SPIIR online coherent pipeline to search for gravitational waves from compact binary coalescences

Qi Chu, Manoj Kovalam, Linqing Wen, Teresa Slaven-Blair, Joel Bosveld, Yanbei Chen, Patrick Clearwater, Alex Codoreanu, Zhihui Du, Xiangyu Guo, Xiaoyang Guo, Kyungmin Kim, Tjonnie G. F. Li, Victor Oloworaran, Fiona Panther, Jade Powell, Anand S. Sengupta, Karl Wette, and Xingjiang Zhu

Phys. Rev. D **105**, 024023 — Published 6 January 2022

DOI: 10.1103/PhysRevD.105.024023

The SPIIR online coherent pipeline to search for gravitational waves from compact binary coalescences

Qi Chu,[1,2] Manoj Kovalam,[1,2,*] Linqing Wen,[1,2,†] Teresa Slaven-Blair,[1,2] Joel Bosveld,[2] Yanbei Chen,[3] Patrick Clearwater,[1,4] Alex Codoreanu,[1,4] Zhihui Du,[5] Xiangyu Guo,[6] Xiaoyang Guo,[6] Kyungmin Kim,[7] Tjonnie G. F. Li,[8,9] Victor Oloworaran,[2] Fiona Panther,[1,2,‡] Jade Powell,[1,10] Anand S. Sengupta,[11] Karl Wette,[1,12] and Xingjiang Zhu[1,13]

[1] *Australian Research Council Centre of Excellence for Gravitational Wave Discovery (OzGrav)*
[2] *Department of physics, University of Western Australia, Crawley WA 6009, Australia*
[3] *Theoretical Astrophysics 350-17, California Institute of Technology, Pasadena, CA 91125, USA*
[4] *Gravitational Wave Data Centre, Swinburne University, Hawthorn VIC 3122, Australia*
[5] *Department of Computer Science, New Jersey Institute of Technology, New Jersey 07102, USA*
[6] *Department of Computer Science and Technology, Tsinghua University, Beijing, China*
[7] *Korea Astronomy and Space Science Institute, 776 Daedeokdae-ro, Yuseong-gu, Daejeon 34055, Republic of Korea*
[8] *Department of Physics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong*
[9] *Centre for Gravitational Waves, Institute for Theoretical Physics, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium*
[10] *Centre for Astrophysics and Supercomputing, Swinburne University, Hawthorn Victoria 3122, Australia*
[11] *Department of Physics, Indian Institute of Technology Gandhinagar, Gujarat 382355, India*
[12] *School of Physics, Australian National University, Acton Australian Capital Territory 2601, Australia*
[13] *School of Physics and Astronomy, Monash University, Clayton, Victoria 3800, Australia*

(Dated: November 27, 2021)

This paper presents the SPIIR pipeline used for public alerts during the third advanced LIGO and Virgo observation run (O3 run). The SPIIR pipeline uses infinite impulse response (IIR) filters to perform extremely low-latency matched filtering and this process is further accelerated with graphics processing units (GPUs). It is the first online pipeline to select candidates from multiple detectors using a coherent statistic based on the maximum network likelihood ratio statistic principle. Here we simplify the derivation of this statistic using the singular-value-decomposition (SVD) technique and show that single-detector signal-to-noise ratios from matched filtering can be directly used to construct the statistic. Coherent searches are in general more computationally challenging than coincidence searches due to extra search over sky direction parameters. The search over sky directions follows an embarrassing parallelization paradigm and has been accelerated using GPUs. The detection performance is reported using a segment of public data from LIGO-Virgo's second observation run. We demonstrate that the median latency of the SPIIR pipeline is less than 9 seconds, and present an achievable roadmap to reduce the latency to less than 5 seconds. During the O3 online run, SPIIR registered triggers associated with 38 of the 56 non-retracted public alerts. The extreme low-latency nature makes it a competitive choice for joint time-domain observations, and offers the tantalizing possibility of making public alerts prior to the merger phase of binary coalescence systems involving at least one neutron star.

Keywords: gravitational waves; low-latency search pipeline; coherent search

I. INTRODUCTION

Gravitational wave (GW) astronomy has been advancing rapidly since the first operation (referred to as O1) of the two Advanced Laser Interferometer Gravitational-wave Observatory (aLIGO) in 2015 [1]. The advanced Virgo detector [2] joined the LIGO detectors from the second observation run (O2). The third and latest run (O3) of aLIGO and Virgo lasted 11 months and finished in March 2020 [3, 4]. There have been over a dozen detections of compact binary coalescences (CBCs) by the LIGO collaboration in the O1 and O2 run [5–7]. The first six months of O3 has seen three times more detections than O1 and O2 combined [8]. With new advanced detectors, the KAGRA detector in operation since 2020 [9] and LIGO-India envisioned to be operational in the next decade [10], it is expected the GW astronomy will see regular frequent detections and possible new breakthroughs.

A high priority for observing runs is to use GW detections as a trigger for electromagnetic (EM) follow-up observations. A successful example is the first detection of GW from merging binary neutron stars (GW170817 [11]) with multiple coincident observations across electromagnetic spectrum [12]. Improved detector sensitivity will expose more binary merger signals, as well as enabling detections ahead of the final coalescence phase, known as the early warning detections. To facilitate the real-time and early-warning detections, a public alert infras-

* manoj.kovalam@research.uwa.edu.au
† linqing.wen@uwa.edu.au
‡ fiona.panther@uwa.edu.au

tructure[1] was established by LIGO-Virgo collaboration before O3 to enable online data streaming, pipeline detection and prompt real-time alert publication.

Five detection pipelines have been used in this infrastructure: one to search for unmodeled signals, the cWB pipeline [13]; four to search for the modeled compact binary coalescence (CBC) signals, including GstLAL [14–16], MBTA [17], PyCBCLive [18–20], and the SPIIR pipeline which is the focus of this paper. Previous work on the SPIIR pipeline development can be found in [21–26] and the SPIIR pipeline started its online trial runs during the O1 and O2 runs.

The SPIIR pipeline is distinguished from other CBC search pipelines in several aspects. It adopts the summed parallel infinite impulse response (SPIIR) method for matched filtering [21–23]. This method is expected to be more efficient computationally than the traditional Fourier method when a filtering delay of less than 10s is intended [21]. It is straightforward to parallelize this algorithm using Graphics Processing Units (GPUs), a popular and cost-effective parallel computing platform [24, 25].

The other main difference is that the SPIIR pipeline selects candidates based on the maximum network likelihood ratio principle which is referred as the coherent method. Coherent methods have been developed for periodic GW searches [27], inspiral searches in the band of the proposed space-based interferometric detector, LISA [28–30], and for GRB-triggered CBC searches [31, 32]. It was proposed for CBC searches [33–35] but not widely used due to computational challenges of searching through additional parameters of source sky directions. A recent work [36] is proposed to reduce the computational cost of parameter search using particle swarm optimization. In this paper, we express the coherent method for CBC signals using singular value decomposition (SVD). SVD and its variation principal component analysis have been applied to many areas of GW research including waveform decomposition [37, 38] and parameter estimation [39]. It has been proposed for general formalization of GW data analysis with a detector network [40]. The SVD derivation here is an extension of [40] and simplifies the expression of the coherent statistic for CBC searches. It shows that output from matched filtering, i.e. the signal-to-noise ratio (SNR) time series, can be directly used and only the two parameters of sky directions need to be searched over. The search on each sky direction is independent and thus has been distributed to parallel GPU threads for acceleration.

This paper is organized as follows: Sec. II gives a detailed explanation of the SPIIR pipeline. Sec. III reports the performance of the SPIIR pipeline using a segment of the public O2 data, including a break-down analysis of contributions to the pipeline latency. It also reports the results of the pipeline on O3 public alerts. Sec. IV gives conclusion and future perspectives of the pipeline.

II. PIPELINE DESCRIPTION

A flowchart of the SPIIR pipeline is shown in Fig. 1. The elements of the flowchart may be grouped into five stages:

(A) A pre-processing stage, where live data streams from detectors LIGO-Livingston (L1), LIGO-Hanford (H1), and Virgo (V1) are read, conditioned using data quality (DQ) channels, downsampled, whitened, and conditioned again.

(B) A filtering stage, where whitened data are convolved with SPIIR filters in the time domain to approximate the matched filtering result – SNR time series.

(C) A coherent search stage, where candidates from individual detectors are searched over companion detectors to form coherent candidates.

(D) Candidate significance estimation, where background (noise) events generated by time-shifted data are used to estimate the false alarm rate (FAR) of a candidate.

(E) Candidate veto and submission, where candidates are tested against a few statistic thresholds and submitted to the GW candidate event database (GraceDB).

A. Data conditioning, down-sampling, and whitening

The first stage of the SPIIR pipeline uses the data acquisition module from the `gstlal` software package[2] to read in live (optionally offline) data. During O3, the live data were delivered in one-second packets. The first step of the pipeline is to "Apply DQ channel" as shown in Fig. 1. The DQ channel uses bit masks to mark the periods of the strain data when not in the lock condition or affected by known loud noises. The pipeline replaces the strain data in these periods with zeros (a.k.a. gating) with a tapering filter. This zeroing method could introduce power spectral leakage in the frequency domain. The data is whitened later in the frequency domain by the pipeline where this leakage will be carried over, which could result in a slightly inaccurate whitening outcome. The leakage can be avoided if data is whitened in the time domain or by adoption of an in-painting method instead

[1] `https://emfollow.docs.ligo.org/userguide/`

[2] `https://git.ligo.org/lscsoft/gstlal`

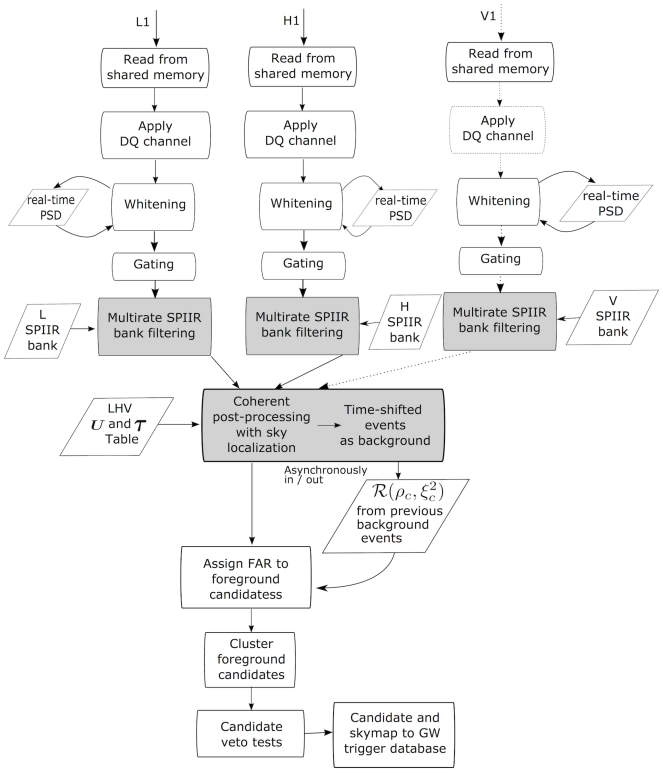


FIG. 1. A flowchart of the online SPIIR pipeline. Rectangular blocks denote pipeline components; those in gray have been accelerated by GPUs. Parallelogram blocks denote input for given components. At the start, the pipeline reads live data from LIGO-Livingston (L1), LIGO-Hanford (H1), and optionally Virgo (V1) (marked by the dashed line) on the computing cluster. The pipeline applies per-detector DQ flags from the DQ channel to gate noise-affected segments in the main strain data. The data are downsampled from 16 kHz to 2 kHz and whitened based on a real-time estimate of the noise power spectral density (PSD). After whitening, a gating process vetoes noise outliers. The gated data are then filtered using SPIIR filters (Sec. II B) and a coherent search is performed (Sec. II C). An asynchronous process collects background events from time shifted data to update the distribution of the ranking statistic $\mathcal{R}$ from which the FAR of a foreground candidate is estimated. Candidates are then clustered and tested against thresholds before submitted along with the sky map of coherent statistics to the GW candidate database (GraceDB).

of the zeroing method shown in the recent work [41]. Both methods will be considered for the pipeline for the next observation run.

