Search for strongly lensed counterpart images of binary black hole mergers in the first two LIGO observing runs

Connor McIsaac, David Keitel, Thomas Collett, Ian Harry, Simone Mozzon, Oliver Edy, and David Bacon
University of Portsmouth, Institute of Cosmology and Gravitation, Portsmouth PO1 3FX, United Kingdom
(Dated: 11th December 2019)

Strong gravitational lensing can produce multiple images of the same gravitational-wave signal, each arriving at different times and with different magnification factors. Previous work has explored if lensed pairs exist among the known high-significance events and found no evidence of this. However, the possibility remains that weaker counterparts of these events are present in the data, unrecovered by previous searches. We conduct a targeted search specifically looking for sub-threshold lensed images of known binary black hole (BBH) observations. We recover candidates matching three of the additional events first reported by Venumadhav et al. (2019), but find no evidence for additional BBH events. We also find no evidence that any of the Venumadhav et al. observations are lensed pairs. We demonstrate how this type of counterpart search can constrain hypotheses about the overall source and lens populations and we rule out at very high confidence the extreme hypothesis that all heavy BBH detections truly come from black holes with intrinsic masses $< 15M_\odot$ at high redshift.

I. INTRODUCTION

Dense accumulations of matter, such as galaxies or galaxy clusters, can bend the path of light from sources behind them, an effect known as gravitational lensing [1]. Similarly, gravitational waves (GWs) can be lensed by masses between the source and observer (see e.g. [2, 3]). In the strong lensing regime, multiple images are produced with a delay between arrival times.

Since 2015, Advanced LIGO and Virgo [4, 5] are regularly detecting GWs [6, 7] from coalescing binary black holes (BBHs) and binary neutron stars (BNSs). Such a signal is described by a set of parameters including the masses, spins, location and orientation of the source. The observed (detector-frame) masses are increased by the cosmological redshift [8]. For galaxy or cluster lenses and GW wavelengths in the aLIGO band, geometric optics apply [9, 10]. This means that for multiple images of the same event, the lensed waveforms will be identical up to different arrival times, phases and amplitudes, and have indistinguishable positions on the sky.

The official LIGO-Virgo event catalog GWTC-1 [6] from the O1 and O2 observing runs contains coalescences of 10 BBHs and 1 BNS. The possibility that some of these are lensed images of a single event has previously been suggested [11, 13]. A systematic study [15] found no evidence for multiple images, or any other lensing effects, among the 10 BBHs.

However, one observational signature of strong lensing that has not yet been systematically tested is that large relative magnification between images of the same event can lead to “subthreshold” [16] counterparts to the known events, which the searches in GWTC-1 were not able to confidently extract from the data. In this letter, we perform 10 separate reanalyses of aLIGO O1 and O2 data [17, 19] searching for faint lensed counterparts. We use a single waveform filter “template” for each GWTC-1 event, allowing new candidates to differ only in arrival time, phase and amplitude. While the expected rate of strongly lensed events at current sensitivity is very low [13, 20], our results already provide astrophysical constraints on the population of BBH sources.

II. SEARCH SETUP

For each BBH event from GWTC-1, we search the entire observing run in which it was found, assuming the break between O1 and O2 (316 days) is too long for astrophysically likely lensing time delays. The total coincident time of publicly available data [17, 19] from the two LIGO detectors [5] is 48 days for O1 and 117 days for O2.

Each search uses a single aligned-spin template waveform corresponding to the maximum of the event’s posterior from [21] after removing the precession parameters, which we generate using the “SEOBNRv4_ROM” model [22, 24]. Details on this procedure, and validation that these templates provide sufficient coverage of the posteriors, can be found in appendix A–C.

We use the PyCBC search pipeline [25, 26], with minor modifications for use with a single template and long data stretches. (The original O2 search used $\sim 4 \times 10^5$ templates and analyzed the data in roughly week long chunks [6, 27].) We record all single-detector triggers above a minimum signal-to-noise ratio (SNR) of 4, a lower threshold than in the original GWTC-1 PyCBC O2 search. This is particularly important for identifying candidates in times when detector sensitivities differ. After finding coincident candidates between the two detectors, we compute their significance by comparing their detection statistic (from [21]) against an empirically measured background as in [24]. Our single-template tar-
FIG. 1. Background distribution for the GW150914 counterpart search as a function of the detection statistic, compared to a full template bank as used in the PyCBC search for [6]. The detection statistic is roughly proportional to the signal strain amplitude.

Several of our single-template searches recover other GWTC-1 events with high significance; details of these cross-matches are provided in Table III in appendix D. This is an expected result of the well-known clustering of high-mass detections in a small part of the overall parameter space [29, 31]. As already demonstrated by [15], no such pairs of GWTC-1 events are viable candidates for strongly-lensed double images after considering their relative time delays and the degree of posterior overlap over all relevant parameters.

Table I lists the most significant candidate events from our search after removing pairs of GWTC-1 events. We separate this table into new candidates and those already reported by [29] and later by [30]. As shown in Fig. 2, the set of new candidates is consistent with the null hypothesis, when ranking either by the inverse false-alarm rate or by the delay-weighted $p$-value.

We recover three of the events from [29] (GW170121, GW170304, GW170727), which were not considered in the test for lensed pairs in [15]. For GW170121 and GW170304, the large time delays from their matching GWTC-1 events already suggest that these pairs are unlikely due to lensing (at least by the more common galaxy lenses), but more probably come from two unrelated sources with similar characteristics. The time delay between GW170727 and GW170729 is relatively short, warranting further investigation.

