CERN-TH-2024-016

Self-interacting dark matter solves the final parsec problem of supermassive black hole mergers

Gonzalo Alonso-Álvarez,[1, 2, *] James M. Cline,[2, 3, †] and Caitlyn Dewar[2, ‡]

[1] *Department of Physics, University of Toronto, Toronto, ON M5S 1A7, Canada*
[2] *McGill University Department of Physics & Trottier Space Institute, 3600 Rue University, Montréal, QC, H3A 2T8, Canada*
[3] *CERN, Theoretical Physics Department, Geneva, Switzerland*

Evidence for a stochastic gravitational wave (GW) background, plausibly originating from the merger of supermassive black holes (SMBHs), is accumulating with observations from pulsar timing arrays. An outstanding question is how inspiraling SMBHs get past the "final parsec" of separation, where they have a tendency to stall before GW emission alone can make the binary coalesce. We argue that dynamical friction from the dark matter (DM) spike surrounding the black holes is sufficient to resolve this puzzle, if the DM has a self-interaction cross section of order cm^2/g. The same effect leads to a softening of the GW spectrum at low frequencies as suggested by the current data. For collisionless cold DM, the friction deposits so much energy that the spike is disrupted and cannot bridge the final parsec, while for self-interacting DM, the isothermal core of the halo can act as a reservoir for the energy liberated from the SMBH orbits. A realistic velocity dependence, such as generated by the exchange of a massive mediator like a dark photon, is favored to give a good fit to the GW spectrum while providing a large enough core. A similar velocity dependence has been advocated for solving the small-scale structure problems of cold DM.

1. Introduction. Nearly 40 years after its theoretical basis was established [1], gravitational wave astronomy has entered a new era with the advent of pulsar timing arrays. Differential time delays of pulsar signals, consistent with a stochastic gravitational wave (GW) background at nanoHertz frequencies, were detected in 2021 by NANOGrav [2], the Parkes Pulsar Timing Array (PPTA) [3] and the European Pulsar Timing Array (EPTA) [4]. The GW interpretation has been reinforced by the $\sim 3\sigma$ evidence for Hellings-Downs correlations in the NANOGrav's recent 15-year data analysis [5], and the compatible measurements by PPTA [6], EPTA [7], and the Chinese Pulsar Timing Array [8]. A plausible origin for the signal are the mergers of supermassive black holes (SMBHs) [9–12] across cosmic time.

One challenge to the SMBH interpretation of the nHz GW background is that the simplest models (assuming GW emission is the only source of energy loss) predict that the timescale for merging once the SMBH separation is of order 1 pc is larger than a Hubble time; this "final parsec" problem suggests that the SMBHs would never merge [13]. At larger distances, three-body interactions with stars allow the SMBH pair to lose energy, "hardening" the binary and driving the inspiral. It was suggested that axisymmetry of the galactic halo profile is sufficient to overcome this problem [14], but this has been debated [15]. Another possibility is that interactions of the SMBHs with an accretion disk accelerate the infall [16–18]. The simulations in [19] show that these astrophysical mechanisms are generally ineffective to reduce the inspiral time below several Gyr, adding motivation to look for others.

A less-explored mechanism for accelerating the infall is the dynamical friction (DF) [20] experienced by the SMBH pair as it rotates through the surrounding dark matter (DM) halo. This effect has been studied for ultralight DM [21–27] and in the context of intermediate- or stellar-mass BH binaries [28–32]. Black holes accumulate surrounding DM overdensities, known as "spikes" [33], which can exceed the galactic DM halo density and enhance the DF damping the BH orbital motion.

Some effects of collisionless cold dark matter (CDM) friction were recently considered for SMBH contributions to PTA signals in Ref. [34]. It was shown that the low-frequency turnover in the spectrum, suggested by the data, can be ascribed to DM frictional energy loss, which dominates over GW losses at intermediate BH separations. The effect of eccentricity of the SMBH orbits was studied in Ref. [35]. The impact on the final parsec problem has however not yet been addressed. Here we show that DM friction drives the binary infall at intermediate separations (see Fig. 1) and can reduce the inspiral timescale to $\lesssim 1$ Gyr, provided that the DM spike is able to absorb the frictional energy without being disrupted. We argue that this is possible for self-interacting dark matter (SIDM), but not for standard CDM.

Self-interacting dark matter (SIDM) has been proposed to address discrepancies between the predictions of CDM and observations of galactic structure on small scales, notably the core versus cusp problem (see [36] for a review). We consider a range of velocity-dependent scattering cross sections, motivated by evidence for scale-dependence of halo cores [37], finding that simple power

* gonzalo.alonso@utoronto.ca; ORCID: 0000-0002-5206-1177
† jcline@physics.mcgill.ca; ORCID: 0000-0001-7437-4193
‡ caitlyn.dewar@mail.mcgill.ca; ORCID: 0009-0002-2191-7224

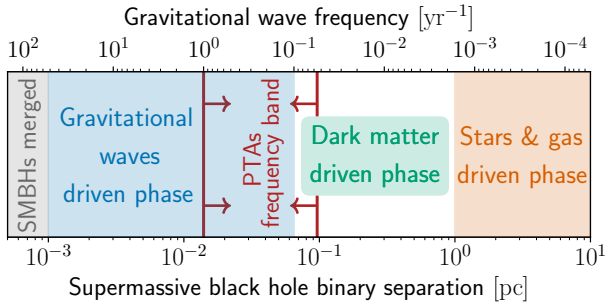

Figure 1. Agents driving the SMBH binary hardening versus separation (bottom axis) or gravitational wave emission frequency (top axis), compared to the bandwidth of current pulsar timing arrays. Regions correspond to a typical merger of two $3 \times 10^9\, M_\odot$ SMBHs at $z = 0$ within an SIDM spike.

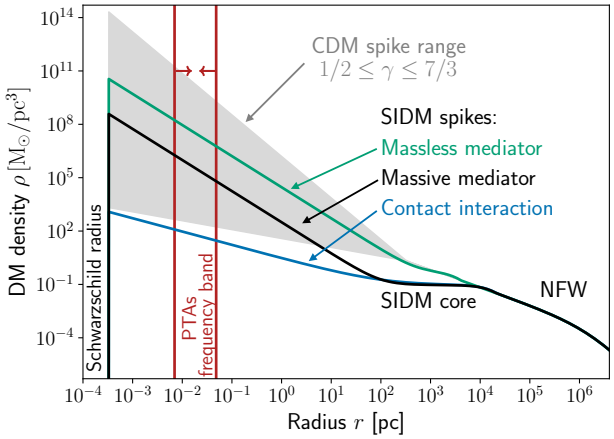

Figure 2. Spike profiles around a SMBH with $M_\bullet = 6 \times 10^9\, M_\odot$ at $z = 0$. NFW parameters of the host galaxy are $r_s \simeq 2$ Mpc, $\rho_s \simeq 3 \times 10^{14}\, M_\odot/\mathrm{Mpc}^3$, and $M_{200} \simeq 2 \times 10^{16}\, M_\odot$. A range of possible CDM spikes is shaded in grey. For SIDM, the blue (green) line corresponds to a contact interaction (massless mediator) with $\sigma_0/m = 3\,\mathrm{cm}^2/\mathrm{g}$ ($30\,\mathrm{cm}^2/\mathrm{g}$). The black line represents a massive mediator with $\sigma_0/m = 3\,\mathrm{cm}^2/\mathrm{g}$ and a transition velocity $v_t = 500\,\mathrm{km/s}$. The age of the core is 100 Myr. Vertical red lines delimit the range for which GW emission is detectable at current PTAs. The profiles are cut off at twice the Schwarzschild radius of the BH [50].

laws do not optimally fit the GW signal. However, a broken power law, as results from a realistic massive force carrier like a dark photon, is able to reproduce the observations, including the hint of reduced power at low frequencies. Our preferred cross section values are compatible with those favored by small-scale structure.

2. DM density profiles. The GW signal from merging SMBHs depends upon their masses M_1 and M_2, parameterized by $q \equiv M_2/M_1 \leq 1$, with final SMBH mass $M_\bullet = M_1(1+q)$. For a given $M_\bullet$ at redshift z, we determine the mean values of the Navarro-Frenk-White (NFW) [38] halo parameters of the DM density profile following [39–41], as described in Appendix A. This provides a starting point for determining the DM spike around the SMBH.

