### Supplemental Material

When performing Bayesian parameter estimation, we used the tempered sampler `eryn`. For all the analyses shown in the main text, we used a mixture of 'stretch moves' and 'adaptive Gaussian moves', with 2 temperatures, which allowed for a thorough exploration of the parameter space. We further checked convergence by inspecting the trace plots for each posterior chain. For completeness, we show in Fig. 3 the one dimensional and two dimensional marginalized posteriors of both the intrinsic and extrinsic vacuum EMRI parameters, for the injection with both noise and migration torques. We do not show the initial separation $r_0/M$ and phase $\Phi_{\phi_0}$ as they are (by definition) different for each observation duration. Each colored histogram corresponds to $T_{\rm obs} = \{0.5, 1, 1.5, 2, 2.5, 3\}$ yr, as discussed in the main text. This supplements Fig. 1 of the main text, where we only showed the posterior inconsistency for the primary mass and spin, for $T_{\rm obs} = \{0.5, 1, 1.5, 2\}$ yr.

To better understand the role of the environmental effect (disk-driven migration) on the EMRI signal, we compute the resulting orbital dephasing (relative to the vacuum case), for the same injected vacuum EMRI parameters. In the top panel of Fig. 1, we show the orbital dephasing as a function of both the observation duration and (equivalently) the orbital separation $r_0/M$. Because the disk-driven migration (in a Type-I $\alpha$-disk) considered in our work is a $-8$PN effect, Fig. 1 shows a significant orbital dephasing $\gtrsim 50$ rad for a 3 yr observation. Conversely, for a shorter observation of $\lesssim 1$ yr, the dephasing is $\lesssim 3$ rad. As discussed in the main text, the inference

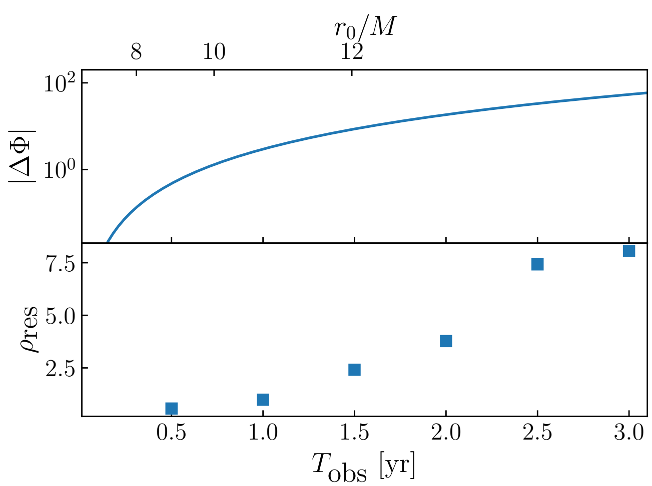


FIG. 1: Top panel shows the orbital dephasing (as a function of duration and orbital separation $r_0/M$) between a noiseless EMRI with migration torques and the corresponding vacuum case. For this injection, the bottom panel shows the residual SNR obtained from the maximum likelihood vacuum EMRI waveform. Note that the injection corresponds to the one in the main text.

inconsistencies become more significant with longer observation duration, because the environmental effects are dominant at lower frequencies.

We also report, in the bottom panel of Fig. 1, the squared residual SNR corresponding to the MLE [18, 23, 90–92], given by $\rho^2_{\rm res} \equiv -\log \mathcal{L}(d|\boldsymbol{\theta}_{\rm MLE})$, which we calculated for the noiseless injections with migration torques. As expected, for longer observations, an increasing amount of signal is lost in the analysis, due to the missing physics in the template. The signal loss is never large enough to prevent detection, but in the context of the LISA global fit, the residual SNR from many sources may accumulate and potentially cause biases and/or missed detections [88].

As a validation of our self-consistency test, we also injected a vacuum EMRI signal with noise. We considered the same source parameters $\boldsymbol{\theta}_{\rm inj}$ and observation durations as in the main text, and inferred posteriors on $\boldsymbol{\theta}$. The results are reported in Table I. We see that the mismatch criterion given by Eq. 4 in the main text is always well satisfied. Note that noise can cause a relative systematic bias $\delta\theta^\alpha > 1$ in the one-dimensional marginalized posteriors, while the mismatch criterion is still satisfied. Indeed, since the mismatch criterion takes into consideration the whole parameter space, it is a more reliable diagnostic of the presence of environmental effects than the one-dimensional posteriors.