The strain data are then downsampled from the original 16384Hz to a lower rate of 2048 Hz to reduce the computational cost downstream. This reduced rate is sufficient to capture CBC signals within the most sensitive band (15-1000 Hz) of the LIGO-Virgo detectors. The pipeline uses the `resample` module from the `gstreamer`[3] library for down-sampling, which implements the method in time-domain using a finite impulse response (FIR) filter.

The data are then whitened (refers to whitening in Fig. 1) in the frequency domain that can be expressed as:

$$d_w(t) = \int_{-\infty}^{\infty} \frac{\tilde{d}(f)}{\sqrt{S_n(|f|)}} e^{i2\pi ft}, \tag{1}$$

where $\tilde{d}(f)$ is data in the frequency domain obtained by Fourier transform and $d_w(t)$ denotes the whitened data. $S_n(|f|)$ is the one-sided noise power spectra density (PSD) defined through ensemble average $E()$ of the noise spectrum $E(\tilde{n}(f)\tilde{n}^*(f')) = 1/2 S_n(f)\delta(f - f'), f > 0$. In an ideal world where the noise is stationary, the past data can be used for noise PSD estimation to whiten the current segment of data. However the noise is known to be non-stationary in the LIGO-Virgo online data. The noise PSD is therefore estimated by tracking the geometric median PSD from overlapping 4-second blocks of immediate past data spanning 56 seconds. This estimation converges quickly and is robust against glitches (See Sec. II. B of [14]). It is implemented by the `lal_whiten` module from the `gstlal` library.

The pipeline applies a second "gating" function after whitening, where outstanding amplitude excursions of the whitened data are replaced with zeros. This is to remove glitches that are not identified by the online DQ channels. We also replace 0.25 second of data on each side of the gated segment with zeros to suppress excess power from the glitch. This step would not cause spectral leakage problems as following steps are performed in the time domain.

B. Time-domain matched filtering with SPIIR filters

1. *SPIIR method*

The matched filtering method is the optimal method to search for known signals from Gaussian noise. It has been used in CBC searches as the CBC GW waveform templates can either be derived from post-Newtonian perturbation theories or from numerical solutions of Einstein's field equations [42–44]. The parameters of the CBC waveforms can be divided into two sets, the intrinsic and the extrinsic parameter sets. The intrinsic set is related to the intrinsic properties of the GW sources — masses and spins. The extrinsic parameters include the distance l, the source location (α, δ), the inclination angle ι and the polarization angle ψ, the phase ϕ_c and the time t_c at the coalescence. The extrinsic parameters can be represented by two parameters effectively — the effective distance $l_{\rm eff}$ and the termination phase ϕ_0 [45].

[3] `https://gstreamer.freedesktop.org/`

The CBC signal $h(t)$ can then be expressed as:

$$h(t) = \frac{l}{l_{\rm eff}} A(t) \cos\left(\phi(t) + \phi_0\right), \tag{2}$$

where $A(t)$ and $\phi(t)$ are the amplitude and phase evolution respectively. The effective distance is a function of the antenna responses to the two GW polarizations ($F^{+,\times}$) and the inclination angle:

$$l_{\rm eff} = \frac{l}{\sqrt{F^{+2}(1+\cos^2\iota)^2/4 + F^{\times 2}(\cos\iota)^2}}, \tag{3}$$

and ϕ_0 is the coalescence phase with modulation from the antenna response functions:

$$\phi_0 = \phi_c - \arctan\left(\frac{F^{\times}(2\cos\iota)}{F^{+}(1+\cos^2\iota)}\right). \tag{4}$$

To search for the unknown phase ϕ_0, a common way is to use a two-phase matched filter with orthogonal phases. This can be implemented as a complex filter [45]:

$$h_T(t) = h(t, \phi_0 = 0) + ih(t, \phi_0 = \pi/2). \tag{5}$$

This filter is then whitened and normalized by its expected value at an effective distance of 1 Mpc, denoted by h_{wT}. Using h_{wT} to cross-correlate the whitened data, one obtains the matched filtering result, the complex SNR time series $z(t)$:

$$z(t) = \int_0^\infty d_w(\tau) h_{wT}(t+\tau) {\rm d}\tau. \tag{6}$$

The detected phase ϕ_0, the coalescence time, and the effective distance can then be obtained from the detected SNR.

The SPIIR method uses a chain of first-order impulse response (IIR) filters to approximate h_{wT}. Each IIR filter is consist of three coefficients to approximate a small segment of the waveform: the feedforward coefficient b_1, the feedback coefficient a_0 and the delay of the filter t_d. The summation of responses from all IIR filters is the approximated waveform denoted by $u(t)$. The SPIIR method is applicable to approximation of any analytical or numerical binary waveforms [21–23]. The metric to measure the approximation accuracy is the overlap O:

$$O = \frac{(h_{wT}, u)}{\sqrt{(u,u)}\sqrt{(h_{wT}, h_{wT})}}, \tag{7}$$

where $(,)$ is the inner product[4]. We optimize the coefficients such that the overlap is over 99% which corresponds to a SNR loss of less than 1% (see Sec. III A however for a realistic scenario). The complex SNR from the SPIIR filtering with $N_{\rm f}$ filters for a given template can be expressed in a discrete form as:

$$z[k] = \sum_{m=0}^{N_{\rm f}} \left(a_{0,m} z[k - 1 - t_{d,m}/\Delta t] + b_{1,m} d_w[k - t_{d,m}/\Delta t]\right), \tag{8}$$

where k is the discrete time, Δt is the interval of time samples and m denotes the index of the filter

2. *Computational cost and GPU acceleration*

A total of 12 floating point operations are required to calculate the SNR with one IIR filter in Eq. 8 [21, 22]. The total computational cost of the SPIIR filtering in a search is proportional to the number of filters over all search templates. Denote the average number of filters per template by $N_{\rm f}$, the number of waveform templates by N_T, the sample rate by N_R and the number of detectors by N_d, the computational cost for SPIIR filtering each second is then $\mathcal{O}(12 N_{\rm f} * N_T * N_R * N_d)$.

A total of 412 000 templates were used by SPIIR during O3 covering the source component mass of $1.1 - 100\ M_\odot$. In a typical setting of O3 where $N_{(m)}$ is around 350, N_R is 2048 Hz and there are 3 detectors, the total computation is about 10.6 Tera floating point operations per second (TFLOPS), requiring 2200 typical 4.8-GFLOPS central processing unit (CPU) cores to process in real-time.

The high demand on CPUs can be mitigated by use of GPUs. The filtering process is a multiple instruction single data process. Filtering operations of templates and of SPIIR filters are independent that they can be distributed in parallel to GPU threads. It is shown that with a moderate GPU, a speed-up of more than 100 over a single-core CPU can be achieved [24–26]. The speed-up can easily scale up with more GPU cores. The frequent release of new massive-core GPU hardwares is likely to accommodate the increased computational demand due to increased detector sensitivity.

C. Coherent trigger generation and localization

1. *Coherent network SNR*

The coincidence search method has long been used to search for event candidates from a detector network where high SNR triggers from individual detectors are selected first and those coincident in time are selected as candidates. The coherent search on the other hand will look for both time and phase coherent triggers from individual detectors based on the maximum network log likelihood ratio (LLR) principle. Previous work can be found in [32–35] and Appx. V A of this paper shows the network LLR derivation can be simplified mathematically using SVD. The coherent statistic from the network

[4] https://en.wikipedia.org/wiki/Dot_product

LLR is referred to the coherent SNR throughout the paper and is expressed with the SVD form as:

$$\rho_{\rm C}^2(\alpha,\delta,t_c,\boldsymbol{\Theta}) = \left\| \boldsymbol{I}\boldsymbol{U}^T \begin{pmatrix} z_1 & 0 & \dots & 0 \\ 0 & z_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_{N_d} \end{pmatrix} \right\|^2, \quad (9)$$

where $\rho_{\rm C}^2$ is the coherent SNR; $\|\|$ is the Euclidean norm; $\boldsymbol{I}$ is a diagonal matrix with the first two elements nonzero $\boldsymbol{I} = \mathrm{diag}\{1,1,0,...,0\}$, and z is the complex SNR of each detector offset by different arrival times of a signal, and N_d is the number of detectors in the network. $\boldsymbol{U}$ is an $N_d \times N_d$ unitary matrix from the SVD of the noise-weighted detector response $\boldsymbol{G}_\sigma$ given in Eq. 27. The coherent SNR dependent on three sets of unknown parameters: the intrinsic source parameters $\boldsymbol{\Theta}$, the sky direction (α,δ), and the coalescence time t_c.

The first two columns of $\boldsymbol{U}$, spans a plane to capture the two signal polarizations from N_d detectors. The complementary statistic, the null SNR which is from the null-space spanned by the remaining $N_d - 2$ columns of $\boldsymbol{U}$, should only include noise contributions. It can be written as:

$$\rho_{\rm NULL}^2 = \sum_I \rho_I^2 - \rho_{\rm C}^2. \quad (10)$$

where ρ_I is the absolute value SNR $\rho_I = |z_I|$. In the true direction, the per-detector signal arrival times and signal phases are coherently matched, the projected signals should preserve the entire signal power and the projection for the null SNR should only retains noise. In other directions where times and phases are not perfectly matched, the coherent SNR will not preserve all signal power, instead some power will leak to the null space. This lends to a localization method that explores the phase information and is expected to be better than the simple triangularization method using arrival time information.

