Incorporating multi-model uncertainty into gravitational-wave Bayesian inference

Charlie Hoy[1], Sarp Akçay[2], Jake Mac Uilliam[2] and Jonathan E. Thompson[3,4]

[1]Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth, PO1 3FX, UK [2]School of Mathematics & Statistics, University College Dublin, Dublin, D4, Ireland [3]Theoretical Astrophysics Group, California Institute of Technology, Pasadena, CA, 91125, U.S.A [4]Mathematical Sciences & STAG Research Centre, University of Southampton, Southampton, SO17 1BJ, UK

Inferring the properties of colliding black holes from gravitational-wave observations is subject to systematic errors arising from modelling uncertainties. Although the accuracy of each model can be calculated through comparison to theoretical expectations from general relativity, Bayesian analyses are yet to incorporate this information. As such, a mixture model is typically used where results obtained with different gravitational-wave models are combined with either equal weight, or based on their relative Bayesian evidence. In this work we present a novel method to incorporate the accuracy of multiple models in gravitational-wave Bayesian analyses. By analysing simulated gravitational-wave signals in zero-noise, we show that our technique uses 30% less computational resources, and more faithfully recovers the true parameters than existing techniques. We envisage that this method will become an essential tool within ground-based gravitational-wave astronomy.

1 Model systematics in gravitational-wave astronomy

Our ability to infer the properties of colliding black holes from an observed gravitational-wave (GW) signal is dependent on our chosen model [1]. Models that poorly describe general relativity will not only yield biased results for individual sources [2–7], but also incorrect inferences for the properties of the underlying astrophysical population – for example, the mass and spin distributions of black holes in the Universe [8–10]. Unbiased results will only be obtained with models that are perfect descriptions of general relativity (assuming a known understanding of the noise in the GW detectors).

Unfortunately, directly computing GW signals from general relativity is a computationally expensive task; numerical relativity simulations, where Einstein's equations of general relativity are solved on high-performance computing clusters, require millions of CPU hours to perform [11]. For this reason only several thousand simulations are currently available [11–18]. As a result, GW models rely on analytical or semi-analytical prescriptions that are calibrated to the numerical relativity simulations [3,19–39], or are based on surrogate modelling techniques [40,41]. However, each modelling approach will incur some degree of approximation errors.

The accuracy of a GW model is typically measured by the mismatch [42] between the model and a fiducial waveform, often a numerical relativity simulation. The mismatch varies between 0, signifying that the model and the true waveform are identical (up to an overall amplitude rescaling), and 1, meaning that the two are completely orthogonal. It is well known that certain models are more faithful to general relativity than others in different regions of parameter space [5].

The standard approach to account for modelling errors when inferring the properties of binary black holes is to construct a mixture model, where results from numerous analyses are combined; a Bayesian analysis is performed for each GW model and the results are either mixed together with equal weights [43] or according to their relative Bayesian evidence [44]. An alternative technique involves sampling over a set of GW models in a single *joint* Bayesian analysis [45,46]. Although widely used, these methods do not account for the known accuracy of the GW model. Other approaches have suggested quantifying the uncertainty in a GW model, and marginalizing over this error in Bayesian analyses [47–51]. However, these methods have either not been demonstrated in practice, or are only suitable for a single model.

In this work we present the first approach to incorporate the accuracy of multiple models into a single GW Bayesian analysis (see Methods). This technique accounts for modelling errors by prioritising the most accurate GW model in each region of the parameter space, thereby mitigating against biased results from using models that are unfaithful to general relativity. For GW signals likely observed by the LIGO–Virgo–KAGRA GW detectors, we demonstrate that current techniques inflate uncertainties and have the potential to produce biased parameter estimates. On the other hand, we show that the method presented here either outperforms current techniques, or in the worst case, gives comparable results.

2 Gravitational-wave Bayesian inference

We first apply our approach to analyse a theoretical GW signal expected from general relativity, specifically, the `SXS:BBH:0926` [13,52] numerical relativity simulation produced by the Simulating eXtreme Spacetimes (SXS) collaboration (https://www.black-holes.org). We assume a total mass of $100\,M_{\odot}$ and we inject this signal into zero noise at a signal-to-noise ratio of 40. The `SXS:BBH:0926` simulation has mass ratio $1:2$ and large spin magnitudes perpendicular to the orbital angular momentum (within the orbital plane of the binary) for both black holes of ~ 0.8. For this system, the general relativistic phenomenon of spin-induced orbital precession [53] is significant, and contributes a signal-to-noise ratio [54,55] ~ 9 to the total power of the signal. This simulation was chosen since the majority of GW models obtain biased results, and disagree on the inferred binary parameters [5,6] A system with significant spin-induced orbital precession has been predicted to be observed once in every 50 GW observations made by the LIGO [56], Virgo [57] and KAGRA [58] GW observatories based on current black hole population estimates [59].

