LIGO SURF Interim Report 2: The impact of astrophysical population model choices on post-Newtonian deviation tests of general relativity

Ruby Knudsen, Mentor: Ethan Payne

August 2024

Abstract

The Laser Interferometer Gravitational-wave Observatory detects gravitational waves and uses them to test the theory that predicts their existence: Einstein's theory of general relativity. Testing general relativity using gravitational waves can be done at the individual event level and the population level. Most current tests of general relativity are not inferred jointly with a population model to describe the astrophysical distribution of the sources. The omission of a population model describing the astrophysical population of binary black hole mergers inscribes an implied population model that may be inaccurate. This inaccurate population model could lead to biases in supposed deviations from general relativity. This investigation used injected signals of simulated binary black hole systems to infer the probability distributions for population model parameters. These probability distributions are constructed using Markov Chain Monte Carlo analysis. A family of mass population models was injected and recovered using each of the different models as the population distribution to determine whether biases in deviations from general relativity and population misspecification can be observed. Further research is needed as more gravitational wave events are detected and our understanding of the mass population distribution evolves.

1 Background

In the early 20th century, Albert Einstein put forward the theory of general relativity (GR), and summarized his theory mathematically through the Einstein field equations. Part of GR theorized the concept of spacetime, which spatially distorts in the vicinity of massive astrophysical objects. Building off of this work, two years later Einstein posited the existence of gravitational waves (GW) — spatial distortions as transverse waves that originate from non-spherically symmetric quadrupolar disturbances before traveling away at the speed of light [2, 9]. Because GR theorizes the existence of GWs, it is possible to use GWs as a method to test GR. There are several tests of GR, but GWs are the best probe for the theory.

Different tests of GR interrogate the theory to different levels, as summarized by Fig. 1. While all astrophysical objects have gravitational potential and cause spacetime curvature — a key insight of GR — the most extreme tests of GR exist in the strong-field regime. The gravitational potential is proportional to M/R^3, and the curvature is proportional to M/R. The perihelion procession of Mercury probes GR, but not strenuously, since it has potential between 10^{-7} and 10^{-8}, and curvature between 10^{-32} and 10^{-33} cm^{-2} [16]. Due to the Schwarzschild radius, the event horizon of a black hole yields potential approximately 0.5, and is the most extreme possible potential. While black holes Sagittarius A* and M87 both have potential near 0.5, their curvature is orders of magnitude lower than the spacetime curvature near merging stellar mass black hole binaries [7]. Curvatures between 10^{-14} and 10^{-10} cm^{-2} are the most extreme for realistic astrophysical signals [7]. Probing GR in this region requires the detection of GWs from black hole and neutron star mergers. Thus, GW tests probe GR at the most extreme curvatures and potentials, and as such are the best tests of GR.

The Laser Interferometer Gravitational-wave Observatory (LIGO) — a joint project between the California Institute of Technology, the Massachusetts Institute of Technology, and the National Science Foundation — was built to detect these GWs and aims to further our understanding of GWs and test GR [1, 4]. LIGO

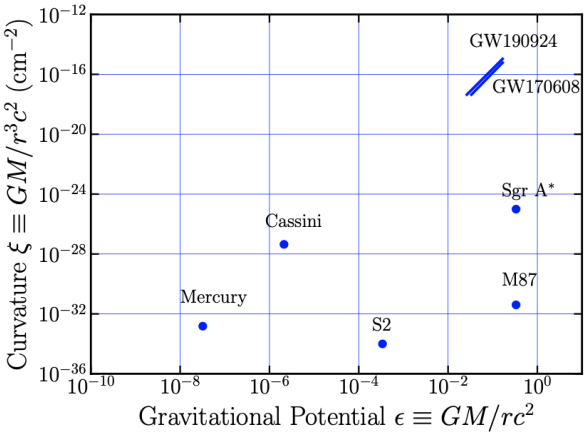

Figure 1: From Ref. [16]. This graphic shows a plot of gravitational potential vs. curvature for various tests of GR. The theoretical limit for the gravitational potential exists between 0.5 and 1.

