoy, Stephen Fairhurst, Mark Hannam, and Vaibhav Tiwari. Understanding How Fast Black Holes Spin by Analyzing Data from the Second Gravitational-wave Catalogue. *Astrophys. J.*, 928(1):75, 2022.

[411] Vishal Baibhav, Zoheyr Doctor, and Vicky Kalogera. Dropping Anchor: Understanding the Populations of Binary Black Holes with Random and Aligned-spin Orientations. *Astrophys. J.*, 946(1):50, 2023.

[412] Richard O'Shaughnessy, Davide Gerosa, and Daniel Wysocki. Inferences about supernova physics from gravitational-wave measurements: GW151226 spin misalignment as an indicator of strong black-hole natal kicks. *Phys. Rev. Lett.*, 119(1):011101, 2017.

[413] Simon Stevenson, Christopher P. L. Berry, and Ilya Mandel. Hierarchical analysis of gravitational-wave measurements of binary black hole spin–orbit misalignments. *Mon. Not. Roy. Astron. Soc.*, 471(3):2801–2811, 2017.

[414] Bruce Allen. χ^2 time-frequency discriminator for gravitational wave detection. *Phys. Rev. D*, 71:062001, 2005.

[415] Samantha A. Usman et al. The PyCBC search for gravitational waves from compact binary coalescence. *Class. Quant. Grav.*, 33(21):215004, 2016.

[416] Derek Davis, Laurel V. White, and Peter R. Saulson. Utilizing aLIGO Glitch Classifications to Validate Gravitational-Wave Candidates. *Class. Quant. Grav.*, 37(14):145001, 2020.

[417] Rich Abbott et al. Search for intermediate-mass black hole binaries in the third observing run of Advanced LIGO and Advanced Virgo. *Astron. Astrophys.*, 659:A84, 2022.

[418] Jonathan Blackman, Scott E. Field, Mark A. Scheel, Chad R. Galley, Christian D. Ott, Michael Boyle, Lawrence E. Kidder, Harald P. Pfeiffer, and Béla Szilágyi. Numerical relativity waveform surrogate model for generically precessing binary black hole mergers. *Phys. Rev. D*, 96(2):024058, 2017.

[419] Alex Correia and Collin D. Capano. Sky marginalization in black hole spectroscopy and tests of the area theorem. *Phys. Rev. D*, 110(4):044018, 2024.

[420] Samaya Nissanke, Daniel E. Holz, Scott A. Hughes, Neal Dalal, and Jonathan L. Sievers. Exploring short gamma-ray bursts as gravitational-wave standard sirens. *Astrophys. J.*, 725:496–514, 2010.

[421] Michele Vallisneri. Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. *Phys. Rev. D*, 77:042001, 2008.

[422] Abhirup Ghosh, Nathan K. Johnson-Mcdaniel, Archisman Ghosh, Chandra Kant Mishra, Parameswaran Ajith, Walter Del Pozzo, Christopher P. L. Berry, Alex B. Nielsen, and Lionel London. Testing general relativity using gravitational wave signals from the inspiral, merger and ringdown of binary black holes. *Class. Quant. Grav.*, 35(1):014002, 2018.

[423] Matteo Breschi, Richard O'Shaughnessy, Jacob Lange, and Ofek Birnholtz. IMR consistency tests with higher modes on gravitational signals from the second observing run of LIGO and Virgo. *Class. Quant. Grav.*, 36(24):245019, 2019.

[424] R. Abbott et al. Tests of General Relativity with GWTC-3. *Phys. Rev. D*, 112(8):084080, 2025.

[425] M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[426] D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

[427] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[428] Solomon Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[429] Michael Boyle, Lawrence E. Kidder, Serguei Ossokine, and Harald P. Pfeiffer. Gravitational-wave modes from precessing black-hole binaries. 9 2014.

[430] Karl Pearson. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242, January 1895.

[431] K. G. Malmquist. On some relations in stellar statistics. *Meddelanden fran Lunds Astronomiska Observatorium Serie I*, 100:1–52, March 1922.

[432] Barry McKernan, K. E. Saavik Ford, Harrison E. Cook, Vera Delfavero, Emily McPike, Kaila Nathaniel, Jake Postiglione, Shawn Ray, and Richard O'Shaughnessy. McFACTS I: Testing the LVK AGN Channel with Monte Carlo for AGN Channel Testing and Simulation (McFACTS). *Astrophys. J.*, 990(2):217, 2025.

[433] Maximiliano Isi, Will M. Farr, and Vijay Varma. The Directional Isotropy of LIGO–Virgo Binaries. *Astrophys. J.*, 962(1):19, 2024.

[434] Bernard F. Schutz. Networks of gravitational wave detectors and three figures of merit. *Class. Quant. Grav.*, 28:125023, 2011.

[435] Samantha A. Usman, Joseph C. Mills, and Stephen Fairhurst. Constraining the Inclinations of Binary Mergers from Gravitational-wave Observations. *Astrophys. J.*, 877(2):82, 2019.

[436] Stephen Fairhurst. Localization of transient gravitational wave sources: beyond triangulation. *Class. Quant. Grav.*, 35(10):105002, 2018.

[437] Maria Okounkova, Maximiliano Isi, Katerina Chatziioannou, and Will M. Farr. Gravitational wave inference on a numerical-relativity simulation of a black hole merger beyond general relativity. *Phys. Rev. D*, 107(2):024046, 2023.

[438] Vijay Varma, Parameswaran Ajith, Sascha Husa, Juan Calderon Bustillo, Mark Hannam, and Michael Pürrer. Gravitational-wave observations of binary black holes: Effect of nonquadrupole modes. *Phys. Rev. D*, 90(12):124004, 2014.

[439] Werner Israel. Event Horizons in Static Vacuum Space-Times. *Phys. Rev.*, 164:1776–1779, 1967.

[440] Keefe Mitman et al. Probing the ringdown perturbation in binary black hole coalescences with an improved quasinormal mode extraction algorithm. *Phys. Rev. D*, 112(6):064016, 2025.