To further study these pairs and to better understand the type of candidates produced by our novel search approach, we perform Bayesian parameter estimation on all candidates in Table I. We obtain posteriors using PyCBC Inference [32, 33] with the aligned-spin IMRPhenomD waveform model [34, 35], using identical priors in every case. We then compute Bayes factors (evidence ratios) $B_{L/U}$ based on posterior overlap integrals [15, 36] to determine whether each pair of events can be efficiently described by a single set of parameters, which could indicate a strongly lensed source (L); or has inconsistent parameters, preferring independent sources (U). See appendix E for details. Results are included in Table I.

The $B_{L/U}$ values (computed over the intrinsic parameters: $m_1, m_2, s_1^z, s_2^z$) are, for most pairs, nominally supportive of the L hypothesis. However, these results need to be considered in context of several mitigating facts. First, the expected lensing rate at O1/O2 sensitivities is very small under standard astrophysical assumptions 13 20, meaning that much higher Bayes factors would be needed to obtain high posterior odds after factoring these in as explicit priors between the two hypotheses. In addition, our search by construction produces candidates with high overlap in the intrinsic parameters, and a high rate of purely coincidentally consistent (i.e. unlensed) events is to be expected when the source population occupies only a small region of parameter space. Also, most new candidates in the first half of the table are low-significance and consistent with background, while the Bayes factors are only meaningful if both candidates are astrophysical. For the stronger events, particularly the previously known ones, most time delays are very long compared to standard lens expectations 13, and most sky localizations are not very consistent.

In summary, the $B_{L/U}$ analysis does not offer clear ev-
TABLE I. Recovered candidates sorted by their delay-weighted \( p \)-value which prefers short time delays (first column), cut at \( \leq 0.5 \). Candidates with an inverse false-alarm rate (second column, derived from the original ranking statistic) \( \geq 1 \) year are also included, regardless of their time delay. The third column gives the candidate end time (UTC) and the fourth column notes if an event has already been published by \([29, 30]\) (listed separately in the lower half of the table). The GWTC-1 event whose counterpart search found the candidate is listed in the fifth column, and the absolute value of the time delay is given in the sixth column. Candidates that are themselves listed in GWTC-1 are excluded; see Table III in the appendix for those. The final two columns are Bayes factors derived from the posterior overlap between each pair: \( B^\text{int}_{L/U} \) for the intrinsic parameters \((m_1, m_2, s_1^z, s_2^z)\) and \( B^\text{sky}_{L/U} \) for the sky localization \((\alpha, \delta)\).

| delay-weighted \( p \)-value | inverse false-alarm rate \( [\text{yr}] \) | UTC time | known event? found by | \( |\Delta t| [\text{d}] \) | \( B^\text{int}_{L/U} \) | \( B^\text{sky}_{L/U} \) |
|-------------------------------|---------------------|----------|---------------------|-----------------|-------------------|-------------------|
| 0.16                         | 0.166               | 2017-07-30 08:05:26.8 | GW170729 | 0.548 | 20 | 0.2 |
| 0.29                         | 0.0066              | 2017-01-04 10:12:57.9 | GW170104 | 0.00687 | 5 | 2 |
| 0.37                         | 0.497               | 2017-08-04 14:57:29.3 | GW170809 | 4.73 | 70 | 0.2 |
| 0.4                          | 0.00550             | 2017-07-29 19:05:05.9 | GW170729 | 0.00598 | 7 | 0.2 |
| 0.46                         | 0.000465            | 2015-09-14 10:04:34.7 | GW150914 | 0.00960 | 30 | 0.2 |
| 0.47                         | 0.000131            | 2017-01-04 10:14:56.3 | GW170104 | 0.00206 | 4 | 0.9 |
| 0.48                         | 0.000241            | 2017-07-29 18:51:31.4 | GW170729 | 0.00345 | 7 | 0.1 |
| 0.86                         | 2.53                | 2017-04-01 08:19:53.0 | GW170809 | 130 | 10 | 5 |
| 0.0066                       | 2246                | 2017-01-21 21:25:36.6 | GW170121 | GW170818 | 208 | 60 | 2 |
| 0.065                        | 191                 | 2017-01-21 21:25:36.6 | GW170121 | GW170823 | 214 | 2 | 0.2 |
| 0.096                        | 1.57                | 2017-07-27 01:04:30.0 | GW170727 | GW170729 | 2.74 | 8 | 0.2 |
| 0.36                         | 15.5                | 2017-03-04 16:37:53.4 | GW170304 | GW170729 | 147 | 20 | 0.4 |
| 0.49                         | 12.4                | 2017-03-04 16:37:53.4 | GW170304 | GW170823 | 172 | 5 | 0.2 |
| 0.99                         | 1.14                | 2017-03-04 16:37:53.4 | GW170304 | GW170818 | 166 | 0.2 | 6 |

FIG. 2. Left pane: The cumulative count of triggers with inverse false-alarm rates less than or equal to a given value. Right pane: The cumulative count of triggers with delay-weighted \( p \)-values greater than or equal to a given value. The dashed line is the distribution of background triggers produced by Poissonian noise. The crosses are foreground triggers, excluding those from GWTC-1 and those reported in \([29]\). The foreground triggers are consistent with the null hypothesis.

IV. IMPLICATIONS OF THE ABSENCE OF CLEAR COUNTERPART CANDIDATES

We now explore the astrophysical implications of this result. First, we evaluate the sensitivity of each search using simulated signals with parameters drawn from the posterior of the corresponding GWTC-1 event. Their strain amplitudes are multiplied by a scale factor \( \sqrt{\mu_{\text{rel}}} = \sqrt{\mu_0/\mu_{\text{inj}}} \), where \( \mu_0 \) and \( \mu_{\text{inj}} \) are the absolute magnifications of the primary and simulated signals respectively. The simulated signals cover a range of time delays with respect to the GWTC-1 events from less than a second to one month (either side) and \( \sqrt{\mu_{\text{rel}}} \in [0, 1, 10] \).
The recovery rate of these signals yields an estimate for the probability of finding a lensed image as a function of magnification ratio $\mu_{\text{rel}}$ and time delay $\Delta t$.

