Uncertainty limits on neutron star radius measurements with gravitational waves

Katerina Chatziioannou¹, ²

¹Department of Physics, California Institute of Technology, Pasadena, California 91125, USA
²LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA

(Dated: April 15, 2022)

Upcoming observing campaigns with improved detectors will yield numerous detections of gravitational waves from neutron star binary inspirals. Rare loud signals together with numerous signals of moderate strength promise stringent constraints on the properties of neutron star matter, with a projected radius statistical uncertainty of 50 – 200m with O(2000) sources. Given this precision we revisit all analysis assumptions and identify sources of systematic errors, quantify their impact on radius extraction, and discuss their relative importance and ways to mitigate them.

I. INTRODUCTION

Astronomical observations constrain the macroscopic properties of neutron stars (NSs) and the behavior of dense, cold matter [1–3]. Among them, gravitational wave (GW) observations of binary NS (BNS) inspirals lead to a measurement of the masses and tidal properties of the stars [4, 5]. The binary masses affect the evolution of the early, minute-long signal, while each NS’s response to its companion’s gravitational tidal field leaves an imprint on the signal during the final coalescence stages.

The mutual tidal interaction induces a quadrupole moment on each NS which accelerates the GW inspiral. The leading-order effect is quantified through the tidal deformability \( \Lambda \equiv (2/3) k_2(R/m)^5 \) [6, 7], where \( k_2 \) is the Love number of a NS with mass \( m \) and radius \( R \). Measuring \( \Lambda \) offers complementary information to traditional radius measurements through \( k_2 \), while the strong mass dependence implies tighter constraints for less massive, and thus more deformable, NSs. The first BNS coalescence detected with GWs, GW170817 [8], had a signal-to-noise ratio (SNR) of \( \rho = 32 \), and a combined tidal deformability \( \Lambda \) [9, 10] of \( 300^{+240}_{-230} \) at the 90% credible level [11, 12]. This result has been extensively shown to disfavor very stiff nuclear matter and large NS radii and tidal deformabilities [13–25] and to be in agreement with further astronomical and terrestrial constraints [23–35].

Scheduled or planned GW detector upgrades and observing campaigns [36] are expected to yield further BNS detections and improve on the overall constraints by both combining information from multiple events of different masses [23, 37–39] and detecting louder signals. The uncertainty in tidal parameters scales as \( \rho^{-1} \) [10], however, the tidal deformability affects the GW signal primarily during the late stages of the coalescence for frequencies \( \gtrsim 400 \text{Hz} \). Lower frequencies are essential for identifying the signal and estimating the masses, but tidal inference primarily relies on the SNR accumulated, and thus the expected detector performance, at this frequency range.

Since the SNR is inversely proportional to the detector sensitivity, we can make rough estimates about the expected constraints on the NS tidal deformability. The LIGO-Livingston [40] strain sensitivity in the relevant frequencies improved by a factor of \( \sim 2 \) between the second [41, 42] and third observing runs [43, 44]. Design sensitivity could bring another factor of \( \sim 1.5 \) [45], with the A+ and Voyager [46, 47] upgrades yielding improvements by factors of \( \sim 2 \) and \( \sim 1.5 \) respectively. Next generation 3G detectors are envisioned to have \( \sim 10 \) times better strain sensitivity than advanced LIGO [48–52]. Though these improvements are not uniform across the frequency band, they roughly suggest that a GW170817-like event would have an SNR (\( \Lambda \) 90% uncertainty) of 100(200) at design sensitivity, 200(100) with A+, 300(66) with Voyager, and 1000(20) with 3G detectors.