We use three of the most accurate and cutting edge models currently available for describing the theoretical GW signals produced by colliding black holes: IMRPHENOMXPHM [34] (with the updated precession formalisation [60]), IMRPHENOMTPHM [37] and SEOBNRV5PHM [2]. All models include the general relativistic phenomenon of spin-induced orbital precession [53] and higher order multipole moments [61]. We analyse 8 seconds of data, and only consider frequencies between [20, 2048] Hz. We generate the numerical relativity simulation from $\approx 10\,\mathrm{Hz}$ to ensure that most higher multipole content is generated prior to our analysis window. Our analysis restricts attention to a two-detector network of LIGO-Hanford and LIGO-Livingston [56], and we assume a theoretical power spectral density for Advanced LIGO's

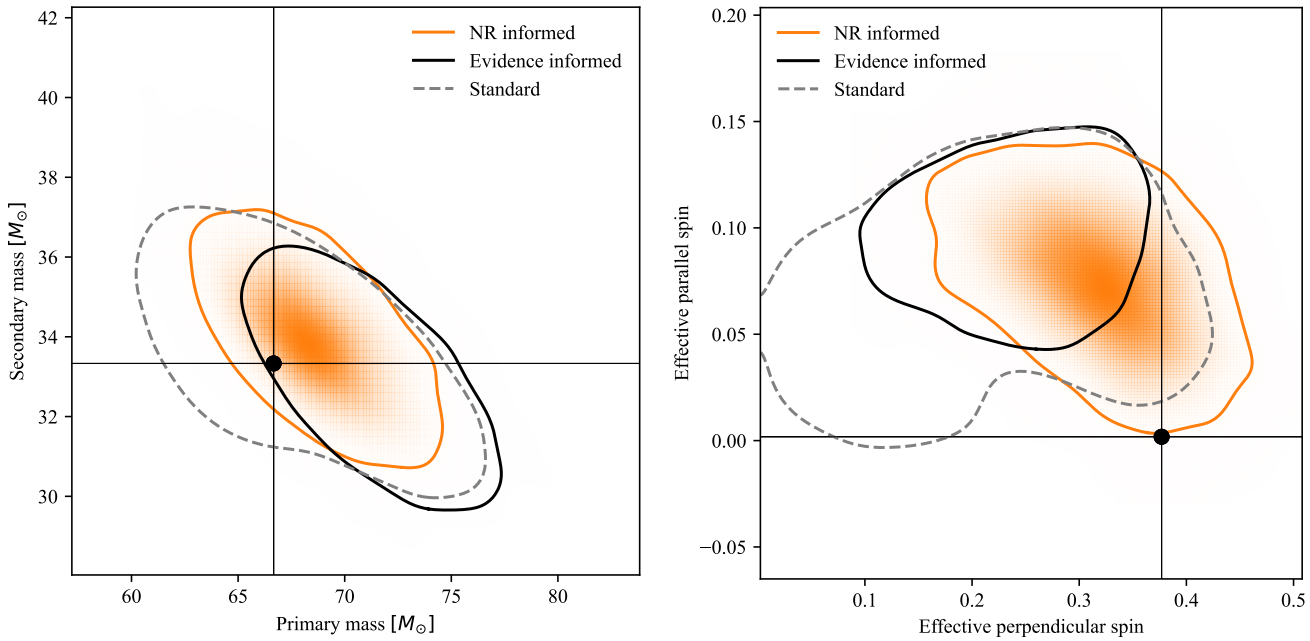

Figure 1 | Two-dimensional posterior probabilities obtained in our analysis of the `SXS:BBH:0926` numerical relativity simulation. The left panel shows the measurement of the primary and secondary mass of the binary, and the right panel shows the inferred effective parallel and perpendicular spin components (as defined in Methods, see Equations 11, 12). An effective perpendicular spin of 0 means that the spin vector lies perpendicular to the plane of the binary. Contours represent 90% credible intervals and the black cross hairs indicate the true value.

design sensitivity [62]. We use the most agnostic priors available for all parameters, identical to those used in all detections made by the LIGO–Virgo–KAGRA collaborations [43]: flat in the component masses and spin magnitude and the cosine of the spin tilt angle. We perform Bayesian inference with the `Dynesty` Nested sampling software [63] via `Bilby` [64], as has been done in all LIGO–Virgo–KAGRA analyses since the third GW catalog [65].

In Figure 1 we compare the results obtained with our method to two widely adopted techniques. The contours labelled *NR informed* utilise the method presented here, *Evidence weighted* combines separate inference analyses obtained with different GW models according to their relative Bayesian evidence [44], and *Standard* combines the results of separate inference analysis with equal weight. Standard is the currently adopted method by the LIGO–Virgo–KAGRA collaboration [43]. When considering the inferred primary and secondary masses of the binary, we see that all three techniques capture the true value within the two-dimensional marginalized 90% credible interval. Both the NR Informed and Standard methods more accurately infer the true values of the binary, with the injected values lying within the 50% credible interval. Given that the Standard method equally combines analyses from the individual GW models, the uncertainty is inflated in comparison to the method presented here and to the Evidence weighted result.

We now turn our attention to the inferred spin on the binary. Since the individual spin components are difficult to measure for binary black holes at present-day detector sensitivities [66], we consider the measurement of effective spin parameters that describe the dominant spin effects of the observed GW signal [67,68]. In Figure 1 we show the measurement of the effective spin parallel and perpendicular to the orbital angular momentum, as defined below in Methods. We see significant differences between the obtained posterior distributions: the NR Informed approach introduced in this work is the least biased as it encompasses the true value within the two-dimensional marginalized 90% credible interval. Although the Evidence informed result has been described as the optimal method in previous work [44], for this simulated signal it produces an inaccurate result. This is because IMRPHENOMTPHM has the largest Bayesian evidence despite not being the most accurate model; it has been shown previously that less accurate models can give large Bayesian evidences due to mismodelling [46]. Our analysis, on the other hand, predominantly uses SEOBNRV5PHM; we use SEOBNRV5PHM 90% of the time, IMRPHENOMTPHM 8% of the time, and IMRPHENOMXPHM 2% of the time. This demonstrates one of the limitations of our method: although we preferentially use the most accurate model in each of the parameter space, there is no guarantee that this model is accurate enough to avoid biases in the inferred parameter estimates [69,70]. However, we highlight that it is the most accurate method of those currently used, and can evolve to include more accurate models when they are developed.