These results are summarized in Fig. 3 using the same thresholds as for Table I; an inverse false-alarm rate $\geq 1$ year or delay-weighted $p$-value $\leq 0.5$. Along the $\mu_{\text{rel}}$ dimension, recovery mostly depends on the strength of the original GW event. Along the time dimension, recovery is mostly limited by the $\lesssim 50\%$ coincident livetime of the LIGO detectors during O1 and O2.

Astrophysical interpretation of the absence of convincing counterpart candidates depends on the choice of priors for the true (unlensed) high-redshift BBH source population and the properties of lenses in the Universe. We provide supplementary data for the sensitivity of all 10 searches so that other authors may explore different choices. Here, we demonstrate one application, testing a strict interpretation of the idea from [12]: What if all of the high-mass GWTC-1 BBH events really came from lighter objects at higher redshifts?

To be more specific, we phrase the test hypothesis as: The intrinsic component masses of any BBH in the Universe cannot be larger than $15 M_\odot$. All apparently heavier GW events are due to lensing. For simplicity, we assume that their primary masses are exactly $m_1 = 15 M_\odot$. This is an unreasonable source distribution, but including a distribution of lower masses would only make the following argument stronger. The ratio of intrinsic and observed masses yields corrected redshifts and luminosity distances, from which the required magnifications are between 100 and 800 with GW150914 having the largest magnification. (See Table IV in appendix F)

Gravitational lensing theory suggests that such highly magnified events are unlikely to appear without a second image of comparable strength and short time delay. Building a model of lensing by the galaxies in the Universe [37], we find that only 2% of images with $\mu > 800$, or 4% with $\mu > 100$, have no counter image with $\mu_{\text{rel}} \leq 3$. The typical time delay for such highly magnified pairs are seconds to minutes. See appendix G for details.

Marginalizing over an ensemble of lenses and highly magnified source positions and using a more conservative threshold (delay-weighted $p$-value $< 0.16$; more significant than anything in the top half of Table I) yields expected recovery fractions ranging from $95\%$ for GW150914 to $84\%$ for GW170104. Combining the seven heaviest events, we find a probability of $1.0 \times 10^{-7}$ for observing these without detecting any counter images. Therefore, our lack of detecting any lensed counter images conclusively rules out the hypothesis that all high-mass detections are lensed events with intrinsic masses below $15 M_\odot$.

V. CONCLUSION AND OUTLOOK

We have performed the first focused search for strongly lensed counterpart images to all binary black holes from the GWTC-1 catalog [6]. We recovered several candidates previously found by [29, 30]. Performing follow-up investigations of these candidate events, we found no evidence that these are lensed counterparts. All other new candidates are consistent with a noise-only background.

The absence of clear candidates constrains astrophysical lensing scenarios. For example, if all observed BBHs had originated from lower mass, highly magnified, high redshift sources, then we should have observed at least one counterpart. We therefore rule out this hypothesis.

Another method to search for sub-threshold lensed events has been proposed in [16]. The single-template searches in this letter provide less freedom to match detector noise fluctuations than the template bank employed in [16]. For future applications, the optimal template bank size per event might lie between our single-template method and the larger banks of [16]. For example one could construct a template bank to obtain a certain minimal match across each event’s posterior. We will explore this in more detail in future work.
A significant improvement could also come from including sky location consistency in the search stage, for example using a multi-detector coherent search [38], to eliminate candidates that match well in intrinsic parameters but have poor overlap on the sky.

Looking ahead, the ongoing LIGO-Virgo O3 run has already yielded a rich crop of additional GW candidates [7], and future observing runs promise many more [9]. It has also been suggested that the first detection of a lensed source is expected within the next 5 years [20, 40]. The framework of targeted sub-threshold searches for lensed counterparts as presented in this letter can be readily applied to new detections in O3 and beyond. Observing strongly lensed BBHs before the detector network reaches design sensitivity [39] would imply that the merger rate increases much more steeply with redshift than expected [13, 31, 41], or challenge the established understanding of lensing statistics.

More generally, once strongly lensed pairs of events can be identified, joint parameter inference on the combined images can significantly improve estimates of the source properties and location. This type of search will become a powerful probe of BBHs at high redshifts beyond the usual detection horizon.

Supplementary data for this letter is available at: https://github.com/icg-gravwaves/lensed-o1-o2-data-release

**ACKNOWLEDGMENTS**

We thank Tjonnie Li, Otto Hannuksela and Rico Ka-Lok Lo for useful discussions. This research has made use of data obtained from the Gravitational Wave Open Science Center, a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. LIGO is funded by the U.S. National Science Foundation. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by Polish and Hungarian institutes. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459, as well as for additional computational resources provided by Cardiff University and funded by STFC grant ST/I006285/1. This paper has been assigned document number LIGO-P1900360.
M. Purrer, Frequency domain reduced order models for gravitational waves from aligned-spin compact binaries, Class. Quant. Grav. 33, 215004 (2016). arXiv:1508.02357 [gr-qc].

T. Dal Canton and I. W. Harry, Improving the template bank to observe compact binary coalescences in Advanced LIGO’s second observing run, arXiv preprints (2017), arXiv:1705.01845 [gr-qc].

A. H. Nitz, Distinguishing short duration noise transients in LIGO data to improve the PyCBC search for gravitational waves from high mass binary black hole mergers, Class. Quant. Grav. 35, 035016 (2018). arXiv:1709.08974 [gr-qc].

