Identifying strongly lensed gravitational waves through their phase consistency

Jose María Ezquiaga,[1,*] Wayne Hu,[2,†] and Rico K. L. Lo[3,‡]

[1] *Niels Bohr International Academy, Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen, Denmark*
[2] *Kavli Institute for Cosmological Physics, Department of Astronomy & Astrophysics, Enrico Fermi Institute, The University of Chicago, Chicago, IL 60637, USA*
[3] *LIGO, California Institute of Technology, Pasadena, CA 91125, USA*

(Dated: August 15, 2023)

Strongly lensed gravitational waves (GWs) from binary coalescence manifest as repeated chirps from the original merger. At the detectors, the phase of the lensed GWs and its arrival time differences will be consistent modulo a fixed constant phase shift. We develop a fast and reliable method to efficiently reject event pairs that are not-lensed copies and appropriately rank the most interesting candidates. Our method exploits that detector phases are the best measured GW parameter, with errors only of a fraction of a radian and differences across the frequency band that are better measured than the chirp mass. The arrival time phase differences also avoid the shortcomings of looking for overlaps in highly non-Gaussian sky maps. Our basic statistic determining the consistency with lensing is the distance between the phase posteriors of two events and it directly provides information about the lens-source geometry which helps inform electromagnetic followups. We demonstrate that for simulated signals of not-lensed binaries with many shared parameters none of the pairs have phases closer than 3σ, and most cases reject the lensing hypothesis by 5σ. Looking at the latest catalog, GWTC3, we find that only 6% of the pairs are consistent with lensing at 99% confidence level. Moreover, we reject about half of the pairs that would otherwise favor lensing by their parameter overlaps and demonstrate good correlation with detailed joint parameter estimation results. This reduction of the false alarm rate will be of paramount importance in the upcoming observing runs and the eventual discovery of lensed GWs. Our code is publicly available and could be applied beyond lensing to test possible deviations in the phase evolution from modified theories of gravity and constrain GW birefringence.

I. INTRODUCTION

Ground-based gravitational-wave (GW) observatories coherently detect the space-time perturbations produced by merging compact objects such as binary black holes. The phase evolution of these signals encode information about their gravitational properties, astrophysical origin and cosmological propagation. LIGO–Virgo–KAGRA (LVK) [1–3] detectors have already accumulated observations of about a hundred compact binaries during the first three observing runs [4].

Compact binary coalescence signals are inevitably affected by the intervening matter along their travel path. The effect of such gravitational interaction is typically negligible. However, for sufficient alignment between the source and the matter distribution, the latter acts as a lens magnifying and possibly distorting the original signal. For large lenses such as galaxies or clusters of galaxies, strong lensing effects can lead to nearly identical chirps of the same signal arriving at the detectors days to months apart. The probability of observing strongly lensed gravitational waves depends heavily on the source and lens populations, but for galaxy lenses rates can be up to 1 lensed event in every 1000 events [5]. As we expect hundreds to thousands of mergers within the fourth and fifth observing runs, the first GW lensing detection could be within the coming years.

Confidently identifying strongly lensed gravitational waves is however a challenging task. Each of the repeated instances of the lensed signals hide within large catalogs in which not-lensed events can easily mimic lensed ones [6]. Past strongly lensed searches [7–10] have focused on computing the overlap between parameters of different events in order to account for the likelihood of lensing [11] before launching a more computationally costly joint parameter estimation [12, 13]. However, with current sensitivities, the physical parameters describing the binary are not well constrained and degenerate among each other, making it common to find large overlaps in not-lensed pairs. Therefore, in order not to miss any lensed event, many false alarms would have to be followed up.

In this work we develop a fast and reliable method to identify strongly lensed candidates through the consistency of the best measured GW quantities, the phases at the detectors. We borrow some of the tension statistics developed in cosmology [14] to compute the confidence level (CL) of agreement between two events. The advantage of our method is that it establishes a well defined measure of how inconsistent (in "tension") a pair of GWs is with the strong lensing hypothesis. Thus, by construction, our method generates a highly complete catalog of lensing candidates while also reducing the number of false alarms compared to previous methods. A code that implements our method, `phazap`, is publicly available.

We begin the paper in Sec. II showing how to effi-

* jose.ezquiaga@nbi.ku.dk
† whu@background.uchicago.edu
‡ kllo@caltech.edu

ciently reconstruct the main GW observables at the detectors — detector phases, time delay phases and polarization states, from pre-existing parameter estimation posteriors. In Sec. III we then show how those observables are affected by strong lensing. In Sec. IV we outline our method to analyze lensed candidates. We apply the method to a set of simulated events and to the latest LVK catalog in Secs. V and VI respectively. Our results are compared with other methods in Sec. VII. We then explore the implications for current and future strong lensing searches in Sec. VIII. We conclude in Sec. IX with the main results and possible extensions of our method including those beyond lensing. In a series of appendices we provide further details on our method and conventions. In particular, we detail our reference frame conventions, demonstrate how to compute the phase at a new frequency, show how to break the sky localization bimodalities, estimate the errors of the detector phases, specify the settings of the simulated GW events and report the detector phases for the real GWs analyzed in Apps. A, B, C, D, E and F respectively.

II. GRAVITATIONAL WAVE OBSERVABLES

The gravitational wave emission of a compact binary coalescence depends on the intrinsic parameters of the source, such as the masses (m_i) and spin vectors ($\vec{S}_i$), and can be described by the two tensorial polarizations h_+ and $h_\times$. The detected GW strain $h_d(t)$ at each detector also depends on the extrinsic parameters: the distance to the source d_L, its sky location as determined by the line of sight $\vec{n} = \{\text{ra}, \text{dec}\}$, the polarization orientation ψ and arrival time t_d, the orbital inclination ι and phase ϕ_{ref} at a reference frequency f_{ref}:

$$\begin{aligned} h_d(t - t_d) &= F_+ h_+ + F_\times h_\times \\ &= \int_{-\infty}^{\infty} \tilde{h}_d(f) e^{2\pi i f(t-t_d)} df, \end{aligned} \tag{1}$$

where $F_{+,\times} = F_{+,\times}(\vec{n}, \psi, t_d)$ are the antenna pattern functions, $\tilde{h}_d(f, m_{1,2}, \vec{S}_{1,2}, \iota, d_L, \phi_{\text{ref}})$ is the complex-valued Fourier transform of the real-valued function $h_d(t)$, and the subscript d labels quantities evaluated at each detector (see App. A for conventions on the different reference frames). For the rest of the paper we will focus on the positive frequency modes, whose negative frequency counterparts can be computed by the reality condition $\tilde{h}(f) = \tilde{h}^*(-f)$.

Before projecting into the detector frame, we can further decompose the frequency domain signal of a given polarization in the radiation frame into a sum of frequency-dependent amplitudes and phases for each multipole mode at emission. For cases with precession, this is usually done in a coprecessing frame and there the dominant mode is $l = |m| = 2$ which we will refer to as the 22-mode. Higher order modes are only sizable when the binary has highly asymmetric component masses or eccentricity. On the other hand since precession and higher modes are part of the standard analysis they impact parameter degeneracies and inference regardless of whether they are detected in a given event or not. For our analysis, we will therefore use waveform models containing these effects.

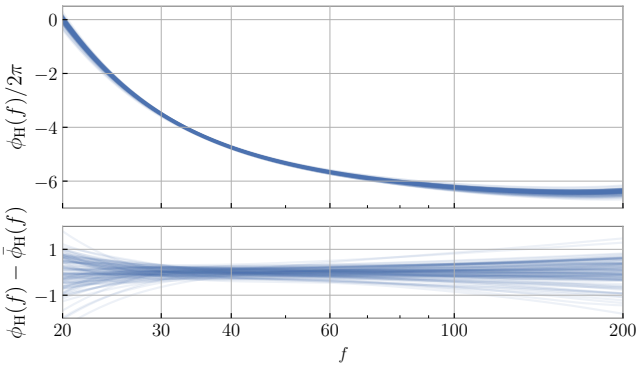


FIG. 1. Reconstructed detector phase as a function of frequency for the first detected event, GW150914 [16], in the Hanford (H) detector. In this frequency range, the phase runs approximately over six cycles with errors of a fraction of a radian with respect to the mean evolution $\bar{\phi}_{\rm H}(f)$. Data is from the latest GWTC-2.1 catalog [17].

From the observation of the amplitudes, phases and arrival times at each detector one can reconstruct the properties of the detected signal with Bayesian parameter estimation. Despite the precise measurement of the frequency-dependent phase of the signal, the high-dimensionality of the parameter space of compact binary coalescences, 15 dimensions for quasi-circular binaries accounting for 8 intrinsic $\{m_1, m_2, \vec{S}_1, \vec{S}_2\}$ and 7 extrinsic $\{\text{ra}, \text{dec}, \psi, \iota, d_L, t_{\text{ref}}, \phi_{\text{ref}}\}$, leads to poorly constrained marginalized posteriors for each of the parameters individually and complicated joint posteriors due to their degeneracies, see e.g. [15]. Still, one can computationally efficiently use the *full* parameter estimation to *derive* the well constrained posterior probability for the phases of the 22-mode at every frequency and detector, which we hereafter refer to as the "detector phases":

$$\phi_d(f) = \phi_{22}(f) + \chi^d_{22}(\vec{n}, \psi', \iota', t_d)\,, \tag{2}$$

where $\phi_{22}(f)$ is the global phase of the 22-mode,[1] $\chi_{lm} = \arctan[F_+, a_{lm}(\iota')F_\times]$, and $a_{lm} = (1 - r_{lm})/(1 + r_{lm})$ with the ratio of the circular polarization amplitudes $r_{lm} = |h^{lm}_L|/|h^{lm}_R|$, see App. B for technical details on how to derive this expression with a waveform-based decomposition of the polarization state. For the 22-mode:

[1] Specifically, for precessing binaries this is associated to the phase of the 22-mode in the co-precessing $\vec{L}$-frame. With our conventions and exploiting the equatorial symmetry, this is built from the phase of the co-precessing $l = 2$, $m = -2$, $+$-polarization mode for positive frequencies.

$a_{22}(\iota) = 2\cos\iota/(1+\cos^2\iota)$. This defines the phase of the 22-component of the detector strain at frequency f at the arrival time $t = t_d$. Note that precession causes a frequency-dependent rotation of the linear polarization angle, $\psi' = \psi + \zeta(f)$, as well as a frequency-dependent inclination angle $\iota' = \iota(f)$ or circular polarization state.

The error in the detected phase scales inversely with the signal-to-noise (SNR) at the detector, $\delta\phi_d \sim 1/\rho_d$ [18, 19]. (see App. D for a simple derivation). While the phase at a given frequency will give us an absolute phase that will be important for the lensing consistency, the phase difference $\Delta\phi_{f,d} = \phi_d(f) - \phi_d(f')$ will be relevant to assess the orbital phase evolution, accounting for the number of cycles, which encodes information about the masses of the binary through the detector frame chirp mass $\mathcal{M}_c = (m_1 m_2)^{3/5}/(m_1+m_2)^{1/5}$.

In Fig. 1 we show the reconstructed phase as a function of frequency at a given detector for GW150914, the first GW [16] using data from the latest analysis [17]. The errors in the phase are a fraction of a radian over a large range of frequencies. Importantly, as we will show later, the phase difference is a better discriminator than the chirp mass due to its smaller relative error and larger range that reduces the probability of consistency by random chance. In addition, it can be seen that the best measured phase is around 40Hz, higher than the fiducial reference frequency of 20Hz but lower than the frequency where the amplitude peaks, $\sim$ 150Hz [16]. This is because at lower frequencies there are more cycles and larger SNR is accumulated as long as the signal is in band. At 40Hz the standard deviation in the phase at the Hanford detector $\phi_{\rm H}$ is 0.15 radians, or $\delta\phi_{\rm H}(40{\rm Hz}) \sim 3/\rho_{\rm H}$. Note that this order unity factor multiplying the $1/\rho_d$ scaling can vary from event to event but we will use ~ 3 as a rough guide for estimation purposes later. In large part, this variation is due to events with high masses and short in-band durations where the 22-phase becomes hard to distinguish especially in the presence of other parameters (see App. D).

In addition to the phase, detected GWs have timing information, which is used to triangulate the sky position of the source. Typically, in the parameter estimation, the arrival time of the signal t_c is defined by the arrival of the maximum of the time domain strain, $\mathcal{A}(t)^2 = |h_+|^2 + |h_\times|^2$, as measured at a common GPS time in the Earth frame. The maximum of the signal is chosen to approximately track the coalescence time of the binary. We can reconstruct the arrival time at each detector as

$$t_d = t_c - \vec{n}\cdot\vec{r}_d/c\,, \tag{3}$$

where $\vec{r}_d$ is the position of each detector. We use the relative arrival time at different detectors to define the additional phases

$$\tau_{d_1 d_2}(f) \equiv 2\pi f \Delta t_{d_1 d_2} = 2\pi f \vec{n}\cdot\vec{r}_{d_2 d_1}/c\,, \tag{4}$$

where $\vec{r}_{d_2 d_1} \equiv \vec{r}_{d_2} - \vec{r}_{d_1}$. The advantage of these time delay phases is that they largely remove degeneracies in localization as we shall now discuss. In general, the arrival time difference between two detectors constrains one angle, $\vec{n}\cdot\vec{r}_{d_1 d_2}$, defining a ring in the sky of possible source locations. With three detectors, two angles are constrained, $\vec{n}\cdot\vec{r}_{d_2 d_1}$ and $\vec{n}\cdot\vec{r}_{d_1 d_3}$, defining two rings in the sky that intersect at two points where the event localization is possible. These two possible localization regions correspond to a reflection symmetry of the time delays across the hemispheres delineated by the plane defined by the position of the three detectors, i.e. distinguished by the sign of $\vec{n}\cdot(\vec{r}_{d_1 d_2}\times\vec{r}_{d_1 d_3})$. As an example, we plot the time delay phase contour lines of $\tau_{\rm HL}$ and $\tau_{\rm HV}$ as a function of right ascension and declination for GW150914 in Fig. 2. The shaded regions corresponds to the 95% CL for the reconstructed time delay phases and their intersection indicates the possible sky localization above and below the plane of the detectors. Note that GW150914 was not measured by Virgo and thus $\tau_{\rm HV}$ is derived from the localization supplied by the two LIGO detectors rather than measured directly. For sufficiently high SNR, the amplitude and polarization information enters into the localization and this information breaks the time delay reflection symmetry. In detector networks with more than three detectors, this degeneracy is also broken and more than two time delay phases can be computed. On the other hand these multiple phases always redundantly parameterize the two sky angles.

Finally, the amplitude information in GW events carries information beyond the localization. The overall amplitude serves to quantify the distance to the source and the lensing magnification while the relative amplitudes

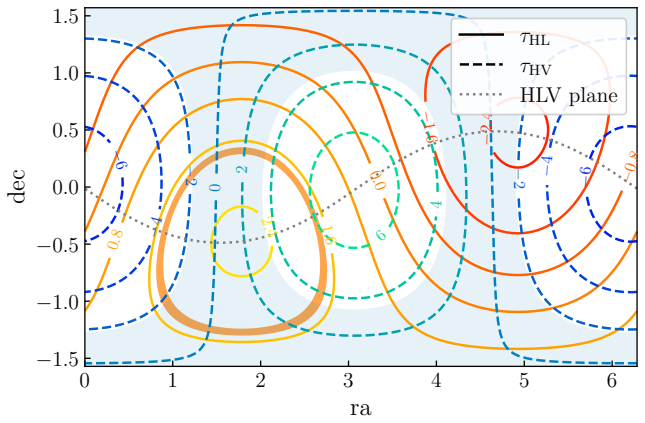


FIG. 2. Time delay phase contour lines as a function of the right ascension (ra) and declination (dec) for GW150914 [16]. Solid lines indicate contours for Hanford and Livingston ($\tau_{\rm HL}$), while dashed lines are for Hanford and Virgo ($\tau_{\rm HV}$). The shaded orange and blue regions correspond to the 95% CL from the reconstructed time delays, $\tau_{\rm HL}$ and $\tau_{\rm HV}$ respectively. The dotted line indicates the plane defined by the position of the HLV detectors. The intersection of both shaded regions occurs in reflection symmetric positions above and below the detector's plane and correspond to a bimodality in the localization from time delays.