We have checked that different noise realizations do not yield statistically different results for both the vacuum and migration injections. In Table II, we show the mismatches obtained with $T_{\rm obs} = 2.5$ yr and $T_{\rm obs} = 3.0$ yr observations against the reference $T_{\rm obs} = 0.5$ yr observation. We considered two different noise realizations for both the vacuum and migration injections, using the same source parameters as those described in the main text. We find that the mismatches obtained for given $T_{\rm obs}$ are comparable across noise realizations, and agree with Tables I of supplemental and Table I of the main text.

Additionally, we tested whether an inconsistent inference across observations can be caused by a slight mismodeling of the noise PSD, instead of a mismodeling of a physical effect. We considered two vacuum injections in which the PSD used to produce the noise and the one in the likelihood calculation differ. In the first case we assume that the acceleration component of the LISA noise curve is underestimated during the global fit. Specifically we increase the acceleration noise component by 10% in the PSD used in the noise generation compared to the one used in the data analysis pipeline. In the second case we consider a slight mismodel in the galactic binaries foreground. Using the empirical description for unresolved galactic binaries we simulate a situation in which the number of extracted GBs is overestimated. Specifically we use the empirical fits introduced in [89] and we use an observation time which is shorter by 2 months in the PSD used for noise generation compared to the one used for parameter estimation. In both cases

TABLE I: Mismatches and systematic biases relative to the reference MLE with $T_{\rm obs} = 0.5\,$yr, for a vacuum injection with noise.

| $T_{\rm obs}$ (yr) | $1 - \mathcal{M}_{\rm net, T_{obs}}$ | $\delta \log_{10} M$ | $\delta a$ |
|---|---|---|---|
| 1.0 | $5.8 \times 10^{-5}$ | 0.07 | 0.04 |
| 1.5 | $1.5 \times 10^{-4}$ | 0.59 | 0.49 |
| 2.0 | $3.2 \times 10^{-4}$ | 0.92 | 0.76 |
| 2.5 | $6.0 \times 10^{-4}$ | 1.16 | 0.87 |
| 3.0 | $4.8 \times 10^{-4}$ | 0.99 | 0.74 |

TABLE II: Mismatch for different noise realizations, for the vacuum and migration injections described in the main text. Note that the vacuum mismatches are always below the threshold discussed in the main text.

| $T_{\rm obs}$ (yr) | Vacuum | | Migration | |
|---|---|---|---|---|
| | Noise 1 | Noise 2 | Noise 1 | Noise 2 |
| 2.5 | $4.8 \times 10^{-5}$ | $8.7 \times 10^{-5}$ | $1.3 \times 10^{-3}$ | $1.3 \times 10^{-3}$ |
| 3.0 | $7.0 \times 10^{-5}$ | $5.7 \times 10^{-5}$ | $1.0 \times 10^{-3}$ | $8.8 \times 10^{-4}$ |

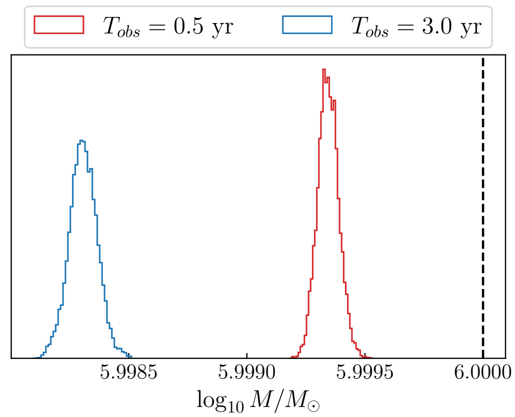


FIG. 2: Comparison of the posterior for $\log_{10} M$ obtained from an observation time of 0.5 yr and 3.0 yr. In black, we report the injection value. The injection coincides with the one in the main text for the vacuum parameters while here we have set $A = 2.0 \times 10^{-4}$ and $n = -1.0$.

no inconsistency between the posteriors has been found and the mismatch values obtained were comparable with the ones reported in table II for different noise realizations. We therefore conclude that our results are robust to different noise realizations and to noise mismodeling.

