Resolution problem of galactic binaries in LISA data

[1]Shao-Dong Zhao[a,b], Xue-Hao Zhang[a,b,c], Soumya D. Mohanty[c,d], Yu-Xiao Liu[a,b]

a Lanzhou Center for Theoretical Physics, Key Laboratory of Theoretical Physics of Gansu Province, and Key Laboratory of Quantum Theory and Applications of MoE, Lanzhou University, Lanzhou, Gansu 730000, China,

b Institute of Theoretical Physics & Research Center of Gravitation, Lanzhou University, Lanzhou 730000, People's Republic of China,

c Morningside Center of Mathematics, Academy of Mathematics and System Science, Chinese Academy of Sciences, 55, Zhong Guan Cun Donglu, Beijing, China, 100190

d Dept. of Physics and Astronomy, University of Texas Rio Grande Valley, One West University Blvd., Brownsville, Texas 78520

1 Introduction

Gravitational wave astronomy has been well-established in the frequency band $\sim [10, 1000]$ Hz using ground-based interferometric detectors. LIGO and Virgo have detected nearly 200 gravitational wave signals since the GW150914 [1]. The fourth observing run (O4), which started in May 2023, has already identified 81 significant signal candidates by January 2024. The total number of detections is expected to surpass 200 by the end of the observing run in February 2025. This catalog includes a diverse range of events, from binary black hole mergers to neutron star-black hole mergers and even double neutron star mergers continuing to expand our understanding of these cosmic phenomena.

The $\sim [10^{-4}, 1]$ Hz band is the target for the space-borne detectors like Laser Interferometer Space Antenna (LISA) [2], Taiji [7] and TianQin [6], all scheduled for launch in the 2030s. Different from ground-based detectors, target signals in space-borne ones are mainly continuous GW signals, emitted by inspiral GW sources. With different components in the inspiral systems, the GW sources can be categorized into extreme-mass ratio inspiral, massive black hole binary, binary black hole, and compact galactic binaries (GBs). Besides, there are stochastic gravitational wave backgrounds originating from both cosmological phases and astrophysical sources.

The galactic compact binaries are the simplest signals in the LISA band, from now on we simply call frequency bands $\sim [10^{-4}, 1]$ Hz LISA band. That is because the components of these systems are stellar-mass compact objects like Neutron stars and white dwarfs, the low mass of components makes it easy to apply approximations for the waveform. But the systems are numerous, around 30 million systems are predicted by the current astrophysical model.

To evaluate the progress and compare data analysis methods, the LISA community has conducted multiple mock LISA data challenges over the past two decades. The latest iteration in this series is the most recent LISA data challenge [3]. LDC1-4 provides single realizations of time-delay interferometry combinations, containing Gaussian stationary instrumental noise added to the GW

[1]120220909001@lzu.edu.cn

signals from 30×10^6 GBs. A catalog of the parameters of the GBs is also provided, allowing the performance of a multisource resolution method to be quantified rigorously.

2 Data description

Time Delay Interferometry (TDI) [8] is a technique that will be employed in space-borne gravitational wave observatories to minimize the impact of noise in laser measurements. In the LISA case, data from different times and positions of spacecraft are combined to effectively compress noise that is not related to the GW signal. Many combinations of TDI are proposed to cancel different noises, like Michelson, Sagnac, Relay, etc. We use the A and E combinations in the Michelson combination, which have mutually independent instrumental noise, corresponding to the first generation of TDIs used in LDC1. (The T combination is dropped because the GW signal in it is highly attenuated.)

For a GW signal, a theoretical waveform is needed before we go further to TDI combination of the signal. We start with plus and cross polarizations in the transverse traceless (TT) gauge of a plane GW incident at the Solar System barycenter origin. For a compact galactic binary, the polarizations are well-modeled in the Solar System barycenter frame as linear chirps,

$$h_+(t) = \mathcal{A}\left(1+\cos^2\iota\right)\cos\Phi(t)\,, \tag{1}$$

$$h_\times(t) = -2\mathcal{A}\cos\iota\sin\Phi(t)\,, \tag{2}$$

$$\Phi(t) = \phi_0 + 2\pi f t + \pi \dot{f} t^2\,, \tag{3}$$

where $\mathcal{A}$ is the amplitude of the wave, ι is the inclination angle between the GB orbital angular momentum and the line of sight from the SSB origin to the GB, ϕ_0 is the initial phase, f is the frequency at the start of observations, and $\dot{f}$ is the secular frequency drift.