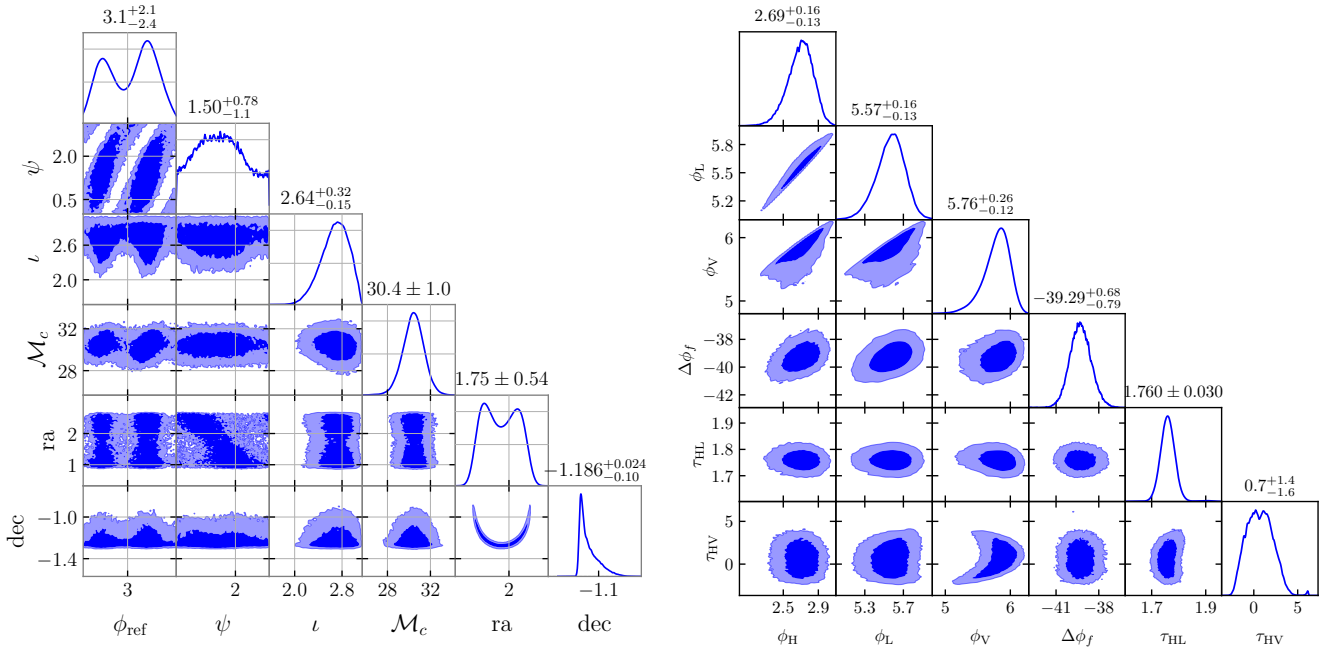


FIG. 3. Comparison between the waveform parameters and the reconstructed phases for GW150914. On the left, the posterior distributions reference phase at the 20 Hz reference frequency ($\phi_{\rm ref}$), the polarization angle (ψ), the inclination at the reference frequency (ι), the right ascension (ra), the declination (dec) and the detector frame chirp mass ($\mathcal{M}_c$). On the right, the detector phases ($\phi_{\rm H,L,V}$) and the time delay phases ($\tau_{\rm HL,HV}$) at our pivot frequency of 40Hz, and the phase evolution between 20 and 100Hz ($\Delta\phi_f$). In both plots the contour lines indicate the 68 and 95% CL, while the quoted ranges in the titles are the median and 68% CL

between detectors help fix the polarization state through the polarization angles β and ζ, respectively (see App. B for details). As with the time delay phases, by deriving the detector phases from the parameter posteriors rather than directly measuring them in each detector, we encode some of this extra information on the polarization state into the inferred parameters.

In summary, for the current network of three detectors LIGO Hanford (H), LIGO Livingston (L) and Virgo (V), we can construct six parameters: three detector phases $\phi_{\rm H}, \phi_{\rm L}, \phi_{\rm V}$ and two arrival-time phases $\tau_{\rm HL}$and τ_{LV} by evaluating (2) and (4) at a given frequency, as well as the orbital phase evolution parameter $\Delta\phi_f$ (our default choice is to reconstruct this phase difference at H, $\Delta\phi_f \equiv \Delta\phi_{f,{\rm H}}$, although the detector dependence of the difference is minimal when derived from the full parameter estimation).

This is to be compared with the five extrinsic parameters ra, dec, $\psi, \iota, \phi_{\rm ref}$ and the intrinsic parameter, the detector frame chirp mass $\mathcal{M}_c$ (as well as the remaining of the original 15 parameters which are marginalized over). By inferring constraints on the former set from the latter set, we efficiently encapsulate the information from parameter estimation into quantities that are closer to the direct observables and hence have reduced degeneracies and multimodality.

To exemplify this transformation, a plot of the reconstructed phases compared to the original parameters for GW150914 is presented in Fig. 3. Here one can also see that the relative error in the orbital phase evolution $\Delta\phi_f$ is smaller than in the chirp mass. Moreover, the errors in the $\phi_{\rm H}$ and $\phi_{\rm L}$ are similar as the SNR in both detectors is comparable and information is shared in the reconstruction, whereas their difference $\phi_{\rm H} - \phi_{\rm L}$ is even better constrained. Similarly despite GW150914 having no Virgo data, the phase that Virgo would have seen, $\phi_{\rm V}$, is still constrained from the parameter estimation, although not as well as H and L due to parameter degeneracies that are not broken by direct measurement. Likewise the time delay $\tau_{\rm HV}$ is also constrained by inference. For this loud event the full sky localization degeneracy is partially broken and the 95% localization region contains only the lower intersection region in Fig. 2 contributes. This also helps in having a well constrained reconstructed phase in Virgo. However, in lower SNR events, as we will see later, the phase in the detectors not online is typically unconstrained.

Overall, though, the better constraints using our derived detector phase parameters vs. the original binary parameters illustrated in Fig. 3 reflects the advantage of our approach in capturing the GW observable and carries over to lensing identification as we shall see next.

III. GRAVITATIONAL WAVE STRONG LENSING

In the regime of strong lensing, multiple instances of the same event are produced, each of them acquiring a change in amplitude, arrival time and phase [20]

$$\begin{aligned} \tilde{h}^j_{\rm lensed}(f) &= \sqrt{|\mu_j|} e^{i n_j \pi/2} \tilde{h}(f)\,, \\ t^j_d &= t_d + \Delta t_j \end{aligned} \tag{5}$$

with $n_j = 0, 1, 2$ for type I, II and III images respectively and Δt_j as the time delay compared to the signal without lensing. Therefore, if the detectors were in the same positions relative to the source for each image (see §IV), the detector phase of two lensed images j and k should only differ by their Morse phase

$$\Delta\phi^d_{jk} = (n_j - n_k)\pi/2\,, \tag{6}$$

meaning that their phase should be identical if they are the same image type or differ by a multiple of $\pi/2$ otherwise. Distances to the source and arrival times will be biased by the magnification and time delay and thus cannot be used to reject the lensing hypothesis in a given pair of events, though with additional astrophysical assumptions both can be informative.

In the regime of geometric optics, the polarization rotation is of the order of the deflection angle and, as a consequence, negligible for ground-based detectors [21]. All instances of the original event should then have the same polarization state as a function of frequency. For example, the degree of circular polarization can be parametrized by

$$r = |h_L|/|h_R|\,, \tag{7}$$

the ratio of the left- and right-handed polarization amplitudes (see App. B for other quantities defining the polarizations). As discussed in the previous section, this information is encoded in the detector phases inferred from parameter estimation. Future observations of the polarization states could be used, for instance, to test the spin consistency among the lensed events.

Altogether, in order to identify strongly lensed GWs we can look at the consistency of the detector phases and time delays. For the current network of detectors, the polarization state is not separately well constrained and for that reason we do not consider its consistency directly. Focusing on the GW detector phases is advantageous over individual waveform parameters since their large degeneracies and measurement uncertainties make them easily overlap by chance, leading to high false alarm rates [6]. This is evident, for example, when looking at Fig. 3. Previous analyses have ranked lensed candidates by computing overlaps in the redshifted component masses, spins and sky positions as we will discuss in Sec. VII.

It is also important to note that for this consistency test to work, the standard parameter estimation should use models that are a good description of the lensed signals. This is the case for the type I and III images, but the phase shift of type II images could induce waveform distortions that are not included within the family of (not-lensed) general relativity waveforms [21]. Therefore, for type II images the parameter estimation of sky positions and inclinations could be biased [22]. However, we have checked with simulated type II images, where the Morse phase shift is applied directly to each frequency, the detected phase is well recovered and consistent with the expected signal-to-noise ratios of current detectors. Extraordinarily loud type II images could be identified directly by including the phase factor in the parameter estimation. We expect similar results to hold for strongly lensed images that are also lensed by smaller lenses inducing waveform distortions: if the distortion is large enough to bias the parameter estimation, it could be identified as lensed by targeted searches.

IV. METHOD

Our goal is to design a fast and reliable method that rejects as many non-lensed events as possible while selecting the most promising events for more detailed study. For that reason, instead of re-analyzing each candidate pair using joint parameter estimation, we efficiently post-process the original posterior samples from the parameter estimation that are performed for every detection.

We analyze the events of a given GW catalog in pairs, aiming to first discard those pairs whose phases are not compatible with the strong lensing hypothesis and then rank the remaining candidates for priority in a joint parameter estimation campaign. We generically label the events in the pair as "event 1" and "event 2". To compare the detector phases of two events, we need first to fix the reference frequency and establish a common detector reference frame.

We fix the reference frequency to 40Hz, which for planar, quasi-circular inspairaling binaries and advanced LIGO and Virgo detectors approximately gives the best measured phase. Note that this choice is different from the standard $f_{\rm ref} = 20$Hz of LVK catalogs, but it is easy to compute the detector phases at a new frequency: see App. B. Unlike Ref. [15], we cannot optimize this frequency per event since we need a common parameterization between pairs. Still, if a more optimal strategy is desired, one could alternately choose the frequency for which the joint errors are minimized at the computational expense of re-processing per pair. As a bonus, at the best measured frequency, the detected phases decorrelate from other intrinsic parameters such as the chirp mass. This is not in general the case at other frequencies. We defer such optimizations to future studies.

In place of the chirp mass we take the phase difference across the widest frequency range where the detected phase is well constrained. Our results are not very sensitive to the precise value since the larger phase errors

of a larger range are compensated by the increase in the number of cycles. Our default choice is from 20 to 100Hz, which matches the well-constrained region of the fiducial case displayed in Fig. 1. Similarly, we compute the orbital phase evolution at Hanford, although the detector's dependence essentially drops out.

As discussed in the previous section, the detector phases between the two lensed events will differ by multiples of $\pi/2$ only if the detectors are in a common reference frame between events. We fix the reference frame to the arrival of event 1. For event 2, this can be achieved by shifting the mean of the arrival of event 2:

$$t_2 \to t_2 - \langle t_2 \rangle + \langle t_1 \rangle \,. \tag{8}$$

Note that this changes the detector phases of event 2 by changing χ_d in (2). With this choice we are comparing the actual "detected phases" of event 1 with the inferred "detector phases" of event 2, defined by what the detectors would have seen, had they been in the orientations of event 1 given the parameter estimation of event 2. The main caveat in this approach is that the detector phases of event 2 will generally have poorer constraints, larger degeneracies, and less Gaussian distributed posteriors than the detected phases of event 2.

Likewise, the time delay phases for event 2 reflect only the localization of event 2 through Eq. (4) and do not directly reflect the arrival times of event 2 at the actual positions of the detectors. Therefore, due to the ring degeneracy of a localization inferred from a phase difference, explained below Eq. (4), event 2 time delay phases will generically inherit a ring-like structure that is typically bimodal. This bimodality can be separated into individual modes by distinguishing the samples that come from above or below the plane of the detectors in their actual positions for event 2 (see App. C for details). By analyzing these two modes separately, we mitigate the non-Gaussianity of the inferred time delays for event 2.

In order to quantify the consistency of a set of posterior samples of parameters of event 1, θ_1, with the same parameters of event 2, θ_2, we focus on the probability density of their difference $\Delta\theta = \theta_1 - \theta_2$ [14]:

$$P(\Delta\theta) = \int P_{\theta_1}(\theta) P_{\theta_2}(\theta - \Delta\theta) \mathrm{d}\theta \,, \tag{9}$$

which is the convolution of the two posterior probabilities P_{θ_1} and P_{θ_1}. Support for large values of $|\Delta\theta|$ indicates that the two parameter sets are not compatible.

For any two posteriors, we can compute the distribution of $\Delta\theta$ numerically and set confidence intervals for consistency [23]. Moreover, if the two posteriors are approximately Gaussian, we can estimate consistency very simply. The width σ of the Gaussian $P(\Delta\theta)$ is determined by the sum of the covariances C_1 and C_2, so that the distance in units of σ is given by [14]

$$D_{12} \equiv D(\theta_1, \theta_2) = \sqrt{\Delta\theta^T (C_1 + C_2)^{-1} \Delta\theta} \,. \tag{10}$$

Inconsistencies between data sets can be defined at different confidence levels (CL). Note that for non-Gaussian posteriors the Gaussian approximation in Eq. (10) will typically lead to overly conservative distances, as residual multi-modality will inflate the covariances leading to shorter distances. An example of this behavior for the time delay phases, which can be multimodal as described above, is presented in App. C.

For a given number of χ^2-distributed parameters, we can relate the distance to the CL. For example, for 1/4/6 degrees of freedom, a 95% CL corresponds to $2.0/3.1/3.5\sigma$, 99% CL corresponds to $2.6/3.6/4.1\sigma$, and 99.9% CL corresponds to $3.3/4.3/4.7\sigma$. If the uncertainties in a given parameter make the data uninformative the effective degrees of freedom $N_{\rm eff}$ that are constrained by the data can be smaller than the total N. In such cases the posterior has support across the whole prior range and we can quantify the effective number of degrees of freedom as

$$N_{\rm eff} = \mathrm{tr}\left[\left(C_{\rm prior} + C_1 + C_2 \right)^{-1} C_{\rm prior} \right] \,, \tag{11}$$

where $C_{\rm prior}$ is the covariance of the priors. For example, for a bounded flat distribution in a single parameter the covariance is $C = (\theta_{\rm max} - \theta_{\rm min})^2/12$ [14]. If the data is informative, $C_{\rm prior} \gg C_1 + C_2$, then $N_{\rm eff} \to N$, the true number, while in the opposite limit, $C_{\rm prior} \ll C_1 + C_2$, one gets $N_{\rm eff} \to 0$. In practice given our 6 actual parameters, we compute $N_{\rm eff}$ for the 3 detector phases only so that the total number of effective parameters is $N_{\rm tot} = N_{\rm eff} + 3$, effectively taking infinite prior covariances for the other parameters. Because detector phases can differ under the lensing hypothesis by multiples of $\pi/2$, we set the detector phase priors ϕ_d for an informative measurement so that $C_{\rm prior} = \delta_{d_1,d_2} (\pi/2)^2/12$ for $d_1, d_2 \in \mathrm{H, L, V}$.