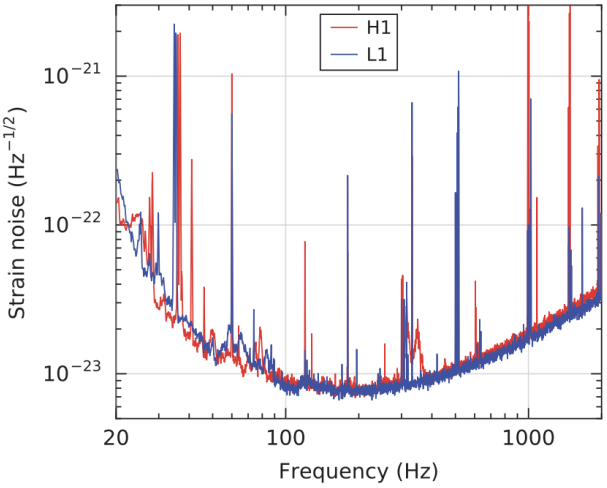

Figure 2: From Ref. [2]. This figure shows the sensitivity curve of LIGO, at different frequencies from the first observing run. LIGO is most sensitive to frequencies between 20 Hz and 1 kHz.

has perpendicular arms that detect compact binary mergers using the interference of light due to the difference in length of the arms that a passing GW causes. This change in length is directly proportional to GW strength [1, 4]. The compact binary mergers, mostly binary black hole mergers, are then analyzed to check for consistency with GR. LIGO is sensitive to GW frequencies between 20 and 1000 Hz, so black holes of stellar mass are easiest to detect [10]. Fig. 2 shows the frequencies associated with orbital velocities near the merger of binary stellar mass black holes, and the lower strain noise corresponds to better detection.

2 Introduction to methods for testing GR

2.1 Modeling modifications to GR with individual observations

There are many different ways of modeling GR deviations by modeling the GW signal. One of these tests is inferring deviations from the post-Newtonian (PN) approximation. The post-Newtonian approximation at the lowest order is the quadrupole formula, which estimates the emitted radiation from a quadrupolar mass distribution [15]. The PN expansion for gravitational-wave emission involves the dimensionless parameter

v/c, where v is the orbital velocity of the binary, with the lowest order being the Newtonian-order v/c order, and higher orders being $(v/c)^2$, $(v/c)^3$, and so on [18, 8]. PN tests are run by first constructing a post-Newtonian description of the GW inspiral in the frequency domain, before making modifications to individual parameters in the phase evolution [6, 17]. This is done with the phase, not the amplitude, because LIGO is more sensitive to phase deviations than those of amplitudes; over the course of an inspiral, phase shifts accumulate, while amplitude differences do not.

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

In this equation, $\Phi(f)$ is the frequency domain GW phase under the stationary phase approximation $\tilde{f}$, the red-shifted chirp-mass multiplied by f (the frequency), divided by c^3. Additionally, t_c and ϕ_c are the coalescence time and phase respectively, η is the symmetric mass ratio, and φ_k, $\varphi_{k,l}$ are PN coefficients for the $k/2$ PN order [6, 17].

After running these PN tests, a posterior is obtained by running Bayesian inference.

Bayesian inference makes statements about the Universe from data by inferring the distribution probable for parameters of each event analyzed, known as the posterior distribution [19]. Calculating this posterior distribution requires a likelihood, to describe the measurement, and a prior, to encode prior beliefs about each event's parameters θ. Bayesian inference uses Bayes' Theorem,

$$p(\theta|d) = \frac{p(d|\theta)\pi(\theta)}{\mathcal{Z}} \tag{2}$$

In this equation, $p(\theta|d)$ is the posterior, $p(d|\theta)$ is the likelihood, and $\pi(\theta)$ is the prior, and the evidence, which normalizes the distribution, is denoted as $\mathcal{Z}$ [19]. This same formalism can be used to study multiple events, with a modified version of the likelihood function.

