Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **464,** 3131–3157 (2017)
Advance Access publication 2016 October 4

doi:10.1093/mnras/stw2452

Massive black hole binary mergers in dynamical galactic environments

Luke Zoltan Kelley,[1]★ Laura Blecha[2] and Lars Hernquist[1]

[1]*Harvard University, Center for Astrophysics, Cambridge, MA 02138, USA*
[2]*University of Maryland, College Park, MD 20742, USA*

Accepted 2016 September 26. Received 2016 September 22; in original form 2016 June 3; Editorial Decision 23 September 2016

ABSTRACT

Gravitational waves (GWs) have now been detected from stellar-mass black hole binaries, and the first observations of GWs from massive black hole (MBH) binaries are expected within the next decade. Pulsar timing arrays (PTA), which can measure the years long periods of GWs from MBH binaries (MBHBs), have excluded many standard predictions for the amplitude of a stochastic GW background (GWB). We use coevolved populations of MBHs and galaxies from hydrodynamic, cosmological simulations ('Illustris') to calculate a predicted GWB. The most advanced predictions so far have included binary hardening mechanisms from individual environmental processes. We present the first calculation including all of the environmental mechanisms expected to be involved: dynamical friction, stellar 'loss-cone' scattering, and viscous drag from a circumbinary disc. We find that MBH binary lifetimes are generally multiple gigayears, and only a fraction coalesce by redshift zero. For a variety of parameters, we find all GWB amplitudes to be below the most stringent PTA upper limit of $A_{\mathrm{yr}^{-1}} \approx 10^{-15}$. Our fairly conservative fiducial model predicts an amplitude of $A_{\mathrm{yr}^{-1}} \approx 0.4 \times 10^{-15}$. At lower frequencies, we find $A_{0.1\,\mathrm{yr}^{-1}} \approx 1.5 \times 10^{-15}$ with spectral indices between -0.4 and -0.6 – significantly flatter than the canonical value of $-2/3$ due to purely GW-driven evolution. Typical MBHBs driving the GWB signal come from redshifts around 0.3, with total masses of a few times $10^9\,\mathrm{M}_\odot$, and in host galaxies with very large stellar masses. Even without GWB detections, our results can be connected to observations of dual active galactic nuclei to constrain binary evolution.

Key words: gravitational waves – galaxies: evolution – galaxies: kinematics and dynamics – galaxies: nuclei – quasars: supermassive black holes.

1 INTRODUCTION

Massive black holes (MBHs) occupy at least the majority of massive galaxies (e.g. Soltan 1982; Kormendy & Richstone 1995; Magorrian et al. 1998) that are also known to merge with each other as part of their typical lifecycles (e.g. Lacey & Cole 1993; Lotz et al. 2011; Rodriguez-Gomez et al. 2015). This presents two possibilities for the MBH of host galaxies that merge: either they also merge or they persist as multiples in the resulting remnant galaxies. Naively, one might expect that BHs *must* undergo mergers for them to grow – as for haloes and to some degree galaxies in the fundamentally hierarchical lambda cold dark matter model (e.g. White & Frenk 1991). On the contrary, the linear growth of black holes (i.e. at most doublings of mass) is known to be woefully inadequate to form the massive quasars observed at high redshifts (e.g. Fan et al. 2006), which require exponential growth from Eddington[1] (or super-Eddington) accretion (e.g. Haiman 2013). Assuming that the energy fuelling active galactic nuclei (AGNs) come from accretion on to compact objects, the integrated luminosity over redshifts requires a present-day mass density comparable to that of observed MBHs (Soltan 1982). Coalescences of MBHs are then neither sufficient nor necessary to match their observed properties (e.g. Small & Blandford 1992).

In the last decade, the search for multi-MBH systems has yielded many with kiloparsec-scale separations ('dual AGNs'; e.g. Comerford et al. 2012) – although they seem to represent only a fraction

[1] The Eddington accretion rate can be related to the Eddington luminosity as $\dot{M}_{\mathrm{Edd}} = L_{\mathrm{Edd}}/\varepsilon_{\mathrm{rad}}c^2$, for a radiative efficiency $\varepsilon_{\mathrm{rad}}$, i.e.

$$\dot{M}_{\mathrm{Edd}} = \frac{4\pi G M m_{\mathrm{p}}}{\varepsilon_{\mathrm{rad}} c \sigma_{\mathrm{T}}},$$

for a proton mass m_{p}, and Thomson-scattering cross-section σ_{T}.

★ E-mail: lkelley@cfa.harvard.edu

of the population (Koss et al. 2012) – and even some triple systems (e.g. Deane et al. 2014). At separations of kiloparsecs, 'dual MBHs' are far from the 'hard' binary phase. A 'hard' binary is one in which the binding energy is larger than the typical kinetic energy of nearby stars (Binney & Tremaine 1987), and relatedly the binding energy tends to increase (i.e. the system 'hardens') with stellar interactions (Hut 1983). 'Hard' is also used more informally to highlight systems that behave dynamically as a bound system. Only one true MBH binary (MBHB, i.e. gravitationally bound) system has been confirmed: a resolved system in the radio galaxy 0402+379, at a separation of ∼7 pc by Rodriguez et al. (2006). Even this system is well outside of the 'gravitational wave (GW) regime', in which the system could merge within a Hubble time due purely to GW emission. There are, however, a growing number of candidates, unresolved systems with possible sub-parsec separations (e.g. Valtonen et al. 2008; Dotti et al. 2009). Detecting – and even more so, excluding – the presence of MBHBs is extremely difficult because activity fractions of AGNs (particularly at low masses) are uncertain, the spheres of influence of MBHs are almost never resolved, and the expected time-scales at each separation are unknown. It is then hard to establish if the absence of MBHB observations is conspicuous.

On the theoretical side, the picture is no more complete. Pioneering work by Begelman, Blandford & Rees (1980, hereafter BBR80) outlined the basic MBH merger process. On large scales (∼kpc), MBHs are brought together predominantly by dynamical friction (DF) – the deceleration of a body moving against a gravitating background. Energy is transferred from the motion of the massive object to a kinetic thermalization of the background medium, in this case the dark-matter (DM), stellar and gaseous environment of MBHB host galaxies. As the binary tightens, stars become the primary scatterers. Once the binary becomes hard (∼10 pc), depletion of the 'loss cone' (LC) – the region of parameter space with sufficiently low angular momentum to interact with the MBHB – must be considered. The rate at which the LC is refilled is likely the largest uncertainty in the merger process and determines the fraction of systems that are able to cross the so-called 'final parsec' (for the definitive review, see Merritt 2013). If binaries are able to reach smaller scales ($\lesssim$0.1 pc), gas drag can contribute significantly to the hardening process in gas-rich systems. Eventually, GW emission inevitably dominates at the smallest scales ($\lesssim 10^{-3}$ pc).

Since BBR80, the details of the merger process have been studied extensively, largely focusing on the LC (for a review, see e.g. Merritt & Milosavljević 2005). Quinlan (1996) and Quinlan & Hernquist (1997) made significant developments in numerical, *N*-body scattering experiments, allowing the 'measurement' of binary hardening parameters. More advanced descriptions of the LC, applied to realistic galaxy density profiles, by Yu (2002) highlighted the role of the galactic gravitational potential – suggesting that flattened and strongly triaxial galaxies could be effective at refilling the LC and preventing binaries from stalling. *N*-body simulations have continued to develop and improve our understanding of rotating galaxies (e.g. Berczik et al. 2006), and more realistic galaxy-merger environments and galaxy shapes (e.g. Khan, Just & Merritt 2011; Khan et al. 2013). The interpretation and usage of simulations have also developed significantly, in step with numerical advancements, allowing for better understanding of the underlying physics (e.g. Sesana & Khan 2015; Vasiliev, Antonini & Merritt 2015).

Smoothed particle hydrodynamic (SPH) simulations of MBH binary dynamics in gaseous environments have also been performed. Escala et al. (2005) showed that dense gaseous regions, corresponding to ULIRG-like[2] galaxies, can be very effective at hardening binaries. Similar SPH studies have affirmed and extended these results to more general environments and MBHB configurations (e.g. Dotti et al. 2007; Cuadra et al. 2009). While modern simulations continue to provide invaluable insights, and exciting steps towards simulating MBHBs over broad physical scales are underway (e.g. Khan et al. 2016), neither hydrodynamic nor purely *N*-body simulations are close to simulating the entire merger process in its full complexity.[3]

With our entrance into the era of GW astronomy (LIGO 2016a), we are presented with the prospect of observing compact objects outside of both the electromagnetic spectrum and numerical simulations. Direct detection experiments for GWs are based on precisely measuring deviations in path length (via light-travel times). While ground-based detectors like the Laser Interferometer Gravitational-Wave Observatory (LIGO; LIGO 2016b) use the interference of light between two orthogonal, kilometre scale laser arms, pulsar timing arrays (PTA; Foster & Backer 1990) use the kiloparsec-scale separations between earth and galactic pulsars (Detweiler 1979). PTA are sensitive to GWs at periods between the total observational baseline and the cadence between observations. These frequencies, roughly 0.1–10 yr^{-1}, are much lower than LIGO – corresponding to steady orbits of MBHBs with total masses between $\approx 10^6$ and 10^{10} $\mathrm{M}_\odot$, at separations of $\approx 10^{-3}$ to 10^{-1} pc (i.e. $1 - 10^6\,R_s$[4]). The parameter space is shown in Fig. A1.

Binaries produce GWs that increase in amplitude and frequency as the orbit hardens, up to the 'chirp' when the binary coalesces. MBHB chirps will be at frequencies below the LIGO band, but above that of PTA. Future space-based interferometers (e.g. eLISA; The eLISA Consortium et al. 2013) will bridge the divide and observe not only the coalescence of MBHBs, but also years of their final inspiral. The event rate of nearby, hard MBHBs that could be observed as individual 'continuous wave' sources is expected to be quite low, and likely the first GW detections from PTA will be of a stochastic GW background (GWB) of unresolved sources (Rosado, Sesana & Gair 2015).

The shape of the GWB spectrum was calculated numerically more than two decades ago (Rajagopal & Romani 1995), but Phinney (2001) showed that the characteristic GWB spectrum can be calculated analytically by considering the total energy emitted as GWs, integrated over redshift. For a complete and pedagogical derivation of the GWB spectrum see e.g. Sesana, Vecchio & Colacino (2008). The 'characteristic strain', $h_c(f)$, can be calculated for a finite number of sources, in some comoving volume V_c (e.g. a computational box), as,

$$h_c^2(f) = \frac{4\pi}{3c^2}\,(2\pi\, f)^{-4/3} \sum_i \frac{1}{(1+z_i)^{1/3}}\,\frac{(G\mathcal{M}_i)^{5/3}}{V_c}. \quad (1)$$

Equation (1) is the simplest way to calculate a GWB spectrum, requiring just a distribution of merger chirp masses and redshifts. This type of relation is often written as

$$h_c(f) = A_0 \left(\frac{f}{f_0}\right)^{-2/3}, \quad (2)$$

[2] Ulta-Luminous Infra-Red Galaxies (ULIRG) are bright, massive, and gas rich – all indicators of favourable MBH merger environments.

[3] Resolving the interaction of individual stars with an MBHB over the course of the entire merger process, for example, would require almost nine orders of magnitude contrast in each mass, distance, and time.

[4] Schwarzschild radii, $R_s \equiv 2GM/c^2$.

Table 1. Representative sample of previous predictions for the GWB, with a basic summary of their implementation. 'SAM' is used loosely to refer to numerical models based on scaling relations and observed populations. 'Cosmo-DM' are cosmological DM-only (*N*-body) simulations, while 'Cosmo-Hydro' are hydrodynamic simulations including baryons. The physical evolution effects included are: DF, LC stellar scattering, VD from a circumbinary disc, and GW radiation. In this study, we use populations of both galaxies and MBHs that coevolved in the cosmological, hydrodynamic 'Illustris' simulations and include all mechanisms of hardening (DF, LC, VD and GW) in our models.

Reference	GWB amplitude[1] $\log A_{\mathrm{yr}^{-1}}$	Populations Galaxies	Black holes	MBHB evolution	Spectral slope [Deviations from $-2/3$]
Jaffe & Backer (2003)	-16	SAM	SAM	GW	–
Wyithe & Loeb (2003)	-14.3	SAM	SAM	GW	–
Kocsis & Sesana (2011)	-15.7 ± 0.3	Cosmo-DM	SAM	VD, GW	Flattened, $f \lesssim 1\ \mathrm{yr}^{-1}$
Sesana (2013b)	-15.1 ± 0.3	SAM	SAM	GW	–
McWilliams et al. (2014)	-14.4 ± 0.3	SAM	SAM	DF, GW	Imposed cutoff, $f \lesssim 0.5\ \mathrm{yr}^{-1}$
Ravi et al. (2014)	-14.9 ± 0.25	Cosmo-DM	SAM	LC-Full,[2] GW	Flattened $f \lesssim 10^{-1}\ \mathrm{yr}^{-1}$, cutoff $f \lesssim 10^{-2}\ \mathrm{yr}^{-1}$
Kulier et al. (2015)	-14.7 ± 0.1	Cosmo-Hydro	SAM	DF, GW	–
Roebber et al. (2016)	$-15.2^{+0.4}_{-0.2}$	Cosmo-DM	SAM	GW	–
Sesana et al. (2016)	-15.4 ± 0.4	SAM	SAM	GW	–

[1]Some values that were not given explicitly in the included references were estimated based on their figures, and thus should be taken as approximate.
[2]In this case, the LC prescription is effectively always 'Full'.

which has become typical for GWB predictions and usually normalized to $f_0 = 1\ \mathrm{yr}^{-1}$ (with some $A_{\mathrm{yr}^{-1}}$). The prediction of a GWB with a spectral slope of $-2/3$ is quite general, but does assume purely GW-driven hardening that produces a purely power-law evolution in frequency. The lack of high- and low-frequency cutoffs is fortuitously accurate at the frequencies observable through PTA, which are well populated by astrophysical MBHB systems. Deviations from pure power-law behaviour within this band, however, are not only possible but expected – the degree of which, determined by how significant non-GW effects are, is currently of great interest.

Many predictions have been made for the normalization of the GWB based on extensions to the method of Phinney (2001). The standard methodology is using semi-analytic models[5] (SAM) of galaxy evolution, with prescribed MBHB mergers to calculate a GWB amplitude. Two of the earliest examples are Wyithe & Loeb (2003) – who use analytic mass functions (Press & Schechter 1974) with observed merger rates (Lacey & Cole 1993), and Jaffe & Backer (2003) – who use observationally derived galaxy mass functions, pair fractions, and merger time-scales. These studies find amplitudes of $\log A_{\mathrm{yr}^{-1}} = -14.3$ and $\log A_{\mathrm{yr}^{-1}} = -16$, respectively, that remain as upper and lower bounds to most predictions since then. Monte Carlo realizations of hierarchical cosmologies (Sesana et al. 2004) exploring varieties of MBHB formation channels (e.g. Sesana et al. 2008; Sesana 2013b; Roebber et al. 2016) have been extremely fruitful in populating and understanding the parameter space, finding GWB amplitudes generally consistent with $\log A_{\mathrm{yr}^{-1}} \approx -15 \pm 1$. Sesana et al. (2016) find that accounting for bias in MBH–host scaling relations moves SAM predictions towards the lower end of this range at $\log A_{\mathrm{yr}^{-1}} = -15.4$.

More extensive models exploring deviations from the purely power-law GWB have also been explored. For example, at higher frequencies ($\gtrsim 1\ \mathrm{yr}^{-1}$) from a finite numbers of sources (Sesana, Vecchio & Volonteri 2009) or at lower frequencies due to eccentric binary evolution (e.g. Sesana 2010). Recently, much work has focused on the 'environmental effects' outlined by BBR80. Kocsis & Sesana (2011) incorporate viscous drag (VD) from a circumbinary gaseous disc (Haiman, Kocsis & Menou 2009, hereafter HKM09) on top of haloes and mergers from the DM-only Millennium simulations (Springel et al. 2005), with MBHs added in post-processing. They find a fairly low-amplitude GWB, $\log A_{\mathrm{yr}^{-1}} \approx -16 \pm 0.5$, with a flattening spectrum below $\sim 1\ \mathrm{yr}^{-1}$. Ravi et al. (2014) explore eccentric binary evolution in an always effectively refilled (i.e. full) LC using the Millennium simulation with the SAM of Guo et al. (2011). They find $\log A_{\mathrm{yr}^{-1}} \approx -15 \pm 0.5$ with a turnover in the GWB below $\sim 10^{-2}\ \mathrm{yr}^{-1}$ and significant attenuation up to $\sim 10^{-1}\ \mathrm{yr}^{-1}$. Recently, both McWilliams, Ostriker & Pretorius (2014) and Kulier et al. (2015) have implemented explicit DF formalisms along with recent MBH–host scaling relations (McConnell & Ma 2013) applied to halo mass functions from Press–Schechter and the Millennium simulations, respectively. McWilliams et al. (2014) find $\log A_{\mathrm{yr}^{-1}} \approx 14.4 \pm 0.3$, and Kulier et al. (2015) $\log A_{\mathrm{yr}^{-1}} \approx 14.7 \pm 0.1$, with both highlighting the non-negligible fraction of binaries stalled at kiloparsec-scale separations. Almost all previous studies had assumed that all MBHBs merge effectively.

These predictions are summarized in Table 1. While far from exhaustive, we believe they are a representative sample, with specific attention to recent work on environmental effects. The amplitudes of the predicted backgrounds are distributed fairly consistently around $A_{\mathrm{yr}^{-1}} \approx 10^{-15}$. Assuming observational baselines of about 10 yr, pulsar TOA accuracies of at least tens of microseconds are required to constrain or observe a GWB with this amplitude (see e.g. Blandford, Romani & Narayan 1984; Rajagopal & Romani 1995). Finding more millisecond pulsars with very small intrinsic timing noise is key to improving GWB upper limits, while increasing the total number (and angular distribution) of pulsars will be instrumental for detections (Taylor et al. 2016).

There are currently three ongoing PTA groups, the North-American Nanohertz Observatory for Gravitational Waves (NANOGrav, The NANOGrav Collaboration et al. 2015), the European PTA (EPTA; Desvignes et al. 2016), and the Parkes PTA (PPTA; Manchester et al. 2013). Additionally, the International PTA (IPTA, Hobbs et al. 2010) aims to combine the data sets from each individual project and has recently produced their first public data release (Verbiest et al. 2016). Table 2 summarizes the current upper limits from each PTA. These are the 2-σ upper bounds, based on both extrapolation to $A_{\mathrm{yr}^{-1}}$ along with that of the specific frequency with the *strongest constraint* assuming a $-2/3$ spectral index. Overall, the lowest bound is from the PPTA, at $A_{\mathrm{yr}^{-1}} < 10^{-15}$, or in terms

[5] We use the term 'semi-analytic model' loosely to refer to a *realized population* constructed by an analytic prescription, as apposed to derived from underlying physical models.

Table 2. Upper limits on the GWB from PTAs. Values are given both at the standard normalization of $f = 1\,\mathrm{yr}^{-1}$ in addition to the frequency and amplitude of the strongest constraint (when given).

		Strongest constraint		
PTA	$A_{\mathrm{yr}^{-1}}$	$A_{f,0}$	f_0 [yr^{-1}]	Reference
European	3.0×10^{-15}	1.1×10^{-14}	0.16	Lentati et al. (2015)
NANOGrav	1.5×10^{-15}	4.1×10^{-15}	0.22	Arzoumanian et al. (2016)
Parkes	1.0×10^{-15}	2.9×10^{-15}	0.2	Shannon et al. (2015)
IPTA	1.5×10^{-15}	–	–	Verbiest et al. (2016)

of the fractional closure density, $\Omega_{\mathrm{GW}}(f = 0.2\ \mathrm{yr}^{-1}) < 2.3 \times 10^{-10}$ (Shannon et al. 2015).

Every existing prediction has been made with the use of SAM – mostly in the construction of the galaxy population, but also in how black holes are added on to those galaxies. SAM are extremely effective in efficiently creating large populations based on observational relations. Higher order, less observationally constrained parameters can have systemic biases however, for example galaxy-merger rates (see e.g. Hopkins et al. 2010) – which are obviously critical to understanding MBHB evolution. Recently, Rodriguez-Gomez et al. (2015) have shown that merger rates from the cosmological, hydrodynamic Illustris simulations (Vogelsberger et al. 2014b) show excellent agreement with observations, while differing (at times substantially) from many canonical SAM.

In this paper, we use results from the Illustris (Genel et al. 2014) simulations (discussed in Section 2) to make predictions for the rates at which MBHs form binaries and evolve to coalescence. Illustris provides the MBH population along with their self-consistently derived parent galaxies and associated stellar, gaseous, and DM components. This is the first time that a hydrodynamic galaxy population, with fully co-evolved MBHs, has been used to calculate a GWB spectrum. We use these as the starting point for post-processed models of the unresolved merger dynamics themselves, including all of the underlying hardening mechanisms: DF, LC stellar scattering, VD, and GW evolution – again for the first time (Section 3) in an MBHB population calculation. From these data, we make predictions for plausible GWBs observable by PTA, focusing on the effects of different particular mechanisms on the resulting spectrum such that future detections and upper limits can be used to constrain the physical merger process (Section 4).

In addition to the GWB, understanding the population of MBHBs is also important for future space-based GW observatories (e.g. eLISA The eLISA Consortium et al. 2013). Solid predictions for binary time-scales at different separations will also be instrumental in interpreting observations of dual and binary AGNs, in addition to offset and 'kicked' BHs (Blecha et al. 2016). Finally, MBHBs could play a significant role in triggering stellar tidal disruption events (TDE; e.g. Ivanov, Polnarev & Saha 2005; Chen et al. 2009) and explaining the distribution of observed TDE host galaxies.

2 THE ILLUSTRIS SIMULATIONS

Illustris are a suite of cosmological, hydrodynamic simulations which have accurately reproduced both large-scale statistics of thousands of galaxies at the same time as the detailed internal structures of ellipticals and spirals (Vogelsberger et al. 2014b). Illustris – hereafter referring to Illustris-1, the highest resolution of three runs – is a cosmological box of 106.5 Mpc on a side, with 1820^3 each gas cells and DM particles. The simulations use the moving, unstructured-mesh hydrodynamic code AREPO (Springel 2010), with superposed SPH particles (e.g. Springel et al. 2005) representing stars (roughly $1.3 \times 10^6\,\mathrm{M}_\odot$ mass resolution, 700 pc gravitational softening length), DM ($6.3 \times 10^6\,\mathrm{M}_\odot$, 1.4 kpc), and MBHs (seeded at $M \approx 10^5\,\mathrm{M}_\odot$), and allowed to accrete and evolve dynamically. Stars form and evolve, feeding back and enriching their local environments, over the course of the simulation that is initialized at redshift $z = 137$ and evolved until $z = 0$ at which point there are over 3×10^8 star particles.

For a comprehensive presentation of the galaxy formation models (e.g. cooling, inter-stellar medium, stellar evolution, chemical enrichment) see the papers Vogelsberger et al. (2013) and Torrey et al. (2014). For detailed descriptions of the general results of the Illustris simulations, and comparisons of their properties with the observed Universe, see e.g. Vogelsberger et al. (2014a), Genel et al. (2014), and Sijacki et al. (2015). Finally, the data for the Illustris simulations, and auxiliary files containing the black hole data[6] used for this analysis, have been made publicly available online (www.illustris-project.org; Nelson et al. 2015).

2.1 The black hole merger population

Black holes are implemented as massive, collisionless 'sink' particles seeded into sufficiency massive haloes. Specifically, haloes with a total mass above $7.1 \times 10^{10}\,\mathrm{M}_\odot$, identified using an on-the-fly Friends-Of-Friends (FOF) algorithm, which do not already have an MBH are given one with a seed mass, $M_{\mathrm{seed}} = 1.42 \times 10^5\,\mathrm{M}_\odot$ (Sijacki et al. 2007). The highest density gas cell in the halo is converted into the BH particle. The BH mass is tracked as an internal quantity, while the particle overall retains a *dynamical* mass initially equal to the total mass of its predecessor gas cell (Vogelsberger et al. 2013). The internal BH mass grows by Eddington-limited, Bondi–Hoyle accretion from its parent gas cell (i.e. the total dynamical mass remains the same). Once the excess mass of the parent is depleted, mass is accreted from nearby gas particles – increasing both the dynamical mass of the sink particle and the internal BH mass quantity.