Then we need to project the GW waveform on the constellation arms, take the example of link $\vec{\mathbf{n}}_{12}$ from spacecraft 1 to spacecraft 2 :

$$H_{12}(t) = h_+(t)\times\xi_+(\hat{\mathbf{u}},\hat{\mathbf{v}},\hat{\mathbf{n}}_{12}) + h_\times(t)\times\xi_\times(\hat{\mathbf{u}},\hat{\mathbf{v}},\hat{\mathbf{n}}_{12}), \tag{4}$$

where $\xi_\times$ and ξ_+ are the *antenna pattern functions* given by

$$\xi_+(\hat{\mathbf{u}},\hat{\mathbf{v}},\hat{\mathbf{n}}_{12}) = (\hat{\mathbf{u}}\cdot\hat{\mathbf{n}}_{12})^2 - (\hat{\mathbf{v}}\cdot\hat{\mathbf{n}}_{12})^2\,, \tag{5}$$

$$\xi_\times(\hat{\mathbf{u}},\hat{\mathbf{v}},\hat{\mathbf{n}}_{12}) = 2(\hat{\mathbf{u}}\cdot\hat{\mathbf{n}}_{12})(\hat{\mathbf{v}}\cdot\hat{\mathbf{n}}_{12}), \tag{6}$$

After appropriate approximations, the relative frequency shift y_{12}^{slr} is

$$y_{12}(t_1) \approx \frac{1}{2\left(1-\hat{\mathbf{k}}\cdot\hat{\mathbf{n}}_{12}(t_1)\right)}\left[H_{12}\left(t_1-\frac{L_{12}(t_1)}{c}-\frac{\hat{\mathbf{k}}\cdot\mathbf{x}_2(t_1)}{c}\right)-H_{12}\left(t_1-\frac{\hat{\mathbf{k}}\cdot\mathbf{x}_1(t_1)}{c}\right)\right]. \tag{7}$$

Then we can get to the first generation TDI X channel data s_1^X by multiplying time delay operators $\mathbf{D}_{ij}x(t) = x(t-L_{ij}(t))$ to the relative frequency shift:

$$s_1^X = y_{13} + \mathbf{D}_{13}y_{31} + \mathbf{D}_{131}y_{12} + \mathbf{D}_{1312}y_{21} - \left[y_{12} + \mathbf{D}_{12}y_{21} + \mathbf{D}_{121}y_{13} + \mathbf{D}_{1213}y_{31}\right]. \tag{8}$$

then Y and Z channel data can be obtained by permuting the subscript for the spacecraft.

The TDI time series $\overline{y}_D^I$ for combination channel I is given by

$$\overline{y}^I = \sum_{k=1}^{N_s}\overline{s}^I(\theta_k) + \overline{n}^I. \tag{9}$$

where $\overline{x} \in \mathbb{R}^N$ denotes a row vector, $\overline{s}^I(\theta)$ denotes a single GB signal corresponding to the source parameters θ, $\overline{n}^I$ is a realization of the instrumental noise, and N_s is the number of GBs, in LDC1-4 the number here is 29857650. For a GB, θ consists of $\{\mathcal{A}, \phi_0, \iota, \psi, \lambda, \beta, f, \dot{f}\}$, where $\mathcal{A}$, ϕ_0, ι, f, and $\dot{f}$ are defined following Eqs. (1) to (3), ψ is the polarization angle defining the orientation of the binary orbit projected on the sky, and the longitude and the latitude of the source in the Solar System barycenter frame are denoted by λ and β, respectively. In the rest of the paper, θ will serve as a stand-in for a GB source itself where convenient. Following LDC1-4 we set the sampling frequency for the uniformly sampled time series $\overline{y}^I_D$, for all I and D, to be $f_s = 1/15$ Hz, and the number of samples to $N = 4194304$ corresponding to an observation period $T_{\rm obs} \approx 2$ yr.