2.2 Testing GR with many observations

Hierarchical inference is a method that uses posterior distributions from individual events to model the overall population of the astrophysical distribution from which the events originated. The analysis returns a distribution for the hyper-parameters of the population model, which is called the hyper-posterior [5, 19]. In this hyper-posterior, the astrophysical population has been inferred under the assumption of the choice of population model. A population model is important because otherwise an incorrect astrophysical population may be implicitly assumed, which in turn leads to bias in supposed measured deviations from GR [14].

To look at the population properties of a collection of events, the prior for the parameters θ is made conditional on hyper-parameters Λ, which encodes the astrophysical distribution from which the θs are drawn [19]. These hyper-parameters are crucial because they parameterize the shape of the inferred astrophysical population, and one goal of population inference is estimating the posterior of these hyper-parameters. One key piece of estimating the hyper-posterior is the population likelihood, which is calculated using Eq. 3.

A population likelihood is necessary for hierarchical inference. The population likelihood is a formula that allows simultaneous astrophysical population inference and GR testing, decreasing this bias [14, 12]. The equation for the population likelihood is

$$p(\{d\}|\Lambda) = \frac{1}{\xi(\Lambda)^N} \prod_{i=1}^{N} \int d\theta_i p(d_i|\theta_i)\pi(\theta_i|\Lambda) \tag{3}$$

Here, $\{d\}$ is the collection of N observations, $\xi(\Lambda)$ accounts for selection biases and is the detectable fraction of observations given the population hyper-parameters, $p(d_i|\theta_i)$ is the likelihood for each individual event, and $\pi(\theta_i|\Lambda)$ are the hyper-priors for each event [19]. Thus, $\pi(\theta_i|\Lambda)$ is where the population distribution is encoded.

This equation shows how individual observations are put together for hierarchical inference. As stated earlier, this hierarchical approach requires a population model, so we therefore must choose one population model to use. Currently, we are not sure in what manner the different choices of population model impact GR tests. This leads to my project.

3 Project motivation

Population distributions depend in part on the choice of population model, thus the choice of population model must be carefully considered. So far, this approach of joint inference has, at the population level, yielded GR deviations more consistent with GR by about 0.4σ, when using a PowerLaw+Peak population distribution for the mass of a black hole [14, 3]. However, it is possible that deviation from general relativity could be absorbed or hidden by an incorrect, assumed astrophysical population, so it is important to study the impact of different astrophysical population models on inferred GR deviation constraints. It is true that at some point that the incorrect choice will also lead to biases, the question is when, and how badly. Testing this using different population models is what I am working on.

4 Methods

This project had three main parts. First, we looked at testing general relativity (TGR) by writing code to analyze GR deviation parameters, modeled as a Gaussian distribution, in the TGR Toy Model section. Next, we built out the model to jointly analyze GR deviation parameters and population characteristics in the Building the Model section. Finally, we injected and recovering each of three primary mass distribution population models in the Model Analysis section. For simplicity's sake, and because time was unfortunately short, the detection fraction was ignored. Further studies should take this into account.

4.1 TGR Toy Model

To work on population inference calculations, I began by recreating the analysis in Ref. [11] using a toy model to "test" general relativity. This required generating a simulated population of 100 events, each with a mean and a standard deviation. Each standard deviation was drawn randomly between 0 and 1, while the means were each drawn from a Gaussian with a mean of 0 and the associated standard deviation. In the beginning, I went straight to the likelihood of the population by using an analytical approximation for the integral of the product of two Gaussian distributions. I ended up with a result consistent with my original distribution, which implied that my toy model agreed with general relativity.