T. Vennumadhav, B. Zachary, J. Roulet, L. Dai, and M. Zaldarriaga, New Binary Black Hole Mergers in the Second Observing Run of Advanced LIGO and Advanced Virgo, arXiv e-print (2019). arXiv:1904.07214 [astro-ph.HE].

A. H. Nitz, T. Dent, G. S. Davies, S. Kumar, C. D. Capano, I. Harry, S. Mazzan, L. Nuttall, A. Lundgren, and M. Tápai, 2-OGC: Open Gravitational-wave Catalog of binary mergers from analysis of public Advanced LIGO and Virgo data, arXiv preprints (2019), arXiv:1910.05331 [astro-ph.HE].

B. P. Abbott et al. (LIGO Scientific Collaboration and Virgo Collaboration), Binary Black Hole Population Properties Inferred from the First and Second Observing Runs of Advanced LIGO and Advanced Virgo, Phys. Rev. Lett. 122, 021106 (2019). arXiv:1811.02453 [astro-ph.HE].

C. M. Biwer, C. D. Capano, S. De, M. Cabero, D. A. Brown, A. H. Nitz, and V. Raymond, PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals, Publ. Astron. Soc. Pac. 131, 024503 (2019). arXiv:1807.10312 [astro-ph.IM].

[33] A. N. Nitz et al., gwastrob/pycbc: Release 1.13.8 of PyCBC (2019).

S. Husa, S. Khan, M. Hannam, M. Purrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal, Phys. Rev. D 93, 044006 (2016). arXiv:1508.07253 [gr-qc].

K. Haris, A. K. Mehta, S. Kumar, T. Vennumadhav, and P. Ajith, Identifying strongly lensed gravitational wave signals from binary black hole mergers, arXiv e-print (2018), arXiv:1807.07062 [gr-qc].

T. E. Collett, The Population of Galaxy-Galaxy Strong Lenses in Forthcoming Optical Imaging Surveys, Astrophys. J. 811, 20 (2015) arXiv:1507.02657 [astro-ph.CO].

T. W. Harry and S. Fairhurst, A targeted coherent search for gravitational waves from compact binary coalescences, Phys. Rev. D 83, 084002 (2011), arXiv:1012.4939 [gr-qc].

B. P. Abbott et al. (LIGO Scientific Collaboration and Virgo Collaboration), Prospects for Observing and Localizing Gravitational-Wave Transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Rev. Rel. 21:3 (2018), 10.1007/s41114-018-0012-9 (2018), arXiv:1304.0670 [gr-qc].

S.-S. Li, S. Mao, Y. Zhao, and Y. Lu, Gravitational lensing of gravitational waves: A statistical perspective, Mon. Not. R. Astron. Soc. 476, 2220 (2018). arXiv:1802.05089 [astro-ph.CO].

M. Fishbach, D. E. Holz, and W. M. Farr, Does the Black Hole Merger Rate Evolve with Redshift?, Astrophys. J. Lett. 863, L41 (2018). [Astrophys. J. Lett.863,L41(2018)], arXiv:1805.10270 [astro-ph.HE].

M. Hannam, P. Schmidt, A. Bohé, L. Haege1, S. Husa, F. Ohme, G. Pratten, and M. Purrer, Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms, Phys. Rev. Lett. 113, 151101 (2014). arXiv:1308.3271 [gr-qc].

B. P. Abbott et al., Multi-messenger Observations of a Binary Neutron Star Merger, Astrophys. J. Lett. 848, L12 (2017) arXiv:1710.05833 [astro-ph.HE].

B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An Algorithm for detection of gravitational waves from inspiraling compact binaries, Phys. Rev. D85, 122006 (2012). arXiv:gr-qc/0509116 [gr-qc].

B. P. Abbott et al. (LIGO Scientific Collaboration and Virgo Collaboration), GW170608: Observation of a 19-solar-mass Binary Black Hole Coalescence, Astrophys. J. 851, L35 (2017) arXiv:1711.05578 [astro-ph.HE].

D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC Hammer, Publ. Astron. Soc. Pac. 125, 306 (2013) arXiv:1202.3665 [astro-ph.IM].

W. D. Vousten, W. M. Farr, and T. Mandel, Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations, Mon. Not. R. Astron. Soc. 455, 1919 (2016) arXiv:1501.05823 [astro-ph.IM].
Appendix A: Template selection

For each of the 10 GWTC-1 BBH events, we have taken the public posterior samples\cite{GWTC1} produced with the precessing IMRPhenomPv2 waveform\cite{LVC2018,LVC2016,Miller2017}. Since there is no evidence for precession in any of these events\cite{Miller2017} and the PyCBC search pipeline in its standard configuration currently relies on aligned-spin waveforms, we select aligned-spin waveforms near the peak of each posterior. To do so, we train a four-dimensional aligned-spin kernel density estimator (KDE) in $\{m_1, m_2, a_1, a_2\}$ on each set of samples and obtain the maximum-posterior (MaP) set of parameters from it. We then generate SEOBNRv4.ROM waveforms\cite{O1,vacalo2017} at these MaP parameters as the search templates. These are listed in Table I. We do not include the BNS event GW170817 in this analysis, since its close distance and extensive electromagnetic observational coverage\cite{2018ApJ...869L..19P} already rule out strong lensing.

We compute the matches

$$m(h_1, h_2) = \max_{t_0, \phi_0} \left( \frac{h_1}{\sqrt{(h_1|h_1)}} \frac{h_2}{\sqrt{(h_2|h_2)}} \right), \quad (A1)$$

In this appendix we provide additional details of the methods used in our analysis, as well as extended listings of search results and a description of available machine-readable data products. We discuss the selection of the single template used for each search in Sec. I and additional aspects of the search setup and data set used in Sec. III.