The total number of sources observed per detector upgrade depends on the BNS merger rate and its distribution with redshift with a current estimate of \( 320^{+240}_{-230} \) Gpc\(^{-3} \) yr\(^{-1} \) mergers uniformly distributed in comoving volume and for uniform NS masses in \( (1, 2.5) M_\odot \) [53]. With these specifications, the median merger rate, and a detection SNR threshold of 12 we use the GW-Toolbox [54] to estimate \( O(10) \) detections per year at design sensitivity, \( O(100) \) for A+, and \( O(500) \) for Voyager. The latter is 7 times lower than the estimate of [46], likely due to differences in the merger rate and SNR threshold. Predictions for 3G detectors are subject to the uncertain redshift distribution of mergers, but estimates suggest \( O(1000) \) sources per day [55, 56].

We estimate the total SNR accumulated by such a catalog of BNS detections by adopting a \( \rho^{-4} \) distribution [57], appropriate for detectors up to Voyager that get the majority of their sources from redshifts up to 0.3 [46]. We simulate 100 source catalogs and compute the total SNR by summing the per-event SNRs in quadrature to find a median total SNR of \( 60(200)[450](650) \) with \( 10(100)[500](1000) \) sources. The total SNR from 3G detectors is expected to be \( O(10^4) \) [58]. These estimates are conservative if dedicated high-frequency detectors join the network [59, 60].

The corresponding NS radius accuracy is shown in Fig. 1 for different NS masses and radii, assuming we can perfectly convert a \((m, \Lambda)\) measurement to \( R \). Since we are interested in the expected radius uncertainty and not the details of the calculation, we use the relations of [61–64] between the masses, tides, and radius. In practice such relations carry additional systematic uncertainty, so approaches that model the whole NS equation...
of state hierarchically would be preferred in the regime of informative measurements. However, our goal here is an estimate of the expected radius uncertainty and not the details of which analysis can achieve it, so such relations are appropriate.

The $\Lambda \sim (R/m)^3$ scaling results in increased radius uncertainty with higher/lower NS mass/radius due to intrinsically weaker binary tidal interactions. An SNR of 1000, achieved either by observing a GW170817-like event with 3G detectors or with 4 years of Voyager operation, would result in a measurement of $\Lambda$ to 20 and $R$ to 50–200m for different NS masses. A more moderate total SNR of 200 from $\sim$100 sources would lead to a $\Lambda (R)$ uncertainty of 100 (1km) at 1.6$M_\odot$, consistent with the more detailed simulations of [23]. The final constraint achieved on NS radii will be a combination of these per-mass estimates depending on the astrophysical NS mass distribution.

The above are not detailed predictions about the expected NS constraints from future detectors; such estimates would require a precise treatment of –among others– the BNS merger rate, its redshift distribution, the NS mass distribution, the broadband detector performance, the network duty cycle, etc. However, they provide a projection that GWs could result in a $\sim$100m radius measurement within the decade with Voyager and beyond with 3G detectors. This radius constraint would also improve if effects such as dynamical tides [65–67] are detected as they are qualitatively different than the standard adiabatic tides considered here and not captured by the $\rho^{-1}$ scaling. Reaching this projected precision relies on ascertaining that every aspect of the GW analysis induces potential systematic errors that are fully quantified and brought below statistical uncertainties.

### II. GRAVITATIONAL WAVE ANALYSIS

Analysis of GW data $d$ to extract source parameters $\theta$ relies on modeling the signal with a waveform template $h(\theta)$ under some model for the detector noise. The likelihood function in the frequency domain is [68, 69]

$$\log L \sim -\frac{1}{2}(d - h(\theta))d - h(\theta)),$$

with the noise-weighted inner product

$$\langle a | b \rangle = 2 \int \frac{a^*(f)b(f) + b^*(f)a(f)}{S_n(f)}df,$n$$

where an asterisk denotes complex conjugation and $S_n(f)$ is the power spectral density (PSD) of the noise. The likelihood and a prior for $\theta$ give the posterior probability.