After creating this simple version of the problem, I made the test more consistent with current analyses of LIGO data by drawing posterior distributions and running the population likelihood from these using Markov-Chain Monte Carlo sampling methods. From each event I drew 10,000 samples, which made up the posterior distribution for each event. To make the computation easier, I used the natural logarithm of the likelihood instead of the likelihood — a simplified version of the likelihood is shown in Eq. 4. After some adjustments to correct errors, I ended up with results consistent with general relativity in my toy model again, where (0,0) for the mean and standard deviation was present in the plot. One issue was that the minimum effective sample size is quite small, so I worked on testing whether this is an issue that was biasing my results, and if so how to fix it.

$$\mathcal{L}(\{d\}|\Lambda) \propto \prod_{i}^{N} \sum_{k}^{n_s} \frac{\pi(\theta_i^k|\Lambda)}{\pi(\theta_i^k|\varnothing)} \tag{4}$$

After verifying the results with multiple methods, I determined that the probability of inconsistency with GR plotted to Fig. 3. I used MCMC and the analytical logarithm of the likelihood in order to verify this. The verification came from Ref. [13], which had similar results as shown in Fig. 4. That said, the curve present on Fig. 3 is still larger than expected, and my mentor and I are looking into possible reasons for this.

Once I had verified that my testing general relativity (TGR) code was working, I moved onto more parameters.

4.2 Building the Model

Binary black hole (BBH) mergers have the following astrophyscially interesting parameters: primary mass, mass ratio, two spin magnitudes, two tilts, and redshift. Each of these parameters had to be coded into my

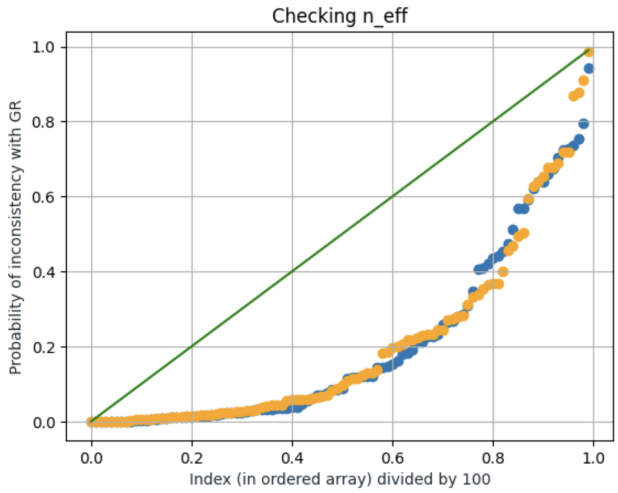

Figure 3: The cumulative distribution function (CDF) for inconsistency with general relativity. This was tested using two methods: blue is MCMC, while orange is the analytic equation for likelihood. The blue and orange line up very well. The green line is a plot of the line $y = x$.

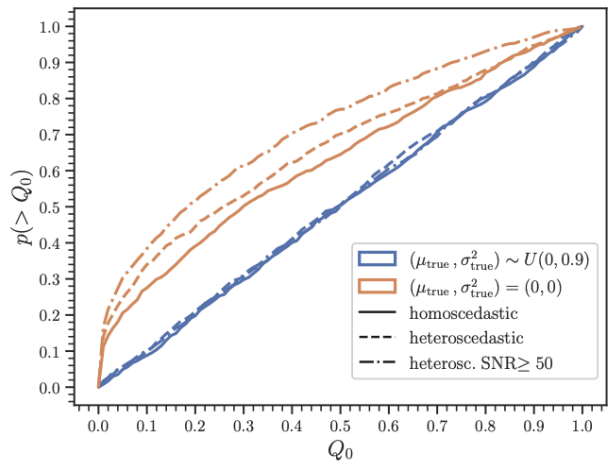

Figure 4: From Ref. [13]. The cumulative distribution function for consistency with general relativity using various methods. The blue lines approximately follow the line $y = x$. The orange lines are similar to those in Fig. 3, but are less curved.

