Pulsar Timing Arrays

Luke Zoltan Kelley[a]
[a]University of California, Berkeley, Department of Astronomy, 501 Campbell Hall, Berkeley, CA 94720-3411

Abstract

Pulsar Timing Arrays (PTAs) have recently found strong evidence for low-frequency gravitational waves (GWs) in the nanohertz frequency regime. As GWs pass, they produce deviations in measured lengths and light-travel times. PTA experiments utilize the highly-consistent radio bursts from millisecond pulsars, distributed throughout the local galaxy, to identify miniscule timing deviations indicative of GWs. To distinguish GWs from noise, PTAs search for a particular correlation pattern between different pulsars called Hellings & Downs correlations. The type of GW signal that has recently been identified is a stochastic GW background (GWB), which is observed to have more power at lower GW frequencies. A GWB matching these observations has long been predicted from massive black holes (MBH) binaries. MBHs are known to exist in the centers of galaxies, which can then form binaries when two MBHs are brought together following the merger of galaxies. No example of an MBH binary has confidently been identified to date, and tremendous uncertainties about their formation and evolution remain. Alternative sources of the GWB have also been proposed, based on models for new fundamental physics, particularly in the early Universe. Improved sensitivity of PTAs will eventual lead to the characterization of GWB anisotropy and constraints on GWs from individual MBH binaries, either of which could definitively demonstrate the true origin of the GWB. If the source is MBH binaries, a variety of electromagnetic counterparts are possible, allowing for multimessenger astrophysics with low-frequency GWs.

Key Concepts

- Gravitational waves (GWs) are traveling perturbations in space-time that carry energy and momentum from sources, such as binaries, to distant observers. Binaries from massive black holes (MBHs) can produce very strong gravitational waves in the low-frequency (nanohertz) regime, both as 'continuous waves' (CWs) from individual binaries, and a stochastic 'gravitational wave background' (GWB) from an ensemble of large numbers of binaries.
- GWs can be detected by identifying correlations in the patterns of length of time changes, called strain, ($h = \Delta T/T = \Delta L/L$). Pulsar Timing Arrays (PTAs) examine correlations in the arrival times of radio bursts from millisecond pulsars distributed throughout the local galaxy. The characteristic correlation pattern for PTAs is the Hellings & Downs curve, which describes how similar the signals should be in any two different pulsars across the sky, based only on the angle between them.
- Extracting GW signals from PTA data requires careful modeling of many complex noise processes including motion of the pulsar and the solar system, and relativistic effects at each; dispersion and scattering from the interstellar medium; and noise/spin-evolution intrinsic to each pulsar. Careful statistical analyses which incorporate our noise models and the expected Hellings & Downs correlations are able to decipher GWB amplitudes as low as $\sim 10^{-15}$.
- Recently, PTAs including NANOGrav and many others across the globe, have identified strong evidence for a GWB signal consistent with the predictions from populations of MBH binaries (MBHBs). Additional data will soon confirm if the signal is truly from GWs; and, if so, determine whether or not the GWB is indeed produced by MBHBs or instead from new, 'beyond the Standard Model' physics. The 'GWB spectrum' (strain vs. frequency curve) can encode significant information about MBHB populations and evolution.
- MBH binaries are formed following the merger of galaxies, each of which are known to contain a central MBH with properties closely correlated with their host galaxy. Environmental interactions between binaries and their host galaxies are always required for binaries to reach the small separations at which nanohertz GWs are produced. MBHB evolution is highly uncertain, but dynamical friction and stellar scattering are always required, while circumbinary accretion disks and a number of other processes could also be important at times.
- In addition to GWs, MBHBs which are actively accreting can produce bright active galactic nucleus (AGN) emission across the electromagnetic (EM) spectrum. GW emitting binaries cannot typically be resolved in images, but their binary motion can be encoded in a variety of time- or frequency- dependent EM signatures. While many candidates have been identified, no MBHBs have been confirmed. Multi-messenger observations combining GWs and EM emission offer tremendous promise in understanding MBHB formation and evolution.

Glossary

Binary Hardening. 'Hardening' refers to the process of binaries gradually shrinking their separation due to the loss of orbital energy. For massive black-hole binaries, this can result from gravitational wave emission, or 'environmental interactions' between the binary and host galaxy. See Secs. 2.3.1 & 4.1.
Black Hole (BH) / Massive Black Hole (MBH) / Massive Black-Hole Binary (MBHB). BHs are the massive remnants of objects whose extreme gravitational force leads to indefinite collapse, and the formation of an 'event horizon' from which nothing can escape. MBHs have masses of $10^5 - 10^{10}$ $M_\odot$ and are observed to reside in the centers of galaxies. MBHBs can form following the merger of two galaxies, when the MBHs become gravitationally bound to each-other in the post-merger galaxy. See Secs. 1 & 4.
Burst With Memory. A GW signature produced at the final coalescence of binaries due to a near-discontinuity, or 'DC offset', in the space-time metric from before to after coalescence. See Sec. 2.3.4.
Chirp Mass. A quantity calculated from the two masses of a binary (Eq. 23) which directly determines the amplitude of binary GW signals, and the rate of binary inspiral due to GW emission.
Continuous Waves (CWs). Nearly monochromatic GWs produced by individual binaries, where the GW strain pattern is roughly

sinusoidal in time. Contrasted with a 'GWB'. See Sec. 2.3.
Detection Statistic. A quantity calculated from data which can be used as a quantitative measure of how confidently a signal can be detected in the data. See Sec. 3.3.
Environmental Interactions. Dynamical processes between an MBHB and its local galactic environment which serve to change (typically extract) orbital energy, leading to changes in the binary separation (typically shrinking it). See Sec. 4.1.
Final-Parsec Problem. The possible stalling of MBHB inspiral at parsec-scale separations, due to the depletion of stars that participate in stellar scattering, called 'loss-cone' stars. See Sec. 4.1.2.
Gravitational Wave (GW) a self-sustained perturbation of spacetime that carries energy and momentum, and produces measurable deviations in time and distance. See Sec. 2.1.
Gravitational Wave Background (GWB) / GWB Spectrum. The stochastic summation of GWs from large numbers of individual binaries each emitting at different frequencies. The GWB is typically described by a characteristic strain spectrum over frequencies $h_c(f)$, or a power spectral-density $S_h(f)$. See Sec. 2.3.3.
GW Strain / Characteristic Strain. Strain is the dimensionless fractional deviation in proper/measured distances and light-travel times, $h \approx \Delta L/L \approx \Delta T/T$, which is produced by passing GWs. Characteristic strain is a measure of strain that can be used in signal-to-noise ratio calculations (Eqs. 34 & 35). See Sec. 2.3.2.
Hellings & Downs (HD) correlations. The spatial correlation pattern between GW detectors at different angles, when the light-travel distance is comparable or longer than the GW wavelength, such as in PTAs (Eqs. 20).
Pulsar Timing Array (PTA). A collection of pulsars whose timing residuals are cross-correlated against each-other to search for correlated signatures such as GWs. See Sec. 3.
Times of Arrival (TOAs) / Timing Residuals. TOAs are the measured times at which radio pulses from pulsars are measured by an observer. Timing residuals are the difference between observed TOAs and the predictions/expectations from a timing model (Eqs. 48 & 61). See Secs. 3.1 & 3.3.
Timing Model. A model that predicts the pulse TOAs from a pulsar, typically accounting for a variety of 'noise' processes which can modify the TOAs. See Sec. 3.1 & 3.2.

Units & Constrants

$M_\odot$	Solar mass, unit of mass, equal to 1.9884×10^{30} kg
pc	Parsec, unit of distance, equal to 3.0857×10^{16} m
G	Newton's gravitational constant, 6.6743×10^{-8} m^3 kg^{-1}s^{-2}
c	Speed of light in a vacuum, equal to 2.998×10^8 m s^{-1}
H_0	Hubble constant at redshift zero, ≈ 70 km s^{-1} Mpc^{-1} $\approx 2 \times 10^{-18}$ s^{-1}.

1 Introduction and Overview

Einstein's theory of general relativity provides a purely-geometric description of spacetime and gravity, and a means for reconciling measurements made from different reference frames. After this theory developed, it was quickly realized that it implied two, nearly-incredible possibilities. First, that objects could exist which are so compact that no force is able to prevent their indefinite collapse, producing regions within which nothing can escape. Second, that the geometric structure allows for traveling perturbations in the shape of spacetime that cary energy to large distances. Evidence for the first objects, called **black holes**, grew consistently after the early 20th century, culminating in the first direct images of black holes in 2019 by the Event Horizon Telescope. Similarly, the second phenomenon, called **gravitational waves (GWs)**, were indirectly measured in the 1980s, and directly detected in 2015 by the LIGO-Virgo experiments.
Pulsar Timing Arrays (PTAs) are experiments to detect GWs in a new regime: at very-low frequencies. While LIGO-Virgo is sensitive to kilohertz GW frequencies, with thousands of oscillations per second, PTAs are sensitive to nanohertz (nHz) frequencies: with oscillations only every few years. The underlying principle behind detecting GWs is mostly the same between these two very-different frequency regimes. As GWs pass by the observer and their measurement devices, distances in spacetime are modulated in predictable patterns. By very precisely measuring deviations in light-travel times, and comparing these deviations between different places, those patterns of length variations can be measured and compared to GW predictions.

LIGO-Virgo, now with KAGRA, use laser beams running along two, perpendicular kilometers-long laser beams which are interfered together to produce a null signal in the absence of GWs. As GWs pass, the two laser beams traverse differing distances which leads to the interference becoming imperfect, and a signal is then measurable by photo-detectors. In these 'laser interferometers', noise sources (e.g. seismic vibrations) also produce imperfect interference. By comparing, or 'correlating', the signals between detectors at different locations across the Earth, noise can be filtered out, and subtle GW signals can be detected. The amplitude of GWs is typically characterized in terms of the GW **strain**: the fractional change in length over some baseline distance, or fractional changes in light-travel time, $h = \Delta L/L = \Delta T/T$ (Sec. 2). Terrestrial laser interferometers are able to achieve strain sensitivities on the order of $h \sim 10^{-24}$.

Instead of lasers, PTAs use **pulsars**: spinning neutron stars[1] which produce pulsed bursts of radio emission as their lighthouse-like beams of radiation pass by the observer (Earth) during each rotation. **Millisecond pulsars** in particular, which spin hundreds of times per second, can produce particularly reliable pulses. Changes in light-travel distance are measured as changes in the **times of arrival (TOAs)** of these pulses: either coming earlier or later than they should have in the absence of GWs (Sec. 3.1), shown schematically in the lower-panel of Fig. 1. A wide variety of noise sources can also produce timing deviations (changes in TOAs) that obscure GW signals, but again the specific patterns of timing delays can be correlated between multiple pulsars distributed across the sky/galaxy to distinguish GWs from noise. The GW frequencies at which PTAs are sensitive is determined by the timing characteristics of pulsar observations. Data has been collected for $10s$ yr, so the lowest accessible frequency (the 'Rayleigh' frequency) is $f_{\rm lo} \sim (10\,{\rm yr})^{-1} \sim$ nHz. The most sensitive frequencies for PTA are near the lowest bin, reaching strain sensitivities on the order of $h \sim 10^{-15}$. Because pulsar data is unevenly sampled and dominated by noise at high frequencies, the highest sensitive frequency is poorly defined. In practice, there are currently meaningful constraints up to $f_{\rm hi} \sim 100$ nHz.

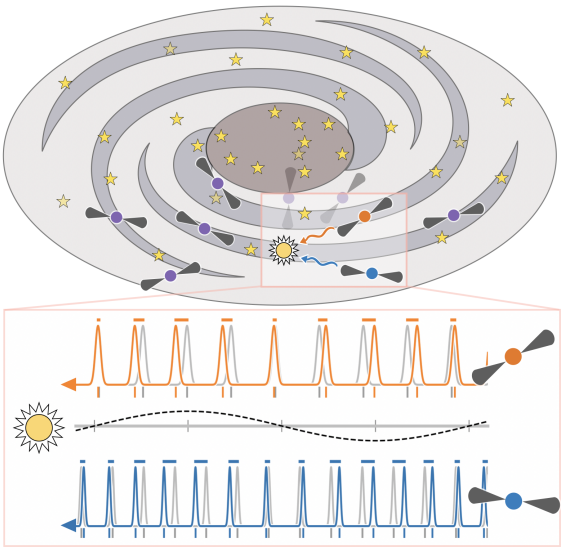

Fig. 1 Pulsar Timing Arrays are galaxy-scale GW detectors. Deviations in TOAs are correlated between pulsars to detect GWs. (Lower:) modeled pulses without GWs (*grey*) vs. measured pulses with GWs (*blue/orange*); vertical ticks show corresponding TOAs. Colored horizontal bars are the deviations/residuals proportional to GW amplitude (*black dashed*).

The 'smoking gun' for detecting GWs in PTAs is a particular pattern of correlations between different pulsars across the sky: the **Hellings & Downs (HD) correlations** (Sec. 2.2). The HD correlation predicts exactly the amount of correlation (or anti-correlation) between the timing delays in two pulsars, based only on their angular separation on the sky. Pulsars with small angular separations ($\gamma \lesssim 10s\,{\rm deg}$) will be strongly correlated (orange and blue pulsars in Fig. 1), those at nearly right angles ($\gamma \sim 90\,{\rm deg}$) will be anti-correlated, and pairs on opposite sides of the sky ($\gamma \sim 180\,{\rm deg}$) will again be strongly correlated.

Uncorrelated signals consistent with GWs were first measured by the North American Nanohertz Observatory for Gravitational waves (NANOGrav) PTA in 2019 (Arzoumanian et al., 2020). The predicted Hellings & Downs correlations were then measured with varying confidence by NANOGrav (Agazie et al., 2023), the Parkes PTA (PPTA; Reardon et al., 2023) centered in Australia, and the European PTA (EPTA) and Indian PTA (InPTA) (EPTA Collaboration et al., 2023). The most-recently published confidence levels correspond to a false-alarm probability of $10^{-3} - 10^{-4}$, corresponding to a Gaussian 'sigma' level of $3-4$; still below the '5σ level' commonly used for a definitive 'detection'. Unlike terrestrial interferometers, PTAs become more sensitive primarily by increasing their data volumes: through the number of pulsars being observed, and the number of TOA measurements from each pulsar[2]. Thus our confidence in whether or not the signal(s) being observed are definitively GWs will gradually improve over the next few years, and particularly through the combination of regional dataset into combined International PTA (IPTA) datasets (Antoniadis et al., 2022).

The idea of pulsar timing and kindred 'doppler tracking' were first formulated in the 1970s and 1980s (Estabrook and Wahlquist, 1975; Sazhin, 1978; Detweiler, 1979; Hellings and Downs, 1983). At the same time, GWs were first 'indirectly' measured by observing the inspiral of a binary pulsar—the Hulse-Taylor pulsar—which yielded the 1993 Nobel prize in physics to Hulse and Taylor. Gravitational waves are optimally produced by two, orbiting, massive, point-like objects—like neutron stars or black holes (Sec. 2.3). Those GWs carry away energy which has to come from the binary orbit, thus leading the orbit itself to contract until eventual coalescence. The presence of a pulsar in this double neutron-star system allowed the orbital period to be measured to incredible precision, such that the inspiral rate of $\approx 2.4 \times 10^{-12}$ s/s could be observed (Weisberg et al., 2010). GWs from dozens of stellar-mass binaries, containing NSs and black holes with masses of $\sim 10{\rm s} - 100\,{\rm M}_\odot$ have now been *directly* measured by LIGO-Virgo (Abbott et al., 2023). These kilohertz GWs are produced during the final handful of orbits before these binaries coalesce.

At roughly the same time as pulsar timing was being devised, and GWs indirectly measured, the scientific community was becoming confident in the existence of **massive black holes (MBHs)** in the centers of galaxies. These behemoth objects have masses of $\sim 10^6 - 10^{10}\,{\rm M}_\odot$, but event horizon radii of only $10^{-2} - 10^{+2}$ AU, i.e. only the size of solar systems[3]. Bright radiation had long been observed across the electromagnetic (EM) spectrum coming from the centers of galaxies: so called **active galactic nuclei (AGNs)**, with the brightest versions (quasars) observed from even the distant Universe. It was quickly realized that one of the only viable power sources for this emission is accretion onto a massive compact object. Careful measurements of the motion of stars in our own galactic nucleus (Sagittarius A-star) demonstrated that only an MBH could possess such a large mass in such a small space. This discovery yielded the 2020 Nobel Prize in Physics to Ghez, Genzel, and Penrose. In 2019, the MBH in the galaxy M87 was directly observed by the Event Horizon Telescope

[1] Neutron stars are the collapsed remnants of moderately-massive stars (initial stellar masses of very roughly $\sim 10 - 20\,{\rm M}_\odot$) after undergoing supernovae. These compact objects have masses of $\approx 2 - 3{\rm M}_\odot$, radii of $\approx 10km$, and often enormous magnetic fields ($\approx 10^{12} - 10^{14}$Gauss $= 10^8 - 10^{12}$Tesla).

[2] Higher sensitivity instruments with better noise characteristics, and the discovery of pulsars with low intrinsic-noise can also produce jumps in sensitivity, however significant data must still be collected to probe sufficiently low frequencies.

[3] The Astronomical Unit (AU) is the average distance from the Earth to the Sun, with $1\,{\rm AU} = 1.5 \times 10^{11}\,{\rm m} = 4.8 \times 10^{-6}\,{\rm pc}$.

(Event Horizon Telescope Collaboration et al., 2019), with the MBH in our own galaxy, Sagittarius $A*$, following in 2022 (Event Horizon Telescope Collaboration et al., 2022).

While the existence of MBHs was still being debated, it was proposed that pairs of MBHs could form binaries that also produce GWs (e.g. Sazhin, 1978; Begelman et al., 1980). These **MBH Binaries (MBHBs)** would be produced following the merger of galaxies, each containing an MBH (Sec. 4). Modeling of MBHB formation and evolution suggests that primarily the most-massive MBHBs in the Universe, $M \gtrsim 10^9\,\mathrm{M}_\odot$, are detectable by PTAs. Kepler's law allows us to relate the binary separation (a) to orbital frequency:

$$f_{\rm orb} = \frac{1}{2\pi}\left(\frac{GM}{a^3}\right)^{1/2}, \tag{1a}$$

$$\approx 19\,\mathrm{nHz}\left(\frac{M}{3\times10^9\,\mathrm{M}_\odot}\right)^{1/2}\left(\frac{a}{10^{-2}\,\mathrm{pc}}\right)^{3/2}. \tag{1b}$$

Thus MBHBs with total masses of $M \sim 10^9\,\mathrm{M}_\odot$ emit nHz GWs when separated by $\sim 10^{-2}$ pc. In this regime, detectable MBHBs are still orbiting in the Keplerian regime: only very slowly inspiralling such that they will persist for $\sim 10^6$ yr before finally coalescing. Despite these long lifetimes, the galaxies containing such massive MBHs are quite uncommon, and they only experience a handful of comparable-mass galaxy mergers in their lifetimes (Sec. 4). This makes detectable MBHBs exceedingly rare. A large number of **electromagnetic (EM)** surveys searching for signatures of MBHBs (Sec. 4.2) have been carried out; but, despite hundreds of candidates identified to date, not a single MBHB has been confirmed.