BH sink particles typically have masses comparable to (or within a few orders of magnitude of) that of the nearby stellar and DM particles. Freely evolving BH particles would then scatter around their host halo instead of dynamically settling to their centre – as is the case physically. To resolve this issue, BH particles in Illustris are repositioned to the potential minima of their host haloes. For this reason, their parametric velocities are not physically meaningful. Black hole 'mergers' occur in the simulation whenever two MBH particles come within a particle smoothing length of one another – typically on the order of a kiloparsec. This project aims to fill in the merger process unresolved in Illustris. In our model, an Illustris 'merger' corresponds to the *formation* of an MBH binary system, which we then evolve. To avoid confusion, we try to use the term

[6] The black hole data files were made public in late September 2016.

'coalescence' to refer to the point at which such a binary would actually collide, given arbitrary resolution.

Over the course of the Illustris run, 135 'snapshots' were produced, each of which includes internal parameters of all simulation particles. Additional black-hole-specific output was also recorded at every time step, providing much higher time resolution for black hole accretion rates[7], local gas densities and, most notably, merger events. The entire set of mergers – a time and pair of BH masses – constitutes our initial population of MBH binaries.

The distribution of BH masses is peaked at the lowest masses. Many of these black holes are short-lived: their small, usually satellite, host-of-matter haloes often quickly merge with a nearby neighbour – producing a BH 'merger' event. Additionally, in some cases, the identification of a particular matter overdensity as a halo by the FOF halo finder, while transient, may be sufficiently massive to trigger the creation of a new MBH seed particle. This seed can then quickly merge with the MBH in a nearby massive halo. Due to the significant uncertainties in our understanding of MBH and MBH-seed formation, it is unclear if and when these processes are physical. For this reason we implement a mass cut on merger events, to ensure that each component of BHs has $M_\bullet > 10^6\,\mathrm{M}_\odot \approx 10\,M_{\rm seed}$. Whether or not these 'fast mergers' are non-physical, the mass cut is effective at excluding them from our analysis. The entire Illustris simulation has 23 708 MBH merger events; applying the mass cut excludes 11 291 (48 per cent) of those, leaving 12 417. We have run configurations without this mass cut, and the effects on the GWB are always negligible.

There is very small population of MBH 'merger' events that occur during close encounters (but not true mergers) of two host haloes. During the encounter, the halo finder might associate the two constituent haloes as one, causing the MBHs to merge spuriously due to the repositioning algorithm. These forced mergers are rare and, we believe, have no noticeable effect on the overall population of thousands of mergers. They are certainly negligible in the overall MBH–halo statistical correlations which are well reproduced in the Illustris simulations (Sijacki et al. 2015).

2.2 Merger host galaxies

To identify the environments that produce the DF, stellar scattering, and VD which we are interested in, we identify the host galaxies of each MBH involved in the merger in the snapshot preceding it, in addition to the single galaxy that contains the 'binary' (at this point a single, remnant MBH) in the snapshot immediately following the merger event. A total of 644 mergers (3 per cent) are excluded because they do not have an associated galaxy before or after the merger. To ensure that each host galaxy is sufficiently well resolved (especially important for calculating density profiles), we require that a sufficient number of each particle type constitute the galaxy. Following Blecha et al. (2016), we use a fiducial cut of 80 and 300 star and DM particles, respectively, and additionally require 80 gas cells. This excludes 54 of the remaining binaries. We emphasize here that this remnant host galaxy, as it is in the snapshot following the Illustris 'merger' event, forms the environment in which we model the MBHB merger process. In the future, we plan to upgrade our implementation to take full advantage of the dynamically evolving merger environment: including information from both galaxies as they merge and the evolving remnant galaxy once it forms.

[7] These self-consistently derived mass accretion rates ($\dot{M}$) are used in our implementation of gas drag (discussed in Section 3.3) as a way of measuring the local gas density.

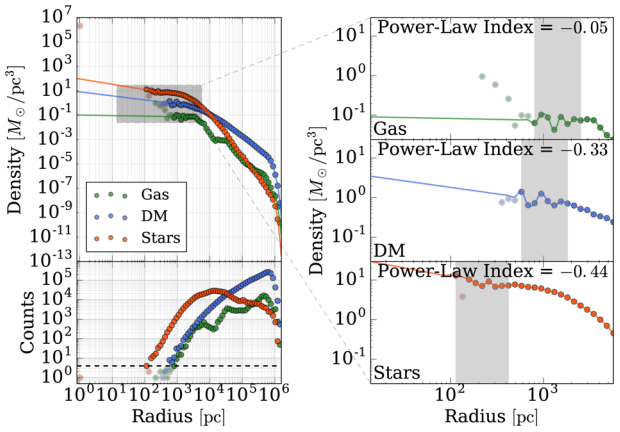

Figure 1. Density profiles from a sample Illustris MBHB host galaxy. Binned densities for each particle type are shown (upper left) along with the number of particles/cells in each (bottom left). Semi-transparent points are bins with less than four particles – the number required for consideration in calculating fits. Zoom-ins are also shown separately for each particle type (right), with the eight innermost bins with at least four particles shown in the shaded region. Those bins were used to calculate fits, which are overplotted. The resulting power-law indices used to extrapolate inwards are also shown. Any galaxy without enough (eight) bins is excluded from our sample, in addition to the MBHB it contains. This was the case for roughly 1 per cent of our initial population, almost entirely containing MBH very near our BH mass threshold ($10^6\,\mathrm{M}_\odot$).

We construct spherically averaged, radial density profiles for each host galaxy and each particle/cell type (star, DM, gas). Because the particle smoothing lengths are larger than the MBHB separations of interest, we extrapolate the galaxy density profiles based on fits to the inner regions. Our fits use the innermost eight radial bins that have at least four particles in them. Out of the valid binaries, 347 (1 per cent) are excluded because fits could not be constructed – generally because the particles are not distributed over the required eight bins. Successful fits typically use $\sim$100 particles, with gas cell sizes $\sim 10^2$ pc and SPH smoothing lengths for stars and DM $\sim 10^3$ pc.

Density profiles for a sample Illustris MBHB host galaxy are shown in Fig. 1. The left panels show the binned density profiles for each particle type (top) and the number of particles in each bin (bottom). The semi-transparent points are those with less than the requisite four particles in them. The right panels show zoom-ins for each particle type, where the shaded regions indicate the eight bins used for calculating fits. The resulting interpolants are overplotted, with the power-law index indicated. While the four particles per bin, and eight inner bins, generally provide for robust fits, we impose a maximum power-law index of -0.1 – i.e. that densities are at least gently increasing, and a minimum index of -3 – to ensure that the mass enclosed is convergent. Using these densities we calculate all additional galaxy profiles required for the hardening prescriptions (Section 3), e.g. velocities, binding energies, etc. When calculating profiles using our fiducial parameters, 2286 binaries (10 per cent) are excluded when calculating the distribution functions (Section 3.2), usually due to significant non-monotonicities in the radial density profile which are incompatible with the model assumptions. Overall, after all selection cuts, 9270/23 708 (39 per cent) of the initial Illustris 'merger' events are analysed in our simulations.

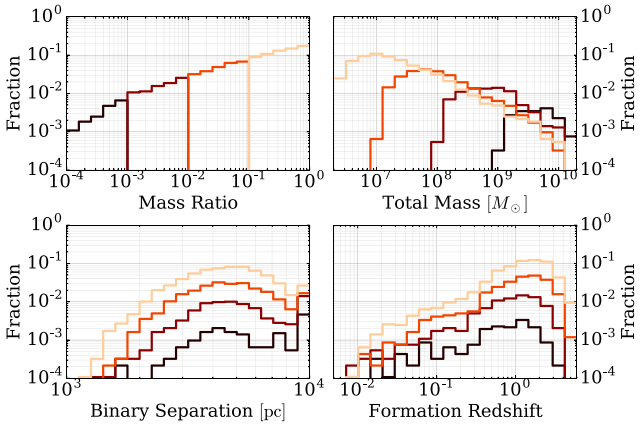

Figure 2. Properties of the MBHBs from the Illustris population passing our selection cuts. After selecting for MBH masses $M > 10^6\,\mathrm{M}_\odot$, and requiring the binary host galaxies to have sufficiently well-resolved density profiles, 9270 of 23 708 (39 per cent) systems remain. Distributions of mass ratio, total mass, initial binary separation (determined by MBH particle smoothing lengths), and formation redshifts (determined as the time at which particles come within a smoothing length of one another) are shown. The different lines (colours) correspond to different mass ratios that are strongly anti-correlated with total mass.

Fig. 2 shows the properties of MBHBs passing our selection cuts, grouped by mass ratio. Mass ratio (upper-left panel) is strongly anti-correlated with total mass (upper-right panel) due to both selection effects (e.g. at total masses just above the minimum mass, the mass ratio must be near unity) and astrophysical ones (e.g. the most massive MBHs, in large, central galaxies tend to merge more often with the lower mass MBH in small satellite galaxies). Binary separations (lower-left panel) are set by the smoothing length of MBH particles in Illustris. Once two MBH particles come within a smoothing length of one another, Illustris considers them a 'merger' event – which corresponds to the 'formation' (lower-right panel) of a binary in the simulations of this study.

3 BINARY HARDENING MODELS

Black hole encounters from Illustris determine the initial conditions for the binary population which are then evolved in our merger simulations. Throughout the 'hardening' process, where the binaries slowly coalesce over millions to billions of years, we assume uniformly circular orbits. In our models, we use information from the MBHB host galaxies to implement four distinct hardening mechanisms (BBR80): DF, stellar scattering in the 'LC',[8] VD from a circumbinary gaseous disc, and GW radiation.

3.1 Dynamical friction

DF is the integrated effect of many weak and long-range scattering events, on a gravitating object moving with a relative velocity through a massive background. The velocity differential causes an asymmetry that allows energy to be transferred from the motion of the massive object to a kinetic thermalization of the background population. In the case of galaxy mergers, DF is the primary mechanism of dissipating the initial orbital energy to facilitate coalescence of the galaxies, generally on time-scales comparable to the local dynamical time ($\sim 10^8$ yr). BHs present in the parent galaxies will tend to 'sink' towards each other in the same manner (BBR80) due to the background of stars, gas and DM. For a detailed review of DF in MBH systems, see Antonini & Merritt (2012).

The change in velocity of a massive object due to a single encounter with a background particle at a fixed relative velocity v and impact parameter b is derived (e.g. Binney & Tremaine 1987) following the treatment of Chandrasekhar (1942, 1943) by averaging the encounters over all possible angles to find

$$\Delta v = -2v \frac{m}{M+m} \frac{1}{1+(b/b_0)^2}, \tag{3}$$

where the characteristic (or 'minimum') impact parameter $b_0 \equiv G(M+m)/v^2$, for a primary object of mass M, in a background of bodies with masses m. The net deceleration on a primary mass is then found by integrating over distributions of stellar velocity (assumed to be isotropic and Maxwellian) and impact parameters (out to some maximum effective distance $b_{\max}$) which yields

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{2\pi G^2 (M+m)\rho}{v^2} \ln\left[1+(b_{\max}/b_0)^2\right], \tag{4}$$

for a background of mass density ρ. The impact parameters are usually replaced with a constant – the 'Coulomb Logarithm', $\ln\Lambda \equiv \ln\left(\frac{b_{\max}}{b_0}\right) \approx \frac{1}{2}\ln\left(1+\Lambda^2\right)$, such that,

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{2\pi G^2 (M+m)\rho}{v^2} \ln\Lambda. \tag{5}$$

In the implementation of equation (5), we use spherically averaged density profiles from the Illustris, remnant host galaxies. Modelling a 'bare', secondary MBH moving under DF through these remnants would clearly drastically underestimate the effective mass – which at early times is the MBH secondary in addition to its host galaxy. Over time, the secondary galaxy will be stripped by tidal forces and drag, eventually leaving behind the secondary MBH with only a dense core of stars and gas directly within its sphere of influence. We model this mass 'enhancement' by assuming that the effective DF mass is initially the sum of the MBH mass (M_2), and that of its host galaxy ($M_{2,\mathrm{host}}$), decreasing as a power law over a dynamical time τ_{dyn}, until only the MBH mass is left, i.e.

$$M_{\mathrm{DF}} = M_2 \left(\frac{M_2 + M_{2,\mathrm{host}}}{M_2}\right)^{1-t/\tau_{\mathrm{dyn}}}. \tag{6}$$

We calculate the dynamical time using a mass and radius from the remnant host galaxy. Specifically, we use twice the stellar half-mass radius $2R_{\star,1/2}$ and define $M_{2,\mathrm{host}}$ as the total mass within that radius, i.e.

$$\tau_{\mathrm{dyn}} = \left[\frac{4\pi\,(2R_{\star,1/2})^3}{3\,G M_{2,\mathrm{host}}}\right]^{1/2}. \tag{7}$$

The galaxy properties and derived dynamical times for all MBHB host galaxies we consider are shown in Fig. 3. Galaxy masses and radii are peaked at about $10^{11}\,\mathrm{M}_\odot$ and 10 kpc, respectively, with corresponding dynamical times around 100 Myr. For comparison, we also perform simulations using a fixed dynamical time of 1 Gyr for all galaxies, i.e. less-efficient stripping of the secondary galaxy.

[8] LC scattering and DF are different regimes of the same phenomenon, we separate them based on implementation.

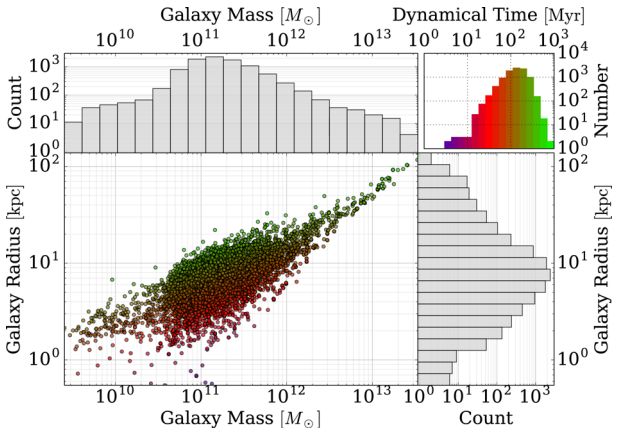

Figure 3. Stellar half-mass radii ($R_{\star,1/2}$) and total mass within $2 \cdot R_{\star,1/2}$ for all Illustris remnant host galaxies. Values for each galaxy are coloured by their dynamical times, calculated using equation (7), which are used for the 'enhanced' DF masses. The histogram in the upper right shows the distribution of dynamical times. Galaxy masses and radii are peaked at about 10^{11} $M_\odot$ and 10 kpc, respectively, with corresponding dynamical times around 100 Myr.

3.1.1 *Impact parameters and explicit calculations*

We have explored calculating the Coulomb logarithm explicitly, following BBR80 for the maximum impact parameter such that

$$b_{\max}(r) = \begin{cases} R_{\rm s} & R_{\rm s} < r, \\ (r/R_{\rm b})^{3/2}\, R_{\rm s} & R_{\rm h} < r < R_{\rm s}, \\ r\, R_{\rm h} & r < R_{\rm h}. \end{cases} \quad (8)$$

This effective maximum impact parameter is a function of binary separation[9] r – to account for the varying population of stars available for scattering and varying effectiveness of encounters. Equation (8) also depends on the characteristic stellar radius $R_{\rm s}$ ('$r_{\rm c}$' in BBR80), radius at which the binary becomes gravitationally bound, $R_{\rm b} = [M/(Nm_\star)]^{1/3} R_{\rm s}$, and radius at which the binary becomes 'hard', $R_{\rm h} \equiv (R_{\rm b}/R_{\rm s})^3 R_{\rm s}$.

Not only is this formalism complex, but it often produces unphysical results. For example, with this prescription the 'maximum' impact parameter not infrequently becomes less than the 'minimum' or larger than the distances that interact in the characteristic time-scales. After imposing a minimum impact parameter ratio of $b_{\max}/b_0 \geq 10$ (i.e. $\ln \Lambda \geq 2.3$), the results we obtained are generally consistent with using a constant Coulomb logarithm, with negligible effects on the resulting merger rates and GWB. We have also implemented an explicit integration over stellar distribution functions (see Section 3.2) and found the results to again be entirely consistent with equation (5) that is both computationally faster and numerically smoother. We believe the explicit impact parameter calculation is only valuable as a heuristic, and instead we use $\ln \Lambda = 15$, consistent with detailed calculations (e.g. Antonini & Merritt 2012). Similarly, in the results we present, we take the local stellar density as that given by spherically symmetric radial density profiles around the galactic centre instead of first determining, then marginalizing over, the stellar distribution functions.

[9] Note that we use the term 'binary separation' loosely, in describing the separation of the two MBHs even before they are gravitational bound.

3.1.2 *Applicable regimes*

There is a critical separation at which the back reaction of the decelerating MBH notably modifies the stellar distribution, and the DF formalism is no longer appropriate. Beyond this radius, the finite number of stars in the accessible region of parameter space to interact with the MBH(B) – the 'LC' (see Merritt 2013) – must be considered explicitly, discussed more thoroughly in Section 3.2. The 'LC' radius can be approximated as (BBR80)

$$\mathcal{R}_{\rm lc} = \left(\frac{m_\star}{M}\right)^{1/4} \left(\frac{R_{\rm b}}{R_{\rm s}}\right)^{9/4} R_{\rm s}. \quad (9)$$

Stars and DM are effectively collisionless, so they can only refill the LC on a slow, diffusive scattering time-scale. Gas, on the other hand, is viscous and supported thermally and by turbulent motion that can equilibrate it on shorter time-scales. In our fiducial model, we assume that for separations $r < \mathcal{R}_{\rm lc}$ the DF due to stars and DM is attenuated to low values, but that of gas continues down to smaller separations. We set the inner edge of gaseous DF based on the formation of a (circumbinary) accretion disc on small scales (as discussed in Section 3.3). The attenuation prescription given by BBR80 increases the DF time-scales by a factor

$$f_{\rm DF,LC} = \left(\frac{m}{M_\bullet}\right)^{7/4} N_\star \left(\frac{R_{\rm b}}{R_{\rm s}}\right)^{27/4} \left(\frac{\mathcal{R}_{\rm lc}}{r}\right), \quad (10)$$

where $N_\star = \frac{1}{M_\star} \int_0^r 4\pi r'^2 \rho_\star dr'$ is the number of stars available to interact with the binary. For all intents and purposes, this negates the effectiveness of DF for $r \lesssim \mathcal{R}_{\rm lc}$, such that without other hardening mechanisms (which become important at smaller scales), no MBHBs would coalesce within a Hubble time.

3.1.3 *DF hardening rates*

The resulting hardening time-scales, $\tau_{\rm h} = a/(da/dt)$, for our different DF implementations, are shown in Fig. 4. We show evolution for 'bare' MBH secondaries (red), in addition to effective masses enhanced ('enh') by the secondary's host galaxy for a dynamical time calculated using twice the 'stellar' half-mass radii (blue) or with a fixed 'Gyr' time-scale (green). In each case, we also compare between letting gas DF continue below $\mathcal{R}_{\rm lc}$ ('Gas Continues': darker colours, dotted lines) versus attenuating gas along with stars and DM ('Gas Cutoff': lighter colours, dashed lines). As a metric of the varying outcomes, we calculate the fraction of 'stalled' major mergers $\mathcal{F}_{\rm stall}$, defined as the number of major mergers (mass ratios $\mu \equiv M_2/M_1 > 0.1$, which are 6040 out of the full sample of 9270, i.e. ∼65 per cent) remaining at separations larger than 100 pc at redshift zero, divided by the total number of major mergers. Attenuation of the DF begins below 100 pc, so $\mathcal{F}_{\rm stall}$ are unaffected by whether the gas DF is also cutoff.

For $r \gtrsim 100$ pc, where the density of stars and especially DM dominate that of gas, the hardening rates are the same with or without a separate treatment of gas. Hardening differs significantly, however, between the 'bare' and enhanced models – with the latter hardening more than an order of magnitude faster at the largest separations[10] (∼10^4 pc). After a dynamical time, the 'stellar' enhancement runs out and the hardening rate approaches that of a bare MBH secondary by ∼10^3 pc. Still, the enhanced mass over

[10] Recall that binary separations are initialized to the MBH particle smoothing lengths, distributed between about 10^3 and 10^4 pc, so the total number of systems being plotted decreases over the same range.

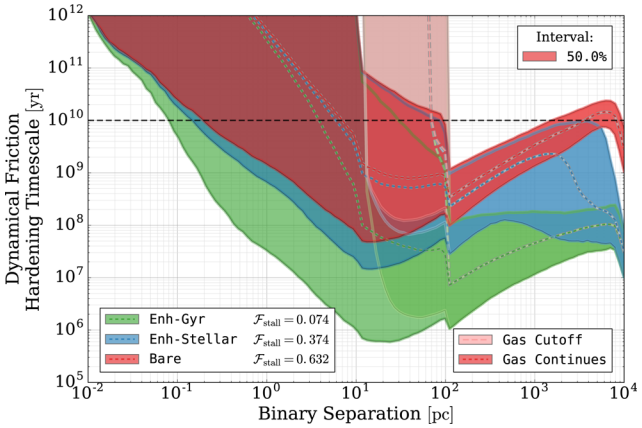

Figure 4. DF hardening time-scales for 50 per cent of our MBHBs around the median, under a variety of implementations. Cases in which a 'bare' secondary MBH migrates through the remnant host galaxy is compared to ones where the effective mass is *enhanced* to the secondary's host galaxy, decreasing as a power law over the course of a dynamical time ('Enh'; see equation 6). The dynamical time is calculated using twice the 'stellar' half-mass radius (see equation 7) shown in blue, or using a fixed 1'Gyr' time-scale shown in green. Allowing gaseous DF to continue below the attenuation radius $\mathcal{R}_{\rm lc}$ ('Gas Continues': darker regions, dotted lines) is compared to cutting off the gas along with stars and DM ('Gas Cutoff': lighter regions, dashed lines). $\mathcal{F}_{\rm stall}$ is the fraction of mergers with mass ratio $\mu > 0.1$, remaining at separations $r > 10^2$ pc, at redshift zero.

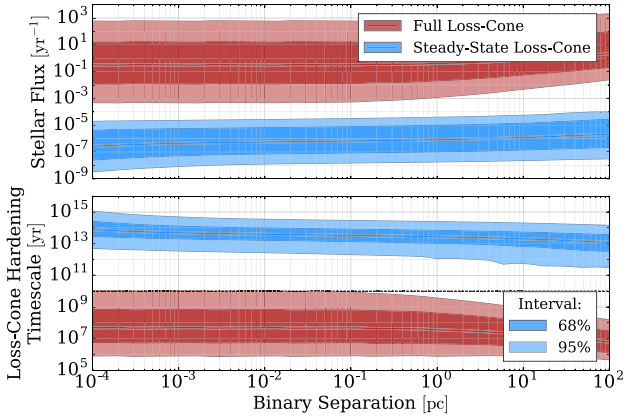

Figure 5. Scattering rates and hardening time-scales for full (red) and steady-state (blue) LCs. The bands represent 68 and 95 per cent of the population around the median. The difference between the two extremes of LC states is a stark six orders of magnitude, illustrating how strong an effect the LC can have on MBHB mergers. We use a simple, single parameter prescription to describe the state of the LC: the fraction, in log-space, between steady-state and full, $\mathcal{F}_{\rm refill}$ (see equation 11).

this time leads to a decrease in the fraction of stalled binaries from ∼63 to ∼37 per cent. When the mass enhancement persists for a gigayear – about a factor of 10 longer than typical dynamical times – a large fraction of MBHBs are able to reach parsec-scale separations before tidal stripping becomes complete, leading to only ∼7 per cent of major mergers stalling at large separations. The particular fraction of stalled systems is fairly sensitive to the total mass and mass-ratio cutoff, which we return to in Section 4.3.