Before we perform coherent searches, we first select candidates from each detector. A candidate is selected if $\rho_I > 4$ (the same threshold used by the GstLAL pipeline [14]). The threshold is less than the fiducial detection threshold of 8 to allow for sub-threshold trigger and background formation. The candidates are grouped by each template bank (A template bank is a set of around 10^3 templates grouped by chirp mass values for computational convenience.) over each second, and are filtered with the following procedure:

- Find the maximum SNR across templates of the template bank at each time sample and select those with $\rho_I > 4$.
- If a template triggers multiple high SNRs within one second, only select the candidate with the highest ρ_I.

This procedure will ensure that there is at most one trigger from each template of a bank at different time samples each second.

For each candidate, we then search for maximum coherent SNR using SNR time series from companion detectors. To search over sky directions, we use HEALPix [46] to divide the whole sky into 3072 equal-area curvilinear tiles (a.k.a. pixels). The centre coordinates for each tile are used as sky direction samples. We estimate the maximum error on coherent SNR using this sampling. The worst-case scenario is when the signal lies at the boundary of a sky tile. The angular distance between any tile centre and its boundary is 1.9^o. This corresponds to a maximum 0.7 ms shift in time from the true time and a SNR loss of 9% from the peak SNR of a typical $1.4\mathrm{M}_\odot + 1.4\mathrm{M}_\odot$ system with the O2 L1 detector sensitivity. The major coherent SNR loss therefore comes from the misidentified peak SNR. The misidentified individual SNRs will be projected to the coherent SNR using the projection matrix $\boldsymbol{U}$ (Eq. 29). The SNR loss from this inaccurate projection due to the inaccurate sky direction is 0.05%. For more accurate sky tiling, readers are referred to [31, 32, 35].

The coherent SNR calculation depends on not only the source sky directions relative to the detectors but also detector sensitivities. We assume that the noise in one detector does not change much during an observing run so that a representative horizon distance is used for the sensitivity representation. To save run-time calculation of the projection matrix $\boldsymbol{U}$, we sample $\boldsymbol{U}$ every half hour up to 24 hours and update it every day to accommodate the change of detector locations due to Earth's rotation. Relative time offsets of GW arrival times for different detectors are also sampled at the same interval along with $\boldsymbol{U}$. The pipeline will calculate the coherent SNR using the two information that are closest in the time to the triggering candidate. The error introduced in this implementation of approximation of $\boldsymbol{U}$ to the coherent SNR value is subject to the sky sampling error explained in the last paragraph but it could affect the detected sky direction up to about 4 degrees in the East-West direction.

2. ξ^2 test for signal consistency

In addition, the pipeline uses another statistic [14, 47] to test the consistency of the detected SNR series from each detector with expectation in time domain. The expected SNR series is projected from the autocorrelation of a template. The discrete form of the statistic ξ_I^2 for detector I is given by:

$$\xi_I^2 = \frac{\sum_{j=-N_j}^{N_j} |z_I[j] - z_I[0]A_I[j]|^2}{\sum_{j=-N_j}^{N_j}(2 - 2|A_I[j]|^2)}, \quad (11)$$

where $A_I[j]$ is the correlation function of a template with itself and N_j is the number of time samples for comparison. The numerator is summation of a group of χ^2 statistics and the denominator is the overall degree of freedom.

The fraction is then a reduced χ^2 with a mean value of 1 [14].

We also compute the average value of ξ_I^2, denoted as $\xi_{\rm C}^2$:

$$\xi_{\rm C}^2 = \frac{1}{N_d}\sum_I \xi_I^2. \tag{12}$$

This average statistic is used along with the coherent SNR to to form a ranking statistic used to rank coherent candidate events shown in Sec. II D 1.

3. Background events from time-shift

To quantify the significance of our coherent candidate by its false alarm rate, incoherent events are constructed as background events. This is done by applying the conventional time shift technique. For every SNR >= 4 candidate from Sec. II C 1, we apply time shifts on other detector SNR time series with sufficiently large time offsets and use them as the background data. The minimum offset is 0.1 seconds, which is much longer than the GW travel time between any two detectors. The number of time shifts is chosen to be 100, as limited by the GPU memory to store past data for the statistics. This sets the lower limit of our FAR from the 100 time shifts of one-week data to be around 0.5/yr. FAR values beyond this limit are extrapolated using K-nearest-neighbour (KNN) techniques shown in Sec. II D 1.

4. Computational cost and GPU acceleration

The computational cost for coherent search and background collection using Eq. 29 is estimated here. For each candidate, $\mathcal{O}(4N_d^2)$ floating point operations are needed to compute the coherent SNR and the null SNR with a pre-calculated look-up table for each sky direction. Here 4 is account for one complex multiplication and one complex addition. N_d is the number of detectors. The look-up tables are prepared every day and the computing of them are negligible. The maximum number of candidates per detector per second is the number of templates N_T. The maximum total FLOPS is therefore $\mathcal{O}(4N_d^3 N_T N_p)$ where N_p is the number of sky locations, For computational efficiency, the number of sky areas searched for background events is reduced to 768 which corresponds to a maximum SNR loss of 13% compared to the optimal SNR. Note we perform 100 times calculation for the background than the foreground. The computation of ξ^2 for each trigger is negligible compared to the coherent network SNR computing. For a typical case of 412 000 templates as in O3 and three detectors, the worst-case cost is about 3.5-TFLOPS including both foreground and background calculations, comparable to the cost of the SPIIR filtering. However, in practice the number of candidates is usually one or two orders of magnitudes smaller than N_T. Therefore the computational cost of coherent search could be significantly less than the SPIIR filtering. The use of GPUs to accelerate this stage is described in [26].

D. Ranking statistic and false alarm rate estimation

1. Ranking statistic

A coherent trigger is selected based on a list of statistics: $\{\rho_1, \xi_1^2, \rho_2, \xi_1^2, ...\rho_{N_d}, \xi_{N_d}^2, \rho_C, \xi_C^2\}$. The current coherent trigger selection criteria is to require that the SNR contribution from non-triggering detectors is reasonably significant so that $\rho_C \geq \rho_{\rm I} + \sqrt{2}$ ($\rho_{\rm I}$ is the SNR of a single-detector candidate at detector I (Sec. II C 1)). This requirement is expected to be removed in the future to allow single-detector candidates.

To construct the ranking statistic, we use the coherent network SNR ρ_C which reflects the overall signal strength in the detector network and ξ_C^2 for the average score from the signal morphology test. The coherent statistic and the null statistic are most useful in signal and noise classification when there are at least three detectors with comparable sensitivities. For a network of three detectors with one much less-sensitive detector as in the case of O2 and O3, the values of coherent statistics would be close to the network SNR (quadrature sum of individual SNRs) from the two dominating detectors, L1 and H1. Besides V1's SNR contribution is mostly random noise given the low sensitivity of V1. To deal with this, the pipeline provides the option to use the dominating detectors for the overall signal strength. ρ_C in the remainder of this section would be replaced with $\sqrt{\rho_L^2 + \rho_H^2}$.

The coherent candidates are ranked mathematically by integrating the background probability $P(\rho_C, \xi_C^2|n)$ as:

$$\mathcal{R}(\rho_C^{'}, \xi_C^{2'}) = \int_{\rho_C \geq \rho_C^{'}} \int_{\xi_C^2 \leq \xi_C^{2'}} P(\rho_C, \xi_C^2|n)\, {\rm d}\xi_C^2\, {\rm d}\rho_C \tag{13}$$

where $P(|)$ is the conditional probability and n denotes the background events constructed by time shifts. In implementation, the background events are collected into 300×300 bins of a 2-dimension histogram.

2. False alarm rate estimation

The significance of a candidate is quantified by the candidate's false alarm probability, which is determined by calculating the cumulative distribution of the ranking statistic $\mathcal{R}$:

$$P(\mathcal{R} < \mathcal{R}^{'}|n) = \sum_{\mathcal{R}<\mathcal{R}'} P(\mathcal{R}|n), \tag{14}$$

where $P(\mathcal{R}|n)$ is the discrete probability of $\mathcal{R}$, calculated by integrating the probability $P(\rho_C, \xi_C^2|n)$.

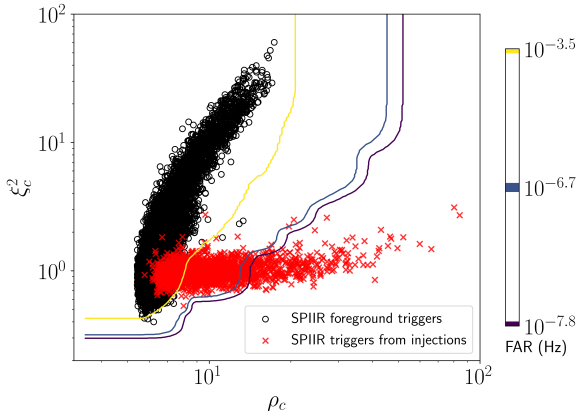


FIG. 2. Triggers from one-week of O2 noise data (black circles), triggers from the binary black hole injections described in Sec. III A (red crosses), and false alarm rate (FAR) estimations using background events of one-week O2 data. Three representative FAR lines are shown corresponding to significance levels of 25.6/day (GW database internal submission threshold), one per two months corresponding to the O3 open public alert threshold, and 0.5/yr which is our FAR limit from data.

THe FAR of a given trigger is calculated by its false alarm probability and the observed noise trigger rate:

$$\mathrm{FAR}(\mathcal{R}') = \frac{P(\mathcal{R} < \mathcal{R}'|n)N_b}{T_b}. \tag{15}$$

where N_b is the number of background events and T_b is the background collection time which is equivalent to the foreground candidate collection time multiplied by the number of time shifts performed. Obviously, different collection of background events will result in a different FAR value. During O3, background events were grouped in each computing node based on chirp mass, with 4000 templates processed per node. We normalized the final FAR of a candidate across the total number of computing nodes to ensure the number of triggers uploaded is consistent with the expectation from the FAR.

For the O2 injection test and the O3 run, the pipeline collects one week of background events for FAR estimation. This sets the lower bound of FAR from data to be 0.5/yr, satisfying the threshold of online alerts. To enable the comparisons of significances between candidates beyond this threshold, a KNN kernel density estimation method [48] was used to extrapolate the significances of SPIIR detections. An empirical K value of 11 is chosen meaning the probability of data within the range specified by a bin is assigned to be the Gaussian smoothed average of probabilities of the nearest 11 histogram bins. The pipeline also collects two hours and one day of background events to capture potentially non-stationary noise behaviours in short-term and medium-term. In addition, the pipeline collects individual statistics, the single SNR and the single ξ_I^2, from the background events and apply the same FAR estimation method for single-detector FAR values to be used in the veto stage below. The pipeline requires at least one million background events before any FAR value assignment, corresponding to a collection time of a few hours, to ensure sufficient data points for reliable FAR estimations. Fig. 2 demonstrates how the signal triggers can be separated from the noise triggers using the calculated FARs from one-week of the O2 data.