The recent detection of low-frequency GWs by PTAs, appears to be in the form of a **stochastic gravitational wave background (GWB)**: the superposition of GW signals from thousands-to-millions of individual MBHBs. This GWB provides the first strong evidence that MBHBs are actually able to form, produce GWs, and eventually coalesce. However, there are a number of alternative explanations for the GWB that do not involve MBH binaries at all: signals from so-called 'new' or 'Beyond the Standard Model (BSM)' physics in the very early Universe[4]. The binary model is currently strongly favored, but far from proven. At the very least, PTA observations still provide a new, independent constraint on models that attempt to describe fundamental properties of our Universe such as the nature of cosmic inflation, the origin of the baryon asymmetry (dominance of matter over anti-matter), or provide a quantum description of gravity. With more data and more sensitivity, PTAs will be able to definitively establish the origin of the GWB, and constrain the origin and evolution of MBH binaries and/or BSM physics.

Currently, PTAs are able to measure the basic shape of the **GWB spectrum**—the amplitude and slope of the GW-strain versus frequency curve. As PTA sensitivities improve, we will be able to characterize the full shape of the spectrum, including any low-frequency and/or high-frequency deviations from a constant slope, and stochastic variations from one frequency-bin to the next. We will also be able to observe how the GWB power is distributed over the sky: whether all of the energy is uniformly distributed (isotropic), or coming from some regions more than others (anisotropy), and eventually we will be able to distinguish individual loud binaries, called '**continuous wave** (CW)' sources, which are promising targets for EM counterparts.

There are tremendous theoretical uncertainties about nearly every aspect of MBHB formation and evolution, due to the lack of previous observational constraints. How are two MBHs able to find each-other and form binaries in turbulent, post-merger galaxies? Once large-separation binaries form, how are they able to reach the relatively small separations necessary for detectable GW emission? Are these binaries able to accrete gas like single MBHs, and what are the dynamics of that accretion? If and when accretion occurs, do MBHBs produce 'normal' AGN emission? What are the EM signatures unique to MBHBs as opposed to single MBHs? How do MBHBs effect their host galaxies, and what happens following their coalescence? These are the questions that we hope to answer using PTA observations of low-frequency GWs, and eventually their EM counterparts.

2 Gravitational Waves (GWs)

2.1 Waves in General Relativity

For additional details, we refer the reader to Misner et al. (1973).

The structure and shape of spacetime is described by the **metric tensor** $g_{\mu\nu}$, such that the invariant length element of spacetime can be calculated as,

$$ds^2 = -c^2 d\tau^2 = g_{\mu\nu}\,dx^\mu dx^\nu \equiv \sum_{\nu=0}^{3}\sum_{\mu=0}^{3} g_{\mu\nu}\,dx^\mu dx^\nu = dx_\nu dx^\nu. \tag{2}$$

The metric is a 4×4 symmetric tensor where each index runs over the four spacetime dimensions[5], for example $x_\mu = (t, x, y, z)$ could denote cartesian coordinates.

We will assume that we are always far from sources of strong gravity, such that we can express our spacetime metric as a perturbation ($h_{\mu\nu}$) to flat spacetime, i.e. $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$. Flat spacetime is typically described by the Minkowski metric $\eta_{\mu\nu}$, such that $\eta_{\mu\nu}\,dx^\mu\,dx^\nu =$

[4] While mostly outside the scope of this document, we briefly comment discuss non-standard models in Sec. 5

[5] The final two equalities above show: (i) Einstein summation notation, where repeated indices imply a summation, and (ii) that metrics are used to 'raise' and 'lower' indices, transforming between co-variant and contra-variant vector components.

$-dt^2 + dx^2 + dy^2 + dz^2$. We will only consider a *transverse traceless gauge* where $h_{\mu 0} = 0$, and $\partial_i h_{ij} = 0$. Einstein's equations determines the relationship between the structure and curvature of spacetime, and its mass/energy content. The linearized equations in a vacuum admit plane-wave solutions of the form[6],

$$h_{\mu\nu} = A_a \, \epsilon^a_{\mu\nu} \, \exp\Big(i \, k_\rho x^\rho\Big). \tag{3}$$

Here, the wave four-vector $k_\rho = (2\pi f/c, k_1, k_2, k_3)$ describes the propagation of a GW with frequency f traveling in a spatial direction with components[7] k_i. GWs possess two independent polarizations, here indexed by a, described by polarization tensors: $\epsilon^a_{\mu\nu}$. Let us define two unit vectors orthogonal to each other and to the wave-vector (i.e. GW propagation direction): $\hat{u}_\mu$, $\hat{v}_\mu$, such that $\hat{u}_\mu \, \hat{v}^\mu = \hat{u}_\mu \, \hat{k}^\mu = \hat{v}_\mu \, \hat{k}^\mu = 0$. We can then define linear polarization tensors as[8],

$$\epsilon^+_{ij} \equiv \hat{u}_i \hat{u}_j - \hat{v}_i \hat{v}_j, \tag{4a}$$

$$\epsilon^\times_{ij} \equiv \hat{u}_i \hat{v}_j + \hat{v}_i \hat{u}_j, \tag{4b}$$

each corresponding to an amplitude $A_a \in \{A_+, A_\times\}$. These waves follow null geodesics, such that $k_\mu \, k^\mu = \eta_{\mu\nu} \, k^\mu \, k^\nu = 0$, and thus $(2\pi f/c)^2 = k_i k^i$. In cartesian coordinates with the GW traveling along the $\hat{z}$ direction (i.e. $k_\mu = \{2\pi f/c, 0, 0, k_z\}$), we can write,

$$h_{\mu\nu}(t, z) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & A_+ & A_\times & 0 \\ 0 & A_\times & -A_+ & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \exp\left[2\pi i f (t - z/c)\right]. \tag{5}$$

The effects of each polarization are '**quadrupolar**', i.e. repeating every 180 deg, with the × polarization equal to the + rotated by 45 deg. The effects of the two polarizations are a ring of test-masses are shown in Fig. 2(c).

In general relativity, the 'geodesic equation' governs motion due to gravity: $\frac{\mathrm{d}^2 x^\alpha}{\mathrm{d}\tau^2} + \Gamma^\alpha_{\mu\nu} \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} \frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} = 0$, where $\Gamma^\alpha_{\mu\nu}$ are 'Christoffel symbols'—functions of the metric and its derivatives. Given a GW metric, in the transverse traceless gauge and linearized gravity, all of the Christoffel symbols vanish, and thus the geodesic equations become $\frac{\mathrm{d}^2 x^\alpha}{\mathrm{d}\tau^2} = 0$. This tells us that the *coordinate* motion of test masses remain constant before, during, and after GWs pass. The same is not true for the *proper distance* between two test masses. From the metric (Eq. 2), we can write for any spacetime interval:

$$s = \int_{x(t)} \left[g_{\mu\nu} \frac{\mathrm{d}x^\mu}{\mathrm{d}t} \frac{\mathrm{d}x^\nu}{\mathrm{d}t} \right]^{1/2} dt. \tag{6}$$

Consider test-mass A to be at coordinate position $x^A_i = 0$, and test-mass B to be at a coordinate position $x^B_i = L_c \hat{x}_1$. If we reuse the coordinate system with a GW traveling along the $\hat{x}_3$ direction (Eq. 5), and recall that proper distance L is the spacetime interval defined with $dx^0 = 0$, we can find that:

$$L = \int_0^{L_c} [1 + h_{11}(t)]^{1/2} dx^1 \approx \left(1 + \frac{1}{2} h_+ \, \exp[2\pi i f (t - z/c)]\right) L_c. \tag{7}$$

The approximation comes from a first-order Taylor expansion of the integrand. If we consider the 'strain'—the fractional proper-distance change—over a full wave cycle, we find $h \equiv \Delta L/L = h_+$. We can then identify the GW metric perturbation (h_{ij}) as a direct proxy for the local, fractional deformation, or the **gravitational wave strain**. The local deformations are thus quadrupolar, just like the waves themselves.

It becomes clear that the proper-distance is what's relevant for experiments[9] if we consider that we could identically calculate the light-travel time in A's reference frame, for a null geodesic traveling from A to B, by replacing the left-hand side of Eq. 6 with $c\Delta T$. We can thus equivalently identify GW strain as $h \equiv \Delta T/T$.

2.2 Timing Deviations and Spatial Correlations

For additional details, we refer the reader to Anholm et al. (2009) and Romano and Allen (2024). Below, we closely follow the latter.

The preceding calculation (Eq. 7) requires the GW to be spatially uniform along the direction of interest (i.e. $h_{ij}[x^1, t] = h_{ij}[x^1 = 0, t]$). In general this is not the case, and both the amplitude and frequency of the GW could be changing. Consider a pulsar located at a spatial position p^i, which is a total coordinate distance from the observer $L_c = \left(p_i \, p^i\right)^{1/2}$. The fully general solution, using light travel time, is:

$$c \, \Delta T(t) = \frac{1}{2} \hat{p}^i \hat{p}^j \int_0^{L_c} h_{ij}[\xi(s)] \, ds, \tag{8}$$

where the metric is integrated along a path specified by $\xi(s)$. The configuration is shown in Fig. 2(a). We will see below that the same GW

[6] It will be assumed that measurables correspond to the real part of complex quantities.

[7] We adopt the convention that Greek indices $(\mu, \nu, \ldots)$ imply the four space-time dimensions, while Latin indices $(i, j, k, \ldots)$ span only the three space dimensions.

[8] The reader may notice that $\hat{u}_\mu$ or $\hat{v}_\mu$ are not fully specified: they can be rotated about $\hat{k}$ by any angle. This degree of freedom is often specified by the polarization angle, usually denoted by ψ.

[9] Light-travel measurements determine delay times (e.g. pulsar/Doppler tracking) and equivalently phase shifts (e.g. laser interferometers), but also dynamical forces such as beads on a string or equivalently masses on a spring (e.g. resonant detectors).

effects can be considered more conveniently with respect to a 'redshift':

$$z(t) \equiv \frac{\mathrm{d}\Delta T}{\mathrm{d}t} = \frac{1}{2}\frac{\hat{p}^i\hat{p}^j}{c}\int_0^{L_c}\frac{\partial h_{ij}[\xi(s)]}{\partial t}\,ds. \tag{9}$$

Without a loss of generality, the path that light takes from the pulsar to the observer can be parameterized as, $\xi(s) = \left(t - \frac{1}{c}[L_c - s], [L_c - s]\hat{p}^i\right)$. For a plane wave, we can re-parameterize in terms of 'retarded coordinates', $h[\xi(t,s)] = h[u(t,s)]$, where $u(t,s) \equiv t(s) - \hat{k}^i\,\xi_i(s)/c = t - (L_c - s)\left(1 + \hat{k}^i\,\hat{p}_i\right)/c$. It is easy to evaluate Eq. 9 by expressing $\partial h_{ij}/\partial t$ in terms of $\partial h_{ij}/\partial s$; i.e. we want to find the coordinate-time rate of change of the metric strain in terms of the rate of change along the trajectory. If we assume that the frequency is constant along the trajectory, this is straightforward because h_{ij} only depends explicitly on the 'retarded coordinate' u, which depends linearly on t and s. Thus, we can write:

$$\frac{\partial h}{\partial s} = \frac{\mathrm{d}h}{\mathrm{d}u}\frac{\partial u}{\partial s} = \frac{1}{c}\left(1 + \hat{k}^i\hat{p}_i\right), \tag{10a}$$

$$\frac{\partial h}{\partial t} = \frac{\mathrm{d}h}{\mathrm{d}u}\frac{\partial u}{\partial t} = \frac{\mathrm{d}h}{\mathrm{d}u} = \frac{c}{\left(1 + \hat{k}^i\hat{p}_i\right)}\frac{\partial h}{\partial s}. \tag{10b}$$

Finally, plugging this into Eq. 9 we get:

$$z(t) = \frac{1}{2}\frac{\hat{p}^i\hat{p}^j}{1 + \hat{k}^i\hat{p}_i}\Big[h_{ij}(t,0) - h_{ij}(t - L_c/c, p)\Big]. \tag{11}$$

Thus, the redshift is simply the difference of metric strains between the pulsar (the *pulsar term*) and the observer (the *Earth term*)[10].

Equation 11 describes the response of timing measurements to a passing GW plane-wave, with constant frequency. Note the factor of $1 + \hat{k}^i\hat{p}_i$ which appears in the denominator, and also in the 'phase' of the GW strain. The situations in which the pulsar is aligned or anti-aligned with the source, $\hat{p} = \pm\hat{k}$, are worth particular consideration. When the pulsar is in the direction that the GW is coming from ($\hat{p} = -\hat{k}$), the redshift goes to zero despite divergence in the denominator, because the numerator is also zero. In a way, the numerator is zero for two separate reasons. First, there is no component of the GW strain along the observer-pulsar direction for this configurations. Second, the propagating photons are always 'surfing' the same phase of the GW (Anholm et al., 2009), so it necessarily arrives at the observer at the same GW phase as it left the pulsar. However, in the case that the source and pulsar are anti-aligned, $\hat{p} = +\hat{k}$, the phase of the GW changes as the photons propagate to the observer, and thus the metrics differ between the Earth and Pulsar. This nicely illustrates that the phase information is not symmetric to flipping the GW propagation direction. While the GWs are purely quadrupolar, the response of a detector is *not purely quadrupolar*. See Romano and Allen (2024) for an excellent discussion.

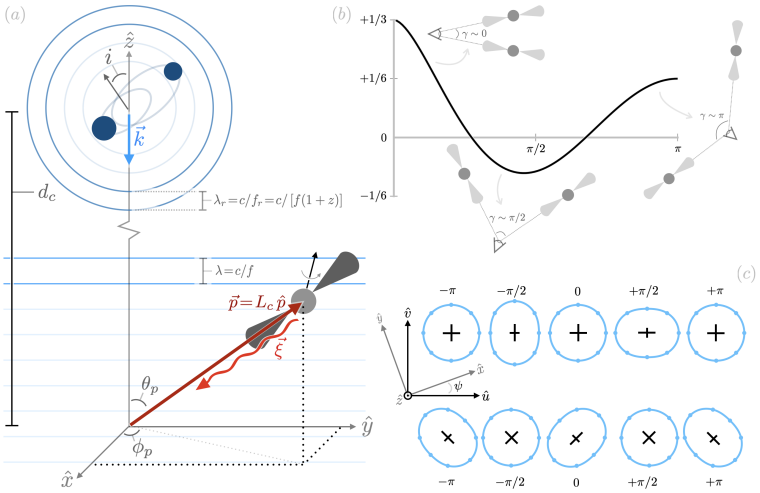

Fig. 2 (a) Binary at comoving distance d_c along the $\hat{z}$ axis, emitting GWs along the $\vec{k}$ vector; and pulsar at position $\vec{p}$ emitting pulses along the path $\vec{\xi}$. (b) 'Hellings-Downs Curve': signal correlation between pulsars vs. angle between them, $\mu_{\rm HD}(\gamma)$. (c) Two polarization patterns vs. GW phase from $[-\pi, +\pi]$, for 'plus +' (upper row) and 'cross ×' (lower row) polarizations.

It will be convenient to rearrange Eq. 11 by grouping the phase and amplitude terms such that $h^a_{ij} = h^a\epsilon^a_{ij}$, and also introducing *geometrical projection factors* (or *antenna response functions*) $F^a(k_i)$:

$$z(t) = \sum_a F^a(k_i, p_i)\cdot(h^a_E - h^a_P), \tag{12a}$$

$$F^a(k_i, p_i) \equiv \frac{1}{2}\frac{\hat{p}^i\hat{p}^j}{1 + \hat{k}^i\hat{p}_i}\epsilon^a_{ij}, \tag{12b}$$

$$h^a = A_a\,\exp\left(i\,k_\alpha x^\alpha\right) = A_a\exp\Big[2\pi i\left(ft - k_i x^i\right)\Big]. \tag{12c}$$

Here we have substituted $h^a_E = h^a(t,0)$ for the Earth term in each polarization, and $h^a_P = h^a(t - L_c/c, p)$ for the pulsar term. The amplitude

[10]Note that, even for constant frequency, this is not the case for time-delays: the expression in Eq. 8 cannot be reduced in general as there is an additional build-up of time/phase change between the endpoints of the trajectory. Under the long-wavelength approximation, $L_c \ll k^{-1} = \lambda$, the integrand can be taken as constant, and we obtain a projected version of Eq. 7 which is appropriate for high-frequency, ground-based GW interferometers.

of a GW signal embedded in pulsar timing data will typically be far smaller than the amplitude of noise (Sec. 3.2). The *measured* redshift for a pulsar p, including noise $n_p(t)$ for that pulsar, is:

$$z_p(t) = \sum_a F_p^a \, h_E^a - F_p^a \, h_P^a + n_p(t). \tag{13}$$

The key to pulsar timing *arrays* is examining the cross-correlations between a pair of pulsars p and s (Sec. 3.3):

$$\rho_{ps}(\tau) = \Big\langle z_p(t)\, z_s(t-\tau) \Big\rangle_T \equiv \frac{1}{T_{\rm obs}} \int_{-T_{\rm obs}/2}^{T_{\rm obs}/2} z_p(t)\, z_s(t-\tau)\, dt, \tag{14}$$

specifically at zero time-lag, $\rho_{ps} \equiv \rho_{ps}(\tau = 0)$. Here, we are integrating over a total observing-duration $T_{\rm obs} \sim 10\,{\rm yr}$. Because the number of GW wave-cycles between the Earth and each pulsar (and also between different pulsars) is large, $fL/c \gg 1$, this means that the pulsar term and the Earth term will be effectively independent, and the pulsar terms from different pulsars will also be nearly independent[11]. If the noise is also independent between different pulsars, then,

$$\rho_{ps} \approx \sum_a \frac{F_p^a \, F_s^a}{T} \int_{-T/2}^{T/2} \big|h_E\big|^2 dt \approx P_h \sum_a F_p^a \, F_s^a. \tag{15}$$

Here, P_h is approximately the (average) signal-'power' of $h(t)$ in the data-stream, which becomes exact in the limit that $T_{\rm obs} \to \infty$.