In Figure 2 we present the ratio of mismatches obtained with the different GW models used in this work. We see that SEOBNRV5PHM has the smallest mismatch in the region of parameter space containing the simulation parameters, and is therefore the most faithful to general relativity in this region. SEOBNRV5PHM obtains mismatches $\sim 3\times$ and $\sim 1.8\times$ smaller than IMRPHENOMXPHM and IMRPHENOMTPHM, respectively.

Since our method chooses the GW model based on its accuracy to numerical relativity in each region of the parameter space, rather than combining finalized results from each GW model individually, we also see a significant decrease in computational cost. Our analysis uses 30% less computational resources than the Standard and Evidence weighting analyses. The analysis completed in 230 CPU days, compared with 35 CPU days, 118 CPU days and 181 CPU days for the individual IMRPHENOMXPHM, IMRPHENOMTPHM and SEOBNRV5PHM analyses, respectively. In the worst case scenario, we expect our method to use the same computational resources as the Standard and Evidence weighting analyses.

Our technique is free to use any combination of GW models. If SEOBNRV5PHM were removed from this analysis, we find consistent results between our method and the Evidence informed result,

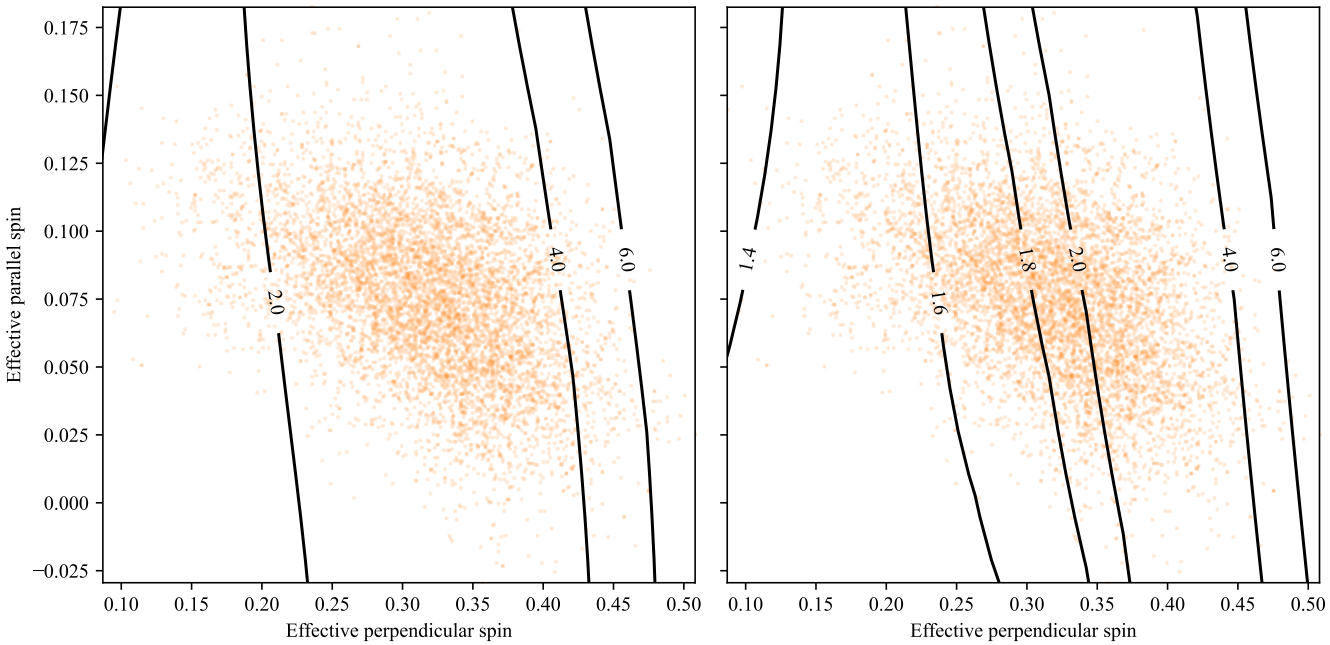

Figure 2 | Contour plot showing the ratio of mismatches to numerical relativity simulations for different effective parallel and perpendicular spin components when averaging over different mass configurations. The left panel compares IMRPhenomXPHM and SEOBNRv5PHM and the right panel compares IMRPhenomTPHM and SEOBNRv5PHM. In orange we show samples obtained from our analysis of the SXS:BBH:0926 numerical relativity simulation. In both cases, a ratio of mismatches greater than unity implies that SEOBNRv5PHM is more faithful to general relativity.

with overlapping two-dimensional marginalized 90% confidence intervals. The reason is because IMRPhenomTPHM now has the largest Bayesian evidence and is the more accurate of the two remaining GW models considered in the region of parameter space.

An alternative approach to the method we propose is to perform a single analysis with the model which is, on average, the most accurate in the parameter space on interest. [JT: is this really an "alternative approach"? This is just using one model to do the analysis.] The issue with this technique is that the mismatch varies considerably across different regions of the parameter space, particularly for the spins which are often not well measured. For instance, when averaging across the parameter space consistent with SXS:BBH:0926, SEOBNRv5PHM is the most accurate model. However, for effective parallel spins > 0 and perpendicular spins < 0.05, we find that IMRPhenomTPHM is more accurate than SEOBNRv5PHM, and IMRPhenomXPHM is of comparable accuracy to SEOBNRv5PHM. By simply averaging the mismatch across the parameter space, we neglect this information, resulting in using a less accurate model in certain regions of the parameter space. However, we emphasise that method presented in this work uniquely incorporates this information.