Previous studies (e.g. Ravi et al. 2014) have assumed that DF is very effective at bringing MBHBs into the dynamically 'hard' regime (i.e. instantly in their models), after which stellar interactions must be calculated explicitly to model the remaining evolution. For comparison, in our results we also include a 'Force-Hard' model in which we assume that all binaries reach the hard regime ($r = R_{\rm h}$) over the course of a dynamically time.[11]

It has long been suggested that MBHBs could stall at kiloparsec-scale separations (e.g. Yu 2002), but only recently has this effect been incorporated into population hardening models and studied specifically (e.g. McWilliams et al. 2014; Kulier et al. 2015). By better understanding the time-scales over which stalled systems could be observable, we can use observed dual-AGNs to constrain the hardening process and the event rates of MBHB encounters.

3.2 Stellar loss-cone scattering

The population of stars that are able to interact (scatter) with the MBHB is said to occupy the 'LC' (Merritt 2013), so named because it describes a conical region in parameter space. When stars are scattered out of the LC faster than they can be replenished, the LC becomes depleted and the DF description, which considers a relatively static background, becomes inconsistent. One approach to compensate for this is to add an 'attenuation' factor, as described in the previous section (Section 3.1). Physically, a steady state must be dynamically realized in which stars are diffused into the outer edges of the LC via two-body relaxation at the same rate at which stars are scattering out by the central MBH(B). The largest uncertainty in the MBHB merger process is likely understanding the nature of this equilibrium state, and how it is affected by realistic galaxy-merger environments.

The LC has been extensively explored in both the context of MBH binary hardeningand TDE. The two cases are almost identical, differing primarily in that for TDE calculations, only impact parameters small enough to cause disruptions are of interest – while for binary hardening, weaker scattering events are still able to extract energy from the MBH or MBHB. For binary hardening, there is an additional ambiguity in two subtly distinct regimes: first, where stars scatter with individual BHs, decelerating them analogously to the case of DF (but requiring the LC population to be considered explicitly). Secondly, for a truly *bound* binary, stars can interact with the combined system – in a three-body scattering – and extract energy from the binary pair together. In our prescriptions we do not distinguish between these cases, considering them to be a spectrum of the same phenomenon instead.

We use the model for LC scattering given by Magorrian & Tremaine (1999), corresponding to a single central object in a spherical (isotropic) background of stars. We adapt this prescription simply by modifying the radius of interaction to be appropriate for scattering with a binary instead of being tidally disrupted by a single MBH. This implementation is presented in pedagogical detail in Appendix C. Scattering rates are calculated corresponding to both a 'full' LC (equation C6), one in which it is assumed that the parameter space of stars is replenished as fast as it is scattered, and a 'steady-state' LC (equation C7), in which diffusive two-body scattering sets the rate at which stars are available to interact with the binary.

The interaction rates (fluxes) of stars scattering against all MBHBs in our sample are shown in the upper panel of Fig. 5 for both full (red, equation C6) and equilibrium (blue, equation C7) LC configurations. The interaction rates for full LC tend to be about six orders of magnitude higher than equilibrium configurations. The

[11] Calculated using the 'Stellar' prescription: twice the stellar half-mass radius, and the mass within it.

resulting binary hardening time-scales are shown in the lower panel of Fig. 5 – reaching four orders of magnitude above and below a Hubble time. Clearly, whether the LC is in the relatively low equilibrium state or is more effectively refilled has huge consequences for the number of binaries that are able to coalesce within a Hubble time.

Many factors exist which may contribute to quickly refilling the LC. In general, any form of asymmetry in the potential will act as an additional perturber – increasing the thermalization of stellar orbits. The presence of an MBHB is premised on there having been a recent galaxy merger – implying that significant asymmetries and aspherical morphologies may exist. Even ignoring galaxy mergers, galaxies themselves are triaxial (e.g. Illingworth 1977; Leach 1981), many have bars (e.g. Sellwood & Wilkinson 1993), and in star-forming galaxies there are likely large, dense molecular clouds (e.g. Young & Scoville 1991) which could act as perturbers. Finally, because binary lifetimes tend to be on the order of the Hubble time while galaxies typically undergo numerous merger events (e.g. Rodriguez-Gomez et al. 2015), subsequent merger events can lead to triple MBH systems (see Section 4.4) that could be very effective at stirring the stellar distribution. While there is some evidence that for galaxy-merger remnants the hardening rate can be nearly that of a 'full' LC (e.g. Khan et al. 2011), the community seems to be far from a consensus (e.g. Vasiliev, Antonini & Merritt 2014), and a purely numerical solution to the LC problem is currently still unfeasible.

In the future, we plan on incorporating the effects of triaxiality and tertiary MBHs to explore self-consistent LC refilling. In our current models, we introduce an arbitrary dimensionless parameter – the logarithmic 'refilling fraction' $\mathcal{F}_{\rm refill}$ (in practice, but not requisitely, between [0.0, 1.0]) – to logarithmically interpolate between the fluxes of steady-state ($F_{\rm lc}^{\rm eq}$) and full LC ($F_{\rm lc}^{\rm full}$), i.e.

$$F_{\rm lc} = F_{\rm lc}^{\rm eq} \cdot \left(\frac{F_{\rm lc}^{\rm full}}{F_{\rm lc}^{\rm eq}} \right)^{\mathcal{F}_{\rm refill}} . \quad (11)$$

3.3 Viscous hardening by a circumbinary disc

The density of gas accreting on to MBHs can increase significantly at separations near the accretion or Bondi radius, $R_{\rm b} \equiv GM_\bullet/c_s^2$, where the sound-crossing time is comparable to the dynamical time. The nature of accretion flows on to MBHs near and within the Bondi radius is highly uncertain, as observations of these regions are currently rarely resolved (e.g. Wong et al. 2011). If a high-density, circumbinary disc is able to form, the VD can be a significant contribution to hardening the binary at separations just beyond the GW-dominated regime (BBR80; Gould & Rix 2000; Escala et al. 2005). Galaxy mergers are effective at driving significant masses of gas to the central regions of post-merger galaxies (Barnes & Hernquist 1992), enhancing this possibility.

We implement a prescription for VD due to a circumbinary accretion disc following HKM09 based on the classic thin-disc solution of Shakura & Sunyaev (1973), broken down into three, physically distinct regions (Shapiro & Teukolsky 1986). These regions are based on the dominant pressure (radiation versus thermal) and opacity (Thomson versus free–free) contributions such that the regions are defined as

(1) $r < r_{12}$, radiation pressure and Thomson-scattering opacity dominated;

(2) $r_{12} < r < r_{23}$, thermal pressure and Thomson-scattering opacity dominated;

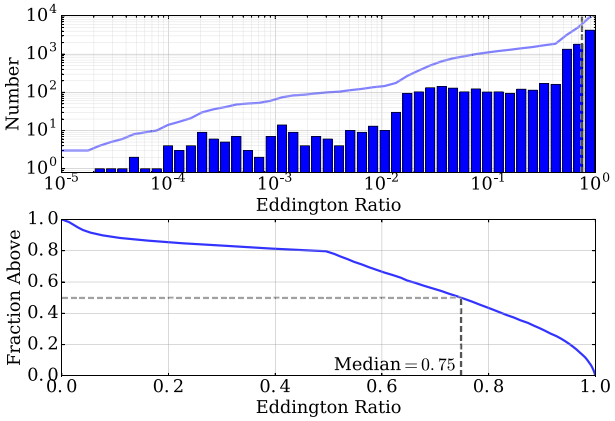

Figure 6. Accretion rates at the time of binary formation for all MBHBs in our analysis. Values are measured as a fraction of the Eddington accretion rate, $\dot{M}_{\rm Edd} \equiv L_{\rm Edd}/\varepsilon_{\rm rad}c^2$, where we use $\varepsilon_{\rm rad} = 0.1$. Recall that in Illustris MBHs merge when they come within a smoothing length of one another – corresponding to the formation of a binary in our models. The accretion rates from Illustris are those of the resulting remnant MBH and are limited to Eddington ratios of unity. The upper panel shows the distribution of accretion rates (bars) and cumulative number (line), which are strongly biased towards near-Eddington values. The lower panel shows the cumulative distribution of accretion rates above each value (note the different x-axis scaling). The median Eddington ratio of 0.75 is overplotted (grey, dashed line).

(3) $r_{23} < r$, thermal pressure and free–free opacity dominated.

Recall that in Illustris, 'mergers' occur when MBH particles come nearer than a particle smoothing length, after which the MBHs are combined into a single, remnant MBH. We track these remnant particles and use their accretion rates ($\dot{M}$) to calibrate the circumbinary disc's gas density. The distribution of Eddington ratios ($\dot{M}/\dot{M}_{\rm Edd}$) for these remnants, at the time of their formation, is presented in Fig. 6, showing a clear bias towards near-Eddington accretion rates. MBH remnants tend to have enhanced accretion rates for a few gigayear after merger and have higher average accretion than general BHs (Blecha et al. 2016). In Illustris, MBH accretion (and thus growth in mass) is always limited to the Eddington accretion rate. We introduce a dimensionless parameter $f_{\rm Edd}$ to modulate those accretion rates, i.e. $\dot{M} = {\rm Min}\left[\dot{M}_{\rm ill}, f_{\rm Edd}\dot{M}_{\rm Edd}\right]$.

Otherwise, in the formalism of HKM09, we use their fiducial parameter values[12] and assume an α-disc (i.e. the viscosity depends on total pressure, not just thermal pressure as in a so-called β-disc). The outer disc boundary is determined by instability due to self-gravity (SG) – measured as some factor times the radius, r_Q, at which the Toomre parameter reaches unity, i.e. $\mathcal{R}_{\rm SG} = \lambda_{\rm sg} r_Q$. In our fiducial model, $\lambda_{\rm sg} = 1$, and variations in this parameter have little effect on the overall population of binaries. After marginalizing over all systems, changes to the different viscous disc parameters tend to be largely degenerate: shifting the distribution of hardening time-scales and the GWB amplitude in similar manners.

VD hardening time-scales tend to decrease with decreasing binary separations. They thus tend to be dominated by *Region-3* – at larger separations. For near-Eddington accretion rates, however, *Region-2* and especially *Region-3* tend to be SG unstable, fragmenting the disc and eliminating VD altogether. For this reason, when

[12] Mean mass per electron, $\mu_{\rm e} = 0.875$; viscosity parameter $\alpha = 0.3$; radiative efficiency, $\varepsilon_{\rm rad} = 0.1$; temperature-opacity constant, $f_{\rm T} = 0.75$; and disc-gap size, $\lambda_{\rm gap} = 1.0$ ('λ' in HKM09).

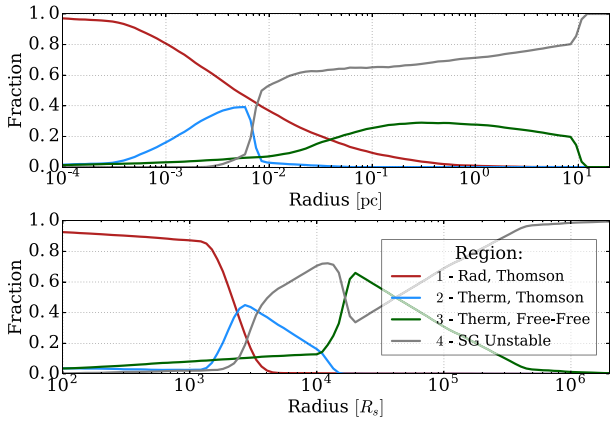

Figure 7. Fraction of binaries in each circumbinary disc region as a function of radius. Radii are given both physical units (upper panel) and Schwarzschild radii (R_s, lower panel), the latter highlighting the intrinsic scalings. *Region-4* are locations in the disc which are unstable to 'SG', defined using the Toomre parameter for each of *Region-2* and *Region-3*.

high accretion rate systems have dynamically important discs, they tend to be in *Region-1*. In these cases, *Region-1* extends to large enough radii such that for most masses of interest, GW emission will only become significant well within that region of the disc. Lower accretion rate systems are stable out to much larger radii, allowing many binaries to stably evolve through *Region-2* and *Region-3*. These regions also cutoff at smaller separations, meaning that GW emission can become significant outside of *Region-1*.

Decreased disc densities mean less drag, but at the same time sufficiently high densities lead to instability, making the connection between accretion rate and VD effectiveness non-monotonic. This is enhanced by gaseous DF, with an inner cutoff radius determined by the SG radius (see Section 3.1.2). In other words, gaseous DF is allowed to continue down to smaller radii when the outer disc regions become SG unstable. We impose an additional, absolute upper limit to the SG instability radius of $\mathcal{R}_{\rm SG,Max} = 10$pc, i.e. $\mathcal{R}_{\rm SG} = {\rm Min}\left[\lambda_{\rm sg}\, r_Q,\ \mathcal{R}_{\rm SG,Max}\right]$, to keep the outer edge of discs physically reasonable.

Fig. 7 shows the fraction of Illustris binaries in the different regions of the disc for our fiducial model. Only a fraction of MBHBs spend time in *Region-2* and *Region-3* discs, and even that is only for a small region of log-radius space. While almost all systems do enter *Region-1* by about $10^3\ R_s$, GW hardening has, in general, also become significant by these same scales.

In addition to the spatially distinct disc regions, different types of migration occur depending on whether the disc or the secondary MBH is dynamically dominant (analogous to the distinction between 'planet-dominated' and 'disc-dominated', Type II migration in planetary discs – see HKM09). If disc-dominated, the system hardens on the viscous time-scale $\tau_{\rm v}$, whereas if the secondary is dominant – as is the typical case in our simulations, the time-scale is slowed by a factor related to the degree of secondary dominance.

The resulting hardening time-scales due to VD from a circumbinary disc are shown in Fig. 8. The more massive binaries ('heavy', $M > 10^9\ {\rm M}_\odot$) have orders of magnitude shorter VD hardening times, but are quite rare. The overall trend (grey, cross-hatched) follows the less massive ('light') systems. The changes in slope of the 'light' populations (especially 'Major', $\mu \equiv M_2/M_1 > 1/4$) at separations larger than 10^{-3} pc are due to transitions in disc regions. The 'heavy' systems tend to have SG unstable Regions 2 and 3, and thus harden more smoothly, predominantly due to *Region-1*.

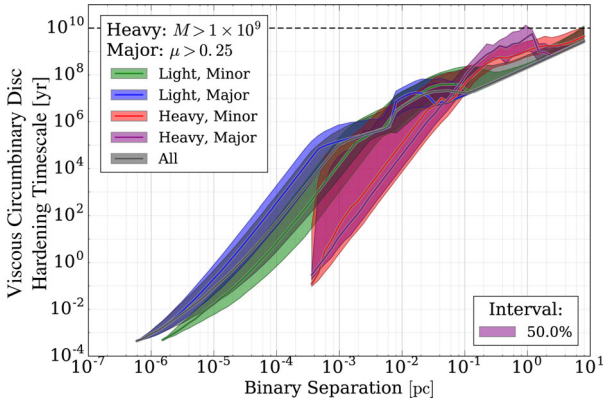

Figure 8. Hardening time-scales due to circumbinary disc drag for binaries grouped by total mass and mass ratio. Light and heavy MBHB are separated by total masses below and above $10^9\ {\rm M}_\odot$ respectively; and minor and major based on mass ratios (μ) below and above 1/4. The median and surrounding 50 per cent intervals for all MBHB systems are shown in grey, showing that 'light' systems dominate the bulk of the binary population. Heavy, and especially heavy-major, systems tend to harden orders of magnitude faster than lighter ones. The light population (especially major) exhibits non-monotonicities at intermediate separations ($\sim 10^{-3}$ to 10^{-1} pc) indicative of changes between disc regions. Heavy systems (especially minor), on the other hand, show much smoother hardening rates consistent with moving through primarily *Region-1*.

Fig. 9 compares median hardening rates in simulations including VD (solid) with those without a disc (dashed). In the former case, the purely VD hardening rates are also shown (dotted) – with the maximum disc cutoff $\mathcal{R}_{\rm SG,Max}$, clearly apparent at 10 pc. Different DF prescriptions are shown, with mass enhancements over dynamical times calculated using the 'stellar' method in blue and fixed 1 Gyr time-scales in red (see Section 3.1). The upper panel shows a moderately refilled LC ($\mathcal{F}_{\rm refill} = 0.6$), while in the lower panel the LC is always full ($\mathcal{F}_{\rm refill} = 1.0$). The effects of VD are clearly apparent below a few 10^{-2} pc in all models and up to $\mathcal{R}_{\rm SG,Max} = 10$pc when $\mathcal{F}_{\rm refill} = 0.6$. In the $\mathcal{F}_{\rm refill} = 1.0$ case, LC hardening dominates to much smaller separations, making the VD effects minimal for the overall hardening rates. This is echoed in the changes in coalescing fractions[13] between the VD and No-VD cases: for $\mathcal{F}_{\rm refill} = 0.6$, VD increases $\mathcal{F}_{\rm coal}$ by $\sim$10 per cent, while for $\mathcal{F}_{\rm refill} = 1.0$, it is only increased by $\sim$2 per cent.

The circumbinary disc is SG unstable for many systems, and thus the median hardening rates including VD are often intermediate between the purely VD time-scales and simulations without VD at all (e.g. seen between $\sim 5 \times 10^{-3}$ and 1 pc in the upper panel of Fig. 9). At smaller separations ($\lesssim 10^{-3}$ pc), where LC is almost always subdominant to VD and/or GW hardening, the VD decreases the median hardening time-scales by between a factor of a few and an order of magnitude. Notably, Fig. 9 shows that the scaling of hardening rate with separation below about 10^{-3} pc is very similar between that of GW (which is dominant in the No-VD case) and VD. At these small scales, $\gtrsim 80$ per cent of our binaries are in disc *Region-1* (see Fig. 7), which has a viscous hardening rate, $\tau_{\rm v,1} \propto r^{7/2}$ (HKM09, equation 21a), compared to a very similar scaling for GWs, $\tau_{\rm gw} \propto r^4$ (see Section 3.4). Thus, even when VD dominates hardening into the mpc-scale regime, we do not expect the GWB

[13] $\mathcal{F}_{\rm coal}$, the fraction of systems with mass ratio $\mu > 0.1$ which coalesce by $z = 0$.

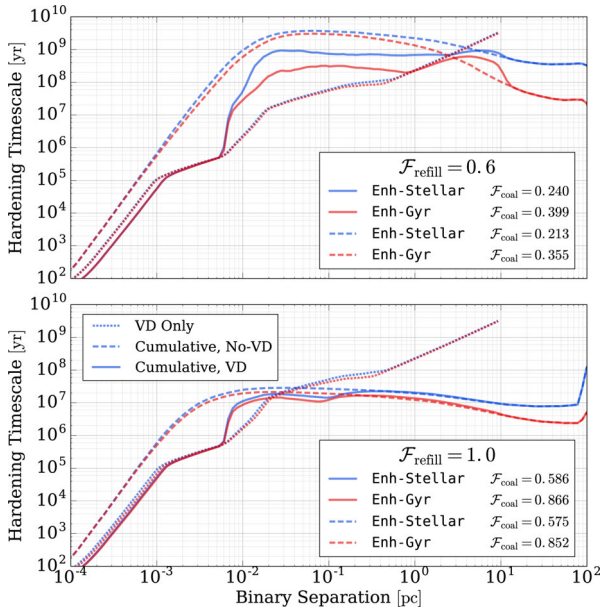

Figure 9. Median hardening time-scales with and without drag from a circumbinary Viscous Disc (VD) for different hardening models. Different LC refilling parameters ($\mathcal{F}_{\rm coal} = 0.6$, upper; and $\mathcal{F}_{\rm coal} = 1.0$, lower) are compared against DF models ('Enh-Stellar', blue; 'Enh-Gyr', red). Each line type shows a different hardening rate: VD only (dotted), and cumulative hardening rates with VD (solid) and without VD (dashed). With $\mathcal{F}_{\rm refill} = 0.6$, VD effects are apparent up to the disc cutoff, $\mathcal{R}_{\rm SG,Max} = 10$pc, whereas for $\mathcal{F}_{\rm refill} = 1.0$, LC scattering dominates down to $\sim 10^{-2}$ pc. Similarly, the presence of a circumbinary disc has a much more pronounced effect on the fraction of high mass-ratio ($\mu > 0.1$) systems which coalesce by redshift zero ($\mathcal{F}_{\rm coal}$), which are indicated in the legends. At very low separations the cumulative (with-VD) hardening rate is very nearly the purely VD rate, showing its importance down to very small scales. At intermediate separations the 'cumulative, VD' rate is intermediate between the 'cumulative, No-VD' and the 'VD only' model, showing that the disc is only present in some fraction of systems at those scales.

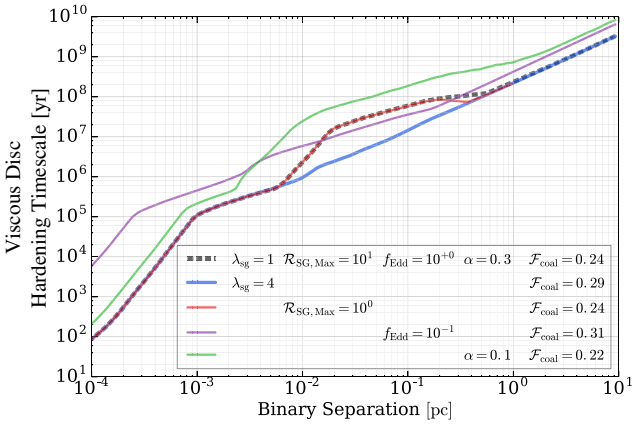

Figure 10. Median hardening time-scales comparing our fiducial Viscous Disc (VD) parameters (black, dashed) with other configurations. The radius at which the disc becomes SG unstable is $\mathcal{R}_{\rm SG} = {\rm Min}\left[\lambda_{\rm sg}\, r_Q,\ \mathcal{R}_{\rm SG,Max}\right]$, where r_Q is the radius at which the Toomre parameter reaches unity. $\lambda_{\rm sg}$ scales the SG-unstable radius, while $\mathcal{R}_{\rm SG,Max}$ is a maximum cutoff radius. α is the standard disc viscosity parameter, and $f_{\rm Edd}$ limits the maximum accretion rate, i.e. $\dot{M} = {\rm Min}\left[\dot{M}_{\rm ill},\ f_{\rm Edd}\dot{M}_{\rm Edd}\right]$. While this variety of VD parameters produces hardening rates varying by two orders of magnitude, the resulting changes to the coalescing fraction $\mathcal{F}_{\rm coal}$ is fairly moderate as VD is often subdominant to LC scattering at larger radii and to GW emission at smaller radii. Effects on $\mathcal{F}_{\rm coal}$ can be counterintuitive, for example decreasing the accretion rate (purple line) increases the median hardening time-scale, but increases the coalescing fraction because the disc becomes SG-stable for a larger fraction of binaries. Each model uses $\mathcal{F}_{\rm refill} = 0.6$, and the 'Enh-Stellar' DF.

spectrum from binaries in *Region-1* to deviate significantly from the canonical $-2/3$ power law.

Differences in median hardening time-scales, solely due to VD, are compared for a variety of VD parameters in Fig. 10. A simulation with our fiducial disc parameters is shown in dashed black, and each colour shows variations in a different parameter. Decreasing the viscosity of the disc (α, green) amounts to a proportional increase in the hardening time-scale and a decrease in the coalescing fractions ($\mathcal{F}_{\rm coal}$). Decreasing the maximum disc radius ($\mathcal{R}_{\rm SG,Max}$, red) decreases the overall effectiveness of VD, but because gaseous DF continues in its place, the coalescing fraction remains unchanged. While $\mathcal{R}_{\rm SG,Max}$ changes the maximum disc radius, $\lambda_{\rm sg}$ changes the radii at which *Region-2* and *Region-3* become SG-unstable directly (i.e. even well within the maximum cutoff radius). Increasing $\lambda_{\rm sg}$ by a factor of 4 (blue) significantly increases the number of MBHBs with SG-stable *Region-2*[14] between $\sim 10^{-2}$ and 10^{-1} pc, increasing the overall coalescing fraction.