[441] Lisa Barsotti, Lee McCuller, Matthew Evans, and Peter Fritschel. The $A+$ design curve, March 2018.

[442] Kevin Kuns and Peter Fritschel. $A\sharp$ Strain Sensitivity, 2023.

[443] Kevin Kuns, Paul Fulda, Lisa Barsotti, and Matthew Evans. Cosmic Explorer Strain Sensitivity, October 2023.

[444] Alex Nitz, Ian Harry, Duncan Brown, Christopher M. Biwer, Josh Willis, Tito Dal Canton, Collin Capano, Thomas Dent, Larne Pekowsky, Gareth S Cabourn Davies, Soumi De, Miriam Cabero, Shichao Wu, Andrew R. Williamson, Bernd Machenschalk, Duncan Macleod, Francesco Pannarale, Prayush Kumar, Steven Reyes, dfinstad, Sumit Kumar, Márton Tápai, Leo Singer, Praveen Kumar, veronica villa, maxtrevor, Bhooshan Uday Varsha Gadre, Sebastian Khan, Stephen Fairhurst, and Arthur Tolley. gwastro/pycbc: v2.3.3 release of pycbc, January 2024.

[445] D. M. Macleod, J. S. Areeda, S. B. Coughlin, T. J. Massinger, and A. L. Urban. GWpy: A Python package for gravitational-wave astrophysics. *SoftwareX*, 13:100657, 2021.

[446] Matthew Mould. h5ify: Save python dictionaries into hdf5 files; load hdf5 files into python dictionaries., 2023.

[447] Abraham. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[448] Lawrence E. Kidder. Using full information when computing modes of post-Newtonian waveforms from inspiralling compact binaries in circular orbit. *Phys. Rev. D*, 77:044016, 2008.

[449] Eugene P. Wigner. *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*. Academic Press, 1959. Originally published in German as *Gruppentheorie und ihre Anwendungen auf die Quantenmechanik der Atomspektren*, Braunschweig: Vieweg Verlag, 1931. English translation by J. J. Griffin. Reprinted by Elsevier in 2013. ISBN: 978-1-4832-7576-5.

[450] Chandra Kant Mishra, Aditya Kela, K. G. Arun, and Guillaume Faye. Ready-to-use post-Newtonian gravitational waveforms for binary black holes with nonprecessing spins: An update. *Phys. Rev. D*, 93(8):084054, 2016.

[451] Alexander H. Nitz, Tito Dal Canton, Derek Davis, and Steven Reyes. Rapid detection of gravitational waves from compact binary mergers with PyCBC Live. *Phys. Rev. D*, 98(2):024050, 2018.

[452] Johan Samsing and Enrico Ramirez-Ruiz. On the Assembly Rate of Highly Eccentric Binary Black Hole Mergers. *Astrophys. J. Lett.*, 840(2):L14, 2017.

[453] Zhoujian Cao and Wen-Biao Han. Waveform model for an eccentric binary black hole based on the effective-one-body-numerical-relativity formalism. *Phys. Rev. D*, 96(4):044028, 2017.

[454] Xiaolin Liu, Zhoujian Cao, and Lijing Shao. Validating the Effective-One-Body Numerical-Relativity Waveform Models for Spin-aligned Binary Black Holes along Eccentric Orbits. *Phys. Rev. D*, 101(4):044049, 2020.

[455] Danilo Chiaramello and Alessandro Nagar. Faithful analytical effective-one-body waveform model for spin-aligned, moderately eccentric, coalescing black hole binaries. *Phys. Rev. D*, 101(10):101501, 2020.

[456] Alessandro Nagar, Piero Rettegno, Rossella Gamba, and Sebastiano Bernuzzi. Effective-one-body waveforms from dynamical captures in black hole binaries. *Phys. Rev. D*, 103(6):064013, 2021.

[457] Mohammed Khalil, Alessandra Buonanno, Jan Steinhoff, and Justin Vines. Radiation-reaction force and multipolar waveforms for eccentric, spin-aligned binaries in the effective-one-body formalism. *Phys. Rev. D*, 104(2):024046, 2021.

[458] Antoni Ramos-Buades, Alessandra Buonanno, Mohammed Khalil, and Serguei Ossokine. Effective-one-body multipolar waveforms for eccentric binary black holes with nonprecessing spins. *Phys. Rev. D*, 105(4):044035, 2022.

[459] Alessandro Nagar, Rossella Gamba, Piero Rettegno, Veronica Fantini, and Sebastiano Bernuzzi. Effective-one-body waveform model for noncircularized, planar, coalescing black hole binaries: The importance of radiation reaction. *Phys. Rev. D*, 110(8):084001, 2024.

[460] Tousif Islam, Avi Vajpeyi, Feroz H. Shaik, Carl-Johan Haster, Vijay Varma, Scott E. Field, Jacob Lange, Richard O'Shaughnessy, and Rory Smith. Analysis of GWTC-3 with fully precessing numerical relativity surrogate models. 9 2023.

[461] L. Santamaria et al. Matching post-Newtonian and numerical relativity waveforms: systematic errors and a new phenomenological model for non-precessing black hole binaries. *Phys. Rev. D*, 82:064016, 2010.

[462] Nihar Gupte et al. Evidence for eccentricity in the population of binary black holes observed by LIGO-Virgo-KAGRA. *Phys. Rev. D*, 112(10):104045, 2025.

[463] Andrea Antonelli, Konstantinos Kritos, Ken K. Y. Ng, Roberto Cotesta, and Emanuele Berti. Classifying the generation and formation channels of individual LIGO-Virgo-KAGRA observations from dynamically formed binaries. *Phys. Rev. D*, 108(8):084044, 2023.

[464] Christian Adamcewicz, Paul D. Lasky, and Eric Thrane. Evidence for a Correlation between Binary Black Hole Mass Ratio and Black Hole Spins. *Astrophys. J.*, 958(1):13, 2023.

