Gravitational wave tests of general relativity with the parameterized post-Einsteinian framework

Neil Cornish, Laura Sampson, Nicolás Yunes, and Frans Pretorius

Phys. Rev. D **84**, 062003 — Published 30 September 2011

DOI: 10.1103/PhysRevD.84.062003

Gravitational Wave Tests of General Relativity with the Parameterized Post-Einsteinian Framework

Neil Cornish,[1] Laura Sampson,[1] Nicolás Yunes,[1,2] and Frans Pretorius[3]

[1]*Department of Physics, Montana State University, Bozeman, MT 59717, USA.*
[2]*Department of Physics and MIT Kavli Institute, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.*
[3]*Department of Physics, Princeton University, Princeton, NJ 08544, USA.*

Gravitational wave astronomy has tremendous potential for studying extreme astrophysical phenomena and exploring fundamental physics. The waves produced by binary black hole mergers will provide a pristine environment in which to study strong field, dynamical gravity. Extracting detailed information about these systems requires accurate theoretical models of the gravitational wave signals. If gravity is not described by General Relativity, analyses that are based on waveforms derived from Einstein's field equations could result in parameter biases and a loss of detection efficiency. A new class of "parameterized post-Einsteinian" (ppE) waveforms has been proposed to cover this eventuality. Here we apply the ppE approach to simulated data from a network of advanced ground based interferometers (aLIGO/aVirgo) and from a future space based interferometer (LISA). Bayesian inference and model selection are used to investigate parameter biases, and to determine the level at which departures from general relativity can be detected. We find that in some cases the parameter biases from assuming the wrong theory can be severe. We also find that gravitational wave observations will beat the existing bounds on deviations from general relativity derived from the orbital decay of binary pulsars by a large margin across a wide swath of parameter space.

PACS numbers: 04.80.Cc,04.80.Nn,04.30.-w,04.50.Kd

I. INTRODUCTION

Einstein's theory of gravity has been subject to a wide array of experimental tests and has passed them all with flying colors [1]. None of these tests, however, has probed the strong field, dynamical regime that pertains to the final inspiral and merger of compact objects. The Hulse-Taylor binary pulsar PSR B1913+16 [2] and the double binary pulsar PSR J0737-3039A [3, 4] have provided convincing evidence for the existence of gravitational waves, and have served as unique laboratories to test general relativity (GR), but these objects have relatively small orbital velocities, $v/c \sim 10^{-3}$, a mere factor of 10 faster than the Earth's orbit around the Sun. The parameter space covered by black hole mergers, where orbital velocities $v/c \gg 10^{-3}$ and can approach $v/c \sim 0.7$, is currently *terra incognita* - Dragons may yet lurk there.

If not accounted for, the possibility that Einstein's theory of gravity may not correctly describe the production and propagation of gravitational waves could have dire consequences for gravitational wave astronomy. In the case of ground-based detectors, the detection of weak signals buried below the instrument noise requires accurate models of the gravitational waveforms. Errors in the modeling of these waveforms can lead to a loss in detection efficiency. When the signals are stronger, as will often be the case with space-based observations of black hole mergers, waveform templates will no longer be needed for detection, but a waveform model will be required to infer the physical parameters of the system, such as the masses and spins of the black holes, and the distance to the system. Waveform models based on an incorrect theory of gravity will lead to *fundamental bias* [5] in the recovered parameters. Because these waveforms would not accurately describe nature, the parameters that maximize the fit of such a waveform to data would not correspond to the true physical values of the system. This bias is distinct from that caused by imperfect modeling of GR, as explored in [6], as it reflects a fundamental lack of knowledge about the true nature of gravity, and not simply the use of inaccurate physical assumptions – see [5] for more details.

Turning the problem around, the discovery that Einstein's theory is flawed would be the greatest result to come out of gravitational wave astronomy [7]. This has served as the motivation for the development of a wide range of tests of GR that use gravitational wave observations. These tests can be broadly classified as "extrinsic" or "intrinsic". Extrinsic tests are possible when there is a concrete alternative theory, such as massive gravitons [8–14], or Brans-Dicke theory [9, 10, 14–16]. Intrinsic tests work within the confines of GR, and take the form of internal consistency checks, such as measuring the multipolar structure of the metric [17, 18], or multi-modal spectroscopy of BH inspiral and ringdown waveforms [19, 20]. These tests are valuable, but they do not cover the full spectrum of possibilities. The existing extrinsic tests are limited by the lack of viable alternative models, while the intrinsic tests do not so much test GR, as "test the nature of massive compact bodies within GR" (to quote [21]).

Convincing alternative models to GR are hard to find because none of the currently proposed alternatives can satisfy key criteria that physicists would like to require. On the observational front, one wishes that any GR al-

ternative passes all Solar System and binary pulsar tests with flying colors, only predicting deviations from GR in the strong-field regime, where tests are currently lacking. Many theories, such as Brans-Dicke theory [9, 10, 14–16], are heavily constrained by this requirement [1]. On the theoretical front, one would wish viable GR alternatives to lead to well-posed theories, with a positive definite Hamiltonian and free of instabilities. All perturbative string theory and loop quantum gravity low-energy effective theories [22, 23] currently lead to higher-derivative theories, which might violate this theoretical criteria.

The paucity of concrete alternative models to GR [24] has impacted other testing grounds, such as those based on solar system observations, or the aforementioned binary pulsar systems. In those instances the standard approach has been to develop models that parameterize a wide class of possible departures from GR - the parameterized post-Newtonian formalism [25–28] and the parameterized post-Keplerian formalism [29]. It is natural to adopt the same strategy when analyzing gravitational wave data, which leads to the parameterized post-Einsteinian (ppE) formalism introduced in Ref. [5].

To motivate this approach, consider the standard post-Newtonian (PN) expression for the dominant contribution to the stationary phase waveform describing the Fourier transform of the time-domain gravitational wave strain signal of the inspiral of two non-spinning black holes on circular orbits (see e.g. [10]):

$$\tilde{h}_{\rm GR}(f) = \sqrt{\frac{5}{24}}\frac{\mathcal{C}}{\pi^{2/3}}\mathcal{A}(f)\frac{\mathcal{M}^{5/6}}{D_L}e^{i\Psi(f)}\,, \qquad (1)$$

where f is frequency, $\mathcal{M} = \eta^{3/5}M$ is the chirp mass, $M = m_1 + m_2$ is the total mass, $\eta = m_1 m_2/M^2$ is the dimensionless, symmetric mass ratio, D_L is the luminosity distance and $\mathcal{C}$ is a geometric factor that depends on the relative orientation of the binary and the detector (its average for LISA is $\bar{\mathcal{C}} = 2/5$). The amplitude $\mathcal{A}(f)$ and phase $\Psi(f)$ are developed as a series in $u = \pi\mathcal{M}f = \eta^{3/5}v^3$, where v is the relative velocity between the two bodies [30] :

$$\mathcal{A}(f) = \sum_{k=0}^{\infty}\gamma_k u^{(2k-7)/6}\,. \qquad (2)$$

and

$$\Psi(f) = 2\pi f t_c - \Phi_c + \sum_{k=0}^{\infty}\left[\psi_k + \psi_{kl}\ln u\right]u^{(k-5)/3}\,. \qquad (3)$$

The coefficients $\gamma_k(\eta)$, $\psi_k(\eta)$ and $\psi_{kl}(\eta)$ are currently known up to $k = 7$ in the post-Newtonian expansion of GR.

In the simplest proposal of Yunes and Pretorius [5], the phase and amplitude are modified by only one ppE term each, but as pointed out by the authors there is no reason to believe that an alternative theory of gravity will predict such a restricted deviation from GR. In view of this, Yunes and Pretorius proposed four different parameterizations that differed in their level of complexity, one of the most complicated of which is (see Eq. (46) in [5])

$$\mathcal{A}(f) \to \left(1 + \sum_i \alpha_i u^{a_i}\right)A_{\rm GR}(f)\,,$$
$$\Psi(f) \to \left(\Psi_{\rm GR}(f) + \sum_i \beta_i u^{b_i}\right)\,, \qquad (4)$$

where the coefficients α_i and β_i may depend on the symmetric mass ratio η (and in more general cases, also on the spin angular momenta and the difference between the two masses) and $A_{\rm GR}$ and $\Psi_{\rm GR}$ are the standard expressions in Eqs. (2) and (3). This is in essence the ppE approach.

In an earlier study, Arun *et.al.* [31–33] considered what can now be interpreted as a restricted version of the ppE formalism in which the exponents a_i and b_i are required to match those found in GR. This amounts to asking how well the standard PN expansion coefficients could be recovered from gravitational wave observations. They also developed internal self-consistency checks based on the observation that each coefficient $\psi_k(\eta)$ provides an independent estimate of the mass ratio η. While interesting, these tests are limited in scope as few of the well known alternative theories of gravity (Brans-Dicke [9, 10, 14–16], Massive Graviton [8–14], Chern-Simons [22, 34–37], Variable G [38], TeVeS [39]*etc.*) have corrections with exponents a_i and b_i that match those of GR [5]. The full ppE formalism allows us to look for a much wider and realistic set of possible departures from GR.

Our goal here is to study how the ppE formalism can be used to search for waveform deviations from GR using data from the next generation of ground based interferometers (aLIGO/aVirgo) and future space based interferometers (*e.g.* LISA). Bayesian model selection is used to determine the level at which departures from GR can be detected (See Ref.[40] for a related study that uses Bayesian inference to study constraints on Massive Graviton theories). Advanced Markov Chain Monte Carlo (MCMC) techniques are used to map out the posterior distributions for the models under consideration. From these distributions, we are able to quantify the degree of fundamental bias in parameter extraction, and in particular, if the fundamental bias can be significant in situations where there is no clear indication that there are departures from GR.

Recently, Pozzo et.al. [37] performed a similar study that applied Bayesian model selection to estimate the bounds that could be placed on massive graviton theory. As such, their work is a sub-case of the ppE framework, *i.e.* a particular choice of (b, β). Their implementation differed from ours in that they used Nested Sampling while we used MCMC techniques, but as we will show, our results are in agreement with theirs for the relevant sub-case.

We find that gravitational wave observations will allow us to extend the existing bounds derived from pulsar orbital decay [41] into the region of parameter space that covers strong field departures from GR ($a_i > 0$ and $b_i > -5/3$) (see Fig. 2–1 in Sec. IV A). As expected, we find that the strength of the bounds on the ppE parameters are inversely proportional to the signal-to-noise ratio (SNR), and the extent to which deviations between GR templates and non-GR signals can be detected (the departure of the "fitting factor" from unity) scales as $1/\mathrm{SNR}^2$. The logarithm of the odds ratio used to decide if a signal is described by GR or some alternative theory follows the same $1/\mathrm{SNR}^2$ scaling. A more surprising result is the possibility of "stealth bias" whereby the parameters recovered using GR templates can be significantly biased even when the odds ratio shows no clear preference for adopting an alternative theory of gravity.