Decreasing the accretion rates ($f_{\rm Edd}$, purple), and thus disc densities, increases the hardening time-scales similarly to changing the viscosity (green). At the same time, significantly more systems have stable outer discs. This has the effect of increasing coalescing fractions noticeably despite the increased median hardening time-scales. In addition to increased outer-disc stability, the transition between disc regions is also inwards. A large number of MBHBs at small separations ($\lesssim 10^{-3}$ pc) remain in disc *Region-2* instead of transitioning to *Region-1*. This softens the scaling of hardening rate with separation to $\tau_{\rm v,2} \propto r^{7/5}$ (HKM09, equation 21b), which differs much more significantly from purely GW-driven evolution.

[14] See the transition in Fig. 7 between *Region-2* (light-blue) and *Region-4* (grey) at $\sim 10^{-2}$ pc.

3.4 Gravitational-wave emission

GW radiation will always be the dominant dissipation mechanism at the smallest binary separations – within hundreds to thousands of Schwarzschild radii. GW hardening depends only on the constituent masses (M_1 and M_2) of the MBHB, their separation, and the system's eccentricity. The hardening rate can be expressed as (Peters 1964)

$$\frac{{\rm d}a}{{\rm d}t} = -\frac{64\,G^3}{5\,c^5}\,\frac{M_1\,M_2\,(M_1+M_2)}{a^3}\,\frac{\left(1+\frac{73}{24}e^2+\frac{37}{96}e^4\right)}{\left(1-e^2\right)^{7/2}}, \qquad (12)$$

where a is the semi-major axis of the binary and e is the eccentricity. In our treatment we assume that the eccentricities of all MBHBs are uniformly zero, in which case, equation (12) can easily be integrated to find the time to merge,

$$t_{\rm GW} = \frac{5c^5}{64G^3}\,\frac{a_0^4 - \mathcal{R}_{\rm crit}^4}{M_1 M_2 M},$$

$$\approx 10^{10}\,{\rm yr}\left(\frac{a_0}{0.01\,{\rm pc}}\right)^4\left(\frac{M}{2\times 10^7\,{\rm M}_\odot}\right)^{-3}\left(\frac{2+\mu+1/\mu}{4}\right), \qquad (13)$$

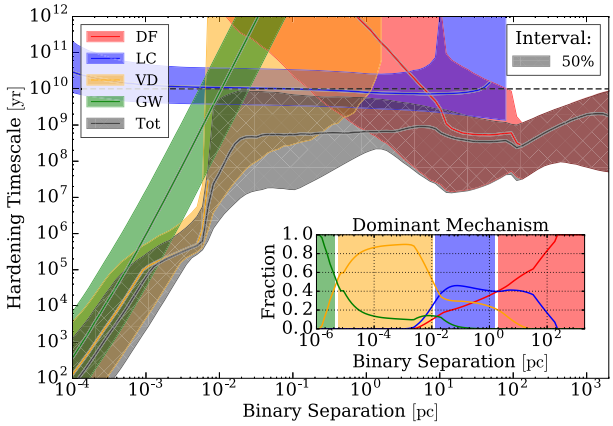

Figure 11. Binary hardening time-scales versus binary separation by mechanism. Coloured lines and bands show the median and 50 per cent intervals for DF, LC scattering, VD, and GW emission with the total hardening rate shown by the grey, hatched region. The inset panel shows the fraction of binaries dominated by each mechanism. This simulation uses our fiducial parameters (e.g. $\mathcal{F}_{\rm refill} = 0.6$), with 'Stellar' DF-mass enhancement. The binary hardening landscape is very similar to that outlined by BBR80, but the details are far more nuanced. For a comparison with alternate models, see Fig. A3.

for a total mass $M = M_1 + M_2$, mass ratio $\mu \equiv M_2/M_1$, initial separation a_0, and critical separation $\mathcal{R}_{\rm crit}$. In practice, we assume that the GW signal from binaries terminates at the Innermost Stable Circular Orbit (ISCO), at which point the binary 'coalesces'; i.e. $\mathcal{R}_{\rm crit} = \mathcal{R}_{\rm isco}(J = 0.0) = 3R_{\rm s}$. For an equal-mass binary, with median Illustris MBH masses[15] of about 10^7 M$_\odot$, the binary needs to come to a separation of $\sim$0.01 pc ($\sim 10^4\,R_{\rm s}$), to merge within a Hubble time. Characteristic time-scales and separations for (purely) GW-driven inspirals across total mass and mass-ratio parameter space are plotted in Fig. A2. While the absolute most-massive MBHB can merge purely from GW emission starting from a parsec, the bulk of physical systems, at 10^6–10^8 M$_\odot$, must be driven by environmental effects to separations of the order of 10^{-3} to 10^{-2} pc ($\sim$500–5000 $R_{\rm s}$) to coalesce by redshift zero.

4 RESULTS

The hardening time-scales for all MBHBs are plotted against binary separation in Fig. 11, broken down by hardening mechanism. This is a representative model with a moderate LC refilling fraction $\mathcal{F}_{\rm refill} = 0.6$ (see Section 3.2), using the 'Enh-Stellar' DF (see Section 3.1). This is the fiducial model for which we present most results, unless otherwise indicated. The inset shows the fraction of binaries with hardening rates dominated by each mechanism. DF is most important at large radii soon after binaries form, until LC scattering takes over at $\sim$1 pc. The median hardening time-scale remains fairly consistent at a few times 100 Myr, down to $\sim 10^{-2}$ pc at which point VD drives the bulk of systems until GW emission takes over at separations below 10^{-5} pc, where the typical hardening time-scale reaches years. The landscape of hardening time-scales for alternative DF prescriptions and LC refilling fractions is shown in the Appendix (Fig. A3).

[15] After typical selection cuts, described in Section 2.

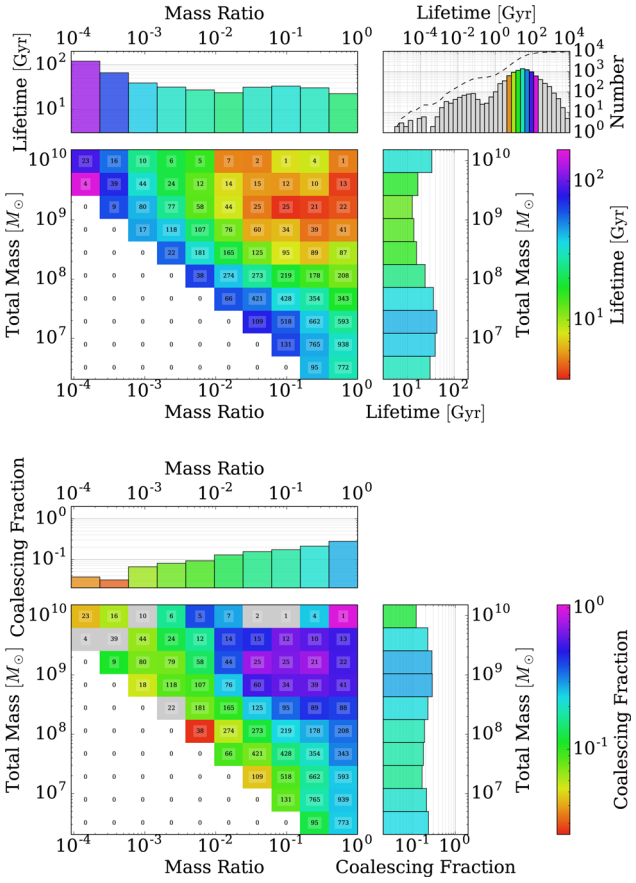

Figure 12. Binary lifetimes (upper) and coalescing fractions (lower) for our fiducial model with a moderate DF and LC refilling ('Enh-Stellar' and $\mathcal{F}_{\rm refill} = 0.6$, respectively). The overall distribution of MBHB lifetimes is shown in the upper-rightmost panel, with the cumulative distribution plotted as the dashed line. The median lifetime is $\sim$30 Gyr overall, but is significantly shorter for MBHBs with either high total masses, or nearly equal mass ratios. For this group, the coalescing fractions are near unity. Grey bins in the lower panel correspond to those with no binaries that coalesce by redshift zero. While systems with the highest masses and mass ratios tend to have much shorter lifetimes, they also form at low redshifts with less time to coalesce.

4.1 Binary lifetimes

Characteristic hardening time-scales are often many 100 Myr, and MBHBs typically need to cross eight or nine orders of magnitude of separation before coalescing. The resulting lifetimes of MBHBs can thus easily reach a Hubble time. Fig. 12 shows binary lifetimes (upper panels) and the fraction of systems that coalesce by $z = 0$ (lower panels) for our fiducial model. Systems are binned by total mass and mass ratio, with the number of systems in each bin indicated. The plotted lifetimes are median values for each bin, with the overall distribution shown in the upper-rightmost panel. Grey values are outside of the range of binned medians, and the cumulative distribution is given by the dashed line.

The lifetime distribution peaks near the median value of 29 Gyr, with only $\sim$7 per cent of lifetimes at less than 1 Gyr. About 20 per cent of all MBHBs in our sample coalesce before redshift zero. Systems involving the lowest mass black holes[16] (i.e. down and left) tend towards much longer lifetimes. Overall, lifetimes and

[16] Recall we require MBH masses of at least 10^6 M$_\odot$.

coalescing fractions are only mildly correlated with total mass or mass ratio, when marginalizing over the other. For systems with total masses $M > 10^8\,\mathrm{M}_\odot$, the coalescing fraction increases to 23 per cent, and for mass ratios $\mu > 0.2$, only slightly higher to 26 per cent. In general, examining slightly different total-mass or mass-ratio cutoffs has only minor effects on lifetimes and coalescing fractions.

There is a strong trend towards shorter lifetimes for simultaneously high total masses and massratios (i.e. up and right), where median lifetimes are *only* a few gigayear. Considering both $\mu > 0.2$ and at the same time $M > 10^8\,\mathrm{M}_\odot$, coalescing fractions reach 45 per cent. A handful of MBHBs which coalesce after $\lesssim$1 Myr ($\sim$0.3 per cent) tend to involve MBHs in overmassive galaxies (i.e. galaxy masses larger than expected from MBH–host scalings) with especially concentrated stellar and/or gas distributions. There are a handful of high mass ratio, and highest total mass MBHB ($M \sim 10^{10}\,\mathrm{M}_\odot$) systems showing a noticeable decrease in coalescing fraction. These systems form at low redshifts and do not have time to coalesce despite relatively short lifetimes.

Lifetimes and coalescing fractions for an always full LC ($\mathcal{F}_{\rm refill} = 1.0$) are shown in Fig. A4. The median value of the lifetime distribution shifts down to $\sim$8 Gyr, with $\sim$24 per cent under 1 Gyr. The coalescing fractions increase similarly, and systems which are *either* high massratio ($\mu \gtrsim 0.2$) *or* high total mass ($M \gtrsim 10^8\,\mathrm{M}_\odot$) generally coalesce by redshift zero. Specifically, the coalescing fractions are 50 and 61 per cent for all systems and those with $\mu > 0.2$ respectively. Considering only $M > 10^8\,\mathrm{M}_\odot$, fractions increase to 54 and 99 per cent.

Cumulative distributions of MBHB lifetimes are compared in Fig. 13 for a variety of LC refilling factors (colours) and our three primary DF prescriptions (panels; see Section 3.1). The first two panels correspond to prescriptions where the effective masses used in the DF calculation are the sum of the secondary MBH mass and the mass of its host galaxy. To model stripping of the secondary galaxy during the merger process, the effective mass decreases as a power law to the 'bare' MBH mass after a dynamical time. The 'Enh-Stellar' model (upper) calculates the dynamical time at twice the stellar half-mass radius and the mass there enclosed. The 'Enh-Gyr' model (middle), on the other hand, uses a fixed 1 Gyr time-scale – almost a factor of 10 longer than the median 'stellar'-calculated value. Finally, the 'Force-Hard' model (lower) uses the 'bare' secondary MBH mass, but the binary is forced to the hard binary regime (generally 1–10 pc) over the course of a dynamical time (calculated in the 'stellar' manner). Each colour of line indicates a different LC refilling fraction, from always full ($\mathcal{F}_{\rm refill} = 1.0$; blue) to the steady state ($\mathcal{F}_{\rm refill} = 0.0$; orange). The fraction of high mass-ratio ($\mu > 0.1$) systems which coalesce by redshift zero ($\mathcal{F}_{\rm coal}$) is also indicated in the legends.

The high mass-ratio coalescing fractions tend to vary by almost a factor of 4 depending on the LC state, while the varying DF prescriptions have less than a factor of 2 effect. Median lifetimes change considerably, however, even between DF models, for example with $\mathcal{F}_{\rm refill} = 1.0$, the median lifetime for the 'Enh-Stellar' model is 7.7 Gyr, while that of 'Enh-Gyr' is only about 0.42 Gyr. Apparently, with a full LC, DF at large scales tends to be the limiting factor for most systems. While the highest overall $\mathcal{F}_{\rm coal}$ occurs for 'Force-Hard' and $\mathcal{F}_{\rm refill} = 1.0$, it takes almost an order of magnitude longer for the first $\sim$10 per cent of systems to coalesce than in either of the 'Enh' models. The effects of DF on the lifetimes of the first systems to merge are fairly insensitive to the LC state. There are thus cases where DF can be effective at driving some systems to coalesce very rapidly. At the same time, for the bulk of systems,

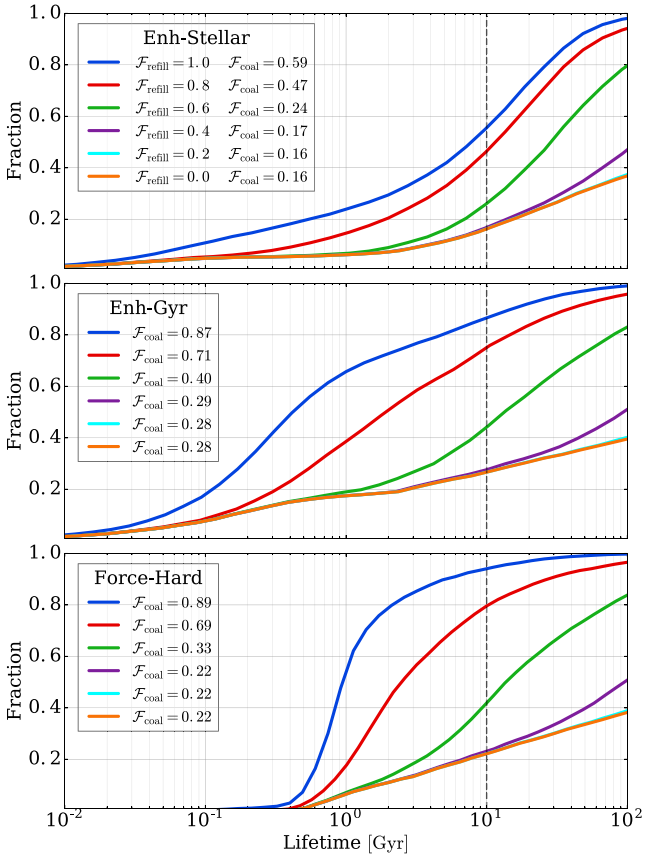

Figure 13. Cumulative distributions of binary lifetimes for a variety of DF and LC parameters. The 'Enh-Stellar', 'Enh-Gyr', and 'Force-Hard' DF models are shown in each panel, and the colours of lines indicate the LC refilling fraction. The fraction of high mass-ratio ($\mu > 0.1$) systems coalescing by $z = 0$ are given in the legend ($\mathcal{F}_{\rm coal}$). $\mathcal{F}_{\rm refill}$ is the dominant factor determining the lifetime distribution, but the DF model significantly affects the earliest merging systems, and overall fraction of coalescing systems.

after hardening past kiloparsec scales the remaining lifetime can be quite substantial. For $\mathcal{F}_{\rm refill} \lesssim 0.6$, neither the precise LC refilling fraction nor the DF model makes much of a difference after the first 10–30 per cent of systems coalesce. In these cases, the most massive systems with high mass ratios coalesce fairly rapidly regardless, but the smaller more extreme mass-ratio systems take many Hubble times to merge.

Fig. 14 shows the distribution of formation (black) and coalescence (coloured) redshifts resulting from a variety of binary evolution models. Each panel shows a different DF prescription, and two LC refilling parameters are shown: always full, $\mathcal{F}_{\rm refill} = 1.0$ (blue), and our fiducial, moderately refilled value of $\mathcal{F}_{\rm refill} = 0.6$ (green). A handful of events have been cutoff at low redshifts ($z < 10^{-3}$) where the finite volume of the Illustris simulations and cosmic variance becomes important. Median redshifts for each distribution are overplotted (dashed), along with their corresponding look-back times. The median formation redshift for our MBHBs is $z = 1.25$ (look-back time of $\sim$8.7 Gyr). For a full LC and the stronger DF models, 'Enh-Gyr' and 'Force-Hard', the median coalescence redshifts are delayed to $z \sim 1.0$ and $z \sim 0.9$, respectively – i.e. by about a gigayear. For our more modest, fiducial DF prescription, 'Enh-Stellar', even the full LC case still delays the median coalescence redshift to $z \sim 0.6$, about 3 Gyr after the peak of MBHB formations.

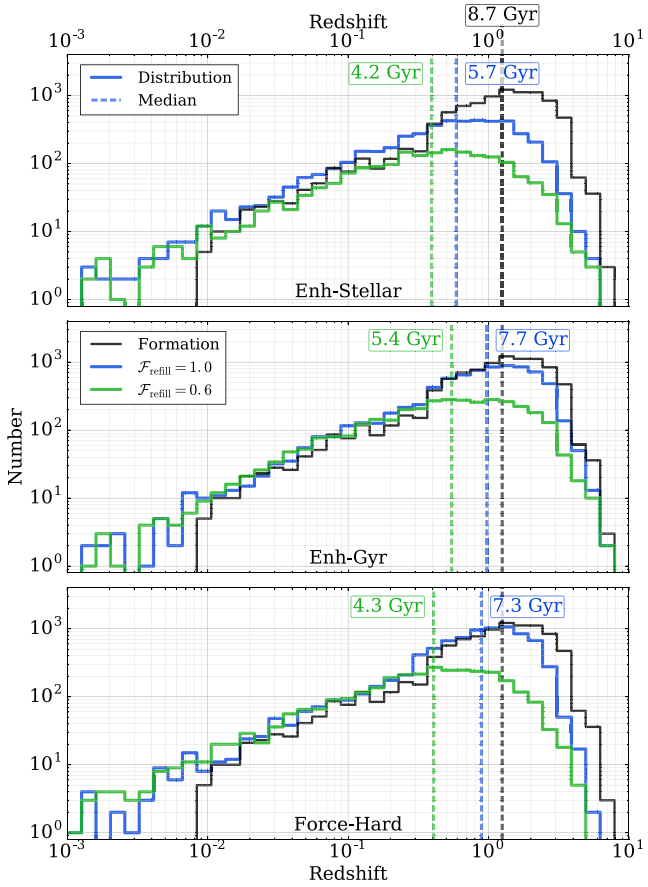

Figure 14. Distribution of MBH binary formation (black) and coalescence (coloured) redshifts for different DF models (panels) and two LC refilling parameters: always full ($\mathcal{F}_{\rm refill} = 1.0$; blue) and our fiducial, moderately refilled ($\mathcal{F}_{\rm refill} = 0.6$; green) value. The median redshift for each distribution is also plotted (dashed), with the corresponding look-back time indicated. The minimum delay time between medians of formation and coalescence is 1 Gyr, but up to 4.5 Gyr for our fiducial LC state and DF model ('Enh-Stellar').

If the LC is only moderately refilled, the median redshifts are much lower: between $z \sim 0.4$ and 0.6.

4.2 The gravitational-wave background

In Section 1 we have outlined the theoretical background for the existence of a stochastic GWB and introduced the formalism for calculating pure power-law spectra. Fig. 15 shows the purely power-law spectrum derived from Illustris MBH binaries, assuming that all systems (passing our selection cuts outlined in Section 2) reach the GW-dominated regime and evolve purely due to GW emission. Other representative power-law predictions (see Section 1) and recent PTA upper limits are included for comparison. The Illustris prediction is completely consistent with the existing literature and about 30 per cent below the most recent PTA upper limits. These consistencies validate the Illustris MBHB population, and the prescriptions for the growth and evolution of individual MBHs.

Almost all of the details of binary evolution are obscured in purely power-law predictions (i.e. equation 2). In particular, they imply that all MBHBs instantly reach the separations corresponding to the frequencies of interest and evolve purely due to GW emission. In reality, we have shown that the delay time distribution can be

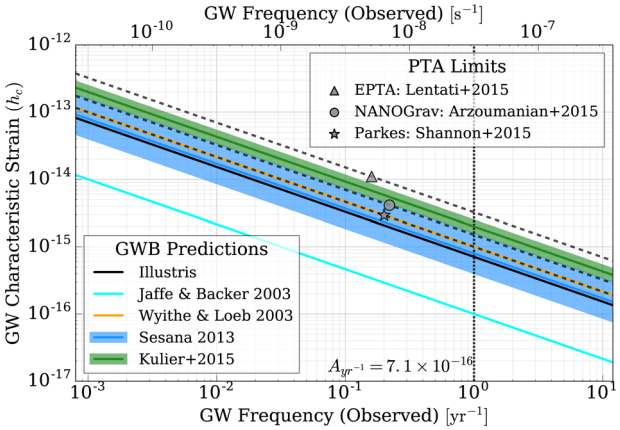

Figure 15. Stochastic GWB spectrum produced by Illustris MBH binaries, assuming purely power-law evolution with all systems efficiently reaching the GW regime. Power-law predictions from the literature (described in Section 1) are presented for comparison, along with the most recent PTA upper limits. The power-law spectrum resulting from the Illustris simulations is very consistent with previous results, and about 30 per cent below the most stringent observational upper limits.

significant at fractions of a Hubble time. This has the important consequence that not all MBHBs coalesce (or even reach the PTA band) before redshift zero. At the same time, the environmental effects (e.g. LC scattering) that are *required* to bring MBH binaries to the relevant orbital frequencies also decrease the time they emit in each band, attenuating the GW signal.

4.2.1 *Full GWB calculation formalism*

The GWB can be calculated more explicitly by decomposing the expression for GW energy radiated per logarithmic frequency interval,

$$\frac{{\rm d}\varepsilon_{\rm GW}}{{\rm d}\ln f_r} = \frac{{\rm d}\varepsilon_{\rm GW}}{{\rm d}t_r}\frac{{\rm d}t_r}{{\rm d}\ln f_r}, \tag{14}$$

where the right-hand-side terms are the GW power radiated and the time spent in each frequency band. The latter term can be further rewritten using Kepler's law as

$$\frac{{\rm d}t_r}{{\rm d}\ln f_r} = f_r\left(\frac{{\rm d}f_r}{{\rm d}t_r}\right)^{-1} = \frac{3}{2}\frac{a}{{\rm d}a/{\rm d}t_r} \tag{15}$$

where 'a' is the semi-major axis of the binary. In this expression, we can identify the binary 'hardening time,'[17] $\tau_{\rm h} \equiv a/(da/dt_r)$. For reference, the binary separations corresponding to each GW frequency are shown in Fig. A1. While the GW power radiated is determined solely by the binary configuration (chirp mass and orbital frequency), the hardening time is determined by both GW emission and the sum of all environmental hardening effects. For more generalized binary evolution we can write

$$\frac{{\rm d}\varepsilon_{\rm GW}}{{\rm d}\ln f_r} = \left.\frac{{\rm d}\varepsilon_{\rm GW}}{{\rm d}\ln f_r}\right|_{\rm GW}\frac{\tau_{\rm h}}{\tau_{\rm gw}}. \tag{16}$$

[17] Sometimes called the 'residence time' in the context of GW spectra.