For CDM, the NFW profile is superseded by the DM spike contribution at radii $r < r_{\rm sp}$, with spike radius $r_{\rm sp} \cong 0.2\, r_{\rm 2M}$, where $r_{\rm 2M}$ is the radius at which the mass enclosed by the NFW profile is $2M_\bullet$ [42, 43]. For an NFW halo, $r_{\rm 2M}^2 \cong M_\bullet/\pi\rho_s r_s$, assuming $r_{\rm 2M} \ll r_s$. The spike profile has the form

$$\rho_{\rm sp}(r) = \rho_{\rm sp}\,(r_{\rm sp}/r)^\gamma\,, \tag{1}$$

where $\rho_{\rm sp} = \rho_{\rm NFW}(r_{\rm sp})$, and the exponent γ is subject to astrophysical uncertainties [44–46], including evolution during the merger [47]. We thus consider γ as a free parameter, with physically motivated values varying between 7/3 for an adiabatically grown spike [33] and 1/2 for a spike formed right after a galaxy merger [44]. The spike can be tapered off by DM annihilations [48, 49], but this will not play a role here.

For SIDM, the galactic halo is flattened within a distance r_1, the radius of the core, due to SI-driven thermalization [37]. The interaction cross section σ determines r_1 by demanding at least one scattering per DM particle for $r < r_1$ during the age of the core,

$$\frac{\langle\sigma v\rangle}{m} \cdot \rho_{NFW}(r_1) \cdot t_{\rm age} \sim 1\,. \tag{2}$$

For $r < r_1$, the NFW profile is replaced by an isothermal one, which satisfies the Poisson equation $v_0^2\nabla^2\rho = -4\pi G\rho$, with one boundary condition (b.c.) being regularity at the origin [37]. Here, v_0^2 is the DM velocity dispersion, which is constant in the isothermal region [51]. For the other b.c. of the Poisson equation, we follow Ref. [37], choosing $\rho'(r_1)$ such that the mass enclosed within r_1 is the same as in the original NFW profile. Satisfying both b.c.'s fixes the value of v_0^2. Technical details about this procedure are given in Appendix B.

As Eq. 2 shows, the self-interaction cross section enters our results only in the combination $\langle\sigma v\rangle\, t_{\rm age}$. Since $t_{\rm age}$ includes the unknown time that the binary takes to reach a ~ 10 pc separation through interactions with the baryonic environment, we consider it as a nuisance parameter in our analysis, normalizing to $t_{\rm age} = 1$ Gyr. It is straightforward to rescale our results to other values of $t_{\rm age}$, which could be predicted in a more detailed analysis including the effects of stars and gas.

The SIDM halo can be matched to the spike similarly to CDM. Ref. [50] considered several dependences of the SI cross section on the DM relative velocity v,

$$\langle\sigma_i(v)v\rangle = \sigma_0 v_0\,(v_{\rm ref}/v_0)^a = \sigma_0 v_{\rm ref}\,(v_{\rm ref}/v_0)^{a-1}\,, \tag{3}$$

with $a = 0, 1, \ldots, 4$. For example, $a = 4$ corresponds to Coulomb scattering, such as mediated by a massless force carrier, while $a = 0$ represents isotropic scattering, resulting from a contact interaction. The choice of $v_{\rm ref}$ in Eq. (3) is arbitrary; here we take $v_{\rm ref} = 100\,\mathrm{km/s}$, and quote values of σ_0/m, as is standard in the literature.

Ref. [50] identifies the spike radius with the radius of influence of the final BH: $r_{\rm sp} = GM_\bullet/v_0^2$, so that setting

$\rho_{\rm sp} = \rho_0$ in Eq. (1) gives the SIDM spike profile. The mass of the spike is typically comparable to that of the SMBH host. The spike density exponent depends on a as $\gamma = (3+a)/4$. We only consider values of σ_0/m large enough so that $r_{\rm sp} \leq r_1$. The resulting $a = 0$ and $a = 4$ profiles are shown in Fig. 2. For $a = 0$, the isothermal core is larger and has a larger $v_0 \simeq 500\,{\rm km/s}$ than for $a = 4$, for which $v_0 \simeq 220\,{\rm km/s}$. Thus, the spike is steeper and more extended for a massless mediator.

A more realistic cross section need not have such simple behavior. Interactions mediated by a massive force carrier, such as dark photons of mass m_γ, behave as $a = 0$ for $v < v_t$ and as $a = 4$ for $v > v_t$, with a transition velocity $v_t \sim c\, m_\gamma/m$. Although the DM velocity dispersion in the core is constant, it starts to rise as $v/v_0 \sim 4/11\,(r_{\rm sp}/r)^{1/2}$ within the spike [50] and can enter the $a = 4$ regime, thereby increasing γ. We model the spike profile as Eq. (1) with $\gamma = 3/4$ in the outer region and $\gamma = 7/4$ in the inner region where $v > v_t$ [52]. These two regimes meet at the transition radius r_t at which $v = v_t$ [53]. If $v_t < v_0$, the $a = 4$ regime is applicable throughout the whole core and spike. An example is shown in Fig. 2, where it is clear that the massive mediator interpolates between $a = 0$ at large radii and $a = 4$ in the inner region.

In summary, for either CDM or SIDM, the outer NFW halo parameters are determined by the BH mass $M_\bullet$, while the details of the BH spike depend on additional parameters: γ for CDM; $\sigma_0 v_{\rm ref}/m$ and a (or v_t for a massive mediator) for SIDM.

3. SMBH merger dynamics. For simplicity, we assume the SMBHs to be in a circular orbit with separation R, angular frequency $\omega = (GM_\bullet/R^3)^{1/2}$, and that for separations $\lesssim 10\,{\rm pc}$ the orbit decays solely due to GW emission and dynamical friction with the DM. The power in GWs is given by $P_{\rm gw} = (32/5) q^2 (1+q) G^4/c^5 (M_1/R)^5$ [54], while the frictional power loss in the common-envelope spike is (see Appendix C)

$$P_{\rm df} = 12\pi\, q^2 \sqrt{1+q}\,(GM_1)^{3/2} R^{1/2} \times \left[\frac{N_1(q)}{q^3}\rho_{\rm sp}\left(\frac{qR}{1+q}\right) + N_2(q)\rho_{\rm sp}\left(\frac{R}{1+q}\right)\right], \quad (4)$$

where $N_{1,2} = 1$ for CDM and $N_1 = N_2 \simeq 0.2$ for SIDM and $q = 1$. $R(t)$ is fixed by equating $P_{\rm df} + P_{\rm gw}$ to the rate of change of the orbital energy, $\dot{E}_{\rm orb} = qGM_1^2\dot{R}/(2R^2)$.

Simple analytic solutions for $R(t)$ exist when either of the two loss terms dominate. Since the evolution due to GW emission is well known, we focus on the dynamical friction. Defining the characteristic timescale $t_{\rm sp} = (r_{\rm sp}^3/GM_1)^{1/2}$ and dimensionless time and radial variables $\tau = t/t_{\rm sp}$, $x = R/(2r_{\rm sp})$, the equation of motion takes the form $dx/d\tau = -Bx^p$, where $p = 5/2 - \gamma$ and $B = f(q,\gamma)\,\rho_{\rm sp}\, r_{\rm sp}^3/M_1$ with $f(q,\gamma) = 96\pi\, q[(1+q)/2]^{\gamma+1/2}(N_2 + N_1 q^{-3-\gamma})$.

The timescale for hardening due to DF is then

$$t_{\rm df} \equiv \partial t/\partial \ln R|_{\rm df} = (t_{\rm sp}/B)\, x_{\rm crit}^{1-p}, \quad (5)$$

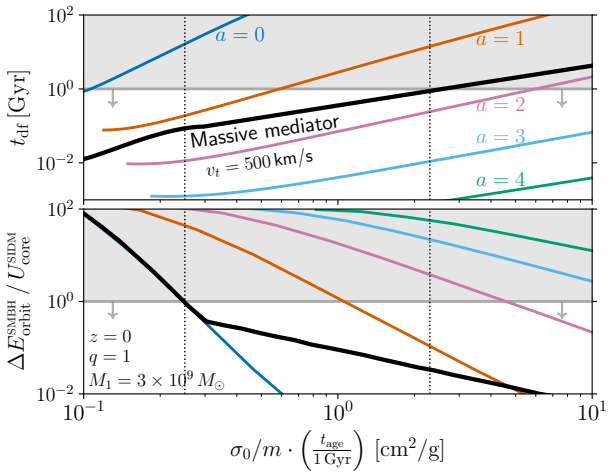

Figure 3. *Upper*: Time for SIDM dynamical friction to bring the SMBH separation below 0.1 pc where GW emission dominates, versus SI cross section. *Lower*: Ratio of orbital energy transmitted by the binary to the DM spike via dynamical friction, to the gravitational binding energy of the SIDM isothermal core. Colors identify different velocity dependence of $\sigma(v)$, the black line corresponding to a massive mediator with $v_t = 500\,{\rm km/s}$. Shaded regions are disfavored and dotted lines delineate the viable range for the massive mediator.

with critical separation $x_{\rm crit}$ where DF is weakest within the DM spike. If $\gamma \geq 3/2$, this occurs at the outer edge, $x_{\rm crit} = R_\star/(2r_{\rm sp})$, where $R_\star$ is the separation beyond which hardening by interactions with stars and gas is efficient; we conservatively take $R_\star = 10$ pc. For shallow spikes with $\gamma < 3/2$, DF weakens at small separations; then $x_{\rm crit} = R_{\rm gw}/(2r_{\rm sp})$, where $R_{\rm gw}$ is the separation at which GW emission becomes sufficiently strong to complete the merger. In the marginal case $\gamma = 3/2$, $x_{\rm crit}^{1-p} \to \ln(R_*/R_{\rm gw})$. For SIDM with a massive mediator, $x_{\rm crit}$ must be evaluated at the intermediate separation r_t at which DM particles have velocity v_t (see Appendix B). Using the GW hardening timescale [54]

$$t_{\rm gw} \equiv \left.\frac{\partial t}{\partial \ln R}\right|_{\rm gw} = \frac{5c^5}{64G^3}\frac{(2R_{\rm gw})^4}{M_1^3\, q\,(1+q)}, \quad (6)$$

we find that $R_{\rm gw} = 0.1-0.2\,{\rm pc}$ for binaries with $M_1 = 3\times 10^9\,M_\odot$ to merge within $0.1-1$ Gyr. We conservatively set $R_{\rm gw} = 0.1$ pc for our numerical evaluations.