Although not presented here (see Supplementary Material and Figures 6 and 7), we also analysed the SXS:BBH:0143 [13,52] and SXS:BBH:1156 [13,52] numerical relativity simulations produced by the SXS collaboration. SXS:BBH:0143 was chosen since it resides in a region of the parameter space where we expect our method to give comparable results to the Standard and Evidence informed analyses. SXS:BBH:1156 was chosen since it has largely asymmetric mass components and lies in the extrapolation regime of our technique (see Methods for details). For the case of SXS:BBH:0143, we find largely overlapping posteriors between all 3 methods, with most one-dimensional marginalized 90% confidence intervals containing the true value. Our analysis uses SEOBNRv5PHM 80% of the time, IMRPhenomTPHM 15% of the time and IMRPhenomXPHM 5% of the time. This represents the worse case scenario: our method performs the same as existing techniques. For the case of SXS:BBH:1156, we find that our method out performs the Standard and Evidence informed analyses despite being in the extrapolation regime of our technique: we more accurately capture the true parameters of the binary. Similar to SXS:BBH:0926, the Evidence informed analysis preferred IMRPhenomTPHM owing to the larger Bayesian evidence, while our analysis preferred SEOBNRv5PHM since it is the more accurate model in this region of the parameter space. Our analysis used SEOBNRv5PHM 78%, IMRPhenomTPHM 9% and IMRPhenomXPHM 13% of the time. [JT: Why not have the 1D figures for 0926?] SA: This is so funny. I had asked why not present 143 and 1156 as 2D plots

3 Conclusions

In this work we present a method which incorporates model uncertainty into gravitational-wave Bayesian inference for the first time. We apply this method to theoretical GW signals expected from general relativity and show that our method (i) marginalizes over model uncertainty by prioritising the most accurate model in each region of the parameter space, and (ii) outperforms widely used techniques that use Bayesian model averaging. The method presented in this work is independent of the models chosen and can, in principle, be used with any combination. GW models are also continuously being developed, and will likely improve in accuracy across the parameter space. Once available, these more accurate models can be incorporated into this method. The method presented here is applicable to ground-based GW parameter estimation analyses and we highly encourage its use in the future.

4 Methods

4.1 Multi-model Bayesian inference: The parameters of a binary are inferred from a gravitational-wave signal through Bayesian inference. Here, the model dependent posterior distribution for parameters $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, ..., \lambda_j\}$ is obtained through Bayes' theorem,

$$p(\boldsymbol{\lambda}|d,\mathfrak{M}_i) = \frac{\Pi(\boldsymbol{\lambda}|\mathfrak{M}_i)\,\mathcal{L}(d|\boldsymbol{\lambda},\mathfrak{M}_i)}{\mathcal{Z}}, \tag{1}$$

where $\Pi(\boldsymbol{\lambda}|\mathfrak{M}_i)$ is the probability of the parameters $\boldsymbol{\lambda}$ given the model $\mathfrak{M}_i$, otherwise known as the prior, $\mathcal{L}(d|\boldsymbol{\lambda},\mathfrak{M}_i)$ is the probability of observing the data given the parameters $\boldsymbol{\lambda}$ and model $\mathfrak{M}_i$, otherwise known as the likelihood, and $\mathcal{Z}$ is the probability of observing the data given the model $\mathcal{Z} = \int \Pi(\boldsymbol{\lambda}|\mathfrak{M}_i)\,\mathcal{L}(d|\boldsymbol{\lambda},\mathfrak{M}_i)\,d\boldsymbol{\lambda}$, otherwise known as the evidence. It is often not possible to trivially evaluate the model dependent posterior distribution; the challenge is evaluating the evidence since it involves computing the likelihood times prior for all points in the parameter space. Thankfully, nested sampling was developed to estimate the evidence through stochastic sampling, and return the model dependent posterior distribution as a by-product [71]. Here, a set of *live points* are randomly drawn from the prior, and the point with the lowest likelihood is stored and replaced with another point randomly drawn from the prior; the new point is accepted through the the Metropolis-Hastings algorithm [72]. This iterative process continues until the highest likelihood region(s) is identified.

When there are multiple models, Bayesian model averaging can be used to marginalize over the model uncertainty,

$$\begin{aligned} p(\boldsymbol{\lambda}|d) &= \sum_{i=1}^{N} p(\boldsymbol{\lambda}|d,\mathfrak{M}_i)\,p(\mathfrak{M}_i|d) \\ &= \sum_{i=1}^{N}\left[\frac{\mathcal{Z}_i\,\Pi(\mathfrak{M}_i)\,p(\boldsymbol{\lambda}|d,\mathfrak{M}_i)}{\sum_{j=1}^{N}\mathcal{Z}_j\,\Pi(\mathfrak{M}_j)}\right], \end{aligned} \tag{2}$$

where $p(\mathfrak{M}_i|d)$ is the probability of the model $\mathfrak{M}_i$ given the data, and $\Pi(\mathfrak{M}_i)$ is the discrete prior probability for the choice of model. For the case of uniform priors for the model, i.e., $\Pi(\mathfrak{M}_i) = 1/N$, Bayesian model averaging simply averages the model-dependent posterior distributions, weighted by the evidence.