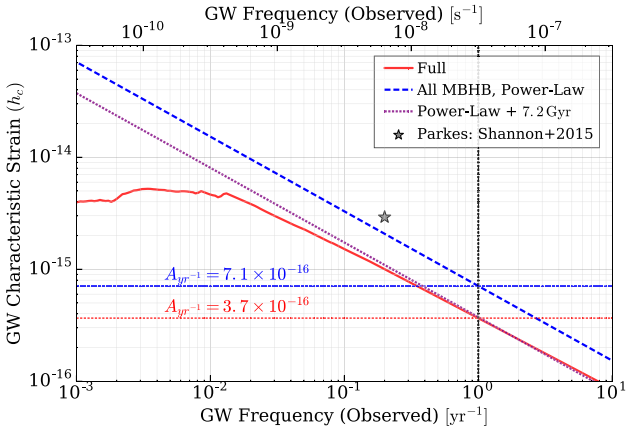

Figure 16. Stochastic GWB calculated from Illustris MBH binaries. The 'Full' calculation, shown in red, includes environmental effects from DF ('Enh-Stellar'), stellar scattering ($\mathcal{F}_{\rm refill} = 0.6$), and a viscous circumbinary disc. Purely power-law models are also shown, for all Illustris MBHBs (blue, dashed) and only the MBHBs that coalesce by redshift zero after being delayed for 7.2 Gyr (purple, dotted). The GWB strain amplitudes at the standard frequency of 1 yr^{-1} are given, showing that a complete model of MBHB evolution leads to an ∼50 per cent decrease of the signal. The most stringent observational upper limits are also shown.

This can be used to reformulate the GWB spectrum calculation[18] (i.e. equation 1) as

$$h_c^2(f) = \frac{4\pi}{3c^2}(2\pi f)^{-4/3} \int \frac{(G\mathcal{M})^{5/3}}{(1+z)^{1/3}} \frac{\tau_{\rm h}}{\tau_{\rm gw}} \frac{{\rm d}^3 n}{{\rm d}z\, d\mathcal{M}\, d\mu} dz\, d\mathcal{M}\, d\mu, \tag{17}$$

or for discrete sources,

$$h_c^2(f) = \frac{4\pi}{3c^2}(2\pi f)^{-4/3} \sum_i \frac{(G\mathcal{M}_i)^{5/3}}{V_{\rm c}\,(1+z_i)^{1/3}} \frac{\tau_{{\rm r},i}}{\tau_{{\rm gw},i}}. \tag{18}$$

Additional hardening mechanisms will decrease the hardening timescale, i.e. $\tau_{\rm h}/\tau_{\rm gw} \leq 1$, decreasing the GWB. The purely power-law expression in equation (1) (and the Illustris spectrum in Fig. 15) thus represents an upper limit to the GWB amplitude. While non-GW mechanisms are required to bring MBH binaries close enough to effectively emit GWs, they also attenuate the amplitude of the GWB.

4.2.2 *Fiducial model predictions*

The stochastic GWB resulting from our fiducial model is presented in Fig. 16. The 'Full' calculation (red, solid) uses equation (18), including the effects of DF, LC scattering, and VD in addition to GW emission. This is compared to a purely power-law model (blue, dashed), calculated with equation (1) and assuming that all Illustris MBHBs reach the PTA-band rapidly, and evolve solely due to GW emission. The amplitudes at 1 yr^{-1} are indicated, showing that the full hardening calculation with an amplitude of $A_{\rm yr^{-1}} \approx 3.7 \times 10^{-16}$ amounts to an almost 50 per cent decrease from the naive, power-law estimate of $A_{\rm yr^{-1}} \approx 7.1 \times 10^{-16}$.

The amplitude of the full GWB calculation can be matched at 1 yr^{-1} using the power-law model by introducing a uniform delay time of ∼7.2 Gyr – such that the systems that formed within a look-back time of 7.2 Gyr do not coalesce or reach the relevant frequency ranges. This is shown in Fig. 16 (purple, dotted) as a heuristic comparison. At frequencies of the PTA band (∼0.1 yr^{-1}) and higher, our full calculation very nearly matches the $A_{\rm yr^{-1}} \propto f^{-2/3}$ power law. A significant flattening of the spectrum is apparent at and below a few 10^{-2} yr^{-1}, where environmental effects (e.g. LC-scattering) significantly increase the rate at which MBHBs move through a given frequency band, decreasing $\tau_{\rm h}$ and thus attenuating the amplitude of the GWB. The particular location and strength of the spectral flattening (or turnover) depend on the details of the DF and especially LC models.

[18] For a more complete derivation, see Kocsis & Sesana (2011).

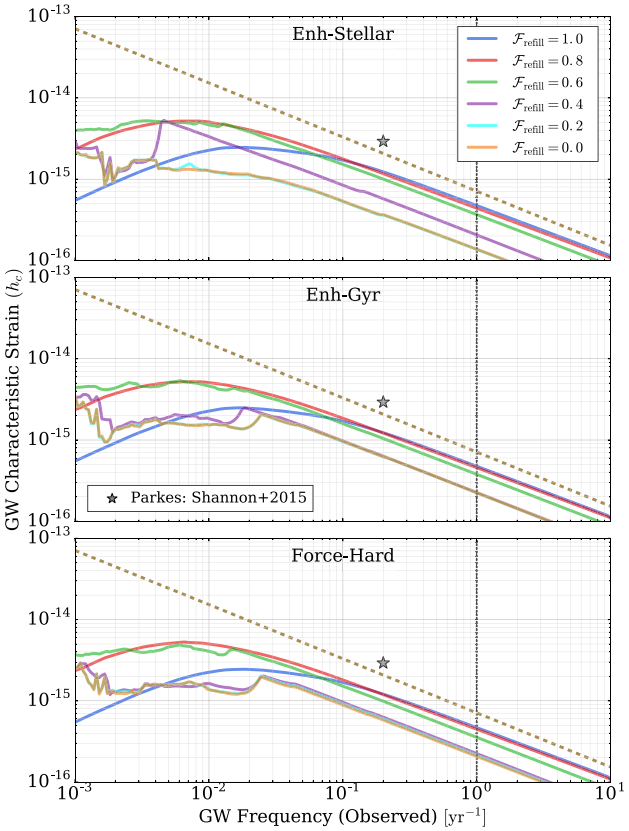

Figure 17. Comparison of GWB spectrum with variations in the LC refilling parameter and DF model. Each panel shows a different DF model, and each line colour a different $\mathcal{F}_{\rm refill}$. Solid lines show the full GWB calculation while the dashed lines show the power-law model using all Illustris MBHBs. Variations in $\mathcal{F}_{\rm refill}$ have much stronger effects on the spectrum than the DF model, changing the location and strength of the spectral break at lower frequencies. There tends to be a substantial jump in GWB amplitude between $\mathcal{F}_{\rm refill} = 0.4$ and 0.6, with more gradual variations on either side. The full range of amplitudes at 1 yr^{-1} and 10^{-1} yr^{-1} is $A_{\rm yr^{-1}} = 0.14$–0.47×10^{-15} and $A_{0.1{\rm yr}^{-1}} = 5.4$–$17 \times 10^{-15}$, respectively.

4.2.3 *GWB variations with dynamical friction and loss-cone model parameters*

Fig. 17 compares the GWB spectrum for different DF prescriptions (panels) and LC refilling fractions (line colours). The naive, power-law model is shown as the dashed line for comparison, along with the most stringent PTA upper limit. Effects from variations in the DF prescription are strongly subdominant to changes in the LC state. The spectral shape is determined almost entirely, and at times sensitively, to $\mathcal{F}_{\rm refill}$. For $\mathcal{F}_{\rm refill} < 0.8$, the spectrum flattens at low frequencies, whereas for higher values it becomes a turnover. Even

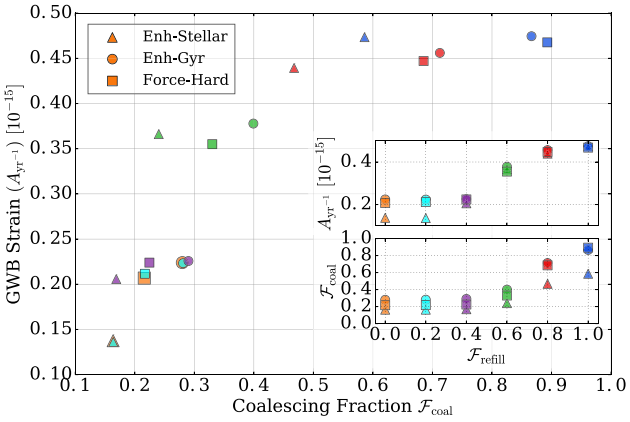

Figure 18. Dependence of GWB strain amplitude and binary coalescing fraction on LC refilling parameter and DF model. The GW strain is measures at the canonical $f = 1\,\mathrm{yr}^{-1}$, and the coalescing fraction is defined using the population of high mass-ratio $\mu > 0.1$ systems. Each symbol represents a different DF model, and each colour a different LC refilling parameter. $A_{\rm yr^{-1}}$ tends to increase monotonically with $\mathcal{F}_{\rm coal}$, but plateaus above $\mathcal{F}_{\rm coal} \gtrsim 0.5$. The insets show how each $A_{\rm yr^{-1}}$ and $\mathcal{F}_{\rm coal}$ change with $\mathcal{F}_{\rm refill}$. The strongest changes in $\mathcal{F}_{\rm refill}$ and $A_{\rm yr^{-1}}$ occur at slightly different values of the refilling fraction. This is because for an increase in $\mathcal{F}_{\rm refill}$, the additional MBHBs that are then able to coalesce tend to be the most massive of those which were previously persisting. At $\mathcal{F}_{\rm refill} \approx 0.5$ there is a significant change in $A_{\rm yr^{-1}}$, due to more massive and stronger GW emitting MBHB merger at that point. At $\mathcal{F}_{\rm refill} \approx 0.7$, $\mathcal{F}_{\rm coal}$ changes significantly due to less massive MBHBs then being able to merger, and them constituting a larger portion of the binary population.

then, the location of the peak amplitude of the spectrum changes by more than a factor of 2 between $\mathcal{F}_{\rm refill} = 0.8$ and $\mathcal{F}_{\rm refill} = 1.0$.

The cutoff seen in the full LC case ($\mathcal{F}_{\rm refill} = 1.0$) is very similar to that found by Sesana (2013a) (with ours approximately five times lower amplitude), who show that in the scattering-dominated regime the GWB turns into a $h_{\rm c} \propto f$ spectrum. McWilliams et al. (2014) also find a spectral cutoff, but at an order of magnitude higher frequency and amplitude. Unlike the results of Ravi et al. (2014), the cutoffs in our predicted GWB spectra are always at lower frequencies than will be reached by PTA in the next decade or so, likely because we assume zero eccentricity in the binary evolution. In the near future we hope to present results expanded to include eccentric evolution, in addition to exploring 'deterministic' or 'continuous' GW sources – i.e. sources individually resolvable by future PTA observations.

For each DF case in Fig. 17, the GWB spectrum is almost identical between $\mathcal{F}_{\rm refill} = 0.0,\ 0.2,$ and 0.4, with very little change in the coalescing fractions. This is consistent with changes in the distribution of lifetimes from varying DF and LC parameters. Looking at $f = 1\,\mathrm{yr}^{-1}$, there is a sudden jump in amplitude with $\mathcal{F}_{\rm refill} = 0.6$, and a modest increase in the coalescing fraction. Between $\mathcal{F}_{\rm refill} = 0.6$ and $\mathcal{F}_{\rm refill} = 0.8$, on the other hand, there tends to be a more modest increase in GWB amplitude, but a roughly factor of 2 increase in $\mathcal{F}_{\rm coal}$. This contrast arises from the changing population of MBHBs that are brought to coalescence from each marginal change in refilling fraction. For an increase in $\mathcal{F}_{\rm refill}$, the additional MBHBs that are then able to coalesce tend to be the most massive of those which were previously persisting. Those, more massive systems, then have a larger effect on the GWB.

Fig. 18 shows the strain at $1\,\mathrm{yr}^{-1}$ ($A_{\rm yr^{-1}}$) versus coalescing fraction for the same set of DF and LC models. The colours again show different $\mathcal{F}_{\rm refill}$, and now symbols are used for different DF prescriptions. The GWB amplitude is strongly correlated with coalescing fraction, but plateaus once roughly 50 per cent of high mass-ratio MBHBs are coalescing. Different DF parameters have little effect on $A_{\rm yr^{-1}}$ but more noticeably affect $\mathcal{F}_{\rm coal}$, in both cases this is especially true at higher $\mathcal{F}_{\rm refill}$. The inset panels show, independently, how $A_{\rm yr^{-1}}$ and $\mathcal{F}_{\rm coal}$ scale with $\mathcal{F}_{\rm refill}$ and DF model, reinforcing the previous points. In general, as $\mathcal{F}_{\rm refill}$ increases, lower total-mass MBHB systems are able to reach the PTA band, contribute to the GWB, and coalesce effectively. At $\mathcal{F}_{\rm refill} \approx 0.5$, the large increase in $A_{\rm yr^{-1}}$ is driven by massive MBH coming to coalescence. At $\mathcal{F}_{\rm refill} \approx 0.7$, on the other hand, a large number of lower-mass MBHs are driven together, significantly increasing $F_{\rm coal}$, but only marginally increasing $A_{\rm yr^{-1}}$.

At the higher frequencies just discussed, the GWB strain increases monotonically with $\mathcal{F}_{\rm refill}$ and coalescing fraction. This is intuitive as increasing effectiveness of the LC means more MBHBs are able to reach the GW regime and then coalesce. Fig. 17 shows that this trend is not the case at lower frequencies (i.e. $f \lesssim 10^{-1}\,\mathrm{yr}^{-1}$) – where the highest $\mathcal{F}_{\rm refill}$ show a decrease in the GWB amplitude. This can be seen more clearly in Fig. A5, which shows the GWB amplitude at $f = 10^{-2}\,\mathrm{yr}^{-1}$ versus coalescing fraction. The trend is generally the same – strain increasing with $\mathcal{F}_{\rm refill}$ – until $\mathcal{F}_{\rm refill} = 1.0$ at which point the GWB amplitude drops significantly. At these low frequencies, LC stellar scattering is effective enough to significantly attenuate the GWB amplitude. This reflects a fundamental tradeoff in the realization of environmental effects: on one side bringing more MBHBs into given frequency bands, at the same time as driving their evolution rapidly through it, and attenuating the GW signal.

4.2.4 *Effects of circumbinary viscous drag on the GWB*

The effects of VD from a circumbinary disc are more subtle than those of DF and LC. Fig. 19 compares the GWB from our fiducial model (black, dashed) with a variety of VD parameter modifications (coloured lines) and to a simulation with VD turned off (grey, dotted). All of these models use the 'Enh-Stellar' DF, but because there is virtually no overlap in between the VD and DF regimes, the results are very similar. The same VD parameters are explored as in the hardening rates shown in Fig. 9: modifying the SG instability radius ($\lambda_{\rm sg}$), the maximum SG radius ($\mathcal{R}_{\rm SG,Max}$), the maximum accretion rate ($f_{\rm Edd}$), and the alpha viscosity parameter (α). The upper panel shows a moderately refilled LC ($\mathcal{F}_{\rm refill} = 0.6$), while in the lower panel the LC is always full ($\mathcal{F}_{\rm refill} = 1.0$). The inset panels show the ratio of GWB strain from each model to that of a 'VD: Off' (i.e. no disc) model.

The overall shape of the GWB spectrum and the location of the spectral turnover is again determined almost entirely by the LC. The circumbinary disc does affect an additional 10–40 per cent amplitude modulation, tending to increase the amplitude at low frequencies ($\lesssim 10^{-2}\,\mathrm{yr}^{-1}$) and decrease it at higher frequencies ($\gtrsim 10^{-1}\,\mathrm{yr}^{-1}$). This reflects the same tradeoff between bringing more MBHBs into each frequency band and driving them more rapidly through them. In the moderately (completely) refilled LC case, our fiducial VD model amounts to an ∼20 per cent (∼30 per cent) decrease in $A_{\rm yr^{-1}} = h_{\rm c}(f = 1\,\mathrm{yr}^{-1})$ and similarly at $h_{\rm c}(f = 10^{-1}\,\mathrm{yr}^{-1})$.

At frequencies near the PTA band, the relationship between $\mathcal{F}_{\rm coal}$ and the GWB amplitude can be non-monotonic for VD variations, like with variations to the LC at low frequencies. For example, a comparison of Figs 9 and 19 shows that the $\alpha = 0.1$ (green) model has the lowest fraction of high mass-ratio coalescences

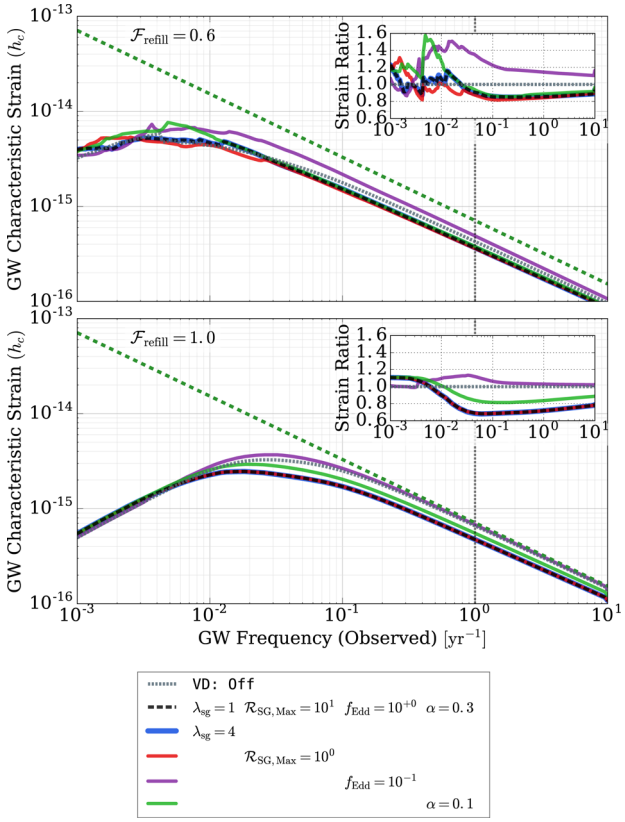

Figure 19. Gravitational-wave background from varying VD parameters. Simulations with a variety of VD models are compared, with our fiducial model in dashed black, a model with no-VD in dotted grey, and each colour of line showing changes to a different parameter. For comparison, the power-law model using all Illustris MBHBs is also shown. The parameters modified are the SG instability radius ($\lambda_{\rm sg}$), the maximum SG radius ($\mathcal{R}_{\rm SG,Max}$), the maximum accretion rate ($f_{\rm Edd}$), and the alpha viscosity parameter (α). The hardening rates for each of these models are shown in Fig. 10. The upper and lower panels show simulations for different LC refilling fractions. The inset panels show the ratio of GWB amplitude from each model to the 'VD: Off' case, as a function of GW frequency. Different VD parameters change $A_{\rm yr^{-1}}$ by 10–40 per cent, and the spectral slope at 1 yr^{-1} by up to ∼10 per cent.

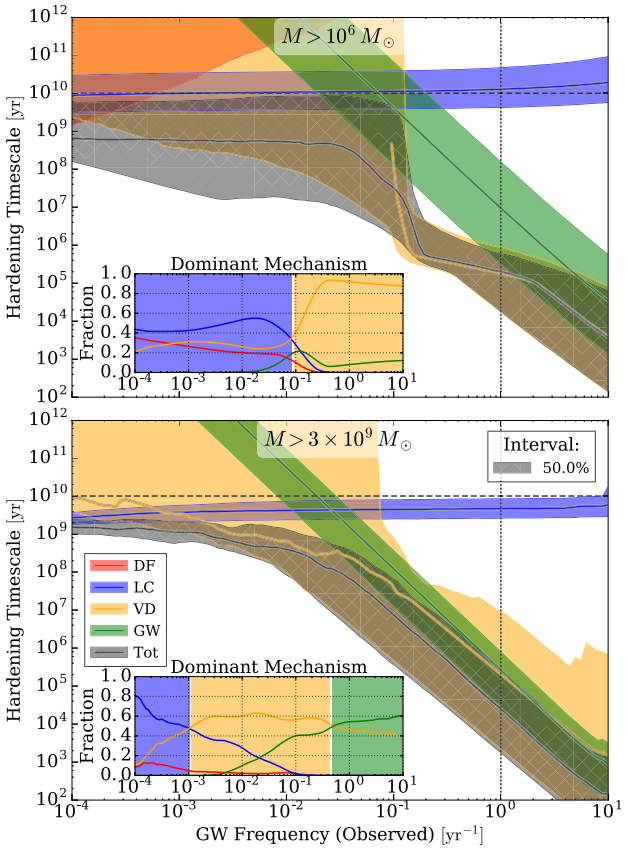

Figure 20. Binary hardening time-scales versus GW frequency, by mechanism, for our fiducial model ($\mathcal{F}_{\rm refill} = 0.6$, DF: 'Enh-Stellar'). Lines and bands show the median and 50 per cent intervals for individual mechanisms (colours), along with the total hardening rate (grey, hatched). The inset panels show the fraction of binaries dominated by each mechanism, again versus frequency. The upper panel shows all MBHBs in our sample, while the lower panel includes only systems with total mass above 3×10^9 M$_\odot$, roughly where the bulk of the GWB amplitude comes from. Only for the high-mass systems do the majority become dominated by GW emission at high frequencies, with VD still contributing substantially to the overall hardening time-scales.

(with $\mathcal{F}_{\rm coal} = 0.22$, versus $\mathcal{F}_{\rm coal} = 0.24$ for the fiducial model, and $\mathcal{F}_{\rm coal} = 0.31$ for the $f_{\rm Edd} = 0.1$ case) but an intermediate $A_{\rm yr^{-1}}$.

One striking feature of the GWB strain ratios is the clear variations in spectral index, even at high frequencies. This is especially true for the always full LC, where the slope of the GWB can deviate by almost 10 per cent from the canonical $-2/3$ power law. The disc-less model (grey, dotted) deviates by about 4 per cent (3 per cent) for $\mathcal{F}_{\rm refill} = 0.6$ ($\mathcal{F}_{\rm refill} = 1.0$) at $f = 1\,{\rm yr}^{-1}$, due to a combination of residual LC scattering effects and some binaries stopping emitting after coalesce at varying critical frequencies. In our fiducial model (black, dashed), the deviations are more significant at 6 per cent (8 per cent). As different parameters make VD hardening more important at this frequency, the GWB amplitude decreases, and the spectral index tends to flatten. Our fiducial VD model tends to have among the strongest spectral deviations. Towards lower frequencies, where PTA are heading, the turnover in the GWB spectrum becomes more significant, especially if the LC is effectively refilled. At $f = 10^{-1}\,{\rm yr}^{-1}$, for example, our fiducial model ('Enh-Stellar', $\mathcal{F}_{\rm refill} = 0.6$) gives a spectral index of about -0.6, while for $\mathcal{F}_{\rm refill} = 1.0$ it becomes slightly flatter than -0.4. A summary of GWB amplitudes and spectral indices is presented in Table B1, for a variety of configurations.

For a given binary system, GW radiation will *always* dominate at some sufficiently small separation (high frequency) where the circumbinary disc dynamically decouples from the hardening MBHB. This does *not necessarily* mean, however, that after considering a full ensemble of MBHB systems, with a variety of masses, that there is any frequency band with a spectral slope identical to the purely GW-driven case ($h_{\rm c} \propto f^{-2/3}$). Hardening rates as a function of GW frequency are shown in Fig. 20. The upper panel includes all MBHBs in our sample[19], for which we see that VD remains dominant well above the PTA frequency band. The high total-mass systems ($M > 3 \times 10^9$ M$_\odot$) – which contribute the bulk of the GWB signal – are shown in the lower panel. These binaries tend to be driven in roughly equal amounts by VD and GW hardening at the frequencies where PTA detections should be forthcoming.