E. Candidate veto and submission

A single GW signal could trigger several templates, a scenario avoided by using a clustering function at the last stage of the pipeline. Candidates are clustered based on their coherent network SNR values, within a time window, set to be 0.5 seconds for O3.

A few more tests were designed to veto possible transient glitches in O3, where O1 and O2 data were used to tune the veto thresholds.

- Background events were collected by three time scales of two-hour, one-day and one-week. The most conservative (maximum) FAR of the three time scales is assigned to the candidate to account for possible transient noise fluctuations in any of these three time scales.

- There were a few loud single-detector glitches in the O1 and O2 data that generated high significance for one detector but very low significance for other detectors. Therefore we set the FAR threshold for single detectors to be 0.5 Hz to reject such loud single-detector glitches while not affecting any detections of known events.

- We submit a subsequent trigger to a latest submission within the last 50 seconds only if its FAR is a factor of two less than the latest submission. This helped remove multiple submissions of a real signal or periodic transient glitches.

- In January 2020, two SPIIR triggers[5] associated with loud glitches were uploaded after being ranked with high significance despite extremely high ξ^2 values. As a result, an additional test was implemented to veto any triggers with a single $\xi_I^2 > 3$. This threshold was chosen based on trials using O1 and O2 data, and we find that all signals previously detected by the pipeline are unaffected, including the GW170817 event despite a significant glitch in the L1 data close to the event time. This veto remained effective at reducing glitch-based triggers for the remainder of O3.

[5] https://gracedb.ligo.org/superevents/S200106au/view/, https://gracedb.ligo.org/superevents/S200106av/view/

If a candidate passes all tests, it is uploaded to GraceDB with a table of the network SNR, the coherent SNR, individual SNR, ξ_I^2 statistics and mass and spin parameters as well as the time of the merger. For each significant candidate where SNR is over 7, the skymap that plots coherent SNRs computed for each searched sky direction is generated.

III. PIPELINE PERFORMANCE AND O3 ONLINE RUN

A. Data, pipeline, and injection setup

Here we show the performance of the pipeline using the O2 open data [49]. Category 1 and 2 (CAT1, CAT2) flags were used to flag periods of poor quality. Compared to the online data, open data is of better quality as it has been further processed with glitch removal and calibrations. The CAT1 and CAT2 data quality flags benefited from post-run measurements of noise-witness channels and provided more data quality information that are not necessarily available online [49, 50]. The data used here span from 1186248818 GPS seconds (Aug 13, 2017 at 02:00:00 UTC) to 1187312718 GPS seconds (Aug 21, 2017, 05:00:00 UTC). The total data duration is 687900 seconds, i.e. 7.96 days. For this period of time, the two LIGO detectors were mostly in duty with a joint duty cycle of about 70%. For about 25% of the time one of the LIGO detectors was also joined by the Virgo detector.

The CBC templates used during the SPIIR O3 run were obtained from [51]. The original templates cover binary neutron star (BNS), neutron star and black hole (NSBH), and intermediate-mass binary black hole (BBH) systems where the total masses are between 2 $M_\odot$ and 400 $M_\odot$ and the mass ratios are between 1.0 and 98.0. The spin parameter of the system is set to be the non-negative aligned spin on the z-component. For component mass less than 2 $M_\odot$, the spin is set within ± 0.05. Otherwise, it is set to be within ± 0.999. For the SPIIR O3 run, the templates were down selected by restricting the component mass to be over $1.1 M_\odot$ as a NS mass less than that is unlikely from estimation [52]. We further constrained the upper bound of the component mass to be less than 100 $M_\odot$ due to computing resource consideration. This gave 412,000 templates which fit into 103 computing machines on the LIGO-Caltech cluster. Each computing machine is equipped with a Quad-Core AMD Opteron(tm) 2376 CPU and a Nvidia GTX 1050Ti GPU.

Two injection sets were used to test the detection performance of the pipeline, the BNS injection set and the BBH injection set. We expect the pipeline detection performance of NSBH injections to be between the performances of the two injection sets here. Injections were placed every 1000 seconds in the O2 data. The BNS injection parameters were sampled as follows: a uniform distribution for component masses with the range of 1.1 $M_\odot$ to 2.3 $M_\odot$, an isotropic distribution up to 0.4 for spin, a uniform distribution for sky directions, and a uniform volume distribution and up to redshift 0.2 for distance. The BBH injection parameters are drawn from the same distributions for sky directions and distances respectively except the redshift range was increased to 0.7. The primary mass for the BBH injection was drawn from Salpeter IMF distribution between 5 $M_\odot$ and 50 $M_\odot$ and the secondary mass was drawn uniformly between 5 $M_\odot$ and the primary mass. The BNS injection set was used to plot the detection triggers in Fig 2.

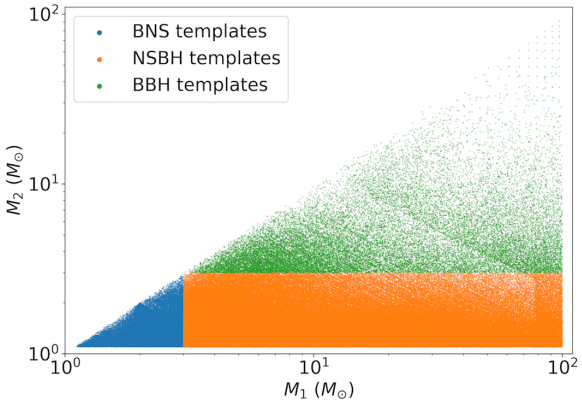


FIG. 3. Component mass of templates used for the O3 SPIIR online search. The templates are divided into BNS, NSBH, and BBH categories for performance demonstration using BNS and BBH injections. The BNS templates are chosen so that both component masses are less than 3 $M_\odot$. The boundary for the BBH category is that both component masses are more than 3 $M_\odot$.

In order to demonstrate the pipeline's performance with injections, the templates were divided into three categories to detect BNS, NSBH and BBH injections respectively (see Fig. 3). We can place as many SPIIR filters for high approximation to the original template. However, in practice, we limit the number of SPIIR filters to be no more than 350 as a tradeoff between the filtering computation cost and the approximation accuracy. Fig. 4 shows that the majority of overlaps from SPIIR filters are larger than 97%, meaning that SNR loss would be less than 3% from the SPIIR filtering. This is comparable to the SNR loss from template placement which is 3%. In general the overlaps for heavier systems are higher than that of low-mass systems, e.g. BNS systems. The same number of filters are used to patch shorter signals from BBH systems, yielding higher overlaps. However there are around 5% of the NSBH and BBH systems with overlaps less than 97%. For comparison, only less than 0.1% of the BNS template overlaps are less than 97%. This is due to large asymmetry of some binary systems causing fast variations in signals that the limited number of filters are insufficient to capture the profile.

Median PSD of the O2 data segment was used for data whitening, template whitening, and SNR calcula-

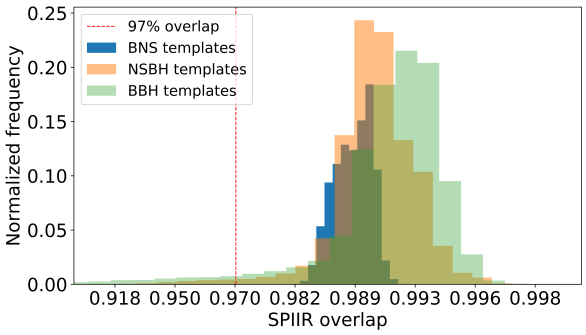


FIG. 4. Overlaps between the responses of SPIIR filters and the original templates for each BNS, NSBH, and BBH category. Overlap of 97% is marked by the red dashed line.

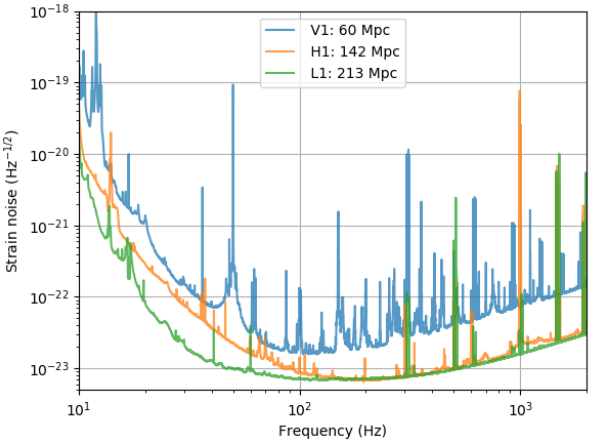


FIG. 5. Representative amplitude spectral density of the noise in H1, L1, and V1 detectors from Aug.13 2017 to Aug. 21 2017. The horizon distance is the maximum distance a detector can observe an optimally oriented binary source of 1.4 $M_\odot$ and 1.4 $M_\odot$ with the signal-to-noise ratio no less than 8.

tions of the injected signals. In reality, the detector noise is known to be non-stationary in very short time scales and the effect of using the median PSD than the true PSD can cause up to 5% discrepancy in detected SNRs for significant events [41]. The PSD fluctuation is expected to be larger in an online run where the pipeline needs to track the PSDs as mentioned in Sec. II A for whitening. Fig. 5 shows the amplitude of the noise spectral density from the median PSD. The figure also shows the BNS horizon distances computed from each PSD as the conventional method to represent the sensitivity of a single detector. The pipeline used the combined network SNR from H1 and L1 for the ranking statistic as explained in Sec. II D 1. The Virgo sensitivity is low compared to the LIGO detectors so that it is not used for ranking but for sky localization. To prepare background events for immediate significance estimation, the pipeline was run for a whole week before the injection started and ran uninterruptedly for the injection period.