For CDM, $12\pi\rho_{\rm sp} r_{\rm sp}^3/M_1 \simeq (1+q)/2$ and thus $B = \mathcal{O}(1)$, since $q \simeq 1$ for SMBH mergers contributing most strongly to the stochastic GW background. For binaries with $M_1 \gtrsim 10^9\,M_\odot$, one finds $t_{\rm sp} \simeq 2\times10^{-3}$ Gyr. Since for a shallow spike $x_{\rm crit} \simeq 1.5\times10^{-4}$, it follows that $t_{\rm df} \lesssim 1$ Gyr as long as $\gamma \gtrsim 0.7$. One would then conclude that CDM is able to solve the final parsec problem; however, the back-reaction of the black hole motion on the spike must be taken into account [55].

The energy lost by the SMBH binary when its orbit shrinks from R_* to $R_{\rm gw}$ is

$$\Delta E_{\rm orb} = q\, G\, M_1^2\left(R_{\rm gw}^{-1} - R_*^{-1}\right)/2. \quad (7)$$

This energy heats the DM particles and should be compared to the gravitational binding energy of the spike, U, to estimate whether it can absorb that much energy:

$$\frac{U}{G} = M_\bullet \int d^3x \, \frac{\rho_{\rm sp}(\vec{x})}{|\vec{x}|} + \int d^3x_1 d^3x_2 \frac{\rho_{\rm sp}(\vec{x}_1)\rho_{\rm sp}(\vec{x}_1)}{2|\vec{x}_1 - \vec{x}_2|}$$
$$= 4\pi M_\bullet \rho_{\rm sp} r_{\rm sp}^2 \left(\frac{1-\epsilon^{2-\gamma_{\rm sp}}}{2-\gamma_{\rm sp}}\right) + 4\pi^2 \rho_{\rm sp}^2 r_{\rm sp}^5 \, g(\gamma), \quad (8)$$

where ϵ is the ratio of the Schwarzschild radius to the spike radius, and $g(\gamma \lesssim 2) \sim 3$, growing to $g(7/3) \sim 20$. Because the mass of the spike is of order $M_\bullet$, the two contributions are comparable, and they are $4-5$ orders of magnitude smaller than $\Delta E_{\rm orb}$ for typical $M_\bullet \sim 6 \times 10^9 \, M_\odot$ contributing to the GW signal. Hence the CDM spike is completely disrupted by the DF energy deposited.

This obstacle can be overcome if DM has self-interactions that rethermalize and replenish the spike sufficiently fast. The SIDM dynamical friction timescale follows from Eq. (5) with $r_{\rm sp} = GM_\bullet/v_0^2$, which depends indirectly upon σ_0/m. For our numerical evaluations, rather than the timescale we use the inspiral duration resulting from integrating the equation of motion from $R_\star$ to $R_{\rm gw}$, see Appendix B for more details. Fig. 3 (upper) shows $t_{\rm df}$ versus σ_0/m for $a = 0, 1, 2, 3, 4$ and for a massive mediator, for a representative SMBH binary with the same combined mass used in Fig. 2, and thus the same host DM halo NFW parameters. Larger values of a produce more pronounced spikes that exert more DF. SIDM yields sub-Gyr inspiral times for values of σ_0/m below an a-dependent threshold. As σ_0/m grows, the isothermal core becomes larger and less dense, resulting in a weaker spike that applies less DF on the SMBHs.

The energy injected in the SIDM spike by the BHs heats and disperses the DM in it, but at the same time the self-interactions repopulate the spike with DM from the isothermal core. If this occurs fast enough, the spike survives the back-reaction and continues to harden the binary. Since we take the spike to be in equilibrium, $t_{\rm df}$ should be larger than the SIDM core relaxation time scale $t_{\rm r} \simeq [\rho_c(\sigma/m)v_0]^{-1}$ (the mean time between particle collisions), which coincides with $t_{\rm age}$ in Eq. (2). Thus, for our approximations to be self-consistent, we demand that $t_{\rm df} = t_{\rm age}$, although this technical assumption could be relaxed in a more general approach.

If large enough, the isothermal core acts as a reservoir whose total binding energy can be sufficient to absorb the orbital energy lost by the binary. This puts an a-dependent lower limit on σ_0/m, shown in Fig. 3 (lower). Lower values of a are favoured since they produce larger cores. Combined with the upper limit from solving the final parsec problem, there is a range of viable σ_0/m values that are listed in Table II in the Appendix. The best-performing model is the massive mediator, which combines a large core in the $a = 0$ regime with a steep spike in the $a = 4$ phase (see Fig. 2), and prefers $\sigma_0/m \sim \mathcal{O}({\rm cm}^2/{\rm g})$, as highlighted by the dotted vertical lines in Fig. 3. The preferred range of SI cross sections, which is compatible with small-scale structure constraints [36], can be tightened by studying how GW emission from the SMBH binaries is modified by the DM.

4. GW spectrum. To calculate the total GW energy emitted by a single binary at each frequency, recall that the frequency of GWs at the source is twice the orbital frequency,

$$f_s = \omega/\pi = (2\pi t_{\rm sp})^{-1} \sqrt{(1+q)/2} \, x^{-3/2}, \quad (9)$$

using the characteristic timescale and dimensionless separation $x = R/(2r_{\rm sp})$ defined in the previous section. The differential GW energy spectrum is (see Appendix C)

$$\frac{{\rm d}E_{\rm gw}}{{\rm d}f_s} = \frac{q}{6 f_s x} \frac{GM_1^2}{r_{\rm sp}} \frac{P_{\rm gw}}{P_{\rm gw} + P_{\rm df}}, \quad (10)$$

where x is understood to be a function of f_s via Eq. (9). GW emission finishes when the two BH horizons merge at a separation $R_{\rm min} = 2GM_1(1+q)$, resulting in a sharp cutoff of the spectrum at high frequency. The GW signal produced by a population of cosmological SMHB mergers is described by the characteristic strain [56, 57]

$$h_c^2(f) = \frac{4G}{\pi c^2 f} \int {\rm d}z \, {\rm d}M \, {\rm d}q \, \frac{{\rm d}^3 n}{{\rm d}z \, {\rm d}M \, {\rm d}q} \frac{{\rm d}E_{\rm gw}}{{\rm d}f_s}, \quad (11)$$

where $f = f_s/(1+z)$ is the frequency of the GW at the detection point.

The NANOGrav analysis [10] includes a phenomenological parameter for the total hardening timescale. Its posterior distribution is peaked at the lower bound of the range considered, 0.1 Gyr, signaling a preference for fast coalescence, with a 1-σ region extending to a few Gyr. We therefore approximate the inspiral as being instantaneous compared to the Hubble time, and explore values of $t_{\rm age}$ in the 10 Myr - 1 Gyr range. This is consistent with Section 3, which showed that DF from the DM spike can yield a sub-Gyr merger time for typical SMBHs from separations as large as 10 pc, beyond which interactions with stars and gas are assumed to be effective. We neglect possible effects of ambient stars or gas on the GW waveform within the PTA frequency range, to clearly illustrate the effects of DM.

The parametrization of ${\rm d}^3n$, described in Appendix D, is based on Ref. [58], used by NANOGrav [10] and EPTA [11]. We expect the astrophysical parameters determining it to be largely unchanged by DM effects, except for the overall normalization of the signal that is sensitive to the hardening timescale. We therefore fix them to the best-fit values found by Ref. [10] and allow only the normalization parameter ψ_0 to vary.