The remainder of this paper is organized as follows. Section II introduces the analysis framework in more detail, including a discussion of the waveform model, noise spectrum, and Bayesian tools used. Section III describes in detail the computational techniques used to to implement the analysis. Section IV presents the results of our analysis. Section V closes with a discussion of how our results might change as the degree of realism is increased, and identifies key questions to be addressed in future work. Throughout this paper we use geometric units with $G = c = 1$.

II. ANALYSIS FRAMEWORK

A. Bayesian Inference

Questions of model selection and parameter biases can be addressed very naturally in the framework of Bayesian inference. This approach is now well established in the field of gravitational wave data analysis, as are the tools used to carry out the analysis. To avoid unnecessary repetition, we will focus on those aspects of the analysis that are new, and refer the reader to Ref. [42] for a detailed description of the techniques used.

We are interested in comparing the hypothesis $\mathcal{H}_0$ that gravity is described by GR with the hypothesis $\mathcal{H}_1$ that gravity is described by an alternative theory belonging to the ppE class. Here we are dealing with nested hypotheses, as the ppE models include GR as a limiting case. When new data d becomes available, our prior belief $p(\mathcal{H})$ in hypothesis $\mathcal{H}$ is updated to give the posterior belief $p(\mathcal{H}|d)$. Bayes' theorem tells us that

$$p(\mathcal{H}|d) = \frac{p(d|\mathcal{H})p(\mathcal{H})}{p(d)}\,, \tag{5}$$

where $p(d|\mathcal{H})$ is the (marginal) likelihood of observing the data d if the hypothesis holds, and $p(d)$ is a normalization constant. For hypotheses described by models with continuous parameters, the likelihood $p(d|\mathcal{H})$ is found by marginalizing the likelihood $p(d|\vec{\theta},\mathcal{H})$ of observing data d for model parameters $\vec{\theta}$:

$$p(d|\mathcal{H}) = \int d\vec{\theta}\; p(\vec{\theta},\mathcal{H}) p(d|\vec{\theta},\mathcal{H})\,, \tag{6}$$

where $p(\vec{\theta},\mathcal{H})$ is the prior distribution of the parameters. The marginalized likelihood, $p(d|\mathcal{H})$, is also known as the evidence for a given model. Hypotheses are compared by computing the odds ratio, or Bayes factor:

$$BF = \mathcal{O}_{1,0} \equiv \frac{p(\mathcal{H}_1|d)}{p(\mathcal{H}_0|d)} = \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}\frac{p(d|\mathcal{H}_1)}{p(d|\mathcal{H}_0)}\,, \tag{7}$$

which gives the "betting odds" of $\mathcal{H}_1$ being a better description of Nature than $\mathcal{H}_0$. The normalization constant $p(d)$ cancels in the odds-ratio. The prior odds ratio $p(\mathcal{H}_1)/p(\mathcal{H}_0)$ gets updated by the likelihood ratio, $p(d|\mathcal{H}_1)/p(d|\mathcal{H}_0)$, which is also known as the evidence ratio. In Bayesian analysis "today's posterior is tomorrow's prior" [43], and $p(\mathcal{H}|d)$ is used in place of $p(\mathcal{H})$ in subsequent analyses. While a single black hole inspiral event may not yield strong evidence for a departure from GR, several such observations can be combined to make a more compelling case.

In addition to simply detecting deviations from GR, we are also interested in studying how departures from GR might affect parameter estimation. This can be assessed by looking at the posterior distribution function $p(\vec{\theta}|d,\mathcal{H})$, which describes the probability distribution for parameters $\vec{\theta}$ under the assumption that the signals are described by model $\mathcal{H}$ given data d. The posterior distribution is given by the product of the prior and the likelihood, normalized by the evidence:

$$p(\vec{\theta}|d,\mathcal{H}) = \frac{p(\vec{\theta},\mathcal{H})p(d|\vec{\theta},\mathcal{H})}{p(d|\mathcal{H})}\,. \tag{8}$$

Once the prior distribution and the likelihood function have been specified we are left with the purely mechanical task of computing the posterior distributions and odds ratio for competing hypotheses.

B. Waveform Model

The original ppE waveforms were for non-spinning, equal mass binaries in quasi-circular orbits, and included a description of the dominant harmonic through inspiral, merger and ringdown. In the current analysis we restrict our attention to the inspiral portion of the waveform, but our signals come from unequal mass binaries. We have examined the generalization of the ppE framework for unequal mass systems, and find that for a single detection it is indistinguishable from the equal mass case. Including multiple detectors, and the merger and ringdown phases, which increase the signal-to-noise ratio, can help break parameter degeneracies that exist in the inspiral phase, but these benefits come at the cost of having to consider

Theory	a	α	b	β
Brans-Dicke [9, 10, 14–16]	–	0	-7/3	β
Parity-Violation [22, 34–37]	1	α	0	–
Variable $G(t)$ [38]	-8/3	α	-13/3	β
Massive Graviton [8–14]	–	0	-1	β
Quadratic Curvature [23, 44]	–	0	-1/3	β
Extra Dimensions [45]	–	0	-13/3	β
Dynamical Chern-Simons [46]	+3	α	+4/3	β

TABLE I: Leading ppE corrections in several alternative theories of gravity (GR corresponds to $\alpha = \beta = 0$). In dynamical Chern-Simons gravity, (α, β) are proportional to the spin-orbital angular momentum coupling. For non-spinning binaries, the last row would simplify to $(\alpha, \beta) = (0, 0)$, but we include it here for completeness.

additional ppE parameters. We will consider this in a separate publication.

In the stationary phase approximation, our ppE waveforms are parameterized as follows

$$\tilde{h}(f) = \tilde{h}_{\rm GR}(f)\,[1 + \alpha\, u^a]\, e^{i\beta\, u^b} \quad f < f_{\rm max}\,, \qquad (9)$$

where (α, a) are amplitude ppE parameters and (β, b) are phase ppE parameters. As noted previously, both α and β can depend on the spin angular momenta and mass difference of the two bodies, as well as the symmetric mass ratio of the system. With a single detection, however, these dependencies are impossible to determine, and so we defer an analysis of them to future work. Here $\tilde{h}_{\rm GR}(f)$ is the usual GR waveform quoted in Eq. (1). We set the maximum frequency cut-off at twice the innermost stable circular orbit frequency of a system described by GR. A more consistent choice would be to use the minimum of the ppE energy function, but the results were found to be fairly insensitive to the choice of $f_{\rm max}$. To simplify the analysis we restrict our attention to the lowest PN order in the amplitude of Eq. (2), setting $\gamma_k = 0$ for $k > 0$. The GR phase terms in Eq. (3) are kept out to $k = 7$. Furthermore, we limit the range of the ppE parameters a and b to not be greater than these corresponding highest order PN terms, namely $a < 2/3$ and $b < 1$. [1]

As discussed in the Introduction, the ppE framework introduces i sets of ppE theory parameters $(\alpha_i, a_i, \beta_i, b_i)$ that modify the amplitude and phase, but we here work to leading order, keeping only the $i = 0$ set. This approach will tend to over-estimate how well the ppE parameters $(\alpha_0, a_0, \beta_0, b_0) \equiv (\alpha, a, \beta, b)$ can be constrained by the data. A better approach, which we intend to pursue in future studies, is to marginalize over the higher order terms.

Table I lists the leading ppE corrections that have been computed for several alternative theories of gravity. Generally, the exponents a and b are pure numbers fixed by the theory, while the amplitudes α and β are free parameters that relate to the unknown coupling strengths of the modified/additional gravitational degrees of freedom.

C. Instrument Response

The aLIGO/aVirgo analysis was performed using simulated data from the 4 km Hanford and Livingston detectors and the 3 km Virgo detector. The time delays between the sites and the antenna beam patterns were computed using the expression quoted in Ref. [47]. Since the detectors barely move relative to the source during the time the signal is in-band, the antenna patterns can be treated as fixed and the time delays Δt between the sites can be inserted as phase shifts of the form $2\pi f \Delta t$. For the instrument noise spectral density, we assumed all three instruments were operating in a wide-band configuration with

$$S_n(f) = 10^{-49}\left(x^{-4.14} - 5x^{-2} + 111\frac{(2 - 2x^2 + x^4)}{2 + x^2}\right), \qquad (10)$$

and $x = (f/215{\rm Hz})$.

The space based (LISA) analysis was performed using the A and E Time Delay Interferometry channels [48] in the low frequency approximation [49, 50]. It is known that this approximation can lead to biases in some of the recovered parameters, such as polarization and inclination angles. This, however, is an example of a modeling bias introduced by inaccurate physical assumptions, and not of a fundamental bias resulting from incomplete knowledge of the theory describing gravity. In our current study the modeling bias is avoided by using the same low frequency response model to produce the simulated data and to perform the analysis.

In contrast to the ground based detectors, the signals seen by LISA are in-band for an extended period of time, and the motion of the detector needs to be taken into account. The time dependent phase delay between the detector and the barycenter and the time dependent antenna pattern functions are put into a form that can be used with the stationary phase approximation waveforms by mapping between time and frequency using $t(f) = (d\Phi/df)/2\pi$. Details of this procedure can be found in Ref. [51]. The noise spectral density model includes instrument noise and an estimate of the foreground confusion noise from unresolved galactic binaries, matching those quoted in Ref. [52].

[1] It is certainly conceivable that the *leading order* deviation arising from an alternative theory comes in at some high order, and has a much larger magnitude than the nearest exponent term in the PN expansion. Thus it is not *a priori* inconsistent to allow a range of exponents outside of that of the PN expansion used for the GR signal in the ppE waveforms, though this would require more complicated priors on the amplitudes, and so for simplicity in this study we restrict to the stated range.

D. Likelihood Function

Under the assumption that the noise is Gaussian, the likelihood that the data d would arise from a signal with parameters $\vec{\theta}$ is given by

$$p(d|\vec{\theta}) = Ce^{-\chi^2(\vec{\theta})/2}\,, \tag{11}$$

where C is a constant that depends on the noise level. Here

$$\chi^2(\vec{\theta}) = (d-h(\vec{\theta})|d-h(\vec{\theta}))\,, \tag{12}$$

and the brackets denote the noise weighted inner product

$$(a|b) = 2\int \frac{\tilde{a}(f)\tilde{b}^*(f)+\tilde{a}^*(f)\tilde{b}(f)}{S_n(f)}\,df\,. \tag{13}$$

For a theoretical study that assumes the noise is Gaussian and has a known spectrum, there is no need to add simulated noise to the data - the appropriate spread in the parameter values and overall topography of the likelihood surface follow from the functional form of the signal and the noise weighting in Eq. (13). Thus, we may write $d = h(\vec{\theta}')$ where $\vec{\theta}'$ are the true source parameters.

Many alternative theories of gravity predict the existence of polarization states beyond the usual "plus" and "cross" polarizations of GR that complicate the treatment of the instrument response, whose Fourier transform is

$$\begin{aligned}\tilde{h}_{inst} &= F_+\tilde{h}_+ + F_\times\tilde{h}_\times + F_S\tilde{h}_S \\ &\quad + F_L\tilde{h}_L + F_{V1}\tilde{h}_{V1} + F_{V2}\tilde{h}_{V2}\,,\end{aligned} \tag{14}$$

Here $\tilde{h}_{+\times}$ are the usual plus and cross-polarization states, $\tilde{h}_S$ is a scalar (breathing) mode, $\tilde{h}_L$ is a scalar longitudinal model and $\tilde{h}_{V1,V2}$ are two vectorial modes [53], while the F's are the detector antenna patterns [54], which depend on the sky location (θ,ϕ) and polarization angle ψ of the signal.