B. Injection performance

The performance of the injection runs is demonstrated in scatter plots (Fig. 6) in terms of missed and found against effective distance (Eq. 3) and chirp mass. The chirp mass is a function of the individual masses and is the leading term in the inspiral phase evolution, expressed as:

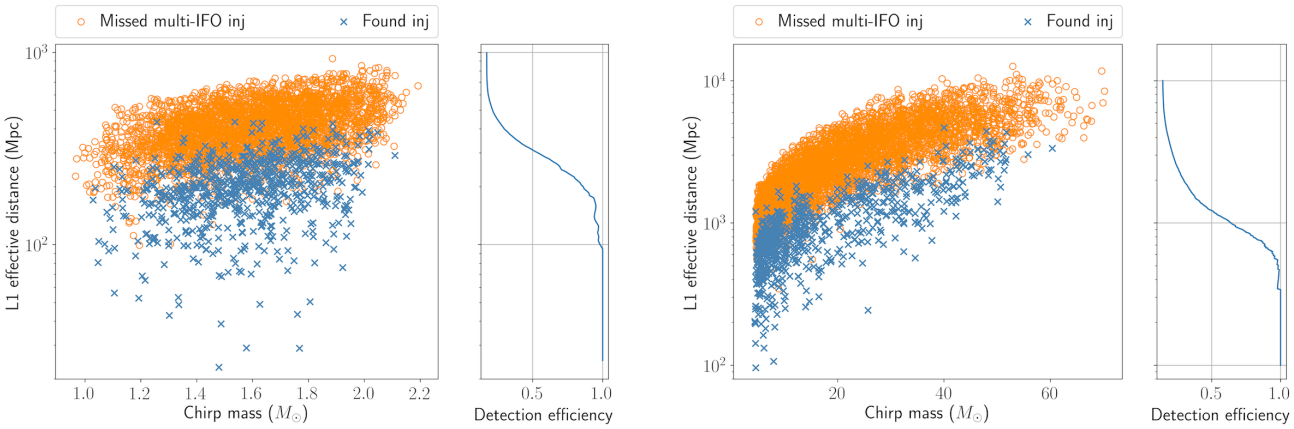
$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}, \tag{16}$$

where m_1 and m_2 are component masses. An injection is considered detected if the pipeline trigger is within ± 0.9 seconds of the injected time and the trigger significance is better than one per two months. The most sensitive detector, L1, has been used to demonstrate the performance in Fig. 6.

The pipeline can detect 100% BNS events at distances less than 100 Mpc and more than 50% of BNS events if the source distances are less than 300 Mpc. For BBH injections, the pipeline can detect 99% sources when the source distances are less than 500 Mpc and more than 50% of sources when the distances are less than 1 Gpc. This is consistent with the expectation that heavier systems which generate stronger GWs could be detected farther away. Two injections within 500 Mpc were missed by the pipeline because the locations and orientations of the sources were disfavored by the H1 detector resulting a SNR below the selection threshold of 4 and our O3 pipeline was not prepared to detect GWs from one detector only.

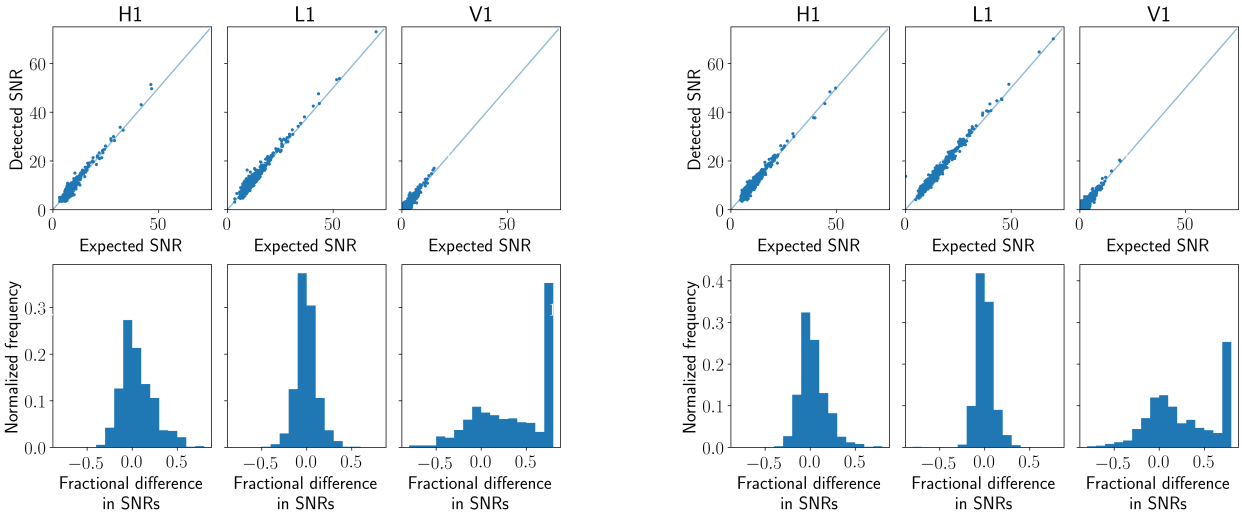
Fig. 7 shows the SNR recovery accuracy with H1, L1 and V1 detectors. The expected SNRs are computed from perfectly matching templates using the same median PSD. There is a 11% error on average in H1 SNR recovery and 7% error in L1 SNR recovery. Up to 5% SNR deviation is expected as we used a fixed PSD instead of a true PSD [41]. Other contributors to the error are the mismatch between the search template and the injection templates, and by the SPIIR approximation to the search templates. There is a large uncertainty in V1 SNR estimations due to the low sensitivity of Virgo in O2. A total of 40% of V1 SNRs are expected to be less than 2, meaning they are not detectable by V1. This caused large SNR discrepancies as shown in the last bin of V1 SNR error in the figure. The V1 SNR estimation should be improved by the sensitivity improvement in O3 and beyond.

Fig. 8 shows the accuracy of detected chirp mass for BNS and BBH injections. For BNS detections, the de-



(a) (b)

FIG. 6. (a) Missed (orange) and found (blue) BNS injections and the SPIIR detection efficiency in terms of the LIGO-Livingston (L1) effective distance and the binary chirp mass. (b) Missed and found BBH injections by the pipeline. Heavier systems that could generate stronger GWs can be detected farther away. The detection efficiency of NSBH sources are expected to between these two types of sources.



(a) (b)

FIG. 7. Top panel: expected SNRs vs. detected individual SNRs of (a) BNS detections, (b) BBH detections using O2 data. Bottom panel: fractional difference of detected SNRs and expected SNRs from (a) BNS detections, (b) BBH detections. SNR fractional differences are averagely around 11% for H1 detections and 7% for L1 detections. Around 40% of V1 detected SNRs are at least 50% off the expected SNRs as shown in the last bins of the histograms. Due to low sensitivity of Virgo in O2, most of the expected SNRs are less than 2 that the large uncertainty in SNR detection is expected.

tected chirp masses are within 0.4% of true values. The errors for BBH chirp mass detections are much larger in the high chirp mass region. This is partly due to that the BBH templates are placed more sparsely for high chirp masses and partly due to that the uncertainty is larger in high mass region[53].

C. O2 event search results

1. *GW170814, GW170817, and GW170818*

The pipeline used the entire bank set (BNS+NSBH+BBH) to search over the same chunk of data without injections. It reported two significant triggers corresponding to the known events GW170817

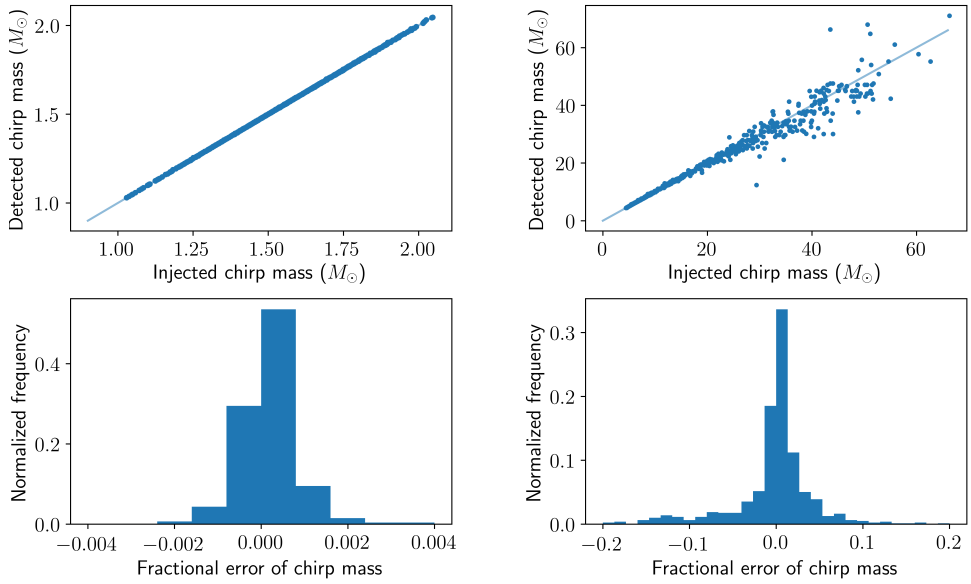


FIG. 8. Injected vs. detected chirp mass (top panel) and fractional error (bottom panel) of (a) BNS detections, (b) BBH detections. The errors for BBH chirp mass detections are larger in the high chirp mass region but are still within 20%.

and GW170814 as shown in Fig. 9. These two events were also reported online by other pipelines during the O2 run [5].

The 3rd event GW170818 within the same data chunk was only identified offline using GstLAL [5] and PyCBC searches [6]. There was a GstLAL online trigger matching this event but with a marginal significance [5]. SPIIR reported a trigger with 4.4×10^2/yr in this search. The significance of GW170818 SPIIR trigger was affected by the background contaminated by the precedent GW170814 detection. If we remove the GW170814 event trigger from the GW170818 background collection, the GW170818 trigger becomes significant at a FAR of 11/yr. In the future runs where detections are more frequent, the SPIIR online pipeline is planned to adopt strategies to remove influences of detections from background. The significance of GW170818 trigger is not affected by itself as the online pipeline only collects background from history.

Table I gives the individual, the network, the null SNRs and the significance for the three SPIIR detections. The network SNRs are consistent with that of the GstLAL (within 1%) and the PyCBC pipelines (within 1%) [5].The null SNRs are not used in the SPIIR pipeline for trigger ranking but are presented here for comparison to the expected null SNR of 2 from Gaussian noise assumption. The null SNR information is not explored by the O3 pipeline as the O3 detector network is equivalently a 2-detector network given the low-sensitivity of Virgo. This information is considered for future versions of the pipeline when there are at least three detectors with comparable high sensitivity.

The pipeline outputs a sky map of coherent SNR values for each significant candidate. Fig. 10 shows the coherent SNR skymap for the most significant detection during this search, the GW170817 SPIIR detection. The optical discovery of the event is highlighted and well captured by the high SNR area. The `Bayestar` program has been used to rapidly construct source sky localizations for CBC triggers during online runs [54]. The localization of GW170817 using the `Bayestar` program with the SPIIR detection is shown in Fig. 10. The 90% localization area from the SPIIR detection is 30 deg^2, consistent with the published 31 deg^2 using the `Bayestar` program with a PyCBC trigger[6]. Comparing the two sky maps in the figure, the coherent SNR map computes the network likelihood ratio optimized over four extrinsic parameters, while the `Bayestar` method takes prior information of all extrinsic parameters and calculates the marginalized posterior distribution.