We also explored how our self-consistency test performs for injections containing a different environmental effect, or corresponding to different source parameters. For the same source parameters mentioned in the main text, which corresponds to an SNR of 100 with a 3 yr observation, we modified the environmental effect by choosing different values of $n$ (keeping the relative torque magnitude $A$ fixed). For event A, we set $n = 4$ (that contributes at -4PN relative to the GW flux), which corresponds to environmental effects such as gas accretion and center-of-mass acceleration, or modification to gravity such as a time dependent Newton's constant [62,64,87]. For events B and C, we chose $n = 5.5$ (that contributes at -5.5PN relative to the GW flux), which corresponds to non-relativistic dynamical friction in a uniform collisionless medium [23, 93, 94]. We also chose different source parameters for event C, setting $M = 5 \times 10^5\ M_\odot$, $\mu = 30\ M_\odot$, $a = 0.7$, which yields an SNR of 60 for a 3yr observation. In Tables III and I, we show the results of the self-consistency test for each event.

For events A and B, the posteriors become inconsistent with the reference MLE for $T_{\rm obs} \geq 2.5$ yr, in agreement with the accumulated dephasing relative to the vacuum template. Moreover, at fixed $T_{\rm obs}$, event B exhibits larger mismatches than event A. We attribute this to the larger dephasing in event B at given $T_{\rm obs}$, since the environmental effect enters at a more negative PN order. For event C, we find that the mismatch threshold is $2.2 \times 10^{-3}$, due to its lower SNR. We observe that event C is fully incompatible with vacuum at $T_{\rm obs} = 2.0$ yr. Also note that for a longer observation time $T_{\rm obs} = 2.5$ yr, the mismatch is comparable to the threshold, and then exceeds it at $T_{\rm obs} = 3.0$ yr. Even though the behavior of the mismatch as a function of $T_{\rm obs}$ is not strictly monotonic for event C, unlike for events A and B, the fact that it becomes comparable to and then exceeds the threshold for $T_{\rm obs} \geq 2$ yr signals an inconsistency due to using the vacuum template.

Since the main focus of this work is on the detection of unmodeled contribution coming from the astrophysical environment we considered negative PN modification to the gravitational fluxes. However, positive PN modifications to the flux can instead arise due to modified gravity [95]. We consider here a 1PN effect that would correspond to a propagation effect due to massive gravity [95], in order to contrast against the systematics of negative PN effects. To see the impact on the self-consistency test due to neglecting such a 1PN correction, we pick the fiducial values $A = 2 \times 10^{-4}$ and $n = -1$ for the injection, together with the same EMRI parameters considered in the main text. In Fig. 2, we show the posteriors for the primary mass obtained with a vacuum EMRI model, corresponding to a 0.5 yr (red histogram) and 3 yr (blue histogram) observation. We see that there is significant inconsistency between the two observations, thus indicating that the self-consistency test also works for positive PN effects. However, we also note that the inference with 0.5 yr observation is biased relative to the true injected value, unlike what we saw with negative PN order environmental effects. This is somewhat expected as the modification becomes more significant close to merger, while in the previous cases the late inspiral modification was negligible. The additional bias between the observation close to merger and the one with the full inspiral is due to the accumulated dephasing between observation durations, which is why this applies to both positive and negative PN effects.

TABLE III: Mismatches, systematic biases relative to the reference MLE at $T_{\rm obs} = 0.5\,{\rm yr}$ and dephasing from the vacuum trajectory, for different injections. Events A and B have the same source parameters as the source in Fig. 1 in the main text, but a different power law for the environmental torque. Event C has different source parameters ($M = 5 \times 10^5\ M_\odot$, $\mu = 30\ M_\odot$, $a = 0.7$, SNR $=$ 60.0) and the same torque as event B.

| Event | $T_{\rm obs}$ (yr) | $1 - \mathcal{M}_{\rm net, T_{obs}}$ | $\delta \log_{10} M/M_\odot$ | $\delta a$ | $\Delta\Phi$ |
|---|---|---|---|---|---|
| A | 2.0 | $2.5 \times 10^{-5}$ | 1.22 | 1.00 | 10.7 |
| | 2.5 | $6.5 \times 10^{-4}$ | 1.76 | 1.77 | 15.4 |
| | 3.0 | $1.3 \times 10^{-3}$ | 2.38 | 2.43 | 20.7 |
| B | 2.0 | $4.6 \times 10^{-4}$ | 1.55 | 1.53 | 12.8 |
| | 2.5 | $1.2 \times 10^{-3}$ | 2.69 | 2.73 | 19.9 |
| | 3.0 | $1.8 \times 10^{-3}$ | 3.10 | 3.15 | 28.7 |
| C | 2.0 | $3.6 \times 10^{-3}$ | 1.93 | 2.00 | 34.3 |
| | 2.5 | $2.0 \times 10^{-3}$ | 1.48 | 1.51 | 53.2 |
| | 3.0 | $2.2 \times 10^{-3}$ | 1.50 | 1.54 | 76.2 |