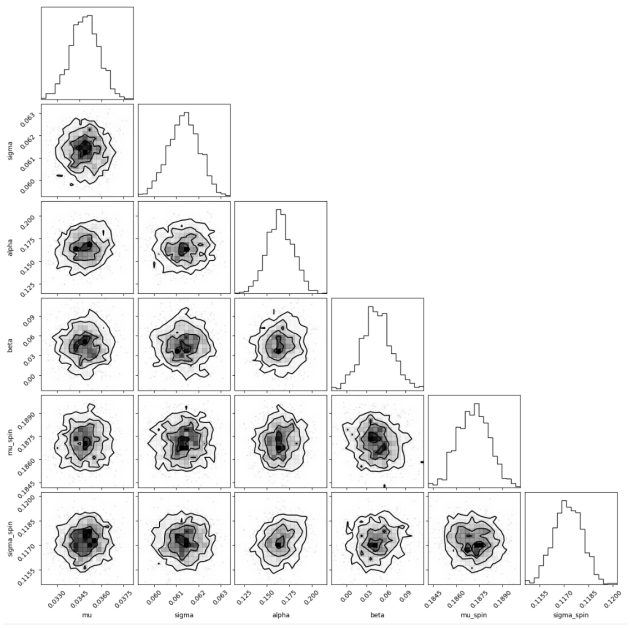

Figure 5: This is a corner plot from an MCMC run where a Gaussian mass distribution was injected, and then recovered with a powerlaw. Because of the missmatch, it is essentially nonsense. Consistency with general relativity is not established, since (0,0) is not present on the (mu, sigma) plot. Furthermore, the code was summed along the wrong axis, making the nonsense worse.

model appropriately. And each parameter had its own distribution that it was drawn from and analyzed with.

At first, I was just focused on making sure that the MCMC analysis would actually run. As a result, I coded up some crude parameter spaces. I ended up with a Gaussian primary mass distribution being analyzed with a Power law model — this resulted in nonsense, as shown in 5. Additional nonsense ensued because the likelihoods were being summed along the incorrect axis.

When my code would run with all eight parameters, I switched to making sure that my injected population was being recovered via MCMC correctly. This has been a challenge. For almost a week I was getting back incorrect values for powerlaw index, GR deviation, and more. It turns out that one list summation was done over the incorrect axis, which accounted for most of the issues. After this problem was (finally) resolved, there are still some problems, but things are getting better.

4.3 Model Analysis

I will be working on this for the last few weeks of the program. Looking forward, after I fix my code I will code up different population models for mass (keep spin magnitude, tilt, redshift, and deviation the same) and compare the results for each population model. Each population model will be injected and then recovered using each of the three population models. Finally, if there is time, I will use injected signals from the pipeline, instead of my crude estimates.

4.4 Challenges

Over the last few weeks, I have been expanding my code from just testing deviation parameters to also incorporating primary mass, mass ratio, spin magnitudes, cosine of tilts, and redshift. This required expanding my previous code and model, and adding other likelihood functions. One big challenge has been recovering the correct value for each parameter. Currently, I am injecting a Power law Evolution model for redshift with an index of 3, but again have not recovered that injected value.