It is convenient to consider the Fourier transform of signals, and their cross-correlations. The Fourier transform is,

$$\tilde{h}(f) \equiv \lim_{T_{\rm obs}\to\infty} \int_{-T_{\rm obs}/2}^{+T_{\rm obs}/2} h(t)\, e^{-2\pi i f t}\, dt. \tag{16}$$

Here we use $\tilde{h}(f)$ to emphasize that this is a Fourier transform, however we will later follow the convention used in the literature that both the time-domain and frequency-domain GW-strain use the same symbol $h : h(t) \leftrightarrow h(f)$, which is motivated by sinusoidal waveforms. We can use Parseval's theorem to relate the signal-power in time and frequency, and thereby define the (one-sided) power spectral density $S_h(f)$:

$$P_h = \frac{1}{T_{\rm obs}} \int_{-T_{\rm obs}/2}^{+T_{\rm obs}/2} |h(t)|^2\, dt = \frac{1}{T_{\rm obs}} \int_{-\infty}^{+\infty} \big|\tilde{h}(f)\big|^2\, df \equiv 2\int_0^{+\infty} S_h(f)\, df. \tag{17}$$

For a stationary process, i.e. one whose statistical properties are independent of time, each frequency is uncorrelated such that,

$$\Big\langle \tilde{h}(f)\, \tilde{h}(f') \Big\rangle = S_h(f)\, \delta(f - f'), \tag{18}$$

where $\delta(x{=}0) = 1$ and $\delta(x{\neq}0) = 0$. This allows us to define frequency-specific cross-correlations as,

$$\rho_{ps}(f) \approx S_h(f) \sum_a F_p^a \, F_s^a. \tag{19}$$

Hellings and Downs (1983) showed that the cross-correlation could be averaged over a large number of unpolarized GW sources, distributed uniformly across the sky, to obtain[12]:

$$\langle \rho_{ps} \rangle_\Omega = \langle \big|h_E\big|^2 \rangle_T \, \mu_{\rm HD}(\gamma) = P_h \, \mu_{\rm HD}(\gamma), \tag{20a}$$

$$\langle \rho_{ps}(f) \rangle_\Omega = S_h(f)\, \mu_{\rm HD}(\gamma), \tag{20b}$$

$$\mu_{\rm HD}(\gamma) \equiv \frac{1}{3} - \frac{1}{6}\left(\frac{1-\cos(\gamma)}{2}\right) + \left(\frac{1-\cos(\gamma)}{2}\right) \ln\left(\frac{1-\cos(\gamma)}{2}\right). \tag{20c}$$

The **Hellings-Downs correlation**, $\mu_{\rm HD}(\gamma)$, depends entirely on the angle between pulsars: $\cos(\gamma) = \hat{p}_{p,i}\, \hat{p}_s^i$. The Hellings-Downs curve, shown in Fig. 2(b), is the 'smoking gun' for evidence of a GW signal in PTA data.

For a comprehensive discussion of HD correlations, see Romano and Allen (2024), which we have largely followed. A few additional points are worth noting. Cornish and Sesana (2013) show that Eqs. 20 also holds for a single GW source, averaging over a large number of uniformly distributed pulsars[13], i.e. $\langle \rho_{ps} \rangle_\Omega = \langle \rho_{ps} \rangle_{ps}$. Allen (2023) shows that the standard deviation in the HD correlations can be comparable to the mean, even in the limit of large numbers of sources; and derive expressions for the mean and variance of the correlations with and without polarization, and with and without the pulsar terms.

2.3 GWs from Binaries

For additional details, we refer the reader to Misner et al. (1973, Part VIII), Flanagan and Hughes (1998, Ch. 4), and Enoki and Nagashima (2007).

[11] For individual GW plane-waves, the pulsar terms will be correlated at time-lags corresponding to their light-travel distances (for the Earth terms) or the differences in light-travel times (between pulsar-terms). Including the pulsar term(s) can thus double the amount of signal when searching for individual sources, but this is outside of the scope for this discussion.

[12] We designate the average over the sky as $\langle \ldots \rangle_\Omega$ and the average over time as $\langle \ldots \rangle_T$.

[13] c.f. Romano and Allen (2024) however, who show that averaging over pulsars can be more nuanced than averaging over sources.

2.3.1 Single Binary GW Power and Evolution

In general, GW production must be calculated numerically. However, an analytic solution exists in the limits of (i) 'slow motion'[14], i.e. $v \ll c$; and (ii) that the gravitating objects are moving pseudo-periodically. A full derivation is an involved process, but the results for a binary in nearly-Keplerian orbit[15] are surprisingly simple (Peters and Mathews, 1963). The power radiated, or luminosity L, into a solid angle Ω from each polarization a, can be expressed in terms of the third time derivative of the quadrupole moment tensor (Q_{ij}),

$$\frac{dL^a}{d\Omega} = \frac{G}{8\pi c^5}\left[\frac{d^3Q_{ij}}{dt^3}\epsilon^a_{ij}\right]^2, \tag{21a}$$

$$Q_{ij}(t) \equiv \int\left[x_i\,x_j - \frac{1}{3}\delta_{ij}x_k\,x^k\right]\rho(x,t)\,d^3x, \tag{21b}$$

where ρ is the mass density. For a set of point-masses m_k we can write, $Q_{ij} = \sum_k m_k\,x_{k,i}\,x_{k,j}$. For a binary in a circular orbit, GWs are emitted at twice the orbital frequency. In this case, the rest-frame total luminosity, summed over both polarizations and integrated over all angles, is:

$$L_{\rm circ} = \frac{32}{5Gc^5}\,(G\mathcal{M})^{10/3}\,(2\pi f_{r,\rm orb})^{10/3}, \tag{22a}$$

$$\approx 8.4\times10^{39}\,{\rm W}\left(\frac{\mathcal{M}}{10^9\,{\rm M}_\odot}\right)^{10/3}\left(\frac{f_{r,\rm orb}}{3\,{\rm nHz}}\right)^{10/3} \tag{22b}$$

where $f_{r,\rm orb}$ is the rest-frame orbital frequency. Much of GW emission is determined by a particular combination of the two component masses m_1 and m_2, called the '**chirp mass**',

$$\mathcal{M} = \frac{(m_1m_2)^{3/5}}{M^{1/5}} = M\frac{q^{3/5}}{(1+q)^{6/5}}, \tag{23}$$

where the total mass is $M = m_1 + m_2$, and the mass ratio is $q = m_2/m_1 \le 1$ We define all masses to be in the rest frame. From Eqs. 22, we see that the amount of energy radiated by GWs is enormous: comparable to the EM energy radiated from a large galaxy or bright quasar. However, the dynamical effects of these GWs, and thus their detectability, is quite small due to the intrinsically weak coupling of the gravitational force.

Binaries with an eccentricity e emit GWs at all integer harmonics n of the orbital frequency. This allows us to decompose the luminosity into contributions from each harmonic n,

$$L_{\rm gw} = \sum_{n=1}^{\infty}L_n = \sum_{n=1}^{\infty}L_{\rm circ}\,g(n,e) = L_{\rm circ}\,F(e), \tag{24a}$$

$$g(n,e) \equiv \frac{n^4}{32}\Bigg(\Big[J_{n-2}(ne) - 2eJ_{n-1}(ne) + \frac{2}{n}J_n(ne) + 2eJ_{n+1}(ne) - J_{n+2}(ne)\Big]^2 \tag{24b}$$

$$+\left(1-e^2\right)\Big[J_{n-2}(ne) - 2eJ_n(ne) + J_{n+2}(ne)\Big]^2 + \frac{4}{3n^2}\Big[J_n(ne)\Big]^2\Bigg),$$

$$F(e) \equiv \frac{1 + \frac{73}{24}e^2 + \frac{37}{96}e^4}{\left(1-e^2\right)^{7/2}}. \tag{24c}$$

Here, $F(e)$ is the 'eccentricity function', $g(n,e)$ is the 'frequency distribution function', and $J_n(x)$ is the nth-order Bessel function of the first kind. Note that for zero eccentricity: $g(n=2, e=0) = 1$ and $g(n\neq2, e=0) = 0$.

During the psuedo-Keplerian phase, the energy emitted from a binary as gravitational waves comes from the orbital energy[16]. This allows us to calculate the 'hardening'[17] rate, at which the binary inspirals. Let us define the '**hardening timescale**' (or '**residence timescale**') with respect to frequency as: $\tau_f \equiv dt/d\ln(f) = f/(df/dt) = -\frac{2}{3}a/(da/dt)$. For gravitational wave emission (Peters, 1964),

$$\tau_{f,\rm GW} \equiv \frac{f}{[df/dt]_{\rm gw}} = \frac{5}{96}\left(\frac{G\mathcal{M}}{c^3}\right)^{-5/3}\frac{(2\pi f_{r,\rm orb})^{-8/3}}{F(e)}, \tag{25a}$$

$$\approx 4.5\times10^5\,{\rm yr}\left(\frac{\mathcal{M}}{10^9\,{\rm M}_\odot}\right)^{-5/3}\left(\frac{f_{r,\rm orb}}{3\,{\rm nHz}}\right)^{-8/3}[F(e)]^{-1}. \tag{25b}$$

Note that the total binary lifetime to go from $f_{r,\rm orb}$ to coalescence ($f\to\infty$) is notably smaller: $\approx (3/8)\,\tau_{f,\rm GW}$. The same chirp-mass at a relatively high frequency of ≈ 30 nHz would still have a residence time of 970 yr. Thus, for typical binary parameters, PTA sources can be

[14] For gravitating systems this is equivalent to the 'quadrupole approximation': that the size-scale of the objects and their motion is small compared to the emitted wavelength.

[15] 'Nearly-keplerian' in the sense that the orbital elements change on timescales much longer than the binary orbital period.

[16] During the 'chirp' and final coalescence of the binary, this is no longer entirely true, and up to $\sim 5\%$ of the rest-mass energy of the binary can be emitted as GWs.

[17] The term 'hardening' is adopted from the stellar binary literature where 'soft' binaries at larger separations tend to separate further due to three body interactions, while 'hard' binaries at smaller separations tend to come closer together: the so-called 'Heggie' or 'Heggie-Hills' law.

considered as monochromatic over human lifetimes[18]. When discussing MBH binary evolution (Sec. 4.1), it is convenient to write the total binary lifetime in terms of binary semi-major axis: $\tau_{\rm life} = \int_{a_0}^{a_1} (da/dt)^{-1} da$. Due to GW emission, this is $\approx (a/4)/(da/dt)$ or exactly,

$$\tau_{\rm life,GW} = \frac{5\,c^3}{256\,G^3}\left(\frac{a_0^4 - a_1^4}{M\,m_1\,m_2}\right) \tag{26a}$$

$$\approx 2.3\ {\rm Gyr}\left(\frac{\mathcal{M}}{10^9\ {\rm M}_\odot}\right)^{5/3}\left(\frac{M}{3\times10^9\ {\rm M}_\odot}\right)^{4/3}\left(\frac{a}{0.4\ {\rm pc}}\right)^4. \tag{26b}$$

Due to the steep dependence on separation (a^4), the 'final' separation a_1 can be taken as zero (c.f. Sec. 2.3.4). GWs also carry away angular momentum, which can be related to the binary eccentricity. The eccentricity-decay timescale is,

$$\tau_{e,\rm GW} \equiv \frac{e}{-[de/dt]_{\rm gw}} = \frac{15\,c^5}{304\,G^3}\frac{a^4}{M\,m_1\,m_2}\left(1 + \frac{121}{304}e^2\right)^{-1} \tag{27}$$

which is very similar to the lifetime in terms of separation.

It will be convenient to know the spectrum of GW energy emitted by a binary. We will denote the rest-frame GW frequency of the binary as it evolves over time as $f_r'(t)$, and use a delta-function to pick-out the time at which the binary is emitting at the rest-frame frequency of interest f_r:

$$\frac{dE_{\rm gw}(f_r)}{d\ln f_r} = \int dt\ \frac{d^2E_{\rm gw}}{dt\,d\ln f_r'}\,\delta[f_r'(t) - f_r], \tag{28a}$$

$$= \sum_{n=1}^{\infty} L_{\rm circ}(f_{r,\rm orb})\,\frac{dt}{d\ln f_r}\,\frac{g(n,e)}{n}\bigg|_{f_{r,\rm orb}=f_r/n}. \tag{28b}$$

The second equality requires the binary to *reach* the orbital frequency corresponding to the GW frequency of interest. This is not the case when the orbital frequency is: (i) below the frequency at which the binary becomes bound, (ii) above the frequency at which the binary coalesces, or otherwise (iii) that the binary evolution stalls before reaching that frequency. Given those caveats, we can calculate the GW energy spectrum for a circular binary, assuming purely GW-driven binary evolution (i.e. $dt/d\ln f_r = \tau_{f,\rm GW}$) to find,

$$\frac{dE_{\rm gw}(f_r)}{d\ln f_r}\bigg|_{\rm gw} = \frac{(G\mathcal{M})^{5/3}}{3\,G}(2\pi f_r)^{2/3}\bigg|_{f_r=2\,f_{r,\rm orb}}. \tag{29}$$

2.3.2 GW Strain

The GW strain can be calculated directly from the second time-derivative of the quadrupole moment tensor as,

$$h_{ij}(t) = \frac{2G}{c^4\,d_{\rm com}}\frac{\partial^2 Q_{ij}(t - d_{\rm com}/c)}{\partial t^2}, \tag{30}$$

where we have noted explicitly that the quadrupole-moment tensor must be evaluated at the 'retarded time' when the GW was emitted. Here $d_{\rm com}$ is the 'comoving distance' to the source. The GW power can be related to the strain of each harmonic as,

$$h_{s,n}^2(f_{\rm GW}) = \frac{G}{c^3}\left(\frac{2}{n}\right)^2\frac{L_n}{(2\pi\,f_{r,\rm orb})^2\,d_{\rm com}^2} = h_{s,\rm circ}^2\,g(n,e)\left(\frac{2}{n}\right)^2. \tag{31}$$

In Eq. 31 we have shifted to describing the strain as a function of frequency instead of time, with the two related by the Fourier transform (Eq. 16). Note also that the observer-frame GW frequency is $f_{\rm GW} = n\,f_{r,\rm orb}/(1+z)$, for a source at a cosmological redshift z. The average strain for a binary in a circular orbit can be expressed as,

$$h_{s,\rm circ}(f_{\rm GW}) = \frac{8}{10^{1/2}}\frac{(G\mathcal{M})^{5/3}}{c^4\,d_{\rm com}}(2\pi f_{r,\rm orb})^{2/3}, \tag{32}$$

$$\approx 7.9\times10^{-16}\left(\frac{\mathcal{M}}{10^9\ {\rm M}_\odot}\right)^{5/3}\left(\frac{f_{r,\rm orb}}{3\ {\rm nHz}}\right)^{2/3}. \tag{33}$$

This 'source strain', defined for each harmonic, is a scalar value: the strain averaged over both polarizations and over all solid angles (c.f. the strain tensor of Eq. 3 or Eq. 30). The time-dependent strain of each polarization at all viewing angles and non-zero eccentricity can be found in Barack and Cutler (2004).

When considering the detectability of GW signals, it is common to utilize the GW '**characteristic strain**' h_c instead of the raw GW strain h or h_s (for thorough discussions, see: Flanagan and Hughes, 1998; Moore et al., 2015). The goal is to construct a measure of GW strength (h_c) that will be proportional to the square-root of the **signal-to-noise ratio (SNR)** for a given detector. The square-root ensures that the GW energy is still $\propto h_c^2$. This measure of strain is ill-defined instantaneously: it is intrinsically *integrated*, over some observing

[18] Because the light-travel time from a pulsar can be $\sim 10^3$ yr, the pulsar may see a noticeably different GW frequency than the Earth. Thus inclusion of the 'pulsar term' (i.e. Eq. 11) can be used to measure the 'chirp', and thus directly constrain the MBHB mass, independently of its distance.

time and possible also frequency bandwidth. For nearly-monochromatic sources, the appropriate quantity is,

$$h_c^2(f) = f\, T_{\rm obs}\, h_s^2(f), \tag{34}$$

where the power of the signal is increased by the number of cycles over an observing duration $T_{\rm obs}$. This is usually the quantity most relevant for PTAs. The characteristic strain for frequency-evolving sources is instead,

$$h_{c,\rm chirp}^2(f) = \frac{2f^2}{df/dt} h_s^2(f), \tag{35}$$

where the power is increased by the number of cycles near the frequency f. More generally, the characteristic strain can be related to the (one-sided) power spectral density (S_h) as, $h_c^2 = f\, S_h(f)$. This is sometimes taken as the definition of characteristic strain. In practice, the power spectral density would be calculated by measuring the time-dependent strain $h(t)$, and taking the Fourier transform to find $\tilde{h}(f)$, which can then be used to estimate $S_h(f)$ using Eq. 17. In this way, the power spectral density and characteristic strain are still well defined for stochastic signals (discussed below). Particularly in cosmological contexts, it is common to describe signals in terms of the GW energy density ($\mathcal{E}_{\rm gw}$) per logarithmic frequency interval,

$$\Omega_{\rm gw}(f) \equiv \frac{1}{\rho_c\, c^2} \frac{d\mathcal{E}_{\rm gw}}{d\ln f} = \frac{2\pi^2}{3H_0^2} f^3 S_h(f). \tag{36}$$

Here, the GW-energy spectrum[19] is normalized by the cosmological critical density, $\rho_c = 3H_0^2/8\pi G$, for a redshift-zero Hubble constant H_0.

2.3.3 The GW Background from a Population of Binaries

As shown in Eqs. 25, the lifetime of MBH binaries in the PTA band is long ($\gtrsim 10^5$ yr), and we might expect a number of binaries to be emitting in each frequency bin (discussed further in Sec. 4). The combination of a large number of individual sources with different frequencies and phases leads to a '**stochastic GW Background (GWB)**'. Phinney (2001) presented an elegant method to calculate the local/present-day energy density of GWs based on the integrated history of GW emission from binaries over cosmic time[20]. Specifically, we can write that,

$$\frac{d\mathcal{E}_{\rm gw}(f)}{d\ln f} = \frac{\pi\, c^2}{4\, G} f^2\, h_c^2(f) = \int_0^\infty \frac{dz}{1+z} \frac{d^2n}{d\mathcal{M}\, dz} \frac{dE_{\rm gw}(\mathcal{M}, f_r)}{d\ln f_r}\bigg|_{f_r = f(1+z)}. \tag{37}$$

The comoving number density of binaries is $n \equiv dN/dV_c$, where $V_c(z)$ is the comoving volume at a redshift z, and the $1+z$ term in Eq. 37 accounts for the redshifting of GW energy from emission to observation. Here $\mathcal{E}_{\rm gw}$ is the local energy per unit comoving volume of GWs, while $E_{\rm gw}$ is the total energy emitted in GWs from a particular binary, and its spectrum is given by Eqs. 28. Assuming GW-only evolution of a smooth continuum of circular binaries (Eq. 29) we find,

$$h_c^2(f) = \frac{4\pi}{3c^2} (\pi f)^{-4/3} \int\int dz\, d\mathcal{M} \frac{d^2n}{d\mathcal{M}\, dz} \frac{(G\mathcal{M})^{5/3}}{(1+z)^{1/3}}, \tag{38a}$$

$$h_c(f) \approx 1.1\times 10^{-15} \left(\frac{f_{\rm GW}}{1\ {\rm yr}^{-1}}\right)^{-2/3} \left(\frac{n_{\rm eff}}{10^{-4}\ {\rm Mpc}^3} \left(\frac{\mathcal{M}_{\rm eff}}{10^9\ {\rm M}_\odot}\right)^{5/3} (1+z)_{\rm eff}^{-1/3}\right)^{1/2}. \tag{38b}$$

The second relation includes the insight that each merger leads to a redshift-zero remnant MBH, so that the local number-density (or mass-density) of MBHs can be directly tied to the GWB amplitude. The second relation in Eq. 38a relies on *ad hoc* effective/integrated parameter values. We have neglected terms that describe the fraction of local mass built-up my mergers. A precise and insightful treatment can be found in Sato-Polito et al. (2024).