3 GBSIEVER

`GBSIEVER` (Galactic Binary Separation by Iterative Extraction and Validation using Extended Range) is an iterative pipeline where the parameters for one single source are determined at one time, and the estimated signal is then removed from the data. With the premise that the data comprises a single source superimposed on Gaussian, stationary noise, we construct a log-likelihood function and estimate the parameters of the signal using maximum likelihood estimation. In the GB parameter estimation case, $\mathcal{F}$-statistic can be helpful by reducing the search space from 8 waveform parameters to 4 intrinsic parameters and obtaining the left 4 extrinsic parameters by analytical derivation. In `GBSIEVER`, the former maximization is carried out using particle swarm optimization (PSO) [5].

When discussing our results, it is helpful to define several types of GB sources: (i) *True*: sources in the LDC1-4 catalog. (ii) *Reported*: the final list of estimated sources returned by `GBSIEVER`. (iii) *Identified*: the initial list of sources generated by the single-source search step in `GBSIEVER` before applying various selection criteria to form the set of reported sources. (iv) *Confirmed*: the reported sources that match true sources as determined by a prescribed metric for association. The fraction of confirmed sources in the set of reported ones is reffered the *detection rate*.

3.1 Single sources estimation

In the estimation of compact GB signals, $\mathcal{F}$-statistic is a commonly used simplification, we can write GW signal as

$$\bar{s}^I(\theta) = \bar{a}\mathbf{X}^I(\kappa). \tag{10}$$

where $\bar{a} = (a_1, a_2, a_3, a_4) \in \mathbb{R}^4$ is the *extrinsic* parameters obtained by reparameterizing $\mathcal{A}$, ϕ_0, ψ, and ι. $\mathbf{X}^I(\kappa)$ is *template* waveforms that depend on the *intrinsic* parameter set $\kappa = \{\lambda, \beta, f, \dot{f}\}$.

Then the estimator, $\hat{\theta}_M$, of the parameters θ of a single sources in iteration $M \geq 1$ is given by

$$\hat{\theta}_M = \underset{\theta}{\operatorname{argmin}} \sum_{I \in \mathcal{I}} \left(\|\overline{y}^I_M - \overline{s}^I(\theta)\|^I \right)^2, \tag{11}$$

where $\overline{y}^I_M = \overline{y}^I_{M-1} - \overline{s}^I(\theta_{M-1})$ and $\overline{y}^I_1 = \overline{y}^I$. Here $\|\cdot\|$ denotes the norm induced by the noise-weighted inner product

$$\langle \overline{x}, \overline{z} \rangle^I = \frac{1}{N f_s} (\widetilde{x}./\overline{S}^I)\widetilde{z}^\dagger, \tag{12}$$

where f_s is the sampling frequency, $\tilde{x}^T = \mathbf{F}\bar{x}^T$ is the discrete Fourier transform of time domain data $\overline{x}$.

Then we define two quantities to get $\mathcal{F}$-statistic

$$\mathbf{W}(i,j) = \sum_{I \in \mathcal{I}} \mathbf{W}^I(i,j) = \sum_{I \in \mathcal{I}} \langle \overline{X}^T(i), \overline{X}(j) \rangle^I, \mathbf{U}(i) = \sum_{I \in \mathcal{I}} \mathbf{U}^I(i) = \sum_{I \in \mathcal{I}} \langle \overline{y}^T, \overline{X}(i) \rangle^I, \tag{13}$$

the estimator of $\bar{a}$ in Eq. (10) analytically becomes

$$\hat{a}^T = \mathbf{W}^{-1}\mathbf{U}, \tag{14}$$

and with using $\mathcal{F}$-statistic the maximum likelihood estimator of κ is given as

$$\hat{\kappa} = \underset{\kappa}{\operatorname{argmax}}\ \mathcal{F}(\kappa) = \underset{\kappa}{\operatorname{argmax}}\ \mathbf{U}^T\mathbf{W}^{-1}\mathbf{U}. \tag{15}$$

To represent the signal strength with respect to the instrumental noise, we use the signal-to-noise ratio(SNR) defined below

$$\mathrm{SNR}^2 = \sum_{I\in\mathcal{I}} \left(\|\overline{s}^I(\theta)\|^I\right)^2. \tag{16}$$

In the postprocessing of the output of the multisource resolution, a metric is required to quantify the degree of association between a given pair of sources. We choose to use the correlation coefficient, $R(\theta,\theta')$, between the TDI signals corresponding to a given pair of sources:

$$R(\theta,\theta') = \frac{\sum_{I\in\mathcal{I}}\langle\overline{s}^I(\theta),\overline{s}^I(\theta')\rangle^I}{[\langle\overline{s}^I(\theta),\overline{s}^I(\theta)\rangle^I\langle\overline{s}^I(\theta'),\overline{s}^I(\theta')\rangle^I]^{1/2}}. \tag{17}$$