To simplify the analysis we assume the usual polarization content for a circular binary viewed at inclination angle ι and neglect the other contributions:

$$\begin{aligned}\tilde{h}_+ &= (1+\cos^2\iota)\Re(\tilde{h}) + 2\cos\iota\,\Im(\tilde{h})\,, \\ \tilde{h}_\times &= (1+\cos^2\iota)\Im(\tilde{h}) - 2\cos\iota\,\Re(\tilde{h})\,.\end{aligned} \tag{15}$$

In other words, we have assumed that the signal in the detector has the form $\tilde{s}(f) = F(\theta,\phi,\psi,\iota)\;\tilde{h}(f)$ with the function $F(\theta,\phi,\psi,\iota)$ given by the usual GR expression. If additional polarization states were present, this assumption would result in a reduction in detection efficiency and biases in the recovery of the extrinsic parameters (θ,ϕ,ψ,ι).

The justification for making this simplification is that we are primarily interested in how well the intrinsic parameters (α,a,β,b) can be constrained, and we expect these parameters to be only weakly correlated with the extrinsic parameters. The presence of additional polarization states will provide an additional handle on detecting departures to GR [55–57], and we plan to explore this possibility in the context of the ppE formalism in future work.

Defining $A_+ = |F^+\tilde{h}_+(f;\vec{\theta})|$ and $A_\times = |F^\times\tilde{h}_\times(f;\vec{\theta})|$, and similarly for $\vec{\theta}'$, the chi-squared goodness of fit of Eq. (12) can be re-expressed as

$$\begin{aligned}\chi^2(\vec{\theta}) &= 4\int\frac{df}{S_n(f)}\Big[A_+^2 + A_\times^2 + A'^2_{\ +} + A'^2_{\ \times} \\ &\quad - 2(A_+A'_+ + A_\times A'_\times)\cos\Delta\Psi \\ &\quad - 2(A_\times A'_+ - A_+A'_\times)\sin\Delta\Psi\Big]\,,\end{aligned} \tag{16}$$

where $\Delta\Psi = \Psi(\vec{\theta}) - \Psi(\vec{\theta}')$. As noted in Ref. [58], in the regime of interested where χ^2 is small, all the terms in the above integrand are slowly varying functions of frequency, so it is possible to compute the likelihood very cheaply using an adaptive integrator.

E. Priors

As we shall see, the choice of priors on the ppE parameters has a significant effect on the results, especially when it comes to model selection. The natural priors on the ppE parameters are those that come from existing data on binary pulsars, but these turn out to range from very restrictive to wide open depending on what sector of the ppE parameter space is being examined. To simplify the analysis we adopt uniform priors for the ppE parameters and seek to determine where direct GW observations would prove more constraining than the existing binary pulsar observations.

The priors on the exponents a and b are taken to be uniform across the ranges $a\in[-3,2/3]$ and $b\in[-4.5,1]$. The upper end of the range is chosen so that the ppE corrections to the amplitude and the phase do not go to higher order in the expansion parameter u than the post-Newtonian order of the reference GR waveforms. The lower end of the range is chosen to cover all known alternative theories, though in any case, the low end of the range turns out to be far better constrained by binary pulsar observations.

The priors on α, and β are more difficult to set. Lacking any theoretical or experimental guidance, we assign uniform priors for the amplitudes $\alpha,\beta\in[-1000,1000]$. The range in α,β is set such that it is sufficiently large that at the most positive end of the prior ranges on a,b, the exploration of possible values of α,β is not restricted by prior bounds. That is, even in the most poorly-constrained region of the ppE parameter-space, the constraints are not due to an overly restrictive prior.

The parameters used to describe the black hole binary were the log of the total mass M and the log of the chirp mass $\mathcal{M}$, the sky location $(\cos\theta,\phi)$, orbital plane orientation $(\cos\theta_L,\phi_L)$, merger phase Φ_c, merger time t_c, and luminosity distance D_L. The angular parameters

are taken to have uniform priors that covered their natural range. For the aLIGO studies, we assign uniform priors: $\ln(M/M_\odot) \in [1.3, 5.3]$; $\ln(\mathcal{M}/M_\odot) \in [0.55, 4.5]$; $t_c/\mathrm{s} \in [1, 16]$; $D_L/\mathrm{Mpc} \in [0.1, 10^4]$. For the LISA studies, we assign uniform priors: $\ln(M/M_\odot) \in [12.2, 16.8]$; $\ln(\mathcal{M}/M_\odot) \in [11.4, 16]$; $t_c/\mathrm{s} \in [1, 6 \times 10^7]$; $D_L/\mathrm{Gpc} \in [0.01, 1000]$. While we could use more physically motivated priors for the black hole parameters (such as distance priors that scaled with D_L^2), these choices have little effect on the model comparison between GR and ppE waveforms.

III. COMPUTATIONAL TECHNIQUES

Posterior distribution functions for the alternative hypotheses were computed using the Markov Chain Monte Carlo (MCMC) implementation described in Ref. [42], additionally enhanced by adding Differential Evolution [59, 60] to the mix of proposal distributions. The evidence for the competing hypotheses was calculated using the volume tessellation algorithm [61] and cross-checked using thermodynamic integration [62].

The ppE waveforms introduce a number of complications that make parameter estimation and model selection challenging. These complications can be seen when using the quadratic Fisher matrix approximation $\Gamma_{ij} = -\partial_i \partial_j \langle \ln p(\vec{\theta}|d) \rangle$ to estimate the parameter correlation matrix $C^{ij} = \langle \Delta\theta^i \Delta\theta^j \rangle \approx \Gamma^{-1}_{ij}$. When evaluated at the GR limit point $(\alpha, \beta) = (0, 0)$, the quadratic approximation to the Fisher matrix is singular, and it is necessary to include higher order derivatives to obtain a finite covariance matrix. The situation is worse when $a = 0$, as then α is fully degenerate with D_L, and when $b = 0$, as then β is fully degenerate with Φ_c. Partial degeneracies also exist whenever the a or b exponents match the exponents found in the post-Newtonian expansion of GR.

The various degeneracies and parameter correlations do not constitute a fundamental problem with the ppE formalism, but they do demand that we use very effective MCMC samplers that are able to fully explore the parameter space. The algorithm described in Ref. [42] uses parallel tempering with multiple, coupled chains, with each chain exploring a tempered likelihood surface $p(d|\vec{\theta})^{1/T}$. The high temperature chains explore more widely, and can communicate this information via parameter exchange to the $T = 1$ chain that is used for parameter estimation. Parallel tempering helps the Markov chains explore complicated posterior distributions, but convergence can still be slow if the proposal distributions are not well chosen.

The ultimate proposal distribution is the posterior distribution itself, but since that is unavailable in advance, we have to make do with approximations to this ideal. The covariance matrix C_{ij} provides a local approximation to the posterior distribution. It can be estimated semi-analytically using the Fisher information matrix, or more directly from the recent past history of the Markov chain itself. The latter approach introduces hysteresis into the chains, but so long as the covariance matrix is only updated occasionally the chains are asymptotically Markovian. In the present study, we continued to use the Fisher matrix based proposal distributions described in Ref. [42], but found that the convergence time of the chains was very long until we augmented these techniques with proposals based on Differential Evolution.

Differential Evolution (DE) provides an approximation to the posterior distribution based on the past history of the chains. Unlike methods based on the covariance matrix, DE works extremely well with highly correlated parameters. In its original formulation, DE [59] was designed to work with a population of N parallel chains (all with temperature $T = 1$). The idea is very simple and can be coded in a few lines: Chain i is updated by randomly selecting chains j and k with $j \neq k \neq i$, forming the difference vector $\vec{\theta}_j - \vec{\theta}_k$ and proposing the move

$$\vec{y}_i = \vec{\theta}_i + \gamma(\vec{\theta}_j - \vec{\theta}_)\,. \tag{17}$$

For D-dimensional multivariate normal distributions, the optimal choice for the scaling is $\gamma = 2.38/\sqrt{2D}$. Since the difference vector points along the D-dimensional error ellipse, the jumps are usually "in the right direction." It is a good idea to occasionally (e.g. 10% of the time) propose jumps with $\gamma = 1$, which act as mode-hopping jumps when the samples (j, k) come from separate modes of the posterior.

The original formulation of DE is not very practical since it requires $N > 2D$ parallel chains for each rung on the temperature ladder. A more economical approach is to use samples from the past history of each chain [60]. It can be shown that this approach is asymptotically Markovian in the limit as one uses the full past history of the chain. We have implemented a variant of the DE algorithm as follows:

- Create a history array for each parallel chain. Initialize a counter M. Store every 10^{th} sample in the history array and add to the counter each time a sample is added. DE moves are more effective if points during the burn-in phase of the search are discarded from the history array.

- Draw two samples from the history array: $j \in [1, M]$, $k \in [1, M]$ and repeat if $k = j$. Propose the move to

$$\vec{y} = \vec{\theta} + \gamma(\vec{\theta}_j - \vec{\theta}_k)\,. \tag{18}$$

Here we draw γ from a Gaussian of width $2.38/\sqrt{2D}$ for 90% of the DE updates and set $\gamma = 1$ for the rest.

The standard DE proposal seeks to update all the parameters at once, but it is often more effective to update smaller sub-blocks of highly correlated parameters. We did this in $\sim 30\%$ of the DE proposals.

The fraction of all proposed moves that use DE is a tunable parameter. We used 60% DE proposals, 30% Fisher matrix based proposals, 5% draws from the prior distribution and 5% uniform draws with width $\sim 10^{-6}$ of the prior range. Notice that even though the Fisher matrix might be singular in certain regions of the parameter manifold, one can still propose jumps with it. In those regions, the proposed jumps will not lead to a better likelihood, and will simply be rejected.

With the mix of proposal distributions described above, and using $\sim$ 10 parallel chains geometrically spaced with $T_{i+1} = 1.3T_i$, our MCMC implementation converges quickly to a stationary distribution. The chains are typically run for 500,000 samples, with the first 100,000 discarded based on a conservative estimate of the burn-in length.

The marginal likelihood, or evidence, $p(d|\mathcal{H})$ is computed using independent codes supplied by Martin Weinberg and Will Farr that implement Weinberg's volume tessellation algorithm (VTA) [61]. The VTA uses the posterior samples from the Markov chain to assign probability to a partition of the sample space and performs the marginal likelihood integral directly. The samples are partitioned using a kd-tree, and volume elements containing m samples (we use $m = 32$ or $m = 64$) are used to provide a discrete approximation to the integral in Eq. (6). The integrand in each volume element is approximated using either the average posterior density (Farr's code) or the median posterior density (Weinberg's code) of the m samples in the volume element. The VTA is applied to a sub-sample of the full chain, and by repeating the calculation with different subsamples in a process called bootstrapping, it is possible to compute statistical errors bars on the evidence caused by using finite length Markov chains.