An alternative solution to marginalizing over model uncertainty is to simultaneously infer the model and model properties in a single *joint* analysis [46]. Here, the parameter set $\boldsymbol{\lambda}$ is expanded to include the model m: $\tilde{\boldsymbol{\lambda}} = \{\lambda_1, \lambda_2, ..., \lambda_j, m\}$, and a discrete set of models can be sampled over during standard Bayesian inference analyses: for each step in, e.g., a nested sampling algorithm, an N-dimensional vector of model parameters is drawn from the prior, including an integer for the model, m. The integer m is mapped to a gravitational-wave model, and the likelihood is evaluated by passing the remaining model parameters, and the selected model, to the standard gravitational-wave likelihood [1]. It was demonstrated that this joint analysis will be at most N$\times$ faster to compute compared to performing Bayesian model averaging [46].

For the case of gravitational-wave astronomy, defining a discrete prior probability for the model is challenging since the accuracy of each model varies across the parameter space $\boldsymbol{\lambda}$ [5]. This makes it difficult to perform Bayesian model averaging; a uniform prior probability is often assumed for the choice of model [44,73] or in some cases, the model accuracy is averaged over the parameter space of interest [46]. However, a parameter-space dependent prior for the choice of model may solve this problem [46]. For instance, an N-dimensional vector of model parameters can be drawn from the prior, $\Pi(\mathfrak{M}_i|\boldsymbol{\lambda})$ can be evaluated for all models, i.e., the prior probability of the model given the parameter set $\boldsymbol{\lambda}$. The most probable model can then be determined, and the gravitational-wave likelihood subsequently evaluated. Although other priors have been suggested [46], we use the following model prior conditional on the parameters $\boldsymbol{\lambda}$,

$$\Pi(\mathfrak{M}_i|\boldsymbol{\lambda}) = \frac{\mathcal{M}_i(\boldsymbol{\lambda})^{-4}}{\sum_j \mathcal{M}_j(\boldsymbol{\lambda})^{-4}} \tag{3}$$

where $\mathcal{M}(\boldsymbol{\lambda})$ is the mismatch [42] between the model $\mathfrak{M}_i$ and a numerical relativity simulation with parameters $\boldsymbol{\lambda}$. Equation (3) implies that the most accurate GW model will more likely be used to evaluate the likelihood in each region of the parameter space. Of course, owing to computational limitations, we do not have numerical relativity simulations for all possible regions of the parameter space. For this reason, we compare the model $\mathfrak{M}_i$ against numerical relativity waveform surrogate models [40,41] as a proxy for full numerical relativity simulations.

4.2 Estimating waveform accuracy: As discussed in Section 1, the accuracy of a theoretical GW model is often assessed by comparing the signals produced by the model against numerical relativity simulations. We introduce a noise-weighted inner product between the model representation of a signal and the signal itself [42],

$$\langle h_{\rm m}|h_{\rm s}\rangle = 4\Re\int_{f_{\rm min}}^{f_{\rm max}} \frac{\tilde{h}^*_{\rm m}\,\tilde{h}_{\rm s}}{S_n(f)}\,{\rm d}f, \tag{4}$$

where a tilde denotes a Fourier transform, * denotes complex conjugation, and $S_n(f)$ is the noise power spectral density. The mismatch between two signals is computed by optimizing the normalised inner product o ver a set of (intrinsic or extrinsic) model parameters $\lambda_{\rm m}$,

$$\mathcal{M} = 1 - \max_{\lambda_{\rm m}} \frac{\langle h_{\rm m}|h_{\rm s}\rangle}{\sqrt{\langle h_{\rm m}|h_{\rm m}\rangle\langle h_{\rm s}|h_{\rm s}\rangle}}. \tag{5}$$

The intrinsic parameter space for a generic compact binary system is comprised of two masses, $m_{1,2}$, and two spin vectors, $\mathbf{S}_{1,2}$, adding up to eight degrees of freedom. Additionally, a quasi-circular binary comes with seven more extrinsic parameters: the right ascension, declination and the luminosity distance $\{\alpha, \delta, d_L\}$, respectively, to the binary's center of mass; the inclination of the orbit and its relative polarization $\{\iota, \psi\}$; and overall constant time phase shifts $\{t_{\rm c}, \varphi_{\rm c}\}$ of the GW.

For binaries where the spins of the compact bodies are aligned with the system's orbital angular momentum, several of the binary parameters become constant in time, and the intrinsic and extrinsic parameters decouple, thus reducing the dimensionality of the model space to four degrees of freedom: $\{m_1, m_2, \mathbf{S}_{1z}, \mathbf{S}_{2z}\}$, where the individual components of the spin vectors are specified at any fixed frequency. To compute the matches for the aligned-spin configurations that follow, we hold all extrinsic parameters fixed and optimize the match over the set of model parameters $\lambda_{\rm m} = \{t_{\rm c}, \phi_{\rm c}\}$. We maximize over $t_{\rm c}$ via an inverse fast Fourier transform and over $\phi_{\rm c}$ using the Nelder-Mead optimization algorithm found in SciPy's `minimize` function.