The most important feature of the GWB from MBH binaries is a characteristic-strain spectrum $h_c(f) \propto f^{-2/3}$. Often the spectrum is characterized/fit as a power-law, normalized at a frequency of 1 yr$^{-1} \approx 31$ nHz, i.e. $h_c(f) \approx A_{\rm yr^{-1}} \left(\frac{f}{1\ {\rm yr}^{-1}}\right)^{-2/3}$. Idealized, power-law GWB spectra are plotted as dashed lines in Fig. 3; purple in panel (a) and grey in panels (b),(c),(d). Realistic models of MBH binary populations show considerable deviations from this idealized power-law. GWB spectra from five random realizations of binary populations are shown as purple lines in Fig. 3(a), while 50%, 90%, &98% confidence intervals are shown with purple contours.

There are four primary causes of deviation from the idealized power-law of Eq. 38:

1. The GWB is formed by a discrete number of binaries, with variations in the number of binaries from bin-to-bin ('cosmic variance', 'Poisson variations', or 'shot noise'), which will produce deviations from a uniform spectral index. This is shown in Fig. 3(a).
2. At sufficiently high frequencies (typically $f \gtrsim 30$nHz), the expected number of binaries contributing to the GWB will drop below unity, and the spectrum will steepen. Higher-mass binaries will also coalesce, and no longer contribute GWs at the highest frequencies. Fig. 3(b) shows the effects on the spectrum of changing the relative number of binaries vs. their masses, contributing to the GWB.

[19] The terminology is confusing here. Note that the 'power' spectral density is a generalized power of a (dimensionless) signal in data, with units of inverse frequency. The GW-energy density, $\mathcal{E}_{\rm gw}$ is a *physical*-energy per unit volume, and ρ_c is a mass-density: mass per unit volume.

[20] As Phinney (2001) points out, this is a GW-version of the '**Soltan argument**': if quasar luminosity is accretion powered, then the integrated energy in quasar light over the history of the Universe, divided by the fraction of rest-mass energy radiated away, must match the present day mass-density of massive black holes (the remnants of past quasars).

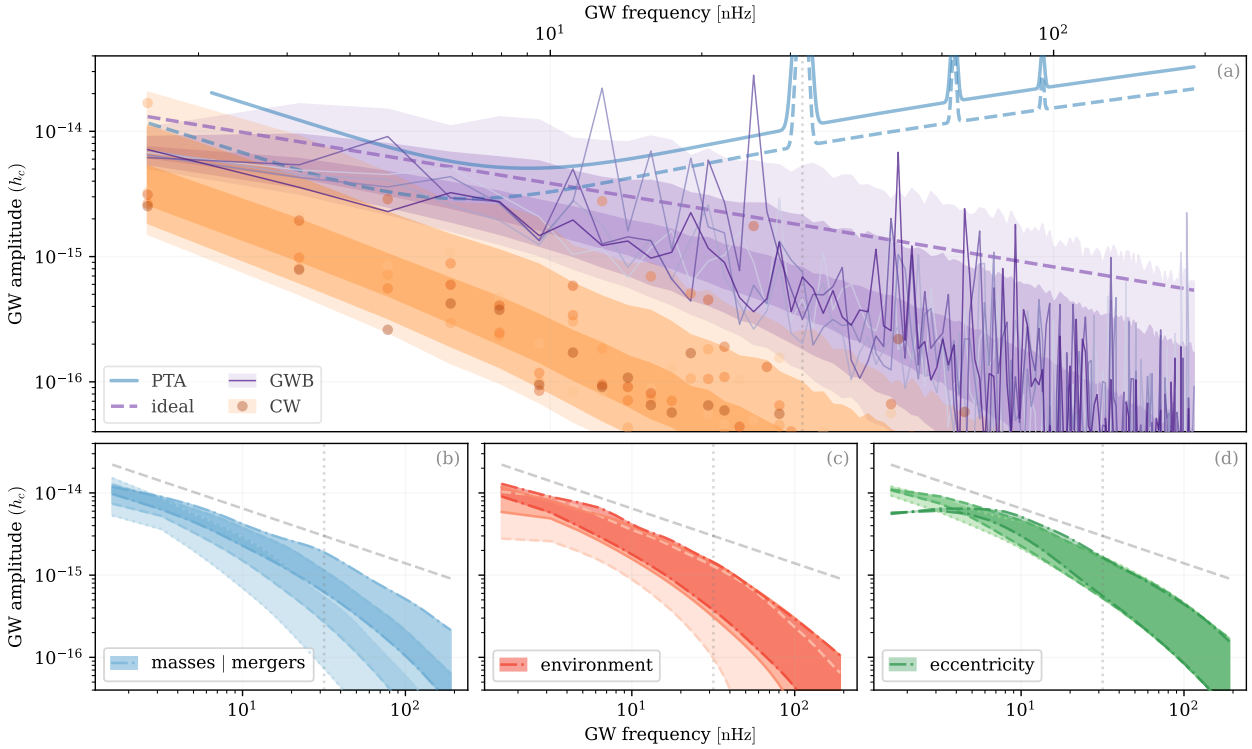

Fig. 3 (a) GWs from binary populations. GWB (*blue*): with $50\%, 90\%, 98\%$ intervals (contours), and five random realizations (*lines*); loudest CW in each frequency bin (*orange*): same contours and random realizations (*circles*). Idealized GWB spectrum (*black dashed*) from Eq. 38a. Schematic PTA sensitivity curves (blue) for 15yrs (solid) and 20yrs (dashed) of data. **(b) Effects of MBH masses and merger rates** on GWB (blues), showing 50% intervals for three population models, each calibrated to the same idealized amplitude - decreasing merger rates as masses increase. **(c) Effects of environmental hardening** on GWB (reds), 50% intervals for models with three different strengths of environmental interactions at sub-parsec separations. **(d) Effects of binary eccentricity** on GWB (greens), 50% intervals of varying binary eccentricities $(0.5, 0.9, 0.95$ initialized at 0.1 pc).

3. At low frequencies (large separations), binary evolution is not entirely GW-driven, with noticeable contributions from 'environmental' processes (Sec. 4.1). This increases the rate of binary inspiral, which decreases the total GW deposition, and thus attenuates/decreases the GWB spectrum at low frequencies (see Eq. 28). The effectiveness of environmental interactions also determines the fraction of binaries able to reach the PTA band. Varying environmental interaction strengths are shown in Fig. 3(c).
4. Eccentric binaries distribute their emitted GW energy across a range of frequencies, instead of only at twice the orbital frequency. In populations that have considerable eccentricity in the PTA band, this decreases the GW energy at low frequencies, and increase the GW energy at high frequencies. At the same time, eccentric binaries generally inspiral more quickly than circular binaries. These effects are expected to be relatively small. Fig. 3(c) compares the GWB spectrum from different eccentricity binary populations.

Each type of deviation, when measured by PTAs, provides significant additional information about the underlying MBHB population. Thus it is largely the *deviations* from the idealized power-law behavior which can provide the most information from low-frequency GW observations[21].

These effects can be taken into account by: first, recasting Eq. 37 in terms of the number of binaries at each frequency ($dN/d\ln f_r$), instead of their number-density (Sesana et al., 2008; Sesana, 2013); and second, by self-consistently evolving binaries from formation (large separations) to the frequencies of interest (smaller separations) while including binary eccentricity (Kelley et al., 2017a,b). The latter point will be discussed further in Sec. 4.1. The key argument to the former is explicitly connecting each binary's evolution over frequencies to the time evolution of the Universe:

$$\frac{\partial^3 N}{\partial \mathcal{M}\, \partial z\, \partial \ln f_r} = \frac{\partial^3 N}{\partial \mathcal{M}\, \partial z\, \partial V_c} \frac{\partial V_c}{\partial z} \frac{\partial z}{\partial t_r} \frac{\partial t_r}{\partial \ln f_p} \tag{39a}$$

$$= \frac{\partial^2 n}{\partial \mathcal{M}\, \partial z}\, 4\pi c\, d_c^2\, (1+z)\, \tau_f. \tag{39b}$$

[21] While outside the scope of this article, yet more information is encoded in the occurrence rate of individually detectable 'CW' binary sources, and the closely-related anisotropy of GWB emission across the sky.

Plugging this expression into Eq. 37, along with the circular and GW-only approximations, gives us the result that,

$$\left\langle h_c^2(f) \right\rangle = \int d\mathcal{M}\, dz\, \frac{\partial^3 N}{\partial \mathcal{M}\, \partial z\, \partial \ln f_r}\, h_s^2(\mathcal{M}, z, f_r)\Big|_{f_r = f(1+z)/2}. \tag{40}$$

In other words, the GWB characteristic strain is given by the quadrature sum of the strains from all binaries per logarithmic frequency interval. We have added angle brackets on the left-hand side of Eq. 40 to emphasize that this yields an expectation value, or average GWB strain spectrum. This comes both from extrapolating a number-density of binaries to a total number in the Universe (Eq. 39b), and also from adding GW strains in quadrature which neglects interference between individual binaries and different phases.

2.3.4 Coalescence Rates and Memory Effects

During the final few orbits of binary inspiral, the 'slow motion' approximation becomes increasingly inaccurate and the GW emission relations presented in Sec. 2.3.1 break down. This happens near the mutual 'innermost stable circular orbit (ISCO)' of the binary. The ISCO location for MBH binaries, is often approximated as that of an extreme mass-ratio inspiral with zero spin(s):

$$a_{\rm ISCO} \equiv \frac{6\,GM}{c^2} \approx 8.6\times10^{-4}\,{\rm pc}\left(\frac{M}{3\times10^9\,{\rm M}_\odot}\right) \qquad f_{\rm ISCO} \approx 730\,{\rm nHz}\left(\frac{M}{3\times10^9\,{\rm M}_\odot}\right)^{-1}. \tag{41}$$

While, technically, such high frequencies could be probed by PTAs, in practice the sensitivity is entirely swamped by noise and the detection of a chirp is unfeasible. Additionally, the coalescence rate of such high-mass binaries is expected to be vanishingly small (Sec. 4). The coalescence rate can be calculated in a similar manner to the number of binaries (Eq. 39b),

$$\frac{\partial N}{\partial t} = \int dz\, \frac{\partial^2 N}{\partial z\, \partial V_c}\, \frac{\partial V_c}{\partial z}\, \frac{\partial z}{\partial t_r}\, \frac{\partial t_r}{\partial t} \tag{42a}$$

$$= \int dz\, \frac{\partial n}{\partial z}\, 4\pi\, c\, d_c^2. \tag{42b}$$

When such coalescences occur, they produce not only a GW chirp, but also a shift or 'DC offset' in the space-time metric which returns to a different rest state than it started (Thorne, 1992). This produces a broad-band signal called a GW '**burst with memory**'. The characteristic amplitude can be estimated as (*Ibid.*),

$$h_{c,{\rm BWM}} \approx 10^{-17}\left(\frac{\epsilon_{\rm BWM}}{10^{-2}}\right)\left(\frac{\mathcal{M}}{10^9\,{\rm M}_\odot}\right)\left(\frac{d_{\rm com}}{500\,{\rm Mpc}}\right), \tag{43}$$

where $\epsilon_{\rm BWM}$ is an efficiency parameter that depends on orientation and BH spins. For more information, see, for example, Favata (2010).

3 Detection of Gravitational Waves with Pulsar Timing Arrays

Before considering the detection of GWs, consider the much simpler case of placing limits on the amount of GW power at the Earth, using pulsar measurements. We can limit the local GW spectral strain based on the fractional, maximum deviation in pulse **times-of-arrival (TOAs)**. We will see below that (i) we will be highly noise dominated, and (ii) that we will be interested in the superposition of many GWs which can interfere with each other. For these reasons, instead of relying on a single maximum deviation, we measure the mean-squared deviations which are proportional to the signal power (e.g., Eq. 17), and then limit the GW power to the same amount:

$$S_h \propto \langle \Delta h^2 \rangle = \sigma_h^2 \lesssim \frac{\langle \Delta t^2 \rangle}{T_{\rm obs}^2} = \frac{\sigma_{\Delta t}^2}{T_{\rm obs}^2} \propto S_{\rm TOA}. \tag{44}$$

Here, Δt is the TOA deviation, $T_{\rm obs}$ is the total observing duration, and σ_h and $\sigma_{\Delta t}$ are the standard-deviations in strain and TOA deviations respectively[22].

Typically we will characterize GW power in terms of a power spectrum over frequencies $S_h(f)$ (Sec. 2.2). Recall that the measurable GW frequencies are limited to falling between the Rayleigh and Nyquist frequencies, i.e. $T_{\rm obs}^{-1} \lesssim f \lesssim 1/(2\,\Delta T_{\rm cad})^{-1}$, for a typical time between TOA measurements ('sampling interval' or 'cadence'), $\Delta T_{\rm cad}$. Specifically, the Fourier frequency basis is $f_i = i/T$, for integer values $i \in \{1, \ldots, N_f\}$. The minimum frequency, and the frequency bin-width, is the 'Rayleigh' frequency, $f_1 = T_{\rm obs}^{-1}$. The maximum frequency for evenly sampled data is the 'Nyquist' frequency, $f_{N_f} = (2\,\Delta T_{\rm cad})^{-1}$, such that the number of independent frequencies is $N_f \approx T/(2\,\Delta T_{\rm cad})$, although f_{N_f} becomes poorly defined for irregularly sampled data, which is typically the case for PTAs.

3.1 Pulsar Timing

For additional details, we refer the reader to Romano and Cornish (2017) and Taylor (2021).

The accuracy by which TOAs can be measured is determined by a combination of the intrinsic pulse stability (determined by largely-unknown pulsar physics) and the radiometer measurement accuracy, i.e. $\sigma_{\Delta t}^2 \sim \sigma_p^2 + \sigma_{\rm radio}^2$. For general information on pulsar astronomy, we

[22] The variance of a quantity x is, $\sigma_x^2 \equiv \langle \Delta x^2 \rangle - \langle \Delta x \rangle^2$, and both strain and timing residuals have zero means s.t. $\sigma_x^2 \approx \langle \Delta x^2 \rangle$.

direct the reader to Lorimer and Kramer (2012). Empirically, the best millisecond pulsars can have stabilities of ~ 100 ns. The measurement accuracy can be estimated as,

$$\left(\frac{\sigma_{\rm radio}}{w}\right)^2 \approx \left(\frac{w}{P-w}\right)\left(\frac{S_{\rm pulsar}}{S_{\rm radio}}\right)^{-1} N_{\rm radio}^{-1}, \tag{45}$$

i.e. the precision of the pulsar TOA ($\sigma_{\rm radio}$) over the pulse width (w), is the ratio of the pulse width to non-pulse ('noise') for pulsar period P, divided by the ratio of power from the pulsar signal to the instrument noise, divided by the number of measurements. This is called the 'radiometer equation'. The number of measurements (samples) can be expressed as a product of observing duration, bandwidth (i.e. data sampling rate), and number of independent polarizations (two): $N_{\rm radio} = T_{\rm TOA} \cdot \Delta f \cdot 2$. We can then rewrite the radiometer equation adopting typical values as[23]:

$$\sigma_{\rm radio} \approx 100\ {\rm ns} \left(\frac{w}{1\ {\rm ms}}\right)^{3/2}\left(\frac{S_{\rm pulsar}}{1\ {\rm mJy}}\right)^{-1/2}\left(\frac{S_{\rm radio}}{10\ {\rm Jy}}\right)^{1/2}\left(\frac{T_{\rm TOA}}{30\ {\rm min}}\right)^{-1/2}\left(\frac{P}{10\ {\rm ms}}\right)^{-1/2}\left(\frac{\Delta f}{100\ {\rm Mhz}}\right)^{-1/2}. \tag{46}$$

For millisecond pulsars, the single-burst signal-to-noise ratio is often on the order of $S_{\rm pulsar}/S_{\rm radio} \sim 10^{-4}$, and thus large numbers of pulses must be stacked or 'folded' to become detectable. Choosing an observing duration of ~ 30 min gives us an instrumental precision matching the intrinsic stability of millisecond pulsars. The optimal PTA 'sensitivity' to gravitational waves over a decade of observing time is then $h_{\rm PTA} \approx \langle \Delta h^2 \rangle^{1/2} \approx 10^{-16} - 10^{-15}$. For a careful analysis of PTA sensitivities, see Siemens et al. (2013, which includes sensitivity forecasting) and Hazboun et al. (2019, which includes the construction of sensitivity curves).

It is convenient to express pulse TOAs in terms of different components: a deterministic component (which may include any GW signals of interest) and a stochastic 'noise' component[24],

$$\delta t(t) = \overline{\Delta t}(t) + n(t). \tag{47}$$

The goal of pulsar timing is to capture the deterministic component through a **timing model**, $\Delta t(t|\alpha) \approx \overline{\Delta t}(t)$, with model parameters α. The '**timing residuals**', the differences between observations and the timing model, are then:

$$r(t|\alpha) \equiv \delta t(t) - \Delta t(t|\alpha). \tag{48}$$

The timing model is fit to the data by constructing a likelihood that assumes the residuals to be distributed as multidimensional Gaussian noise:

$$p(r|\Delta t) = \left[\det\left(2\pi\, N_{ij}\right)\right]^{-1/2} \exp\left[-\frac{1}{2} r_i\, N_{ij}^{-1}\, r_j\right]. \tag{49}$$

Because TOAs are evaluated at discrete times t_i, we adopt the notation that $r_i \equiv r(t_i)$. Here N_{ij} is a 'noise covariance matrix'[25]:

$$N_{ij} = {\rm cov}(n_i, n_j) \equiv \left\langle [n_i - \langle n_i \rangle][n_j - \langle n_j \rangle] \right\rangle = \langle n_i\, n_j \rangle - \langle n_i \rangle \langle n_j \rangle. \tag{50}$$

In practice, the timing model is typically first calculated for each individual pulsar alone, ignoring any GW components because $h(t) \ll \langle \Delta t \rangle / T$. This can be done with 'maximum likelihood' estimates determined by maximizing Eq. 49, or equivalently the 'log likelihood', $\ln\left[p(r|\Delta t)\right]$. Alternatively, full probability distributions ('posteriors') for timing parameters can be determined using Bayes' theorem, in this context:

$$p(\alpha|\delta t, \Upsilon) = \frac{p(\delta t|\alpha, \Upsilon)\, p(\alpha|\Upsilon)}{p(\delta t|\Upsilon)}, \tag{51}$$

where we have made explicit that this requires a model Υ. One significant benefit of the latter approach is that uncertainties in the timing model parameters can be marginalized over, at times analytically. In the Bayesian framework, multiple possible models can be compared with the 'odds ratio',

$$O_{\mu\nu}(\delta t) \equiv \frac{p(\Upsilon_\mu|\delta t)}{p(\Upsilon_\nu|\delta t)} = \frac{p(\delta t|\Upsilon_\mu)}{p(\delta t|\Upsilon_\nu)} \frac{p(\Upsilon_\mu)}{p(\Upsilon_\nu)}. \tag{52}$$

This provides a quantitative means of determining what components should be included in a given timing model. Both the maximum likelihood and Bayesian methods yield a timing model, and can give us a first estimate of the timing model parameters: $\alpha^{\rm noise}$.