3.2 Narrow frequency band search and undersampling

In `GBSIEVER`, the search for individual sources is performed concurrently across several narrow frequency intervals, each spanning 0.02 mHz. However, only those sources found within the central 0.01 mHz—referred to as the acceptance zone—are considered valid. This restriction is designed to reduce the high rate of false positives near the edges of the bands, a problem that can arise when the energy from strong sources spreads into adjacent bands. The edges of each frequency interval are determined by applying a Tukey window to the discrete Fourier transform (DFT) of a TDI time series. The flat central part of this window is marginally wider, at 0.015 mHz, than the acceptance zone itself. Sources not accepted in one frequency band are still detected in neighboring bands due to the overlap, ensuring that the acceptance zones connect seamlessly.

Once the Tukey window is applied to a specific frequency band, an inverse DFT is used to revert the band-passed TDI data back into the time domain. Following this, a key procedure is the *undersampling* [4] of the time series. This process drastically reduces the sample count while preserving all critical information. Undersampling takes advantage of aliasing errors, which happen when sampling below the Nyquist rate, thereby shifting the data's informational content to lower frequencies. Since the $\mathcal{F}$-statistic for each frequency band is computed under the assumption of white noise, this method allows the inner product in Eq. (12) to be efficiently calculated in the time domain.

3.3 Cross-validation

Cross-validation plays a crucial role in `GBSIEVER`. The entire search for single sources is repeated on the same dataset, maintaining all settings except for extending the frequency derivative search range by a factor of a hundred. This process generates two sets of identified sources. The search with the narrow frequency derivative range is termed the *primary search*, while the one with the broader range is referred to as the *secondary search*. For each source $\hat{\theta}_{1,i}$ identified in the primary search, we calculate the correlation with sources found in adjacent frequency bands during the secondary search. An identified source $\hat{\theta}_{1,i}$ is classified as a reported source only if its correlation value $R_{\mathrm{ee}}(\hat{\theta}_{1,i})$ exceeds a predefined threshold. This cross-validation method is highly effective in

filtering out false positives. The predefined Ree threshold depends on both the frequency and the signal-to-noise ratio (SNR) of the identified source, as follows:

$$R_{\rm ee} = \begin{cases} 0.9, & \nu \in [0,3]\text{mHz,SNR} \leq 25 \\ 0.5, & \nu \in [0,3]\text{mHz,SNR} > 25 \end{cases}, R_{\rm ee} = \begin{cases} 0.9, & \nu \in [3,4]\text{mHz,SNR} \leq 20 \\ 0.5, & \nu \in [3,4]\text{mHz,SNR} > 20 \end{cases}. \tag{18}$$

For $f > 4$mHz, cross-validation is not required, that is because the number of spurious sources decreases rapidly.

4 Results

Until now, we have successfully implemented `GBSIEVER` in various scenarios, including the single LISA detector, the combined LISA-Taiji network detector, and configurations that incorporate astrophysical priors. In each of these applications, the algorithm has consistently performed at the forefront of the field, demonstrating state-of-the-art effectiveness and reliability.

4.1 Single LISA detector

In Table 1 we show the number of identified, reported, and confirmed sources in each block, which is divided according to SNR and frequency of the sources. And also the detection rate in each block and overall information. It can be seen in the low-SNR low-frequency region (block 1 and block 3), due to not only the high overlapping between the signals but also the faintness of the signals, the detection rate is low. The other results like subtraction residual and parameter estimation will be shown when comparing the single LISA detector results with improved settings.

	$f = [0,3]$		$f = [3,4]$		$f = [4,15]$
	SNR $[0,25]$	SNR $[25,\infty]$	SNR $[0,20]$	SNR $[20,\infty]$	SNR $[10,\infty]$
$R_{\rm ee}$	0.9	0.5	0.9	0.5	-1
Identified	23231	2106	3696	1526	4279
Reported	2767	2073	1622	1510	4279
Confirmed	1760	1892	1303	1394	4039
Detection rate	63.61%	91.27%	80.33%	92.32%	94.39%

Table 1: Performance of `GBSIEVER` for the single-detector LDC1-4 data. The total number of reported sources is 12251, and the total number of confirmed sources is 10388, and the total detection rate is 84.79%.