Fig. 20 (and Fig. 11) shows that the typical hardening rates for VD are very similar to that of GW radiation. Indeed, as discussed

[19] Recall that we select only MBH with masses $M > 10^6$ M$_\odot$.

in Section 3.3, the innermost disc region has hardening times $\tau_{v,1} \propto r^{7/2}$, while that of purely GW emission is $\tau_{gw} \propto r^4$. Hardening rates for farther-out disc regions tend to deviate more strongly from that of purely GW evolution, which could become more important for lower density discs.

The MBH accretion rates, which set the density of the circumbinary discs in our models, are perhaps one of the more uncertain aspects of the Illustris simulations, given that the accretion disc scale is well below the resolution limit and must therefore rely on a sub-grid prescription. Additionally, out of all possible configurations, the fiducial disc parameters we adopt tend to produce fairly strong effects on the GWB. If, for example, a β-disc model is more accurate, or the α-viscosity should be lower, the effects in the PTA band will be more moderate (see e.g. Kocsis & Sesana 2011). None the less, we consistently see GWB spectral indices between -0.6 and -0.65 at 1 yr^{-1}, for a wide variety of model parameters. While these $\lesssim 10$ per cent deviations may be entirely unobservable in PTA observations (especially after taking stochastic variations into account; e.g. Sesana et al. 2008), it may need to be considered when using priors or match filtering for detecting a GWB. More stringent observational constraints on specifically post galaxy-merger AGN activity could be used to better calibrate the VD model.

4.3 The populations of MBHB

For the first time, we have used cosmological, hydrodynamic models which self-consistently evolve DM, gas, stars, and MBH, to more precisely probe the connection between MBHB mergers and their environments. Previous calculations (see Section 1) of the GWB using SAM prescribe MBH on to their galaxies based on scaling relations. The MBH population in Illustris, on the other hand, co-evolves with, and shapes, its environment. These data are then much better suited to analyse the details of MBHB and GW source populations, and their hosts.

Fig. 21 shows the distribution of properties for sources contributing to the GWB, from top to bottom: total mass (M), mass ratio (μ), and redshift (z). In the left column, these properties are weighted by squared strain[20] for each source, and the resulting one-, two-, and three-sigma contours are shown as a function of GW frequency. The right column shows the cumulative distribution over the same source properties, weighted by $A^2_{\rm yr^{-1}}$ (solid), compared to the unweighted distribution of all sources contributing at $f = 1$ yr^{-1}. Strain-weighted sources tend to be at higher mass ratios and much higher masses. While the fraction of all binaries rises fairly smoothly with total masses between 10^7 and 10^9 M$_\odot$ (dashed, black line; top-right panel), 90 per cent of the GWB is contributed (solid, black line) by binaries with total mass $\gtrsim 10^9$ M$_\odot$ – simply showing the strong dependence of the GW strain on the total system mass.

The core contribution over all three parameters tends to remain fairly constant over GW frequency, with median values around $M \approx 4 \times 10^9$ M$_\odot$, $\mu \approx 0.3$, and $z \approx 0.3$. The tails of the distribution drop to noticeably lower values when moving to higher frequencies. This is especially pronounced in the redshift distribution, where at frequencies of a few times 10^{-3} yr^{-1} virtually all GWB-weighted sources come from $z > 10^{-2}$, while at $f = 1$ yr^{-1}, almost 10 per cent are below that redshift. While $\sim$20 per cent of binaries that reach $f = 1$ yr^{-1} come from redshift above $z = 1$, they only contribute $\sim$0.5 per cent of the GWB amplitude. Lower redshift and higher mass-ratio systems do contribute somewhat disproportionately to the GWB amplitude, but their distributions are altogether fairly consistent with the overall population. The presence of a non-negligible fraction of low-redshift sources motivates the need to explore populations of MBHBs in the local Universe that could be resolvable as individual 'stochastic' sources or contribute to angular anisotropies in the GW sky. An analysis of our results in this context is currently underway, and the results will be presented in a future study.

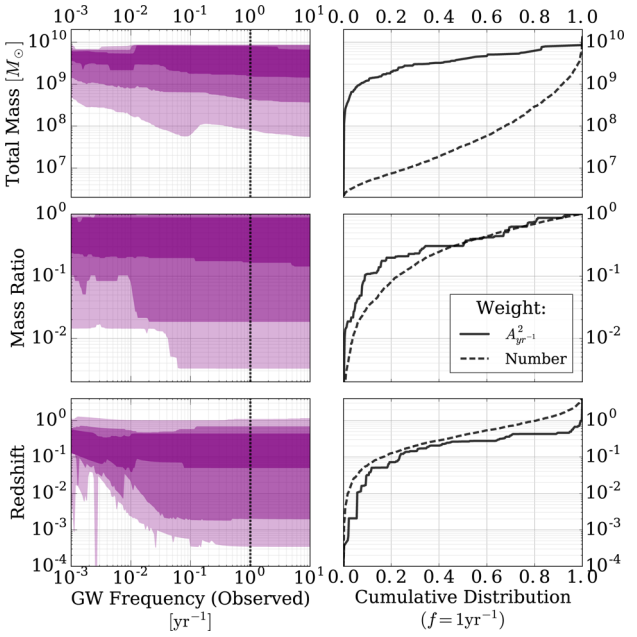

Figure 21. Population of MBHB contributing to the GWB. The left column shows, from top to bottom, the MBHB total mass, mass ratio, and redshift, weighted by each system's contribution to the GWB amplitude. Contours represent one-, two-, and three-sigma intervals. The right column shows cumulative distributions, at a frequency of 1 yr^{-1}, for the same parameters. The solid line weights by contribution to the GWB amplitude ($A^2_{\rm yr^{-1}}$) and the dashed line is the distribution of the number of sources *contributing at 1 yr*$^{-1}$. The median values by GWB contribution are roughly constant over GW frequency, at $M \approx 4 \times 10^9$ M$_\odot$, $\mu \approx 0.3$, and $z \approx 0.3$ for this, our fiducial model. The overall distribution of sources moves noticeably to include lower masses, mass ratios, and redshifts at higher GW frequencies. The contribution from redshifts above $z \approx 0.4$ drops sharply, with $\lesssim 1$ per cent of the GWB signal coming from $z > 1.0$, while still $\sim$20 per cent of all binaries emit there.

As we move into the forthcoming era of PTA *detections*, it will be increasingly important to use self-consistent hydrodynamic models to better understand the coupling of the MBH populations to their host galaxies and merger environments. The Illustris host–galaxy properties of our MBHBs, at the time of binary formation, are presented in Fig. 22. We show stellar radius, stellar mass, and 'subhalo mass', and each of these properties[21] is strongly biased towards higher values when weighting by GW strain. In particular, the median, strain-weighted subhalo, and stellar masses are each more than an order of magnitude larger than the median of the host-galaxy

[20] As seen in equation (18), binaries contribute to the strain spectrum in quadrature.

[21] The stellar radius is measured as the stellar half-mass radius ($R_{\star,1/2}$); the stellar mass is the mass of star particles within $R = 2\,R_{\star,1/2}$; and the subhalo mass is the combined mass of all particles and cells associated with the host galaxy. While these simulation measurements are, of course, non-trivial to relate to their observational counterparts, they are useful for relative comparisons.

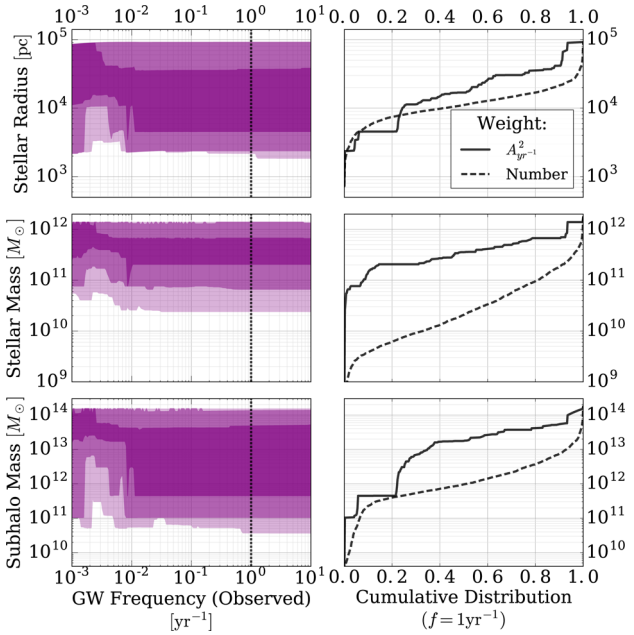

Figure 22. The properties of host galaxies for the population of MBHBs contributing to the GWB. From top to bottom, rows show the stellar (half-mass) radius, stellar mass (within twice the stellar half-mass radius), and subhalo mass (mass of all particles associated with the galaxy). The left column shows these parameters weighted by their resident MBHB's contribution to the GWB as a function of frequency. The right column shows the cumulative distribution at $f = 1\,\mathrm{yr}^{-1}$, both for contribution to GWB amplitude (solid) and by overall number (dashed). The GWB comes from MBHBs predominantly in galaxies that are oversized and significantly overmassive – especially in stars.

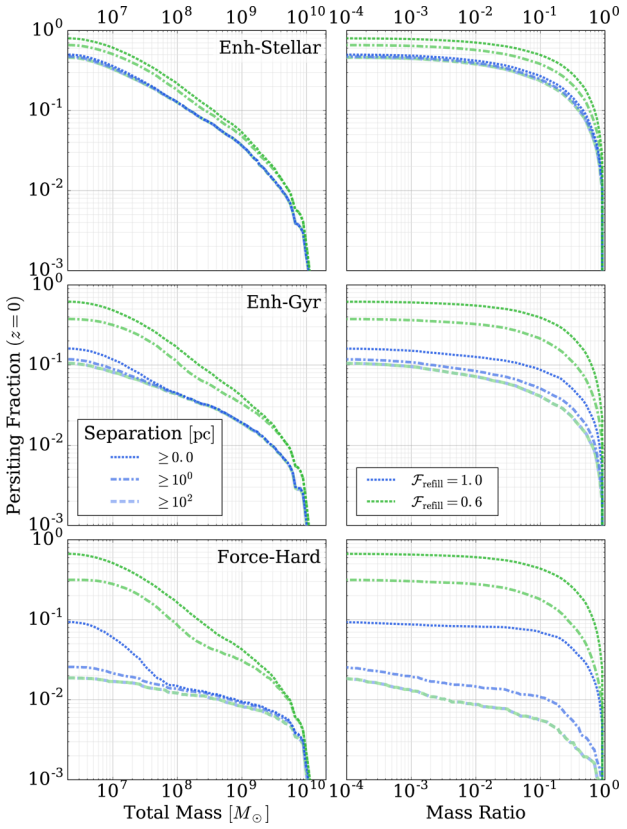

Figure 23. Fraction of binaries that persist at redshift zero as a function of total mass (left column) and mass ratio (right column). Three different DF models are compared, from top to bottom: 'Enh-Stellar', 'Enh-Gyr', and 'Force-Hard'; and in each case, LC refilling parameters of $\mathcal{F}_{\rm refill} = 0.6$ (green) and $\mathcal{F}_{\rm refill} = 1.0$ (blue) are compared. Different line patterns show binaries with different separations: all separations ($r \geq 0.0$, dotted), $r \geq 1$ pc (dot–dashed), and $r \geq 10^2$ pc (dashed). Note that in each panel the $r \geq 10^2$ pc distributions are indistinguishable between LC parameters as the LC only takes effect at and below about 10^2 pc. The fraction of persisting systems is very strongly dependent on both DF and LC models. For 'Enh-Stellar', the most conservative DF case, a large fraction of systems remains in the DF regime ($r \sim 10^2$ pc), before LC scattering can have a significant effect. 'Force-Hard', on the other hand, represents an approximately optimal DF at large scales and shows a corresponding dearth of wide separation systems. Observations of the true fraction of systems at these separations could strongly constrain the efficiency of these hardening mechanisms.

population by number. The bias is exceedingly strong for stellar mass, where ∼90 per cent of the GWB amplitude is contributed by only ∼20 per cent of MBHB host galaxies.

Following the galaxies that host MBHBs to observe their parameters at the times they contribute significantly to the GW spectrum will be important for any future multi-messenger observations using PTA or predicting and deciphering anisotropies in the GWB (Mingarelli et al. 2013; Taylor & Gair 2013; Taylor et al. 2015). Better understanding host galaxy properties as they evolve in time could also be useful in understanding whether 'offset' AGNs (those distinctly separated from the morphological or mass-weighted centre of their galaxies) are due to binarity (i.e. a recent, or perhaps not so recent, merger) or possible due to post-coalesce GW 'kicks' (e.g. Blecha et al. 2016).

Of great observational interest is the presence (e.g. Comerford et al. 2015), or perhaps conspicuous absence (e.g Burke-Spolaor 2011), of dual and binary AGNs. The observational biases towards finding or systematically excluding MBH binaries with electromagnetic observations are extremely complex. None the less, understanding the characteristic residence times of binaries at different physical separations, the types of host galaxies they occupy, and the probability they will be observable (e.g. via the amount of gas available to power AGN activity) is crucial to backing out the underlying population and placing empirical constraints on models of MBHB inspiral. A systematic study of this topic using these data is currently underway (Kelley et al., in preparation).

The fraction of MBHBs that persist (i.e. remain uncoalesced) at redshift zero are shown in Fig. 23 as a function of total mass (left column) and mass ratio (right column). Three different separation criteria are shown in each panel: $r > 0.0$ (i.e. any persisting MBHBs; dark, dotted), $r > 1$ pc (medium, dot–dashed), and $r > 10^2$ pc (light, dashed). Each row corresponds to a different DF model, and line colours vary by LC refilling fractions. In general, persisting fractions fall rapidly with increasing total mass and moderately with increasing mass ratio, until nearly equal-mass systems where the persisting fractions plummet.

The specific persisting fraction depends quite sensitively on both $\mathcal{F}_{\rm refill}$ and DF model. The 'Enh-Stellar' model has by far the most persisting systems, and relatively slight variance with either separation criteria or $\mathcal{F}_{\rm refill}$. For our fiducial model with $\mathcal{F}_{\rm refill} = 0.6$, 80 per cent of all binaries persist, with only weak trends with either total mass or mass ratio: 77 per cent with $M > 10^8\,\mathrm{M}_\odot$, and 74 per cent with $\mu > 0.2$. For systems fulfilling both requirements, the persisting fraction drops more noticeably to 55 per cent.

The $r > 10^2$ pc population in particular is almost solely determined by DF, as the other hardening mechanisms take effect only at smaller scales. At these large separations, persisting fractions for our fiducial model are 46 per cent (all), 45 per cent ($M > 10^8\,\mathrm{M}_\odot$), and

33 per cent ($\mu > 0.2$), but for both high total mass and mass ratio, the widely separated persisting fraction drops dramatically to only 1 per cent. If the DF is more effective, as in the 'Enh-Gyr' model, these fractions decrease significantly to 11 per cent (all), 15 per cent ($M > 10^8\,\mathrm{M}_\odot$), 6 per cent ($\mu > 0.2$), and 1 per cent ($M > 10^8\,\mathrm{M}_\odot$ and $\mu > 0.2$). A summary of persisting fractions at both $r > 10^2$ pc and $r > 1$ pc, for mass combinations and DF & LC models, is presented in Table B1.

4.4 MBH triples

The long characteristic lifetimes we see in our MBHB populations, and the (at times) substantial number of systems that remain at large separations, immediately beg the question of how often a third MBH (i.e. second galaxy merger) could become dynamically relevant. For $\mathcal{F}_{\rm refill} = 1.0$, the median lifetimes of our MBHBs tend to be comparable to the median time between binary formation events, and for $\mathcal{F}_{\rm refill} = 0.6$, they are almost an order of magnitude longer. After selection cuts (see Section 2), 37 per cent of our MBH binaries have subsequent 'merger' events (i.e. a second 'merger' is recorded by Illustris involving an MBH meeting our selection criteria). In our implementation, each of those binary systems is evolved completely independently, even if parts of their evolution are occurring simultaneously.[22] With this caveat, we can still consider, very simplistically, in how many systems the *second* binary overtakes the first as they harden. Out of the binaries with subsequent events, 83 per cent (31 per cent of all binaries) are overtaken, 76 per cent (28 per cent overall) of those before redshift zero, and 42 per cent (16 per cent) with $z > 0.0$ and mass ratio $\mu > 0.2$.

The tendency for subsequent binaries to cross in our simulations likely reflects systems' ability to increase noticeably in total mass over the course of the merger process. We emphasize that this is a very simplistic and preliminary investigation. If, for example, MBH remnants tend to receive significant 'kicks' after merger, the resulting fractions could change significantly. None the less, the apparent commonality of candidate multiples suggests that the role of triples should be investigated more thoroughly.

It is unclear how such triple systems should be treated, even in a simple semi-analytic manner (see, however, Bonetti et al. 2016). The conventional wisdom of triple system dynamics is that the lowest mass object will be ejected, while the more massive pair becomes bound in a binary (e.g. Hills 1975). Such 'exchange' interactions are motivated primarily from stochastic scattering events, like those which may occur between stars in dense stellar environments. In these cases, the system can be viewed as nearly dissipationless, and their initial encounter is effectively stochastic. It is our premise, however, that the environments and dynamics of MBH multiples are heavily dissipational. For example, consider an initial pair of MBHs that encounter at kpc scales, on a hyperbolic orbit. If the system quickly circularizes, and hardens to scales of 1–100 pc, then a third MBH that encounters the system – again at kpc scales – may similarly settle into an outer, roughly circular orbit forming a hierarchical system. In such a situation, secular instead of scattering dynamics, such as the Kozai–Lidov mechanism (Kozai 1962; Lidov 1962) or resonant migration, may be more appropriate than traditional three-body scatterings. In this case, the outer MBH in the triple system may accelerate the hardening of the inner binary, driving it to coalescence (e.g. Blaes, Lee & Socrates 2002). This may be less likely in gas-rich environments that could effectively damp eccentric evolution, but here gas-driven inspiral will likely cause rapid coalescence in any case.

MBH triples forming hierarchically with low to moderate eccentricities may evolve in a resonant fashion. If, on the other hand, environmental effects sufficiently enhance (or preserve initially high) eccentricities of MBHBs, then the resulting highly radial orbits may strongly intersect. In that case, a more stochastic-scattering-like regime may indeed still be appropriate. Numerous studies have suggested that environmental effects can indeed enhance MBHB eccentricity (e.g. Quinlan 1996; Sesana 2010; Ravi et al. 2014). If the interaction between MBH triples (or even higher order multiples) is indeed most similar to scattering, then the simplest prescription of removing the lowest mass BH, with or without some additional hardening of the more massive pair, may still be appropriate (e.g. Hoffman & Loeb 2007). An ejected MBH that may later fall back to the galactic centre, while of great observational interest in and of itself, is likely less important for GW emission per se.

The observation of a triple-AGN system could provide insight into the type of system they form (i.e. hierarchical versus scattering) and their lifetimes. Additionally, MBHs ejected by three-body interactions could be observable as offset AGNs, and possibly confused with binary MBHs, or ones 'recoiling' from previous coalescences (e.g. Blecha et al. 2011). We have assumed that recoiling systems do not significantly affect our populations, effectively assuming that kicks are small – which is expected for spin-aligned MBHBs. This is motivated by studies which have shown that gravitational torques from circumbinary discs, such as those we consider, can be effective at aligning spins on time-scales significantly shorter than a viscous time (e.g. Bogdanović, Reynolds & Miller 2007; Dotti et al. 2010; Miller & Krolik 2013).

The MBH populations from the Illustris simulations are well suited for this problem, as they accurately follow the histories and large-scale environments of MBHB systems and host galaxies. As we are currently working on implementing eccentric evolution into our simulations, we plan to explore multi-MBH systems in more detail. This framework will also allow for the treatment of kicked MBHs resulting from random spin orientations, if for example the spins of a substantial fraction of MBHBs occur in gas-poor environments in which they may not be aligned.

5 CONCLUSIONS AND SUMMARY

For the first time, we have used the results of self-consistent, hydrodynamic cosmological simulations, with a co-evolved population of MBHs to calculate the plausible stochastic GWB soon to be detectable by PTA. We have also presented the first simultaneous, numerical treatment of all classes of MBHB hardening mechanisms, discussing the effects of each: DF, stellar (LC) scattering, gas drag from a viscous circumbinary disc, and GW emission.

The most advanced previous studies have included only individual environmental effects, for example, calculating DF time-scales to determine which systems will contribute to the GWB (McWilliams et al. 2014) or attenuating the GWB spectrum due to LC stellar scattering (Ravi et al. 2014). We explicitly integrate each of almost ten thousand MBH binaries, from galactic scales to coalescence, using self-consistently derived, realistic galaxy environments and MBH accretion rates. We thoroughly explore a broad parameter space for each hardening mechanism to determine the effects on the MBHB merger process, the lifetimes of systems, and the resulting GWB spectrum they produce.

[22] Recall that in Illustris, after the initial 'merger' event, only a single remnant MBH particle remains.

The resulting lifetimes of MBHBs that coalesce by redshift zero are usually gigayears, while that of low total-mass and extreme mass-ratio systems typically extend well above a Hubble time. In our fiducial model, with a modest DF prescription ('Enh-Stellar') and moderately refilled LC ($\mathcal{F}_{\rm refill} = 0.6$), the median lifetime of MBHBs with total masses $M > 10^8\,{\rm M}_\odot$ is 17 Gyr, with 23 per cent coalescing before redshift zero. Massive systems that also have high mass ratios, $\mu > 0.2$, merge much more effectively, with a median lifetime of 6.9 Gyr and 45 per cent coalescing at $z > 0$. Increasing the effectiveness of the LC drastically decreases system merger times. For an always full LC ($\mathcal{F}_{\rm refill} = 1.0$), the lifetime of massive systems decreases to 4.9 and 0.35 Gyr for systems with $M > 10^8\,{\rm M}_\odot$ and all mass ratios, and those with $\mu > 0.2$, respectively. The coalescing fractions in these cases doubles to 54 and 99 per cent, respectively. A summary of lifetimes and coalescing fractions for different models is presented in Table B1.

The growing number of dual-MBH candidates (e.g. Deane et al. 2014; Comerford et al. 2015) presents the opportunity to constrain binary lifetimes and coalescing fractions observationally. For most of our models, only about 1 per cent of MBHBs with total masses $M > 10^8\,{\rm M}_\odot$ and mass $\mu > 0.2$ remain at separations $r > 10^2$ pc at redshift zero. At smaller separations, $r > 1$ pc, the fractions are dependent on model parameters, but in general between 1 and 40 per cent. Tabulated persisting fractions are included in Table B1 for a variety of models and situations. Observational constraints on these fractions can narrow down the relevant parameter space of hardening physics. Accurate predictions for dual-MBH observations must fold in AGN activity fractions and duty cycles, and their correlations with binary merger lifetimes. A comprehensive study of dual-AGN observability predicted by our models, over redshift and different observational parameters, is currently underway.

In addition to measuring the fraction of MBHs in associations (e.g. dual AGNs) as a function of separation, the redshift distribution of dual MBHs can be useful in understanding their evolution. The Illustris simulations, for example, give a median MBHB formation redshift[23] of $z \approx 1.25$. Depending on the parameters of the hardening models, the median coalescing redshift can be anywhere between $z \approx 0.4$ and 1.0, with $z \approx 0.6$ suggested by our fiducial model.