We extract central h_c values and upper and lower error bars for the first five NANOGrav [5], ten PPTA [6], and six EPTA [11] frequency bins, by fitting two one-sided gaussians to the probability distributions represented by the violin plots; see Table III in the Appendix. From these, χ^2 values are calculated for model predictions using Eq. (11). The characteristic strain spectrum for the

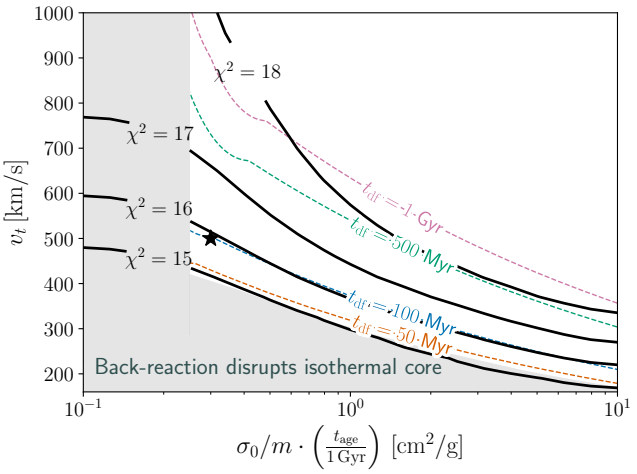

Figure 4. Contours of χ^2 from the fit to the characteristic strain spectra of NANOGrav [10], PPTA [6], and EPTA [11], for a SIDM model with a massive mediator, in the plane of v_t (velocity at which the SI transitions from contact-like to long-range) versus interaction cross section per DM mass at 100 km/s. In the grey shaded region, the orbital energy lost by SMBHs with equal $3 \times 10^9\, M_\odot$ masses at $z = 0$ is larger than the gravitational binding energy of the SIDM core. Dashed lines show contours of the DF timescale, assuming $t_{\rm age} = t_{\rm df}$.

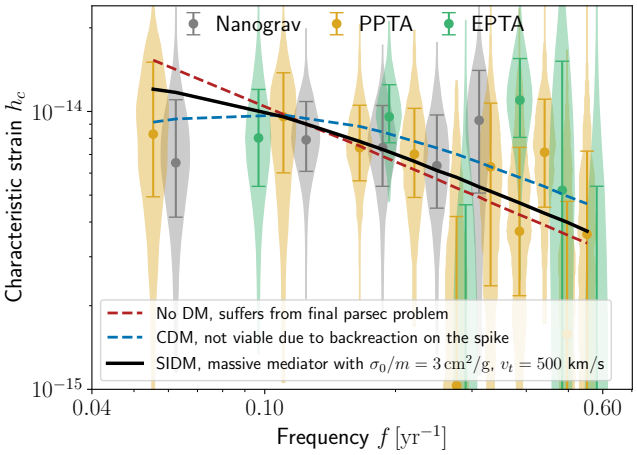

Figure 5. Selected characteristic strain spectra compared to PTA data. The SIDM model corresponds to the black star in Fig. 4, for which $t_{\rm df} = 100$ Myr.

best-fitting model in each category is shown in Fig. 5. The best fits among them give $\chi^2 \lesssim 15$, which is lower than the expected ~ 21 due to correlations between the different frequency bins in the data [59]. The normalization is treated as a nuisance parameter, and for our best-fitting models is $\psi_0 \sim -2.5$. This is significantly smaller than that found in the fiducial NANOGrav analysis [10] and better matches astrophysical expectations [58, 60], but the conclusion may vary when the finite duration of the mergers is taken into account.

As is visible in Fig. 5, the presence of DM improves the fit by softening the gravitational wave spectrum at low frequencies, where energy is being lost to DF rather than emitted in GWs. Although a CDM spike with $\gamma \simeq 1.5$ appears to give a good fit, the destructive back-reaction undermines this result. SIDM with a single power-law velocity dependent cross section can only produce a moderate softening due to the requirement of having a large enough isothermal core to survive back-reaction. The minimum χ^2 value within the viable σ_0/m range can be found in Table II in the Appendix for these models.

SIDM with a massive mediator is the only model among the ones studied that can simultaneously absorb the DF heat and produce a noticeable softening in the GW spectrum. As shown in Fig. 4 as a function of the two free parameters of the model, we find good fits to the PTA data, compatible with merging well within 0.1 Gyr, for $v_t \sim 300\text{-}600\,{\rm km/s}$ and $\sigma_0/m \sim 2.5\text{-}25\ {\rm cm}^2/{\rm g} \cdot (100\,{\rm Myr}/t_{\rm age})$. Recall that we impose $t_{\rm age} = t_{\rm df}$ as given by the colored dashed contours in Fig. 4.

5. Conclusions. Despite astrophysical uncertainties about their detailed nature, there is no doubt that dark matter spikes exist around supermassive black hole binaries, and thus contribute to the dynamical friction accelerating the decay of their orbit. We have shown that well-motivated models of velocity-dependent self-interacting dark matter have two correlated and desirable effects: robustly resolving the long-standing final parsec problem, thus allowing GW emission to finish the inspiral process, and softening the GW strain spectrum at low frequencies, matching the feature hinted at by PTA data. In contrast, collisionless CDM spikes are incapable of absorbing the frictional heat and are destroyed by the merger.

It is encouraging that the properties of the self-interaction cross section favored by the PTA signal is compatible with what was already proposed to solve the small-scale structure problems of CDM. In particular, the required magnitude and broken power law velocity dependence can reconcile the different cross section values needed to explain observations of galaxy and galactic cluster core sizes [37].

Our preliminary study opens the way to using gravitational wave signals from supermassive black hole mergers as a probe of dark matter microphysics. Ways to sharpen the predictions include accounting for the finite duration of the inspiral in an improved statistical analysis, using a more realistic velocity dependence for the SIDM scattering cross section, and performing numerical simulations to validate the analytical calculations for the back-reaction on the DM profile.

Acknowledgments. We thank A. Benson, L. Combi, D. Curtin, C. Mingarelli, M. Reina-Campos, J. Sievers, S. Tulin, and G.-W. Yuan for helpful discussions. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix A: SMBH and halo relations

The effect of the DM spike is correlated with $M_\bullet$ by a series of relations involving the (pseudo)bulge mass $M_{\rm bul}$ of the host galaxy and its stellar mass $M_\star$. At a given redshift z, the halo-to-stellar mass relation is a function $[M_{200}/M_\star](z)$, which for our fiducial results we take from Ref. [39]:

$$\frac{M_{200}}{M_\star}(z) = \frac{1}{2A(z)}\left[\left(\frac{M_{200}}{M_A(z)}\right)^{-\beta(z)} + \left(\frac{M_{200}}{M_A(z)}\right)^{\gamma(z)}\right], \tag{A1}$$

where polynomial fits to the functions A, M_A, α, β, γ of z are given. The galaxies of interest for the PTA signal have $M_\star \gtrsim 10^{11}\,M_\odot$, at the heavy end of the distribution where uncertainties are large. We have confirmed that our results are robust against using a different stellar-to-halo mass relation, in particular that of Ref. [61], for which

$$\frac{M_{200}}{M_\star} = \frac{(M_1/M_{200})^\alpha + (M_1/M_{200})^\beta}{10^{\epsilon+\gamma\exp[-(x/\delta)^2/2]}}, \tag{A2}$$

where $\log M_1$, α, β, γ are linear functions of z and $x = \log_{10}(M_{200}/M_1)$. As can be seen in Fig. 6, the latter relation predicts significantly smaller DM haloes at the heavy end of the spectrum at low redshift.

To relate the SMBH mass to the stellar mass of the galaxy, we use the black-hole-to-bulge mass relation from Ref. [40],

$$\log_{10}\left(\frac{M_\bullet}{M_\odot}\right) = 8.7 + 1.1\log_{10}\left(\frac{M_{\rm bulge}}{10^{11}M_\odot}\right). \tag{A3}$$

This phenomenological fit has a normally distributed scatter of ~ 0.3 dex, which we do not try to simulate here. The bulge mass is related to the stellar mass by [58]

$$M_{\rm bulge} = f_{\star,\rm bulge} M_\star, \tag{A4}$$

where $f_{\star,\rm bulge} = 0.615 + df_\star$, with

$$df_\star = \begin{cases} 0, & \text{if } M_\star \le 10^{10} M_\odot \\ \frac{\sqrt{6.9}\exp\left(\frac{-3.45}{\log_{10}M_\star - 10}\right)}{(\log_{10}M_\star - 10)^{1.5}}, & \text{if } M_\star > 10^{10} M_\odot \end{cases} \tag{A5}$$

The small correction $df_\star$ peaks around $M_* \sim 10^{12}\,M_\odot$, corresponding to $M_\bullet \sim$ few $\times\, 10^9\,M_\odot$.