Since strong lensing introduces a constant phase shift of the detected phases, we compute the distance for all the possible phase shifts:

$$D^n_{12} = D(\phi_1, \phi_2 + n\pi/2) \,, \tag{12}$$

for $n = 0, \pm 1, \pm 2$. A true lensed pair will have consistent parameters in both orderings and so we define a pair ordered distance that maximizes over the ordering:

$$D^n_J = \max(D^n_{12}, D^{-n}_{21}) \,, \tag{13}$$

where the relative phase shift and hence n switches sign under a change in ordering. Conversely since each n is a possible lensing outcome, we minimize the pair ordered distance over lensing types to obtain the final joint distance

$$D_J = \min_n (D^n_J) \,. \tag{14}$$

This D_J will be our basic metric in the analysis. Note that by virtue of this formalism, we get directly the phase shift n most consistent with the lensing hypothesis, which carries information about the image types and therefore

the lens model. This is not possible with methods that compute the posterior overlaps that do not include phase information (see Sec. VII).

Finally, with the joint distance D_J and the total number of effective parameters $N_{\rm tot}$ we compute the confidence level of consistency with the lensing hypothesis and produce a catalog of pairs that are consistent with lensing. In our analysis of simulated and real events in Secs. V and VI we draw the line at 99% CL but our same analysis could trivially output results at different levels, trading purity for completeness, though the probabilistic inference will depend on how Gaussian the tails of the parameter posterior are.

Since pairs that pass the distance consistency threshold may do so simply because the parameter errors are so large as to encompass lensing as a possibility, we can further rank them according to a statistic that measures how well the parameters are constrained. Because the errors in the parameters for a given event scale inversely with the SNR, one possibility is to rank pairs in terms of expected scaling of total errors with the network SNR

$$R = \sqrt{\rho_{\rm ntw,1}^{-2} + \rho_{\rm ntw,2}^{-2}} \tag{15}$$

where

$$\rho_{\rm ntw}^2 = \rho_{\rm H}^2 + \rho_{\rm L}^2 + \rho_{\rm V}^2 \tag{16}$$

is the median "optimal SNR" across the posterior parameter distribution (see Eq. D3), combining those of each detector. The drawback is that there can be events with high SNR and poor parameter errors when compared relative to different events with very different binary parameters. In fact, as discussed above when the 22-mode is the dominant detection, all detector phase errors should scale with R with the same order unity coefficient (~ 3) so that a parameter error based ranking can be constructed using just the parameter volume of the three detector phases $\theta \in \{\phi_{\rm H}, \phi_{\rm L}, \phi_{\rm V}\}$

$$V_\theta = \sqrt{\det(C_\theta)}\,. \tag{17}$$

For a pair we take the sum over each volume

$$V_J = V_1 + V_2\,. \tag{18}$$

Unlike for the parameter distance and $N_{\rm eff}$, this covariance and the associated volume is computed in the reference frame of individual events with the detectors in their true orientations so as to better reflect the scaling with SNR of all of the derived parameters.

In our default analysis we compute the parameter distance and consistency of the six phases: $\{\phi_{\rm H}, \phi_{\rm L}, \phi_{\rm V}, \Delta\phi, \tau_{\rm HL}, \tau_{\rm LV}\}$. This can be extended to include any additional parameter, properly including their correlations. In particular, in the future the polarization amplitude ratio $r(f)$ could be an additional discriminator that measures spin or precession consistency, though it is currently not independently well measured.

As we will show in the next sections, since the phase difference is the more constraining parameter at the moment, when constructing a candidate lensing catalog we recommend employing both D_J and $\Delta\phi_f$, rejecting those pairs inconsistent with either at the chosen CL.

Our method is implemented in a public code: `phazap`[2], which includes both the post processing tools to rapidly obtain the detector phases as well as the tension statistics that are relevant for lensing searches.

V. ANALYSIS OF SIMULATED EVENTS

In order to test our method we simulate a set of lensed and not-lensed GW detections, which we refer to as injections. Our set of injections contains a reference GW event with zero phase shift denoted "Type I", an image of this event (demagnified and phase shifted by $\pi/2$) denoted "Type 2", and a third unrelated event which is not-lensed, denoted "Not-lensed". This accounts for 1 lensed pair (type I & II) and 2 not-lensed pairs (type I/II & Not-lensed). Within each set we tune their luminosity distance so that there is a low-SNR configuration ($\rho_{\rm ntw} \sim 12 - 16$) and a high-SNR configuration ($\rho_{\rm ntw} \sim 22 - 30$) across the detector network. We study a LIGO-only two detector network (HL) and a LIGO-Virgo three detector network (HLV). In all the cases we perform the standard parameter estimation that is applied to real data in LVK catalogs using `bilby` [24] and the projected sensitivities of LIGO and Virgo for the fourth observing run, see App. E for details.

We first consider the case of a "vanilla" binary black hole with masses comparable to GW150914, i.e. $\mathcal{M}_c \sim 30 M_\odot$. In order to test the method in a situation in which standard posterior overlap analysis could lead to false alarms, we consider the case in which the non-lensed injection has the same intrinsic parameters as the lensed injection. We only change the phase, polarization angle and sky position. The sky position is chosen so that for a LIGO-only detector network the sky maps of the lensed and not-lensed events overlap, i.e. on the degeneracy ring of the time delay $\tau_{\rm HL}$ (see Fig. 2). Details of these injections are given in Tab. II, Injections 1-12.

To get a sense of the method, we first consider the most optimistic scenario with high SNR and an HLV network. We present the posterior distributions for the different phases of the three injection in Fig. 4. The lensed pair, Type I - Type II, displays agreement when shifting the detector phases by $-\pi/2$, while it is clear by eye that the not-lensed event shows inconsistency with them. To quantify the (in)consistency of the pairs, we compute the distance D_J of the simulated lensed and not-lensed pairs. For the lensed pair, the minimum distance in the six dimensional phase space is 0.6 when shifting the type II image phases by $-\pi/2$. Any other phase

[2] https://github.com/ezquiaga/phazap

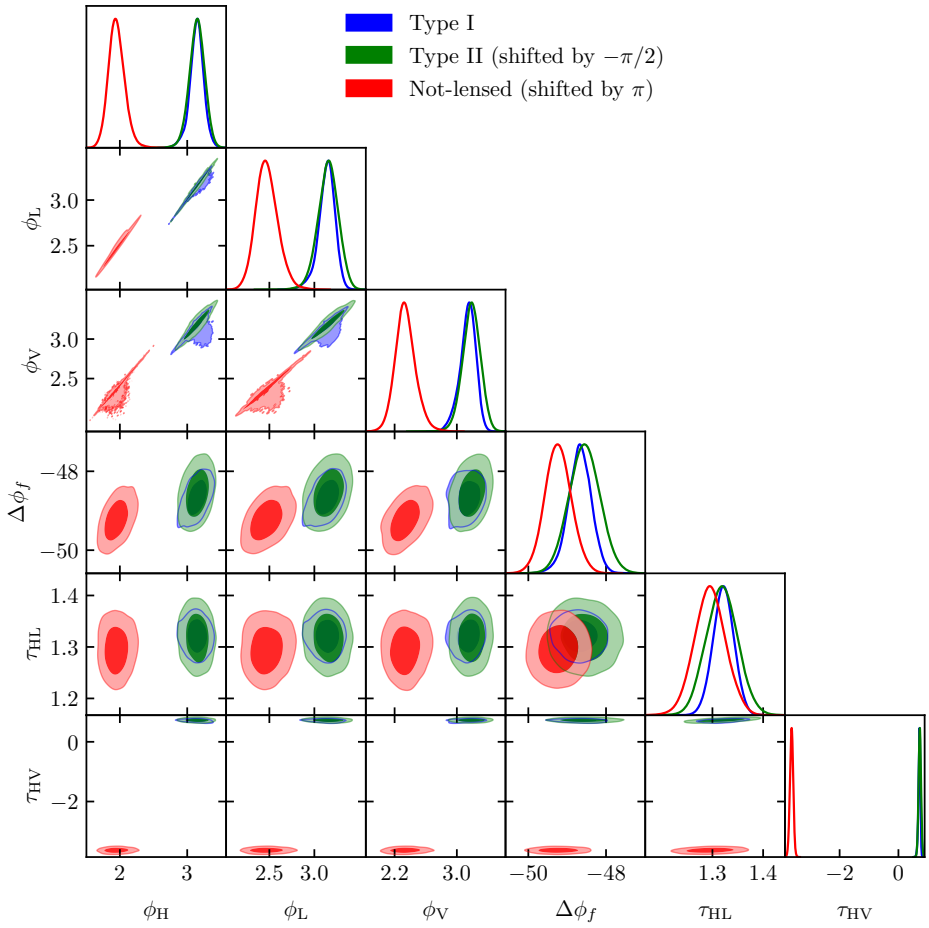


FIG. 4. Lensing consistency test for a simulated pair of lensed events (type I and type II images) and a not-lensed event. The reconstructed phases at each detector (ϕ_d) and the time delay phases ($\tau_{d_1 d_2}$) are computed at 40Hz, while the orbital phase evolution ($\Delta\phi_f$) is between 20 and 100Hz. We are comparing the consistency of both the type I image and the not-lensed event to the type II image. For the three injections the posteriors correspond to the high signal-to-noise, 3-detector configuration (specifically Injections 10-12 described in App. E). In this six dimensional phase space, the lensed pair is consistent within a distance of 0.6, while the not-lensed pairs are rejected with distances > 70. Contours are drawn at 68 and 95% CL.

shift, $0, \pi/2, \pm\pi$, would lead to distances > 11. On the other hand, for the not-lensed pairs, all possible phase shift configurations lead to disagreements with $D_J > 70$. This large distance is dominated by the time delay phase $\tau_{\rm HV}$ that alone rejects the lensing hypothesis with a distance of 67. However, the three detector phases alone would also strongly reject lensing, with a three dimensional distance $D_{\phi_d} > 15$, thanks to the well constrained direction $\phi_{\rm H} - \phi_{\rm L}$. On the other hand, in this case where we have intentionally set the mass to be the same the orbital phase evolution $\Delta\phi_f$ alone would agree within a one dimensional distance of $D_{\Delta\phi_f} \sim 1$, consistent with noise. In any of the pairs, all three detector phases are informative, with a total effective number of detector phases $N_{\rm eff} = 2.7$ and a total number of effective parameters $N_{\rm tot} = 5.7$. For this number of parameters both $D_J > 70$ and $D_{\phi_d} > 15$ indicate an entirely negligible probability of being a lensed pair.

Moving forward, our goal is to test how well we can reject the non-lensed hypothesis for pairs that share many common parameters in different configurations. For this first set of 12 injections we study all the possible low/high-SNR and HL/HLV configurations for a total of $12 \times 11/2 = 66$ possible pairs. Each injection has a different noise realization and their parameters are summarized in App. E. We find that the not-lensed pairs are *always* rejected with distances $D_J > 3$ in the six dimensional phase space or, equivalently, $> 94\%$ CL when taking into account the effective number of informative degrees of freedom. Only 2 of the 32 not-lensed pairs are consistent with lensing at 99% CL even in this situation in which most of the parameters of the simulated GWs are the same. For injections with high-SNR or HLV configurations (9 out 12), non-lensed pairs are *always* rejected with distances larger than 5. On the other hand, we find all lensed pairs to be consistent within a distance of 2.5, where recall that the median expectation for $N_{\rm tot} = 6$ is ~ 2.3.

The distribution of distances for all simulated pairs is presented in Fig. 5. There we also plot the detector phase volume in the vertical axis. The distance itself is the one determining the consistency with the lensing hypothesis. V_J, on the other hand, provides the ranking of the events that pass the lensing consistency test for follow-up analyses. In particular, we clearly see that HLV pairs have smaller volumes as expected.

To understand how the masses of the binary play a role in the method, we then study a lower mass binary with $\mathcal{M}_c \sim 12M_\odot$ (corresponding to Injections 13-15 in Tab. II). The relation between the three injections is the same as before, but in order to enlarge the cases tested we choose a different inclination and sky position to the previous case for injection 3 and change the arrival times in all cases. This defines the "low mass" version of the Type I, Type II and Not-lensed injections. In this case we only consider the low-SNR, HLV configuration. Therefore, when putting this together with the previous injections we analyze a total of 15 injections and 105 possible pairs. We find similar results as before, rejecting the not-lensed pairs with distances > 3, see Fig. 5. No additional not-lensed pair is consistent with lensing at 99% CL, reflecting the fact that the original set was designed to be difficult to distinguish, and so the false-alarm rate goes down to 2 out of 70 of the injection pairs. Moreover, this new set of injections allow us to compare with the previous injections to see how two events with all equal parameters but masses and arrival times could be distinguished. Thanks to incorporating the phase difference $\Delta\phi_f$, events with different masses are rejected at higher significance. Moreover, we find that the phase difference can be a more stringent discriminator than the chirp mass. In this case $D_{\Delta\phi_f} \sim 10 D_{\mathcal{M}_c} \sim 100$, though as we shall see in the next section a more typical number for $D_{\Delta\phi_f}/D_{\mathcal{M}_c}$ is $2-3$ for distances that are more consistent with the lensing hypothesis.

Finally, we look at the effect of inclination. We do so by simulating high-SNR HLV triggers with the same parameters of Type I and II, but with a different inclination, $\theta_{JN} = 2.3$ compared to $\theta_{JN} = 0.8$, cf. Injections 10,11 to 16,17 in Table II for details. We find that for all pairs with these different inclinations, the lensing hypothesis is rejected through the phases at the detectors. This implies that ϕ_d encodes the inclination information, or circular polarization ratio r, as expected. We also look for the agreement in the polarization ratio r alone. We find that only HLV-injections are able to discern between the two cases due to parameter degeneracies. In particular, the inclination will be degenerate with other polarization state parameters in generic precessing systems. An example of this behavior is presented in Fig. 15 in App. B.

Altogether, for the 17 injections and 136 pairs, including both HL and HLV configurations, we find that the detector phases can carry relevant information. When computing the number of effective degrees of freedom, we find that all pairs have at least one effective phase, and $> 40\%$ have 2 effective phases, with $> 25\%$ having $N_{\rm eff} > 2.5$ (a compilation of the detector phases posteriors is presented in App. E). However, given our specific choices for the injections, this does not represent a full characterization of an astrophysical population and observing run. For example, in this catalog of simulated events, $\sim 40\%$ of pairs are HLV detections. A detailed study with a fair selection of simulated not-lensed events from an astrophysical population for different observing scenarios is left for future work, though as we shall see next, the known real events provide a proxy.

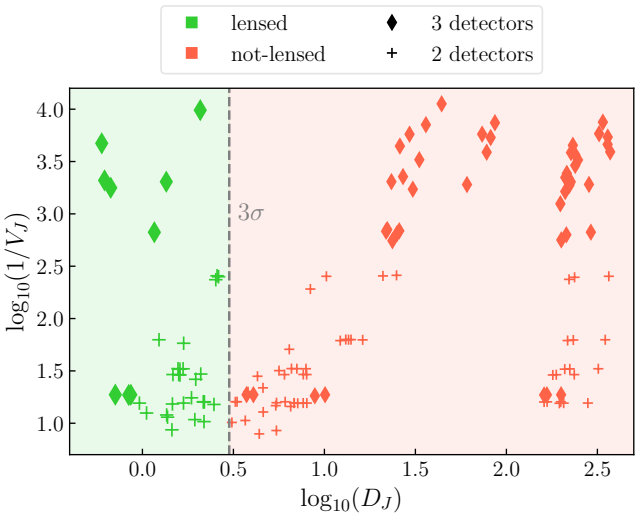


FIG. 5. Analysis of simulated gravitational wave events. We compare the distance (D_J) and phase volume (V_J) statistics for all 136 pairs from 17 injections. Lensed and non-lensed pairs are indicated with green and red colors respectively. Point shapes indicate pairs in which in both events there were HL (+) or HLV-detectors (♦). The vertical dashed line delineates 3σ in distance ($D_J = 3$) and note that all lensed pairs lie below this threshold (shaded green) and all non-lensed pairs above (shaded red). The volume serves to quantify how well constrained the detector phases are compared to their prior volume and can be used to rank candidates for followup analyses, from larger to smaller $1/V_J$.