There is a trade-off in the choice of the boxing number m, with large values of m providing better estimates of the average or mean posterior density in each cell, and small values of m providing better resolution to features in the posterior. In our experience, the statistical error found from the bootstrap procedure is usually smaller than the systematic error that we estimate by varying the boxing size from $m = 16$ to $m = 64$.

As a cross check we applied thermodynamic integration [62] to a few test cases using the implementation described in the appendix of Ref. [63]. In tests on distributions where the evidence can be calculated analytically, such as multi-variate Gaussians, we found that thermodynamic integration gave more accurate results. On the other hand, thermodynamic integration requires many more chains (upwards of 50 for the ppE studies) and a careful tuning of the temperature ladder in order to resolve the integrand. This tuning necessitates a long pilot run, or complicated adaptive tuning of the temperature ladder. So while thermodynamic integration produces more accurate results, it requires careful tuning and is far more computationally intensive. Based on the tests described in Appendix A, we estimate that the errors in the (natural) log Bayes factors computed using the VTA algorithm are of order ± 2.

IV. RESULTS

We explore a range of questions concerning the application of the ppE formalism to detecting departures from GR using gravitational wave observations from both LISA and the three-detector network of aLIGO/aVIRGO interferometers. First, we derive simple estimates of how well the ppE parameters can be constrained by gravitational wave data by using ppE templates to detect GR signal injections. The spread in the recovered ppE parameters establishes the range that is consistent with GR, and values outside of this range would point towards a departure from GR. We then compare these simple bounds to the more rigorous (and computationally expensive) bounds that can be derived from Bayesian model selection. Finally, we explore how searching for gravitational waves using GR templates can lead to biases in the recovered parameters if Nature is described by an alternative theory of gravity. We find that these biases can become significant before the evidence disfavors GR.

A. Cheap Bounds and Comparison with Pulsar Bounds

The first question we seek to address in this paper is how well the four ppE parameters (α, a, β, b) can be determined. One approach to answering this question is to examine how a search using ppE templates would look when used to characterize a signal that is consistent with GR. That is, if the signal observed is described by GR to the given level of accuracy of our detectors, what values for the ppE parameters will be recovered from a search with ppE templates? Because we know that in GR the values of α and β should be 0 for all values of a and b, we wish to determine the typical spread in the recovered value of (α, β), centered at zero. The standard deviation in this spread then gives us a constraint on the magnitude of the deviation that is still consistent with observations, ie. deviations that are 'inside our observational error bars.'

Cheap constraints will be defined as the (3σ)-bound on the posterior distribution of ppE parameters α or β, while keeping a or b fixed and marginalizing over all other system parameters. These bounds are 'cheap' because we do not have to re-run a search with pure GR templates and then compute the evidence, via integration of the posterior, to compute the Bayes factor (the latter is particularly computationally expensive). These cheap bounds are similar to constraints studied by looking at the (α, α) or (β, β) elements of the variance-covariance matrix. Our cheap constraints, however, are 3σ ones, in contrast to the more standard 1σ bounds quoted from variance-covariance matrix studies.

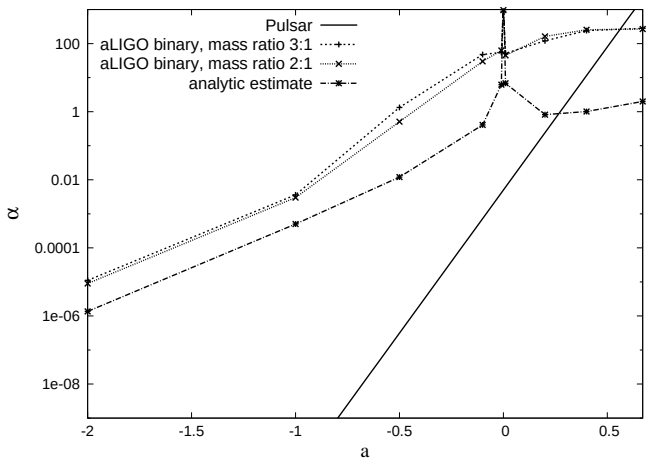

FIG. 1: UPPER PANEL:Bounds on α for different values of a, found using two different aLIGO sources. The two sources had different mass ratios, total masses, and sky locations, but were scaled to have a network SNR of 20. The rough estimate for the α bound from equation (20) is shown for comparison. Also included is the bound on α derived from the golden pulsar (PSR J0737-3039) data.
LOWER PANEL: Bounds on α for different values of a, found using two LISA sources at redshift $z = 1$ and $z = 3$. The pulsar bound is shown for comparison. The sources injected had the same parameters as those from the lower panel in Figure 2 .

Rough analytic estimates for the bounds on (α, β) can be derived by considering how the the ppE terms affect the overall amplitude $\mathcal{A}$ and phase Ψ of the signal:

$$\begin{aligned} \Delta \ln \mathcal{A} &\simeq \alpha(u^a_{\rm min} - u^a_{\rm max}) \\ \Delta\Psi &\simeq \beta(u^b_{\rm min} - u^b_{\rm max}). \end{aligned} \tag{19}$$

Here $u_{\rm min}$ and $u_{\rm max}$ are the minimum and maximum values of the u parameter. For the aLIGO sources $u_{\rm min} \sim 3 \times 10^{-3}$, while for the LISA sources $u_{\rm min} \sim 10^{-3}$. The ISCO cut-off in the frequency evolution sets $u_{\rm max} \sim 3 \times 10^{-2}$ for moderate mass ratios. Combining these estimates with a crude Fisher matrix estimate for how well the amplitude and phase are constrained: $\Delta \ln \mathcal{A} \sim \Delta\Psi \sim 1/{\rm SNR}$ yields the 3σ bounds

$$\begin{aligned} |\alpha| &\leq \frac{3}{{\rm SNR}\,|u^a_{\rm min} - u^a_{\rm max}|} \\ |\beta| &\leq \frac{3}{{\rm SNR}\,|u^b_{\rm min} - u^b_{\rm max}|}. \end{aligned} \tag{20}$$

These estimates reproduce the overall shape of the exclusion plots in the (a, α) and (b, β) planes, but they tend to over estimate the strength of the bounds as they do not take into account covariances with other parameters. The α bounds turn out to be a factor of ~ 10 weaker due to covariances between α and the distance and inclination, while the bounds on β come out a factor of ~ 100 weaker due to covariances between β and the chirp mass and mass ratio.

Figures 1 and 2 show these cheap constraints on the ppE amplitude parameters as a function of the exponents a and b for a variety of aLIGO/aVirgo and LISA detections. To generate these plots, we injected GR signals and then searched on them with ppE templates. For each search, either a or b was held fixed at a specific value, while the other three ppE parameters (and all other system parameters) were allowed to vary. We then calculated the standard deviation of the posterior distribution of the relevant amplitude parameter α or β, and used three times this value as the cheap bound shown on the plots.

A natural course of action might seem to be the following: marginalize over a and b as well, instead of keeping them fixed, and calculate constraints on α and β this way. Looking at Figures 2 and 1, however, show why this analysis would not be particularly helpful. The uncertainty in α and β is so much higher at the positive ends of the prior ranges on a and b than at the negative ends that the Markov chains would spend almost all of their iterations exploring this area of parameter space if a and b were allowed to change. Thus, to get any knowledge about the uncertainties in α and β for negative values of a and b, we need to fix a and b.

The aLIGO systems were chosen to have network SNR = 20, but different masses and sky locations. One system had masses $m_1 = 6M_\odot$, $m_2 = 18M_\odot$ ($\eta = 0.1875$), $D_L = 258$ Mpc, while the other had $m_1 = 6M_\odot$, $m_2 = 12M_\odot$ ($\eta = 0.2222$), $D_L = 462$ Mpc. The LISA sources were at different redshifts and had different masses and SNRs. The system at redshift $z = 1$ had $m_1 = 1 \times 10^6 M_\odot$, $m_2 = 3 \times 10^6 M_\odot$ ($\eta = 0.1875$) and SNR = 879, while the system at redshift $z = 3$ had $m_1 = 2 \times 10^6 M_\odot$, $m_2 = 3 \times 10^6 M_\odot$ ($\eta = 0.24$) and SNR = 280.

Figures 1-2 are 'exclusion' plots, showing the region (above the curves) which could be excluded with a 99.73% confidence. These figures also plot the bound on the ppE parameters that have already been achieved through analysis of the 'golden pulsar' system, PSR J0737-3039 [41]. Observe that for the amplitude parameter α, the pulsar bounds beat the aLIGO bounds through almost the entire range of a; LISA can improve upon the pulsar bounds for $a > 0$. For the phase parameter β, however, both aLIGO and LISA do better than the pulsar analysis through a significant portion of the range. As expected, gravitational wave observations tend to do better in the strong field regime, corresponding to high post-Newtonian terms ($b > -5/3$ and $a > 0$), while the reverse is true for binary pulsar observations.

Vertical lines in Figs. 1 and 2 can be mapped to bounds on specific alternative theories, which we can then compare to current Solar System constraints. For example, consider the following cases:

- Brans-Dicke $[(\alpha, b, \beta_{\rm BD}) = (0, -7/3, \beta_{\rm BD})]$: The tracking of the Cassini spacecraft [64] has constrained $\omega_{\rm BD} > \bar{\omega}_{\rm BD} \equiv 4 \times 10^3$, which then forces $\beta_{\rm BD} < (5/3584) 4^{-2/5} (s_1 - s_2)^2 / \bar{\omega}_{\rm BD}$, where $s_{1,2}$ are

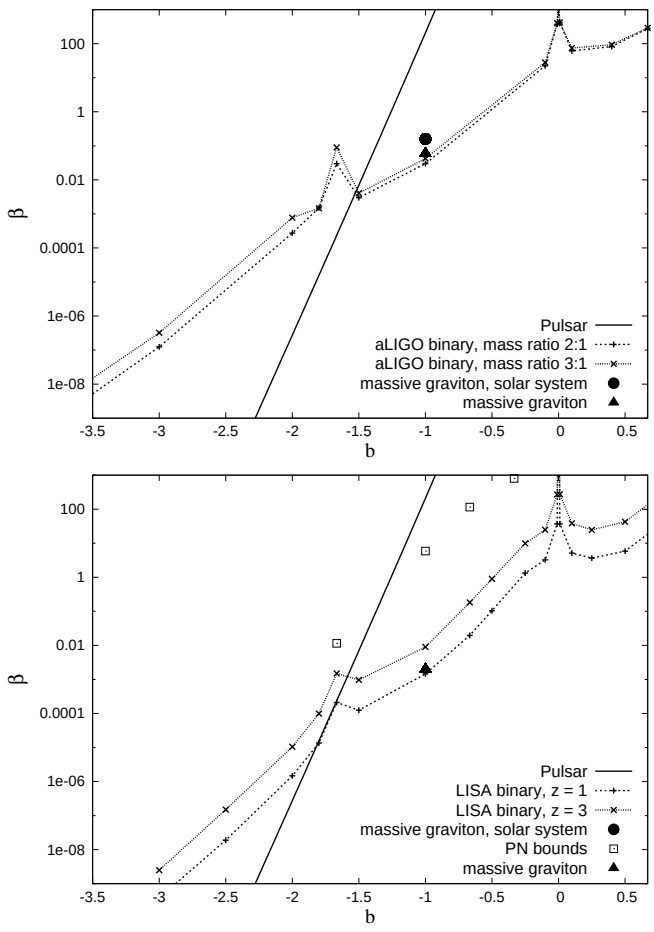

FIG. 2: UPPER PANEL: Bounds on β for different values of b for a single SNR = 20 aLIGO/aVirgo detection. Plotted here is a (3σ) constraint, where σ is the standard deviation of the β parameter derived from the Markov chains. The sources injected had the same parameters as those from the upper panel in Figure 1. Also included is the bound on β derived from the golden pulsar (PSR J0737-3039) data, as well as bounds found from solar system experiments and other aLIGO analyses for massive graviton theory.
LOWER PANEL: Bounds on β for different values of b found using two LISA sources at redshift $z = 1$ and $z = 3$. The pulsar bound is shown for comparison, as well as bounds found from solar system experiments and other LISA analyses for massive graviton theory. These other bounds are scaled to a system with $z = 1$.

the sensitivities of the binary components (for BHs $s_{\rm BH} = 1/2$, and for NSs $s_{\rm NS} \approx 0.2 - 0.3$).