For binary systems in which the spins contain non-zero components orthogonal to the orbital angular momentum, the intrinsic and extrinsic parameters couple and evolve in time. Our aim is to isolate the intrinsic parameter space, accordingly the mismatch to which we intend to fit should somehow be independent of the extrinsic parameters. For this purpose, we first map $\{\alpha, \delta, \psi\}$ into a single parameter known as the effective polarizability κ [74]. We then prepare an evenly-spaced signal grid over the $\{\kappa, \varphi_{\rm c}, \iota\}_{\rm s} \in [0, \pi/2] \otimes [0, 2\pi) \otimes [0, \pi]$ space with $7 \times 6 \times 7 = 294$ elements. At each point on this signal grid, we compute the sky-optimized mismatch [5,38,74] between the signal and the model template from Eqn. (5), where the parameter set we optimize over is $\lambda_{\rm m} = \{t_{\rm c}, \varphi_{\rm c}, \kappa, \varphi_{\rm spin}\}$. Here $\varphi_{\rm spin}$ represents the freedom to rotate the in-plane spin angles ϕ_1, ϕ_2 by a constant amount. κ is optimized over analytically and optimizations over $\varphi_{\rm c}$ and $\varphi_{\rm spin}$ are performed numerically using dual annealing algorithms [3,34,38]. With these optimizations, we arrive at the maximum possible match between the template and the signal at a given point $\{\kappa, \varphi_c, \iota\}_{\rm s}$ in the signal grid. We repeat this procedure at every point of the 294-element grid then

compute the mean of this set as our final result for the mismatch

$$\mathcal{M}_{\rm av} := \frac{1}{294}\sum_{\rm s=1}^{294}\mathcal{M}(\kappa_{\rm s},\varphi_{\rm c,s},\iota_{\rm s}). \quad (6)$$

This is done to marginalize over any dependence of the mismatch on the sky position and inclination, thus obtain values which depend exclusively on the intrinsic parameters of the source. We additionally retain the standard deviation σ of the 294-mismatch set and use this as our error bar when needed. Note that our mean match, $1-\mathcal{M}_{\rm av}$, is a discretely averaged version of the sky-and-polarization averaged faithfulness given by Equation (35) of Ref. [2].

4.3 Constructing a match interpolant: Mismatch computations are fast, taking $O(\text{ms})$ per evaluation, for simplified models of the GW signal, such as those for aligned-spin configurations with only dominant quadrupolar emission. With increased model complexity, the computation can take a significantly longer time to evaluate, and producing the mismatch $\mathcal{M}(\boldsymbol{\lambda})$ will be a limiting cost in a Bayesian analysis since the likelihood is evaluated $O(10^8)$ times during a typical nested sampling analysis. For this reason, we construct an interpolant for the mismatch across the parameter space, $\mathcal{M}(\boldsymbol{\lambda})$, based on a discrete set of K mismatches. We describe how we construct an interpolant for binaries with spins aligned with the orbital angular momentum in Section 4.4. We further test this interpolant by comparing the posterior samples obtained from a Bayesian inference analyses that is guided by an actual mismatch computation at every step vs. a Bayesian inference analyses guided by the interpolant. In Section 4.5 we describe how we generalise this to build a generic spin interpolant. Due to computational cost, we use the Bayesian inference verification analysis in Section 4.4 to justify using an interpolant-guided analysis for systems with generic spins.

4.4 Interpolant for aligned-spin waveform mismatches: We begin with a test of the method using aligned-spin gravitational wave models containing higher signal multipoles. We evaluate mismatches using the numerical relativity hybrid surrogate model NRHYBSUR3DQ8 [41] as a proxy for numerical relativity simulations, and compare against the IMRPHENOMXHM [33] and IMRPHENOMTHM [36] waveform models, two of the leading frequency and time-domain models available for aligned-spin binaries, respectively. We do not use the state of the art EOB models [2,39] for this proof-of-principle test because IMRPHENOM models are one to two orders of magnitude faster to evaluate.

To simplify the construction of the mismatch interpolant for this test application, we reduce the dimensionality of the mismatch parameterization by artificially fixing several signal and model parameters. We choose to fix the total mass of the binary to $M = 90M_\odot$ and the inclination angle to $\theta_{\rm JN} = \pi/3$,where $\theta_{\rm JN}$ spans the angle between the line of sight and the total angular momentum vectors. This choice leaves three remaining free parameters in each model: the mass ratio $q = m_2/m_1 \leq 1$ and the component spins of the primary and secondary masses aligned with the orbital angular momentum, χ_1 and χ_2, respectively, defined from $\chi_i = \mathbf{S}_{iz}/m_i^2$ for $i = 1, 2$ with $-1 \leq \chi_i \leq 1$.

The 3-dimensional mismatch interpolants are constructed from mismatches computed on a uniform grid of 8 points in $0.125 \leq q \leq 1$ and 17 points in each $-0.8 \leq \chi_{1,2} \leq 0.8$, providing 2312 total mismatch points for each model. The interpolants are produced as polynomial fits of the form

$$\mathcal{M}(q,\chi_1,\chi_2) = \sum_{\substack{0\leq a\leq 6\\ 0\leq b,c\leq 8}} f_{abc}\, q^a \chi_1^b \chi_2^c, \quad (7)$$

[JT: the interpolant was built to test its use in the sampler, so I didn't care much for how it was built. I just chose a number of points in each dimension less than the number of eval points. We can run the notebook again to get the statistics from `Fit` if we want, but the aligned-spin fit wasn't supposed to be used for analysis (at least not when I made it).] SA: All of this are well known to us, but I wonder if the text in this section emphasizes this point well enough computing the fit coefficients f_{abc} using MATHEMATICA's `Fit` function and exported to PYTHON using `FortranForm`.