This is followed by an extended discussion of search validation tests in Sec. II. In Sec. III we provide a list of recovered candidates, extending Table I from the main letter. We then describe the method used to calculate the posterior overlap Bayes factors in Sec. IV. The details of the extreme lensing hypothesis test are discussed in Sec. V followed in Sec. VI by a full description of the astrophysical priors on magnification and time delay used in the test.

APPENDIX

In this appendix we provide additional details of the methods used in our analysis, as well as extended listings of search results and a description of available machine-readable data products. We discuss the selection of the
defined using the usual inner product \((h_1|h_2)\) [44], of these (aligned) MaP waveforms against any draws from the whole set of (precessing) posterior samples. The matches are high for the higher-SNR events (e.g. a worst match of 95% for GW150914) and for the bulk of the posteriors for all events. They fall off significantly for some outliers in the far tails of lower-SNR events (e.g. worst match of 50% for GW151012, though 90% of its samples still have matches > 89%).

Hence, an additional step of validation is required to show that these template choices are good enough for single-template lensing counterpart searches: as described in the following section, we test the recovery for simulated signals (“injections”) with parameters drawn from the full precessing posteriors.

For further reference, Fig. 4 shows the cross-matches between the selected templates for the 3 events from O1 and 7 events from O2.

**Appendix B: Details on data set and search implementation**

Our data set includes GW strain data from the two LIGO detectors [5] in Hanford (Washington, USA) and Livingston (Louisiana, USA). The O1 data set [18] from September 2015 to January 2016 has a total length of 130 days, with 48 days of coincident science mode data from both detectors. The O2 data set [19] from December 2016 to August 2017 covers a total of 268 days, containing 117 days of usable coincident data.

We do not include the short data stretch around GW170608 when the Hanford detector was nominally out of observing mode [45]. For robust statements on any candidates from this stretch, we would have needed to include and characterize all data in similar states to produce a consistent background. Instead, we accept this as another blind period similar to any other when not both of the detectors were online in nominal observing state. Hence, there is also no coverage of the \(|\Delta t| < 1\) h set of simulated signals (see Fig. 3) for this event. However, due to its low mass [6, 45], GW170608 is not relevant for the hypothesis test in Sec. II anyway.

Here we also note a technical aspect of the search implementation: with standard settings, PyCBC reports no more than one candidate signal within a ±10s window [25], so that we would not have been able to recover any counterpart signals with time-delays shorter than 10s around each GWTC-1 event, regardless of their strength. To ensure that we do not miss anything, we perform additional reanalyses with this clustering criterion disabled over these narrow time windows, for all nine events besides GW170608, and find no additional candidates.

**Appendix C: Search validation and sensitivity estimation**

To validate template selection and search setup we evaluate the recovery of simulated signal injections. While each of the 10 searches uses a single aligned-spin SEOBNRv4_ROM template, for the main injection sets used throughout the paper we draw parameters from the full GWTC-1 posterior samples [21] and use the same precessing IMRPhenomPv2 waveform model that those samples come from to perform the actual injections.

For each of the 10 searches we run two sets of injections, one uniformly covering a full month of data either side of the original GWTC-1 event, and one focusing on a smaller window 2 hours either side. From the broader injection set we recover the sensitive distances of each search, as listed in Table II. The sensitive distance is calculated by applying a detection threshold on the false-alarm rate of 1 per year and finding the detection efficiency in a number of distance bins. The volume of each distance bin is multiplied by the search efficiency before being summed and converted to a sensitive distance.

As seen in Sec. A the single template will have a greater mismatch when moving away from the MaP values. Therefore, we investigate how much the sensitive distance could be improved by using additional templates across each event’s posterior. Instead of analyzing each injection with an extended template bank that covers the whole posterior we do a simpler test with a separate set of injections with parameters identical to the MaP values, still searching for them with the same fixed single template.

As an example, we consider GW151012, the lowest-
SNR event. Fig. 5 shows the difference in recovery for the two injection sets, corresponding to $\sim 13\%$ in sensitive distance. This is the largest possible increase in sensitive distance out of the 10 searches. The sensitive distances for the alternative injection set for all 10 searches are also listed in Table I.

These provide an estimate of the gains in sensitivity possible by using a realistic template bank covering the full posterior. However, in practice a search with a wider template bank will also increase the rate of background triggers, reducing the sensitivity. Additionally, in this test the injected signals are a perfect match to the search template, a practical template bank will still have some mismatch due to: (i) the discrete placement of the templates; (ii) the current search not including precessing templates in the analysis; (iii) mismatch between the real signal and waveform model. This test therefore represents an upper limit on the sensitivity that could be gained using a larger template bank.

For the rest of the paper we will always refer to the injection SNR defined as the smallest of the two optimal SNRs in each detector.

FIG. 5. Fraction of recovered signal injections for GW151012 for two injection sets: one (‘single’) with all parameters fixed to those of the search template, and one (‘posterior’) with parameters drawn from the full GWTC-1 posterior. An injection is counted as recovered if found with an inverse false-alarm rate greater than one year. The horizontal axis gives the injection SNR defined as the smallest of the two optimal SNRs in each detector.

FIG. 6. Example of the detailed injection recovery results going into Fig. 5. Probability of finding a lensed image of GW150914 as a distribution over time delay $|\Delta t|$ and the scale factor $\sqrt{\mu_0/\mu_0}$. Calculated using injections drawn from the posterior of GW150914, with a threshold on the inverse false-alarm rate of one year.

which do not themselves correspond to another GWTC-1 event. An extended version of the full search results is provided here as Table III, which includes all candidates above either of those two thresholds from any of the ten searches, and both sets of known events from GWTC-1 and from $\psi_9$ are mixed in with the new candidates.

We also provide the full set of search results, without thresholds on $p$-value or false-alarm rate, as machine-readable supplementary data files.