The GW signals we search for are small, so we can generally assume that the timing model parameters in the presence of GWs, $\alpha^{\rm GW}$,

[23] The instrument noise is often expressed as a product of a 'system temperature' and 'instrument gain': $(T_{\rm radio}/50\ {\rm K}) = (S_{\rm radio}/10\ {\rm J}) \cdot (G_{\rm radio}/5\ {\rm K\ Jy^{-1}})$. In radio astronomy the 'Jansky' unit of spectral flux density is sadly common, $1\ {\rm Jy} \equiv 10^{-23}$ erg s^{-1} cm^{-2} Hz^{-1} $= 10^{-26}$ J s^{-1} m^{-2} Hz^{-1}.

[24] Particular effects are often grouped in either 'noise' or 'deterministic' depending on the particular formalism. Ultimately the requirements is that the noise components can be described by a noise covariance matrix (described below).

[25] The term 'noise correlation matrix' is often also used in the literature. The difference is only in their normalizations: ${\rm corr}(x,y) = {\rm cov}(x,y)/\sigma_x\sigma_y$, but for likelihood calculations this has no effect, so the two types of matrices can be used interchangeably.

can be found with a slight perturbation ($\delta\alpha$) from their values assuming noise alone:

$$\Delta t\left(\alpha^{\rm GW}\right) \approx \Delta t\left(\alpha^{\rm noise}\right) + M\delta\alpha, \tag{53a}$$

$$M \equiv \left.\frac{\partial \Delta t}{\partial \alpha}\right|_{\alpha=\alpha^{\rm GW}}, \quad \delta\alpha \equiv \alpha^{\rm GW} - \alpha^{\rm noise}. \tag{53b}$$

Here M is usually expressed as an $N_{\rm TOA} \times N_{\rm par}$ matrix for each pulsar, called the 'design matrix', which gives the dependency of each TOA ($t_i; i \in [0, N_{\rm TOA}-1]$) on each timing-model parameter ($\alpha_j : j \in [0, N_{\rm par}-1]$), i.e. $M_{ij} \equiv \partial \Delta t_i / \partial \alpha_j$. This is the 'linearized' timing model which is drastically faster to optimize and can often be performed (semi-)analytically.

3.2 Noise Sources

For additional details, we refer the reader to Backer and Hellings (1986), Cordes and Shannon (2010) and Condon and Ransom (2016, Ch. 6).

Constructing a noise model, and performing pulsar timing more broadly, is actually quite complicated. Individual pulses are usually far too faint to detect (Eq. 46), so large numbers need to be stacked to make a measurement. Intrinsically, pulses are typically far from delta functions, with substantial substructure. That substructure can vary over time, sometimes stochastically, which is sometimes referred to as 'jitter', sometimes with systematic trends, and sometimes abruptly. Even a delta-function pulse is subject to the orbit of the pulsar if it is in a binary (which is common), changes in its spin, and propagation effects through the ionized 'inter-stellar medium (ISM)'—particularly 'dispersion' and 'scintillation' which are both strongly radio-frequency dependent.

Typically, the roughly-dozen different physical components of noise can be treated as independent, additive components, such that we can write:

$$\Delta t = \Delta t_{\rm DM} + \Delta t_{\rm sc} + \Delta t_{\rm spin} + \Delta t_{\rm bary} + \Delta t_{\rm rel} + \ldots. \tag{54}$$

These timing model components can be incorporated either in the time-domain or in the frequency/Fourier domain. Timing components in the frequency basis can still be part of the 'linearized' timing-model because the *coefficients* are the model parameters of interest:

$$\Delta t_j^{\rm freq} = F_{jk}^{\rm freq} \alpha_k^{\rm freq} = \sum_{k=0}^{N_f} \alpha_k^{\rm freq} \exp\left(-2\pi i k t_j / T_{\rm obs}\right), \tag{55}$$

where $F_{jk}^{\rm freq}$ is the timing model for Fourier-basis frequency k at time t_j and $\alpha_k^{\rm freq}$ is the Fourier coefficient for frequency k. Many of the timing-model components are periodic, and most of the noise components are stationary (i.e. depending not on individual times, but on time lags: $|t_i - t_j|$), making the Fourier basis particularly convenient. For a much more thorough description of the construction of a timing model including time-domain and frequency-domain components, see Taylor (2021). A subset of important components of the timing model are briefly described here.

- **ISM dispersion**: a plasma propagation effect which delays EM waves in a EM-frequency (ν) dependent manner (Draine, 2011):

$$\Delta t_{\rm DM}(\nu) \approx \frac{e^2}{2\pi m_e c}\frac{D_m}{\nu^2} \qquad D_m \equiv \int_0^L n_e(l)\, dl \tag{56a}$$

$$\approx 4.6\,{\rm s}\left(\frac{D_m}{100\,{\rm pc\,cm^{-3}}}\right)\left(\frac{\nu}{300\,{\rm MHz}}\right)^{-2}. \tag{56b}$$

  Measurements at multiple radio frequencies are required to determine the 'Dispersion Measure (DM; D_m)', which is the column-density of free electrons from the observer to the source, and then convert measured arrival times to some reference frequency, often taken as $\nu \to \infty$. DM is particularly challenging to model as it has substantial spatial and temporal structure that can be common across many different pulsars due to the solar wind.

- **Scintillation / scattering**: Scintillation or scattering is the accumulation of phase changes (or path changes) due to interaction with localized over-densities in the ISM (often modeled as clumpy 'screens'). The frequency dependence of scintillation depends on the density spectrum of the scattering material, but is typically steeper than dispersion with a characteristic power-law index of ≈ 4. For example, for a Kolmogorov spectrum[26] (Rickett, 1990):

$$\Delta t_{\rm sc}(\nu) \approx 10^{-6}\,{\rm s}\left(\frac{C}{10^{-17}\,{\rm cm^{-20/3}}}\right)^{6/5}\left(\frac{\nu}{300\,{\rm MHz}}\right)^{-22/5}\left(\frac{D}{1\,{\rm kpc}}\right)^{11/5}. \tag{57}$$

  The normalization (C) and slope of actual density power-spectrum must be measured from observations and can vary significantly. Additional effects depend on the relative velocities of the observer, source, and scattering screen and thus vary systematically-in-time with additional structure.

- **Pulsar evolution**: pulsars tend to spin down due to loss of spin energy, including from dipole-like emission from their large magnetic fields. Spin changes are typically parameterized in terms of time-derivatives of the spin period, $\dot{P}$, $\ddot{P}$, etc. For magnetic dipole radiation,

[26] i.e. a PSD with a size-scale power-law index of $-5/3$ for energy, or $-11/3$ for density.

the $\dot{P}$ effect on timing is,

$$\Delta t_{spin} \approx \dot{P} \cdot T_{\rm obs} \approx 1.9 \times 10^{-5}\,{\rm s} \left(\frac{B}{10^{12}\,{\rm G}}\right)^2 \left(\frac{R}{10\,{\rm km}}\right) \left(\frac{T_{\rm obs}}{10\,{\rm yr}}\right) \left(\frac{M_p}{2\,{\rm M}_\odot}\right)^{-1} \left(\frac{P}{10\,{\rm ms}}\right)^{-1}, \tag{58}$$

where B, R, M_p, P are the pulsar magnetic field, radius, mass, and spin period, respectively. Additionally, astrometric motion on the sky, and binary motion[27] must also be accounted for.

- **Solar-system motion**: The motion of the Earth (observatory) contributes significantly to TOA measurements. To make comparisons across time (decades) and space (different observatories on Earth, and orbital positions), pulse TOAs are typically moved to the 'solar-system barycenter' reference frame which can be taken as an inertial frame over long periods of time. This requires accounting for Earth's position ($\approx 500\,{\rm s}$; the 'Roemer delay') and motion ($\approx 10^{-6}\,{\rm s}$) but also that of the other planets using measured ephemerides (Vallisneri et al., 2020). For example, if the uncertainty in Jupiter's orbit is $\delta x_{\rm Jup} \approx 100\,{\rm km} \approx 10^{-3}\,R_{\rm Jup}$, then the added uncertainty in the solar-system barycenter position's light-travel time is,

$$\Delta t_{\rm bary} \approx 3.2 \times 10^{-7}\,{\rm s} \left(\frac{M_{\rm Jup}}{{\rm M}_\odot}\right) \left(\frac{\delta x_{\rm Jup}}{100\,{\rm km}}\right). \tag{59}$$

- **Relativistic effects**: As pulses pass near massive objects (e.g. binary companions or our Sun), their propagation is affected by the 'Shapiro Delay': the added propagation time across curved space-time. For a source at an angle θ to a massive object, the Shapiro delay is (Backer and Hellings, 1986):

$$\Delta t_{\rm rel} \approx -\frac{2GM}{c^3}(1-\cos\theta) \approx -9.8 \times 10^{-6}\,{\rm s} \left(\frac{M}{{\rm M}_\odot}\right)(1-\cos\theta). \tag{60}$$

  Additional relativistic effects can also be important. For example, for very small angular separations, gravitational lensing can be larger than the Shapiro delay; and motion of the pulsar and Earth in their gravitational potentials changes their gravitational redshifts ('Einstein delays').

An epoch of observations for a single pulsar will typically be tens of minutes long, collecting $\sim 10^4 - 10^5$ pulses, measuring the DM, scattering, and other measurable noise processes and then constructing a single average/effective pulse arrival time t_i along with a measured uncertainty σ_i. After subtracting the timing model (Eq. 48), we are left with any GW signals in addition to unmodeled noise—both intrinsic noise and residuals due to imperfect fitting of the deterministic components. The noise typically has a component that is 'white', having uniform power with respect to signal-frequency f, and also significant 'red'-noise with more power at lower frequencies. Different pulsars are observed to have very different red-noise characteristics, often referred to as 'spin noise', which is typically modeled as a power-law spectrum.

3.3 The Optimal Statistic & GW Searches

For additional details, we refer the reader to Anholm et al. (2009), Romano and Cornish (2017) and Taylor (2021).

We saw in Sec. 2 that a distinct feature of GWs is the correlations between detector data-streams. For PTAs, these are the Hellings-Downs correlations (Eqs. 20; Fig. 2) which depend only on the angular separation between pairs of pulsars on the sky. It is then natural to utilize these correlations as a method, or as a 'statistic' to detect them. In this context, a **detection statistic** is a quantity that can be calculated from the measured data, which can be directly related to our confidence in a signal being present—for example a signal-to-noise ratio (SNR) or Bayes factor.

Let us express our pulsar TOAs as including a GW signal component s, a timing model that approximates deterministic physical effects $\Delta t \approx \overline{\Delta t}$ (Eq. 54), and a noise component $n(t)$:

$$\delta t(t) = \overline{\Delta t}(t) + s(t) + n(t), \tag{61}$$

$$r(t) \equiv \delta t - \Delta t \approx s(t) + n(t). \tag{62}$$

Consider two different data streams of residuals, $r_i(t)$ and $r_j(t)$, from two different pulsars. We can construct the cross-correlation (Eq. 14), and consider it's expectation value at zero lag[28]:

$$\langle \rho_{ij} \rangle = \langle s^2 \rangle + \langle s\, n_i \rangle + \langle s\, n_j \rangle + \langle n_i\, n_j \rangle \approx \langle s^2 \rangle. \tag{63}$$

The approximation is valid as long as any non-GW correlations between two pulsars have correctly been subtracted in the timing models, and thus the strains and both noise streams are uncorrelated[29]: $\langle s\, n_i \rangle \approx \langle s\, n_j \rangle \approx \langle n_i\, n_j \rangle \approx 0$. If our data streams are pulsar redshifts, then we can identify Eq. 63 with Eqs. 20, i.e. $\langle \rho_{ij} \rangle_{i \neq j} = \langle |h_E|^2 \rangle_T\, \mu_{\rm HD}(\gamma) = P_h\, \mu_{\rm HD}(\gamma)$. The cross-correlation gives us the GW power, filtered by the detector response: the Hellings-Downs correlations.

To make estimates from data, it is convenient to construct a likelihood function. Let us again consider two, discretely sampled data

[27] Most millisecond pulsars are also in binaries (or higher-order multiples) as their short spin-periods generally require *spinning up* through mass-transfer from a binary companion.

[28] As mentioned previously: for CW signals, the pulsar terms (Eq. 11) can be recovered by considering time-lags matching the light-travel times to each pulsar.

[29] Noises and strains are un-correlated both in time and ensembles: $\lim_{T\to\infty} \rho_{ij} = \langle s^2 \rangle_T$, and also $\langle \rho_{ij} \rangle = \langle s^2 \rangle_E$. However, $\langle s^2 \rangle_T \neq \langle s^2 \rangle_E$, because our 'ensemble' is composed of different population realizations, in which case the GWs are not ergodic. Variances will also differ in the two cases.

streams $r_{1i} \equiv r_1(t_i)$, and $r_{2i} \equiv r_2(t_i)$ each with $N_{\rm samp}$ samples[30], which we concatenate into a single array $r_i = \{r_{11}, r_{12}, \ldots, r_{1N_{\rm samp}}, r_{21}, r_{22}, \ldots, r_{2N_{\rm samp}}\}$. We can write the likelihoods in the presence of a signal, and in the presence of only noise as:

$$p(r|H) = \mathrm{Det}(2\pi H)^{-1/2}\,\mathrm{Exp}\left[-\frac{1}{2} r_i\, H^{-1}_{ij}\, r_j\right], \qquad p(r|N) = \mathrm{Det}(2\pi N)^{-1/2}\,\mathrm{Exp}\left[-\frac{1}{2} r_i\, N^{-1}_{ij}\, r_j\right]. \tag{64}$$

Here the covariance matrices are, $H_{ij} = \langle (s_i + n_i)\,(s_j + n_j)\rangle$ with GWs, and $N_{ij} = \langle n_i\, n_j\rangle$ with only noise. If we again assume that the noise is uncorrelated in time and between pulsars, and further that the GW signal is uncorrelated in time, these matrices will be block diagonal and diagonal respectively,

$$H = \begin{bmatrix} \begin{bmatrix} S_{n_1} + S_s & 0 \\ 0 & \ddots \end{bmatrix} \begin{bmatrix} S_s & 0 \\ 0 & \ddots \end{bmatrix} \\ \begin{bmatrix} S_s & 0 \\ 0 & \ddots \end{bmatrix} \begin{bmatrix} S_{n_2} + S_s & 0 \\ 0 & \ddots \end{bmatrix} \end{bmatrix}, \qquad N = \begin{bmatrix} \begin{bmatrix} S_{n_1} & 0 \\ 0 & \ddots \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \ddots \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \\ 0 & \ddots \end{bmatrix} \begin{bmatrix} S_{n_2} & 0 \\ 0 & \ddots \end{bmatrix} \end{bmatrix}. \tag{65}$$

Each of the four 'blocks' of these matrices are shaped $N_{\rm samp} \times N_{\rm samp}$. The power in the signal and in the noise of each pulsar is: $S_h = \langle s^2\rangle$, $S_{n_1} = \langle n_1^2\rangle$, and $S_{n_2} = \langle n_2^2\rangle$. The likelihoods can be analytically maximized to find the maximum likelihood estimates of the powers: $\hat{S}_h = \hat{S}_{12}$, $\hat{S}_{n_1} = \hat{S}_{11} - \hat{S}_{12}$, and $\hat{S}_{n_2} = \hat{S}_{22} - \hat{S}_{12}$, where

$$\hat{S}_{12} = \frac{1}{N}\sum_{i=1}^{N} r_{1i}\, r_{2i}, \qquad \hat{S}_{11} = \frac{1}{N}\sum_{i=1}^{N} r_{1i}^2, \qquad \hat{S}_{22} = \frac{1}{N}\sum_{i=2}^{N} r_{2i}^2, \tag{66}$$

We can also analytically determine the maximum-likelihood ratio,

$$\Lambda_{\rm ML} \equiv \frac{\mathrm{Max}[p(r|H)]}{\mathrm{Max}[p(r|N)]} = \left[1 - \frac{\hat{S}_h^2}{\hat{S}_1\,\hat{S}_2}\right]^{-N_{\rm samp}/2}. \tag{67}$$

It is often convenient to define another statistic to be twice the logarithm of the maximum-likelihood ratio which, in the weak signal limit, becomes the SNR of the cross-correlation:

$$\Lambda \equiv 2\,\ln(\Lambda_{\rm ML}) \approx \frac{N\,\hat{S}_h^2}{\hat{S}_{n_1}\,\hat{S}_{n_2}}. \tag{68}$$

Typically noise will, however, be correlated in time. In this case the likelihoods (Eq. 64) cannot be maximized analytically. However, an optimal statistic can still be calculated numerically and applied directly to sets of TOAs from an arbitrary number of pulsar pairs (see Ch. 7 of Taylor, 2021). While the noise may be correlated in time, it is typically stationary, such that it is uncorrelated in frequency space. In that case, a frequency-domain version of Eq. 63 will still hold with, $\langle \rho(f)\rangle = \langle \tilde{r}_1(f)\,\tilde{r}_2^*(f)\rangle \approx \langle \tilde{s}(f)\,\tilde{s}^*(f)\rangle$, where $\tilde{r}^*$ is the complex conjugate of the Fourier transform $\tilde{r}(f) \equiv \mathcal{F}(r(t))$, given in Eq. 16. In this case, the frequency-space covariance matrices look very similar to Eq. 65, and analytic solutions analogous to Eq. 67 and Eq. 68 can again be found (see Sec. 4.3 of Romano and Cornish, 2017).

We wish to construct an '**optimal statistic**' ($S_{\rm opt}$) defined to maximize the SNR. Following Anholm et al. (2009, Sec. III)[31], we define the SNR as $\mathrm{SNR}^2 = \langle S_{\rm opt}(\mathrm{GW})\rangle^2 / \langle S^2_{\rm opt}(\mathrm{noise})\rangle$, which is the power of the statistic in the presence of GWs, divided by the variance of the statistic in the absence of GWs, such that $\mathrm{Var}\big(S_{\rm opt}(\mathrm{noise})\big) = \langle S^2_{\rm opt}(\mathrm{noise})\rangle - \langle S_{\rm opt}(\mathrm{noise})\rangle^2 = \langle S^2_{\rm opt}(\mathrm{noise})\rangle$. This definition of the SNR is appropriate for weak GW signals, i.e. when the SNR is small. The optimal statistic can then be calculated as,

$$S_{\rm opt} = \sum_{i=1}^{j}\sum_{j=1}^{N_p}\sum_{k=1}^{(N_f)_{ij}} \frac{\Gamma_{ij}\, S_{h0}(f_k)}{P_i(f_k)\,P_j(f_k)}\, s_{ijk}. \tag{69}$$

The summations should be followed from the inside out, where the number of frequency bins for a particular pair of pulsars is $\big(N_f\big)_{ij}$, and the number of pulsars is N_p. To simplify the notation, we will replace the triple summation in Eq. 69 with $\sum_{ijk}$. The filtered and noise-weighted cross-correlation between pulsars i and j, at a frequency f_k is,

$$s_{ijk} = 2\int_0^{+\infty} \delta_T(f_k - f')\,\tilde{s}_i(f_k)\,\tilde{s}_j(f')\,\frac{S_h(f_k)\,\Gamma_{ij}}{P_i(f_k)\,P_j(f')}\,df'. \tag{70}$$

This optimal statistic is also a 'matched filter', comparing the data to a template signal spectrum S_{h0}. The Fourier transform of the window function, or the approximation to a delta function for a finite time-span $T_{\rm obs}$, is:

$$\delta_T(f) \equiv \frac{\sin(\pi f T_{\rm obs})}{\pi f}. \tag{71}$$

[30] In practice, some sort of interpolation or binning must be performed to get samples at matching times, unless we are working in frequency-space (see below).
[31] See also Rosado et al. (2015), whose notation we employ.