4.2 LISA-Taiji network

We extended our original `GBSIEVER` from the single LISA detector to network detectors by extending the definition of $\mathcal{F}$-statistic, SNR, and also generated the TDI data of 30 million GBs for the Taiji detector [9]. By extending the single LISA detector to the LISA-Taiji network detector case, we get the total number of confirmed GBs increase from 10388 to 18151, a significant improvement of 74.73%.

4.3 Astrophysical prior

In the LDCs GB catalog, there are two kinds of sources, ones with mass transfer between two components of the system called semi-detached, and another set of sources without mass transfer

	$f = [0, 3]$		$f = [3, 4]$		$f = [4, 15]$
	SNR $[0, 25]$	SNR $[25, \infty]$	SNR $[0, 20]$	SNR $[20, \infty]$	SNR $[10, \infty]$
R_{ee}	0.9	0.5	0.9	0.5	-1
Identified	33144	3941	3461	2528	4687
Reported	8420	3920	2440	2526	4687
Confirmed	5561	3605	1876	2406	4536
Detection rate	66.05%	91.96%	83.73%	95.25%	96.78%

Table 2: Performance of `GBSIEVER` for the LISA-Taiji network using LDC1-4 and Taiji-mod data. The total number of reported sources is 21993 and the total number of confirmed sources is 18151, and the total detection rate is 82.53%.

called detached. These two kinds of sources are both well-constrained in frequency derivative when we have clear information on frequency.

We here use a Tukey window $w(\dot{f})$ in frequency derivative with window position determined by frequency. Then the $\mathcal{F}$-statistic can be modified as

$$\mathcal{F}_{prior} = \mathcal{F}(\kappa) * w(\dot{f}). \tag{19}$$

We can see the effect of this Tukey window on the single parameter fitness in Fig. 1, where the big blue dot shows the location and fitness value of the original fitness function and the yellow star shows the true location and fitness value, and the red diamond is the location of max windowed fitness.

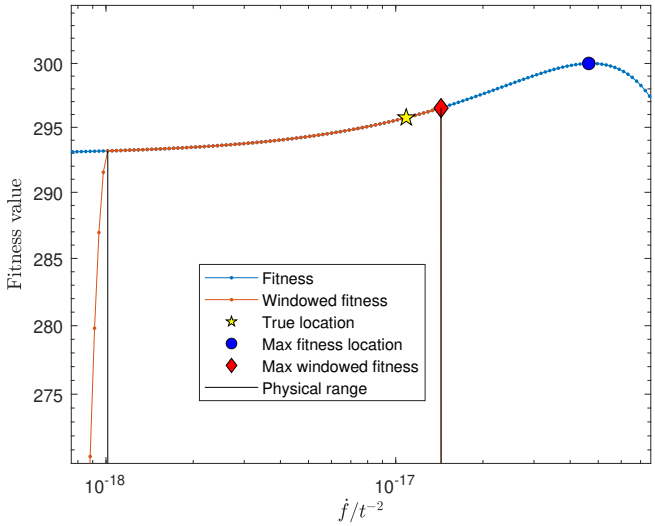

Figure 1: Illustration for the effect of frequency derivative prior on fitness function when fixing other parameters but the derivative. Here is the case with ($f \in [2.30, 2.32]$ mHz). The data is a superposition of instrumental noise, all weaker signals, and the signal itself.

We implement this astrophysical-equipped GBSIEVER pipeline into a simple dataset that contains 55 verification galactic binaries(VGBs) and all the neighboring sources (3084738 in total) distributed in the narrow frequency bins. As a result, we found 41 VGBs with 23 high correlations when the prior setting is activated and 43 VGBs with 22 high correlations when the prior setting is deactivated. With the help of the Fisher Information Matrix (FIM) we can get a theoretical