[465] Isobel M. Romero-Shaw, Paul D. Lasky, and Eric Thrane. Signs of Eccentricity in Two Gravitational-wave Signals May Indicate a Subpopulation of Dynamically Assembled Binary Black Holes. *Astrophys. J. Lett.*, 921(2):L31, 2021.

[466] Serguei Ossokine et al. Multipolar Effective-One-Body Waveforms for Precessing Binary Black Holes: Construction and Validation. *Phys. Rev. D*, 102(4):044055, 2020.

[467] Rhiannon Udall and Derek Davis. Bayesian modeling of scattered light in the LIGO interferometers. *Appl. Phys. Lett.*, 122(9):094103, 2023.

[468] Chris Pankow et al. Mitigation of the instrumental noise transient in gravitational-wave data surrounding GW170817. *Phys. Rev. D*, 98(8):084016, 2018.

[469] Katerina Chatziioannou, Neil Cornish, Marcella Wijngaarden, and Tyson B. Littenberg. Modeling compact binary signals and instrumental glitches in gravitational wave data. *Phys. Rev. D*, 103(4):044013, 2021.

[470] Sophie Hourihane, Katerina Chatziioannou, Marcella Wijngaarden, Derek Davis, Tyson Littenberg, and Neil Cornish. Accurate modeling and mitigation of overlapping signals and glitches in gravitational-wave data. *Phys. Rev. D*, 106(4):042006, 2022.

[471] Sudarshan Ghonge, Joshua Brandt, J. M. Sullivan, Margaret Millhouse, Katerina Chatziioannou, James A. Clark, Tyson Littenberg, Neil Cornish, Sophie Hourihane, and Laura Cadonati. Assessing and mitigating the impact of glitches on gravitational-wave parameter estimation: A model agnostic approach. *Phys. Rev. D*, 110(12):122002, 2024.

[472] Neil J. Cornish and Tyson B. Littenberg. BayesWave: Bayesian Inference for Gravitational Wave Bursts and Instrument Glitches. *Class. Quant. Grav.*, 32(13):135012, 2015.

[473] Tyson B. Littenberg and Neil J. Cornish. Bayesian inference for spectral estimation of gravitational wave detector noise. *Phys. Rev. D*, 91(8):084034, 2015.

[474] Neil J. Cornish, Tyson B. Littenberg, Bence Bécsy, Katerina Chatziioannou, James A. Clark, Sudarshan Ghonge, and Margaret Millhouse. BayesWave analysis pipeline in the era of gravitational wave observations. *Phys. Rev. D*, 103(4):044006, 2021.

[475] Alfredo Carpineti. An unimaginable breakthrough: Loudest ever gravitational wave collision proves stephen hawking correct. https://www.iflscience.com/an-unimaginable-breakthrough-loudest-ever-gravitational-wave-collision-proves-stephen-hawking-correct-80760, 2025.

[476] S. W. Hawking. Gravitational radiation from colliding black holes. *Phys. Rev. Lett.*, 26:1344–1346, May 1971.

[477] Demetrios Christodoulou. Reversible and irreversible transformations in black-hole physics. *Phys. Rev. Lett.*, 25:1596–1597, Nov 1970.

[478] Demetrios Christodoulou and Remo Ruffini. Reversible transformations of a charged black hole. *Phys. Rev. D*, 4:3552–3555, Dec 1971.

[479] R. PENROSE and R. M. FLOYD. Extraction of rotational energy from a black hole. *Nature Physical Science*, 229(6):177–179, 1971.

[480] Erik Curiel. A Primer on Energy Conditions. *Einstein Stud.*, 13:43–104, 2017.

[481] S. W. HAWKING. Black hole explosions? *Nature*, 248(5443):30–31, 1974.

[482] Jacob D. Bekenstein. Generalized second law of thermodynamics in black-hole physics. *Phys. Rev. D*, 9:3292–3300, Jun 1974.

[483] S. W. Hawking. Particle creation by black holes. *Communications in Mathematical Physics*, 43(3):199–220, 1975.

[484] Roger Penrose. Gravitational Collapse: the Role of General Relativity. *Nuovo Cimento Rivista Serie*, 1:252, January 1969.

[485] J. M. Bardeen, B. Carter, and S. W. Hawking. The four laws of black hole mechanics. *Communications in Mathematical Physics*, 31(2):161–170, 1973.

[486] Tyson B. Littenberg and Neil J. Cornish. Bayesian inference for spectral estimation of gravitational wave detector noise. *Phys. Rev. D*, 91(8):084034, 2015.

[487] B. P. Abbott, R. Abbott, T. D. Abbott, et al. Tests of general relativity with gw150914. *Phys. Rev. Lett.*, 116:221101, May 2016.

[488] Clifford M. Will. The Confrontation between General Relativity and Experiment. *Living Rev. Rel.*, 17:4, 2014.

[489] Nicolas Yunes, Kent Yagi, and Frans Pretorius. Theoretical Physics Implications of the Binary Black-Hole Mergers GW150914 and GW151226. *Phys. Rev. D*, 94(8):084002, 2016.

[490] Marc Davis, Remo Ruffini, and Jayme Tiomno. Pulses of gravitational radiation of a particle falling radially into a schwarzschild black hole. *Phys. Rev. D*, 5:2932–2935, Jun 1972.

[491] Emanuele Berti, Vitor Cardoso, and Andrei O. Starinets. Quasinormal modes of black holes and black branes. *Class. Quant. Grav.*, 26:163001, 2009.

[492] Hermann Bondi, M. G. J. Van der Burg, and A. W. K. Metzner. Gravitational waves in general relativity, vii. waves from axi-symmetric isolated system. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 269(1336):21–52, 08 1962.

[493] R. K. : Sachs. Gravitational waves in general relativity viii. waves in asymptotically flat space-time. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 270(1340):103–126, 10 1962.

[494] Marco Raveri and Wayne Hu. Concordance and Discordance in Cosmology. *Phys. Rev. D*, 99(4):043506, 2019.