The mean statistic in the presence of a GW signal is,

$$\langle S_{\rm opt}({\rm GW})\rangle = \sum_{ijk} \frac{\Gamma_{ij}^2\, S_h(f_k)\, S_{h0}(f_k)}{P_i(f_k)\, P_j(f_k)}, \tag{72}$$

and the variance in the absence of a GW is,

$$\langle S_{\rm opt}^2({\rm noise})\rangle = \sum_{ijk} \frac{\Gamma_{ij}^2\, S_{h0}^2(f_k)}{P_i(f_k)\, P_j(f_k)}. \tag{73}$$

For a (correctly) matched filter, i.e. $S_{h0} = S_h$, the SNR is then $\langle S_{\rm opt}({\rm GW})\rangle^{1/2}$.

4 Massive Black Hole Binaries

As introduced in Sec. 1, **massive black hole binaries (MBHBs)** have long been proposed as sources of the loudest gravitational waves in the Universe. MBHs, when accreting sufficient amounts of gas, are visible as '**active galactic nuclei (AGN)**', with many quasars (the brightest examples of AGNs) observed from even the very distant and early Universe. More locally, where the kinematics of gas and/or stars can be resolved in the central $\sim$ pc of galaxies, dynamical modeling demonstrates the presence of MBHs in galactic nuclei even when they're inactive. Observations are consistent with virtually all galaxies hosting an MBH in their center. MBH masses are observed to be tightly correlated with host-galaxy properties (Kormendy and Ho, 2013). Many empirical relationships have been derived, with the MBH-mass vs. stellar velocity-dispersion relation ($M_{\rm BH}$–$\sigma_\star$) typically regarded as the most precise, but the MBH-mass vs. stellar bulge-mass relation ($M_{\rm BH}$–$M_{\rm bulge}$) often being the most convenient. MBH masses in the local Universe ($z \approx 0$) are well described by a log-normal distribution with mean and standard deviation (*Ibid.*),

$$\log_{10}\!\left(\frac{M_{\rm BH}}{\rm M_\odot}\right) \sim \mathcal{N}\!\left(\mu + \alpha_\mu\, \log_{10}\!\left(\frac{M_{\rm bulge}}{10^{11}\,\rm M_\odot}\right), \epsilon_\mu\right) \tag{74}$$

$$\mu \approx 8.69 \pm 0.05, \qquad \alpha_\mu \approx 1.17 \pm 0.08, \qquad \epsilon_\mu \approx 0.28. \tag{75}$$

An effective heuristic is that MBH masses are 200 times less than the galaxy stellar mass[32], with a standard deviation close to a factor of two.

Structure formation in the Universe is fundamentally hierarchical in nature, with massive galaxies growing through the merger of many smaller galaxies. While galaxy mergers are abundantly identified in galaxy surveys, the difficulty of associating a lifetime to those mergers makes the calculation of a galaxy merger-rate non-trivial. Typical merger rate estimates are either calculated directly from cosmological hydrodynamic simulation, or utilize their merger timescales to normalize observational counts of galaxy pairs. A comprehensive fit to simulated galaxy merger rates are provided by Rodriguez-Gomez et al. (2015). For parameters near those most important for PTAs, their results can be *very-roughly* approximated as:

$$\frac{d^2N}{dq\,dz} = 0.92\,{\rm Gyr}^{-1} \left(\frac{M}{6\times10^{11}\,\rm M_\odot}\right)^{0.41} \left(\frac{q}{1/4}\right)^{-1.3} \left(\frac{1+z}{2.0}\right)^{2.0}. \tag{76}$$

Most massive galaxies then have the opportunity to host multiple pairs of MBHs over their cosmic histories.

Galaxy mergers bring two MBHs into a common, post-merger host galaxy, with a separation of $a \sim 10^3$ pc. Based on the typical central densities of massive galaxies, it is only at separations of $a \lesssim 10$ pc that the two MBHs will enter their mutual 'sphere of influence' and become gravitationally bound as a true 'binary'. As binaries emit GWs, the system loses energy, and the semi-major axis a decreases. Recall that it is only at $a \lesssim 1$ pc before GW-emission alone is sufficient to drive binaries to coalesce: the 'coalescence time' $\tau_{\rm life} \equiv \int (df/dt)^{-1} df \sim 1 - 10\,{\rm Gyr}$ (from $\sim$parsec separations; Eqs. 26). The separations corresponding to PTA-sensitive frequencies are only slightly smaller, $a \lesssim 10^{-2}$ pc (Eq. 1a); and thus, after galaxy merger, binaries need to traverse $\sim$five orders of magnitude in separation to produce detectable GW emission. Recall also that once in the PTA band, the binary '**hardening timescale**' is $\tau \equiv f/(df/dt) \sim 10^5 - 10^6$ yr (Eqs. 25).

These characteristic time scales already tell us a good deal about MBHB populations. First, because the time-scale between galaxy mergers is comparable to the time-scale of binary coalescence, a typical massive galaxy can host of order unity dual-/binary- MBHs at any given time! Due to the steep lifetime scaling ($\tau_{\rm life} \propto a^4$), there are drastically fewer binaries at smaller separations. For PTA-band MBHBs specifically, we expect $\lesssim 10^{-3}\, f_{\rm coal}$ galaxy^{-1} hosting binaries in the low-redshift Universe. Here, $f_{\rm coal} = f_{\rm coal}(M, q, z, \dots)$ is the coalescing fraction—the fraction of binaries which will coalesce before redshift zero (discussed more below), which is nearly identical to the fraction of binaries reaching PTA-frequencies. The fraction of AGN containing binaries is more difficult to estimate. While AGN are known to be triggered by mergers, the difference between typical AGN lifetimes and the delay time between AGN activity and binaries reaching the PTA band are unconstrained[33]. However, if we assume that activation and reaching-the-PTA-band are uncorrelated, and that AGN lifetimes are $10^7 - 10^8$ yr (or $f_{\rm AGN} \sim 10^{-2}$, e.g. Hopkins et al., 2008, and references therein), we can estimate that $\lesssim 10^{-5}\, f_{\rm coal}$ AGN^{-1} host PTA-band binaries in the low-redshift Universe.

[32] The fraction of stellar mass in the stellar bulge, the 'bulge fraction' $f_{\rm bulge}$, varies by a factor of a few, but for the high-mass MBHs of interest to PTAs, $f_{\rm bulge} \gtrsim 0.7$

[33] Fast mergers may preferentially produce binaries in AGN, but it's also possible that binaries reach the PTA band specifically after AGN activity tends to cease.

4.1 MBH Binary Evolution

For additional details, we refer the reader to Begelman et al. (1980), Yu (2002), Milosavljević and Merritt (2003a), and Kelley et al. (2017a).

A variety of '**environmental interactions**' between the binary and components of the host galaxy are required to bring the two MBHs from galaxy scales to the post-merger galactic nucleus, and then the GW regime. The process for MBH binary evolution was first outlined by Begelman et al. (1980). At large separations ($a \sim 10^2 - 10^4$ pc), 'dynamical friction' governs the galaxy-galaxy merger itself, and the early inspiral of the two MBHs in the post-merger host galaxy. Near the spheres of influence of the MBHs ($a \sim 10$ pc), where they become gravitationally bound as true 'binaries', the dynamical friction formalism breaks down and individual three-body 'stellar scattering' events must be considered instead. In gas-rich mergers, which are more common in lower-mass galaxies, a '**circumbinary accretion disk**' can form around the binary at similar scales $a \sim$ pc, and torques from this disk can further influence the binary evolution. Finally, as shown above, GW emission becomes dominant at the smallest separations ($a \lesssim$ pc) leading to eventual binary coalescence (described in Sec. 2.3). The broad picture of this evolution is still believed to hold true, with at least some fraction of binaries able to reach the GW-driven regime, and eventually coalesce. The details in every phase, however, remain highly uncertain, as are the distribution of typical binary lifetimes and coalescing fractions.

4.1.1 Dynamical Friction

In the dynamical friction regime, both MBHs are initially fully embedded within their host galaxies. In the early stages, the dominant mass-constituents are the two dark-matter halos, although stars and sometimes gas are non-negligible. Conceptually, and in simplified calculations, we often consider the secondary MBH/galaxy inspiraling into the halo/galaxy of the primary. The standard, idealized model of dynamical friction developed by Chandrasekhar (1943) considers a point-mass moving through a uniform, gravitating background that then becomes perturbed. This model wouldn't seem to hold at all: the two galaxies are comparable in size, and thus fundamentally extended; similarly, there is no uniform background of material; and galaxy mergers are typically highly disruptive.

Oddly then, the simplified formalism of dynamical friction still agrees fairly well with detailed simulations. Two key considerations are, however, important. First, accounting for 'stripping' of the secondary galaxy: as it inspirals, the furthest-out and least-bound material is gradually removed and deposited into the outer regions of the primary galaxy. This is induced both by tidal forces and ram-pressure ablation from the head wind that the secondary experiences. After a few dynamical times, the secondary galaxy will be stripped from the secondary MBH, leaving only a tightly bound core of stars and gas. Second, accounting for the multi-component radially-stratified density and velocity profiles of the primary galaxy. As the secondary transitions from the primary's halo to the galaxy proper, the effective 'background' material transitions from dark matter to stars and gas. More generally, the density and relative velocities of each components change significantly with radius which must be accounted for.

It is expected that the dynamical friction phase is not always successful in that the secondary MBH can '**stall**' in the outskirts of the primary galaxy, typically at $\sim$ kpc scales. This is more common in more-extreme mass-ratio mergers ($q \lesssim 10^{-1}$) and/or less penetrating initial orbits (initial pericenter distances $\gtrsim 1$ kpc), which produce the satellite/dwarf galaxies and tidal streams seen in the Milky Way. In these cases, the secondary never reaches sufficiently high densities, and/or tidal stripping decreases the secondary mass too quickly. The resulting 'wandering' or 'orphan' MBHs in galaxy outskirts would be difficult to detect or constrain even within the Milky Way as these MBHs are far less likely to be accreting. Lensing surveys of the halo, and possibly 'tidal disruption events' (TDEs; described below) are likely the best channels.

4.1.2 Stellar Scattering

Once the energy of the two MBHs becomes comparable to the local stellar background, typically soon after they form a bound binary, the dynamical friction formalism breaks down[34] Eventually, individual three-body 'stellar scattering' events must be considered. The parameter space of stars that are able to interact with the MBHB is called the 'loss cone' (LC), this includes stars with a range of energies (semi-major axes), and small angular momenta (high eccentricities). Scattered stars tend to extract energy from the binary, allowing it to continue hardening, while those stars are then ejected from the LC. For every e-folding of binary MBH separation, a mass of stars roughly equal to that of the MBHB must be scattered. For a $\sim 10^9\,M_\odot$ binary, a scattering rate of $\gtrsim 1 - 10\,\mathrm{yr}^{-1}$ must be sustained for $\sim$ Gyr to merge the system. The population of ejected stars can thus be a substantial fraction of a galaxy's stellar core. The resulting 'core scouring' may be observable, and indeed observed, as the flattened or 'cored' stellar density profiles in some massive galaxies.

The efficiency of stellar scattering in merging the MBHB is determined by two effects: the maximum number of stars in each particular galaxy's LC, and how effectively the LC can be replenished as stars are ejected. The challenge to refill the LC, and the subsequent stalling of MBHBs in the stellar scattering regime, is referred to as the '**final-parsec problem**' (Milosavljević and Merritt, 2003b). In idealized models, particularly spherically-symmetric ones, the rate at which LC stars are replenished is very slow: governed by two-body diffusion at the edge of the LC. In this case, the MBHB inspiral will often stall. Over the last decade, a general consensus has emerged that more realistic galaxies have strongly asymmetric, 'tri-axial' gravitational potentials (particularly following galaxy mergers) which act to stir the stellar phase-space distribution. This can rapidly refill the LC to an effectively-full steady-state. In this case, the stellar scattering rate can remain high, and in many cases the binary will continue to inspiral effectively.

Despite this growing consensus, stellar scattering remains very difficult to model. Scattering events themselves require resolving $\lesssim pc$ scale sizes, and $\lesssim$ yr timescales, while the refilling of the loss-cone depends on the galaxy at $\sim$ kpc sizes and $\gtrsim 10^6$yr timescales. Even

[34] At this point the 'back reaction' on the background can no longer be treated as perturbative.

if the LC remains full, the effectiveness of stellar scattering still depends on the central mass, density, and velocity profile of stars which are usually unresolvable in EM observations. Some fraction, possibly substantial, of binaries could still stall at $\sim$ pc separations. If the nanohertz GWB measured by PTAs is confirmed to be produced by MBHBs, it would be the first direct evidence that the final-parsec problem is indeed solved, at least for a large portion of the most-massive MBHBs.

4.1.3 Circumbinary Disks and Torques

[Cross-reference Diego's Circumbinary disk chapter.]

In accretion disks, internal 'viscous' stresses transport angular momentum outwards which allows material to move inwards. Around compact objects, the viscous stresses are produced by the 'magneto-rotational instability' (MRI). When a second, massive object is added as a binary companion within an accretion disk, three distinct regions form, depicted in Fig. 4(a). At large radii, much larger than the binary separation ($r \gg a$), the '**circumbinary disk**' (grey) behaves as if it were around a single-object, now with the combined mass of both components. At small radii surrounding each component—specifically well-within each's Hill surface ($r \lesssim r_{\rm hill}$)—two 'circumsingle disks' behave like perturbed single accretion disks around the primary (red) and secondary (blue). At intermediate radii ($r_{\rm hill} < r \lesssim 2a$), orbiting material is no longer stable and a 'gap' is cleared.

The overall accretion rate through the circumbinary disk and through both circumsingle disks is relatively unaltered by the presence of the gap. An over-density tends to form at the inner-edge of the circumbinary disk (Fig. 4(a); darker grey), which then overflows and feeds the circumsingle disks through streams. Even when the binary system begins to inspiral rapidly due to GW emission, more rapidly than viscous torques within the accretion disks can feed material, the accretion continues to be driven by dynamical torques. Thus, the current consensus is that accretion will continue until nearly the time of coalescence[35], which offers promising opportunities for EM counterparts throughout the inspiral process (Sec. 4.2).

While the net accretion rate is unchanged by the presence of the binary, the amount of material accreting onto each component can be highly variable. Specifically, the accretion rate is often periodically modulated at the binary orbital period, and/or the orbital period of the inner-edge of the circumbinary disk (typically $\approx 5-10$ times longer). Simulations show that accretion tends to favor the lower-mass secondary, at times by large factors. For some orbital configurations, particularly near-equal-mass binaries, this 'preferential accretion' can alternate between the two components on secular timescales of hundreds to thousands of orbits. The detailed sharing of material seems to be affected by not only the mass-ratio and eccentricity, but also the thermal structure of the disk, all topics of ongoing study.

The biggest surprise in the recent study of circumbinary disks lies in the gravitational interaction between the disk(s) and binary. It was long believed that circumbinary disks would always extract angular momentum from binaries, hastening their inspiral. A classic derivation common in planetary dynamics literature is to calculate torques from Lindblad (spiral wave) resonances that result from a first-order expansion of the binary gravitational potential. Those torques on the binary are negative, extracting angular momentum. A large number of hydrodynamic simulations, starting with Miranda et al. (2017), have now shown that this is not always the case. The density distribution, particularly in the circumsingle disks, from a full solution to the equations of motion are such that the net torques can be positive: depositing additional angular momentum into the binary, acting against their inspiral (Muñoz et al., 2019).

The implications of this paradigm shift are still being studied. As for the partitioning of accretion material, the detailed torque balance appears to depend on the binary parameters and also the thermal structure of the accretion disks. Both theoretical and observational arguments suggest that in many situations (for example in extreme mass-ratios) the torques should indeed be negative, or otherwise relatively small. Three-dimensional, non-idealized (e.g. radiative and/or magnetic) accretion disk simulations are currently only computational feasible for relatively small numbers of orbits. Similarly, large explorations of parameter space remain intractable. While the picture is still evolving, there are two reasons to expect that disk 'softening' is less important for the highest-mass MBHBs that PTAs are most sensitive to (Bortolas et al., 2021). First, high accretion rates are required for disk densities to become dynamically important, and the highest-mass MBHs tend to have lower accretion rates. Second, if such high accretion rates are achieved and maintained, the MBHBs may grow sufficiently to enter the GW-dominated regime and still coalesce effectively.

4.1.4 Other Effects

The most important environmental processes for PTA-detectable MBHB evolution are dynamical friction and stellar scattering. Circumbinary disk torques may also play a role, but are likely less important for the highest-mass MBHBs that PTAs are most sensitive to. A number of other processes may be non-negligible in the evolution of some MBHBs.

- **MBH triple interactions**. As we've seen, the expected lifetime of MBHBs is comparable to the time between galaxy mergers. This means that it may be common for a subsequent galaxy merger to deliver a third MBH component. Particularly when the first pair of MBHs stall, triple interactions could proceed. Indeed, more careful analysis suggests that up to tens of percent of MBHBs could be affected by a third MBH. In the context of idealized triple interactions, e.g. as applicable to stellar clusters, the lowest mass component will typically be ejected[36] while hardening a binary of the two heavier components. It has been proposed that this serves as a mechanism to induce stalled binaries to coalesce and thus establish some minimum level of coalescences, and thus GW signals, regardless of uncertainties in environmental interactions (Ryu et al., 2018; Bonetti et al., 2018). These 'strong' triple interactions are likely to occur in some fraction of systems, particularly when all three MBHs have small distances of closest approach. More common will be 'weak'

[35] At some point, possibly as late as the final few orbits, the circumbinary accretion disk will decouple and no longer be able to feed material as rapidly as the binary inspirals.

[36] These ejected triples can produce 'offset' or 'wandering' MBHs. See Sec. 4.2.4.

triple interactions at larger separations, where the galaxy potential and dissipative environmental processes are still relevant. In these cases the dynamics may resemble planetary-like systems. Kozai-Lidov type interactions could still drive mergers through eccentricity excitation, but in many cases they may be damped by the environment, or simply take longer than a Hubble time. These systems are particularly challenging to model realistically and require further study.