The above allows us to identify the ingredients of parameter estimation:

(i) the data $d$,

(ii) the noise PSD $S_n(f)$,

(iii) the waveform model $h(\theta)$,

as well as the main assumptions:

(iv) the detector noise is stationary, leading to a diagonal noise covariance matrix and an inner product that is a one-dimensional frequency integral, and

(v) the detector noise is gaussian, which leads to the Gaussian functional form of the likelihood.

Each of the above introduces systematic uncertainties that will affect inference at some level.

### III. ASSUMPTION: GAUSSIAN NOISE

The functional form of the likelihood is dictated by the assumption of gaussian detector noise. Gaussianity can be violated by instrumental artifacts, known as glitches, or multiple GW signals temporally overlapping. Glitches are a common occurrence, with a rate of $\lesssim 1$ per minute in the LIGO detectors in O3a [44] and already coinciding with signals, notably GW170817 [8, 70]. Given the typical duration of BNS signals in the LIGO band of a few minutes and the rate of glitches, we expect such BNS-glitch overlaps to be quite common. Overlapping signals are expected to be rare in advanced LIGO but a possibility for 3G detectors [71–74].

The temporal coincidence of glitches and signals has led to the development of mitigation techniques that simultaneously model the signal and the glitch [75] or use auxiliary channel information [76–79]. In the context of
Posterior Density

FIG. 2. Marginalized posterior for $\tilde{\Lambda}$ examining the effect of non-Gaussian residuals (top) and misestimated noise PSD (bottom). Black vertical lines give the injected values. Purple (orange) shade distributions correspond to simulated signals of SNR 140 (420) in data with glitches. Green shade distributions correspond to signals analyzed with different misestimated high-frequency PSDs. Legends give the glitch SNR and the identification probability, defined as the percentage of posterior samples that model the glitch (top), and the level of PSD misestimation (bottom).

... tidal inference, glitches are relevant when overlapping with the signal at frequencies $\gtrsim$400Hz. Though O3a was dominated by glitches with peak frequencies below 100Hz [44], those glitches extend to higher frequencies and improved detector sensitivity could bring new glitch families. A prominent glitch will be modeled together with the signal [75], leading to unbiased signal parameters. However, this does not preclude the possibility of a stealth bias [80], where the glitch is not loud enough to be identified but could still affect tidal inference.

We explore this possibility by simulating BNS signals with SNR 140 and 420 in a zero-noise realization. We break the Gaussianity assumption by adding a glitch in the LIGO-Livingston data given by a sine-gaussian that overlaps with the BNS signal at 700Hz, the most interesting frequency region for tidal inference. The quality factor of the glitch is 20, resulting in an instrumental transient that resembles the underlying signal but is present in only one detector. We vary the SNR of the glitch and simultaneously model the signal with waveform templates and the glitch with wavelets [75].

Figure 2 shows the resulting $\tilde{\Lambda}$ posterior also compared to the case of no glitch and purely Gaussian noise. Uncertainties are consistent with the signal SNR and the projections of Fig. 1. In all cases the correct value is recovered regardless of the SNR of the glitch. The skewed shape of the posterior is due to a correlation between $\tilde{\Lambda}$ and the binary mass ratio causing the mean of the posterior to be to the left of the injected value; we have verified that the two-dimensional posterior peaks at the injected value. For glitches with $\rho \gtrsim 6$ the probability of glitch identification is $\sim 1$. For lower glitch SNR the probability drops, with only 3% of the posterior samples identifying a glitch of $\rho = 4$. Despite this, the $\tilde{\Lambda}$ posterior remains unbiased, suggesting no stealth bias: if the glitch is too quiet to be identified, it is also too quiet to bias tidal inference. This example considers one glitch and signal and it is possible that different glitch morphologies could prove more problematic. However, it shows that glitch mitigation techniques are already in place and able to handle instrumental artifacts in the data.