Without electromagnetic observations, GWB detections and upper limits can also be used to inform our understanding of MBH evolution (e.g. Sampson, Cornish & McWilliams 2015). Even if the fraction of systems that coalesce is quite low, the most massive and high mass-ratio systems, which produce the strongest GW, are difficult to *keep* from merging. In a simulation with the weakest hardening rates ('Enh-Stellar'; $\mathcal{F}_{\rm refill} = 0.0$) only ∼12 per cent of all binaries coalesce by redshift zero, but the GWB amplitude at $f = 1\,{\rm yr}^{-1}$ is still 0.2×10^{-15} – only about a factor of 5 below the most recent upper limits.[24]

In our fiducial model, we use a moderate LC refilling rate ($\mathcal{F}_{\rm refill} = 0.6$) that increases the number of MBHBs contributing to the GWB at $1\,{\rm yr}^{-1}$, producing an amplitude of $A_{\rm yr^{-1}} \approx 0.4 \times 10^{-15}$. Increasing the effectiveness of DF and/or LC scattering tends to increase the amplitude further. Our fiducial model also includes fairly strong VD from circumbinary discs, which decreases the time MBHBs spend emitting in each frequency band, and thus attenuating the GWB. This effect tends to be more subtle, producing GWB attenuation of about 15 per cent. In general, for a fairly broad range of parameters, our simulations yield GWB amplitudes between ∼0.3 and 0.6×10^{-15}. A GWB amplitude of $A_{\rm yr^{-1}} \approx 0.4 \times 10^{-15}$ is less than a factor of 3 below current detection limits – a parameter space that will likely be probed by PTA within the next decade.

The most stringent PTA upper limits of $A_{\rm yr^{-1}} \lesssim 10^{-15}$ (Shannon et al. 2015) have already excluded a broad swath of previous predictions. Many of those models assume that binary hardening is very effective, with all MBHBs quickly reaching the PTA band and emitting an unattenuated signal – i.e. evolving purely due to GW emission, without additional environmental hardening effects. Following the same procedure, to calculate an upper limit to the GWB based on our population of MBHBs, we find a GWB amplitude of $A_{\rm yr^{-1}} \approx 0.7 \times 10^{-15}$ – slightly below the PTA limit. The Illustris simulation volume is very large for a hydrodynamic simulation, but it lacks the very rare, most massive MBHs in the Universe ($\gtrsim 10^{10}\,{\rm M}_\odot$) which could slightly increase our predicted GWB amplitude – although likely a correction of the order of ∼10 per cent[25] (Sesana, private communication). None the less, our upper limit suggests that the current lack of PTA detections should not be interpreted as a missing signal.

Our upper-limit value of $A_{\rm yr^{-1}} \approx 0.7 \times 10^{-15}$ falls just within the lower end of some recent studies (e.g. Ravi et al. 2014; Roebber et al. 2016), but is generally lower than much of the previous literature (see e.g. Table 1, and fig. 2 of Shannon et al. 2015). Likely, this is at least partly because the MBH merger rates derived from Illustris are based directly on simulated galaxy–galaxy merger rates. The bulk of existing calculations have either used inferences from (DM-only) halo–halo mergers that may have systematic issues (see e.g. Rodriguez-Gomez et al. 2015), or observations of galaxy-merger rates that have uncertain time-scales. This upper limit is based on optimistic, GW-only evolution. In our fiducial model, the signal is lower by ∼50 per cent due primarily to the moderately refilled LC and mildly due to VD attenuation.

Variations in the rate at which the stellar LC is refilled has the strongest effect on the shape and amplitude of the GWB spectrum in our simulations, especially at low frequencies. PTA observations are moving towards these frequencies, as the duration of their timing baselines increases. Unlike at higher frequencies where scattering increases the number of MBHBs contributing to the GWB, at $10^{-1}\,{\rm yr}^{-1}$, for example, effective LC refilling leads to attenuation of the GWB from accelerated binary hardening. Here, our spectra tend to lie at amplitudes between 1.5 and 2.5×10^{-15}, with spectral indices between about −0.4 and −0.6 – a significant deviation from the canonical −2/3 power law. At frequencies lower still ($f \lesssim 10^{-2}\,{\rm yr}^{-1}$), the effective LC scattering produces a strong turnover in the GWB spectrum. A summary of GWB amplitudes and spectral indices is presented in Table B1 for both $f = 1$ and $10^{-1}\,{\rm yr}^{-1}$, and numerous hardening models.

In our fiducial simulation, we find that the median contribution to the GWB comes from binaries at a redshift of $z \approx 0.3$, with total masses $M \approx 10^9\,{\rm M}_\odot$, and mass ratios $\mu \approx 0.3$. The co-evolved

[23] Recall that MBHB 'formation', in this context, corresponds to two MBHs coming within a few kpc of each other.

[24] Previous studies have shown that including eccentric evolution can significantly decrease the GWB amplitude (Ravi et al. 2014), so we caution that the weakest GWB observed in our simulations, which *do not* include eccentric evolution, may not be a robust lower limit. We are currently exploring the effects of eccentricity and altered MBH-host scaling relations on the minimum plausible GWB – to be presented in a future study.

[25] The effects of simulation volume on the predicted GWB amplitude should be studied more carefully to confirm this estimate.

population of MBHs and galaxies in Illustris allows us to also examine typical host–galaxy properties for the first time. Galaxies containing MBHBs contributing strongly to the GWB are noticeably larger and more massive galaxies. The median stellar mass of galaxies, weighted by GWB contribution, is about $3 \times 10^{11}\,M_{\odot}$ – more than an order of magnitude larger than the median stellar mass for all galaxies.

Based on the merger trees and binary lifetimes produced from our simulations, we have also shown that the presence of higher order MBH multiples could be a non-negligible aspect of MBH evolution. The simplest examination suggests that triples could be important in about 30 per cent of MBHBs in our simulations. In future work, we will explore these triple systems in more detail, as well as the effects of non-zero eccentricity and post-merger recoils. We also hope to implement more self-consistent LC refilling and more comprehensive tracking of the changing galactic environment.

The summary is as follows.

(i) *MBH binary lifetimes tend to be multiple Gyr, even for massive systems*. While massive and high mass-ratio systems are likely rare at very large separations, observations of dual MBHs at $r > 1$ pc can be used to constrain the merger physics.

(ii) *The GWB amplitude predicted by our models is* $A_{1\,\mathrm{yr}^{-1}} \approx 0.4 \times 10^{-15}$*, with a range of about 0.3–0.6* $\times\ 10^{-15}$ *for different hardening parameters.* At lower frequencies, we find $A_{0.1\,\mathrm{yr}^{-1}} \approx 1.5\text{–}2.5 \times 10^{-15}$, with spectral indices between -0.4 and -0.6 – a noticeable deviation from the canonical $-2/3$ power law.

(iii) *We find that the lack of PTA detections so far is entirely consistent with our MBH population* and does not require environmental effects. At the same time, our most conservative models yield a GWB amplitude of $A_{\mathrm{yr}^{-1}} = 0.2 \times 10^{-15}$. While incorporating non-zero eccentricities may further suppress our GWB predictions, our simulations suggest that if PTA limits improve by a factor of 3–4 and no detection is made, our understanding of galaxy and MBH evolution may require revision.

(iv) *The median redshift and total mass of MBHB sources contributing to the GWB are* $z \approx 0.3$ *and a few* $10^9\,M_{\odot}$*, while the median coalescence time of all systems tends towards* $z \approx 0.6$. Observations of the redshift distribution and host galaxy properties of dual MBHs can be informative for our understanding of binary evolution.

(v) *Our simulations suggest that up to 30 per cent of binaries could involve the presence of a third MBH.* The role of MBH triples is currently unclear, but should be explored and included in future simulations.

The environments around MBHBs form a complex and interwoven parameter space with additive and often degenerate effects on the GWB. Better constraints on MBH–host correlations, combined with increasingly strict upper limits on the GWB amplitude, will soon tightly constrain the efficiency with which MBHBs are able to coalesce. That efficiency then determines the fraction of galaxies that should have observable dual or binary AGNs, providing an additional test of our most fundamental assumptions of MBH/galaxy growth and co-evolution. We believe that our results, and similar analyses, can be used to leverage GWB observations along with dual and offset AGNs to comprehensively understand the MBH population and their evolution. These exotic binaries involve a plethora of dynamical processes that are still poorly understood, but affect our most fundamental assumptions of black holes in the Universe and thus the evolution of galaxies over cosmological time. Right now, we are entering the era of GW astronomy, and with it a direct view of BH physics and evolution on all scales.

ACKNOWLEDGEMENTS

We are grateful to Julie Comerford, Jenny Greene, Daniel Holz, Maura McLaughlin, and Vikram Ravi for fruitful suggestions and discussions. Advice from Alberto Sesana has been extremely helpful throughout, and especially in beginning this project. We are also thankful to the facilitators, organizers, and attendees of the NANOGrav, spring 2016 meeting at CalTech, especially Justin Ellis, Chiara Mingarelli, and Stephen Taylor. We also thank the anonymous referee for numerous, very constructive comments on the manuscript.

This research made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013), in addition to SciPy (Jones et al. 2001), ipython (Pérez & Granger 2007), and NumPy (Van Der Walt, Colbert & Varoquaux 2011). All figures were generated using matplotlib (Hunter 2007).

REFERENCES

Antonini F., Merritt D., 2012, ApJ, 745, 83
Arzoumanian Z. et al., 2016, ApJ, 821, 1
Astropy Collaboration et al., 2013, A&A, 558, A33
Barnes J. E., Hernquist L., 1992, ARA&A, 30, 705
Begelman M. C., Blandford R. D., Rees M. J., 1980, Nature, 287, 307 (BBR80)
Berczik P., Merritt D., Spurzem R., Bischof H.-P., 2006, ApJ, 642, L21
Binney J., Tremaine S., 1987, Galactic Dynamics. Princeton Univ. Press, Princeton, NJ
Blaes O., Lee M. H., Socrates A., 2002, ApJ, 578, 775
Blandford R., Romani R. W., Narayan R., 1984, JA&A, 5, 369
Blecha L., Cox T. J., Loeb A., Hernquist L., 2011, MNRAS, 412, 2154
Blecha L. et al., 2016, MNRAS, 456, 961
Bogdanović T., Reynolds C. S., Miller M. C., 2007, ApJ, 661, L147
Bonetti M., Haardt F., Sesana A., Barausse E., 2016, MNRAS, 461, 4
Burke-Spolaor S., 2011, MNRAS, 410, 2113
Chandrasekhar S., 1942, Principles of Stellar Dynamics. Univ. Chicago Press, Chicago
Chandrasekhar S., 1943, ApJ, 97, 255
Chen X., Madau P., Sesana A., Liu F. K., 2009, ApJ, 697, L149
Comerford J. M., Gerke B. F., Stern D., Cooper M. C., Weiner B. J., Newman J. A., Madsen K., Barrows R. S., 2012, ApJ, 753, 42
Comerford J. M., Pooley D., Barrows R. S., Greene J. E., Zakamska N. L., Madejski G. M., Cooper M. C., 2015, ApJ, 806, 219
Cuadra J., Armitage P. J., Alexander R. D., Begelman M. C., 2009, MNRAS, 393, 1423
Deane R. P. et al., 2014, Nature, 511, 57
Desvignes G. et al., 2016, MNRAS, 458, 3341
Detweiler S., 1979, ApJ, 234, 1100
Dotti M., Colpi M., Haardt F., Mayer L., 2007, MNRAS, 379, 956
Dotti M., Montuori C., Decarli R., Volonteri M., Colpi M., Haardt F., 2009, MNRAS, 398, L73
Dotti M., Volonteri M., Perego A., Colpi M., Ruszkowski M., Haardt F., 2010, MNRAS, 402, 682
Escala A., Larson R. B., Coppi P. S., Mardones D., 2005, ApJ, 630, 152
Fan X. et al., 2006, AJ, 131, 1203
Foster R. S., Backer D. C., 1990, ApJ, 361, 300
Frank J., Rees M. J., 1976, MNRAS, 176, 633
Genel S. et al., 2014, MNRAS, 445, 175
Gould A., Rix H.-W., 2000, ApJ, 532, L29
Guo Q. et al., 2011, MNRAS, 413, 101
Haiman Z., 2013, in Wiklind T., Mobasher B., Bromm V., eds, Astrophysics and Space Science Library, Vol. 396, The First Galaxies. Springer-Verlag, Berlin, Heidelberg, p. 293

Haiman Z., Kocsis B., Menou K., 2009, ApJ, 700, 1952 (HKM09)
Hills J. G., 1975, AJ, 80, 809
Hobbs G. et al., 2010, Class. Quantum Gravity, 27, 084013
Hoffman L., Loeb A., 2007, MNRAS, 377, 957
Hopkins P. F. et al., 2010, ApJ, 724, 915
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Hut P., 1983, ApJ, 272, L29
Illingworth G., 1977, ApJ, 218, L43
Ivanov P. B., Polnarev A. G., Saha P., 2005, MNRAS, 358, 1361
Jaffe A. H., Backer D. C., 2003, ApJ, 583, 616
Jones E. et al., 2001, SciPy: Open Source Scientific Tools for Python, 2001-. Available at: http://www.scipy.org/
Khan F. M., Just A., Merritt D., 2011, ApJ, 732, 89
Khan F. M., Holley-Bockelmann K., Berczik P., Just A., 2013, ApJ, 773, 100
Khan F. M., Fiacconi D., Mayer L., Berczik P., Just A., 2016, ApJ, 828, 2
Kocsis B., Sesana A., 2011, MNRAS, 411, 1467
Kormendy J., Richstone D., 1995, ARA&A, 33, 581
Koss M., Mushotzky R., Treister E., Veilleux S., Vasudevan R., Trippe M., 2012, ApJ, 746, L22
Kozai Y., 1962, AJ, 67, 591
Kulier A., Ostriker J. P., Natarajan P., Lackner C. N., Cen R., 2015, ApJ, 799, 178
Lacey C., Cole S., 1993, MNRAS, 262, 627
Leach R., 1981, ApJ, 248, 485
Lentati L. et al., 2015, MNRAS, 453, 2576
Lidov M. L., 1962, Planet. Space Sci., 9, 719
Lightman A. P., Shapiro S. L., 1977, ApJ, 211, 244
LIGO 2016a, Phys. Rev. Lett., 116, 061102
LIGO 2016b, Phys. Rev. Lett., 116, 131103
Lotz J. M., Jonsson P., Cox T. J., Croton D., Primack J. R., Somerville R. S., Stewart K., 2011, ApJ, 742, 103
Magorrian J., Tremaine S., 1999, MNRAS, 309, 447
Magorrian J. et al., 1998, AJ, 115, 2285
Manchester R. N. et al., 2013, Publ. Astron. Soc. Aust., 30, e017
McConnell N. J., Ma C.-P., 2013, ApJ, 764, 184
McWilliams S. T., Ostriker J. P., Pretorius F., 2014, ApJ, 789, 156
Merritt D., 2013, Class. Quantum Gravity, 30, 244005
Merritt D., Milosavljević M., 2005, Living Rev. Relativ., 8
Miller M. C., Krolik J. H., 2013, ApJ, 774, 43
Mingarelli C. M. F., Sidery T., Mandel I., Vecchio A., 2013, Phys. Rev. D, 88, 062005
Nelson D. et al., 2015, Astron. Comput., 13, 12
Pérez F., Granger B., 2007, Comput. Sci. Eng., 9, 21
Peters P. C., 1964, Phys. Rev., 136, 1224
Phinney E. S., 2001, preprint (arXiv:e-print)
Press W. H., Schechter P., 1974, ApJ, 187, 425
Quinlan G. D., 1996, New Astron., 1, 35
Quinlan G. D., Hernquist L., 1997, New Astron., 2, 533
Rajagopal M., Romani R. W., 1995, ApJ, 446, 543
Ravi V., Wyithe J. S. B., Shannon R. M., Hobbs G., Manchester R. N., 2014, MNRAS, 442, 56
Rodriguez C., Taylor G. B., Zavala R. T., Peck A. B., Pollack L. K., Romani R. W., 2006, ApJ, 646, 49
Rodriguez-Gomez V. et al., 2015, MNRAS, 449, 49
Roebber E., Holder G., Holz D. E., Warren M., 2016, ApJ, 819, 163
Rosado P. A., Sesana A., Gair J., 2015, MNRAS, 451, 2417
Sampson L., Cornish N. J., McWilliams S. T., 2015, Phys. Rev. D, 91, 084055
Sellwood J. A., Wilkinson A., 1993, Rep. Progress Phys., 56, 173
Sesana A., 2010, ApJ, 719, 851
Sesana A., 2013a, Class. Quantum Gravity, 30, 224014
Sesana A., 2013b, MNRAS, 433, L1
Sesana A., Khan F. M., 2015, MNRAS, 454, L66
Sesana A., Haardt F., Madau P., Volonteri M., 2004, ApJ, 611, 623
Sesana A., Vecchio A., Colacino C. N., 2008, MNRAS, 390, 192
Sesana A., Vecchio A., Volonteri M., 2009, MNRAS, 394, 2255
Sesana A., Shankar F., Bernardi M., Sheth R. K., 2016, MNRAS, 463, L6
Shakura N. I., Sunyaev R. A., 1973, A&A, 24, 337
Shannon R. M. et al., 2015, Science, 349, 1522
Shapiro S. L., Teukolsky S. A., 1986, Black Holes, White Dwarfs and Neutron Stars: The Physics of Compact Objects. Wiley, New York
Sijacki D., Springel V., Di Matteo T., Hernquist L., 2007, MNRAS, 380, 877
Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, MNRAS, 452, 575
Small T. A., Blandford R. D., 1992, MNRAS, 259, 725
Soltan A., 1982, MNRAS, 200, 115
Springel V., 2010, MNRAS, 401, 791
Springel V. et al., 2005, Nature, 435, 629
Taylor S. R., Gair J. R., 2013, Phys. Rev. D, 88, 084001
Taylor S. R. et al., 2015, Phys. Rev. Lett., 115, 041101
Taylor S. R., Vallisneri M., Ellis J. A., Mingarelli C. M. F., Lazio T. J. W., van Haasteren R., 2016, ApJ, 819, L6
The eLISA Consortium et al., 2013, preprint (arXiv:1305.5720)
The NANOGrav Collaboration et al., 2015, ApJ, 813, 65
Torrey P., Vogelsberger M., Genel S., Sijacki D., Springel V., Hernquist L., 2014, MNRAS, 438, 1985
Valtonen M. J. et al., 2008, Nature, 452, 851
Van Der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., preprint (arXiv:1102.1523)
Vasiliev E., Antonini F., Merritt D., 2014, ApJ, 785, 163
Vasiliev E., Antonini F., Merritt D., 2015, ApJ, 810, 49
Verbiest J. P. W. et al., 2016, MNRAS, 458, 2
Vogelsberger M., Genel S., Sijacki D., Torrey P., Springel V., Hernquist L., 2013, MNRAS, 436, 3031
Vogelsberger M. et al., 2014a, MNRAS, 444, 1518
Vogelsberger M. et al., 2014b, Nature, 509, 177
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
Wong K.-W., Irwin J. A., Yukita M., Million E. T., Mathews W. G., Bregman J. N., 2011, ApJ, 736, L23
Wyithe J. S. B., Loeb A., 2003, ApJ, 590, 691
Young J. S., Scoville N. Z., 1991, ARA&A, 29, 581
Yu Q., 2002, MNRAS, 331, 935

APPENDIX A: ADDITIONAL FIGURES OF GRAVITATIONAL-WAVE SCALINGS AND ALTERNATIVE MODEL RESULTS

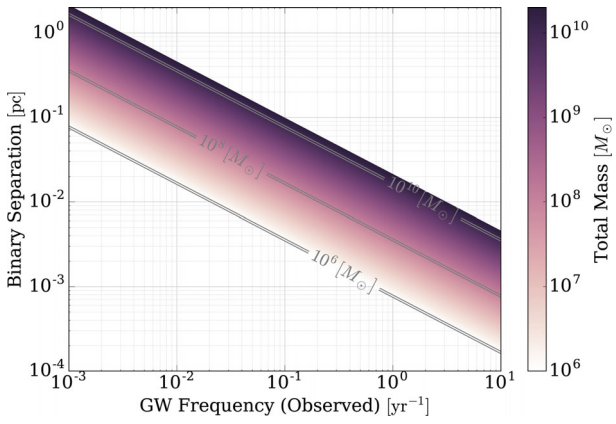

Figure A1. Binary separation versus GW frequency for the total MBHB masses used in our simulations. GW frequency is twice the orbital frequency and is redshifted travelling from the source to observer; the values plotted are at $z = 0$. For the PTA band, roughly 0.1–1 yr^{-1}, binaries contribute anywhere between about 10^{-3} and 10^{-1} pc. Most of the GWB amplitude come from higher mass systems, towards larger separations.

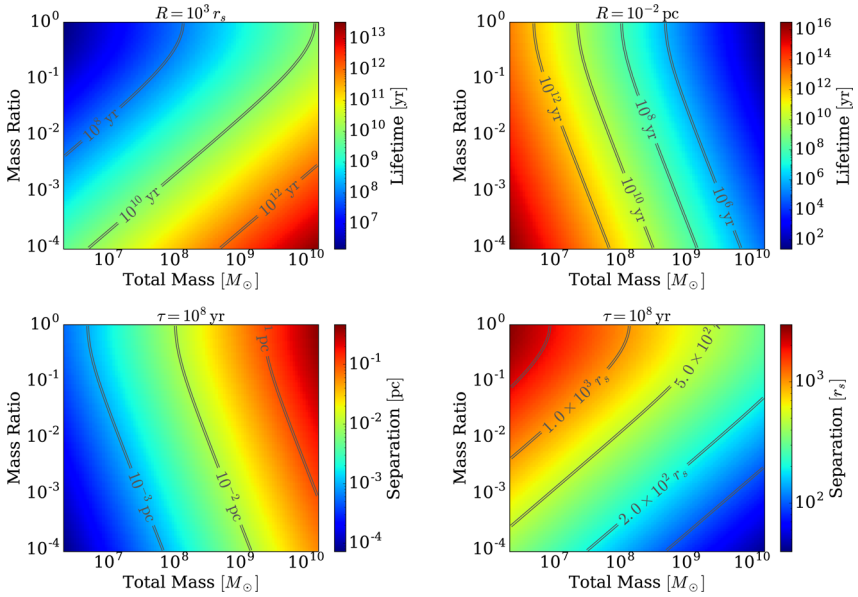

Figure A2. Characteristic lifetimes and separations for coalescence due to purely GW emission over the relevant parameter space of mass ratio and total mass. The upper panels are coloured by time to coalescence for fixed separations: $R = 10^3\, R_s$ (left) and $R = 10^{-2}$ pc (right). The lower panels are coloured by separation to coalesce within 10^8yr, in units of parsec (left) and Schwarzschild radii (R_s, right).