D. Latency

The pipeline latency is defined as the time between receiving data and producing a trigger. The latency in gen-

[6] `https://dcc.ligo.org/LIGO-G1701985/public`

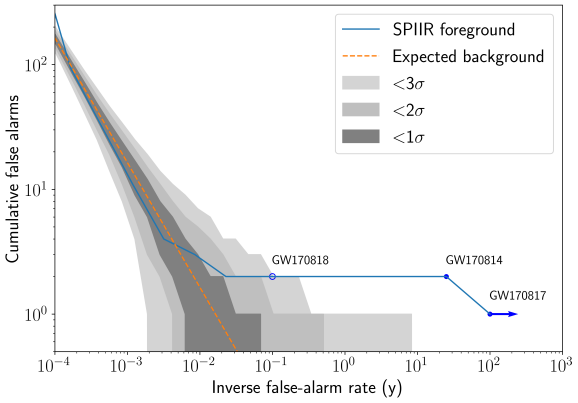


FIG. 9. Search results of H1, L1, and V1 data from Aug.13 to Aug.21, 2017 using the O3 SPIIR pipeline. Orange line is the expected number of false alarms given the threshold of inverse false alarm rate (IFAR). Shaded area shows the Gaussian standard deviations for the number of the false alarms. Blue curve is the number of SPIIR foreground candidates from this survey. The GW170818 detection has a significance of 11/yr if removing GW170814 influence from the background.

Event detection	$\rho_{\rm L}$	$\rho_{\rm H}$	$\rho_{\rm V}$	Network SNR	$\rho_{\rm NULL}$	FAR ($\rm y^{-1}$)
GW170814	13.8	8.9	3.8	16.9	0.3	0.04[a]
GW170817	24.4	19.4	3.7	31.5	1.3	$< 10^{-10}$[a]
GW170818	9.9	4.4	4.1	11.6	1.0	4.4×10^2 (11[b])

[a] 0.5/yr is our FAR limit from data, the significant FARs here are extrapolated using KNN for the purpose of comparison.
[b] If GW170814 influence is excluded from background.

TABLE I. Individual, network, null SNRs and significance of three SPIIR detections in the search of H1, L1, and V1 data from Aug.13 to Aug.21, 2017. The expected distribution of the null SNR square in Gaussian noise is a central χ^2 distribution with the degree of freedoms 2.

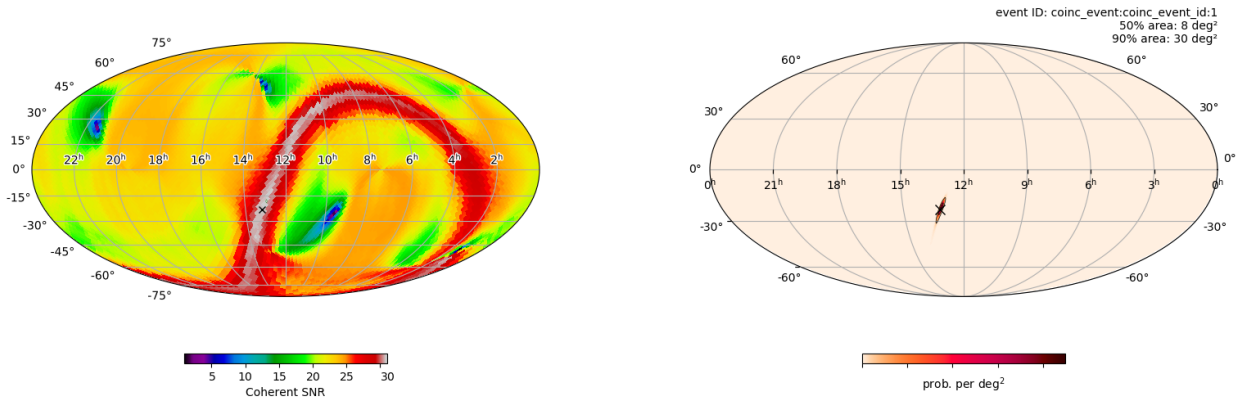
eral comes from two sources: the intrinsic delay to collect required size of data, and the time spent on computing. Tab. II provides a list of main latency sources and possible future improvement. $N_{\rm rate}$ is the data sample rate (2048 Hz for O3). At the beginning of the pipeline, the data is packed in one-second packets causing a waiting time of one second. The size of a data packet in theory can be reduced to as small as one sample. Next the downsampling module uses a FIR filter that introduces a waiting time equaling the filter length $N_{\rm FIR,D}$ which has a typical length of 192. The downsampling precision is proportional to the length of this filter. We assume this filter does not change for simplicity. In the next stage, segments of the data are collected and Fourier transformed into the frequency domain. The inverse of the noise PSD is then applied to this as a weight to perform the spectral whitening. The whitening latency is dependent on the segment length which is set to two seconds in O3. A recent work proposed a time-domain whitening that can reduce the latency to be zero-second [55]. The next stage gating applies a 0.25 second of side window and we simply assume the same for future runs. In the coherent search stage, it calculates ξ_I^2 over a series of data samples that requires a collection of N_j data samples (typically 175). In the last stage of candidate clustering and veto process, a 0.5 second window is applied to cluster triggers. This could be removed by applying submission thresholds. Though data comes in one-second packets, they would be sliced and combined at different stages of the pipeline and this process introduces latencies. For example, the downsampling module needs to reserve a fraction of $N_{\rm FIR,D}/N_{\rm rate}$ seconds to process the data at different sampling rates, and then the downstream module working on integral seconds, needs to wait this extra time to synchronize and combine these data streams at different sampling rates. A latency of five-second is feasible by using the zero-latency whitening [55] and advanced computing hardware. The latency could be reduced to sub-second in an ideal scenario when all pipeline components reach the best latency solutions.

To measure the latency of the pipeline in reality, we ran the pipeline over one-day stream replay of the O2 online data. We compared the pipeline latency on two computing platforms. As the CBC search is computationally intensive and most of the computation has been carried out on GPUs, we listed the GPU and its host CPU for each platform. Platform 1 is equipped with one Quad-Core AMD Opteron(tm) 2376 CPU and one Nvidia GTX 1050Ti GPU. Platform 1 was used by the SPIIR pipeline in O3. Platform 2 is equipped with a more advanced CPU, the Intel(R) Xeon(R) E5-2630v4 @ 2.20GHz and a more advanced GPU, the Nvidia Ti-tan V100. Just by using advanced hardware from Platform 2, the latency has been improved by nearly one second as shown by Fig. 11. At the time of writing only the PyCBCLive pipeline published its O3 latency which is in the range of 11 to 15 seconds [20]. The median latency of the SPIIR pipeline is less than 9 seconds.

E. O3 public alerts

During O3, 80 public alerts were issued from candidates reported by the five online pipelines [7]. The rate of glitches, predominantly at frequencies below 100 Hz, has increased significantly during O3 [8]. This resulted in 24 immediate retractions of submitted public alerts based on online investigation of noise association with the candidates. Two retractions were from the SPIIR pipeline on the same day of Jan. 6, 2020 due to high-amplitude scattered noise in the L1 detector on that day. It is expected the offline searches using cleaned data will find some alerts to be insignificant and in the meantime

[7] https://gracedb.ligo.org/superevents/public/O3/



(a) (b)

FIG. 10. (a) The all-sky map of coherent SNRs searched by the SPIIR pipeline for the GW170817 detection. (b) Rapid sky localization, `Bayestar`, using the SPIIR detection. The black cross marks the sky direction from the optical discovery [12].

.

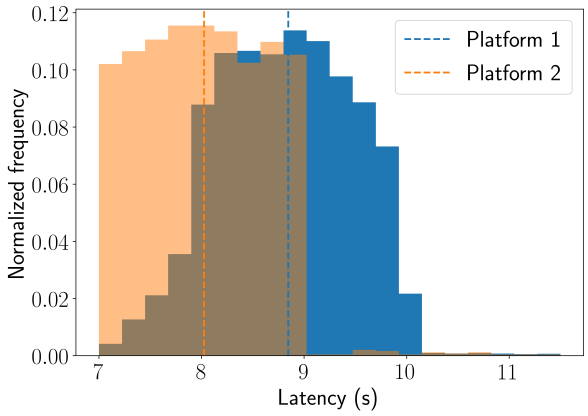


FIG. 11. Latency of the O3 SPIIR pipeline on two hardware platforms (see Sec. III D for description of the platforms). Dashed lines are median latencies. Platform 1 was used for the SPIIR pipeline for O3 online search.

Component	Current latency (s)	Future latency (s)
Data package	$1^{(1)}$	$N_{\rm PACK}/N_{\rm RATE}{}^{(1)}$
Downsampling	$N_{\rm FIR,D}/N_{\rm RATE}{}^{(1)}$	$N_{\rm FIR,D}/N_{\rm RATE}{}^{(1)}$
Whitening	$2^{(1)}$	0
Gating	$0.25^{(1)}$	$0.25^{(1)}$
SPIIR filtering	$< 1^{(2)}$	$< 1^{(2)}$
Coherent search	$< 1^{(2)} + N_j/N_{\rm RATE}{}^{(1)}$	$< 1^{(2)} + N_j/N_{\rm RATE}{}^{(1)}$
Trigger generation	$< 1^{(2)} + 0.5^{(1)}$	$< 1^{(2)}$
Data irregular slicing	$< 1^{(1)}$	~ 0
Computing time from components other than afore mentioned	$< 1^{(2)}$	~ 0
Overall	< 9	< 5

TABLE II. Breakdown of the pipeline latency. $^{(1)}$ marks the intrinsic delay associated with data collection and $^{(2)}$ marks the latency associated with computing time. $N_{\rm rate}$ is the data sample rate. $N_{\rm PACK}$ is the number of samples for each data packet. $N_{\rm FIR,D}$ is the filter length for downsampling. N_j is the number of data samples used for ξ^2 calculation in the coherent search.

rediscover some sub-threshold online triggers with improved significance. Of the 56 online alerts, the SPIIR pipeline registered 37 of them as published in LIGO-Virgo GCN notices [8] and one online trigger associated with the GW190814 event [56]. 37 of them were reported with other pipelines and one alert was reported only by the SPIIR pipeline (S190910d) which is proved no longer significant by the offline deep searches [8].