VI. ANALYSIS OF REAL EVENTS

Having validated our method with simulations in the previous section, we proceed to analyze the latest LVK catalog, GWTC-3 [4].[3] We focus on binary black holes with a high probability of having astrophysical origin, $p_{\rm astro} > 0.8$, and a network SNR > 8, for a total of 67 events or 2211 pairs.

[3] It is to be noted that GWTC-3 includes the catalog GWTC-2.1 [17] that reanalyzed the events in the first two observing run with the newest waveforms used during the third observing run.

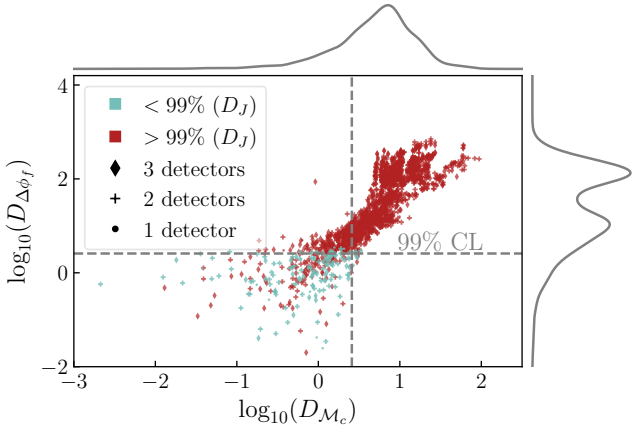


FIG. 6. Comparison of the one dimensional distance in the derived orbital phase evolution $\Delta\phi_f$ and the detector frame chirp mass $\mathcal{M}_c$ for real gravitational wave events. $\Delta\phi_f$ efficiently rejects 89% of the pairs as being not-lensed at 99% CL, while the same threshold on $\mathcal{M}_c$ more than doubles the number of the pairs. Colors denote the 99% CL cut on D_J alone and note that some pairs are ruled out by $D_{\Delta\phi_f}$ but not D_J (see text for discussion). Point shapes indicate pairs in which both events had *at least* 3 detectors (♦), 2 detectors (+) or 1 detector (•) online.

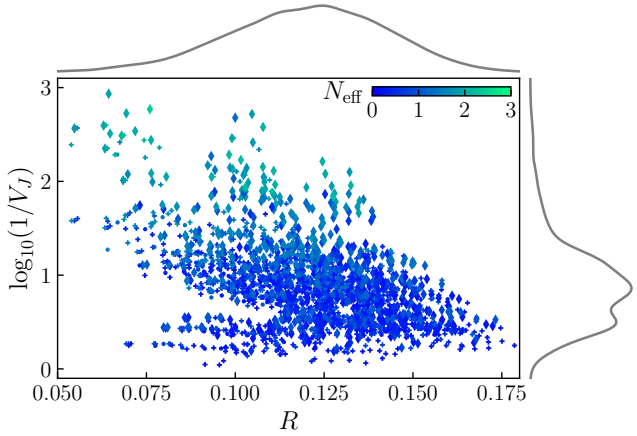


FIG. 7. Detector phase volume V_J as a function of R for real gravitational wave events. R weights the inverse network SNR of the pair, see Eq. (15). We color the pairs by their number of effective detector phases $N_{\rm eff}$. Smaller R tends to correspond to larger $1/V_J$ and a higher $N_{\rm eff}$ albeit with large scatter, mainly associated with the poorer phase measurements for high chirp mass events. Point shapes indicate the number of detectors as in Fig. 6.

With the current catalog, where the chirp masses are distributed over a wide range of values and the Virgo detections are low SNR at best, we expect the main component of the distance discriminator to be the phase difference between frequencies $\Delta\phi_f$. As previously discussed, $\Delta\phi_f$ is a better constrained parameter than the detector-frame chirp mass, leading to distances which are typically at least a factor of 2-3 larger as shown in Fig. 6. This leads to a significant advantage when testing the lensing hypothesis. If we compute the 1D distance consistency $D_{\Delta\phi_f}$ determined by $\Delta\phi_f$ alone, we find that only 11% of the pairs are consistent with the lensing hypothesis at 99% CL. This is to be compared with the $\mathcal{M}_c$ consistency alone where more than double that number 23% are consistent with lensing at 99% CL.

While the joint distance D_J automatically accounts for $D_{\Delta\phi_f}$, the detector phase differences are only significant in a fraction of the pairs in the current catalog. Quantitatively, only 31% of the pairs have more than one effectively constrained detector phase at 40Hz. In other words, for most of the pairs, the detector phases have errors larger than $\pi/2$, a precision that is necessary to distinguish between all the possible lensing phase shifts (and that is a factor of 4 smaller than the naive 2π prior range of the phases, requiring a factor of 4 higher SNR to be informative). The number of pairs with $N_{\rm eff} > 1$ increases to 45% when focusing on pairs with three detectors online. As shown in Fig. 7, pairs with a low number of effective phases are those with large volumes and R statistic (see Eq. 15). Note though that the scatter around this trend is large and that the pairs in Fig. 7 with small R but large volume are associated to events with high chirp mass.

The large scatter to higher V_J also implies that there can be cases where lensing is excluded by $D_{\Delta\phi_f}$ but marginally allowed by D_J. Within the current catalog and a 99% CL threshold, this occurs in 16 pairs. Recall that we compute the p-value for D_J using the $N_{\rm eff}$, see Eq. (11). In order to avoid cases where a strong rejection in the one dimensional space is diluted by poor and potentially non-Gaussian measurements in the higher dimensional phase space, we require both $D_{\Delta\phi_f}$ and D_J to pass the 99% CL consistency test in the current catalog. This brings our total to 131 out of 2211 pairs or $\sim 6\%$ that are consistent with lensing.

We display these cuts and the final lensing catalog in Fig. 8. As expected, 3 detector (HLV) events with higher SNR display smaller volumes and reject the lensing hypothesis more strongly. Notice also that some of the structure in the distribution of distances and volumes in Fig. 8 is inherited from the astrophysical population of chirp masses discussed in Fig. 7.

For each pair in the final lensing catalog, our method provides information about the image types through the phase shift which gives the minimum distance. The pair GW170104–GW170814 stands out as the most interesting candidate given its small phase volume, $\log_{10}(1/V_J) = 1.2$, and distance, $D_J = 1.3$, for a phase shift of $\pm\pi$. This would correspond to type I-type III image pairs. This pair was identified early on by [25, 26], and remains the pair with highest coherence ratio [9] (see also Sec. VII). When including astrophysical priors about the lens model and population of lenses, the lensing hypothesis is however disfavored [8].

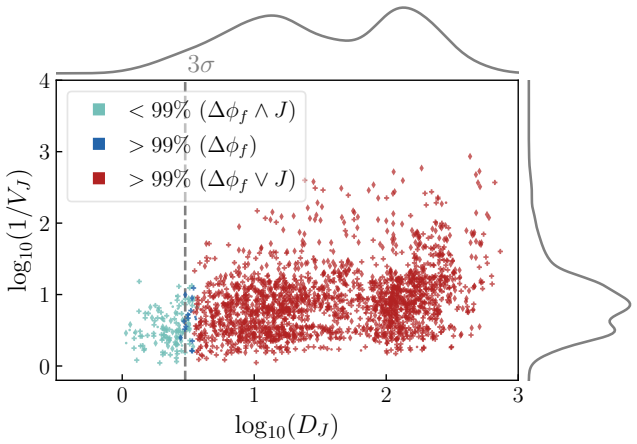


FIG. 8. Analysis of real gravitational wave events. We compare the distance (D_J) and detector phase volume (V_J) statistics for 67 binary black holes in GWTC3 with probability of having astrophysical origin larger than 0.8 and network SNR > 8. We use different symbols to indicate the number of detectors online and different colors to show the confidence level of agreement with the lensing hypothesis. 94% of pairs are rejected at 99% CL. Point shapes indicate the number of detectors as in Fig. 6.

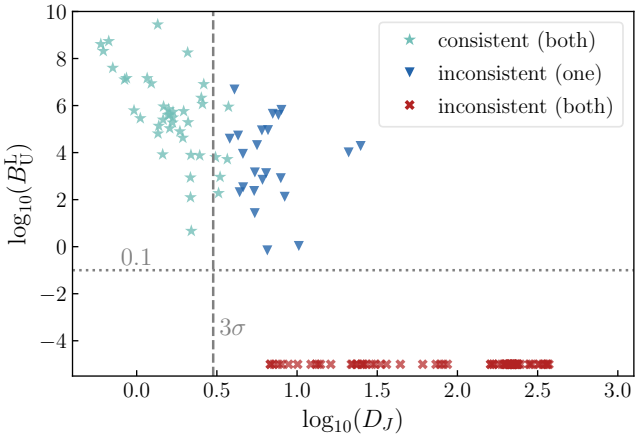


FIG. 9. Comparison between the phase consistency and the posterior overlap methods for simulated gravitational wave events. Phase consistency is determined by the distance D_J, while posterior overlap is quantified by the $B^{\rm L}_{\rm U}$ statistic. We follow the same plotting conventions of Fig. 5. The minimum $B^{\rm L}_{\rm U}$ is set to 10^{-5}.

VII. COMPARISON WITH OTHER METHODS

GW strongly lensed candidate pairs can be ranked by the amount of overlap of their parameters. This posterior overlap method [11] uses

$$B^{\rm L}_{\rm U} = \int {\rm d}\Theta \frac{p(\Theta|{\rm event}_1)p(\Theta|{\rm event}_2)}{p(\Theta)} \qquad (19)$$

as its basic statistic where Θ are the set of parameters over which overlaps are computed, $p(\Theta|{\rm event})$ are their posterior distributions and $p(\Theta)$ their prior distribution. Previous analyses [7–10] have computed overlaps over 8 parameters: detector-frame masses, dimensionless spin magnitudes, the cosine of spin tilt angles, the cosine of orbital inclination θ_{JN} and sky position. The $B^{\rm L}_{\rm U}$ statistic is therefore a ratio of marginalized posteriors that is agnostic to astrophysical assumptions. The $B^{\rm L}_{\rm U}$ is however sensitive to the prior range of its parameters, in particular the detector frame masses. Moreover, being a statistic that it is not normalized or calibrated intrinsically, the interpretation of a $B^{\rm L}_{\rm U}$ is subject to prior knowledge of the expected values for a population of lensed and not-lensed sources. Aided by background studies, one can then translate a $B^{\rm L}_{\rm U}$ into an odds ratio or p-value [8, 9]. Since we want to compare the posterior overlap method with our phase consistency approach that it is agnostic to astrophysical modeling or assumptions about the likelihood of the lensed and not-lensed hypotheses, we therefore restrict to $B^{\rm L}_{\rm U}$ and interpreting it as a Bayes factor. Correspondingly, for illustration purposes, we assume that only pairs with $B^{\rm L}_{\rm U} < 0.1$ reject the lensing hypothesis.

In the past $B^{\rm L}_{\rm U}$ has been used to select interesting candidates to follow with joint parameter estimation. In the first half of the third observing run 19 pairs of super-threshold events with $B^{\rm L}_{\rm U} > 50$ were followed up [8], although 4 of them involved a GW event (GW190424_180648) that was later downgraded to sub-threshold [17]. In [8], the O2 pair GW170104-GW170814 was also analyzed in joint PE. In the second half of O3 pairs with a false positive probability (FPP) of $< 1\%$ were further analyzed [9], with the FPP computed after a large injection campaign. This accounted for another 14 joint parameter estimation pairs. In total there are 30 pairs of super-threshold events that we can compare with. For all these pairs, the joint parameter estimation provides the coherence ratio $C^{\rm L}_{\rm U}$, defined as the ratio of the lensed and not-lensed evidences, and it is (mostly) agnostic to astrophysical priors on the population of sources and lenses [8]. LVK analyses [8, 9] also computed the Bayes factor that accounts for the final likelihood of lensing taking into account reasonable astrophysical expectations.

Following [7–9], we compute the posterior overlap of a given pair approximating the posterior distributions with a kernel density estimator (KDE) for the first six parameters and computing the sky overlap independently. Importantly, the sky overlaps in this approach are not normalized. Alternative sky map overlap statistics have been studied in [27]. We note that the numerical calculation of the six-dimensional KDE is significantly more demanding than the distance calculation, which is typically 1000 times faster.

We begin by analyzing our injection set. Fig. 9 presents the distance statistic D_J versus the posterior overlap $B^{\rm L}_{\rm U}$ for all possible pairs. Noticeably most lensed pairs display a large $B^{\rm L}_{\rm U}$. However, there are also many not-lensed pairs with similar overlap that are rejected by their phase

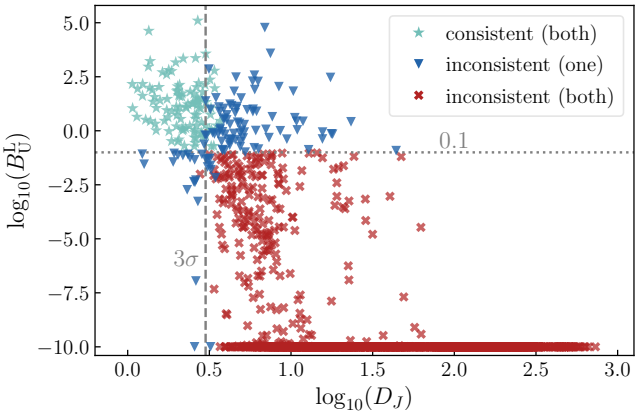


FIG. 10. Comparison between the phase consistency and the posterior overlap methods for real gravitational wave events. We follow the same plotting conventions of Fig. 8. The minimum $B^{\rm L}_{\rm U}$ is set to 10^{-10}.

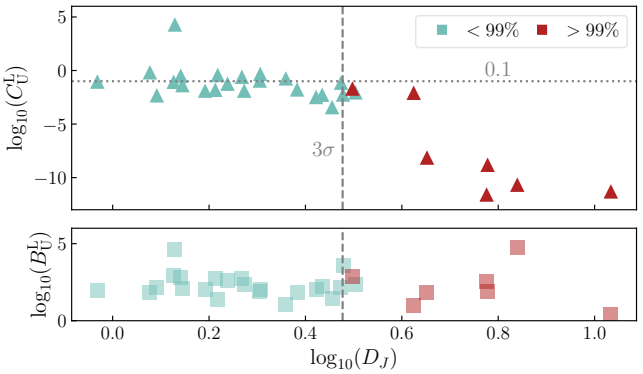


FIG. 11. Comparison between the phase consistency distance (D_J), the joint parameter estimation coherence ratio ($C^{\rm L}_{\rm U}$) and the posterior overlap statistic ($B^{\rm L}_{\rm U}$) for real gravitational wave events. The joint parameter estimation results are taken from [8, 9].

consistency. Moreover there are no pairs with $B^{\rm L}_{\rm U} < 0.1$ and $D_J > 3$ which would represent a more efficient rejection via overlap. This demonstrates the potential of our method to better reduce the number of false alarms. Because of not being normalized, the overlap statistics show a tendency to predict HLV-pairs with higher $B^{\rm L}_{\rm U}$, regardless of lensing.