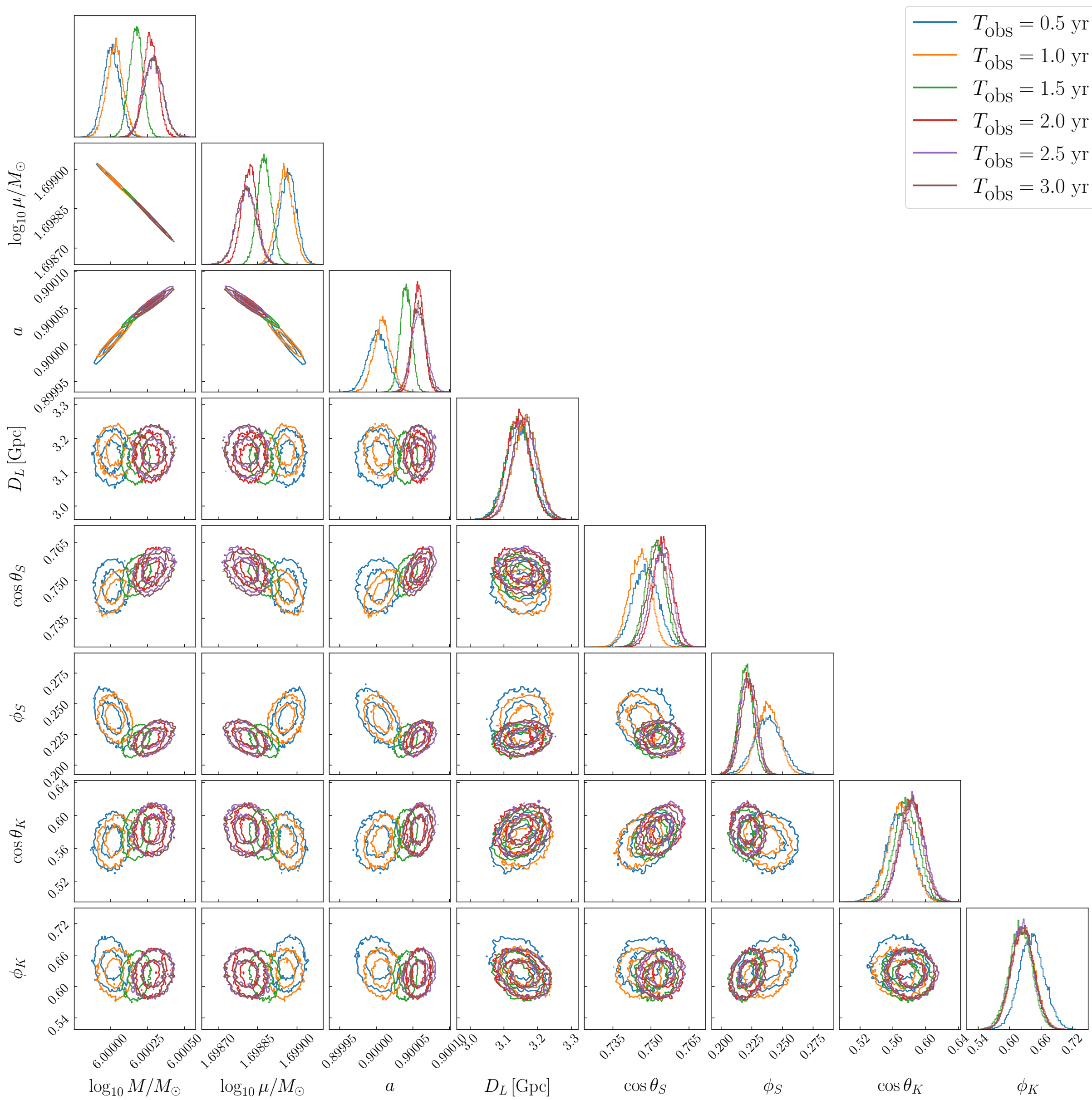


FIG. 3: Posteriors for the migration + noise injection, for all the observation times considered. We omit the posteriors of $r_0/M$ and $\Phi_{\phi_0}$, as their true values change with observation time. Note that the use of a wrong template has little effect on the extrinsic parameters posteriors, which are always consistent.