[8] `https://gcn.gsfc.nasa.gov/lvc_events.html`

IV. CONCLUSION AND DISCUSSION

This paper presents a low-latency pipeline used in the third LIGO-Virgo observation (O3) run, the SPIIR pipeline. It uses the efficient time-domain GPU-accelerated SPIIR method to perform time-domain matched filtering covering templates from binary neutron stars to intermediate binary black holes. It is the first low-latency CBC pipeline that employs the coherent search method with the help of GPU acceleration. The median latency of the pipeline for O3 is less than nine

seconds which is the fastest among online pipelines. It has the potential to be reduced to be below five seconds.

For the next coming run of O4, it will use data from new detector KAGRA and explore new algorithms to utilize the null-SNR information. Though the pipeline is using the novel coherent statistic for multi-detector detections, the single-detector significance estimation has been implemented for vetoing and can be adapted to single detector triggering. This can help increase the number of detections up to 30% [8]. The early-warning configuration of the pipeline to detect CBC events before the merger has been tested recently [57]. Given that the detector sensitivity is expected to improve significantly in this decade, early-warning detections have the potential to lead to major breakthroughs in the near future.

Additional information of the pipeline document and code can be found in https://git.ligo.org/lscsoft/spiir/.

ACKNOWLEDGMENTS

This work was funded by the Australian Research Council (ARC) Centre of Excellence for Gravitational Wave Discovery OzGrav under grant CE170100004. KK is partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1C1C1005863). TGFL was partially supported by grants from the Research Grants Council of the Hong Kong (Project No. CUHK14306419 and CUHK14306218), Research Committee of the Chinese University of Hong Kong and the Croucher Foundation of Hong Kong. A.Sengupta thanks the Department of Science and Technology for their ICPS cluster grant DST/ICPS/Cluster/Data_Science/2018/General/T-150. We wish to acknowledge Tom Almeida, Andrew Munt, Zhaohong Peng, Han-Shiang Kuo, Fengli Lin, Guo Chin Liu for helpful discussions on the improvement of the work.

This work used the computer resources of the LIGO CIT (Caltech) computer cluster and OzStar computer cluster at Swinburne University of Technology. LIGO CIT cluster is funded by National Science Foundation Grants PHY-0757058 and PHY-0823459. The OzSTAR program receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government. We wish to thank Stuart Anderson, Jarrod Hurley for the great help to use the clusters. This research used data obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org), a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. 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. This research used the injection sets generated by the rates and population group of the LIGO Scientific Collaboration. We acknowledge the GstLAL Team for the GstLAL library for several modules used in this work.

V. APPENDIXES

A. Network log likelihood ratio

A GW signal is a function of two sets of parameters, the intrinsic parameters ($\mathbf{\Theta}$) including the masses and spins of the components, and extrinsic parameters including the sky location (right ascension α and declination δ), the coalescence time t_c, the luminosity distance l, the polarization angle ψ, the inclination angle ι, and the GW coalescence phase ϕ_c. Previous work have shown that four extrinsic waveform parameters (l, ψ, ι, and ϕ_c) can be analytically maximized for the network log likelihood ratio (LLR) statistic [32–35] which leaves a reduced set of parameters for the network LLR representation. Here we simplify the reduced LLR using the SVD technique and show that the SNRs from matched filtering can be directly used to construct the network LLR statistic which is referred to the coherent SNR throughout the paper.

We first show the GW signal expression with an interferometric detector I:

$$h_I(t) = F_I^+(t)h^+(t) + F_I^\times(t)h^\times(t), \tag{17}$$

where $h^{+,\times}$ are the two polarizations and $F^{+,\times}$ are beampattern functions describing the responses of a detector to the two polarizations.

By rearranging the extrinsic parameters, the signal can then be expressed with the direction-induced modulations $G^{+,\times}$ which is dependent on the source sky location α and δ, and the detector location and orientation $\boldsymbol{s}(t)$; the a_{jk} matrix pertaining to the source luminosity distance l, the polarization angle ψ, the inclination angle ι, and the GW coalescence phase ϕ_c; and the h_c and h_s waveforms which only depend on the intrinsic parameters and the coalescence time.

$$h_I(t;\mathbf{\Theta},\alpha,\delta,t_c,l,\psi,\iota,\phi_c,\boldsymbol{s}) = \begin{pmatrix} G_I^+(\alpha,\delta,\boldsymbol{s}(t)) & G_I^\times(\alpha,\delta,\boldsymbol{s}(t)) \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} h_c(t;\mathbf{\Theta},t_c) \\ h_s(t;\mathbf{\Theta},t_c) \end{pmatrix}. \tag{18}$$

The $G^{+,\times}$ expressions can be found in Eq.12 and Eq.13 of [27] (G^+ equivalent to $a(t)$ and $G^\times$ equivalent to $b(t)$ in

cited equations) or Eq.1.53 and Eq.1.54 in [58]. They are related to the beam-pattern functions by the polarization angle ψ:

$$\begin{pmatrix} F_I^+(t) \\ F_I^\times(t) \end{pmatrix} = \begin{pmatrix} \cos 2\psi & \sin 2\psi \\ -\sin 2\psi & \cos 2\psi \end{pmatrix} \begin{pmatrix} G_I^+(t) \\ G_I^\times(t) \end{pmatrix}.$$

The a_{jk} matrix can be expressed as:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \frac{1}{l} \begin{pmatrix} \cos 2\psi & \sin 2\psi \\ -\sin 2\psi & \cos 2\psi \end{pmatrix} \begin{pmatrix} \frac{1}{2}(1+\cos^2\iota) & 0 \\ 0 & \cos\iota \end{pmatrix} \begin{pmatrix} \cos\phi_c & \sin\phi_c \\ -\sin\phi_c & \cos\phi_c \end{pmatrix}. \tag{19}$$

The solutions to $\{l, \psi, \iota, \phi_c\}$ from a_{jk} can be found in [34]. h_c and h_s are in quadrature and the strength of each at a distance of 1 Mpc seen from a detector is defined by σ_I:

$$\begin{aligned} (h_c \mid h_s) &= 0, \\ \sigma_I^2 \equiv \tfrac{1}{1\mathrm{Mpc}}(h_c \mid h_c) &= \tfrac{1}{1\mathrm{Mpc}}(h_s \mid h_s). \end{aligned} \tag{20}$$

The operator $(\cdot \mid \cdot)$ is defined as:

$$(a \mid b) = 4\mathrm{Re}\int_0^\infty \frac{\tilde{a}(f)\tilde{b}^*(f)}{S_{n_I}(f)}\mathrm{d}f, \tag{21}$$

where $S_{n_I}(f)$ is the noise PSD in this detector.

The network LLR is the sum of single LLRs assuming the noises in individual detectors are independent:

$$\begin{aligned} \ln \mathcal{L}_{\mathrm{NW}} &= \sum_I (d_I|h_I) - \frac{1}{2}(h_I|h_I), \\ &= (\boldsymbol{d}^T|\boldsymbol{h}) - \frac{1}{2}(\boldsymbol{h}^T|\boldsymbol{h}), \end{aligned} \tag{22}$$

where the subscript NW stands for the network, $\boldsymbol{d} = (d_1, ..., d_{N_d})^T$ with N_d being the number of detectors. $\boldsymbol{h} = (h_1, ..., h_{N_d})^T$ depends on the detector location. The operator on the matrix is defined as:

$$(\boldsymbol{D} \mid \boldsymbol{B}) = \sum_{p=1}^{n} (D_{jp} \mid B_{pk})_p. \tag{23}$$

We group the a_{ij} into two entities $\boldsymbol{A}_c = (a_{11}, a_{21})^T$ and $\boldsymbol{A}_s = (a_{12}, a_{22})^T$ for expression convenience. $\boldsymbol{G}$ represents the modulation for each detector. The network LLR can be written as:

$$\ln \mathcal{L}_{\mathrm{NW}} = \left(\boldsymbol{d}^T \mid \boldsymbol{G}\boldsymbol{A}_c h_c + \boldsymbol{G}\boldsymbol{A}_s h_s\right) - \frac{1}{2}\left(\boldsymbol{A}_c^T\boldsymbol{G}^T h_c + \boldsymbol{A}_s^T\boldsymbol{G}^T h_s \mid \boldsymbol{G}\boldsymbol{A}_c h_c + \boldsymbol{G}\boldsymbol{A}_s h_s\right). \tag{24}$$

Maximizing LLR over a_{jk} is equivalent to maximization over $\boldsymbol{A}_c$ and $\boldsymbol{A}_s$ respectively. The solution is then:

$$\begin{aligned} \boldsymbol{A}_x\Big|_{x=\{c,s\}} &= \left(\boldsymbol{G}^T h_x \mid \boldsymbol{G}h_x\right)^{-1}\boldsymbol{G}^T(\boldsymbol{d}^T \mid h_x), \\ &= \frac{1}{1\mathrm{Mpc}}(\boldsymbol{G}_\sigma^T\boldsymbol{G}_\sigma)^{-1}\boldsymbol{G}_\sigma^T\left(\overline{\boldsymbol{H}}_x \mid \boldsymbol{d}^T\right), \end{aligned} \tag{25}$$

where

$$\overline{\boldsymbol{H}}_x\Big|_{x=\{c,s\}} = \begin{pmatrix} h_x/\sigma_1 & 0 & \dots & 0 \\ 0 & h_x/\sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & h_x/\sigma_{N_d} \end{pmatrix}. \tag{26}$$

$\boldsymbol{G}_\sigma$ is the noise-weighted modulation and its SVD has the form:

$$\boldsymbol{G}_\sigma = \begin{pmatrix} G_1^+\sigma_1 & G_1^\times\sigma_1 \\ G_2^+\sigma_2 & G_2^\times\sigma_2 \\ \vdots & \vdots \\ G_{N_d}^+\sigma_{N_d} & G_{N_d}^\times\sigma_{N_d} \end{pmatrix} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T, \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix}. \tag{27}$$

The SVD decompose $\boldsymbol{G}_\sigma$ into a $N_d \times N_d$ unitary matrix $\boldsymbol{U}$, a $N_d \times 2$ pseudo-diagonal matrix $\boldsymbol{\Lambda}$ with decreasing positive singular values, and the transpose of a 2×2 unitary matrix $\boldsymbol{V}$. This form can be used obtain the Moone-Penrose pseudo inverse $(\boldsymbol{G}^T h_x \mid \boldsymbol{G}h_x)^{-1}$ and simplify Eq. 25.