We then analyze the catalog of real binary black holes in Fig. 10. We find that with the posterior overlap statistic 9% of pairs are consistent with the lensing hypothesis as determined by $\log_{10} B^{\rm L}_{\rm U} > -1$. This is to be compared to the 6% found with the phase consistency at 99% CL. The rejection of the lensing hypothesis in the posterior overlap analysis is mainly triggered by non-overlapping sky maps, with 30% of pairs having no sky-overlap. This is due mostly to the majority of events in the catalog being relatively heavy (due to selection effects) leading to larger errors in their masses, and agrees with previous expectations [6]. When excluding the sky localization information, 18% of pairs are consistent with lensing.

In terms of the purity of the $B^{\rm L}_{\rm U} > 0.1$ pair catalog, the phase consistency method disfavors (at 99% CL) 44% of candidate pairs, for a total of 82. In Fig. 10, these are the upper right blue points. From them, some exhibit strong consistency with the lensing hypothesis through their overlaps, $B^{\rm L}_{\rm U} > 100$, but are rejected by the phase consistency at high confidence. This is because in the posterior overlap method a single parameter can outweigh the rest, while the phase consistency accounts more equally from the consistency in all the phases. This demonstrates the complementarity and advantages of our new method compared to existing analyses.

Conversely, in terms of the purity of the phase consistency catalog, 19% of pairs consistent with lensing by their phases have $B^{\rm L}_{\rm U} < 0.1$. In Fig. 10, these are the 25 lower left blue points, i.e. less than a third of the false alarms of $B^{\rm L}_{\rm U}$. These are cases in which the Gaussian distance estimator provides an overly conservative agreement for an intrinsically non-Gaussian posterior, especially in the time delay phases of two detector events where the degeneracy forms rings (see App. C). In this case, spot checking our phase-based pair catalog for sky overlap consistency can be helpful. More generally, one can automate tests for significantly non-Gaussian distributions such as a difference between the mean and mode. For such cases, a non-Gaussian tension estimation for phase consistency can be efficiently achieved using machine learning methods [14]. We leave the implementation of such tools for future work.

Finally, we compare our phase consistency distance statistic D_J with the joint parameter estimation coherence ratio $C^{\rm L}_{\rm U}$ and the posterior overlap $B^{\rm L}_{\rm U}$ for the 30 pairs in which this information is available. As shown in Fig. 11 all pairs display a large $B^{\rm L}_{\rm U}$ since this was the criteria for joint parameter estimation follow up. We can also see that only GW170104-GW170814 favors lensing with $\log_{10} C^{\rm L}_{\rm U} > 0$. In fact, many of them are rejecting the lensing hypothesis $\log_{10} C^{\rm L}_{\rm U} < -1$. Noticeably, our distance statistic is able to reject 6 pairs at 99% CL with large $B^{\rm L}_{\rm U}$ but small $C^{\rm L}_{\rm U}$. The anticorrelation between $D_J > 3$ and $C^{\rm L}_{\rm U}$ demonstrates the benefit of incorporating the phase consistency in future strong lensing searches. Our method also correctly identified the pair GW170104-GW170814 as the most significant given the consistent distance and small volume, as discussed in the previous section.

VIII. CURRENT AND FUTURE STRONG LENSING SEARCHES

Our analysis of real GW events in the first three observing runs together with the simulations of lensed and not-lensed events with the sensitivity of the fourth observing run allow us to draw some expectations for the upcoming and future multiple image searches. As

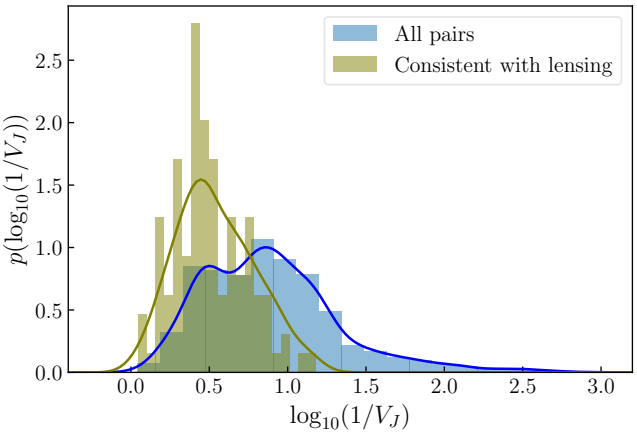


FIG. 12. Distribution of detector phase volume for all the pairs analyzed in GWTC3, and those consistent with the lensing hypothesis. Consistency with lensing is defined at 99% CL as in the main text. This represents 6% of the pairs.

we have seen in the previous sections, our ability to test the lensing hypothesis depends on how well we can reconstruct the different phases. Past detector sensitivities were such that the distance measuring the (in)consistency of a given pair with the lensing hypothesis was dominated by the orbital phase evolution information, as the detector phases were typically not well enough constrained to distinguish between the possible lensing phase shifts (see Fig. 19 in App. F for the actual posterior distributions of ϕ_d). In fact, when looking at the distribution of detector phase volumes, see Fig. 12, one can see that pairs consistent with the lensing hypothesis are skewed towards larger volumes, making ranking by V_J even more important. In other words, most of the pair catalog is composed of events with poor parameter constraints and/or low SNR.

The situation with our simulated events shows more promise for the future. During the fourth observing run there should be more events where the phases are better constrained and provide a better discriminant for the lensing hypothesis. Still, we observed large difference in the constraining power of our optimistic, high SNR simulated events to our more common, low SNR injections and in any given observing run low SNR events will always be more numerous than high SNR events.

In general, for the detector phases to be informative we need to distinguish between the possible $\pi/2$ phase shifts within the pair. Thus, the effective σ from the lensing prior is $\pi/2\sqrt{12}$. From data SNR per detector is $3\sqrt{2}/\rho_d$, so nominally for the detector phases to be informative it is necessary that $\rho_d > 12\sqrt{6}/\pi \sim 10$ or $R < \pi/12\sqrt{6} \sim 0.1$ for a two detector network. Within the real events that we analyzed in GWTC3, that is true for only about 20% of pairs, see Fig. 7. The distribution of SNRs for a given detector network is (quasi-)universal [28].[4] It follows $p(\rho_{\rm ntw}) = 3\rho_{\rm th}^3/\rho_{\rm ntw}^4$, where $\rho_{\rm th}$ is the threshold for detection. Therefore, it is to be anticipated that for a fixed network, the majority of events will be of low significance close to the threshold. Still, future observing runs will improve their sensitivity, making distant events with higher chances of experiencing lensing better measured and easier to identify. Within our current data, if we increase the network SNR cut to 12, the catalog size reduces to 36 events (630 pairs) and the fraction of pairs consistent with lensing also reduces to 3%, or 18 pairs. Given the current computational cost of joint parameter estimation analyses, reducing the number of pairs to follow up in a "golden catalog" of lensed candidates would be advantageous and with our technique can be achieved through cuts in V_J. As demonstrated in Fig. 11, our distance measure correlates nicely with the joint parameter estimation coherence ratio and can be used to reduce the candidate list compared to other methods.

Besides the GW information, electromagnetic follow ups of strongly lensed candidates will be essential to conclude the lensed origin of a set of GW events. This can be achieved in different fashions. Targeted follow ups and archival searches can look for lensing signatures of the host galaxy or other galaxies in the line of sight of the joint sky localization of the pair. Similarly, one may search for other lensed transients that might be associated with the GW signals. In any case, information about the lens model is key to achieve an efficient cross-correlation. In that respect, our method is also advantageous as it provides a fast lensing consistency check with additional information about the image types through the preferred lensing phase shift. Image type information together with the time delay and magnification ratios can be used to model the lens.

IX. CONCLUSIONS

The detection of lensed gravitational waves remains elusive in current datasets. The key signature for GWs traveling through a massive lens would be repeated copies of the emitted signal, arriving at different times, sharing the same frequency evolution and polarization state, but with different amplitudes and specific differences in absolute phases. The quest of discovering strongly lensed GWs thus entails identifying these repeated chirps within a large catalog of not-lensed events. The crux is that many not-lensed pairs of events may look alike, simply because their parameters are not well enough constrained to discern the difference.

We have developed a new method to efficiently reject event pairs that are inconsistent with the lensing hypothesis and construct a catalog of candidates for fur-

[4] The strict universality holds in an Euclidean universe where sources are uniformly distributed in distance. This is a good approximation for low-redshift observations.

ther study. Our method's main ingredients are: 1) reconstructing the best measured GW parameters and 2) determining the consistency of the events rather than account for the degree of their parameter overlap. This design allow us to efficiently produce a highly complete catalog that is also more pure than previous methods. Efficiency comes from the fact that we post-process the full parameter estimation posterior samples to obtain the phases that would have been measured by each detector. We demonstrate that at 40Hz such phases are constrained within a fraction of a radian and that its change through the frequency spectrum precisely quantifies the orbital phase evolution with better relative errors than the chirp mass. Moreover, we show that the phases associated to the arrival time differences of the signal at the detectors are also well constrained and reduce the degeneracies of typical sky localizations.

The degree of completeness vs. purity is controlled by computing the distance between two events in the multi-dimensional space of their phases. Because the phases are better constrained, their distributions are more Gaussianly distributed than the original waveform parameters and so we apply a Gaussian approximation to the distance whose evaluation is computationally trivial. From the pairs consistent with lensing, we rank them by their phase volume so as to de-emphasize the remaining candidate pairs with low SNR and poor parameter measurements. The code behind this pipeline, `phazap`, is publicly available.

Because we are working directly with the GW phases, we can determine for a given candidate pair the most probable lensing phase shift and hence their image types. This information together with the ability of our technique to rapidly consider all possible pairs makes our method a valuable tool for electromagnetic follow ups, as it could then inform the potential lens model. Moreover, due to the good correlation of our distance statistic with the full joint parameter estimation inference (as measured by the coherence ratio, cf. Fig. 11), our method is well positioned to handle the ever-growing number of possible pairs in future catalogs.

In its application for searching for lensed GWs, our method could be extended in several ways. Consistency between a pair could not only be evaluated in terms of the detector phases, but it could also include the polarization state as a function of frequency: this would effectively determine spin consistency. In addition, it could be used to forecast the prospects of detecting lensing with future ground- and space-based GW facilities. In this respect, our method is expected to improve rapidly with the number of detectors in the GW network, and aspect that we will explore in future work.

We have also found that the orbital phase evolution is more constraining than the chirp mass, typically by a factor of 2-3 in terms of parameter distance. For any given threshold in detection or distance this substantially enhances the purity of the catalog and may even be useful for subthreshold searches. Similarly, the time delay phases offer an interesting opportunity to improve over the largely unlocalized subthreshold triggers. In terms of its current implementation, `phazap` could be improved with machine learning algorithms to compute the distance in low significance, non-Gaussian posteriors.

Although we focused on its use to discover lensed GWs, the code developed to post-process GW phases has far reaching applications. For example, tracking the phase evolution at different frequencies one could test the theory of gravity, and by looking at the phase difference of the polarization states one could probe birefringence. Finally, while we have focused on the dominant quadrupole mode, comparing the phases of different modes can potentially address waveform systematics.

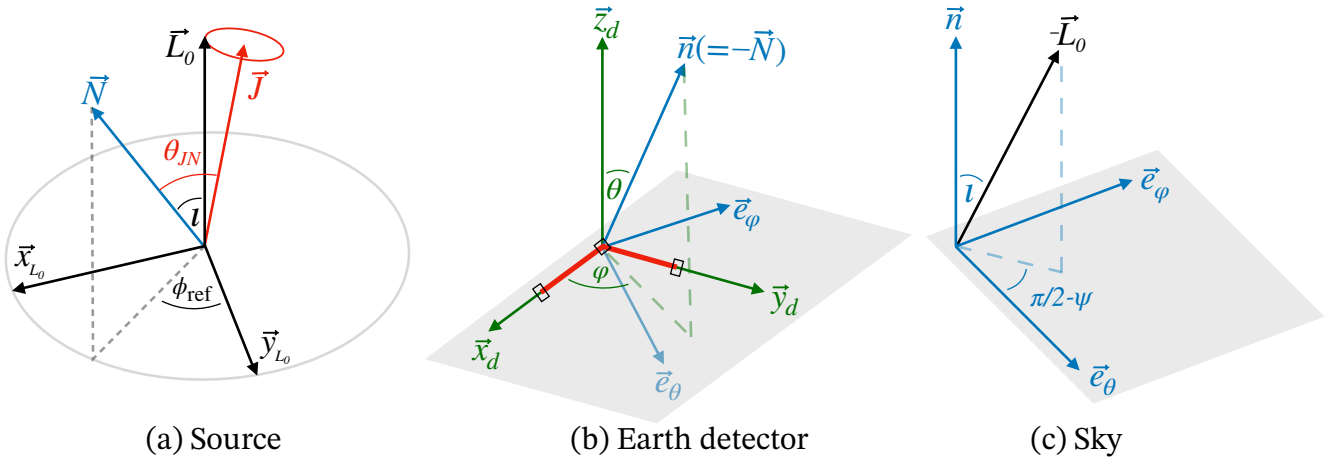
ACKNOWLEDGMENTS

We are grateful to Srashti Goyal for comments on the paper and Geraint Pattern for correspondence about the IMRPhenomX waveform family. JME is supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 847523 INTERACTIONS, and by VILLUM FONDEN (grant no. 53101 and 37766). WH is supported by U.S. Dept. of Energy contract DE-FG02-13ER41958 and the Simons Foundation. RKLL is supported by the National Science Foundation through award number PHY-1912594 and PHY-2207758. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459.

Appendix A: Reference frames and polarization basis conventions

The gravitational wave signal emitted during a compact binary coalescence can be defined in different reference frames. To simplify comparisons with existing codes, we will follow the LALSuite conventions for gravitational wave data analysis [29] (used by LVK, waveform modelers, e.g [30], and numerical relativists [31]).

The *source frame* defines the reference frame at a given epoch where the intrinsic parameters of the binary are inputted. We fix this initial frame to a reference frequency of the GW $f_{\rm ref}$ and anchor it to the orbital angular momentum of the binary at the time the reference frequency is emitted in the 22-mode, $\vec{L}_0 = \vec{L}(f_{\rm ref})$, which defines the z-axis. Therefore, we call the axes of this reference frame $\{\hat{x}_{L_0}, \hat{y}_{L_0}, \hat{z}_{L_0}\}$. The $\hat{x}_{L_0}$-axis connects the compact objects, and $\hat{y}_{L_0} = \hat{z}_{L_0} \times \hat{x}_{L_0}$. The propagation direction $\vec{N}(= -\vec{n})$ pointing from the source to the observer has the spherical-polar angles $(\iota(f_{\rm ref}), \pi/2 - \varphi_{\rm ref})$,



(a) Source (b) Earth detector (c) Sky

FIG. 13. Summary figure of our frame conventions. The source frame (a) defines the coordinate system in which the intrinsic parameters of the binary are defined: masses, spins and phase. It is anchored to the orbital angular momentum at the reference frequency $\vec{L}_0$. The Earth detector frame (b) serves to define the time of arrival and the position of the sky of the binary event for a fiducial detector at the center of the Earth, in order to compute the antenna response function of each detector. The sky frame (c) defines the remaining extrinsic parameters, the inclination ι and the polarization angle ψ. See text for further details.

related to the inclination and phase at the reference frequency. The *radiation frame* (also known as wave frame) is defined by $\vec{N}$ with the x-axis, which defines the + polarization, along the transverse projection of $\vec{L}_0$. Note that for a precessing binary the total angular momentum $\vec{J}$ defines an alternate inertial frame.