- Massive Graviton $[(\alpha, b, \beta_{\rm MG}) = (0, -1, \beta_{\rm MG})]$: Observations of Solar system dynamics [65] have constrained $\lambda_{\rm MG} > \bar{\lambda}_{\rm MG} \equiv 2.8 \times 10^{12}$ km, which then forces $\beta_{\rm MG} < \pi^2 (D/\bar{\lambda}_{\rm MG}) \mathcal{M}(1+z)^{-1}$ km^{-2}, where D is a distance measure to the source [8].

The Solar System constraint on β_{MG} is shown in Fig. 2 with a black circle[2]. Observe that the constraints we could place with aLIGO and particularly LISA can be orders of magnitude stronger than Solar System constraints (below the black circle). This is more easily seen by mapping our projected constraints on $\beta_{\rm MG}$ to constraints on $\lambda_{\rm MG}$; with the aLIGO source, we find $\lambda_{\rm MG} \lesssim 8.8 \times 10^{12}$ km, while for the LISA source, we find $\lambda_{\rm MG} \lesssim 3.763 \times 10^{16}$ km. This is consistent with results from previous Fisher [8–16] and Bayesian studies [40]. Plotted for comparison are the bounds from Pozzo et al. [40] on the upper panel of 2 and from Stavridis and Will [11] on the lower panel of 2 both labeled as "massive graviton." We find that our bound on β for $b = -1$ is quite comparable to those found in these previous studies. Finally, shown on the lower panel of Fig. 2 are the bounds found in the study by Arun et al. [12], which allowed the PN coefficients themselves to vary as parameters. Their bounds on β are somewhat weaker than those we found in our analysis, but this is an expected effect of the covariance between the PN coefficients.

For all comparisons with previous studies, we took into account differences in SNR between the systems we analyzed and those we were comparing to. We also chose systems with the same or very similar total masses and mass ratios as those explored in previous papers. For the LISA systems, we compare the results from previous papers to our results for redshift $z = 1$.

These plots show several other features that deserve further discussion. First, observe that all results show very little dependence on the choice of system parameters. This is quantitatively true for the aLIGO sources, shown in the upper panels of Figs. 2 and 1, as these signals have the same SNR. The LISA sources, shown in the lower panels of Figures 2 and 1, show a factor of ~ 9 offset, since these curves correspond to signals with different SNRs. The SNR difference is a factor of ~ 3, which is a bit surprising as one would expect the spread on a parameter to scale with the SNR, and not the square of the SNR. However, we are working here in a region where the quadratic approximation to the Fisher matrix is singular, so the usual scaling does not hold. The more rigorous bounds derived in the next section do follow a linear scaling with SNR, which is reasonable since they use ppE injections and have non-singular Fisher matrix elements for the ppE parameters.

Another interesting feature in these plots are the spikes at certain values of a and b. These spikes say that for those values of a and b, gravitational wave observations can say little about the magnitude of GR deviations. The reason for such spikes is that for those values of a and b, α and β become completely or partially degenerate with other parameters. For instance, when $a = 0$, α is fully

[2] We don't show similar constraints for Brans-Dicke theory, as here we consider binary BH inspirals, for which the Brans-Dicke correction would vanish due to the no-hair theorem.

degenerate with the luminosity distance, and when $b = 0$, β is fully degenerate with the initial orbital phase ϕ_c.

One can also develop 'cheap' bounds that use ppE instead of GR signal injections. For instance, one could start with injections with a range of values for α and β, and then look to see when the posterior distributions for these parameters no longer show significant support at the GR values of $\alpha = \beta = 0$. These two types of cheap bounds are illustrated in Fig. 3. Given an observation of a non-zero α, a cheap bound calculation as described in this section (solid curve) would indicate a value $|\alpha| < 1.5$ is still consistent with GR. A similar study with ppE injections, however, which produced the dashed-curve posterior distribution for α would indicate a preference for the ppE model over the GR model with a detection of $\alpha > 0.75$. Thus the technique used in this section, which is a variance-covariance study, answers an inherently different question from a model selection study. In the next section, we explore model selection in detail.

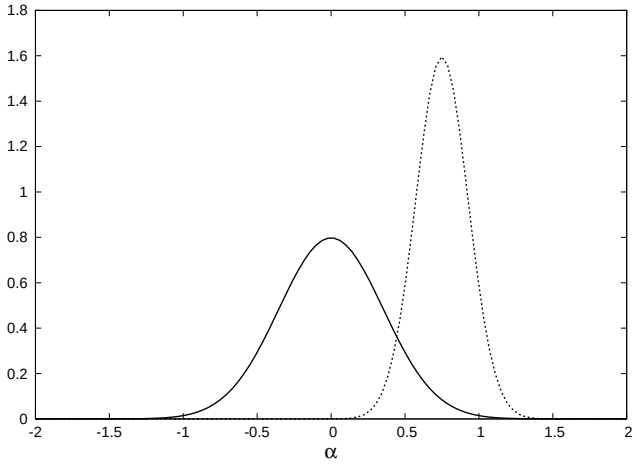

FIG. 3: An illustration of the two approaches for calculating cheap bounds on the ppE amplitude parameters. The solid curve illustrates the bound that can be derived by looking at the spread in the amplitude α when applying the ppE search to GR signals. In this example, values of $|\alpha| > 1.5$ would be taken as indicating a departure from GR. The dashed curve shows the bound that can be derived by starting with ppE signals and determining how large the ppE amplitude needs to be for the posterior distribution to have little weight at the GR value of $\alpha = 0$. In this example, theories with $\alpha > 0.75$ would be considered distinguishable from GR.

B. Rigorous Bounds and Model Selection

In order to see how accurate the cheap bounds found in the previous section are, we next performed a full Bayesian model selection analysis on several different signals. We injected a signal with a given set of ppE parameters and ran a search using both GR and ppE templates. We then calculated the Bayesian evidence for each model and from this the Bayes factor. To compare these results to the cheap bounds, we ran the analysis on several different ppE signals, each with the same injected value of a or b, but with progressively larger values of α or β. This then allows us to determine the values of ppE amplitudes α or β where the evidence for the ppE hypothesis exceeds that of the GR hypothesis by some large factor, which we took to be Bayes factors in excess of 100 (in the Jeffery's classification [66], Bayes factors in excess of 100 represent *decisive* evidence in favor of that model).

We do not expect the cheap bounds to agree precisely with the more rigorous model selection bounds as they are based on quite different reasoning. The cheap bounds simulate what we would find if GR was consistent with observations, and establishes the spread in the ppE amplitude parameters that would remain consistent. If we were to analyze some data and find ppE amplitude parameters outside of this range, it would give us motivation to search more rigorously for departures from GR. With the more expensive model selection bounds, we start with non-GR signals and seek to determine how large the departures from GR have to be for the ppE hypothesis to be preferred. In the first case the distribution of α and β is known to be centered around zero, but in the second case they are not, so the two analyses should not be expected to agree precisely.

One can derive a more detailed connection between the alternative form of the cheap bounds derived using ppE injections (discussed at the end of the previous section) and the more rigorous Bayesian evidence calculations using the Savage-Dickey density ratio [67]. The latter states that for nested hypotheses with separable priors, the Bayes factor is equal to the ratio of the posterior and prior densities evaluated at the parameter values that correspond to the lower dimensional model. If the posterior distribution was a Gaussian with width σ centered at $\alpha = n\sigma$, and we were using a uniform prior with width $N\sigma$, then the Bayes factor would equal $\mathrm{BF} = Ne^{-n^2/2}/\sqrt{2\pi}$, where this Bayes factor shows the odds of the lower dimensional model being correct. For example, with $N = 100$ and $n = 4$ we get a Bayes factor of $\mathrm{BF} = 0.013$, showing strong support for the higher dimensional model. While the cheap bounds that can be derived using ppE signal injections will be stronger than the cheap bounds that can be derived from GR signal injections, the computational cost is higher as multiple simulations have to be run to find the transition point, and this approach is only moderately cheaper than performing the full Bayesian model selection.

Examples of the full model selection procedure are shown in Fig. 4 for aLIGO/aVirgo detections with $\mathrm{SNR} = 20$. Each panel shows Bayes factors for two types of ppE search, one with a or b held fixed at the injected value, and one in which all four ppE values were allowed to vary. The Bayes factor, defined in Eq. (7), is here the odds ratio between the ppE model and the GR model. A larger Bayes factor indicates a stronger preference for the ppE model. The search in which a or b was fixed provides the closest comparison with the cheap bounds of

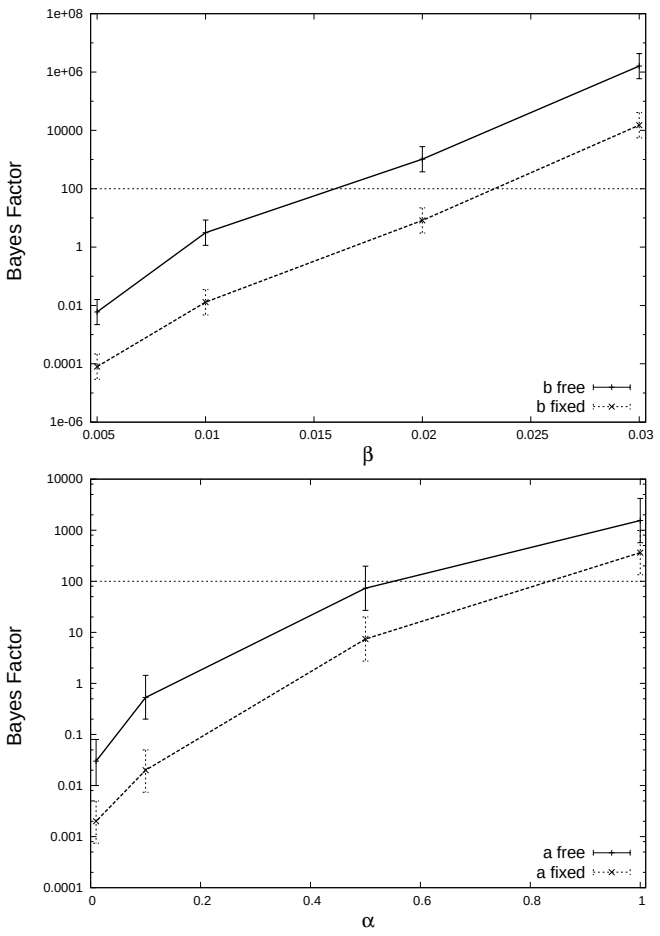

FIG. 4: UPPER PANEL: Bayes factors for a SNR = 20 aLIGO ppE injection with parameters $(a, \alpha, b, \beta) = (0, 0, -1.25, \beta)$. The Bayes factors are the 'betting odds' that ppE (and not GR) is the model that accurately describes the data. As the deviation from GR gets larger, ppE becomes the preferred model.
LOWER PANEL: Bayes factors for a SNR = 20 aLIGO ppE injection with parameters $(a, \alpha, b, \beta) = (-0.5, \alpha, 0, 0)$.

the previous section. The bound on β derived by setting a Bayes factor threshold of 100 are roughly 3 times larger than the cheap bounds when b is held fixed and roughly 2 times larger when b is free to vary. The bounds on α match the cheap bounds when a is held fixed, and is slightly smaller when a is allowed to vary.