- **GW recoils.** GW emission becomes stronger and stronger until binaries coalesce. During the final orbits, GW emission can become asymmetric, particularly when the MBHs have large and misaligned spins. The asymmetric GWs carry momentum, producing a 'recoil' in the merged MBH remnant (Blecha et al., 2016). The spin distributions of single MBHs are extremely uncertain, and those of binaries are completely unconstrained. For typical assumptions, we expect that most recoils will be $\lesssim 100$s kms^{-1}, but some fraction can reach $\gtrsim 10^3$kms^{-1}, ejecting the remnant from the host galaxy. Few, if any, massive galaxies seem to be missing their central MBHs, but no studies have carefully constrained the allowed recoil velocities under these constraints. It remains unclear whether this implies that MBHBs stall, that spins are small or aligned, or that new MBHs can be formed (or brought in) to replace ejected ones.
- **Preferential accretion.** As discussed above, hydrodynamic simulations show that the smaller secondary component tends to accrete more material than the primary. In some cases it can be an order of magnitude higher or more. Additionally, the binary configuration may make it easier to exceed typical Eddington limits as the primary could help to prevent material from being blown off through winds. Combined with the long timescales that MBH binaries may spend at $a \lesssim$ pc scales, this implies that the mass-ratio of the binary, and thus its chirp-mass, could change significantly during inspiral.

4.2 Electromagnetic Counterparts

For additional details, we refer the reader to D'Orazio and Charisi (2023).

Gravitational waves intrinsically probe the central, highest density/mass regions. Electromagnetic (EM) observations, on the other hand, probe the highest temperature gases that, at times, surround those massive objects. Each *messenger* thus provides different types of information about their sources, and is subject to different selection effects and modeling degeneracies. Soon after the first GW detections by LIGO was the first multi-messenger event which firmly demonstrated that having both types of signatures revealed tremendously more than than the sum of each type of observation alone. As millions of MBHs are already observed as bright EM sources, MBHBs also present tantalizing opportunities for '*multi-messenger*' astrophysics.

The majority of MBH growth, particularly at lower masses, is provided by accretion of gas from galactic nuclei. When the accretion rates are high, gas becomes hotter and more dense, and produces observable EM emission as '**active galactic nuclei**' (AGNs; Antonucci, 1993; Netzer, 2015). The emission often spans the EM spectrum, with different types of AGN providing some of the brightest sources of emission in the radio (radio galaxy/quasar, BL Lac objects), optical (quasi-stellar objects, QSOs; quasars; and Seyferts), and X-Ray/γ-Ray (blazars). While the detailed structure and dynamics of AGN accretion flows and emission remains surprisingly unclear, they possess a number of key characteristics. AGN are extremely variable over a wide variety of timescales, in all emission bands. Variability is apparent at timescales as short as roughly the light-crossing time of the event-horizon $\approx 3.0 \times 10^4$ s $\left(M/3 \times 10^9\ \mathrm{M}_\odot\right)$, and on timescales as long as observations exist (many decades). The variability tends to be 'red' in character, i.e. with higher amplitude variations at longer periods, with signs of excess (pseudo-)periodicity. AGN also produce a wide variety of emission lines at a range of radii in their accretion disks, and also their more extended gas distributions. As the orbital velocity increases for material nearer the black-hole, the characteristic width of the emission lines becomes broader and broader. The Narrow Line Region (NLR) is typically at ~ 100 pc; while the broad-line region (BLR) is typically at $\sim 10^{-2}$ pc for optical lines, and up-to the inner edge of the accretion disk for X-Ray lines (e.g. Fe-K-α).

A variety of EM signatures have been suggested to signpost the presence of MBH binaries specifically (discussed below; and shown in Fig. 4). Unfortunately, nearly every EM feature seems to be seen in 'normal' individual AGN as well. A large number of candidate binaries have been identified with these signatures, but significant contamination from false-positives seems necessary.

4.2.1 Dual AGN

The only confirmed examples of two MBHs in the same system are as 'dual AGN/MBHs', where two separate emitting 'cores' are spatially distinguishable in a common system, illustrated in Fig. 3(c). Most are identified in the X-Ray and optical, and thus restricted to relatively large angular separations, corresponding to physical separations of kpc—long before the two MBHs become gravitationally-bound as a binary. These types of observations have clearly demonstrated that MBHs contained in a merging galaxy are much more likely to be highly accreting than those in isolated galaxies. This is consistent with theoretical studies and other galaxy merger observations that show large amounts of gas are funneled into galactic nuclei during the galaxy-merger process. While at first glance this seems promising for the synchronous emission of GWs and bright EM emission, the relative timescales of enhanced accretion and MBHBs reaching the GW-detectable separations remains unclear. If the enhanced accretion rates produce bursts of star formation and AGN activity, both of which then result in feedback which then tends to heat up and eject additional gas, it may also be possible that accretion could be suppressed at the time of GW emission. Post-merger and post-star-burst galaxies are also significantly dustier, which again may make EM observations more difficult.

The closest separation dual-AGN source was identified with radio VLBI observations that resolved a $\sim 10\ pc$ projected-separation pair of MBHs (Rodriguez et al., 2006). It remains unclear if these are close enough to be gravitationally interacting. While radio and sub-mm VLBI allows for drastically better angular resolution, only MBHs in the relatively local Universe are bright enough to observe, significantly limiting the sample of possible host galaxies. Still, searches for dual, or even binary, MBHs using Earth-scale interferometers such as the Event Horizon Telescope (EHT) and next-generation very-large array (ngVLA), is an important endeavour.

Ruling out the presence of a non-accreting (or accreting, but spatially unresolvable) MBH companion to an AGN is surprisingly difficult.

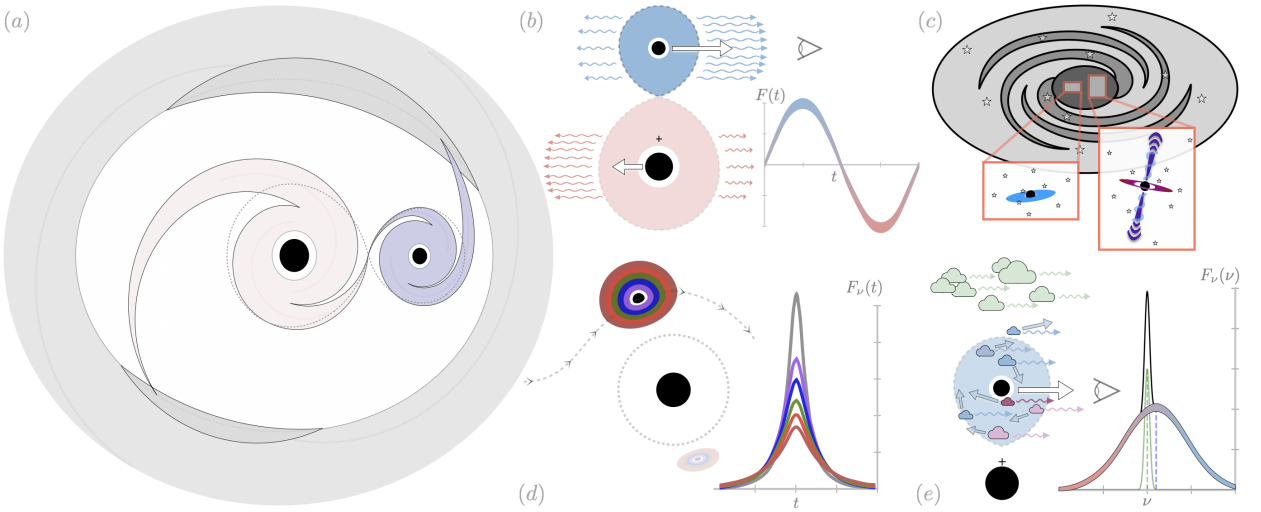

Fig. 4 (a) Accreting binary with circumbinary disk (CBD; grey), and two circumsingle disks (CSDs; red and blue) truncated at Hill surfaces. Streams feed CSDs, preferentially the secondary (blue), from over-densities in the CBD. **(b) Doppler-boosting** induced periodic photometric-variability of secondary MBH. Binaries with orbital planes oriented along the observers line of sight will appear brighter (and bluer) when moving towards the observer, and fainer (and redder) when moving away. **(c) Dual AGN** spatially resolved in post-merger galaxy. Optical and X-ray surveys provide the largest samples, but angular resolution limits most detections to ∼kpc separations. Radio surveys can detected ∼pc binaries, but only those in the nearby Universe. **(d) Gravitational self-lensing** of secondary with accretion disk, producing lensing flare. Thermal stratification of the secondary's disk, combined with a lensing radius (grey dotted) comparable in-size to the disk, can produce a chromatic lensing-flare. **(e) Kinematic/Spectroscopic offsets** from the broad-lines near the secondary (red-to-blue), relative to narrow-lines surrounding the combined binary at larger distances and lower velocities (green).

While MBHBs may produce time-variability signatures (described below), they may also not, and/or those signatures could be on timescales too long to probe. This, in addition to the difficulty in characterizing selection effects in many dual-AGN surveys, makes the incorporation of dual-AGN observations difficult to incorporate in theoretical studies.

4.2.2 Periodic Photometric Variability

Three processes are believed to produce detectable periodic changes in brightness for one or both accreting MBHs that are specifically in a binary system.

- **Binary-induced, hydrodynamic variability.** Simulations of circumbinary accretion, illustrated in Fig. 4(a), show strong periodic modulations of the accretion rate onto one or both components of the binary (Farris et al., 2014). Depending on the detailed configuration of the system, the accretion rate can be predominantly modulated on the orbital period: as one of the MBHs passes near the edge of the circumbinary accretion disk, 'grabbing' material; or at the orbital period of the inner-edge of the circumbinary accretion disk itself (a few times the binary orbital period). If the emission from any of the accretion disks follows the accretion rate changes without too much damping, the overall brightness of the MBHs can retain some degree of periodic modulation. Shocks periodically produces within the inner circumsingle or circumbinary accretion disks may also be observably periodic. More detailed study of the emission produced with variable accretion rates is required to make better predictions for hydrodynamically-induced photometric variability.
- **Gravitational self-lensing.** Binaries oriented nearly edge-on to the observer will have one (or both) components pass in front of the other. The strong gravitational field of the foreground component can then gravitational lense and significantly brighten the emission from the background component (D'Orazio and Di Stefano, 2018), illustrated in Fig. 4(d). Such an event would produce a 'lensing flare' with a duration significantly shorter than the binary orbital period and also typically shorter than much of the typical intrinsic-variability of AGN accretion disks, making them easier to identify. Depending on the detailed configuration of the binary, these lensing events may be chromatic, providing additional means of excluding false positives and additional information about the accretion disk structure.
- **Doppler boosting.** Particularly in unequal-mass binaries, the secondary component can reach mildly-relativistic velocities which can then produce detectable Doppler-induced variations in brightness modulated on the orbital period (D'Orazio et al., 2015), illustrated in Fig. 4(b). Doppler boosting would again preferentially occur in systems that are nearly edge-on. Additionally, because both self lensing and Doppler boosting can be directly connected to the binary orbital phase, these two signatures could at times be identified in tandem.

4.2.3 Kinematic/Spectroscopic Offsets

As described above, the velocity of the MBHs along the observer's line of sight will produce Doppler boosts. In addition to increasing and decreasing the brightness of broadband emission, Doppler shifts will change the frequency/wavelength of emitted spectral lines. Doppler shifted line emission from the circumsingle accretion disks, particularly from the faster-moving secondary, can thus reveal the presence of binary orbital motion. This is illustrated in Fig. 4(e). The optimal signature would be a broad line that is observed to be time-variable, requiring orbital periods shorter than decades. Instantaneously-offset broad lines could also signature the presence of a binary, however

typical broad lines from single AGN are often asymmetric, offset, and time-variable. As for periodic photometric variability this likely explains most of the existing spectroscopically-identified binary candidates. Again, how to filter out false positives remains unclear and an important direction of ongoing research.

Dynamical considerations determine whether a given orbital configuration can plausibly be detected by spectroscopic offsets Kelley (2021). First, the broad-line emitting region must be located at radii that are within the Hill sphere of one or the other binary component. Second, the projected velocity-offset must be larger than the accuracy by which the emission line 'center' can be identified. Instrumentally, this typically requires $v_{\rm orb} \gtrsim 10^3$ km s^{-1}. These two criteria already significantly restrict the viable parameter space of detectable binaries, such that only a small fraction of all binaries would be detectable. Much more challenging, however, is the intrinsic variability of normal AGN broad-lines, both in time and across populations. This implies that a selection criterion closer to $v_{\rm orb} \gtrsim 10^4$ km s^{-1}, and/or requiring time-changing offsets consistent with orbital motion, should be required. Either more stringent criteria makes detections fairly unlikely. Shifted X-Ray broad lines offer a much more promising portion of parameter space, however X-Ray spectroscopic instruments are much more limited in their resolution and sensitivity.

4.2.4 Other possible signatures

A number of other dual/binary signatures have been proposed in the literature which are worth mentioning.

- **Chromatic Deficits**. The presence of a binary within an accretion disk clears out a 'gap' of material, see Fig. 4(a). Due to the temperature stratification of accretion disks, this suggests a corresponding deficit of emission at temperatures corresponding to the binary separation. Such signatures could be masked by: differing thermal structures in the circumbinary vs. circumsingle disks, emission from outside of the plane of the disk, and the presence of general non-thermal emission.
- **Enhanced and double TDEs.** Tidal-disruption events (TDEs) occur when nuclear stars pass within the tidal disruption radius of a black hole. Similar to the stellar-scattering hardening mechanism of binary orbits, the population of stars able to make close-encounter orbits can be depleted over time. However, the presence of a binary, particularly at $\sim$ pc scales, may replenish or otherwise enhance the rate of encounters. This has been proposed as a method of explaining the excess of TDE host galaxies showing signs of recent bursts of star-formation and/or galaxy mergers (so-called 'E+A' galaxies). TDE hosts may then be higher-likelihood sources for additional followup in search of additional binary signatures. In some rare cases, stars in TDEs (or their debris streams) may actually interact-with, or feed, both MBH components of a binary. If this produces observable signatures, it could provide a direct indication of a binary. Post-merger MBH remnants that experience GW recoils could also experience enhanced TDE rates, offering a probe of post-merger systems.
- **Varsitometry.** The combination of a variable-brightness point-source (i.e. AGN), and a steady circularly-symmetric source (i.e. galaxy), leads to a time varying astrometric centroid of the combined light, even if the two sources are well within the point-spread function of the observatory (Shen et al., 2019). Differentiating the two components chromatically and/or spectroscopically, along with carefully modeling the variability, allows for the measurement of an offset between the two sources that (to first order) is insensitive to the size of the point-spread function. For typical sensitivities and configurations this could allow dual AGN to be detected at separations as low as $10s$ pc in the optical, perhaps marginally entering the 'binary' regime. Recently-merged galaxies are often highly asymmetric, making this approach more challenging.
- **Offset/Ejected AGN.** Just as dual AGN can be spatially resolved at $\sim$ kpc separations, an individual accreting MBH may be observably offset from the nucleus/centroid of the overall galaxy. This could happen during the early inspiral when only one component is active (again, recently-merged galaxies pose a challenge). As discussed above (Sec. 4.1.4), three-body interactions may occur in a noticeable fraction of MBHBs, in which case one of the components (typically the least massive) can be dynamically ejected from the system. Additionally, following a binary merger, the combined remnant MBH can receive a recoil 'kick' due to the asymmetric emission of GWs. In many cases this could produce an offset AGN that wanders the galaxy until it sinks back to the nucleus due to dynamical processes. In some cases that kick velocity can exceed the escape velocity of the galaxy, leading to MBHs that are entirely ejected from their host galaxies. Offset and especially ejected MBHs will fairly quickly deplete their reservoirs of gas, and are less likely to continue receiving large inflows, so they will preferentially be dark after some period of time. Even within the Milky Way, identifying an inactive, wandering MBH would be challenging.

5 Non-Binary GW Sources

The standard model of particle physics has proven incredibly successful, but it does not explain a number of fundamental problems, for example: the baryon asymmetry, and the nature of dark matter and dark energy. For these reasons, it is necessary for there to be additional physics 'Beyond the Standard Model' (BSM). Similarly, a quantum theory of gravity is broadly expected, but has yet to be expounded. Many BSM models have been proposed, focusing on the behavior of the Universe at very early times and at very large scales. Many of these models, such as cosmic inflation, generically predict the generation of GWs in the early Universe (Maggiore, 2000). BSM scalar fields could exhibit phase transitions in the early Universe[37] which directly, or via topological defects (such as cosmic strings and domain walls), would produce GWs (Kibble, 1976). The imprints from GWs at ultra-low frequencies (with wavelengths comparable to the observable Universe) in the cosmic microwave background are actively being searched for, typically in the form of 'B-mode' polarization signals (Hu and White, 1997).

[37] Note that quantum chromo-dynamics from the standard model, also predicts an early Universe phase transition that could produce GWs (Witten, 1984).

The recent evidence for a stochastic GWB has also been analyzed in terms of BSM models, and the predictions from a number of models are found to be consistent with current observations (Afzal et al., 2023). It is difficult to assess the *a priori* plausibility of these models as their parameter spaces are often entirely unconstrained, and most models haven't been entirely reconciled with typical cosmological/astronomical observations and/or the Standard Model. At the same time, no confirmed MBHBs have been identified either in EM surveys or GW observations. The identification of individual MBHBs, particularly through GW emission, would strongly suggest they produce the bulk of the GWB. It is of course possible that both MBHBs and cosmological sources could both contribute, even if not at an equal level. A general feature of BSM GW predictions is that the effective number of GW sources is far larger than the expected number of GW emitting MBH binaries, and their emission is in the very early Universe. Both factors suggest that a GWB from cosmological/BSM sources would be far more isotropic on the sky than MBHBs. This motivates the search for anisotropy as a 'smoking gun' for a binary origin to the GWB, whereas stringent limits on the maximum allowable anisotropy could strongly indicate a BSM origin.

Conclusions & Outlook

Pulsar timing arrays (PTAs) have found compelling evidence for a low-frequency (nanohertz), stochastic gravitational-wave background (GWB). Such a signal has long been proposed from the ensemble of GWs produced by many massive black-hole binaries (MBHBs) distributed throughout the Universe. While the idealized prediction is a power-law GWB characteristic-strain spectrum with an index of $-2/3$, realistic GWB spectra are likely to deviate significantly due to the discreteness of emitters and non-GW evolution due to environmental interactions. In fact, these deviations from the power-law behavior encode a wealth of information about MBHB populations and their evolution.

Current data is unable to confirm whether the measured signal is produced by MBHBs. It is also possible that non-standard-model physics in the early Universe could be producing the measured GW signal. Additional data is needed to definitively establish the origin of the GWB. Specifically, the identification of individual, loud binaries producing continuous wave (CW) emission, or similarly significant variations in loudness of the GWB either between different frequency bins or different parts of the sky, would definitively establish MBHBs as the origin of the signal. In this case, a wealth of possible electromagnetic (EM) counterpart signals could allow for multi-messenger astrophysics. Identifying EM signals that uniquely identify MBHBs, as opposed to peculiar normal/individual active galactic nuclei (AGN), remains very challenging.