Next, we validate that our interpolant gives indistinguishable results to computing the mismatch directly in a Bayesian inference analysis. We perform two Bayesian inference analyses, both with the `Dynesty` Nested sampling software [63] via `Bilby` [64]. We use the same priors and sampler settings as those typically used in LIGO–Virgo–KAGRA analyses. The only distinguishing factor between these runs is that in one we use Equation (7) when computing the conditional probabilities of Equation (3), and in the other we directly compute the mismatch between the models and the surrogate at the sample point.

To compare posterior distributions we use the Jensen-Shannon Divergence [75] since it is commonly used in gravitational-wave astronomy [76,77]. The Jensen-Shannon Divergence ranges between 0, statistically identical distributions, and 1, statistically distinct distributions. A general rule of thumb is that a Jensen-Shannon Divergence $< 50\,\text{mbits}$ implies that the distributions are in good agreement [76].

In Table 1 we present Jensen-Shannon Divergences between marginalized posterior distributions obtained when a) calculating the mismatch exactly, and b) using the interpolant. We find that all Divergences are significantly less than $50\,\text{mbits}$, implying that the distributions are close to statistically identical. We find that the Bayesian analysis that used the interpolant completed in ~ 500 CPU hours; $\sim 250\times$ faster than the Bayesian analysis that computed the mismatch exactly. Given that the posteriors were close to statistically identical, and the reduced computational cost, we use the interpolated mismatch for all subsequent analyses.

Primary spin, χ_1/m_1	0.45	mbits
Secondary spin, χ_2/m_2	0.51	mbits
Mass ratio, $q = m_2/m_1$	0.89	mbits
Phase, ϕ	2.1	mbits
Polarization, Ψ	0.82	mbits
Right ascension α	0.99	mbits
Declination δ	1.3	mbits

Table 1 | Jensen-Shannon divergences between posteriors obtained when using true mismatch and the mismatch interpolant for an aligned-spin injection and recovery. These Jensen-Shannon divergences are reported in the base 2 logarithm and reported in millibits (mbits).

4.5 Interpolant for precessing waveform mismatches: When computing interpolants for the mismatches in Equation (6), we choose to fit for the $\log_{10}$ of the sky-averaged, optimized waveform mismatch (6). Accordingly, our error bars become $\sigma_{\log} := |\log_{10}(\mathcal{M}_{\rm av}-\sigma) - \log_{10}(\mathcal{M}_{\rm av}+\sigma)|$.

Next we generate a mismatch dataset to be used for fit construction (training). We could simply select values for the intrinsic parameters $\{m_1, m_2, \mathbf{S}_1, \mathbf{S}_2\}$ and obtain $\mathcal{M}$ via the procedure above, but we find that the brute force use of analytic functions of eight variables to fit to this data set is not the best approach. Instead, we opt to first reduce the dimensionality of the parameter space and then employ functional fitting. Already in Appendix A of MacUiliam *et al.* [5] we had seen encouraging preliminary results of this approach. We also note that generating just a single data point for this mismatch set is computationally expensive because of the four-dimensional optimization over $\lambda_{\rm m} = \{t_{\rm c}, \phi_{\rm c}, \kappa, \varphi_{\rm spin}\}$ that needs to be repeated for every element of

the 294-term sum in Equation (6). For example, depending on mass ratio and total mass, the computation of the average mismatch (6) at a single point in the intrinsic parameter space takes approximately 2-3 CPU hours for IMRPHENOMXPHM, 4.5-11 CPU hours for IMRPHENOMTPHM, and 6 CPU hours for SEOBNRV5PHM.

We start by mapping $m_{1,2}$ to the total mass M and the symmetric mass ratio η via

$$M = m_1 + m_2, \quad \eta := \frac{m_1 m_2}{M^2} \tag{8}$$

with the former quoted in solar masses ($M_\odot$) here and the latter being bounded $0 < \eta \leq 1/4$. Alternatively, we could have worked with the chirp mass $M_c := (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5}$ instead of M, but we opt to work with the total mass as its impact on the mismatch, Eqns. (5, 6) has been well documented[2,5,32,37,38].

The Cartesian components of each spin vector may be written in terms of spherical coordinates with respect to some reference frame, usually taken to be the orbital angular momentum vector at a reference frequency. Thus, we may write $\mathbf{S}_i = |\mathbf{S}_i|(\sin\theta_i \cos\phi_i, \sin\theta_i \sin\phi_i, \cos\theta_i)^T$ for $i = 1, 2$.

We reduce the dimensionality of this eight-dimensional intrinsic parameter space by mapping the six-dimensional spin space to two effective spins that we here label as x and y, representing the effective spin projections perpendicular and parallel to the reference orbital angular momentum vector of the binary, respectively. Two logical candidates for $\{x, y\}$ already exist: $\{\chi_{\rm p}, \chi_{\rm eff}\}$. The former is given by [28,78]

$$\chi_{\rm p} = \max\left(\bar{S}_1 \sin\theta_1, q\,\frac{4q+3}{4+3q}\,\bar{S}_2 \sin\theta_2\right) \tag{9}$$

with the bounds $0 \leq \chi_{\rm p} \leq 1$, and we have introduced $\bar{S}_{1,2} = |\mathbf{S}_{1,2}|/m_{1,2}^2$. A non-zero value for this quantity is an indication of spin precession, with $\chi_{\rm p} = 1$ corresponding to a maximally precessing binary, *i.e.*, all component spins of the binary constituents lie in the orbital plane and take their maximum magnitudes.