[495] L. Verde, T. Treu, and A. G. Riess. Tensions between the Early and the Late Universe. *Nature Astron.*, 3:891, 2019.

[496] Tejaswi Venumadhav, Barak Zackay, Javier Roulet, Liang Dai, and Matias Zaldarriaga. New binary black hole mergers in the second observing run of Advanced LIGO and Advanced Virgo. *Phys. Rev. D*, 101(8):083030, 2020.

[497] Bin Liu and Dong Lai. Black Hole and Neutron Star Binary Mergers in Triple Systems: Merger Fraction and Spin–Orbit Misalignment. *Astrophys. J.*, 863(1):68, 2018.

[498] Giacomo Fragione and Bence Kocsis. Effective spin distribution of black hole mergers in triples. *Mon. Not. Roy. Astron. Soc.*, 493(3):3920–3931, 2020.

[499] Charlie Hoy, Cameron Mills, and Stephen Fairhurst. Evidence for subdominant multipole moments and precession in merging black-hole-binaries from GWTC-2.1. *Phys. Rev. D*, 106(2):023019, 2022.

[500] Krzysztof Belczynski, Ronald E. Taam, Emmanouela Rantsiou, and Marc van der Sluys. Black Hole Spin Evolution: Implications for Short-hard Gamma Ray Bursts and Gravitational Wave Detection. *Astrophys. J.*, 682:474, 2008.

[501] Jim Fuller, Matteo Cantiello, Daniel Lecoanet, and Eliot Quataert. The Spin Rate of Pre-collapse Stellar Cores: Wave-driven Angular Momentum Transport in Massive Stars. *Astrophys. J.*, 810(2):101, September 2015.

[502] Katerina Chatziioannou, Eric Poisson, and Nicolas Yunes. Tidal heating and torquing of a Kerr black hole to next-to-leading order in the tidal coupling. *Phys. Rev. D*, 87(4):044022, 2013.

[503] Sebastian Khan, Katerina Chatziioannou, Mark Hannam, and Frank Ohme. Phenomenological model for the gravitational-wave signal from precessing binary black holes with two-spin effects. *Phys. Rev. D*, 100(2):024059, 2019.

[504] Javier Roulet and Matias Zaldarriaga. Constraints on binary black hole populations from LIGO–Virgo detections. *Mon. Not. Roy. Astron. Soc.*, 484(3):4216–4229, 2019.

[505] Daniel Wysocki, Jacob Lange, and Richard O'Shaughnessy. Reconstructing phenomenological distributions of compact binaries via gravitational wave observations. *Phys. Rev. D*, 100(4):043012, 2019.

[506] Vijay Varma, Sylvia Biscoveanu, Maximiliano Isi, Will M. Farr, and Salvatore Vitale. Hints of Spin-Orbit Resonances in the Binary Black Hole Population. *Phys. Rev. Lett.*, 128(3):031101, 2022.

[507] Vaibhav Tiwari, Stephen Fairhurst, and Mark Hannam. Constraining black-hole spins with gravitational wave observations. *Astrophys. J.*, 868(2):140, 2018.

[508] Horng Sheng Chia, Seth Olsen, Javier Roulet, Liang Dai, Tejaswi Venumadhav, Barak Zackay, and Matias Zaldarriaga. Signs of higher multipoles and orbital precession in GW151226. *Phys. Rev. D*, 106(2):024009, 2022.

[509] Tousif Islam, Scott E. Field, Carl-Johan Haster, and Rory Smith. Improved analysis of GW190412 with a precessing numerical relativity surrogate waveform model. *Phys. Rev. D*, 103(10):104027, 2021.

[510] Avi Vajpeyi, Rory Smith, and Eric Thrane. Deep Follow-up for Gravitational-wave Inference: A Case Study with GW151226. *Astrophys. J.*, 947(1):10, 2023.

[511] Maite Mateu-Lucena, Sascha Husa, Marta Colleoni, Héctor Estellés, Cecilio García-Quirós, David Keitel, Maria de Lluc Planas, and Antoni Ramos-Buades. Parameter estimation with the current generation of phenomenological waveform models applied to the black hole mergers of GWTC-1. *Mon. Not. Roy. Astron. Soc.*, 517(2):2403–2425, 2022.

[512] Seth Olsen, Tejaswi Venumadhav, Jonathan Mushkin, Javier Roulet, Barak Zackay, and Matias Zaldarriaga. New binary black hole mergers in the LIGO-Virgo O3a data. *Phys. Rev. D*, 106(4):043009, 2022.

[513] John Skilling. Nested Sampling. *AIP Conference Proceedings*, 735:395, November 2004.

[514] John Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833 – 859, 2006.

[515] Matthew Mould, Davide Gerosa, Floor S. Broekgaarden, and Nathan Steinle. Which black hole formed first? Mass-ratio reversal in massive binary stars from gravitational-wave data. *Mon. Not. Roy. Astron. Soc.*, 517(2):2738–2745, 2022.

[516] A. Olejak and K. Belczynski. The Implications of High Black Hole Spins for the Origin of Binary Black Hole Mergers. *Astrophys. J. Lett.*, 921(1):L2, November 2021.

[517] Krzysztof Belczynski, Christine Done, Scott Hagen, Jean-Pierre Lasota, and Koushik Sen. Common origin for black holes in both high mass X-ray

binaries and gravitational-wave sources. *Astron. Astrophys.*, 690:A21, 2024.

[518] Yong Shao and Xiang-Dong Li. Stable Mass Transfer Can Explain Massive Binary Black Hole Mergers with a High-spin Component. *Astrophys. J.*, 930(1):26, 2022.

[519] Sylvia Biscoveanu, Thomas A. Callister, Carl-Johan Haster, Ken K. Y. Ng, Salvatore Vitale, and Will M. Farr. The Binary Black Hole Spin Distribution Likely Broadens with Redshift. *Astrophys. J. Lett.*, 932(2):L19, 2022.