The other possibility is that of a secondary signal overlapping with the primary BNS. The most problematic scenario occurs for signals separated by less than 0.1s [71–74] hence the secondary signal could overlap with the tidally-affected region of the primary signal. Biases could occur also if the secondary signal is subthreshold and not individually resolvable [73]. This case requires a joint analysis of the two signals, similar to the joint signal and glitch analysis from above, and also similar to techniques developed for space-based detectors whose data contain millions of overlapping signals [81–83].

In summary, non-Gaussian features in the data are unavoidable, but they can be modeled simultaneously with the primary signal to ensure unbiased parameter recovery. What is more, any systematic bias from non-Gaussian data will be unique to each BNS. The bias will therefore not accumulate when multiple events are combined, though it might affect individual very loud sources.

IV. ASSUMPTION AND INGREDIENT: STATIONARY NOISE WITH A KNOWN PSD

The inner product of Eq. (2) is based on the assumption that the Gaussian detector noise is stationary: its mean and covariance are constant in time for the duration of the analyzed data segment. The frequency-domain covariance matrix is then diagonal with a (assumed known) variance related to $S_\alpha(f)$. Neither noise stationarity nor a perfect knowledge of $S_\alpha(f)$ are strictly true, introducing a potential systematic in parameter estimation.

Tidal effects influence the signal for a few tens of milliseconds, during which the noise is most likely stationary. However, the entire signal lasts for longer as a GW170817-like signal is 2(5)×10$^3$ minutes from merger [44], putting a strain on the stationarity assumption. Efforts to subtract nonstationary noise [84, 85] or correct for it at the analysis level [86] are under way, with a further option of abandoning the frequency domain altogether [87, 88]. Spectral lines in the data [89–91] and finite analysis segments [92] can also introduce nondiagonal terms in the covariance matrix.

Under the assumption of stationarity the noise PSD itself can be computed using either off-source [68, 93] data that assume stationarity from even longer data segments or it can be modeled based on on-source data [89, 94]. Uncertainty in the PSD estimation can also be marginalized over [75, 95–97]. PSD errors could cause parameter
biases [94], however these should primarily affect the amplitude of the GW signal and less so its phase evolution.

To test this we simulate a BNS in Gaussian noise and analyze it with a misestimated noise PSD in the (400, 1000)Hz range. The PSD is based on the LIGO design sensitivity and we alter its strain sensitivity in the relevant frequency range by some percentage compared to the true value used for the simulated data. The resulting Λ posterior is shown in Fig. 2 showing PSD misestimation does not affect tidal parameter recovery for PSD relative errors of up to ±10%.

V. INGREDIENT: DATA

The data d correspond to the relative displacement of the interferometer test masses, tracked through interfering laser light incident on photodetectors. Converting the photodetector output to strain is achieved through a calibration process whose uncertainties could affect parameter estimation if left unaccounted for [98, 99]. Detector calibration relies on detector strain induced by photon calibrators [100], resulting in an estimate for the systematic error and corresponding statistical uncertainty for the detector frequency-domain amplitude and phase response [101–105].

During first half of O3 the calibration uncertainty (systematic and statistical) was determined to be no more than 4° in phase in the LIGO detectors at the 68% level, corresponding to 7° at the 90% level [103], with similar estimates for the second half of O3 [104]. A conservative phase calibration error of 10° is compared in Fig. 3 against the GW phase shift for different NS radii. The phase calibration uncertainty is comparable to the GW dephasing induced by a 100–200m change in the radius.

Astrophysical parameter estimation studies marginalize over calibration uncertainty [106], a procedure that effectively increases measurement uncertainties, though the result is small at current sensitivities. Though calibration uncertainty is typically treated as being uncorrelated between different frequencies, using a physical calibration model that correctly encodes calibration error across frequencies could further mitigate the effect on parameter estimation [107, 108]. Improvements in photon calibration [109], alternative methods such as the Newtonian calibrator [110, 111], and even astrophysical calibration [112, 113] could reduce the impact of calibration error which is currently comparable to the target radius uncertainty of 100m.