In terms of the observer, the Earth *detector frame* is defined by the coordinate system $\{\hat{x}_d, \hat{y}_d, \hat{z}_d\}$ at a fixed time t_d with the origin at the center of the Earth.[5] In the detector frame the sky position of the source is $\{\theta, \varphi\}$. Conversely, the *sky frame* is defined by $\{\hat{e}_\theta, \hat{e}_\varphi, -\vec{N}\}$ and $\pi/2 - \psi$ is the angle of the transverse projection of $\vec{L}_0$ which describes the orientation of the polarization.

Following these LVK conventions [32, 33], the antenna response functions for a fiducial detector at the center of the Earth with arms in the $\vec{x}_d$ and $\vec{y}_d$ directions are given by

$$F_+ = -\frac{1+\cos^2\theta}{2}\cos(2\varphi)\cos(2\psi) - \cos\theta\sin(2\varphi)\sin(2\psi),$$
$$F_\times = \frac{1+\cos^2\theta}{2}\cos(2\varphi)\sin(2\psi) - \cos\theta\sin(2\varphi)\cos(2\psi)\,. \tag{A1}$$

It will be convenient for later to introduce the responses to the left and right circular polarization states

$$F_L = \frac{F_+ + iF_\times}{\sqrt{2}} \equiv |F_L| e^{i\alpha},$$
$$F_R = F_L^* = |F_L| e^{-i\alpha}, \tag{A2}$$

where $\alpha = \arg(F_L) = \text{atan}[F_+, F_\times]$. Notice that under a change in the polarization angle ψ at fixed (θ, φ), F_L and F_R pick up a pure phase of equal and opposite sign.

Since for a precessing binary, spins and inclination are in general frequency dependent, we define them at the source frame fixed by $f_{\rm ref}$ and $\vec{L}_0$. Related to the three vectors $\vec{J}$, $\vec{N}$ and $\vec{L}_0$, there are three relevant angles: the angle between $\vec{J}$ and $\vec{N}$, θ_{JN}, defined by $\cos\theta_{JN} = \vec{J}\cdot N/|\vec{J}||\vec{N}| = N_z$, and the two spherical polar angles of $\vec{J}$ in the $\vec{L}_0$ frame $\{\theta_{JL_0}, \phi_{JL_0}\}$, that is $\cos\theta_{JL_0} = \vec{J}\cdot\vec{L}_0/|\vec{J}||\vec{L}_0|$ and $\cos\phi_{JL_0} = J_{x_{L_0}}/(J^2_{x_{L_0}} + J^2_{y_{L_0}})$. [6] The inclination of the binary ι is defined by the angle between $\vec{L}_0$ and $\vec{N}$. The component spins $\vec{S}_1$ and $\vec{S}_2$ are defined by their 6 Cartesian coordinates. It is common however in parameter estimation to refer instead to their dimensionless spin magnitudes χ_1 and χ_2, their tilt with respect to the Newtonian orbital angular momentum $\vec{L}_N$ (which is always perpendicular to the binary's orbital plane) denoted as ϕ_1 and ϕ_2, and ϕ_{12} the difference in azimuthal angles of $\vec{S}_1$ and $\vec{S}_2$.

To obtain the actual response of each individual detector we rotate from Earth detector frame to the individual detector frames given their orientations and the source

[5] Note that for current detected sources and detector sensitivity one can assume that the detector frame is fixed throughout the duration of the signal. This will change with next-generation ground-based detectors where the Earth rotation should be taken into account, and certainly for LISA where signals could last years in band.

[6] Note that in the GW literature it is common to simply denote the angles θ_{JL_0} and ϕ_{JL_0} as θ_{JL} and ϕ_{JL}, specifying that they are defined at the reference frequency.

position in Earth detector coordinates (θ, φ) to (ra, dec) using the Greenwich Mean Sidereal Time of the observation (ra $= \varphi +$ gmst and dec $= \pi/2 - \theta$). Conveniently, LALSuite has integrated routines to easily change from the Earth detector (also known as Earth fixed) frame to each of the actual detector frames, for our analysis Hanford, Livingston and Virgo.

Appendix B: Measured phase at a new frequency

The GW signal is constructed following a given waveform model with intrinsic and extrinsic parameters defined at a reference frequency $f_{\rm ref}$. Specifying $f_{\rm ref}$ is necessary as for precessing binaries the spin components vary over time. LVK analyses typically fix $f_{\rm ref} = 20$Hz, which for typical signals does not correspond to the frequency of the best measured phase. Our goal here is to demonstrate how to consistently reconstruct the detected phase of a GW at a new frequency. This will also serve to illustrate the working procedure of the `phazap` package.

1. Polarization states

The frequency domain waveform of an emitted GW is described by the real and imaginary parts of the two tensor polarizations.[7] These polarizations can be described in different bases, see [34] for a review, and are fully characterized by four real numbers, two amplitudes and two phases. For example, for the linear polarizations of a monochromatic signal this corresponds to $\tilde{h}_+ = A_+ e^{-i\phi_+}$ and $\tilde{h}_\times = A_\times e^{-i\phi_\times}$. We follow App. C of [35] and perform a Stokes decomposition to construct a more physically intuitive set of parameters: $\{A, \phi, \beta, \zeta\}$, where A and ϕ describe the global amplitude and phase of the GW signal, whereas the angles β and ζ characterize the relation between the two polarization components. In order to connect the new set of parameters to the ones of the polarizations, it is convenient to move to the circular basis

$$\tilde{h}_L = (\tilde{h}_+ - i\tilde{h}_\times)/\sqrt{2}\,, \tag{B1}$$

$$\tilde{h}_R = (\tilde{h}_+ + i\tilde{h}_\times)/\sqrt{2}\,, \tag{B2}$$

which can also be decomposed in an amplitude and a phase $\tilde{h}_L = A_L e^{i\phi_L}$ and $\tilde{h}_R = A_R e^{i\phi_R}$. The global amplitude and phase are defined by

$$\phi = (\phi_L + \phi_R)/2\,, \quad A = \sqrt{A_L^2 + A_R^2}\,. \tag{B3}$$

[7] As in the main text, we implicitly write all our expressions for positive frequency and left the negative frequencies to be defined by the reality condition of the time domain signal.

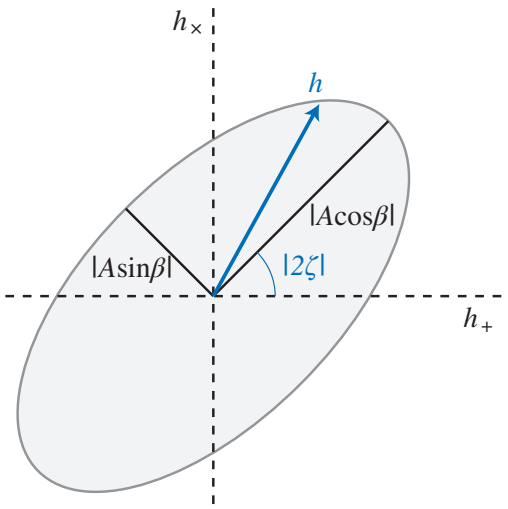


FIG. 14. A general gravitational wave h is elliptically polarized, where the polarization state rotates between the linear polarization states h_+ and $h_\times$ with ζ defining the orientation of the principal axes and β defining their relative amplitude.

The angle β is related to the ratio of the semi-major and minor axes

$$\tan\beta = \frac{1-r}{1+r} \tag{B4}$$

where r is the amplitude ratio $r = A_L/A_R$. The difference in the phases describes the orientation of the semi-major axis or equivalently a rotation of the $+, \times$ basis

$$\zeta = (\phi_L - \phi_R)/4\,. \tag{B5}$$

The circularly-polarized modes can then be written as

$$\tilde{h}_L = h_R r e^{-4i\zeta} = \frac{Ar}{\sqrt{1+r^2}} e^{-i(\phi+2\zeta)}\,, \tag{B6}$$

$$\tilde{h}_R = A_R e^{-i\phi_R} = \frac{A}{\sqrt{1+r^2}} e^{-i(\phi-2\zeta)}\,. \tag{B7}$$

Purely linear polarization corresponds to $\beta = 0$ (or $r = 1$), with h_+ if $\zeta = 0$ and $h_\times$ if $\zeta = \pm\pi/4$. Purely circular polarization corresponds to $\beta = \pm\pi/4$. A graphical representation of the angles defining the polarization state is shown in Fig. 14.

The frequency domain waveform of an emitted GW defined by a given waveform model is outputted in the radiation frame. For generic precessing binaries, the orbital angular momentum $\vec{L}$, polarization angle ψ and inclination ι all change with time and, therefore, frequency. Similarly, all $\{A, \phi, \beta, \zeta\}$ will be frequency dependent. In our default analysis we will use the IMRPhenomXPHM waveform model [30] that was used to analyze the latest GW catalog [4].

2. Detected phase at a new frequency

Our main goal is to compute the detected phase of the GW at any frequency, i.e. Eq. (2). For this we first need the GW global phase $\phi(f)$ in the radiation frame which can be computed from the left and right phase following Eq. (B3). This phase can be constructed for any multipole mode, provided we decompose the signal in its multiple moments $\tilde{h} = \sum_{lm} A_{lm} e^{-i\Phi_{lm}}$ to get $\{A_{lm}, \phi_{lm}, \beta_{lm}, \zeta_{lm}\}$. In our analysis we reconstruct the phase of the 22 mode at different frequencies $\phi_{22}(f)$. As noted in the main text, we define the 22-mode in the coprecessing frame from the +-polarization. The equatorial symmetry of this frame together with the reality condition of the time domain signal implies $\tilde{h}_{lm} = (-1)^l \tilde{h}^*_{l-m}(-f)$ [36]. Therefore only one m-mode defines the phase at positive frequencies. In the conventions of the IMRPhenomX waveform family [30, 36] that we follow, this corresponds to $l = 2$, $m = -2$.

The 22-waveform phase at the reference frequency is then read off directly from the waveform approximant for a given set of binary parameters using Eq. (B3). We remind the reader that in the stationary phase approximation (SPA) it is related to the orbital phase $\phi_{\rm ref}$ as

$$\Phi_{22}(f_{\rm ref})|_{\rm SPA} = 2\phi_{\rm ref} - 2\pi f_{\rm ref}(t_{\rm ref} - t_d) + \text{const} \quad \text{(B8)}$$

where $t_{\rm ref}$ is the arrival time of the frequency $f_{\rm ref}$. We take the global waveform arrival time t_d in Eq. (1) to coincide with this time $t_d = t_{\rm ref}$. Therefore, our "detector phase" ϕ_d is then the phase of frequency f at the time when $f_{\rm ref}$ hits the given detector.

Given $\phi(f) \equiv \Phi_{22}(f)$, in order to obtain the detector phase $\phi_d(f)$ we need to project from the radiation to the detector frame. At a given detector, the frequency domain signal of a given lm-mode (although we do not include the lm label here for notational simplicity) is

$$\begin{aligned}\tilde{h}_d &= F_L \tilde{h}_L + F_R \tilde{h}_R \\ &= \frac{A|F_L|}{\sqrt{1+r^2}} e^{-i\phi} \left(r e^{+i(\alpha-2\zeta)} + e^{-i(\alpha-2\zeta)} \right) \\ &\equiv A_d(f) e^{-i(\phi+\chi)} \,, \end{aligned} \quad \text{(B9)}$$

where we have used Eq. (A2), and

$$\tan\chi = \frac{(1-r)}{(1+r)} \tan(\alpha - 2\zeta) \quad \text{(B10)}$$

Finally, noting that the frequency dependent polarization angle ζ can be reabsorbed into a redefinition of $\psi' = \psi + \zeta$, we arrive at the detector phase

$$\begin{aligned}\phi_d(f) &= \phi(f) + \chi(f) \quad \text{(B11)} \\ &= \phi(f) + \text{atan}\left[F_+(\psi'), \frac{(1-r)}{(1+r)} F_\times(\psi') \right]\end{aligned}$$

and amplitude

$$A_d = \frac{A|F_L|}{\sqrt{1+r^2}} \sqrt{1 + r^2 + 2r\cos(2\alpha - 4\zeta)} \,. \quad \text{(B12)}$$

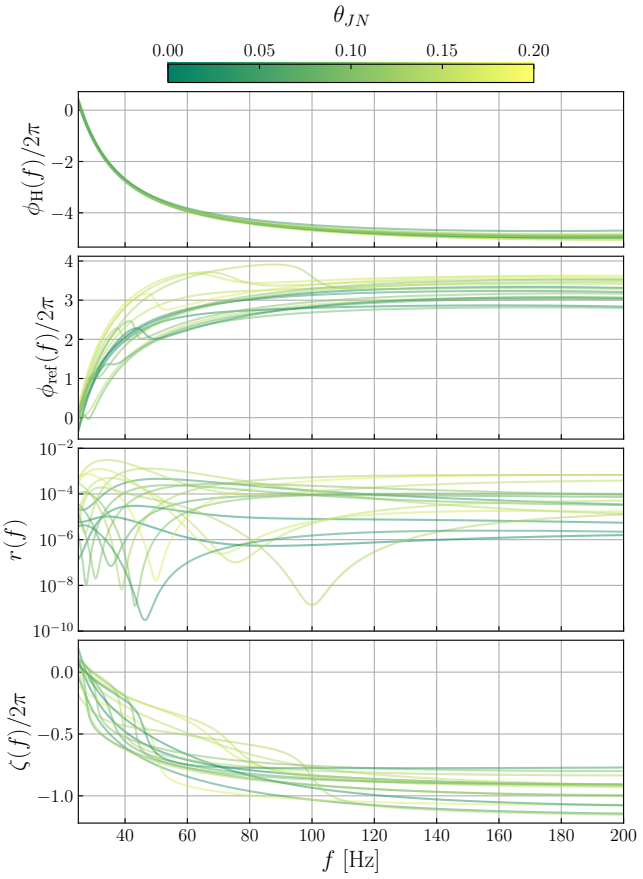


FIG. 15. Frequency evolution for the detector phase ($\phi_{\rm H}$) in comparison to the 22 reference phase ($\phi_{\rm ref}$), the polarization amplitude ratio (r) and the polarization phase difference ζ for random samples close to face-on. The colors indicate the value of the angle between the total angular momentum $\vec{J}$ and the line of sight $\vec{N}$. The samples correspond to the simulated Type I, low SNR, HLV event. The reference frequency is at 20Hz.

With this derivation, the role of the circular polarization ratio r and the linear polarization angle ζ become more transparent. After the multipolar decomposition, the ratio of the polarization amplitudes, r_{lm}, provides a direct way of computing the inclination at a new frequency

$$r_{lm} = \frac{|h^{lm}_L|}{|h^{lm}_R|} = \frac{1 - a_{lm}(\iota)}{1 + a_{lm}(\iota)} \,, \quad \text{(B13)}$$

where a_{lm} gives the relative strength of the linear polarizations in the emission direction. For $l = |m|$ modes one gets

$$a_{l=|m|}(\iota(f)) = \frac{2\cos\iota}{1 + \cos^2\iota} = \frac{1 - r_{lm}}{1 + r_{lm}} \,. \quad \text{(B14)}$$

This shows that a change in the polarization ratio can be mapped to a change in the inclination, and that both quantities are in general frequency dependent.