We were surprised to find that the bounds are tighter for the higher dimensional models, with (a, b) free, than for the lower dimensional models, with (a, b) fixed. To explore this more thoroughly, we performed a study where the prior on b was increased from a very small range to the full prior range. Since holding a parameter fixed is equivalent to using a delta-function prior, we expect the evidence to interpolate between the values found when b was fixed and when b was free to explore the full prior. Figure 5 confirms this expectation, and also provides an explanation for the growth in the evidence.

To understand this plot, it is helpful to look at the Laplace approximation to the evidence [68], which assumes that the region surrounding the maximum of the posterior distribution is well approximated by a multivariate Gaussian. With this assumption, the evidence is given by

$$p(d|\mathcal{H}) \approx p(d|\vec{\theta}, \mathcal{H})|_{\rm MAP} \left(\frac{\Delta V_{\mathcal{H}}}{V_{\mathcal{H}}}\right) . \qquad (21)$$

The first term is the likelihood evaluated at the maximum of the posterior, and the second term is the ratio of the posterior volume ΔV to the prior volume V. The posterior volume can be estimated from the volume of the error ellipsoid containing 95% of the posterior probability. The ratio $\mathcal{O} = \Delta V/V$ is termed the "Occam factor", and the quantity $I = \log_2(V/\Delta V)$ provides a measure of how much information has been gained about the parameters from the data.

Now consider a situation where we have nested hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, with the second hypothesis involving an additional parameter y. If the likelihood is insensitive to y then the first factor in the evidence stays the same, and since y is unconstrained, $\Delta V_y = V_y$ and the Occam factor is also unchanged. Thus, both models have the same evidence, even though one has more parameters than the other. Conversely, if the additional parameter is tightly constrained by the data, $\frac{\Delta V_y}{V_y}$ can be a very small number. In this case, the evidence for $\mathcal{H}_1$ is much reduced by the Occam factor, and the factor is referred to as an "Occam penalty."

The growth in evidence for the ppE model as the prior range for b gets larger is an effect of this Occam factor, which is a ratio of the uncertainty in the recovered value of an extra parameter to the prior volume for that parameter. As the prior range on b expands, this leads to a greater variance in the recovered values for β'. Because the prior volume of β' remains unchanged, the large growth in its variance as the prior range of b is expanded leads to a large growth in the Occam factor - and thus a shrinking of the Occam penalty. As the Occam factor gets larger, so does the evidence for the ppE model. The evidence for the GR model, of course, does not depend on the priors we use for the ppE parameters, and so as the evidence for ppE grows, the Bayes factor indicates a stronger preference for ppE.

Figure 6 shows Bayes factors between the GR and ppE hypotheses for a $z = 1$ LISA source. In the upper panel, the injections where chosen with $a = 0, b = -1$ and variable β, while in the lower panel the injections were chosen with $a = 0.5, b = 0$ and variable α. Because LISA sources have much higher SNR, the ppE parameters are more tightly constrained, and the difference between the Bayes factors when a or b are fixed versus freely varying is less pronounced. The more rigorous bounds on α and β are both a factor of ~ 2 times weaker than those predicted by the cheap bounds, which is in line with what we found for the phase correction β in the aLIGO example.

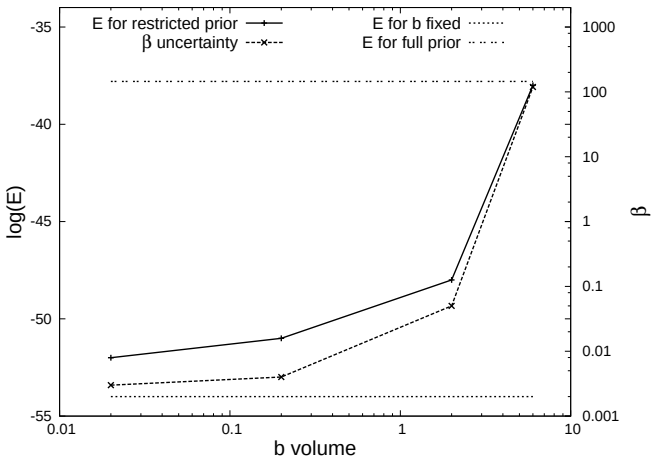

FIG. 5: Here we plot the log of the evidence (E) for the ppE model characterizing a ppE injection as the prior volume on b is increased. The evidence for the ppE model increases with the prior volume on b. The growth in the evidence can be attributed to the growth in the variance of β, which lessens the severity of the 'Occam penalty' for more model parameters.

In summary, the cheap bounds provide a fair approximation to the bounds that can be derived from Bayesian model selection, and can generally be trusted to within an order of magnitude.

C. Fitting Factor

Another quantity of interest is the fitting factor, which measures how well one template family can recover an alternative template family. To define the fitting factor, we must first define the match between two templates h and h' as

$$\mathcal{M} = \frac{(h|h')}{\sqrt{(h|h)}\sqrt{(h'|h')}}\,. \tag{22}$$

The match is related to the metric distance between templates [69] by $\mathcal{M} = 1 - \frac{1}{2} g_{ij}\Delta x^i \Delta x^j$, where the metric is evaluated with the higher-dimensional model (appropriate when dealing with nested models). The fitting factor FF is then defined as the best match that can be achieved by varying the parameters of the h' template family to match the template belonging to the the other family, h.

Another interpretation for the fitting factor is as the fraction of the true signal-to-noise ratio $\mathrm{SNR} = \sqrt{(h|h)}$ that is recovered by the frequentist statistic $\rho = \max[(h|h')/\sqrt{(h'|h')}]$. The imperfect fit leaves behind a residual $(h - h')$ with $\mathrm{SNR}^2_{\rm res} = \chi^2$, which can be minimized by adjusting the amplitude of h' to yield

$$\mathrm{SNR}^2_{\rm res} = (1 - \mathrm{FF}^2)\mathrm{SNR}^2\,. \tag{23}$$

Assuming that a residual with SNR_* is detectable, and

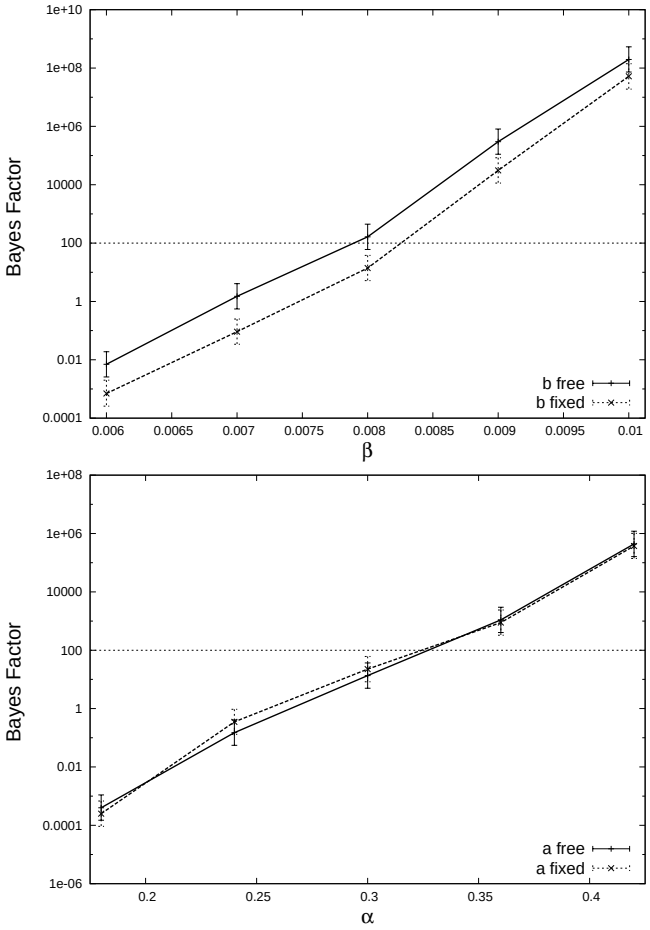

FIG. 6: UPPER PANEL: Bayes factors for a $z = 1$ LISA ppE injection with parameters $(a, \alpha, b, \beta) = (0, 0, -1.0, \beta)$. LOWER PANEL: Bayes factors for a $z = 1$ LISA ppE injection with parameters $(a, \alpha, b, \beta) = (0.5, \alpha, 0, 0)$.

working in the limit where $\mathrm{FF} \sim 1$, we have

$$1 - \mathrm{FF} \simeq \frac{\mathrm{SNR}^2_*}{2\,\mathrm{SNR}^2}\,. \tag{24}$$

We see then that the ability to detect departures from GR scales inversely with the square of the SNR, as given by Eq. (24). On the other hand, the detectable difference between the parameters in the two theories will scale inversely with a single power of the SNR. This is because this detectable difference is proportional to the square-root of the minimized match function and

$$\sqrt{\min(g_{ij}\Delta x^i \Delta x^j)} \simeq \frac{\mathrm{SNR}_*}{\mathrm{SNR}}\,, \tag{25}$$

and the metric is independent of SNR. This reasoning applies to both the additional model parameters of the alternative theory, *e.g.* $\Delta x^i = (\alpha, \beta)$, and the physical source parameters such as the masses and distance. We then expect both the bounds on the ppE model parameters and the biases caused by using the wrong template family to scale inversely with SNR. This scaling is in

keeping with the usual scaling of parameter estimation errors that follows from a Fisher matrix analysis where $\langle \Delta x^i \Delta x^j \rangle \simeq (h_{,i}|h_{,j})^{-1} \sim \mathrm{SNR}^{-2}$. Figure 7 shows that the errors in the recovery of the ppE parameters follows the expected scaling with SNR.