VI. INGREDIENT: WAVEFORM MODEL

The final ingredient of the analysis, and the most commonly considered one in the context of systematics, is the waveform model [5]. The effect of waveform systematics has been investigated for GW170817 by employing a diverse set of waveform models including post-Newtonian [6, 114], effective-one-body [115–120], and phenomenological models calibrated to numerical relativity simulations [121–124]. The main conclusion is that current statistical uncertainties dominate over waveform systematics [12, 125].

Studies of simulated signals using a wide variety of waveforms suggest, however, that waveform systematics could become significant for ρ ∼100 [5, 126–128], corresponding to GW170817 at design sensitivity or 50 sources at the A+ timescale. Waveform biases increase with Λ, and thus less massive or bigger NSs, and could be due to modeling error in the point-particle or the tidal sectors of the waveform [5]. Additionally, numerical errors in the numerical relativity simulations used to calibrate the waveform models could influence results at a similar SNR [126] as in some cases waveform and simulation errors can be comparable [129–134]. This level of systematic error will be comparable to statistical errors for detectors with A+ sensitivity, once the statistical uncertainty in Λ (R) reaches 200 (0.5-1km) at the 90% credible level.

We quantify the waveform accuracy required to achieve the radius measurement projected for Voyager and 3G detectors in Fig. 3. Using the IMRPhenomD_NRtidalv2 [123] waveform model we compute the frequency-domain phase difference between signals emitted by BNSs of 1.4M⊙ as a function of the NS radius relative to 13km. A 0.5km radius difference induces a 3–4 rad dephasing; waveform systematics need to be kept below this level to achieve such radius accuracy. Errors in numerical simulations are typically quoted at ~1rad [133], though these refer to time-domain phase and are not directly comparable.

The dephasing is larger at higher frequencies where the detector sensitivity decreases, thus making absolute phase differences less informative. As an alternative we also plot a noise-weighted phase difference, derived in [135] as the effective cycles of phase that a specific effect (here the tidal deformation) contributes to the waveform. The effective cycles of phase are related to an upper limit on the Bayes factor that the effect in question is detectable in the data [135]. We use the LIGO design sensitivity, though the level of the PSD does not affect the noise-weighted phase difference, only its shape.

The noise-weighted phase difference is a less sensitive function of the upper frequency cutoff, but it still increases for frequencies up to 1400Hz. A 0.5km change in the radius leads to a noise-weighted phase change of ~0.3rad, again setting the threshold for waveform accuracy in other to achieve such a radius measurement.

VII. CONCLUSIONS

The astounding precision expected to be achieved with GW observations in the next decade(s) [136] brings to the forefront all analysis assumptions and ingredients to be examined under the light of systematic errors. In the context of BNSs we study the frontier for measuring...
NS properties by quantifying the limiting effect of various sources of systematics. Improved waveform models are needed for a radius measurement of $\Delta R \sim 1$km at the timescale of $\Lambda +$, while improvements in detector calibration are needed at the timescale of Voyager for $\Delta R \sim 200$m. Both waveform and calibration errors would introduce correlated biases for different sources, and thus increasingly affect combined constraints.

The importance of non-Gaussian noise depends on the type of glitches in future detectors, but methods to simultaneously model signals and glitches or multiple signals will mitigate the effect. The effect of nonstationary noise depends on the detector low frequency performance and hence the signal length. Misestimating the noise PSD has a negligible effect on tidal inference. Non-Gaussianity, nonstationarity, and PSD misestimation would likely affect each signal in a unique way, and thus not accumulate in combined constraints.