By numerically inverting the above relations, one can find M_{200} for a given $M_\bullet$ and redshift z (see Fig. 6). This provides one input for fixing the outer part of the DM halo density profile, which we take to be NFW [38],

$$\rho_{NFW}(r) = \frac{\rho_s}{(r/r_s)\,(1+r/r_s)^2}. \tag{A6}$$

The virial radius R_{200} is defined to be that which contains the mass $M_{200} = 4\pi\int dr\, r^2\rho(r)$, which corresponds to 200 times the mass given by the volume $4\pi R_{200}^3/3$ times the critical density $\rho_c(z)$. The NFW parameters $\rho_s(z)$ and $r_s(z)$ are constrained by the concentration parameter $c_{200}(z) = R_{200}/r_s$, which evolves with z as given in Ref. [41]:

$$c_{200}(z) = \frac{C_c(z)}{(M_{200}/M_{\rm ref})^{\gamma_c(z)}}\left[1 + \left(\frac{M_{200}}{M_c(z)}\right)^{0.4}\right], \tag{A7}$$

with $M_{\rm ref} = 10^{12}h^{-1}M_\odot$, and the functions C_c, γ_c and M_c are tabulated in Ref. [41]. These relations numerically determine the NFW parameters ρ_s and r_s from $M_\bullet$ and z.

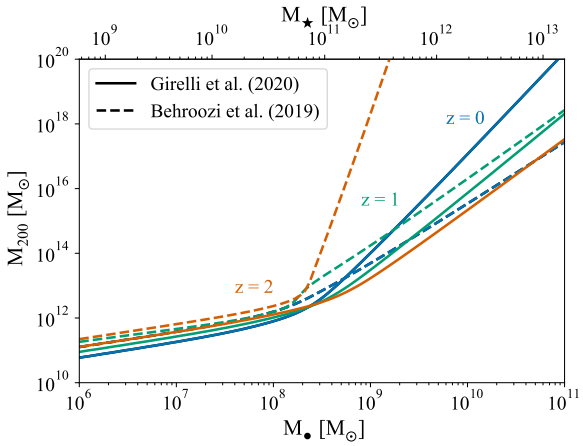

Figure 6. The halo-to-black hole mass relation $M_{200} - M_\bullet$ for several values of redshift z. The solid lines use the stellar-to-halo mass relation from [39], while the dashed lines use the relation from [61]. The upper x-axis shows the resulting stellar mass $M_\star$ from combining equations A3 and A4.

Appendix B: SIDM core profile and merger time

To construct the isothermal core region of the SIDM halo profile, one must solve the Poisson equation

$$v_0^2\nabla^2\ln\rho = -4\pi G\rho \tag{B1}$$

between $r = 0$ and $r = r_1$ with boundary conditions $\rho'(0) = 0$ and $\rho(r_1) \equiv \rho_c = \rho_{NFW}(r_1)$, subject to the constraint that the mass enclosed within r_1 is conserved relative to its value in the original NFW profile. Using the dimensionless variables $w = r/r_1$ and $\Lambda = \ln(\rho/\rho_c)$, Eq. (B1) becomes

$$\Lambda'' + \frac{2}{w}\Lambda' = -Ce^\Lambda, \tag{B2}$$

where $C = 4\pi G\rho_c r_1^2/v_0^2$. The boundary conditions become $\Lambda'(0) = 0$ and $\Lambda(1) = 0$. For a given value of C, this can be solved by shooting from $w = 0$: take $\Lambda'(0) = 0$ and $\Lambda(0) = \Lambda_0$, and vary Λ_0 until $\Lambda(1) = 0$. Next one must satisfy the mass conservation condition:

$$\int_0^1 w^2 e^\Lambda dw = \int_0^1 \frac{w(1+y)^2}{(1+wy)^2}\,dw, \tag{B3}$$

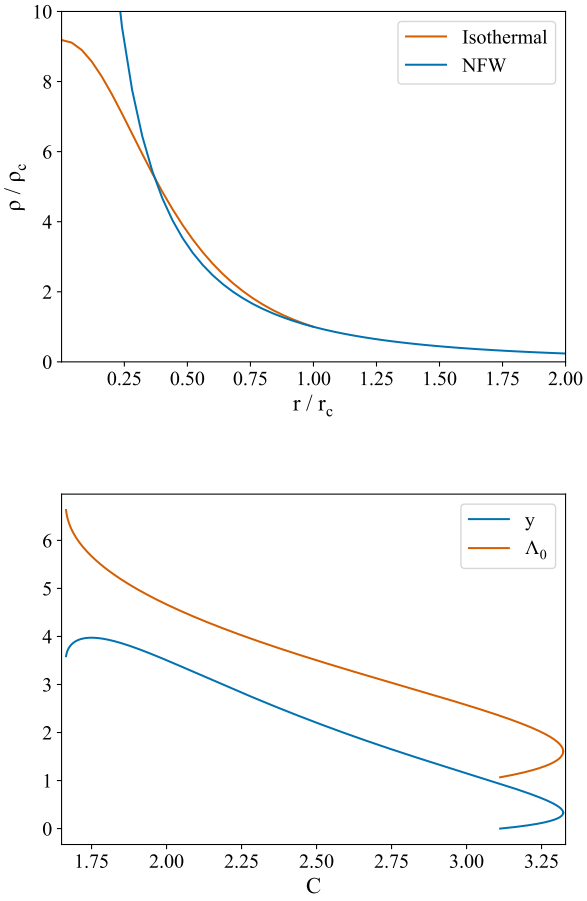

Figure 7. Top: example of cored density profile (orange) versus original NFW profile (blue), for $C = 3.17$ and $y = 0.815$. Bottom: Λ_0 (orange) and $y = r_1/r_s$ (blue) as a function of C for cored halos satisfying the Poisson equation.

where $y = r_1/r_s$ (r_s is the scale radius of the NFW profile.) This requires finding the right value of C, which amounts to determining v_0 since ρ_c and r_1 are fixed by Eq. (2). An example of the resulting profile is shown in Fig. 7 (top).

A subtlety in this procedure is that there can be more than one possible value of Λ_0 satisfying the boundary conditions for a given value of y. For example, with $y = 0.815$ there are two possible solutions, and it is necessary to choose the larger one in order to satisfy the mass conservation requirement. The required values are $\Lambda_0 = 2.22$ and $C = 3.17$. This gives rise to a small discontinuity in ρ' at $w = 1$, from -1.9 to -2.3, which is hardly noticeable in Fig. 7 (top panel).

In practice, instead of shooting, it is numerically efficient to fix C and solve for $y(C)$, and $\Lambda_0(C)$, and then numerically invert the relations. By doing so, one can avoid having to re-solve the Poisson equation for each different set of NFW halo parameters and self-interaction cross section. The resulting functions are shown in Fig. 7 (bottom). Then, a given $M_\bullet$ determines the NFW parameters, while σ_0/m which determines r_1 hence y, and the velocity dispersion in the isothermal core is given by

$$v_0^2 = 4\pi \frac{y}{(1+y)^2 C(y)} G\rho_s r_s^2 \,. \tag{B4}$$

For the massive mediator model with velocity dependent cross section

$$\sigma(v) = \frac{\sigma_0}{1+(v/v_t)^4} \cong \sigma_0 \begin{cases} 1, & v < v_t \\ (v_t/v)^4, & v > v_t \end{cases} , \tag{B5}$$

one must verify that the solution within the core is self-consistent. We first compute v_0 from Eq. (B4) assuming $a = 0$ (i.e., $\sigma = \sigma_0$) in Eq. (2), which determines r_1. If $v_0 > v_t$, then r_1 and the other core parameters must be recomputed for $a = 4$ (i.e., $\sigma = \sigma_0 (v_t/v)^4$).

As stated in the main text, we assume the velocity dispersion of the SIDM particles to be constant in the core and to vary as

$$\frac{v(r)}{v_0} = \frac{7}{11} + \frac{4}{11}\left(\frac{r_{\rm sp}}{r}\right)^{1/2} \tag{B6}$$

within the spike. This slightly deviates from the analytic approximation proposed in [50] to ensure continuity of the velocity dispersion profile at $r = r_{\rm sp}$. Thus, if $v_0 < v_t$, the $a = 0$ to $a = 4$ transition occurs in the spike, at a transition radius r_t such that $v(r_t) = v_t$ in Eq. (B6).