[520] Simone S. Bavera, Maya Fishbach, Michael Zevin, Emmanouil Zapartas, and Tassos Fragos. The χeff $-$ z correlation of field binary black hole mergers and how 3G gravitational-wave detectors can constrain it. *Astron. Astrophys.*, 665:A59, 2022.

[521] Maya Fishbach, Daniel E. Holz, and Will M. Farr. Does the Black Hole Merger Rate Evolve with Redshift? *Astrophys. J. Lett.*, 863(2):L41, 2018.

[522] 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. *Astrophys. J.*, 896(2):L44, June 2020.

[523] Michele Vallisneri, Jonah Kanner, Roy Williams, Alan Weinstein, and Branson Stephens. The LIGO Open Science Center. In *Journal of Physics Conference Series*, volume 610 of *Journal of Physics Conference Series*, page 012021, May 2015.

[524] Reed Essick and Will Farr. Precision Requirements for Monte Carlo Sums within Hierarchical Bayesian Inference. 4 2022.

[525] Isabella Verdinelli and Larry Wasserman. Computing bayes factors using a generalization of the savage-dickey density ratio. *Journal of the American Statistical Association*, 90(430):614, 1995.

[526] Katerina Chatziioannou, Neil Cornish, Antoine Klein, and Nicolás Yunes. Detection and Parameter Estimation of Gravitational Waves from Compact Binary Inspirals with Analytical Double-Precessing Templates. *Phys. Rev. D*, 89(10):104023, 2014.

[527] Linhao Ma and Jim Fuller. Angular momentum transport in massive stars and natal neutron star rotation rates. *Mon. Not. Roy. Astron. Soc.*, 488(3):4338–4355, 2019.

[528] Ilya Mandel and Tassos Fragos. An alternative interpretation of GW190412 as a binary black hole merger with a rapidly spinning secondary. *Astrophys. J. Lett.*, 895(2):L28, 2020.

[529] E. P. J. van den Heuvel, S. F. Portegies Zwart, and S. E. de Mink. Forming short-period Wolf–Rayet X-ray binaries and double black holes through stable mass transfer. *Mon. Not. Roy. Astron. Soc.*, 471(4):4256–4264, 2017.

[530] Coenraad J. Neijssel, Alejandro Vigna-Gómez, Simon Stevenson, Jim W. Barrett, Sebastian M. Gaebel, Floor Broekgaarden, Selma E. de Mink, Dorottya Szécsi, Serena Vinciguerra, and Ilya Mandel. The effect of the metallicity-specific star formation history on double compact object mergers. *Mon. Not. Roy. Astron. Soc.*, 490(3):3740–3759, 2019.

[531] Jakob Stegmann and Fabio Antonini. Flipping spins in mass transferring binaries and origin of spin-orbit misalignment in binary black holes. *Phys. Rev. D*, 103(6):063007, 2021.

[532] Benjamin Farr and Will Farr. On the incompatibility of q–$\chi_{\rm eff}$ correlations and iid spins. LIGO DCC, 2025.

[533] Sylvia Biscoveanu, Maximiliano Isi, Salvatore Vitale, and Vijay Varma. New Spin on LIGO-Virgo Binary Black Holes. *Phys. Rev. Lett.*, 126(17):171103, 2021.

[534] Lillie Szemraj and Sylvia Biscoveanu. Disentangling spinning and non-spinning binary black hole populations with spin sorting. *Class. Quant. Grav.*, 43(8):085005, 2026.

[535] Salvatore Vitale and Matthew Mould. Long road to alignment: Measuring black hole spin orientation with expanding gravitational-wave datasets. *Phys. Rev. D*, 112(8):083015, 2025.

[536] Sharan Banagiri, Thomas A. Callister, Christian Adamcewicz, Zoheyr Doctor, and Vicky Kalogera. Structure and Skewness of the Effective Inspiral Spin Distribution of Binary Black Hole Mergers. *Astrophys. J.*, 990(2):147, 2025.

[537] Maya Fishbach, Chase Kimball, and Vicky Kalogera. Limits on Hierarchical Black Hole Mergers from the Most Negative χ_{eff} Systems. *Astrophys. J. Lett.*, 935(2):L26, 2022.

[538] Vishal Baibhav, Davide Gerosa, Emanuele Berti, Kaze W. K. Wong, Thomas Helfer, and Matthew Mould. The mass gap, the spin gap, and the origin of merging binary black holes. *Phys. Rev. D*, 102(4):043002, 2020.

[539] April Qiu Cheng, Michael Zevin, and Salvatore Vitale. What You Don't Know Can Hurt You: Use and Abuse of Astrophysical Models in Gravitational-wave Population Analyses. *Astrophys. J.*, 955(2):127, 2023.

[540] Samsuzzaman Afroz and Suvodip Mukherjee. Phase space of binary black holes from gravitational wave observations to unveil its formation history. *Phys. Rev. D*, 112(2):023531, 2025.

[541] Manuel Arca Sedda, Albrecht W. H. Kamlah, Rainer Spurzem, Francesco Paolo Rizzuto, Thorsten Naab, Mirek Giersz, and Peter Berczik. The DRAGON-II simulations - II. Formation mechanisms, mass, and spin of intermediate-mass black holes in star clusters with up to 1 million stars. *Monthly Notices of the Royal Astronomical Society*, 526(1):429–442, November 2023.

[542] Christian Adamcewicz and Eric Thrane. Do unequal-mass binary black hole systems have larger χeff? Probing correlations with copulas in gravitational-wave astronomy. *Mon. Not. Roy. Astron. Soc.*, 517(3):3928–3937, 2022.

[543] Davide Gerosa, Geraint Pratten, and Alberto Vecchio. Gravitational-wave selection effects using neural-network classifiers. *Phys. Rev. D*, 102(10):103020, 2020.

[544] Stefano Rinaldi and Walter Del Pozzo. (H)DPGMM: a hierarchy of Dirichlet process Gaussian mixture models for the inference of the black hole mass function. *Mon. Not. Roy. Astron. Soc.*, 509(4):5454–5466, 2021.