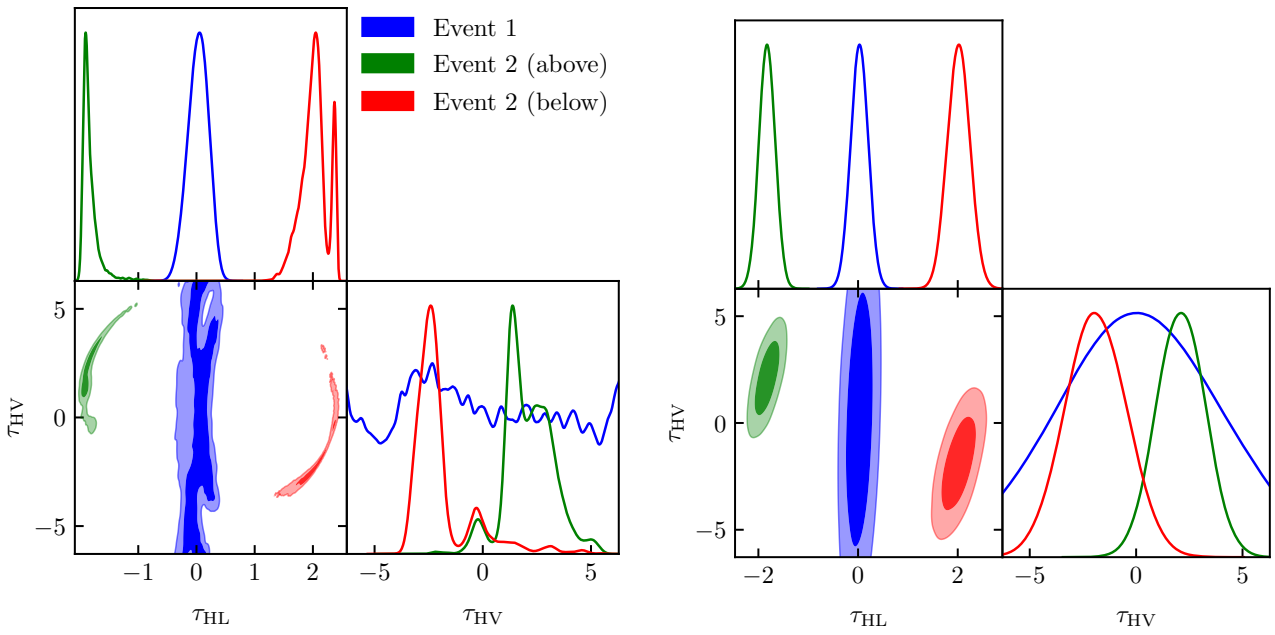


FIG. 16. Time delay phases for a pair of events analyzed under the lensing hypothesis. Event 1 corresponds to GW190706_222641 and Event 2 to GW190519_153544. Event 2 phases are reconstructed at the reference frame determined by Event 1. The above and below samples of Event 2 are defined with respect to its detector plane, i.e. $\vec{n}\cdot(\vec{r}_{\rm HL}\times\vec{r}_{\rm HV})$ negative and positive respectively. The right hand plot corresponds to the Gaussianization of the posteriors.

On the other hand, ζ represents a frequency dependent rotation of the polarization basis as it is clear in Eq. (B9). Note that by defining the radiation frame + polarization through the projection of $\vec{L}_0$ at the reference frequency $\zeta(f_{\rm ref})=0$.

Finally, let us note that the reconstructed detector phase is typically a well-constrained, monotonic function in frequency even when other quantities such as the inferred reference phase as a function of frequency is not. This illustrates how our choice of parameters removes degeneracies between binary parameters, in this case the reference phase and inclination. With precessing binaries, the inclination of the orbital plane can precess from a finite value at the reference frequency to nearly face/edge on at a different frequency and encounter large degeneracies. To exemplify this limiting case we draw the frequency evolution for $\phi_{\rm ref}$, r and ζ in Fig. 15 for random samples with small $\theta_{JN}\ll 1$. It is to be noted that the jumps in $\phi_{\rm ref}$ and ζ compensate with r to give a smooth detector phase in the top panel. Here we can also see how the spread in the reference phase is much larger than in the detector phase.

Appendix C: Breaking the bimodality of the time delay phases

As discussed in the main text, constant time delay phases track rings on the sky which intersect in points for multiple detector pairs. When the time delays themselves are well measured, transforming localization to time delays is advantageous as it collapses these degeneracies. On the other hand when time delays are derived from other localization information, as is the case for analyzing event 2 in the reference frame of event 1, this collapse does not fully occur and constraints on the derived delays inherit the ring like or multiple point intersections of the localization. For three detectors this often leads to a bimodal distribution of time delays. While the information from localization is still retained in the joint time-delays and can be used in parameter distances, bimodality implies non-Gaussianity which degrades our simple Gaussian distance approach.

This bimodality can however be broken when identifying the parameter samples that come from above or below the plane defined by the three detectors since time delays are symmetric under reflection across this plane. Operationally this can be achieved looking at the samples with positive or negative product:

$$\vec{n}\cdot(\vec{r}_{d_1d_2}\times\vec{r}_{d_1d_3})\ , \tag{C1}$$

where $\vec{r}_{d_id_j}\equiv\vec{r}_{d_i}-\vec{r}_{d_j}$ with $\vec{r}_{d_i}$ as the position of detector d_i.

In Fig. 16 we exemplify this construction. We choose GW190706_222641 as event 1 and GW190519_153544 as event 2. Event 1 is an HL-detection event, while event 2 is HLV. Since three detector events are better localized, it is advantageous to set them as event 2 where the time delay phases encode that information. By decomposing the samples of event 2 into those above or below

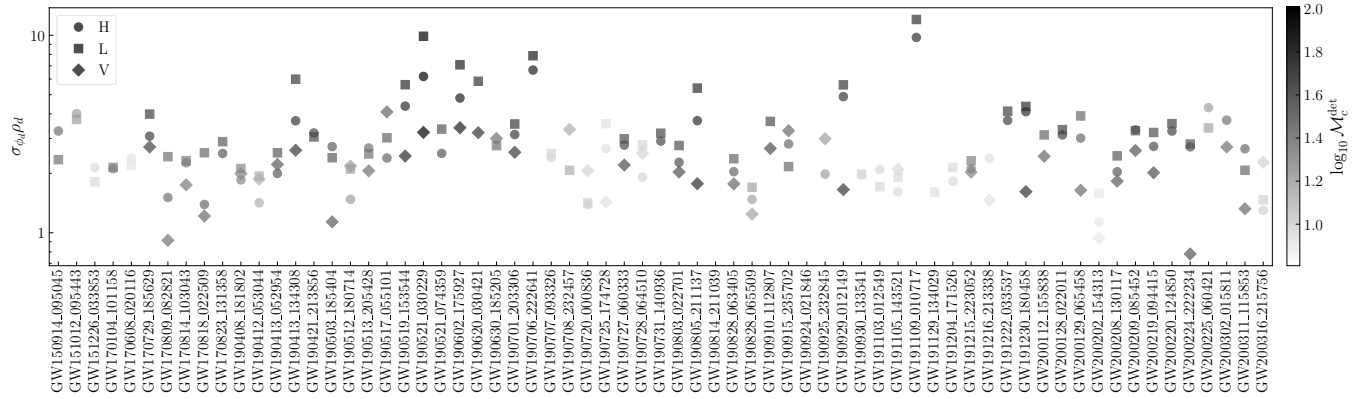


FIG. 17. Detector phase standard deviation (σ_{ρ_d}) times median "optimal" detector SNR (ρ_d) for the binary black holes analyzed in this work (GWTC3 with $p_{\rm astro} > 0.8$ and $\rho_{\rm ntw} > 8$). The color bar indicates the median detector-frame chirp mass. Detector phases are at 40Hz.

its detector plane we can identify the two independent modes. This allow us to asses the (in)consistency of each mode with respect to event 1 by finding the minimum distance between the two. For each of them we can compute the distance using the Gaussian approximation in Eq. (10). If we were not to break the reflection symmetry, the Gaussian approximation would determine that these two events are consistent since its covariance would be forced to cover both modes.

For reference, in the right hand side of Fig. 16 we present the Gaussianization of the same time delay posteriors that is used to compute the distance D_J. This also serve to demonstrate how the Gaussianization procedure tends to be conservative with respect to rejecting not-lensed pairs. Here the thin arcs in the joint time delay phases are converted to full ellipses which then are consistent with a wider range of joint time delays.

Since event 1 and event 2 both have this hemisphere/planar degeneracy, when in our method described in §IV we consider both orderings of the events, we are then taking into account the four sections of the sky that the two orientations of the detector planes define. If in either ordering both modes of the pair is inconsistent with lensing then the lensing hypothesis is rejected.

Appendix D: SNR scaling of phase errors

In order to estimate the errors in the detected phase and its correlation with other parameters one can use the Fisher information matrix, which is a good approximation in the limit of high signal-to-noise (SNR). If we define the noise-weighted inner product

$$(a|b) = 4{\rm Re}\left[\int \frac{a\cdot b^*}{S_n(f)}{\rm d}f\right], \tag{D1}$$

with $S_n(f)$ being the one-sided noise power spectrum density of the noise, we can compute the Fisher matrix

$$F_{ij} = \left(\frac{\partial h}{\partial\theta_i}\middle|\frac{\partial h}{\partial\theta_j}\right), \tag{D2}$$

where θ is the set of intrinsic $\theta_{\rm int}$ and extrinsic $\theta_{\rm ext}$ parameters describing the signal h and the derivatives are evaluated at the true parameters $h = h_T(\theta)$. The covariance matrix is then simply F^{-1}. The diagonal of this matrix give us a measure of the standard deviation $\sigma_{\theta_i} = \sqrt{(F^{-1})_{ii}}$. For a given model h_T given detector d,

$$\rho_d = \sqrt{(h_T|h_T)}\,, \tag{D3}$$

known as the "optimal SNR".

To estimate the errors in phase-like and amplitude-like parameters, we use a simple two-parameter toy model that can be thought to describe the signal at a fixed frequency

$$h_T = \mathcal{A}e^{i\phi}\,. \tag{D4}$$

From here it is easy to derive that the phase and amplitude will be uncorrelated with standard deviation

$$\sigma_\phi = \sigma_{\mathcal{A}}/\mathcal{A} = 1/\rho_d\,. \tag{D5}$$

In general, the signal will be chirping, and one needs to vary over the reference time and the intrinsic parameters describing the phase evolution like chirp mass. At leading post-Newtonian order for the quadrupolar radiation of a quasi-circular inspiralling binary this can be done analytically [18, 19]. Then one finds that the phase is correlated with the time of arrival and the chirp mass, enlarging the errors by a factor of a few. As discussed in the main text, GW150914 has, for example, $\sigma_{\phi_{\rm H}}(40{\rm Hz}) \simeq 3/\rho_{\rm H}$. The precise number depends on the frequency domain signal and the detector's noise. Importantly, heavier binaries will have more weight on their merger and ring-down, making the measured parameter errors larger for a given

ρ_d and this simple inspiral approximation less reliable. Similarly, signals that do not fit well the data where $(h|h)$ differs substantially from $(h_T|h_T)$, for example if the precession is not properly modelled or glitches remain in the data, can downgrade the errors of the 22-phase.

Moreover, Fisher estimates only include statistical errors whereas our parameter based phase inferences include all errors that are modelled in the original binary parameter inference. LVK analyses marginalize over calibration uncertainties, enlarging amplitude and phase errors. Calibration upper limit errors in LIGO detectors during the third observing run are < 0.07 radians (< 4 degrees) within 20-2000 Hz at 68% CL, although generically they are ~ 0.02 radians [37, 38]. In the first and second observing run LIGO detector had an upper calibration error $\lesssim 0.05$ radians ($\lesssim 3$ degrees) [39]. Note however that these errors are correlated across different frequencies and our phase inference at any given frequency combines information from all frequencies.

For reference, we present in Fig. 17 the product of the standard deviation of the detector phase and the (median) optimal SNR over the parameter posteriors of the events considered in our analysis. The full posterior distribution of the detector phases is given in Fig. 19. One can see how for most of the events the error is $\sim$ few$/\rho_d$, although there are some outliers with larger errors. It is to be noted that, GW191109_010717, the heaviest event in the second half of the third observing run has the largest error. However this event had glitches overlapping with the signal in both detectors and has been shown to keep some anomalies even after removing those, see App. A.3. in [40].

Appendix E: Details on the simulated gravitational waves

In this appendix we detail all the configuration settings for the simulated events. A table with the parameters of the injections is presented in Table II, where we follow the standard parameter definition of `bilby` [24]. For triggers 1, 2 and 3 we perform injections for both HL and HLV detector configurations. For the rest of injections we only set HLV configurations. In total we analyze 17 injections. In all our analyses we use the projected sensitivites for the fourth observing run as described by the amplitude spectral densities `aligo_O4high` for both LIGO detector and `avirgo_O4high_NEW` for Virgo in the public LIGO Document T2000012-v1 (https://dcc.ligo.org/LIGO-T2000012-v1/public). The optimal SNR (for the true injected values) for all the injections is summarized in Table I.

In Fig. 18 we present the posterior distributions for the detector phases of all the injections. For most of the cases, except the HL detections, the detector phases 68% CL is comparable or smaller than $\pi/2\sqrt{12}$, indicating that they are informative against the lensing hypothesis.

Appendix F: Details on the gravitational waves analyzed

To complement the analyses of the main text, in Fig. 19 we present the posterior distribution of the detector phases for all the real events that we have analyzed: 67 binary black holes with $p_{\rm astro} > 0.8$ and $\rho_{\rm ntw} > 8$ in GWTC3 [4]. Following the LVK convention, we use the full name for all the events: UTC date with the time of the event given after the underscore.

From Fig. 19 one can realize that the Hanford and Livingston phases are typically within a fraction of a radian, while Virgo phases are less constrained and in many cases cover the whole prior range.

[1] J. Aasi *et al.*, Classical and Quantum Gravity **32**, 074001 (2015).
[2] F. Acernese *et al.*, Classical and Quantum Gravity **32**, 024001 (2014).
[3] T. Akutsu *et al.* (KAGRA), PTEP **2021**, 05A102 (2021), arXiv:2009.09305 [gr-qc].
[4] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), (2021), arXiv:2111.03606 [gr-qc].
[5] F. Xu, J. M. Ezquiaga, and D. E. Holz, (2021), arXiv:2105.14390 [astro-ph.CO].
[6] M. Çalışkan, J. M. Ezquiaga, O. A. Hannuksela, and D. E. Holz, Phys. Rev. D **107**, 063023 (2023), arXiv:2201.04619 [astro-ph.CO].
[7] O. Hannuksela, K. Haris, K. Ng, S. Kumar, A. Mehta, D. Keitel, T. Li, and P. Ajith, Astrophys. J. Lett. **874**, L2 (2019), arXiv:1901.02674 [gr-qc].
[8] R. Abbott *et al.* (LIGO Scientific, VIRGO), Astrophys. J. **923**, 14 (2021), arXiv:2105.06384 [gr-qc].
[9] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), (2023), arXiv:2304.08393 [gr-qc].
[10] J. Janquart *et al.*, (2023), arXiv:2306.03827 [gr-qc].
[11] K. Haris, A. K. Mehta, S. Kumar, T. Venumadhav, and P. Ajith, (2018), arXiv:1807.07062 [gr-qc].
[12] J. Janquart, O. A. Hannuksela, H. K., and C. Van Den Broeck, Mon. Not. Roy. Astron. Soc. **506**, 5430 (2021), arXiv:2105.04536 [gr-qc].
[13] R. K. L. Lo and I. Magana Hernandez, Phys. Rev. D **107**, 123015 (2023), arXiv:2104.09339 [gr-qc].
[14] M. Raveri and W. Hu, Phys. Rev. D **99**, 043506 (2019), arXiv:1806.04649 [astro-ph.CO].
[15] J. Roulet, S. Olsen, J. Mushkin, T. Islam, T. Venumadhav, B. Zackay, and M. Zaldarriaga, (2022), arXiv:2207.03508 [gr-qc].
[16] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 061102 (2016), arXiv:1602.03837 [gr-qc].
[17] R. Abbott *et al.* (LIGO Scientific, VIRGO), (2021), arXiv:2108.01045 [gr-qc].
[18] L. S. Finn and D. F. Chernoff, Phys. Rev. D **47**, 2198

Injection	Characteristics	Network	$\rho_{\rm H}$	$\rho_{\rm L}$	$\rho_{\rm V}$	$\rho_{\rm ntw}$
1	Type I, low SNR	HL	9.0	12.3	-	15.2
2	Type II, low SNR	HL	6.7	9.1	-	11.3
3	Not-Lensed, low SNR	HL	8.0	8.9	-	12
4	Type I, low SNR	HLV	9.0	12.3	6.9	16.7
5	Type II, low SNR	HLV	6.7	9.1	4.8	12.3
6	Not-lensed, low SNR	HLV	8.0	8.9	3.3	12.4
7	Type I, high SNR	HL	17.2	23.5	-	29.1
8	Type II, high SNR	HL	12.8	17.4	-	21.6
9	Not-lensed, high SNR	HL	15.2	16.9	-	22.7
10	Type I, high SNR	HLV	17.2	23.5	13.1	31.9
11	Type II, high SNR	HLV	12.8	17.4	9.1	23.4
12	Not-lensed, high SNR	HLV	15.2	16.9	6.2	23.6
13	Type I, low SNR, low mass	HLV	9.5	11.1	9.9	17.6
14	Type II, low SNR, low mass	HLV	7.1	9.6	6.1	13.4
15	Not-lensed, low SNR, low mass	HLV	7.5	10.0	5.0	13.5
16	Type I, high SNR, new θ_{JN}	HLV	16.8	23.8	13.4	32.1
17	Type II, high SNR, new θ_{JN}	HLV	12.7	17.8	9.3	23.8

TABLE I. Summary of the true optimal signal-to-noise ratio (ρ_d) in each detector and the network SNR for all simulated events. Each injection has a different noise realization. The network SNR is computed as $\rho^2_{\rm ntw} = \rho^2_{\rm H} + \rho^2_{\rm L} + \rho^2_{\rm V}$.