Appendix E: Posterior overlap Bayes factors

For each of the candidates listed in Table I as well as the corresponding GWTC-1 events, we use PyCBC Inference [32, 33] to produce posterior samples with the emcee [34, 35] sampler and the aligned-spin IMRPhenomD waveform model [34, 35]. We use a fixed prior for all events: $t_c$ uniform within $0.2\text{ s}$ around the time reported by the search, component masses uniform within $[5, 90] M_\odot$, spins uniform in $[-0.99, 0.99]$, distance uniform in $[10, 5000] \text{ Mpc}$ and uniform angular priors on $i, \psi, \alpha, \delta$.

For each pair of candidates/events $i \neq 1$, we start from the individual posteriors $P(\theta_i | d_i, I)$ in the parameters $\theta_i$ as sampled over the two data sets $d_i$. As derived by [36], a simple lens hypothesis $H_L$ states that $\theta_i' = \theta_2'$ for some sub-set of parameters $\theta'$, while the unlensed hypothesis $H_U$ allows for independent parameter sets. The Bayes factor between the two is equal to the evidence ratio:

$$B_{L/U} = \frac{Z_L}{Z_U} = \int d\theta' P(\theta' | d_1, I) P(\theta' | d_2, I) \frac{P(\theta' | I)}{P(\theta' | I)}$$

where the numerator is the shared prior and the remaining parameters are ignored. Essentially, this quantifies

Appendix D: Full search results

Table IV in the main part of the paper provides only those candidates found with a delay-weighted $p$-value $< 0.5$ or an inverse false-alarm rate of more than one year.

Appendix E: Posterior overlap Bayes factors

For each of the candidates listed in Table IV as well as the corresponding GWTC-1 events, we use PyCBC Inference [32, 33] to produce posterior samples with the emcee [34, 35] sampler and the aligned-spin IMRPhenomD waveform model [34, 35]. We use a fixed prior for all events: $t_c$ uniform within $0.2\text{ s}$ around the time reported by the search, component masses uniform within $[5, 90] M_\odot$, spins uniform in $[-0.99, 0.99]$, distance uniform in $[10, 5000] \text{ Mpc}$ and uniform angular priors on $i, \psi, \alpha, \delta$.

For each pair of candidates/events $i \neq 1$, we start from the individual posteriors $P(\theta_i | d_i, I)$ in the parameters $\theta_i$ as sampled over the two data sets $d_i$. As derived by [36], a simple lens hypothesis $H_L$ states that $\theta_i' = \theta_2'$ for some sub-set of parameters $\theta'$, while the unlensed hypothesis $H_U$ allows for independent parameter sets. The Bayes factor between the two is equal to the evidence ratio:

$$B_{L/U} = \frac{Z_L}{Z_U} = \int d\theta' P(\theta' | d_1, I) P(\theta' | d_2, I) \frac{P(\theta' | I)}{P(\theta' | I)}$$

where the numerator is the shared prior and the remaining parameters are ignored. Essentially, this quantifies
the amount of overlap between the two posteriors, inversely weighted by the prior — i.e., posteriors that jointly peak in a region of lower prior support would provide stronger preference for a joint origin.

Which parameter set $\theta'$ to evaluate the overlap integral over depends on the waveform model used and on which parameters are expected to be shared in a lensing scenario. Distance, phase and polarization angle are naturally excluded from $B_{L/U}$, since they can be expected to be different between the two signals in a lensed pair. (The apparent inferred distance follows the lensing magnification.) While $|\mathcal{I}|$ evaluated $B_{L/U}$ over a set $\theta' = \{m_1, m_2, a_1, a_2, \alpha, \delta, i\}$ for a precessing waveform model, here we simplify the problem by considering a spin-aligned model, drop-a precessing waveform model, here we simplify the