To compute $t_{\rm df}$ in the dark photon model, three regimes must be considered. If the transition occurs at $r_t < r_{\rm gw}$, the entire spike is in the $a = 0$ region, and the equation of motion for the dimensionless separation $x = R/(2r_{\rm sp})$ is $dx/d\tau = -Bx^{7/4}$ with $B = 192\pi N_i \rho_{\rm sp} r_{\rm sp}^3/M_1$ for the case of $q = 1$ (see above Eq. (5)). The solution is

$$t_{\rm df} = \frac{4t_{\rm sp}}{3B}\left(x_{\rm gw}^{-3/4} - x_*^{-3/4}\right). \tag{B7}$$

If the transition region occurs outside of R_*, then the $a = 4$ solution applies everywhere, and the equation of motion is $dx/d\tau = -2Bx_t x^{3/4}$. The time required to shrink from R_* to $R_{\rm gw}$ is

$$t_{\rm df} = 2\frac{t_{\rm sp}}{Bx_t}\left(x_*^{1/4} - x_{\rm gw}^{1/4}\right). \tag{B8}$$

If the transition region is inside the spike, the equation of motion is $dx/d\tau = -Bx^{7/4}$ in the outer region $x > x_t$, and by continuity $dx/d\tau = -2Bx_t x^{3/4}$ in the inner region $x < x_t$. The resulting time interval is

$$t_{\rm df} = 2\frac{t_{\rm sp}}{B}\left(\tfrac{5}{3}x_t^{-3/4} - \frac{x_{\rm gw}^{1/4}}{x_t} - \tfrac{2}{3}x_*^{-3/4}\right). \tag{B9}$$

The expression (B9) joins continuously with the previous ones if $x_t = x_{\rm gw}$ or x_*.

Appendix C: Derivation of DF power loss formula and single-merger GW spectrum

To derive the dynamical friction power loss Eq. (5), we start from the classical expression for the DF force originally derived by Chandrasekhar [20] (see, e.g. [62])

$$F_{\rm df} = -4\pi G^2 \rho_{\rm sp} M_{\rm bh}^2 \log\Lambda \frac{N(<v_{\rm bh})}{v_{\rm bh}^2}. \tag{C1}$$

Here, $M_{\rm bh}$ and $v_{\rm bh}$ are the black hole mass and velocity, $\log\Lambda$ is the Coulomb logarithm which we take to be 3 [62], and N denotes the fraction of DM particles that have velocities $< v_{\rm bh}$. As a function of the binary separation R, $v_{\rm bh1} = q(GM_1/(1+q)R)^{1/2}$ and $v_{\rm bh2} = (GM_1/(1+q)R)^{1/2}$, for the heavier ($M_{\rm bh} = M_1$) and lighter ($M_{\rm bh} = qM_1$) BH, respectively.

For CDM $N \simeq 1$, but for SIDM the velocity dispersion in the inner core is larger due to the self-interaction driven thermalization and this fraction can be smaller. Assuming a Maxwellian distribution for the SIDM particles, one has

$$N_i = \mathrm{erf}\left(\frac{u_i}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}}\, u_i\, \mathrm{e}^{-u_i^2/2}, \tag{C2}$$

$i = 1, 2$. Here, u_i is the ratio between $v_{\rm bh,i}$ and the DM velocity dispersion for the ith BH,

$$u_i = \begin{cases} \frac{11}{4}\, q^{3/2}\,(1+q)^{-3/2}, & i = 1 \\ \frac{11}{4}\,(1+q)^{-3/2}, & i = 2 \end{cases} \tag{C3}$$

which stays constant throughout the spike since both velocities scale as $r^{-1/2}$. Substituting these expressions into Eq. (C1), together with the radial coordinate of each BH from the center of mass, $r_1 = qR/(1+q)$ for the heavier BH and $r_2 = R/(1+q)$ for the lighter BH, one arrives at the expression for $P_{\rm df} = F_{\rm df}\, v_{\rm bh}$ given in Eq. (5).

The spectrum of gravitational waves, Eq. (10) follows from the power emitted in GWs, defined above Eq. (5), and the relation between frequency and orbital separation, Eq. (9), by using the chain rule:

$$\frac{dE_{\rm gw}}{df}(f) = \frac{P_{\rm gw}}{df/dt} = \frac{P_{\rm gw}}{\frac{3}{2} f \dot{x}/x} = \frac{GM_1M_2}{6fx\, r_{\rm sp}} \frac{P_{\rm gw}}{P_{\rm tot}}\,. \tag{C4}$$

To obtain the last equality, we use $\dot{E} = -(GM_1M_2/4r_{\rm sp})(\dot{x}/x^2) = -P_{\rm tot}$, the total rate of energy loss from the circular orbit. We assume that GW emission and dynamical friction are the only significant sources of energy loss, so that $P_{\rm tot} = P_{\rm gw} + P_{\rm df}$. Alternatively, one can derive Eq. (10) from the standard expression for the GW spectrum from a merging binary [56], as in Eq. (16) of Ref. [34], by averaging over the inclination angle of the orbit; we have confirmed that the two methods lead to the same result.

Appendix D: Parametrization of the SMBH binary population

The density of SMBH mergers as a function of redshift is related to the differential galaxy merger rate,

$$\frac{\mathrm{d}^3 n}{\mathrm{d}z\,\mathrm{d}M\,\mathrm{d}q} = \frac{\mathrm{d}^3 n_{\rm g}}{\mathrm{d}z\,\mathrm{d}M_\star\,\mathrm{d}q_\star} \frac{\mathrm{d}M_\star}{\mathrm{d}M} \frac{\mathrm{d}q_\star}{\mathrm{d}q}\,, \tag{D1}$$

where $M_\star$ is the stellar mass of the primary merging galaxy and $q_\star$ is the mass ratio of the two galaxies involved. We parametrize the differential galactic merger rate as [10, 58]

$$\frac{\mathrm{d}^3 n_{\rm g}}{\mathrm{d}z\,\mathrm{d}M_\star\,\mathrm{d}q_\star} = \frac{\Psi(M_\star, z')}{M_\star} \frac{P(M_\star, q_\star, z')}{T_{\text{g-g}}(M_\star, q_\star, z')} \frac{\mathrm{d}t}{\mathrm{d}z'}, \tag{D2}$$

in terms of the galaxy stellar-mass function Ψ, the galaxy pair fraction P, and the galaxy merger time $T_{\text{g-g}}$. Because the galaxy merger can take a significant time, the quantities on the right-hand side must be evaluated at the advanced redshift z', such that $t(z) - t(z') = T_{\text{g-g}}(z')$. We use

$$\frac{\mathrm{d}t}{\mathrm{d}z} = \frac{1}{(1+z)H(z)}\,, \tag{D3}$$

where

$$H(z) = H_0 \left[\Omega_\Lambda + (1+z)^3 \Omega_m\right]^{1/2} \tag{D4}$$

and $H_0 = 67.4\,\mathrm{km\,s^{-1}Mpc^{-1}}$, $\Omega_m = 0.315$, and $\Omega_\Lambda = 0.685$ [63]. This, together with $t(z=0) = 13.79$ Gyr [63], allows to calculate $z(t)$. The galactic stellar mass function follows

$$\Psi(M_\star, z) = \Psi_0 \left(\frac{M_\star}{M_\Psi}\right)^{\alpha_\Psi} \exp\left(-\frac{M_\star}{M_\Psi}\right), \tag{D5}$$

with

$$\log_{10}\left(\frac{\Psi_0}{\mathrm{Mpc}^{-3}}\right) = \psi_0 + \psi_z \cdot z, \tag{D6}$$

$$\log_{10}\left(\frac{M_\Psi}{M_\odot}\right) = m_{\psi 0} + m_{\psi z} \cdot z, \tag{D7}$$

$$\alpha_\Psi = 1 + \alpha_{\psi 0} + \alpha_{\psi z} \cdot z. \tag{D8}$$

As detailed in Table I, all of this parameters are kept fixed except for ψ_0, which is varied to adjust the total normalization of the differential merger rate. The galaxy pair fraction is described by

$$P(M_\star, q_\star, z) = P_0\,(1+z)^{\beta_{p0}}, \tag{D9}$$

while the galaxy merger time is given by

$$T_{\text{g-g}}(M_\star, q_\star, z) = T_0\,(1+z)^{\beta_{t0}}\, q_\star^{\gamma_{t0}}. \tag{D10}$$

All the parameters in these last two functions are fixed to the values shown in Table I.

Parameter	Value	Parameter	Value
ψ_0	Free	P_0	0.033
ψ_z	-0.6	β_{p0}	1
$m_{\psi 0}$	11.5	T_0	0.5 Gyr
$m_{\psi z}$	0.11	β_{t0}	-0.5
$\alpha_{\psi 0}$	-1.21	γ_{t0}	-1
$\alpha_{\psi z}$	-0.03		

Table I. List of parameters describing the SMBH merger population along with the values used in our numerical evaluations, based on the fiducial analysis of [10].

a	Viable cross section $[\mathrm{cm}^2/\mathrm{g}]$	$\chi^2_{\rm min}$
0	None	–
1	None	–
2	$4.5 \lesssim \sigma_0/m \cdot \left(\frac{t_{\rm age}}{1\,\mathrm{Gyr}}\right) \lesssim 6$	19.2
3	$20 \lesssim \sigma_0/m \cdot \left(\frac{t_{\rm age}}{1\,\mathrm{Gyr}}\right) \lesssim 90$	18.8
4	$50 \lesssim \sigma_0/m \cdot \left(\frac{t_{\rm age}}{1\,\mathrm{Gyr}}\right) \lesssim 1600$	17.6

Table II. Viable values of the dark matter self-interaction cross section with velocity-dependence parametrized by Eq. (3). The lower limit ensures that the isothermal core is large enough to absorb the frictional energy, while the upper limit comes from requiring $\tau_{\rm df} < 1$ Gyr. The last column shows the test statistic of the PTA h_c fit at the best-fit point within each range of cross sections. For reference, the GW-only fit without DM dynamical friction has $\chi^2 = 19.3$.