[545] Colm Talbot and Eric Thrane. Flexible and Accurate Evaluation of Gravitational-wave Malmquist Bias with Machine Learning. *Astrophys. J.*, 927(1):76, 2022.

[546] Konstantin Leyde, Stephen R. Green, Alexandre Toubiana, and Jonathan Gair. Gravitational wave populations and cosmology with neural posterior estimation. *Phys. Rev. D*, 109(6):064056, 2024.

[547] Isobel M. Romero-Shaw, Eric Thrane, and Paul D. Lasky. When models fail: An introduction to posterior predictive checks and model misspecification in gravitational-wave astronomy. *Publ. Astron. Soc. Austral.*, 39:e025, 2022.

[548] Jam Sadiq, Thomas Dent, and Daniel Wysocki. Flexible and fast estimation of binary merger population distributions with an adaptive kernel density estimator. *Phys. Rev. D*, 105(12):123014, 2022.

[549] Ethan Payne and Eric Thrane. Model exploration in gravitational-wave astronomy with the maximum population likelihood. *Phys. Rev. Res.*, 5(2):023013, 2023.

[550] Nir Guttman, Ethan Payne, Paul D. Lasky, and Eric Thrane. Trends in the Population of Binary Black Holes Following the Fourth

Gravitational-wave Transient Catalog: A Data-driven Analysis. *Astrophys. J.*, 996(2):144, 2026.

[551] Andrew Gelman, Xiao-Li Meng, and Hal S. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–807, 1996.

[552] Donald B. Rubin. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151–1172, December 1984.

[553] Irwin Guttman. The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 29(1):83–100, 1967.

[554] Anarya Ray, Shirsha Mukherjee, Michael Zevin, and Vicky Kalogera. On the Astrophysical Origin of Binary Black Hole Subpopulations: A Tale of Three Channels? 3 2026.

[555] Matthew Mould, Jack Heinzel, Sofia Alvarez-Lopez, Cailin Plunkett, Noah E. Wolfe, and Salvatore Vitale. Measurement prospects for the pair-instability mass cutoff with gravitational waves. 2 2026.

[556] Christian Adamcewicz, Shanika Galaudage, Paul D. Lasky, and Eric Thrane. Which Black Hole Is Spinning? Probing the Origin of Black Hole Spin with Gravitational Waves. *Astrophys. J. Lett.*, 964(1):L6, 2024.

[557] Christian Adamcewicz, Nir Guttman, Paul D. Lasky, and Eric Thrane. Do Both Black Holes Spin in Merging Binaries? Evidence from GWTC-4 and Astrophysical Implications. *Astrophys. J.*, 994(2):261, 2025.

[558] Michele Vallisneri, Patrick M. Meyers, Katerina Chatziioannou, and Alvin J. K. Chua. Posterior predictive checking for gravitational-wave detection with pulsar timing arrays. I. The optimal statistic. *Phys. Rev. D*, 108(12):123007, 2023.

[559] Patrick M. Meyers, Katerina Chatziioannou, Michele Vallisneri, and Alvin J. K. Chua. Posterior predictive checking for gravitational-wave detection with pulsar timing arrays. II. Posterior predictive distributions and pseudo-Bayes factors. *Phys. Rev. D*, 108(12):123008, 2023.

[560] Gabriella Agazie et al. The NANOGrav 15 yr dataset: Posterior predictive checks for gravitational-wave detection with pulsar timing arrays. *Phys. Rev. D*, 111(4):042011, 2025.

[561] Rutger van Haasteren. Calculation of p-values for quadratic statistics in pulsar timing arrays. *Phys. Rev. D*, 112(10):103009, 2025.

[562] M. J. Bayarri and J. O. Berger. P values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142, December 2000.

[563] Xiao-Li Meng. Posterior predictive p-values. *The Annals of Statistics*, 22(3):1142–1160, 1994.

[564] James M. Robins, Aad W. van der Vaart, and Valérie Ventura. Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, 95(452):1143–1156, December 2000.

[565] Nils Lid Hjort, Fredrik A. Dahl, and Gunnhildur H. Steinbakk. Post-processing posterior predictive p values. *Journal of the American Statistical Association*, 101(475):1157–1174, 2006.

[566] Tousif Islam, Javier Roulet, and Tejaswi Venumadhav. Factorized Parameter Estimation for Real-Time Gravitational Wave Inference. 10 2022.

[567] Jiawei Li and Jonathan H. Huggins. Calibrated Model Criticism Using Split Predictive Checks. *arXiv e-prints*, page arXiv:2203.15897, March 2022.

[568] Gemma E Moran, David M Blei, and Rajesh Ranganath. Holdout predictive checks for Bayesian model criticism. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):194–214, February 2024.

[569] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. Gwtc-4.0: Population properties of merging compact binaries, 2025.

[570] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. Gwtc-4.0: O4a search sensitivity estimates, 2025.

[571] 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 — o1+o2+o3 search sensitivity estimates, 2021.

[572] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. Gwtc-4.0: Cumulative search sensitivity estimates, 2025.

[573] 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, 2021.

[574] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration. Data behind figures of "gwtc-4.0: An introduction to version 4.0 of the gravitational wave transient catalog", 2025.

[575] S. Soni et al. LIGO Detector Characterization in the first half of the fourth Observing run. *Class. Quant. Grav.*, 42(8):085016, 2025.

[576] Noah E. Wolfe, Matthew Mould, Jack Heinzel, and Salvatore Vitale. Studying the gravitational-wave population without looking that FAR out. 10 2025.

[577] Vaibhav Tiwari. Estimation of the Sensitive Volume for Gravitational-wave Source Populations Using Weighted Monte Carlo Integration. *Class. Quant. Grav.*, 35(14):145009, 2018.

[578] Javier Roulet and Tejaswi Venumadhav. Inferring Binary Properties from Gravitational Wave Signals. 2 2024.

[579] Snehal Tibrewal, Aaron Zimmerman, Jacob Lange, and Deirdre Shoemaker. Misinterpreting Spin Precession as Orbital Eccentricity in Gravitational-Wave Signals. 1 2026.