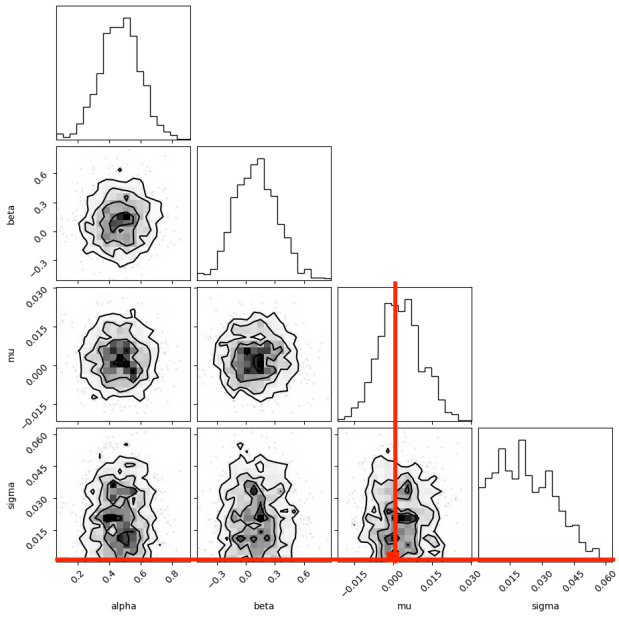

Figure 6: A corner plot for the simultaneous inference of testing general relativity (TGR) and the parameters that govern mass distribution (alpha, beta). The mass model is misspecified. However, consistency with general relativity is still recovered, as the red dot shows.

To try and fix these problems, I have broken down the code into different portions. Since I am treating each parameter as independent from the others, I am able to work through the analysis of each parameter separately. I have also been creating plots to try and figure out where the issue is occurring. I had some trouble with making sure values and posterior distributions were astrophysically reasonable, but after working to understand the *scipy stats truncated normal* function, I am now confident that everything is "in bounds." I also had an issue creating a POWER LAW for the mass to be drawn from, but after some Googling (and help from my mentor) everything is good to go.

I am unsure if my current issue is with the injected populations, the analysis models, or something else. I am working with my mentor to fix my code, and hope to have a breakthrough soon!

5 Results and Analysis

Plot 6 shows crucial results: it is not always possible to tell if a population model is misspecified even when TGR is simultaneously inferred, if deviation parameters are not correlated to the mass model. For this plot, a GAUSSIAN primary mass distribution was injected, but for the MCMC the model assumed a POWER LAW population. Thus, the population model was misspecified. Additionally, a GAUSSIAN TGR deviation distribution was injected and present in the analysis model. These GAUSSIAN deviation distributions were injected in such a way that consistency with general relativity was imbedded. In Fig. 6, consistency with GR was recovered, as the red cross shows. Thus, since a misspecified population model was used, but this did not affect the TGR results, it was not possible to tell after the fact that the mass model was misspecified when only relying on this method. And since we take a positive TGR test to mean that the population model was correct, we would have no way of knowing that this was a false positive without using other methods of checking.

In 7, consistency with GR was injected, as was a POWER LAW primary mass distribution. The primary mass was injected and analyzed with a POWER LAW distribution with an index of 2. As the yellow bar shows, 2 is a possible value for alpha, which refers to the POWER LAW index variable. Thus, the correct value has been recovered. Beta, the variable for q, the mass ratio, was injected to be 0. As the green bar shows, 0 is also the value that was recovered. Therefore, for masses, the hoped-for values were recovered.

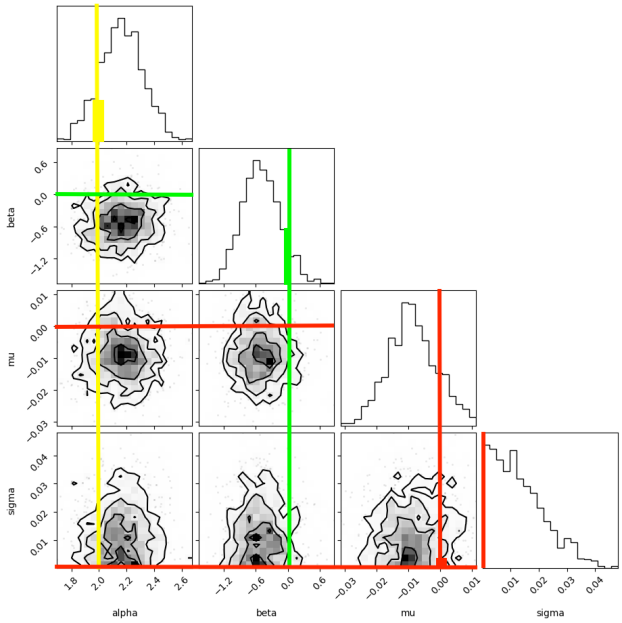

Figure 7: This figure shows a corner plot with mass POWER LAW index, mass ratio, and GR deviation. A primary mass distribution consistent with a POWER LAW of index 2 was injected and then attempted to recover this with the appropriate model.