$f\,[\mathrm{yr}^{-1}]$	$h_c/10^{-15}$	$f\,[\mathrm{yr}^{-1}]$	$h_c/10^{-15}$
NANOGrav [5]		PPTA [6]	
0.062	$6.5^{+4.5}_{-2.4}$	0.055	$8.3^{+6.8}_{-3.4}$
0.12	$7.9^{+3.0}_{-1.8}$	0.11	$9.6^{+4.2}_{-3.6}$
0.19	$7.4^{+3.1}_{-2.0}$	0.17	$7.4^{+3.1}_{-1.8}$
0.25	$6.4^{+3.3}_{-1.9}$	0.22	$6.3^{+4.5}_{-2.5}$
0.31	$9.3^{+4.8}_{-4.2}$	0.28	$1.0^{+3.2}_{-0.8}$
EPTA [11]		0.33	$6.3^{+4.4}_{-4.0}$
0.097	$8.0^{+4.0}_{-2.7}$	0.39	$3.7^{+3.7}_{-1.5}$
0.19	$9.6^{+2.9}_{-1.9}$	0.44	$7.1^{+3.9}_{-2.6}$
0.29	$8.2^{+3.8}_{-8.2}$	0.50	$1.6^{+3.2}_{-1.0}$
0.39	$11.0^{+4.5}_{-2.9}$	0.55	$3.6^{+3.6}_{-2.09}$
0.48	$5.2^{+10.0}_{-5.5}$		
0.58	$0.6^{+4.8}_{-0.4}$		

Table III. Data points from the PTA characteristic spectrum used in our fits. We neglect correlations between different frequency bins and/or experiments.

To relate the SMBH mass to the stellar mass of the galaxy, we use the relations described in Appendix A. Note that in our analysis we do not include any scatter in Eq. (A3). We do not expect this simplification to affect our conclusions since the NANOGrav posterior distribution for the variance in the relation is peaked at zero [10].

[1] R. Hellings and G. Downs, "Upper limits on the isotropic gravitational radiation background from pulsar timing analysis," *Astrophys. J. Lett.* **265** (1983) L39–L42.

[2] K. Aggarwal *et al.*, "The NANOGrav 11-Year Data Set: Limits on Gravitational Waves from Individual Supermassive Black Hole Binaries," *Astrophys. J.* **880** (2019) 2, arXiv:1812.11585 [astro-ph.GA].

[3] B. Goncharov *et al.*, "On the Evidence for a Common-spectrum Process in the Search for the Nanohertz Gravitational-wave Background with the Parkes Pulsar Timing Array," *Astrophys. J. Lett.* **917** no. 2, (2021) L19, arXiv:2107.12112 [astro-ph.HE].

[4] **EPTA** Collaboration, S. Chen *et al.*, "Common-red-signal analysis with 24-yr high-precision timing of the European Pulsar Timing Array: inferences in the stochastic gravitational-wave background search," *Mon. Not. Roy. Astron. Soc.* **508** no. 4, (2021) 4970–4993, arXiv:2110.13184 [astro-ph.HE].

[5] **NANOGrav** Collaboration, G. Agazie *et al.*, "The

NANOGrav 15 yr Data Set: Evidence for a Gravitational-wave Background," *Astrophys. J. Lett.* **951** no. 1, (2023) L8, arXiv:2306.16213 [astro-ph.HE].

[6] D. J. Reardon *et al.*, "Search for an Isotropic Gravitational-wave Background with the Parkes Pulsar Timing Array," *Astrophys. J. Lett.* **951** no. 1, (2023) L6, arXiv:2306.16215 [astro-ph.HE].

[7] **EPTA, InPTA:** Collaboration, J. Antoniadis *et al.*, "The second data release from the European Pulsar Timing Array - III. Search for gravitational wave signals," *Astron. Astrophys.* **678** (2023) A50, arXiv:2306.16214 [astro-ph.HE].

[8] H. Xu *et al.*, "Searching for the Nano-Hertz Stochastic Gravitational Wave Background with the Chinese Pulsar Timing Array Data Release I," *Res. Astron. Astrophys.* **23** no. 7, (2023) 075024, arXiv:2306.16216 [astro-ph.HE].

[9] M. C. Begelman, R. D. Blandford, and M. J. Rees, "Massive black hole binaries in active galactic nuclei," *Nature* **287** (1980) 307–309.

[10] **NANOGrav** Collaboration, G. Agazie *et al.*, "The NANOGrav 15 yr Data Set: Constraints on Supermassive Black Hole Binaries from the Gravitational-wave Background," *Astrophys. J. Lett.* **952** no. 2, (2023) L37, arXiv:2306.16220 [astro-ph.HE].

[11] **EPTA** Collaboration, J. Antoniadis *et al.*, "The second data release from the European Pulsar Timing Array: V. Implications for massive black holes, dark matter and the early Universe," arXiv:2306.16227 [astro-ph.CO].

[12] J. Ellis, M. Fairbairn, G. Hütsi, J. Raidal, J. Urrutia, V. Vaskonen, and H. Veermäe, "Gravitational waves from supermassive black hole binaries in light of the NANOGrav 15-year data," *Phys. Rev. D* **109** no. 2, (2024) L021302, arXiv:2306.17021 [astro-ph.CO].

[13] M. Milosavljevic and D. Merritt, "Long term evolution of massive black hole binaries," *Astrophys. J.* **596** (2003) 860, arXiv:astro-ph/0212459.

[14] F. M. Khan, K. Holley-Bockelmann, P. Berczik, and A. Just, "Supermassive Black Hole Binary Evolution in Axisymmetric Galaxies: The final parsec problem is not a problem," *Astrophys. J.* **773** (2013) 100, arXiv:1302.1871 [astro-ph.GA].

[15] E. Vasiliev, F. Antonini, and D. Merritt, "The final-parsec problem in nonspherical galaxies revisited," *Astrophys. J.* **785** (2014) 163, arXiv:1311.1167 [astro-ph.GA].

[16] B. Kocsis and A. Sesana, "Gas driven massive black hole binaries: signatures in the nHz gravitational wave background," *Mon. Not. Roy. Astron. Soc.* **411** (2011) 1467, arXiv:1002.0584 [astro-ph.CO].

[17] F. G. Goicovic, A. Sesana, J. Cuadra, and F. Stasyszyn, "Infalling clouds on to supermassive black hole binaries – II. Binary evolution and the final parsec problem," *Mon. Not. Roy. Astron. Soc.* **472** no. 1, (2017) 514–531, arXiv:1602.01966 [astro-ph.HE].

[18] F. G. Goicovic, C. Maureira-Fredes, A. Sesana, P. Amaro-Seoane, and J. Cuadra, "Accretion of clumpy cold gas onto massive black hole binaries: a possible fast route to binary coalescence," *Mon. Not. Roy. Astron. Soc.* **479** no. 3, (2018) 3438–3455, arXiv:1801.04937 [astro-ph.HE].

[19] L. Z. Kelley, L. Blecha, and L. Hernquist, "Massive Black Hole Binary Mergers in Dynamical Galactic Environments," *Mon. Not. Roy. Astron. Soc.* **464** no. 3, (2017) 3131–3157, arXiv:1606.01900 [astro-ph.HE].

[20] S. Chandrasekhar, "Dynamical Friction. I. General Considerations: the Coefficient of Dynamical Friction," *Astrophys. J.* **97** (1943) 255.

[21] R. Vicente and V. Cardoso, "Dynamical friction of black holes in ultralight dark matter," *Phys. Rev. D* **105** no. 8, (2022) 083008, arXiv:2201.08854 [gr-qc].

[22] B. C. Bromley, P. Sandick, and B. Shams Es Haghi, "Supermassive Black Hole Binaries in Ultralight Dark Matter," arXiv:2311.18013 [astro-ph.GA].

[23] L. Berezhiani, G. Cintia, V. De Luca, and J. Khoury, "Dynamical friction in dark matter superfluids: The evolution of black hole binaries," arXiv:2311.07672 [astro-ph.CO].

[24] A. Boudon, P. Brax, and P. Valageas, "Supersonic friction of a black hole traversing a self-interacting scalar dark matter cloud," *Phys. Rev. D* **108** no. 10, (2023) 103517, arXiv:2307.15391 [astro-ph.CO].