Over the next few years, PTAs are poised to either (i) answer long-standing questions about the formation and evolution of the Universe's most behemoth MBH binaries, and/or (ii) measure/constrain beyond-the-standard-model physics in the very-early Universe. The hunt is on to increase PTA sensitivity, better characterize the GWB spectrum, and measure the contributions to the GWB from different parts of the sky—determining the true source of these GWs and possibly identifying individually-loud MBHBs, their host galaxies, and their EM counterparts.

References

Abbott R, Abbott TD, Acernese F, Ackley K, Adams C, Adhikari N, Adhikari RX, Adya VB, Affeldt C, Agarwal D and et al. (2023), Oct. GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run. *Physical Review X* 13 (4), 041039. doi:10.1103/PhysRevX.13.041039. 2111.03606.

Afzal A, Agazie G, Anumarlapudi A, Archibald AM, Arzoumanian Z, Baker PT, Bécsy B, Blanco-Pillado JJ, Blecha L, Boddy KK and et al. (2023), Jul. The NANOGrav 15 yr Data Set: Search for Signals from New Physics. *Astrophys. J. Lett.* 951 (1), L11. doi:10.3847/2041-8213/acdc91. 2306.16219.

Agazie G, Anumarlapudi A, Archibald AM, Arzoumanian Z, Baker PT, Bécsy B, Blecha L, Brazier A, Brook PR, Burke-Spolaor S and et al. (2023), Jul. The NANOGrav 15 yr Data Set: Evidence for a Gravitational-wave Background. *Astrophys. J. Lett.* 951 (1), L8. doi:10.3847/2041-8213/acdac6. 2306.16213.

Allen B (2023), Feb. Variance of the Hellings-Downs correlation. *Phys. Rev. D* 107 (4), 043018. doi:10.1103/PhysRevD.107.043018. 2205.05637.

Anholm M, Ballmer S, Creighton JDE, Price LR and Siemens X (2009), Apr. Optimal strategies for gravitational wave stochastic background searches in pulsar timing data. *Phys. Rev. D* 79 (8), 084030. doi:10.1103/PhysRevD.79.084030. 0809.0701.

Antoniadis J, Arzoumanian Z, Babak S, Bailes M, Bak Nielsen AS, Baker PT, Bassa CG, Bécsy B, Berthereau A, Bonetti M and et al. (2022), Mar. The International Pulsar Timing Array second data release: Search for an isotropic gravitational wave background. *Mon. Not. R. Astron. Soc.* 510 (4): 4873–4887. doi:10.1093/mnras/stab3418. 2201.03980.

Antonucci R (1993), Jan. Unified models for active galactic nuclei and quasars. *Annu. Rev. Astron. Astrophys.* 31: 473–521. doi:10.1146/annurev.aa.31.090193.002353.

Arzoumanian Z, Baker PT, Blumer H, Bécsy B, Brazier A, Brook PR, Burke-Spolaor S, Chatterjee S, Chen S, Cordes JM and et al. (2020), Dec. The NANOGrav 12.5 yr Data Set: Search for an Isotropic Stochastic Gravitational-wave Background. *Astrophys. J. Lett.* 905 (2), L34. doi:10.3847/2041-8213/abd401. 2009.04496.

Backer DC and Hellings RW (1986), Jan. Pulsar timing and general relativity. *Annu. Rev. Astron. Astrophys.* 24: 537–575. doi:10.1146/annurev.aa.24.090186.002541.

Barack L and Cutler C (2004), Apr. LISA capture sources: Approximate waveforms, signal-to-noise ratios, and parameter estimation accuracy. *Phys. Rev. D* 69 (8), 082005. doi:10.1103/PhysRevD.69.082005. gr-qc/0310125.

Begelman MC, Blandford RD and Rees MJ (1980), Sep. Massive black hole binaries in active galactic nuclei. *Nature* 287 (5780): 307–309. doi:10.1038/287307a0.

Blecha L, Sijacki D, Kelley LZ, Torrey P, Vogelsberger M, Nelson D, Springel V, Snyder G and Hernquist L (2016), Feb. Recoiling black holes: prospects for detection and implications of spin alignment. *Mon. Not. R. Astron. Soc.* 456 (1): 961–989. doi:10.1093/mnras/stv2646. 1508.01524.

Bonetti M, Sesana A, Barausse E and Haardt F (2018), Jun. Post-Newtonian evolution of massive black hole triplets in galactic nuclei - III. A robust lower limit to the nHz stochastic background of gravitational waves. *Mon. Not. R. Astron. Soc.* 477 (2): 2599–2612. doi:10.1093/mnras/sty874. 1709.06095.

Bortolas E, Franchini A, Bonetti M and Sesana A (2021), Sep. The Competing Effect of Gas and Stars in the Evolution of Massive Black Hole Binaries. *Astrophys. J. Lett.* 918 (1), L15. doi:10.3847/2041-8213/ac1c0c. 2108.13436.

Chandrasekhar S (1943), Mar. Dynamical Friction. I. General Considerations: the Coefficient of Dynamical Friction. *Astrophys. J.* 97: 255. doi:10.1086/144517.

Condon JJ and Ransom SM (2016). Essential Radio Astronomy.

Cordes JM and Shannon RM (2010), Oct. A Measurement Model for Precision Pulsar Timing. *arXiv e-prints* , arXiv:1010.3785doi:10.48550/arXiv.1010.3785. 1010.3785.

Cornish NJ and Sesana A (2013), Nov. Pulsar timing array analysis for black hole backgrounds. *Classical and Quantum Gravity* 30 (22), 224005. doi:10.1088/0264-9381/30/22/224005. 1305.0326.

Detweiler S (1979), Dec. Pulsar timing measurements and the search for gravitational waves. *Astrophys. J.* 234: 1100–1104. doi:10.1086/157593.

D'Orazio DJ and Charisi M (2023), Oct. Observational Signatures of Supermassive Black Hole Binaries. *arXiv e-prints* , arXiv:2310.16896doi:10.48550/arXiv.2310.16896. 2310.16896.

D'Orazio DJ and Di Stefano R (2018), Mar. Periodic self-lensing from accreting massive black hole binaries. *Mon. Not. R. Astron. Soc.* 474 (3): 2975–2986. doi:10.1093/mnras/stx2936. 1707.02335.

D'Orazio DJ, Haiman Z and Schiminovich D (2015), Sep. Relativistic boost as the cause of periodicity in a massive black-hole binary candidate. *Nature* 525 (7569): 351–353. doi:10.1038/nature15262. 1509.04301.

Draine BT (2011). Physics of the Interstellar and Intergalactic Medium.

Enoki M and Nagashima M (2007), Feb. The Effect of Orbital Eccentricity on Gravitational Wave Background Radiation from Supermassive Black Hole Binaries. *Progress of Theoretical Physics* 117 (2): 241–256. doi:10.1143/PTP.117.241. astro-ph/0609377.

EPTA Collaboration, InPTA Collaboration, Antoniadis J, Arumugam P, Arumugam S, Babak S, Bagchi M, Bak Nielsen AS, Bassa CG, Bathula A and et al. (2023), Oct. The second data release from the European Pulsar Timing Array. III. Search for gravitational wave signals. *Astron. Astrophys.* 678, A50. doi:10.1051/0004-6361/202346844. 2306.16214.

Estabrook FB and Wahlquist HD (1975), Oct. Response of Doppler spacecraft tracking to gravitational radiation. *General Relativity and Gravitation* 6 (5): 439–447. doi:10.1007/BF00762449.

Event Horizon Telescope Collaboration, Akiyama K, Alberdi A, Alef W, Asada K, Azulay R, Baczko AK, Ball D, Baloković M, Barrett J and et al. (2019), Apr. First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole. *Astrophys. J. Lett.* 875 (1), L1. doi:10.3847/2041-8213/ab0ec7. 1906.11238.

Event Horizon Telescope Collaboration, Akiyama K, Alberdi A, Alef W, Algaba JC, Anantua R, Asada K, Azulay R, Bach U, Baczko AK and et al. (2022), May. First Sagittarius A* Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole in the Center of the Milky Way. *Astrophys. J. Lett.* 930 (2), L12. doi:10.3847/2041-8213/ac6674.

Farris BD, Duffell P, MacFadyen AI and Haiman Z (2014), Mar. Binary Black Hole Accretion from a Circumbinary Disk: Gas Dynamics inside the Central Cavity. *Astrophys. J.* 783 (2), 134. doi:10.1088/0004-637X/783/2/134. 1310.0492.

Favata M (2010), Apr. The gravitational-wave memory effect. *Classical and Quantum Gravity* 27 (8), 084036. doi:10.1088/0264-9381/27/8/084036. 1003.3486.

Flanagan ÉÉ and Hughes SA (1998), Apr. Measuring gravitational waves from binary black hole coalescences. I. Signal to noise for inspiral, merger, and ringdown. *Phys. Rev. D* 57 (8): 4535–4565. doi:10.1103/PhysRevD.57.4535. gr-qc/9701039.

Hazboun JS, Romano JD and Smith TL (2019), Nov. Realistic sensitivity curves for pulsar timing arrays. *Phys. Rev. D* 100 (10), 104028. doi:10.1103/PhysRevD.100.104028. 1907.04341.

Hellings RW and Downs GS (1983), Feb. Upper limits on the isotropic gravitational radiation background from pulsar timing analysis. *Astrophys. J. Lett.* 265: L39–L42. doi:10.1086/183954.

Hopkins PF, Hernquist L, Cox TJ and Kereš D (2008), Apr. A Cosmological Framework for the Co-Evolution of Quasars, Supermassive Black Holes, and Elliptical Galaxies. I. Galaxy Mergers and Quasar Activity. *Astrophys. J., Suppl. Ser.* 175 (2): 356–389. doi:10.1086/524362. 0706.1243.

Hu W and White M (1997), Oct. A CMB polarization primer. *Nature Astron.* 2 (4): 323–344. doi:10.1016/S1384-1076(97)00022-5. astro-ph/9706147.

Kelley LZ (2021), Jan. Basic considerations for the observability of kinematically offset binary AGN. *Mon. Not. R. Astron. Soc.* 500 (3): 4065–4077. doi:10.1093/mnras/staa3219. 2005.10255.

Kelley LZ, Blecha L and Hernquist L (2017a), Jan. Massive black hole binary mergers in dynamical galactic environments. *Mon. Not. R. Astron. Soc.* 464 (3): 3131–3157. doi:10.1093/mnras/stw2452. 1606.01900.

Kelley LZ, Blecha L, Hernquist L, Sesana A and Taylor SR (2017b), Nov. The gravitational wave background from massive black hole binaries in Illustris: spectral features and time to detection with pulsar timing arrays. *Mon. Not. R. Astron. Soc.* 471 (4): 4508–4526. doi:10.1093/mnras/stx1638. 1702.02180.

Kibble TWB (1976), Aug. Topology of cosmic domains and strings. *Journal of Physics A Mathematical General* 9 (8): 1387–1398. doi:10.1088/0305-4470/9/8/029.

Kormendy J and Ho LC (2013), Aug. Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies. *Annu. Rev. Astron. Astrophys.* 51 (1): 511–653. doi:10.1146/annurev-astro-082708-101811. 1304.7762.

Lorimer DR and Kramer M (2012). Handbook of Pulsar Astronomy.

Maggiore M (2000), Jul. Gravitational wave experiments and early universe cosmology. *Phys. Rep.* 331 (6): 283–367. doi:10.1016/S0370-1573(99)00102-7. gr-qc/9909001.

Milosavljević M and Merritt D (2003a), Oct. Long-Term Evolution of Massive Black Hole Binaries. *Astrophys. J.* 596 (2): 860–878. doi:10.1086/378086. astro-ph/0212459.

Milosavljević M and Merritt D (2003b), Oct., The Final Parsec Problem, Centrella JM, (Ed.), The Astrophysics of Gravitational Wave Sources, American Institute of Physics Conference Series, 686, AIP, 201–210, astro-ph/0212270.

Miranda R, Muñoz DJ and Lai D (2017), Apr. Viscous hydrodynamics simulations of circumbinary accretion discs: variability, quasi-steady state and angular momentum transfer. *Mon. Not. R. Astron. Soc.* 466 (1): 1170–1191. doi:10.1093/mnras/stw3189. 1610.07263.

Misner CW, Thorne KS and Wheeler JA (1973). Gravitation, W. H. Freeman, San Francisco. ISBN 978-0-7167-0344-0, 978-0-691-17779-3.

Moore CJ, Cole RH and Berry CPL (2015), Jan. Gravitational-wave sensitivity curves. *Classical and Quantum Gravity* 32 (1), 015014. doi:10.1088/0264-9381/32/1/015014. 1408.0740.

Muñoz DJ, Miranda R and Lai D (2019), Jan. Hydrodynamics of Circumbinary Accretion: Angular Momentum Transfer and Binary Orbital Evolution. *Astrophys. J.* 871 (1), 84. doi:10.3847/1538-4357/aaf867. 1810.04676.

Netzer H (2015), Aug. Revisiting the Unified Model of Active Galactic Nuclei. *Annu. Rev. Astron. Astrophys.* 53: 365–408. doi:10.1146/annurev-astro-082214-122302. 1505.00811.

Peters PC (1964), Nov. Gravitational radiation and the motion of two point masses. *Phys. Rev.* 136: B1224–B1232. doi:10.1103/PhysRev.136.B1224. https://link.aps.org/doi/10.1103/PhysRev.136.B1224.

Peters PC and Mathews J (1963), Jul. Gravitational Radiation from Point Masses in a Keplerian Orbit. *Physical Review* 131 (1): 435–440. doi:10.1103/PhysRev.131.435.

Phinney ES (2001), Aug. A Practical Theorem on Gravitational Wave Backgrounds. *ArXiv Astrophysics e-prints* astro-ph/0108028.

Reardon DJ, Zic A, Shannon RM, Hobbs GB, Bailes M, Di Marco V, Kapur A, Rogers AF, Thrane E, Askew J and et al. (2023), Jul. Search for an Isotropic Gravitational-wave Background with the Parkes Pulsar Timing Array. *Astrophys. J. Lett.* 951 (1), L6. doi:10.3847/2041-8213/acdd02. 2306.16215.

Rickett BJ (1990), Jan. Radio propagation through the turbulent interstellar plasma. *Annu. Rev. Astron. Astrophys.* 28: 561–605. doi:10.1146/annurev.aa.28.090190.003021.

Rodriguez C, Taylor GB, Zavala RT, Peck AB, Pollack LK and Romani RW (2006), Jul. A Compact Supermassive Binary Black Hole System. *Astrophys. J.* 646 (1): 49–60. doi:10.1086/504825. astro-ph/0604042.

Rodriguez-Gomez V, Genel S, Vogelsberger M, Sijacki D, Pillepich A, Sales LV, Torrey P, Snyder G, Nelson D, Springel V and et al. (2015), May. The merger rate of galaxies in the Illustris simulation: a comparison with observations and semi-empirical models. *Mon. Not. R. Astron. Soc.* 449 (1): 49–64. doi:10.1093/mnras/stv264. 1502.01339.

Romano JD and Allen B (2024), Sep. Answers to frequently asked questions about the pulsar timing array Hellings and Downs curve. *Classical and Quantum Gravity* 41 (17), 175008. doi:10.1088/1361-6382/ad4c4c. 2308.05847.

Romano JD and Cornish NJ (2017), Dec. Detection methods for stochastic gravitational-wave backgrounds: a unified treatment. *Living Reviews in Relativity* 20 (1), 2. doi:10.1007/s41114-017-0004-1. 1608.06889.

Rosado PA, Sesana A and Gair J (2015), Aug. Expected properties of the first gravitational wave signal detected with pulsar timing arrays. *Mon. Not. R. Astron. Soc.* 451 (3): 2417–2433. doi:10.1093/mnras/stv1098. 1503.04803.

Ryu T, Perna R, Haiman Z, Ostriker JP and Stone NC (2018), Jan. Interactions between multiple supermassive black holes in galactic nuclei: a solution to the final parsec problem. *Mon. Not. R. Astron. Soc.* 473 (3): 3410–3433. doi:10.1093/mnras/stx2524. 1709.06501.

Sato-Polito G, Zaldarriaga M and Quataert E (2024), Sep. Where are the supermassive black holes measured by PTAs? *Phys. Rev. D* 110 (6), 063020. doi:10.1103/PhysRevD.110.063020.

Sazhin MV (1978), Feb. Opportunities for detecting ultralong gravitational waves. *Soviet Astronomy* 22: 36–38.

Sesana A (2013), Nov. Insights into the astrophysics of supermassive black hole binaries from pulsar timing observations. *Classical and Quantum Gravity* 30 (22), 224014. doi:10.1088/0264-9381/30/22/224014. 1307.2600.

Sesana A, Vecchio A and Colacino CN (2008), Oct. The stochastic gravitational-wave background from massive black hole binary systems: implications for observations with Pulsar Timing Arrays. *Mon. Not. R. Astron. Soc.* 390 (1): 192–209. doi:10.1111/j.1365-2966.2008.13682.x. 0804.4476.

Shen Y, Hwang HC, Zakamska N and Liu X (2019), Nov. Varstrometry for Off-nucleus and Dual Sub-Kpc AGN (VODKA): How Well Centered Are Low-z AGN? *Astrophys. J. Lett.* 885 (1), L4. doi:10.3847/2041-8213/ab4b54. 1910.02969.

Siemens X, Ellis J, Jenet F and Romano JD (2013), Nov. The stochastic background: scaling laws and time to detection for pulsar timing arrays. *Classical and Quantum Gravity* 30 (22), 224015. doi:10.1088/0264-9381/30/22/224015. 1305.3196.

Taylor SR (2021), May. The Nanohertz Gravitational Wave Astronomer. *arXiv e-prints* , arXiv:2105.13270doi:10.48550/arXiv.2105.13270. 2105.13270.

Thorne KS (1992), Jan. Gravitational-wave bursts with memory: The Christodoulou effect. *Phys. Rev. D* 45 (2): 520–524. doi:10.1103/PhysRevD.45.520.

Vallisneri M, Taylor SR, Simon J, Folkner WM, Park RS, Cutler C, Ellis JA, Lazio TJW, Vigeland SJ, Aggarwal K and et al. (2020), Apr. Modeling the Uncertainties of Solar System Ephemerides for Robust Gravitational-wave Searches with Pulsar-timing Arrays. *Astrophys. J.* 893 (2), 112. doi:10.3847/1538-4357/ab7b67. 2001.00595.

Weisberg JM, Nice DJ and Taylor JH (2010), Oct. Timing Measurements of the Relativistic Binary Pulsar PSR B1913+16. *Astrophys. J.* 722 (2): 1030–1034. doi:10.1088/0004-637X/722/2/1030. 1011.0718.

Witten E (1984), Jul. Cosmic separation of phases. *Phys. Rev. D* 30 (2): 272–285. doi:10.1103/PhysRevD.30.272.

Yu Q (2002), Apr. Evolution of massive binary black holes. *Mon. Not. R. Astron. Soc.* 331 (4): 935–958. doi:10.1046/j.1365-8711.2002.05242.x. astro-ph/0109530.