[580] Thomas Huckle. Computations with gohberg-semencul-type formulas for toeplitz matrices. *Linear Algebra and its Applications*, 273(1):169–198, 1998.

[581] Neha Sharma, Aditya Vijaykumar, and Prayush Kumar. Rapid inference of gravitational-wave signals in the time domain using a heterodyned likelihood. 1 2026.

[582] Lionel London, Deirdre Shoemaker, and James Healy. Modeling ringdown: Beyond the fundamental quasinormal modes. *Phys. Rev. D*, 90(12):124032, 2014. [Erratum: Phys.Rev.D 94, 069902 (2016)].

[583] L. T. London. Modeling ringdown. II. Aligned-spin binary black holes, implications for data analysis and fundamental theory. *Phys. Rev. D*, 102(8):084052, 2020.

[584] Anarya Ray, Sharan Banagiri, and Eric Thrane. GW231123: extreme spins or microglitches? 10 2025.

[585] Gregory Ashton, Colm Talbot, Andrew Lundgren, Ann-Kristin Malz, and Joseph Areeda. Measuring the rate of glitches in interferometric gravitational wave detectors with a hierarchical Bayesian model. 4 2026.

[586] Matteo Di Giovanni. Einstein Telescope an Cosmic Explorer. In *59th Rencontres de Moriond on Gravitation: Moriond 2025 Gravitation*, 5 2025.

[587] Yuetong Zhao, Youjun Lu, Changshuo Yan, Zhiwei Chen, and Wei-Tou Ni. Multiband gravitational wave observations of stellar binary black holes at the low to middle and high frequencies. *Mon. Not. Roy. Astron. Soc.*, 522(2):2951–2966, 2023.

[588] Evan D. Hall and Matthew Evans. Metrics for next-generation gravitational-wave detectors. *Classical and Quantum Gravity*, 36(22):225002, November 2019.

[589] Travis Robson, Neil J. Cornish, and Chang Liu. The construction and use of LISA sensitivity curves. *Class. Quant. Grav.*, 36(10):105011, 2019.

[590] Matthew C. Digman and Neil J. Cornish. LISA Gravitational Wave Sources in a Time-varying Galactic Stochastic Background. *Astrophys. J.*, 940(1):10, 2022.

[591] Tyson B. Littenberg and Neil J. Cornish. Prototype global analysis of LISA data with multiple source types. *Phys. Rev. D*, 107(6):063004, 2023.

[592] Swetha Bhagwat, Costantino Pacilio, Enrico Barausse, and Paolo Pani. Landscape of massive black-hole spectroscopy with LISA and the Einstein Telescope. *Phys. Rev. D*, 105(12):124063, 2022.

[593] Rhondale Tso, Davide Gerosa, and Yanbei Chen. Optimizing LIGO with LISA forewarnings to improve black-hole spectroscopy. *Phys. Rev. D*, 99(12):124043, 2019.

[594] Salvatore Vitale and Matthew Evans. Parameter estimation for binary black holes with networks of third generation gravitational-wave detectors. *Phys. Rev. D*, 95(6):064052, 2017.

[595] Salvatore Vitale and Chris Whittle. Characterization of binary black holes by heterogeneous gravitational-wave networks. *Phys. Rev. D*, 98(2):024029, 2018.

[596] Ken K. Y. Ng, Salvatore Vitale, Will M. Farr, and Carl L. Rodriguez. Probing multiple populations of compact binaries with third-generation gravitational-wave detectors. *Astrophys. J. Lett.*, 913(1):L5, 2021.

[597] Matthew C. Digman and Neil J. Cornish. Parameter estimation for stellar-origin black hole mergers in LISA. *Phys. Rev. D*, 108(2):023022, 2023.

POCKET MATERIAL: TIME-DOMAIN INFERENCE ON GW231123

Following the methodology of Part II, I now briefly present a *preliminary*comparison of spin precession measurements before and after various cutoff times for the gravitational-wave detection, GW231123 [21]—the most massive binary black hole observed to date—and GW190521 (Ch. 7). For GW190521, we need to see quiet pre-inspiral cycle ($-30\,M \lesssim t \lesssim -10\,M$) and loud merger to constrain spin precession. However, GW231123's inference hinges on the data right around merger ($-10\,M \lesssim t \lesssim 10\,M$). Further work is needed to analyze this seemingly distinct morphological imprint.