The GR deviations were modeled as a GAUSSIAN, and analyzed with this prior. As the red cross shows, a deviation parameter with a mean and standard deviation of (0, 0) is present on the plot, inside the tail. Thus, consistency with GR has been recovered jointly with the correct mass distribution variable values, but it should be noted that mu and sigma, though correlated with each other, were not correlated with alpha and beta.

6 Conclusions

In this project, first, we looked at testing general relativity (TGR) by writing code to analyze GR deviation parameters, modeled as a Gaussian distribution. Next, we built out the model to jointly analyze GR deviation parameters and population characteristics. Finally, we injected and recovering each of three primary mass distribution population models. Due to time constraints, the detection fraction was ignored. Further studies should take this into account.

Misspecified population models have not yet led to biases in my project. Joint inference simultaneously testing general relativity and the parameter space is necessary. But as Fig. 6 shows, when there is a model mismatch between the injected population and the prior, consistency with GR can still be recovered. However, it is possible that inconsistency with GR could be obscured in a similar scenario. Further research into population models will help alleviate this worry.

Identifying a misspecified population model for testing GR is hard to tell from uncorrelated simultaneous inferrence. As Fig. 6 shows, realizing a misspecified population model is not always possible, especially when using a deviation distribution that is uncorrelated to mass. In this corner plot, a TRUNCATED GAUSSIAN population was injected, but the population was analyzed under the assumption that it followed a POWER LAW distribution. In one realization of this scenario, Fig. 6, consistency with GR was recovered. In another realization of this scenario, consistency with GR was not recovered, but due to the uncorrelated nature of this test, this was likely an outlier due to the small testing size. Consistency with GR was injected, as was a misspecified population model. This is worrisome; we cannot necessarily tell if we use a misspecified population model for testing GR — especially if we do not simultaneously infer GR — using just this method. However, this is only two of seven parameters in addition to the uncorrelated testing general relativity (TGR)

parameters. Thus, more parameters could further complicate this issue, and is what I am currently working on. But beyond the rest of my project, more research in this area is necessary.

Further research is necessary into population model choices. This project focused on mass distribution models, which evolve as more gravitational wave events are detected. However, the properly specified population distribution for other parameters, such as spin and tilt, should also be investigated. More work also must be done to check for model misspecification using multiple techniques in conjunction, including posterior predictive checks. Finally, other models need to be used for similar simultaneous inference of TGR and astrophysical population distribution.

7 Acknowledgments

This work was supported by the National Science Foundation Research Experience for Undergraduates (NSF REU) program, the LIGO Laboratory Summer Undergraduate Research Fellowship program (NSF LIGO), and the California Institute of Technology Student-Faculty Programs. I would like to thank my mentor, Ethan Payne, my fellow LIGO SURF-ers, my family and friends, and all of the faculty and mentors at the LIGO SURF program, especially Alan Weinstein and Jonah Kanner.

References

[1] Junaid Aasi, BP Abbott, Richard Abbott, Thomas Abbott, MR Abernathy, Kendall Ackley, Carl Adams, Thomas Adams, Paolo Addesso, RX Adhikari, et al. Advanced ligo. *Classical and quantum gravity*, 32(7):074001, 2015.