[25] M. H. Chan and C. M. Lee, "Indirect Evidence for Dark Matter Density Spikes around Stellar-mass Black Holes," *Astrophys. J. Lett.* **943** no. 2, (2023) L11, arXiv:2212.05664 [astro-ph.HE].

[26] S. Mitra, S. Chakraborty, R. Vicente, and J. C. Feng, "Probing the quantum nature of black holes with ultra-light boson environments," arXiv:2312.06783 [gr-qc].

[27] J. C. Aurrekoetxea, K. Clough, J. Bamber, and P. G. Ferreira, "The effect of wave dark matter on equal mass black hole mergers," arXiv:2311.18156 [gr-qc].

[28] K. Eda, Y. Itoh, S. Kuroyanagi, and J. Silk, "Gravitational waves as a probe of dark matter minispikes," *Phys. Rev. D* **91** no. 4, (2015) 044045, arXiv:1408.3534 [gr-qc].

[29] M. Tang and J. Wang, "The eccentricity enhancement effect of intermediate-mass-ratio-inspirals: dark matter and black hole mass," *Chin. Phys. C* **45** no. 1, (2021) 015110, arXiv:2005.11933 [gr-qc].

[30] B. J. Kavanagh, D. A. Nichols, G. Bertone, and D. Gaggero, "Detecting dark matter around black holes with gravitational waves: Effects of dark-matter dynamics on the gravitational waveform," *Phys. Rev. D* **102** no. 8, (2020) 083006, arXiv:2002.12811 [gr-qc].

[31] A. Coogan, G. Bertone, D. Gaggero, B. J. Kavanagh, and D. A. Nichols, "Measuring the dark matter environments of black hole binaries with gravitational waves," *Phys. Rev. D* **105** no. 4, (2022) 043009, arXiv:2108.04154 [gr-qc].

[32] K. Kadota, J. H. Kim, P. Ko, and X.-Y. Yang, "Gravitational Wave Probes on Self-Interacting Dark Matter Surrounding an Intermediate Mass Black Hole," arXiv:2306.10828 [hep-ph].

[33] P. Gondolo and J. Silk, "Dark matter annihilation at the galactic center," *Phys. Rev. Lett.* **83** (1999) 1719–1722, arXiv:astro-ph/9906391.

[34] Z.-Q. Shen, G.-W. Yuan, Y.-Y. Wang, and Y.-Z. Wang, "Dark Matter Spike surrounding Supermassive Black Holes Binary and the nanohertz Stochastic Gravitational Wave Background," arXiv:2306.17143 [astro-ph.HE].

[35] L. Hu, R.-G. Cai, and S.-J. Wang, "Distinctive GWBs from eccentric inspiraling SMBH binaries with a DM spike," arXiv:2312.14041 [gr-qc].

[36] S. Tulin and H.-B. Yu, "Dark Matter Self-interactions and Small Scale Structure," *Phys. Rept.* **730** (2018) 1–57, arXiv:1705.02358 [hep-ph].

[37] M. Kaplinghat, S. Tulin, and H.-B. Yu, "Dark Matter Halos as Particle Colliders: Unified Solution to Small-Scale Structure Puzzles from Dwarfs to Clusters," *Phys. Rev. Lett.* **116** no. 4, (2016) 041302, arXiv:1508.03339 [astro-ph.CO].

[38] J. F. Navarro, C. S. Frenk, and S. D. M. White, "The Structure of cold dark matter halos," *Astrophys. J.* **462** (1996) 563–575, arXiv:astro-ph/9508025.

[39] G. Girelli, L. Pozzetti, M. Bolzonella, C. Giocoli, F. Marulli, and M. Baldi, "The stellar-to-halo mass relation over the past 12 Gyr: I. Standard ΛCDM model," *Astron. Astrophys.* **634** (2020) A135, arXiv:2001.02230 [astro-ph.CO].

[40] J. Kormendy and L. C. Ho, "Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies," *Ann. Rev. Astron. Astrophys.* **51** (2013) 511–653, arXiv:1304.7762 [astro-ph.CO].

[41] A. Klypin, G. Yepes, S. Gottlober, F. Prada, and S. Hess, "MultiDark simulations: the story of dark matter halo concentrations and density profiles," *Mon. Not. Roy. Astron. Soc.* **457** no. 4, (2016) 4340–4359, arXiv:1411.4001 [astro-ph.CO].

[42] D. Merritt, "Single and binary black holes and their influence on nuclear structure," in *Carnegie Observatories Centennial Symposium. 1. Coevolution of Black Holes and Galaxies.* 1, 2003. arXiv:astro-ph/0301257.

[43] D. Merritt, "Evolution of the dark matter distribution at the galactic center," *Phys. Rev. Lett.* **92** (2004) 201304, arXiv:astro-ph/0311594.

[44] P. Ullio, H. Zhao, and M. Kamionkowski, "A Dark matter spike at the galactic center?," *Phys. Rev. D* **64** (2001) 043504, arXiv:astro-ph/0101481.

[45] D. Merritt, M. Milosavljevic, L. Verde, and R. Jimenez, "Dark matter spikes and annihilation radiation from the galactic center," *Phys. Rev. Lett.* **88** (2002) 191301, arXiv:astro-ph/0201376.

[46] O. Y. Gnedin and J. R. Primack, "Dark Matter Profile in the Galactic Center," *Phys. Rev. Lett.* **93** (2004) 061302, arXiv:astro-ph/0308385.

[47] M. Milosavljevic and D. Merritt, "Formation of galactic nuclei," *Astrophys. J.* **563** (2001) 34–62, arXiv:astro-ph/0103350.

[48] E. Vasiliev, "Dark matter annihilation near a black hole: Plateau vs. weak cusp," *Phys. Rev. D* **76** (2007) 103532, arXiv:0707.3334 [astro-ph].

[49] S. L. Shapiro and J. Shelton, "Weak annihilation cusp inside the dark matter spike about a black hole," *Phys. Rev. D* **93** no. 12, (2016) 123510, arXiv:1606.01248 [astro-ph.HE].

[50] S. L. Shapiro and V. Paschalidis, "Self-interacting dark matter cusps around massive black holes," *Phys. Rev. D* **89** no. 2, (2014) 023506, arXiv:1402.0005 [astro-ph.CO].

[51] There can be a transition region between the NFW and the isothermal profiles where v_0 varies, which for simplicity we do not attempt to capture here.

[52] In the inner region, $\rho_{\rm sp}(r) = \rho_{\rm NFW}(r_{\rm sp})(r_{\rm sp}/r_t)^{3/4}(r_t/r)^{7/4}$.

[53] Assuming that $r_t < r_{\rm sp}$; otherwise one should set $r_t = r_1$ since the entire core and spike is governed by $a = 4$.

[54] P. C. Peters, "Gravitational Radiation and the Motion of Two Point Masses," *Phys. Rev.* **136** (1964) B1224–B1232.

[55] We thank J. Sievers for suggesting to check the energetics.

[56] E. S. Phinney, "A Practical theorem on gravitational wave backgrounds," arXiv:astro-ph/0108028.

[57] S. Chen, A. Sesana, and W. Del Pozzo, "Efficient computation of the gravitational wave spectrum emitted by eccentric massive black hole binaries in stellar environments," *Mon. Not. Roy. Astron. Soc.* **470** no. 2, (2017) 1738–1749, arXiv:1612.00455 [astro-ph.CO].

[58] S. Chen, A. Sesana, and C. J. Conselice, "Constraining astrophysical observables of Galaxy and Supermassive Black Hole Binary Mergers using Pulsar Timing Arrays," *Mon. Not. Roy. Astron. Soc.* **488** no. 1, (2019) 401–418, arXiv:1810.04184 [astro-ph.GA].

[59] Taking into account the correlations between different energy bins and/or experiments in our statistical treatment is beyond the scope of this work.

[60] C. J. Conselice, A. Wilkinson, K. Duncan, and A. Mortlock, "The Evolution of Galaxy Number Density at z ¡ 8 and Its Implications," *Astrophys. J.* **830** no. 2, (Oct., 2016) 83, arXiv:1607.03909 [astro-ph.GA].

[61] P. Behroozi, R. H. Wechsler, A. P. Hearin, and C. Conroy, "UniverseMachine: The correlation between galaxy growth and dark matter halo assembly from z = 0−10," *Mon. Not. Roy. Astron. Soc.* **488** no. 3, (2019) 3143–3194, arXiv:1806.07893.

[62] A. Gualandris and D. Merritt, "Ejection of Supermassive Black Holes from Galaxy Cores," *Astrophys. J.* **678** (2008) 780, arXiv:0708.0771 [astro-ph].

[63] **Planck** Collaboration, N. Aghanim *et al.*, "Planck 2018 results. VI. Cosmological parameters," *Astron. Astrophys.* **641** (2020) A6, arXiv:1807.06209 [astro-ph.CO]. [Erratum: Astron.Astrophys. 652, C4 (2021)].