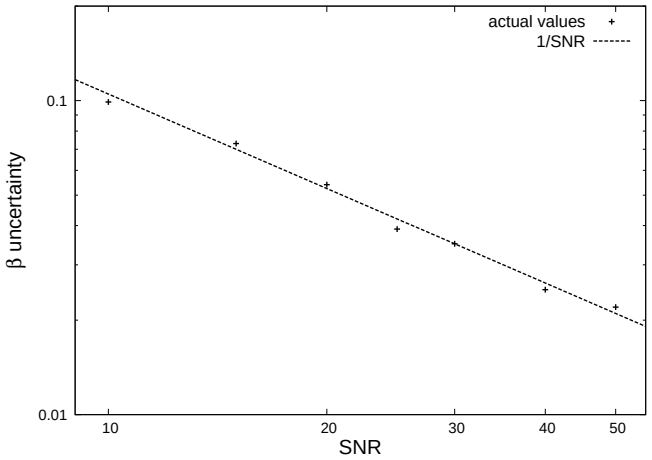

FIG. 7: The scaling of the parameter estimation error in the ppE parameter β for an aLIGO simulation with ppE parameters $(a, \alpha, b, \beta) = (0, 0, -1.25, 0.1)$. The parameter errors follow the usual 1/SNR scaling.

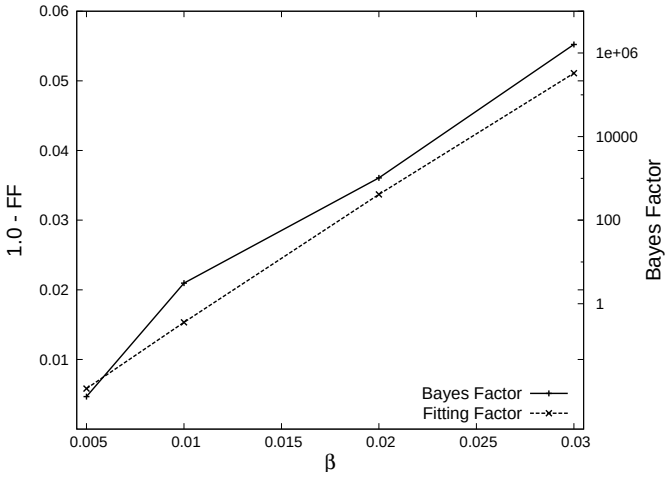

FIG. 8: The log Bayes factors and $(1 - \mathrm{FF})$ plotted as a function of β for a ppE injection with parameters $(a, \alpha, b, \beta) = (0, 0, -1.25, \beta)$. The predicted link between the fitting factor and Bayes factor is clearly apparent.

Alternative models that are not well-fitted by GR will be more easily distinguished than models that can be well-fitted. This suggests that there should be a correlation between the fitting factor and the Bayes factor. The relationship can be established using the Laplace approximation to the evidence [Eq. (21)], from which it follows that the log Bayes factor is equal to

$$\begin{aligned} \log \mathrm{BF} &= \log \frac{e^{-\chi^2(\mathcal{H}_1)/2}}{e^{-\chi^2(\mathcal{H}_0)/2}} \frac{\mathcal{O}_1}{\mathcal{O}_0} \\ &= \frac{\chi^2_{\rm min}}{2} + \Delta \log \mathcal{O} \\ &= (1 - \mathrm{FF}^2) \frac{\mathrm{SNR}^2}{2} + \Delta \log \mathcal{O}\,, \end{aligned} \tag{26}$$

where $\mathcal{O}$ is the Occam factor, defined in the discussion following [Eq. (21)]. Thus, up to the difference in the log Occam factors, the log Bayes factor should scale as $2(1 - \mathrm{FF})$ when $\mathrm{FF} \sim 1$. This link is confirmed in Figure 8.

D. Parameter Biases

If we assume that Nature is described by GR, but in truth another theory is correct, this will result in the recovery of the wrong parameters for the systems we are studying. For instance, when looking at a signal that has non-zero ppE phase parameters, a search using GR templates will return the incorrect mass parameters, as illustrated in Fig. 9. Observe that as the magnitude of β is increased (thus increasing the Bayes factor), the error in the chirp mass parameter extraction grows well beyond statistical errors.

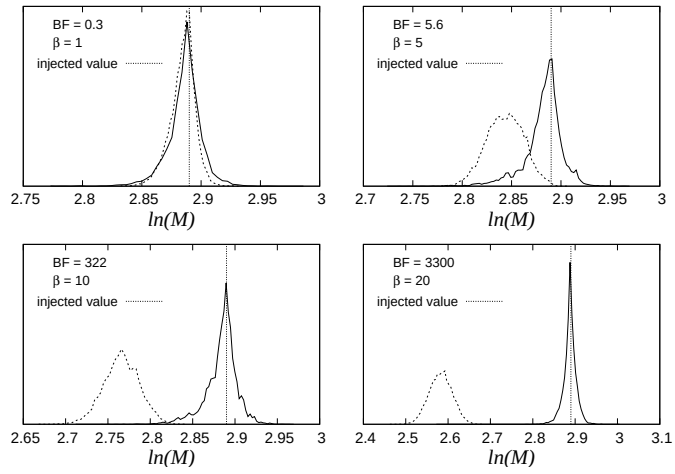

FIG. 9: Histograms showing the recovered log total mass for GR (dashed) and ppE (solid) searches on ppE signals. As the source gets further from GR, the value for total mass recovered by the GR search moves away from the actual value. All signals had injected $b = -0.25$.

Perhaps the most interesting point to be made with this study is that the GR templates return values of the total mass that are completely outside the error range of the (correct) parameters returned by the ppE search, *even for ppE signals that are not clearly discernible from GR*. We refer to this parameter biasing as 'stealth bias', as it is not an effect that would be easy to detect, even if one were looking for it.

As an example, consider stealth bias for non-zero ppE α parameters, as illustrated in Fig. 10. As one would expect, when a GR template is used to search on a ppE signal that has non-zero ppE amplitude corrections, the parameter that is most affected is the luminosity distance. We again see the bias of the recovered parameter becoming more apparent as the signal differs more from GR[3]. For example, the recovered posterior distribution from the search using GR templates has zero weight at the correct value of luminosity distance when the Bayes factor is ~ 50. Even when the Bayes factor is of order unity, the peaks of the posterior distributions of the luminosity distance differ by approximately 10 Gpc.

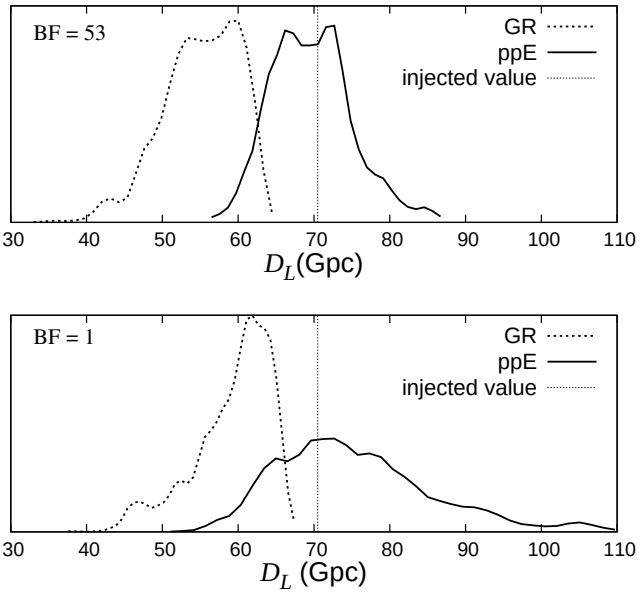

FIG. 10: Histograms showing the recovered values for luminosity distance from GR and ppE searches on a LISA binary at redshift $z = 7$. Both signals have $a = 0.5$, and were injected with a luminosity distance of 70.5 Gpc. The top plot has $\alpha = 3.0$ and the bottom has $\alpha = 2.5$. As the Bayes factor favors the ppE model more strongly, the bias in the recovered luminosity distance from the GR search becomes more pronounced.

V. CONCLUSION

The two main results of this study are that GW observations of binary compact object inspirals using ppE waveforms can constrain higher PN order (i.e. $b > -5/3$ and $a > 0$) deviations from GR much more tightly than binary pulsar observations, and that parameter estimates can be significantly biased if GR templates are used to recover signals when an alternative theory of gravity better describes the event. This latter bias can be significant even in cases where it is *not* obvious that GR is not quite the correct theory of gravity. We also see that the detection efficiency of GR templates can be seriously compromised if they are used to characterize data that is not described by GR.

The current study makes several simplifying assumptions about the waveforms: we consider only the inspiral stage for non-spinning black holes on circular orbits, and include just the leading order ppE corrections to the waveforms. In future work we plan to include a marginalization over these higher order corrections. Including this marginalization will be more realistic, as the ppE formalism allows for many higher order corrections to the waveform. Marginalizing over the higher order terms will weaken the bounds on the leading order ppE parameters, though probably not by that much since they are sub-dominant terms.

Another subject that we will examine in the future is the effect on our analysis of multiple detections. Simultaneously characterizing several systems with different mass ratios should allow us to examine the dependence of the α/β parameters on spin, mass difference, mass ratio, etc.. Furthermore, looking at several systems simultaneously will break the degeneracies between the ppE parameters and the individual system parameters (masses, distances *etc*), and will allow us to detect significantly smaller deviations from GR.

We also plan to perform a study similar to that done by Arun *et al.* [31–33], in which the exponents a_i, b_i are fixed at the values found in the PN expansion of GR, and compare their Fisher matrix based bounds to those from Bayesian inference. We expect a full Bayesian inference study to lead to significantly different conclusions, due to the singularities in the Fisher matrix already observed in the present study.

Finally, we will look at LISA observations of galactic white-dwarf binaries to see if the brighter systems, which may have SNRs in the hundreds, may allow us to beat the pulsar bounds across the entire ppE parameter space. The brightest white-dwarf systems will have $u \sim 10^{-8} \to 10^{-7}$ (for comparison the 'golden' double pulsar system, PSR J0737-3039A has $u = 3.94 \times 10^{-9}$), and these small values for u make the ppE effects, which scale as u^a and u^b, much larger than for black hole inspirals when $a, b < 0$.

The chance to test the validity of Einstein's theory of gravity is one of the most exciting opportunities that gravitational wave astronomy will afford to the scientific community. Without the appropriate tools, however, our ability to perform these tests is sharply curtailed. This analysis has shown that the ppE template family could be an effective means of detecting and characterizing deviations from GR, and also that assuming that our GR waveforms are correct could lead to lessened detection ef-

[3] Here, the uncertainty in the recovered luminosity distance changes considerably between the different systems, because we held the injected luminosity distance constant instead of the injected SNR.

ficiency and biased parameter estimates if gravity is described by an alternative theory (even when choosing parameters at the threshold of what has already been ruled out by Solar System and binary pulsar observations). We have identified several areas of future investigation, and will continue to study this area in depth.