Injection	$m_1\,[M_\odot]$	$m_2\,[M_\odot]$	a_1	a_2	ϕ_1	ϕ_2	ϕ_{12}	ϕ_{JL}	θ_{JN}	ra	dec	ψ	$\phi_{\rm ref}$	d_L [Mpc]	$t_{\rm ref}$ [sec]	Morse phase
1, 4	35.6	30.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	1.0	0.52	0.7	2.0	1900.0	0.0	0
2, 5	35.6	30.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	1.0	0.52	0.7	2.0	2650.0	517988	$\pi/2$
3, 6	35.6	30.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	0.3	0.18	0.5	3.77	2900.0	3451153	0
7, 10	35.6	30.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	1.0	0.52	0.7	2.0	1000.0	0.0	0
8, 11	35.6	30.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	1.0	0.52	0.7	2.0	1394.0	517988	$\pi/2$
9, 12	35.6	30.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	0.3	0.18	0.5	3.77	1526.0	3451153	0
13	15.4	12.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	1.0	0.52	0.7	2.0	823.0	79121.0	0
14	15.4	12.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	1.0	0.52	0.7	2.0	1147.87	428696	$\pi/2$
15	15.4	12.6	0.2	0.1	0.6	0.3	1.2	0.5	0.8	1.41	0.54	0.5	3.77	1256.16	4918029	0
16	35.6	30.6	0.2	0.1	0.6	0.3	1.2	0.5	2.34	1.0	0.52	0.7	2.0	1000.0	0.0	0
17	35.6	30.6	0.2	0.1	0.6	0.3	1.2	0.5	2.34	1.0	0.52	0.7	2.0	1394.0	517988	$\pi/2$

TABLE II. Summary of the physical parameters for the injections described in Table I: detector frame component masses ($m_{1,2}$), dimensionless spin magnitudes ($a_{1,2}$), the azimuthal angle of the spin vectors ($\phi_{1,2}$), the difference between the azimuthal angles of the individual spin vectors (ϕ_{12}), the difference between total and orbital angular momentum azimuthal angles (ϕ_{JL}), the angle between the total angular momentum and the line of sight (θ_{JN}), right ascension (ra), declination (dec), polarization angle (ψ), phase at reference frequency ($\phi_{\rm ref}$), luminosity distance (d_L) and reference time ($t_{\rm ref}$). The reference frequency is at 20Hz. Injections with Morse phase equal to 0, $\pi/2$ and π correspond to type I, II and III images respectively.

(1993), arXiv:gr-qc/9301003.

[19] C. Cutler and E. E. Flanagan, Phys. Rev. D **49**, 2658 (1994), arXiv:gr-qc/9402014.

[20] P. Schneider, J. Ehlers, and E. Falco, *Gravitational Lenses* (Springer-Verlag Berlin Heidelberg, 1992).

[21] J. M. Ezquiaga, D. E. Holz, W. Hu, M. Lagos, and R. M. Wald, Phys. Rev. D **103**, 064047 (2021), arXiv:2008.12814 [gr-qc].

[22] A. Vijaykumar, A. K. Mehta, and A. Ganguly, (2022), arXiv:2202.06334 [gr-qc].

[23] M. Raveri and C. Doux, Phys. Rev. D **104**, 043504 (2021), arXiv:2105.03324 [astro-ph.CO].

[24] G. Ashton *et al.*, Astrophys. J. Suppl. **241**, 27 (2019), arXiv:1811.02042 [astro-ph.IM].

[25] L. Dai, B. Zackay, T. Venumadhav, J. Roulet, and M. Zaldarriaga, (2020), arXiv:2007.12709 [astro-ph.HE].

[26] X. Liu, I. M. Hernandez, and J. Creighton, Astrophys. J. **908**, 97 (2021), arXiv:2009.06539 [astro-ph.HE].

[27] H. W. Y. Wong, L. W. L. Chan, I. C. F. Wong, R. K. L. Lo, and T. G. F. Li, (2021), arXiv:2112.05932 [gr-qc].

[28] B. F. Schutz, Class. Quant. Grav. **28**, 125023 (2011), arXiv:1102.5421 [astro-ph.IM].

[29] LIGO Scientific Collaboration, "LIGO Algorithm Library - LALSuite," free software (GPL) (2018).

[30] G. Pratten *et al.*, Phys. Rev. D **103**, 104056 (2021), arXiv:2004.06503 [gr-qc].

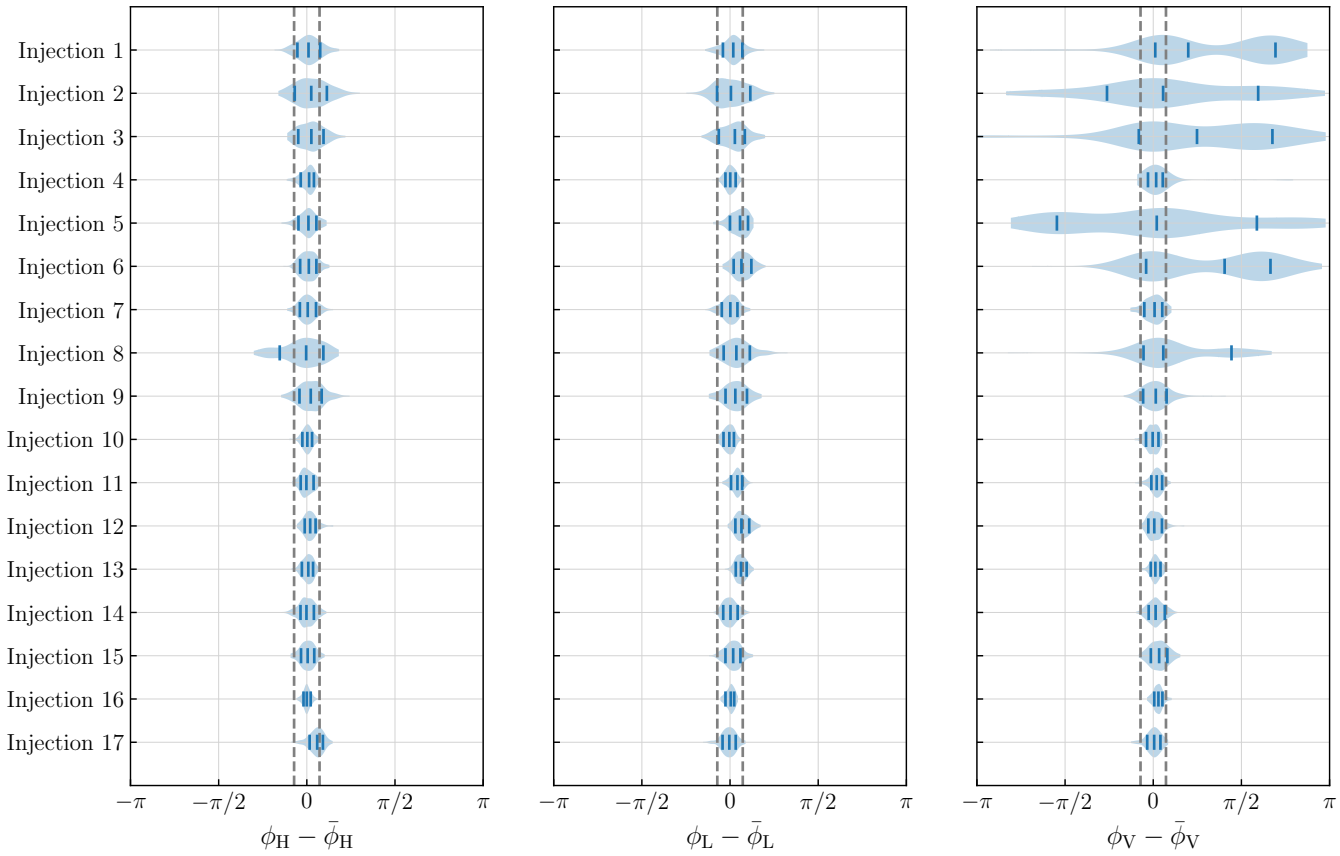


FIG. 18. Posterior distribution of the detector phases centered at their median value for the set of simulated GW events. The blue ticks indicate the 68% CL and the dashed vertical lines indicate the prior width for an informative phase of $\pi/2\sqrt{12}$, see main text for details.

[31] P. Schmidt, I. W. Harry, and H. P. Pfeiffer, (2017), arXiv:1703.01076 [gr-qc].

[32] W. G. Anderson, P. R. Brady, J. D. Creighton, and E. E. Flanagan, Phys. Rev. D **63**, 042003 (2001), arXiv:gr-qc/0008066.

[33] W. Anderson, P. R. Brady, D. Chin, J. D. E. Creighton, K. Riles, and J. T. Whelan, LIGO Internal Document, T010110 https://dcc.ligo.org/LIGO-T010110/public.

[34] M. Isi, (2022), arXiv:2208.03372 [gr-qc].

[35] J. M. Ezquiaga, W. Hu, M. Lagos, and M.-X. Lin, JCAP **11**, 048 (2021), arXiv:2108.10872 [astro-ph.CO].

[36] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, G. Pratten, A. Ramos-Buades, M. Mateu-Lucena, and R. Jaume, Phys. Rev. D **102**, 064002 (2020), arXiv:2001.10914 [gr-qc].

[37] L. Sun *et al.*, Class. Quant. Grav. **37**, 225008 (2020), arXiv:2005.02531 [astro-ph.IM].

[38] L. Sun *et al.*, (2021), arXiv:2107.00129 [astro-ph.IM].

[39] C. Cahillane *et al.* (LIGO Scientific), Phys. Rev. D **96**, 102001 (2017), arXiv:1708.03023 [astro-ph.IM].

[40] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), (2021), arXiv:2112.06861 [gr-qc].

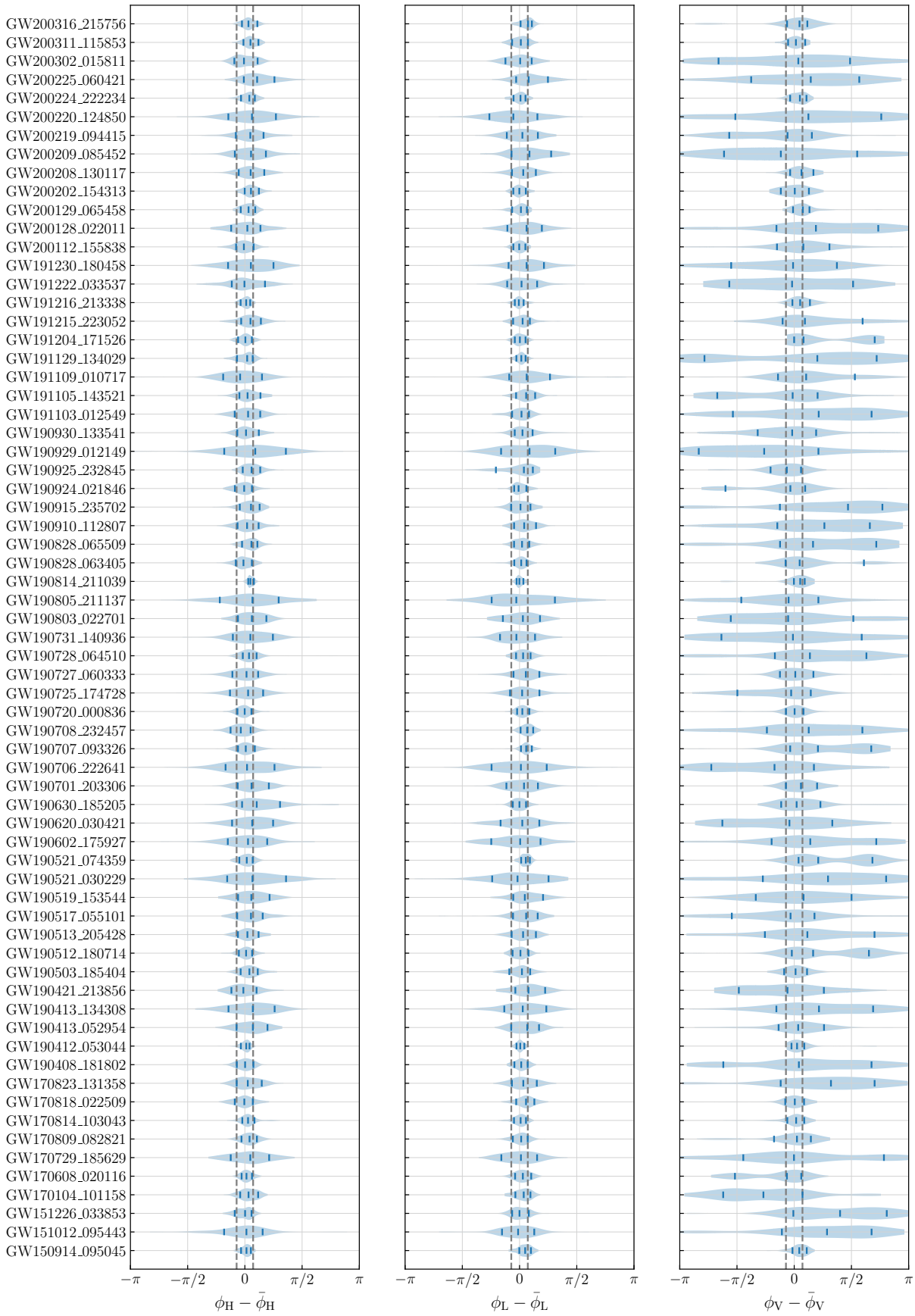


FIG. 19. Posterior distribution of the detector phases centered at their median value for the GW events analyzed in this paper: $p_{\rm astro} > 0.8$ and $\rho_{\rm ntw} > 8$ in GWTC3 [4]. The blue ticks indicate the 68% CL and the dashed vertical lines indicate the prior width for an informative phase of $\pi/2\sqrt{12}$ as in Fig. 18.