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
delay-weighted p-value & false-alarm rate $^{-1}$ [yr] & UTC time & known event? & $|\Delta t|$ [d] \\
\hline
$\ll 0.001$ & $3.26 \times 10^7$ & 2017-08-14 10:30:43.5 & GW170814 & 222 \\
$\ll 0.001$ & $3.26 \times 10^7$ & 2017-01-04 10:11:58.6 & GW170104 & 222 \\
$\ll 0.0001$ & $8.15 \times 10^6$ & 2017-08-23 13:13:58.5 & GW170823 & 5.5 \\
$\ll 0.0001$ & $3.62 \times 10^6$ & 2017-08-23 13:13:58.5 & GW170823 & 25 \\
$\ll 0.0001$ & $1.09 \times 10^6$ & 2017-01-04 10:11:58.6 & GW170104 & 217 \\
$\ll 0.0001$ & $1.78 \times 10^5$ & 2017-08-09 08:28:21.8 & GW170809 & 5.1 \\
$\ll 0.0001$ & $4.89 \times 10^3$ & 2017-08-09 08:28:21.8 & GW170809 & 8.7 \\
0.0028 & 151 & 2017-08-14 10:30:43.5 & GW170814 & 5.1 \\
0.0045 & 3.94 \times 10^3 & 2017-01-04 10:11:58.6 & GW170104 & 226 \\
0.0066 & 2.25 \times 10^3 & 2017-03-21 21:25:36.6 & GW170121 & 208 \\
0.0073 & 1.93 \times 10^3 & 2017-08-09 08:28:21.8 & GW170809 & 214 \\
0.0076 & 30 & 2017-08-14 10:30:43.5 & GW170814 & 3.7 \\
0.014 & 74 & 2017-08-23 13:13:58.5 & GW170823 & 14 \\
0.035 & 43 & 2017-07-29 18:56:29.3 & GW170729 & 25 \\
0.065 & 191 & 2017-03-21 21:25:36.6 & GW170121 & 214 \\
0.096 & 1.6 & 2017-07-27 01:04:30.0 & GW170727 & 2.7 \\
0.14 & 5.3 & 2017-06-09 08:28:21.8 & GW170809 & 14 \\
0.14 & 11 & 2015-09-14 09:59:45.4 & GW150914 & 28 \\
0.16 & 0.17 & 2017-07-30 08:05:26.8 & GW170729 & 0.55 \\
0.17 & 30 & 2017-08-23 13:13:58.5 & GW170823 & 9.1 \\
0.29 & $9.13 \times 10^{-5}$ & 2017-01-04 10:12:57.9 & GW170104 & 6.87 \times 10^{-4} \\
0.36 & 16 & 2017-03-04 16:37:53.4 & GW170304 & 147 \\
0.37 & 0.50 & 2017-08-09 08:28:21.8 & GW170809 & 4.7 \\
0.4 & $5.50 \times 10^{-4}$ & 2017-07-29 19:05:05.9 & GW170729 & 5.98 \times 10^{-3} \\
0.45 & 19 & 2017-01-04 10:11:58.6 & GW170104 & 231 \\
0.46 & $4.65 \times 10^{-4}$ & 2015-09-14 10:04:34.7 & GW150914 & 9.60 \times 10^{-3} \\
0.47 & $1.31 \times 10^{-4}$ & 2017-01-04 10:14:56.3 & GW170104 & 2.06 \times 10^{-3} \\
0.48 & $2.41 \times 10^{-4}$ & 2017-07-29 18:51:31.4 & GW170729 & 3.45 \times 10^{-3} \\
0.49 & 12 & 2017-03-04 16:37:53.4 & GW170304 & 172 \\
0.85 & 4.1 & 2017-08-23 13:13:58.5 & GW170823 & 231 \\
0.86 & 2.5 & 2017-04-01 08:19:53.0 & GW170809 & 8.9 \\
0.91 & 3.3 & 2017-01-04 10:11:58.6 & GW170104 & 206 \\
0.99 & 1.1 & 2017-04-04 16:37:53.4 & GW170304 & 166 \\
\hline
\end{tabular}
\caption{List of all candidates with a delay-weighted p-value (first column) $< 0.5$ or an inverse false-alarm rate (second column) of more than one year. In contrast to Table I all new candidates as well as any known events from both GWTC-1 and 29 are combined together, all sorted by the delay-weighted p-value. The third column gives the candidate end time (UTC) and the fourth column notes if an event at this time has already been published. The GWTC-1 event whose search found the reported candidate is listed in the fifth column, and the absolute value of the time delay between the two is given in the final column.}
\end{table}

As for the practical implementation, we use Gaussian KDEs (in the scipy [49] implementation) to compute the overlaps. For $B_{L/U}$, on each posterior we train a single four-dimensional KDE weighted by $1/\sqrt{P(\theta | L)}$, after applying a logit transform to the samples to deal with boundary effects, then evaluate the product integral (including the transform Jacobian). Since the KDE code internally normalizes the weights, we finally multiply the overlap integral by the mean of the inverse priors to obtain the Bayes factor. (When all priors are uniform, this prescription reduces to a product of unweighted KDEs divided by the total prior volume.)

For $B_{L/U}$, where the posteriors have much more complicated substructure than in the intrinsic parameters, we use a $k$-means clustering algorithm [50] to split each posterior into a number of clusters (adaptively chosen by optimizing the Bayesian Information Criterion (BIC) [51]), compute a weighted KDE for each cluster, then compute the sum of overlap integrals over all pairs of clusters and again divide by the mean inverse prior.
Appendix F: Extreme lensing hypothesis

Let us start from the following quantities under the standard no-lensing hypothesis from GWTC-1 [6]: primary source-frame masses $m_{1,U}$, redshifts $z_U$ and luminosity distances $d_U$. We then define the extreme lensing hypothesis as there being no merging BBHs in the universe with component masses over $15M_\odot$, i.e. fixing the intrinsic (source-frame) primary mass of each event to $m_{1,L} = 15M_\odot$. [12] originally phrased the limit in terms of chirp mass, but since it is motivated by observed terms of chirp mass, but since it is motivated by observed to

\[ m = \frac{m_{1,U}}{15 M_\odot} (1 + z_U) - 1. \]  

(F1)

These are then converted to luminosity distances under the standard Planck cosmology [52], and the corresponding lensing magnification factors are $\mu = (d_L/d_U)^2$. Table [IV] lists the results for the eight heaviest BBH sources (with median $m_1 > 15 M_\odot$), using median results from [6] for simplicity.

GW151226 and GW170608 have median $m_1 < 15 M_\odot$ and hence are irrelevant for the extreme hypothesis test. GW151012 at median $m_1 \approx 23 M_\odot$ is included in the table for completeness but also not used in the test, because its required magnification does not fall into the sufficiently extreme regime for the approximations discussed below and the result of the test is already sufficiently strong from just considering the seven heavier sources.

Appendix G: Astrophysical priors on magnification / time delay

The hypothesis that the observed high mass BBH events are actually magnified low mass BBH events has been invoked to remove the need for BHs with masses above $15 M_\odot$ [12]. For GW150914 this implies a magnification, $\mu$, of at least 800. High magnification events are rare in the Universe ($P(\mu) \sim \mu^{-3}$), but they are possible, particularly for point sources. For example a star at redshift 1.5 has been observed with a magnification of $\sim 2000$ [54]. The catastrophe theory of strong gravitational lensing [55] shows that very high magnification images are formed when the source lies either: (i) just inside a fold catastrophe, forming a pair of images with the same brightness; (ii) just inside a cusp catastrophe, forming a triplet with one image twice as bright as the other two; (iii) or just outside the cusp catastrophe, forming a single highly magnified image [1]. Higher order catastrophes can produce more complicated configurations [56] [57] but are extremely rare [58] [59]. The time delays between the multiple highly magnified images are extremely short, with $\Delta t \sim \mu^{-3}$.