Acknowledgments

We thank Patrick Brady, Curt Cutler, Ben Owen, David Spergel, Xavier Siemens, Paul Steinhardt and Michelle Vallisneri for detailed comments and suggestions. We are very grateful to Martin Weinberg and Will Farr for making their direct evidence integration codes available to us, and for helping us to understand the results. N. J. and L. S. acknowledge support from the NSF Award 0855407 and NASA grant NNX10AH15G. N. Y. and F. P. acknowledge support from the NSF grant PHY-0745779, and FP acknowledges the support of the Alfred P. Sloan Foundation.

[1] C. M. Will, Living Reviews in Relativity **9** (2006), URL `http://www.livingreviews.org/lrr-2006-3`.
[2] R. A. Hulse and J. H. Taylor, Astrophys. J. **195**, L51 (1975).
[3] M. Burgay et al., Nature. **426**, 531 (2003), astro-ph/0312071.
[4] M. Kramer et al., Science **314**, 97 (2006), astro-ph/0609417.
[5] N. Yunes and F. Pretorius, Phys. Rev. **D80**, 122003 (2009), 0909.3328.
[6] C. Cutler and M. Vallisneri, Phys. Rev. **D76**, 104018 (2007), 0707.2982.
[7] B. F. Schutz, J. Centrella, C. Cutler, and S. A. Hughes (2009), 0903.0100.
[8] C. M. Will, Phys. Rev. **D57**, 2061 (1998), gr-qc/9709011.
[9] C. M. Will and N. Yunes, Class. Quant. Grav. **21**, 4367 (2004), gr-qc/0403100.
[10] E. Berti, A. Buonanno, and C. M. Will, Class. Quant. Grav. **22**, S943 (2005), gr-qc/0504017.
[11] A. Stavridis and C. M. Will, Phys. Rev. **D80**, 044002 (2009), 0906.3602.
[12] K. G. Arun and C. M. Will, Class. Quant. Grav. **26**, 155002 (2009), 0904.1190.
[13] D. Keppel and P. Ajith, Phys. Rev. **D82**, 122001 (2010), 1004.0284.
[14] K. Yagi and T. Tanaka (2009), 0906.4269.
[15] C. M. Will, Phys. Rev. **D50**, 6058 (1994), gr-qc/9406022.
[16] P. D. Scharre and C. M. Will, Phys. Rev. **D65**, 042002 (2002), gr-qc/0109044.
[17] F. D. Ryan, Phys. Rev. D **52**, 5707 (1995).
[18] N. A. Collins and S. A. Hughes, Phys. Rev. **D69**, 124022 (2004), gr-qc/0402063.
[19] E. Berti, V. Cardoso, and C. M. Will, Phys. Rev. **D73**, 064030 (2006), gr-qc/0512160.
[20] E. Berti, J. Cardoso, V. Cardoso, and M. Cavaglia, Phys. Rev. **D76**, 104044 (2007), 0707.1202.
[21] S. A. Hughes, AIP Conf. Proc. **873**, 233 (2006), gr-qc/0608140.
[22] S. Alexander and N. Yunes, Phys. Rept. **480**, 1 (2009), 0907.2562.
[23] N. Yunes and L. C. Stein (2011), 1101.2921.
[24] C. F. Sopuerta, GW Notes, Vol. 4, p. 3-47 **4**, 3 (2010).
[25] K. Nordtvedt, Phys. Rev. **169**, 1017 (1968).
[26] C. M. Will, Astrophys. J. **163**, 611 (1971).
[27] C. M. Will and K. J. Nordtvedt, Astrophys. J. **177**, 757 (1972).
[28] K. J. Nordtvedt and C. M. Will, Astrophys. J. **177**, 775 (1972).
[29] T. Damour and J. H. Taylor, Phys. Rev. D **45**, 1840 (1992).
[30] N. Yunes, K. G. Arun, E. Berti, and C. M. Will (2009), 0906.0313.
[31] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Phys. Rev. **D74**, 024006 (2006), gr-qc/0604067.
[32] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Class. Quant. Grav. **23**, l37 (2006), gr-qc/0604018.
[33] C. K. Mishra, K. G. Arun, B. R. Iyer, and B. S. Sathyaprakash (2010), 1005.0304.
[34] S. Alexander, L. S. Finn, and N. Yunes, Phys. Rev. D **78**, 066005 (2008), 0712.2542.
[35] N. Yunes, R. O'Shaughnessy, B. J. Owen, and S. Alexander, Phys. Rev. **D82**, 064017 (2010), 1005.3310.
[36] N. Yunes and F. Pretorius, Physical Review D (Particles, Fields, Gravitation, and Cosmology) **79**, 084043 (pages 14) (2009), URL `http://link.aps.org/abstract/PRD/v79/e084043`.
[37] C. F. Sopuerta and N. Yunes, Physical Review D (Particles, Fields, Gravitation, and Cosmology) **80**, 064006 (pages 24) (2009), URL `http://link.aps.org/abstract/PRD/v80/e064006`.
[38] N. Yunes, F. Pretorius, and D. Spergel (2009), 0912.2724.
[39] J. D. Bekenstein, Phys. Rev. **D70**, 083509 (2004), astro-ph/0403694.
[40] W. Del Pozzo, J. Veitch, and A. Vecchio, ArXiv e-prints (2011), 1101.1391.
[41] N. Yunes and S. A. Hughes, Phys. Rev. D **82**, 082002 (2010), 1007.1995.
[42] T. B. Littenberg and N. J. Cornish, Phys. Rev. **D80**, 063007 (2009), 0902.0368.
[43] L. D. V., *Bayesian statistics: a review* (SIAM, Philadelphia, USA, 1972).
[44] L. C. Stein, N. Yunes, and S. A. Hughes (2010), 1012.3144.
[45] K. Yagi, N. Tanahashi, and T. Tanaka (2011), 1101.4997.
[46] J. Gair and N. Yunes, in progress (2011).
[47] W. Anderson, P. Brady, D. Chin, J. Creighton, K. Riles, and J. Whelan, LIGO Techinal Report **LIGO-T010110-00-Z** (2002).
[48] T. A. Prince, M. Tinto, S. L. Larson, and J. W. Armstrong, Phys. Rev. **D66**, 122002 (2002), gr-qc/0209039.

[49] N. J. Cornish and L. J. Rubbo, Phys. Rev. **D67**, 022001 (2003), gr-qc/0209011.
[50] L. J. Rubbo, N. J. Cornish, and O. Poujade, Phys. Rev. **D69**, 082003 (2004), gr-qc/0311069.
[51] C. Cutler, Phys. Rev. **D57**, 7089 (1998), gr-qc/9703068.
[52] J. S. Key and N. J. Cornish (2010), 1006.3759.
[53] C. M. Will and E. Poisson, book in progress (2011).
[54] C. M. Will, *Theory and experiment in gravitational physics* (Cambridge University Press, Cambridge, UK, 1993).
[55] R. W. Hellings, Phys. Rev. D **17**, 3158 (1978).
[56] K. J. Lee, F. A. Jenet, and R. H. Price, Astrophys. J. **685**, 1304 (2008).
[57] M. Tinto and M. E. da Silva Alves, Phys. Rev. **D82**, 122003 (2010), 1010.1302.
[58] N. J. Cornish (2010), 1007.4820.
[59] C. J. Ter Braak, Statistics and Computing **16**, 239 (2006), ISSN 0960-3174, URL `http://portal.acm.org/citation.cfm?id=1145406.1145416`.
[60] C. J. Ter Braak and J. A. Vrugt, Statistics and Computing **18**, 435 (2008), ISSN 0960-3174, URL `http://dx.doi.org/10.1007/s11222-008-9104-9`.
[61] M. D. Weinberg (2009), 0911.1777.
[62] P. M. Goggans and Y. Chi, AIP Conference Proceedings **707**, 59 (2004), URL `http://link.aip.org/link/?APC/707/59/1`.
[63] T. B. Littenberg and N. J. Cornish, Phys. Rev. **D82**, 103007 (2010), 1008.1577.
[64] B. Bertotti, L. Iess, and P. Tortora, Nature **425**, 374 (2003).
[65] C. Talmadge, J. P. Berthias, R. W. Hellings, and E. M. Standish, Phys. Rev. Lett. **61**, 1159 (1988).
[66] H. Jeffreys, Zeitschrift Naturforschung Teil A **6**, 471 (1951).
[67] I. Verdinelli and L. Wasserman, Journal of the American Statistical Association **90**, pp. 614 (1995), ISSN 01621459, URL `http://www.jstor.org/stable/2291073`.
[68] A. Azevedo-Filho and R. D. Shachter, in *UAI'94* (1994), pp. 28–36.
[69] B. J. Owen, Phys. Rev. D **53**, 6749 (1996), arXiv:gr-qc/9511032.

VI. APPENDIX A

As described in Section III, the VTA method for calculating evidences involves two possible sources of error. One is introduced by the fact that our Markov chains are of finite length. To get an idea of the magnitude of this statistical uncertainty, the implementation of the VTA that we used calculates the evidence many times using different sub-samples of the Markov chain. This process is called bootstrapping, and we find that in general it results in an uncertainty in the log Bayes factor of the order ± 0.5.

The second source of possible error in the VTA techniques comes from the choice of boxing number. The boxing number is the number of points from the chain that are sorted into each volume element. A higher boxing number will return a more accurate number for the mean or median of the posterior in a given volume element, but at the cost of having large volume elements that may not resolve fine features in the posterior distribution. Lower boxing numbers lead to greater variance in the estimate of the posterior density in each cell, but allows for better resolution of sharp features in the posterior landscape. To examine the systematic error in Bayes factors associated with using different boxing numbers, we calculated the Bayes factor between ppE and GR models for a source with injected ppE parameters $(a, \alpha, b, \beta) = (0.5, 75, 0, 0)$. We first used thermodynamic integration with a run using 50 chains, and found the log Bayes factor to equal $\log(B) = 12.0 \pm 1.0$. Because thermodynamic integration performs more accurately than the VTA when integrating posterior distributions for which analytic answers are available, such as a multi-variate Gaussian, we take this value as our reference. We then calculated the log Bayes factor using the VTA with boxing numbers of 16, 32, and 64. The results, including the statistical uncertainty, are shown in Table II.

TABLE II: Bayes factors calculated using the VTA with different boxing numbers.

Boxing Number	GR	ppE	$log(BF)$
16	$-41.50^{+0.2}_{-0.22}$	$-31.04^{+0.88}_{-0.75}$	$10.5^{+1.0}_{-1.0}$
32	$-40.43^{+0.15}_{-0.13}$	$-28.02^{+0.67}_{-0.44}$	$12.4^{+0.8}_{-0.6}$
64	$-39.51^{+0.26}_{-0.31}$	$-25.78^{+0.43}_{-0.30}$	$13.7^{+0.7}_{-0.6}$

The results show that the variation in $\log(B)$ between different boxing sizes is similar to, but slightly larger than the statistical variation introduced by the VTA within one boxing size. The variation due to choice of boxing size is roughly ± 1.5. We therefore use error bars indicating $\log(B) \pm 1$ on our Bayes factor plots. Further, we found that a boxing size of 32 returned the most accurate value for the Bayes factor, and so we used this size for the rest of our analysis.