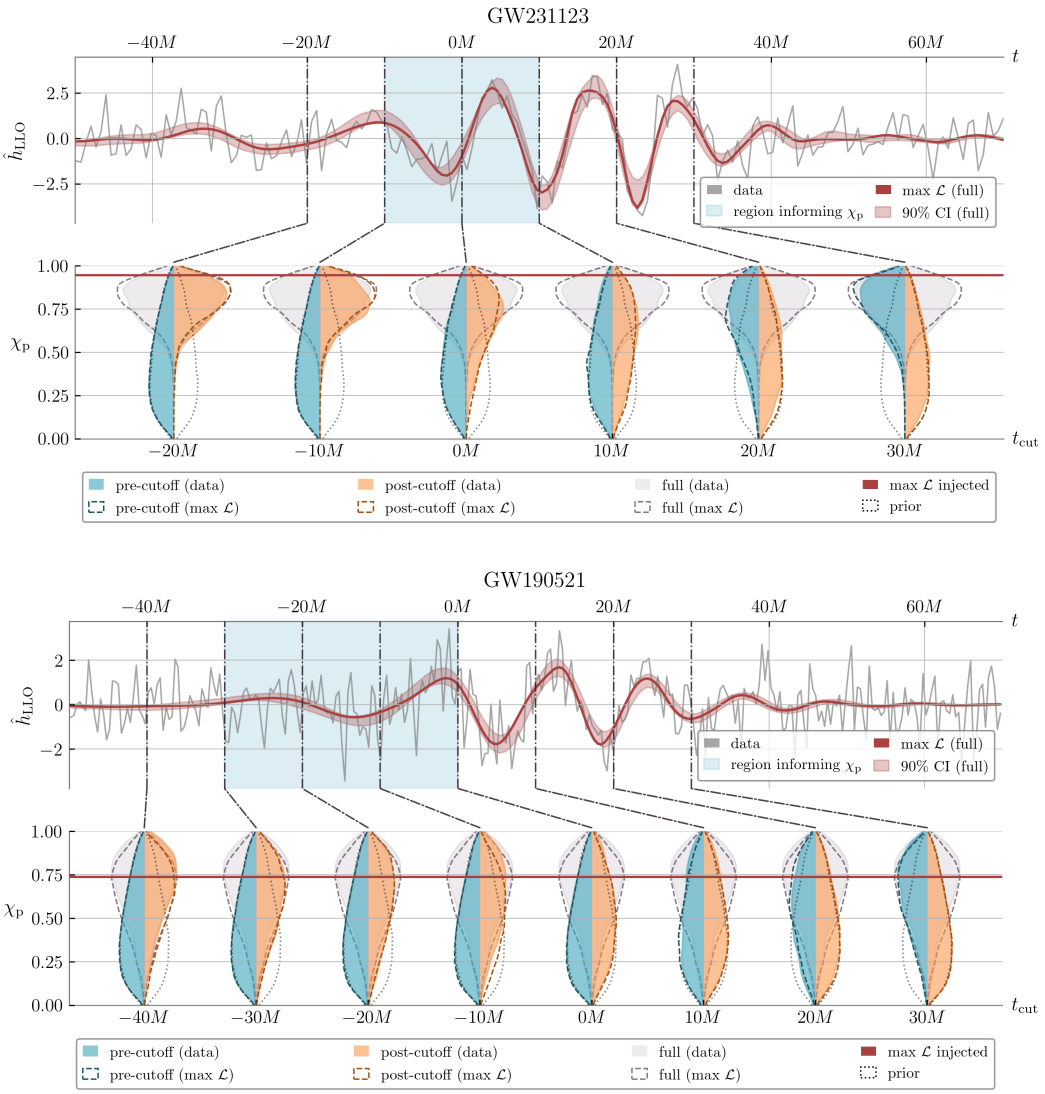


Figure 15.1: Comparison of spin precession measurements before and after various cutoff times for GW231123 (top half of figure) and GW190521 (bottom half). For each signal, we show the whitened data in LLO (gray), recovered 90% credible waveform reconstructions from analyzing that full data (light red), and the max. $\mathcal{L}$ reconstruction (dark red). Cutoff times are indicated with the black dot-dashed lines drawn between the time in each signal and the corresponding χ_{p} posterior distribution below. For both the true signals (filled) and corresponding max. $\mathcal{L}$ injections (empty with dashed outline), posteriors from the pre-cutoff analyses are shown in blue (left half-violins), while post-cutoff analyses are in orange (right half-violins). As a point of comparison, we also include the full-signal posterior (gray filled full-violins), prior (empty full-violins with gray dotted outline), and injected max. $\mathcal{L}$ χ_{p} (dark red horizontal lines) determined from the from the full real-data analysis. From the behavior of the posteriors as a function of cutoff time, we determine the temporal interval that is most informative about spin precession; these are indicated with the light-blue shaded regions in each waveform plot.

POCKET MATERIAL: GW231123 INFOGRAPHIC

For better or for worse, scientific discovery means little if it cannot be communicated. During my graduate studies, I have sought not only to conduct important science, but to make those innovations accessible to the public. On this front, my important contribution was making the headline outreach infographic [Fig. 15.2] for GW231123 and coordinating the graphic's translation into over 10 languages, enabling global circulation. The graphic has appeared in Science News (USA), Coelum Astronomia (Italy), AstroArts (Japan), LIGO-India News (India), the LIGO magazine (international), and many other publications around the world.

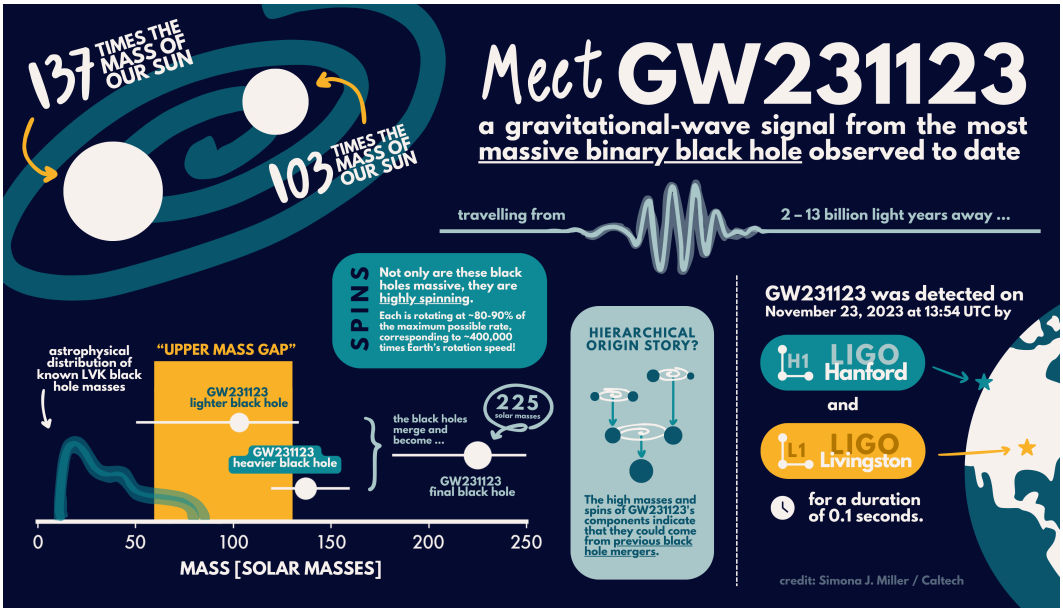


Figure 15.2: "Meet GW231123" Infographic made by ***SJM*** for the event's announcement and press release.