The lensing mass of galaxies is well approximated by singular isothermal ellipsoids [60]. In this case the fraction of highly magnified images without a comparably bright counterimage is given by [61]

\[ P_{\text{single}} = \frac{1}{1 + 4\pi \frac{15}{16\pi^3} \frac{\sqrt{1-q^2}}{1+q} \mu^{1/2}} \]  

(G1)

where $q$ is the axis ratio of the lens. Thus, unless the lens is very close to spherical ($q = 1$), highly magnified images are unlikely to occur without a bright counter image. In the case of a lensed BBH, we do not know the specific lens and cannot measure $q$ directly. Instead we must marginalise over the population of all potential lenses in the Universe. We use the lens population model of [37] to realise the population of gravitational lenses in the Universe; this model includes masses and ellipticities derived from SDSS and accounts for the correlation between mass and ellipticity. The model includes both the mass function of galaxies [62] and the strong lensing cross section for each galaxy. Using Equation (G1) and marginalising over the lens population we find that only 2% of lensed images with a magnification above 800 are without a comparably bright counter image. We define the probability of not having a bright counter image as $P_{\text{single}}$ in Table [IV]. The same lens model allows us to numerically infer the time delay and magnification ratio between counter images. For each putative lensed BBH event, we realise 100000 lens systems weighted by their lensing cross section given the true source redshift. For each lens we draw a random image position and infer the magnification; we do this repeatedly until we find 1000 image positions that have a magnification within 10% of the putative magnification. Each image position is then traced back onto the source plane. We then solve the lens equation numerically to find the counter images, the time delays between counter images and the magnification ratios. This model shows that for the highly magnified images with bright counter images, 90% (99%) of the counter images occur within 5 (45) minutes for GW170823 and 2 (15) seconds for GW150914. For each simulated image, the probability of missing the counter image is given by one minus the recovery fraction for injections with the correct time delay and strain ratio relative to the observed BBH event. Marginalising over all of the simulated image pairs gives the probability of missing a comparably bright counterimage, $P_{\text{missed}}$. The probability of seeing a counter image is thus

\[ P_{\text{found}} = 1 - P_{\text{single}} - (1 - P_{\text{single}})P_{\text{missed}} \approx 1 - P_{\text{single}} - P_{\text{missed}} \]  

(G2)

These values are shown for each event in Table [IV]. The product of $1 - P_{\text{found}}$ is $1 \times 10^{-7}$. This is the probability of missing the counter images for all of the 7 events assuming they are all lensed by the magnification corresponding to a maximum component mass of $15 M_\odot$. If
TABLE IV. Parameters of the eight heaviest GWTC-1 events, reinterpreted under the extreme lensing hypothesis that they all should have intrinsic primary masses of $m_{1,U} = 15 M_\odot$. Unlensed parameters correspond to the median values from [6]. Luminosity distances are obtained under standard Planck cosmology [52]. $P_{\text{single}}$ is the probability of an event of this magnification not having a comparably bright counter image. $P_{\text{missed}}$ is the probability that our search fails to recover the counter image in the LIGO data (with delay-weighted p-value below 0.16). $P_{\text{found}}$ is the probability of our search recovering a counter image under the extreme lensing hypothesis. GW151012 is excluded from the analysis as we only consider systems with magnification $> 100$.

| event     | $m_{1,U}$ [M$_\odot$] | $z_{U}$ | $z_L$ | $d_{U}$ [Gpc] | $d_L$ [Gpc] | $\mu$ | $P_{\text{single}}$ | $P_{\text{missed}}$ | $P_{\text{found}}$ |
|-----------|------------------------|---------|-------|---------------|-------------|------|---------------------|---------------------|---------------------|
| GW150914  | 36                     | 0.09    | 1.6   | 0.4           | 12          | 800  | 0.020              | 0.026               | 0.955               |
| GW151012  | 23                     | 0.21    | 0.9   | 1.1           | 5.8         | 30   | —                  | —                   | —                   |
| GW170104  | 31                     | 0.20    | 1.5   | 1.0           | 11          | 130  | 0.046              | 0.123               | 0.837               |
| GW170729  | 50                     | 0.49    | 4.0   | 2.8           | 37          | 170  | 0.040              | 0.304               | 0.860               |
| GW170809  | 35                     | 0.20    | 1.8   | 1.0           | 14          | 200  | 0.038              | 0.065               | 0.900               |
| GW170814  | 31                     | 0.12    | 1.3   | 0.6           | 9.3         | 260  | 0.034              | 0.039               | 0.928               |
| GW170818  | 35                     | 0.21    | 1.8   | 1.1           | 19          | 200  | 0.038              | 0.050               | 0.914               |
| GW170823  | 40                     | 0.35    | 2.5   | 1.9           | 21          | 130  | 0.046              | 0.118               | 0.841               |

The model presented here does not include lensing by clusters. The mean Einstein radius in the model is 0.7 arcseconds, and the time delays are proportional to the Einstein radius, so even if clusters dominate the lensing cross section, the expected time delays would only increase by a factor of a few as the Einstein radius of typical clusters is $\sim 5$ arcseconds [63]. The model also does not account for deviations from isothermality. This introduces a small change in the constant of proportionality in the time delays between highly magnified image pairs. To assess the potential size of these systematics, we rerun our pipeline, but assuming all time delays are 10 times longer than in the fiducial model: in this scenario the probability of missing all of the counter images increases to $5 \times 10^{-7}$. Thus, the lack of detecting any lensed counterimages conclusively rules out the hypothesis that all of the high mass events are lensed events with intrinsic masses below $15 M_\odot$. 