[2] Benjamin P Abbott, Richard Abbott, TDe Abbott, MR Abernathy, Fausto Acernese, Kendall Ackley, Carl Adams, Thomas Adams, Paolo Addesso, Rana X Adhikari, et al. Observation of gravitational waves from a binary black hole merger. *Physical review letters*, 116(6):061102, 2016.

[3] R Abbott, TD Abbott, F Acernese, K Ackley, C Adams, N Adhikari, RX Adhikari, VB Adya, C Affeldt, D Agarwal, et al. Population of merging compact binaries inferred using gravitational waves through gwtc-3. *Physical Review X*, 13(1):011048, 2023.

[4] Fet al Acernese, M Agathos, K Agatsuma, Damiano Aisa, N Allemandou, Aea Allocca, J Amarni, Pia Astone, G Balestri, G Ballardin, et al. Advanced virgo: a second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity*, 32(2):024001, 2014.

[5] Matthew R Adams, Neil J Cornish, and Tyson B Littenberg. Astrophysical model selection in gravitational wave astronomy. *Physical Review D*, 86(12):124032, 2012.

[6] KG Arun, Bala R Iyer, Bangalore Suryanarayana Sathyaprakash, and Pranesh A Sundararajan. Parameter estimation of inspiralling compact binaries using 3.5 post-newtonian gravitational wave phasing: The nonspinning case. *Physical Review D—Particles, Fields, Gravitation, and Cosmology*, 71(8):084008, 2005.

[7] Tessa Baker, Dimitrios Psaltis, and Constantinos Skordis. Linking tests of gravity on all scales: from the strong-field regime to cosmology. *The Astrophysical Journal*, 802(1):63, 2015.

[8] Luc Blanchet and Thibault Damour. Post-newtonian generation of gravitational waves. *Annales de l'IHP Physique théorique*, 1989.

[9] Albert Einstein and Emil Warburg. *Die Relativitätstheorie*. Springer, 1911.

[10] Davide Gerosa, Geraint Pratten, and Alberto Vecchio. Gravitational-wave selection effects using neural-network classifiers. *Physical Review D*, 102(10):103020, 2020.

[11] Maximiliano Isi, Katerina Chatziioannou, and Will M Farr. Hierarchical test of general relativity with gravitational waves. *Physical Review Letters*, 123(12):121101, 2019.

[12] Ryan Magee, Maximiliano Isi, Ethan Payne, Katerina Chatziioannou, Will M. Farr, Geraint Pratten, and Salvatore Vitale. Impact of selection biases on tests of general relativity with gravitational-wave inspirals. *Phys. Rev. D*, 109(2):023014, 2024.

[13] Costantino Pacilio, Davide Gerosa, and Swetha Bhagwat. Catalog variance of testing general relativity with gravitational-wave data. *Physical Review D*, 109(8):L081302, 2024.

[14] Ethan Payne, Maximiliano Isi, Katerina Chatziioannou, and Will M. Farr. Fortifying gravitational-wave tests of general relativity against astrophysical assumptions. *Phys. Rev. D*, 108(12):124060, 2023.

[15] William H Press and Kip S Thorne. Gravitational-wave astronomy. *Annual Review of Astronomy and Astrophysics*, 10(1):335–374, 1972.

[16] Dimitrios Psaltis, Colm Talbot, Ethan Payne, and Ilya Mandel. Probing the black hole metric: Black hole shadows and binary black-hole inspirals. *Physical Review D*, 103(10):104036, 2021.

[17] Bangalore Suryanarayana Sathyaprakash and SV Dhurandhar. Choice of filters for the detection of gravitational waves from coalescing binaries. *Physical Review D*, 44(12):3819, 1991.

[18] Hideyuki Tagoshi, Masaru Shibata, Takahiro Tanaka, and Misao Sasaki. Post-newtonian expansion of gravitational waves from a particle in circular orbit around a rotating black hole: Up to o (v 8) beyond the quadrupole formula. *Physical Review D*, 54(2):1439, 1996.

[19] 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:e010, 2019.