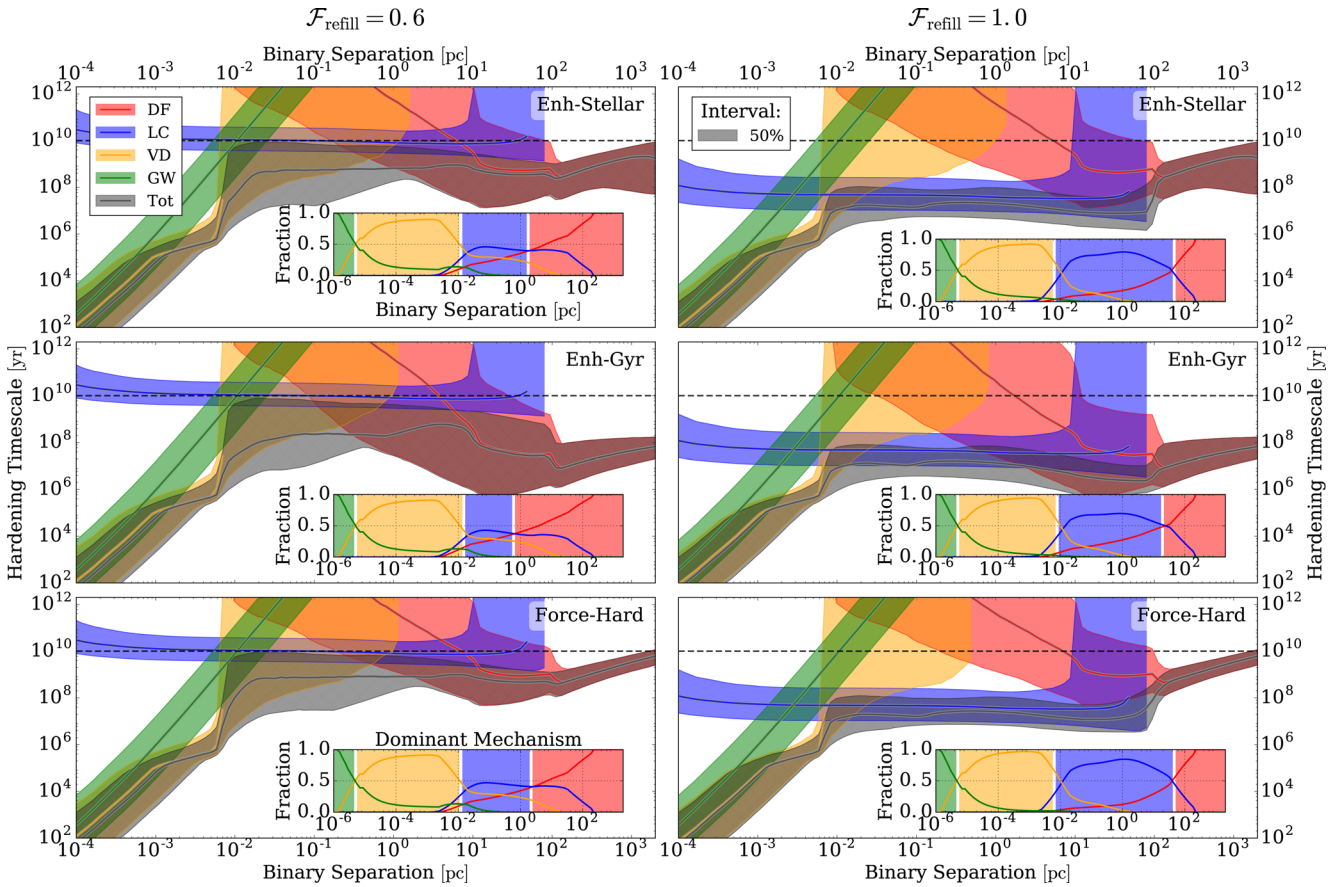

Figure A3. Binary hardening time-scales versus separation for different DF models (rows) and LC refilling fractions (columns). Coloured lines and bands show the median and 50 per cent intervals for each hardening mechanism: DF, LC scattering, VD, and GW emission with the total hardening rate shown by the grey, hatched region. The inset panels show the fraction of binaries dominated by each mechanism, also as a function of separation. **Note:** *in the 'Force-Hard' model, the binary separation is artificially altered faster than the time-scale shown (between about $\gtrsim 10^2$ pc). While the DF hardening time-scale is still relatively long, the binaries are forcibly hardened at a faster rate.*

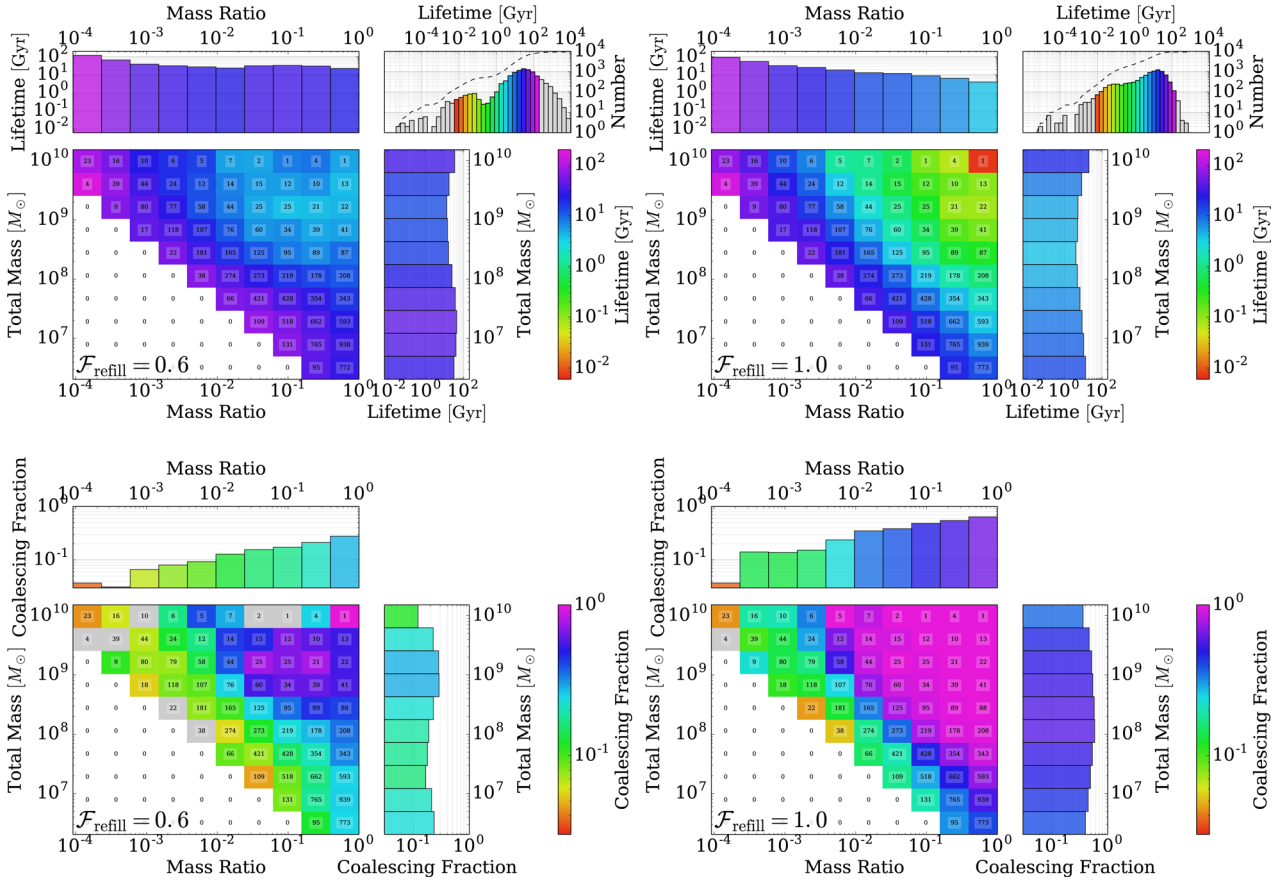

Figure A4. Binary lifetimes (upper row) and coalescing fractions (lower row) for our fiducial, moderately refilled LC ($\mathcal{F}_{\rm refill} = 0.6$; left) and an always full one ($\mathcal{F}_{\rm refill} = 1.0$; right). Both simulations use the 'Enh-Stellar' DF model. The overall distribution of MBHB lifetimes is shown in the upper-rightmost panel for each simulation, with the cumulative distribution plotted as the dashed line. For $\mathcal{F}_{\rm refill} = 0.6$, most binaries need to be both high total mass ($M \gtrsim 10^8\,{\rm M}_\odot$) and moderately high mass ratio ($\mu \gtrsim 10^{-2}$) to have lifetimes short enough to coalesce by redshift zero. In the $\mathcal{F}_{\rm refill} = 1.0$ case, on the other hand, either criteria is sufficient – and the coalescing fraction in that parameter space approaches unity.

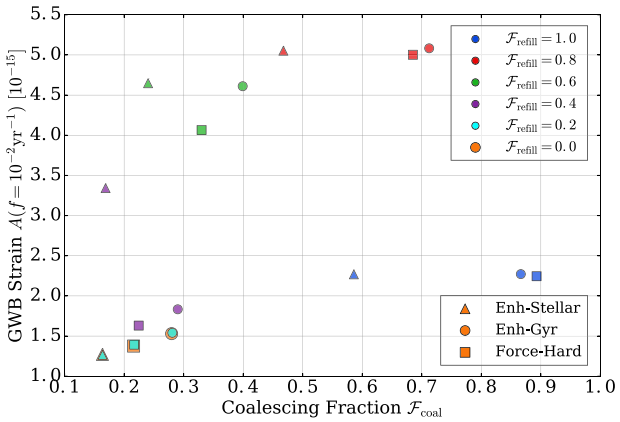

Figure A5. Dependence of GWB strain amplitude and binary coalescing fraction on LC refilling parameter and DF model. The GW strain is measured at very low frequencies, $f = 10^{-2}$ yr^{-1}, and the coalescing fraction is defined using the population of high mass-ratio $\mu > 0.1$ systems. Each symbol represents a different DF model, and each colour a different LC refilling parameter. $A_{\rm yr^{-1}}$ tends to increase monotonically with $\mathcal{F}_{\rm coal}$, until the highest $\mathcal{F}_{\rm refill}$. At $\mathcal{F}_{\rm refill} \approx 1.0$, LC stellar scattering is effective enough at these low frequencies to significantly attenuate the GWB amplitude.

APPENDIX B: SUMMARY OF QUANTITATIVE RESULTS FOR A VARIETY OF PARAMETER CONFIGURATIONS

Table B1. Summary of quantitative results for the GWB, and MBH binary lifetimes and coalescing/persisting fractions. Results are shown for the three DF models described in the text: 'Enh-Stellar', 'Enh-Gyr', and 'Force-Hard'. For our fiducial DF model ('Enh-Stellar'; the most conservative case), results are shown for both a case with no circumbinary, viscous discs ('VD: None') and with a disc using our standard parameters ('VD: Fiducial'). Two different LC scattering states are also compared, in which the LC is always full ($\mathcal{F}_{\rm refill} = 1.0$) and our fiducial case of moderately refilled ($\mathcal{F}_{\rm refill} = 0.6$). Amplitudes and spectral indices are presented at both $f = 1\ {\rm yr}^{-1}$ and $f = 0.1\ {\rm yr}^{-1}$. For lifetime and fractional statistics, binaries are split into four subsets: 'All' binaries included in our analysis, and those with mass ratios $\mu > 0.2$, total masses $M > 10^8\ {\rm M}_\odot$, and both criteria ($\mu > 0.2$, $M > 10^8\ {\rm M}_\odot$). The lifetimes shown are median values of systems in each of the 'Full' subset, and only those which coalesce by redshift zero ('Coal') – the number (and fraction) of such systems is given in the 'Coalescing Fraction' column. Finally, the fraction of systems that remain uncoalesced at separations $r > 1$ pc and $r > 10^2$ pc is shown in the 'Persisting Fraction' column. Results for a simulation using models with all fiducial parameters are shown in bold.

Dynamical friction (viscous disc)	Loss cone	GWB Amplitude $[10^{-16}]$ $1\ {\rm yr}^{-1}$	GWB Amplitude $[10^{-16}]$ $0.1\ {\rm yr}^{-1}$	GWB Spectral index $1\ {\rm yr}^{-1}$	GWB Spectral index $0.1\ {\rm yr}^{-1}$	Subset	Lifetimes [Gyr] Full (Coal)	Coalescing fraction		Persisting fraction >1pc	Persisting fraction $>10^2$pc
Enh-Stellar (VD: None)	$\mathcal{F}_{\rm refill} = 0.6$	4.3	18	−0.64	−0.55	All ($M > 10^6\ {\rm M}_\odot$)	37 (1.3)	1641/9270	(0.18)	0.67	0.46
						$\mu > 0.2$	32 (1.1)	1089/4759	(0.23)	0.57	0.33
						$M > 10^8\ {\rm M}_\odot$	26 (1.7)	504/2610	(0.19)	0.68	0.45
						$M > 10^8\ {\rm M}_\odot$, $\mu > 0.2$	11 (0.89)	184/478	(0.38)	0.37	0.01
	$\mathcal{F}_{\rm refill} = 1.0$	6.6	25	−0.65	−0.39	All	8.0 (1.3)	4556/9270	(0.49)	0.47	0.46
						$\mu > 0.2$	5.1 (1.1)	2840/4759	(0.60)	0.35	0.33
						$M > 10^8\ {\rm M}_\odot$	5.2 (0.80)	1414/2610	(0.54)	0.45	0.45
						$M > 10^8\ {\rm M}_\odot$, $\mu > 0.2$	0.37 (0.35)	469/478	(0.98)	0.01	0.01
Enh-Stellar (VD: Fiducial)	$\mathcal{F}_{\rm refill} = 0.6$	**3.7**	**15**	**−0.63**	**−0.59**	All	**29 (2.7)**	**1875/9270**	**(0.20)**	**0.66**	**0.46**
						$\mu > 0.2$	**26 (2.2)**	**1225/4759**	**(0.26)**	**0.55**	**0.33**
						$M > 10^8\ {\rm M}_\odot$	**17 (4.0)**	**608/2610**	**(0.23)**	**0.64**	**0.45**
						$M > 10^8\ {\rm M}_\odot$, $\mu > 0.2$	**6.9 (3.3)**	**214/478**	**(0.45)**	**0.28**	**0.01**
	$\mathcal{F}_{\rm refill} = 1.0$	4.7	17	−0.61	−0.38	All	7.7 (1.2)	4634/9270	(0.50)	0.47	0.46
						$\mu > 0.2$	4.8 (1.0)	2900/4759	(0.61)	0.35	0.33
						$M > 10^8\ {\rm M}_\odot$	4.9 (0.79)	1422/2610	(0.54)	0.45	0.45
						$M > 10^8\ {\rm M}_\odot$, $\mu > 0.2$	0.35 (0.35)	472/478	(0.99)	0.01	0.01
Enh-Gyr (VD: Fiducial)	$\mathcal{F}_{\rm refill} = 0.6$	3.8	16	−0.63	−0.62	All	13 (1.1)	3536/9270	(0.38)	0.38	0.11
						$\mu > 0.2$	13 (0.33)	1970/4759	(0.41)	0.30	0.06
						$M > 10^8\ {\rm M}_\odot$	8.3 (3.5)	1091/2610	(0.42)	0.39	0.15
						$M > 10^8\ {\rm M}_\odot$, $\mu > 0.2$	6.4 (3.0)	229/478	(0.48)	0.26	0.01
	$\mathcal{F}_{\rm refill} = 1.0$	4.7	17	−0.61	−0.38	All	0.42 (0.30)	7783/9270	(0.84)	0.12	0.11
						$\mu > 0.2$	0.37 (0.28)	4122/4759	(0.87)	0.07	0.06
						$M > 10^8\ {\rm M}_\odot$	0.32 (0.25)	2200/2610	(0.84)	0.15	0.15
						$M > 10^8\ {\rm M}_\odot$, $\mu > 0.2$	0.20 (0.20)	474/478	(0.99)	0.01	0.01
Force-Hard (VD: Fiducial)	$\mathcal{F}_{\rm refill} = 0.6$	3.6	15	−0.64	−0.61	All	14 (3.0)	3089/9270	(0.33)	0.32	0.02
						$\mu > 0.2$	17 (1.9)	1627/4759	(0.34)	0.24	0.01
						$M > 10^8\ {\rm M}_\odot$	7.8 (4.4)	1093/2610	(0.42)	0.30	0.04
						$M > 10^8\ {\rm M}_\odot$, $\mu > 0.2$	7.0 (3.1)	212/478	(0.44)	0.23	0.00
	$\mathcal{F}_{\rm refill} = 1.0$	4.7	17	−0.61	−0.38	All	0.42 (0.30)	7783/9270	(0.84)	0.03	0.02
						$\mu > 0.2$	0.37 (0.28)	4122/4759	(0.87)	0.02	0.01
						$M > 10^8\ {\rm M}_\odot$	0.32 (0.25)	2200/2610	(0.84)	0.05	0.04
						$M > 10^8\ {\rm M}_\odot$, $\mu > 0.2$	0.20 (0.20)	474/478	(0.99)	0.01	0.00

APPENDIX C: STELLAR, LOSS-CONE (LC) SCATTERING CALCULATIONS

The rate at which stars can refill the LC is governed by the 'relaxation time' ($\tau_{\rm rel}$). Following Binney & Tremaine (1987), consider a system of N masses m, with number density n, and characteristic velocities v. The relaxation time can be written as

$$\tau_{\rm rel} \approx \frac{N}{8\ln\Lambda}\tau_{\rm cross} \approx \frac{v^3}{8\pi G^2 m^2 n \ln\Lambda}, \tag{C1}$$

where $\ln\Lambda$ is again the Coulomb logarithm, and $\tau_{\rm cross} \equiv r/v$ is the crossing time. $\tau_{\rm rel}$ represents the characteristic time required to randomize a particle's velocity via scatterings, i.e. equation (C1) can be used to define the diffusion coefficient $\mathcal{D}_{v^2}$ as $\tau_{\rm rel} \approx v^2/\mathcal{D}_{v^2}$. If $t/\tau_{\rm rel} \ll 1$, then two-body encounters (and relaxation) have not been important.

Consider a distribution function (or phase-space density) $f = f(\boldsymbol{x}, \boldsymbol{v})$, such that the number of stars in a small spatial-volume $d^3\boldsymbol{x}$ and velocity-space volume $d^3\boldsymbol{v}$ is given as $f(\boldsymbol{x}, \boldsymbol{v})\, d^3\boldsymbol{x}\, d^3\boldsymbol{v}$. In a spherical system in which there are the conserved energy E and angular momentum $\boldsymbol{L}$, the six independent position and velocity variables can be reduced to these four independent, conserved quantities via the Jeans theorem. Furthermore, if the system is perfectly spherically symmetric – which we assume in our analysis, then the three independent angular momentum components can be replaced with the angular momentum magnitude, i.e. $f = f(E, L)$.

If we define a relative potential and relative energy, $\Psi \equiv -\Phi + \Phi_0$, and, $\mathcal{E} \equiv -E + \Phi_0 = \Psi - \frac{1}{2}v^2$, then we can calculate the number density as[26]

$$\begin{aligned} n(\boldsymbol{x}) = n(x) &= 4\pi \int_0^{\sqrt{2\Psi}} f(x, v) v^2\, dv \\ &= 4\pi \int_0^{\Psi} f(\mathcal{E})\, [2(\Psi - \mathcal{E})]^{1/2}\, d\mathcal{E}. \end{aligned} \tag{C2}$$

Inverting this relationship, the distribution function can be calculated from an isotropic density profile using

$$\begin{aligned} f(\mathcal{E}) &= \frac{1}{\pi^2\sqrt{8}} \frac{\rm d}{{\rm d}\mathcal{E}} \int_0^{\mathcal{E}} \frac{{\rm d}n}{{\rm d}\Psi} \frac{d\Psi}{(\mathcal{E} - \Psi)^{1/2}} \\ &= \frac{1}{\pi^2\sqrt{8}} \left[\mathcal{E}^{-1/2} \left(\frac{{\rm d}n}{{\rm d}\Psi} \right)_{\Psi=0} + \int_0^{\mathcal{E}} \frac{{\rm d}^2 n}{{\rm d}\Psi^2} \frac{d\Psi}{(\mathcal{E} - \Psi)^{1/2}} \right]. \end{aligned} \tag{C3}$$

We have found the latter form of (C3) to be much simpler and more reliable to implement.

We follow the discussion and prescription for LC scattering given by Magorrian & Tremaine (1999), corresponding to a single central object in a spherical (isotropic) background of stars. We adapt this prescription simply by modifying the radius of interaction to be appropriate for scattering with a binary instead of being tidally disrupted by a single MBH. A more extensive discussion of LC dynamics – explicitly considering MBH binary systems and asphericity – can be found in Merritt (2013).

Stars with a pericentre distance smaller than some critical radius $\mathcal{R}_{\rm crit}$ will interact with the binary. For a fixed energy ($\mathcal{E}$) orbit, there is then a critical angular momentum, $J_{\rm lc}(\mathcal{E}) = \mathcal{R}_{\rm crit}\,(2\,[\mathcal{E} - \Psi(\mathcal{R}_{\rm crit})])^{1/2} \approx \mathcal{R}_{\rm crit}\,(2\,[-\Psi(\mathcal{R}_{\rm crit})])^{1/2}$, which defines the LC (Frank & Rees 1976; Lightman & Shapiro 1977). In general, the number of stars with energy and angular momentum in the range $d\mathcal{E}$ and dJ^2 around $\mathcal{E}$ and J^2 can be calculated as

$$N(\mathcal{E}, J^2)\, d\mathcal{E}\, dJ^2 = 4\pi^2\, f(\mathcal{E}, J^2) \cdot P(\mathcal{E}, J^2)\, d\mathcal{E}\, dJ^2, \tag{C4}$$

where $P(\mathcal{E}, J^2)$ is the stellar orbital period. For an isotropic stellar distribution $f(\mathcal{E}, J^2) = f(\mathcal{E})$, and $P(\mathcal{E}, J^2) \approx P(\mathcal{E})$. The total number of stars can be calculated as

$$N_i(\mathcal{E})\, d\mathcal{E} = 4\pi^2\, f(\mathcal{E})\, P(\mathcal{E})\, J_i^2(\mathcal{E})\, d\mathcal{E}. \tag{C5}$$

For the number of stars in the LC $N_{\rm lc}(\mathcal{E})$, this uses the LC angular momentum, $J_i^2(\mathcal{E}) = J_{\rm lc}^2(\mathcal{E})$; and for all stars $N(\mathcal{E})$, the circular (and thus maximum) angular momentum, $J_i^2(\mathcal{E}) = J_{\rm c}^2(\mathcal{E})$. When we initially calculate the distribution function, we use the stellar density profile from Illustris galaxies which have recently hosted an MBH 'merger' event (see Section 2.2). We assume that the resulting distribution function $f(\mathcal{E})$ is (so far) unperturbed by the MBH binary, i.e. it does not take into account stars already lost (scattered). The resulting $N_{\rm lc}(\mathcal{E})$ from equation (C5) then corresponds to the number of stars in the 'full' LC specifically.

Stars in the LC are consumed on their orbital time-scale $\tau_{\rm orb} = P(\mathcal{E})$. The rate of flux of stars to within $\mathcal{R}_{\rm crit}$ is then

$$F_{\rm lc}^{\rm full}(\mathcal{E})\, d\mathcal{E} = 4\pi^2\, f(\mathcal{E})\, J_{\rm lc}^2(\mathcal{E})\, d\mathcal{E}, \tag{C6}$$

coming almost entirely from within the central objects sphere of influence $\mathcal{R}_{\rm infl}$, defined as $M(r < \mathcal{R}_{\rm infl}) \approx M_\bullet$. Refilling of the LC occurs on the characteristic relaxation time-scale $\tau_{\rm rel}$. From equation (C1), it is clear that $\tau_{\rm orb}/\tau_{\rm rel} \approx \tau_{\rm cross}/\tau_{\rm rel} \ll 1$, i.e. the LC is drained significantly faster than it is refilled – and the LC will, in general, be far from 'full'.

To calculate the steady-state flux of the LC, the Fokker–Planck equation must be solved with a fixed (unperturbed) background stellar distribution at the outer edge of the LC and no stars surviving within the scattering region at the inner edge. A full derivation can be found in Magorrian & Tremaine (1999), which yields an equilibrium flux of stars,

$$F_{\rm lc}^{\rm eq}(\mathcal{E})\, d\mathcal{E} = 4\pi^2 P(\mathcal{E}) J_{\rm c}^2(\mathcal{E}) f(\mathcal{E}) \frac{\mu(\mathcal{E})}{\ln R_0^{-1}(\mathcal{E})}, \tag{C7}$$

where the angular momentum diffusion parameter $\mu \equiv 2r^2\mathcal{D}_{v^2}/J_{\rm c}^2$, and

$$\ln R_0^{-1} = -\ln \mathcal{R}_{\rm crit} + \begin{cases} q & q \geq 1 \\ 0.186q + 0.824\sqrt{q} & q < 1 \end{cases} \tag{C8}$$

describes the effective refilling radius depending on which refilling regime ('pin hole' or 'diffusive', see fig. 1 of Lightman & Shapiro 1977) is relevant, for a refilling parameter $q(\mathcal{E}) \equiv P(\mathcal{E})\, \mu(\mathcal{E})/\mathcal{R}_{\rm crit}(\mathcal{E})$.

Equations (C6) and (C7) give the full and steady-state LC fluxes, which are interpolated between using a logarithmic, 'refilling fraction' (equation (11)) which then determines the hardening rate of each binary in our simulations.

[26] Φ_0 is arbitrary, but $\Phi_0 \equiv E(r \to \infty)$ may be convenient.

This paper has been typeset from a TEX/LATEX file prepared by the author.