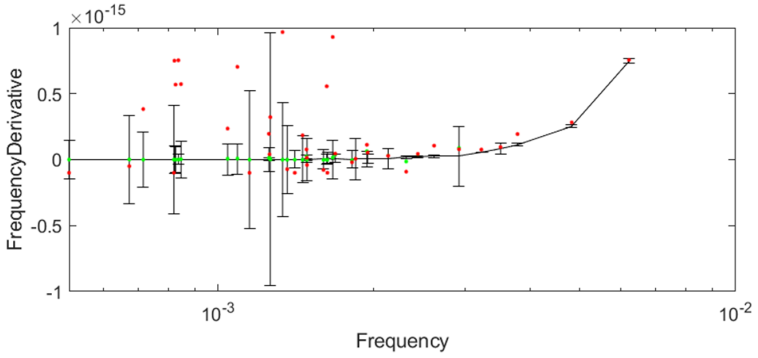

Figure 2: The scatter plot of 40 commonly identified VGBs with error bars gained from FIM value using true parameter values. The black line and the black points are true VGBs parameter values. The error bar is determined by the FIM $\mathbf{I}(\theta_{\rm VGBs})$ We can see clearly in the low f region, astrophysical prior can make a good constraint on the estimated value of frequency derivative.

minimal variance of different parameter estimation methods, defined as:

$$\Gamma_{ij} = \left(\frac{\partial s(t,\theta)}{\partial \theta_i}|\frac{\partial s(t,\theta)}{\partial \theta_j}\right) = \int_{-\infty}^{\infty} \frac{\frac{\partial s(t,\theta)}{\partial \theta_i}\frac{\partial s(t,\theta)}{\partial \theta_i}}{\int_{f_1}^{f_1+\Delta f} S_n(f)df} dt \tag{20}$$

Where $S_n(f)$ is the noise level PSD, and the limits of integration in the denominator is a small frequency region determine by the frequency value of the desired source.

But from Fig. 2, it's clear that with the help of the prior, we can get a better parameter estimation of frequency derivative than expected.

5 Conclusions

The algorithm `GBSIEVER` is one of the state-of-the-art data analysis pipelines for space-borne GW detectors aiming galactic binary. We successfully implement the pipeline with a single LISA detector and LISA-Taiji network detector for a complete dataset, also including the astrophysical prior in parameter estimation for a verification binaries dataset. The results of these analyses are all at the leading level in the context of the multisource GB resolution problem.

Acknowledgments

This work was supported by National Key Research and Development Program of China (Grant No. 2021YFC2203003), the National Natural Science Foundation of China (Grants No. 12247101), the 111 project under (Grant No. B20063). We gratefully acknowledge the use of high perfomance computers at the Supercomputing Cneter of Lanzhou University and State Key Laboratory of Scientific and Engineering Computing, CAS. The support provided by China Scholarship Council (CSC) during a visit to Valencia is acknowledged.

References

[1] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, et al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.*, 116:061102, Feb 2016. doi: 10.1103/PhysRevLett.116.061102. URL `https://link.aps.org/doi/10.1103/PhysRevLett.116.061102`.

[2] Pau Amaro-Seoane, Heather Audley, Stanislav Babak, John Baker, Enrico Barausse, Peter Bender, Emanuele Berti, Pierre Binetruy, Michael Born, Daniele Bortoluzzi, et al. Laser Interferometer Space Antenna. *arXiv preprint arXiv:1702.00786*, 2017.

[3] Quentin Baghi. The LISA Data Challenges. *arXiv preprint arXiv:2204.12142*, 2022.

[4] David L Donoho and Jared Tanner. Precise undersampling theorems. *Proceedings of the IEEE*, 98(6):913–924, 2010.

[5] Russell Eberhart and James Kennedy. Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks*, volume 4, pages 1942–1948. Citeseer, 1995.

[6] Jun Luo, Li-Sheng Chen, Hui-Zong Duan, Yun-Gui Gong, Shoucun Hu, Jianghui Ji, Qi Liu, Jianwei Mei, Vadim Milyukov, Mikhail Sazhin, et al. TianQin: a space-borne gravitational wave detector. *Classical and Quantum Gravity*, 33(3):035010, 2016.

[7] Wen-Hong Ruan, Zong-Kuan Guo, Rong-Gen Cai, and Yuan-Zhong Zhang. Taiji program: Gravitational-wave sources. *International Journal of Modern Physics A*, 35(17):2050075, 2020.

[8] Massimo Tinto and Sanjeev V Dhurandhar. Time-delay interferometry. *Living Reviews in Relativity*, 24(1):1–73, 2021.

[9] Xue-Hao Zhang, Shao-Dong Zhao, Soumya D Mohanty, and Yu-Xiao Liu. Resolving galactic binaries using a network of space-borne gravitational wave detectors. *Physical Review D*, 106 (10):102004, 2022.