Substituting the solutions of $\boldsymbol{A}$ into the network LLR,

the maximized LLR can then be expressed as:

$$\ln \mathcal{L}_{\underset{\max\{\boldsymbol{a}_{jk}\}}{\mathrm{NW}}} = \frac{1}{1\mathrm{Mpc}} \sum_{x=\{c,s\}} \left(\overline{\boldsymbol{H}}_x \mid \boldsymbol{d}\right)^T \boldsymbol{U}\boldsymbol{I}\boldsymbol{U}^T \left(\overline{\boldsymbol{H}}_x \mid \boldsymbol{d}\right)$$

$$= \frac{1}{1\mathrm{Mpc}} \parallel \boldsymbol{I}\boldsymbol{U}^T \left(\overline{\boldsymbol{H}}_c + i\overline{\boldsymbol{H}}_s \mid \boldsymbol{d}\right) \parallel^2 \quad (28)$$

$$= \frac{1}{1\mathrm{Mpc}} \left\| \boldsymbol{I}\boldsymbol{U}^T \begin{pmatrix} z_1 & 0 & \dots & 0 \\ 0 & z_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_{N_d} \end{pmatrix} \right\|^2, \quad (29)$$

where $\| \cdot \|$ is the Euclidean norm, $\boldsymbol{I} = \mathrm{diag}\{1, 1, 0, ..0\}$, and z is the individual SNR from each detector. The maximization procedure can be thought of as a projection of all the signal components in the N_d streams onto the signal plane spanned by the two vectors from $\boldsymbol{U}$, with the noise contributions reduced from N_d Gaussian streams to two Gaussian streams. If the noise is Gaussian in each detector, then this statistic will obey a non-central χ^2 distribution with a 4 degrees of freedom.

For a detector network with more than two detectors, the null stream or the null statistic can then be expressed as:

$$\ln \mathcal{L}_{\mathrm{NULL}} = \frac{1}{1\mathrm{Mpc}} \left\| \boldsymbol{I}^\dagger \boldsymbol{U}^T \begin{pmatrix} z_1 & 0 & \dots & 0 \\ 0 & z_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_{N_d} \end{pmatrix} \right\|^2, \quad (30)$$

where $\boldsymbol{I}^\dagger = diag\{0, 0, 1, \dots, 1\}$. When noise is Gaussian in each detector this statistic follows a central χ^2 distribution with $(N_d \times 2 - 4)$ degrees of freedom.

[1] J. Aasi *et al.*, Classical and Quantum Gravity **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].

[2] F. Acernese *et al.* (VIRGO), Classical Quantum Gravity **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].

[3] M. Tse *et al.*, Phys. Rev. Lett. **123**, 231107 (2019).

[4] F. Acernese *et al.* (Virgo Collaboration), Phys. Rev. Lett. **123**, 231108 (2019).

[5] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. X **9**, 031040 (2019).

[6] A. H. Nitz, T. Dent, G. S. Davies, S. Kumar, C. D. Capano, I. Harry, S. Mozzon, L. Nuttall, A. Lundgren, and M. Tápai, The Astrophysical Journal **891**, 123 (2020).

[7] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, Phys. Rev. D **101**, 083030 (2020).

[8] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. X **11**, 021053 (2021).

[9] T. Akutsu *et al.*, arXiv e-prints (2020), arXiv:2005.05574 [physics.ins-det].

[10] B. Iyer *et al.*, "Ligo-india, proposal of the consortium for indian initiative in gravitational-wave observations (indigo)," https://dcc.ligo.org/ligo-M1100296/public (2011).

[11] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **119**, 161101 (2017).

[12] B. P. Abbott *et al.* (LIGO Scientific, Virgo, Fermi GBM, INTEGRAL, IceCube, AstroSat Cadmium Zinc Telluride Imager Team, IPN, Insight-Hxmt, ANTARES, Swift, AGILE Team, 1M2H Team, Dark Energy Camera GW-EM, DES, DLT40, GRAWITA, Fermi-LAT, ATCA, ASKAP, Las Cumbres Observatory Group, OzGrav, DWF (Deeper Wider Faster Program), AST3, CAASTRO, VINROUGE, MASTER, J-GEM, GROWTH, JAGWAR, CaltechNRAO, TTU-NRAO, NuSTAR, Pan-STARRS, MAXI Team, TZAC Consortium, KU, Nordic Optical Telescope, ePESSTO, GROND, Texas Tech University, SALT Group, TOROS, BOOTES, MWA, CALET, IKI-GW Follow-up, H.E.S.S., LOFAR, LWA, HAWC, Pierre Auger, ALMA, Euro VLBI Team, Pi of Sky, Chandra Team at McGill University, DFN, ATLAS Telescopes, High Time Resolution Universe Survey, RIMAS, RATIR, SKA South Africa/MeerKAT), Astrophys. J. Lett. **848**, L12 (2017), arXiv:1710.05833 [astro-ph.HE].

[13] S. Klimenko, G. Vedovato, M. Drago, F. Salemi, V. Tiwari, G. A. Prodi, C. Lazzaro, K. Ackley, S. Tiwari, C. F. Da Silva, and G. Mitselmakher, Phys. Rev. D **93**, 042004 (2016).

[14] C. Messick *et al.*, Phys. Rev. D **95**, 042001 (2017).

[15] S. Sachdev *et al.*, arXiv e-prints (2019), arXiv:1901.08580 [gr-qc].

[16] C. Hanna *et al.*, Phys. Rev. **D101**, 022003 (2020), arXiv:1901.02227 [gr-qc].

[17] T. Adams, D. Buskulic, V. Germain, G. Guidi, F. Marion, M. Montani, B. Mours, F. Piergiovanni, and G. Wang, Class. Quant. Grav. **33**, 175012 (2016), arXiv:1512.02864 [gr-qc].

[18] A. H. Nitz, T. Dent, T. Dal Canton, S. Fairhurst, and D. A. Brown, Astrophys. J. **849**, 118 (2017), arXiv:1705.01513 [gr-qc].

[19] A. H. Nitz, T. Dal Canton, D. Davis, and S. Reyes, Phys. Rev. D **98**, 024050 (2018), arXiv:1805.11174 [gr-qc].

[20] T. Dal Canton, A. H. Nitz, B. Gadre, G. S. Davies, V. Villa-Ortega, T. Dent, I. Harry, and L. Xiao, arXiv (2020), arXiv:2008.07494 [astro-ph.HE].

[21] J. Luan, S. Hooper, L. Wen, and Y. Chen, Phys. Rev. D **85**, 102002 (2012), arXiv:1108.3174 [gr-qc].

[22] S. Hooper, S. K. Chung, J. Luan, D. Blair, Y. Chen, and L. Wen, Phys. Rev. D **86**, 024012 (2012), arXiv:1108.3186 [gr-qc].

[23] D. McKenzie, *Using the SPIIR method for detection of gravitational waves from spinning neutron star binaries.*, Master's thesis, University of western australia (2014).

[24] Y. Liu, Z. Du, S. K. Chung, S. Hooper, D. Blair, and L. Wen, Classical and Quantum Gravity **29**, 235018 (2012).

[25] X. Guo, Q. Chu, S. K. Chung, Z. Du, L. Wen, and Y. Gu, Computer Physics Communications **231**, 62 (2018).

[26] X. Guo, Q. Chu, Z. Du, and L. Went, in *2018 26th European Signal Processing Conference (EUSIPCO)* (IEEE, 2018) pp. 2638–2642.

[27] P. Jaranowski, A. Królak, and B. F. Schutz, Phys. Rev.

D **58**, 063001 (1998), arXiv:gr-qc/9804014.
[28] A. Rogan and S. Bose, Classical and Quantum Gravity **21**, S1607 (2004).
[29] A. Krolak, M. Tinto, and M. Vallisneri, Phys. Rev. D **70**, 022003 (2004).
[30] N. J. Cornish and E. K. Porter, Class. Quant. Grav. **24**, 5729 (2007), arXiv:gr-qc/0612091.
[31] A. R. Williamson, C. Biwer, S. Fairhurst, I. W. Harry, E. Macdonald, D. Macleod, and V. Predoi, Phys. Rev. D **90**, 122004 (2014).
[32] I. W. Harry and S. Fairhurst, Phys. Rev. D **83**, 084002 (2011).
[33] A. Pai, S. Dhurandhar, and S. Bose, Phys. Rev. D **64**, 042004 (2001), arXiv:gr-qc/0009078.
[34] S. Bose, T. Dayanga, S. Ghosh, and D. Talukder, Classical and quantum gravity **28**, 134009 (2011).
[35] D. Macleod, I. W. Harry, and S. Fairhurst, Phys. Rev. D **93**, 064004 (2016).
[36] M. E. Normandin and S. D. Mohanty, Phys. Rev. D **101**, 082001 (2020).
[37] K. Cannon, C. Hanna, and D. Keppel, Phys. Rev. D **84**, 084003 (2011).
[38] I. S. Heng, Classical and Quantum Gravity **26**, 105005 (2009).
[39] C. Röver, M.-A. Bizouard, N. Christensen, H. Dimmelmeier, I. S. Heng, and R. Meyer, Phys. Rev. D **80**, 102004 (2009).
[40] L. Wen, Int. J. Mod. Phys. D **17**, 1095 (2008), arXiv:gr-qc/0702096.
[41] B. Zackay, T. Venumadhav, J. Roulet, L. Dai, and M. Zaldarriaga, "Detecting gravitational waves in data with non-gaussian noise," (2019), arXiv:1908.05644 [astro-ph.IM].
[42] A. Lundgren and R. O'Shaughnessy, Phys. Rev. D **89**, 044021 (2014), arXiv:1304.3332 [gr-qc].
[43] A. Bohé *et al.*, Phys. Rev. D **95**, 044028 (2017).
[44] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).
[45] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, Phys. Rev. D **85**, 122006 (2012).
[46] K. M. Gorski, E. Hivon, A. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, The Astrophysical Journal **622**, 759 (2005).
[47] B. Allen, Phys. Rev. D **71**, 062001 (2005).
[48] N. S. Altman, The American Statistician **46**, 175 (1992).
[49] R.Abbott *et al.*, SoftwareX **13**, 100658 (2021).
[50] D. Davis, T. Massinger, A. Lundgren, J. C. Driggers, A. L. Urban, and L. Nuttall, Classical and Quantum Gravity **36**, 055011 (2019).
[51] D. Mukherjee *et al.*, arXiv e-prints (2018), arXiv:1812.05121 [astro-ph.IM].
[52] F. Özel and P. Freire, Annual Review of Astronomy and Astrophysics **54**, 401 (2016).
[53] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).
[54] L. P. Singer and L. R. Price, Phys. Rev. D **93**, 024013 (2016), arXiv:1508.03634 [gr-qc].
[55] L. Tsukada, K. Cannon, C. Hanna, D. Keppel, D. Meacher, and C. Messick, Physical Review D **97** (2018).
[56] R. Abbott *et al.*, The Astrophysical Journal **896**, L44 (2020).
[57] R. Magee *et al.*, The Astrophysical Journal Letters **910**, L21 (2021).
[58] Q. Chu, *Low-latency detection and localization of gravitational waves from compact binary coalescences*, Ph.D. thesis, University of Western Australia (2017).