This discussion concerns extracting binary parameters, namely masses and tidal deformability, from GW data. The projected constraints were translated to NS radius assuming a perfect conversion method for illustration. In reality, additional analysis steps are required to combine information from multiple sources and obtain combined constraints on NS radii or features such as phase transitions $\Delta\Phi$ (rad). This is achieved through hierarchical inference $\Delta\Phi$ (rad) which accounts for parameters correlations such as the ones between $\Lambda$ and mass ratio or the distance and the inclination, statistical uncertainties, and selection effects. Additional systematic errors here would be the NS equation of state model employed $\Delta\Phi$ (rad), the estimation of the selection function $\Delta\Phi$ (rad), and the population model for NS spins and masses $\Delta\Phi$ (rad). The latter, if neglected, could bias the NS equation of state after 25-50 events $\Delta\Phi$ (rad) but the effect can be fully addressed by simultaneously inferring the NS mass and spin distribution with the equation of state $\Delta\Phi$ (rad).

Finally, this study examines only GW data and their analysis. X-ray observations $\Delta\Phi$ (rad) will provide additional information about NS properties yielding overall tighter constraints. The relation between systematic errors in the different observations and their effect on the decreasing statistical uncertainty needs to be explored.

**VIII. ACKNOWLEDGEMENTS**

We thank the participants of the Aspen Center program “Exploring Extreme Matter in the Era of Multimessenger Astronomy: from the Cosmos to Quarks” for numerous discussions that motivated this study. We thank Max Isi for comments on the manuscript. This work was initiated at Aspen Center for Physics, which is supported by National Science Foundation grant PHY-1607611. This work was partially supported by a grant from the Simons Foundation. This research has made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org), a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by Polish and Hungarian institutes. This material is based upon work supported by NSF’s LIGO Laboratory which is a major facility fully funded by the National Science Foundation. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459. Software: *gwy* [162], *matplotlib* [163], *gwtoolbox* [54].

\[ \frac{\Delta R}{\text{km}} \]

\[ \Delta\Phi \text{ (rad)} \]

\[ \text{800 Hz, 1200 Hz, 1000 Hz, 1400 Hz} \]

\[ \text{Absolute phase, Noise-weighted phase} \]

\[ \text{FIG. 3. GW frequency-domain phase difference between signals from BNSs with different radii relative to 13km up to different frequencies. Solid (dashed) lines denote absolute (noise-weighted) phase differences. The shaded region marks a 10^6 calibration uncertainty, relevant only for the absolute phase difference. Unbiased tidal and radius inference hinges on waveform models achieving phase evolution accuracies better than the induced tidal effect.} \]

\[ \Delta R \text{ (km)} \]

\[ \Delta\Phi \text{ (rad)} \]

\[ 10^{-2}, 10^{-1}, 10^{0}, 10^{1} \]

\[ \text{800 Hz, 1000 Hz, 1200 Hz, 1400 Hz} \]

\[ \text{Absolute phase, Noise-weighted phase} \]

\[ \text{FIG. 3. GW frequency-domain phase difference between signals from BNSs with different radii relative to 13km up to different frequencies. Solid (dashed) lines denote absolute (noise-weighted) phase differences. The shaded region marks a 10^6 calibration uncertainty, relevant only for the absolute phase difference. Unbiased tidal and radius inference hinges on waveform models achieving phase evolution accuracies better than the induced tidal effect.} \]
[158] T. E. Riley et al., (2021), arXiv:2105.06980 [astro-ph.HE].
[159] S. Bogdanov et al., Astrophys. J. Lett. 887, L25 (2019), arXiv:1912.05706 [astro-ph.HE].
[160] S. Bogdanov et al., Astrophys. J. Lett. 887, L26 (2019), arXiv:1912.05707 [astro-ph.HE].
[161] S. Bogdanov et al., (2021), arXiv:2104.06928 [astro-ph.HE].
[162] D. Macleod, A. L. Urban, S. Coughlin, T. Massinger, M. Pitkin, paulaltin, J. Areeda, E. Quintero, T. G. Badger, L. Singer, and K. Leinweber, “gwpy/gwpy: 1.0.1,” (2020).
[163] J. D. Hunter, Computing In Science & Engineering 9, 90 (2007).