$\chi_{\rm eff}$ is the parallel projection counterpart to $\chi_{\rm p}$. It reads [28,79–81]

$$\chi_{\rm eff} = \frac{1}{1+q}(\chi_1 + q\chi_2)\,. \tag{10}$$

This is a conserved quantity up to 1.5 post-Newtonian (PN) order[80] and its magnitude changes very little over the course of an inspiral, making it very useful for inferring spin information about a compact binary system. It is clear from Equation (10) that $-1 \leq \chi_{\rm eff} \leq 1$ given the Kerr spin limit $|\chi_{1,2}| \leq 1$.

Other perpendicular projections exist in the literature [82,83], but the one which we empirically determined to be the best for fitting is $\chi_\perp$ given by[84]

$$\chi_\perp = \frac{|\mathbf{S}_{1,\perp} + \mathbf{S}_{2,\perp}|}{M^2}, \tag{11}$$

where $\mathbf{S}_{i,\perp} = \bar{S}_i m_i^2(\sin\theta_i \cos\phi_i, \sin\theta_i \sin\phi_i, 0)^T$ for $i = 1, 2$. We also experimented with a generalized version of $\chi_{\rm p}$[85], but found this quantity to be not as well suited for fitting as $\chi_{\rm p}$ or $\chi_\perp$. Given that the mismatches will be maximized over the in-plane spin angle $\varphi_{\rm spin}$, we map $\phi_{1,2}$ to a single azimuthal spin angle $\Delta\phi = \phi_2 - \phi_1$ by rotating our source frame axes such that $\phi_1 = 0$.

Finally, as an alternative to $\chi_{\rm eff}$, we introduce

$$\chi_\parallel := \frac{|\mathbf{S}_{1,\parallel} + \mathbf{S}_{2,\parallel}|}{M^2} = \frac{1}{(1+q)^2}(\chi_1 + q^2\chi_2)\,. \tag{12}$$

We thus have several choices for each perpendicular/parallel scalar: $x = \chi_{\rm p}$ or $\chi_\perp$, $y = \chi_{\rm eff}$ or $\chi_\parallel$, yielding four possible pairings for the dimensional reduction of the spin space. Our preliminary work based on gauging the faithfulness of the fits has, however, compelled us to drop $\chi_{\rm p}$ as it produced less faithful results, partly due to the fact that it does not carry any information about the planar spin angle separation $\Delta\phi$. Thus, we are left with two possible pairings for the reduced spin space: $\{\chi_\perp, \chi_{\rm eff}\}$ and $\{\chi_\perp, \chi_\parallel\}$. Accordingly, we introduce the fit training-set labels $K_1 = \{\chi_\perp, \chi_{\rm eff}, \eta, M\}$ and $K_2 = \{\chi_\perp, \chi_\parallel, \eta, M\}$.

The spin parameters introduced above depend on q, accordingly our fitting variables $\{x, y, \eta\}$ do not form a linearly independent three-dimensional subspace. Our motivation for choosing the particular fit variables above was ultimately empirical: our initial fits, using projections of spins with no q dependence, were less faithful to the data. It seems that mass-ratio dependent spin projections retain more useful information when the dimensionality of the parameter space is reduced. Additionally, we find that $\{x, y, \eta\}$ are either not correlated or weakly correlated for which we present correlation coefficients at the end of this section.

Next, we introduce a discrete parameter grid over the chosen four-dimensional $\{x, y, \eta, M\}$ space that we use for fitting. We limit η to range from $\eta = 0.16$ (corresponding to $q = 1/4$) to $\eta = 0.25$ ($q = 1$) in four even steps, resulting in five distinct values η_j, $j = 1, \ldots 5$. For the total mass, we employ $M = \{75, 117.5, 150\}M_\odot$ as our grid points, chosen because (i) NRSUR7DQ4 has been trained with data from binaries with only $q \geq 1/4$, (ii) NRSUR7DQ4's time length limit of $4300M$ [40] imposes $M \gtrsim 75M_\odot$ in order for the binary to enter the detector bandwidth at a GW frequency of 20 Hz, (iii) binaries with $M > 150M_\odot$ mostly emit merger-ringdown signals in the detection band [5], thus leaving hardly any imprint of precession in the reconstructed waveform from detector data, (iv) model mismatches tend to weakly depend on the total mass [2,5,32,37,38], thus sufficing three grid points in mass space for our current purposes given the computational burden of generating new data.

For better fit performance, the remaining two fit parameters, x and y, should also be placed on a regular grid. However, the quantities that we picked to cover this space, namely the pairings $\{\chi_\perp, \chi_{\rm eff}\}$ and $\{\chi_\perp, \chi_\parallel\}$ are not intrinsic parameters of the binary system. In order to construct a regular grid in $\{x, y\}$, we therefore start from first a regular grid of roughly 50,000 elements in $\{\bar{S}_1, \bar{S}_2, \theta_1, \theta_2, \Delta\phi\}$ space and use this to populate the $\{x, y\}$ space with values of q already determined by the η_j grid. The resulting grid in, e.g., the $\chi_\perp$-$\chi_{\rm eff}$ plane is scatter-plotted as the blue dots in the left panel of Figure 3, where we observe that the parameter space seems to be bounded by a half *prolate* ellipse drawn as the orange curve. The horizontal/vertical axes of the ellipse are given by

$$a = \max(x), \quad b = \max(y). \tag{13}$$

Guided by this observation, we construct a regular, *elliptical* grid in $\{x, y\}$ space as follows. First, we introduce the elliptical coordinates A, Φ with oblate/prolate-ness parameter $\mu > 0$

$$x = A \sinh\mu \cos\Phi, \tag{14a}$$

$$y = A \cosh\mu \sin\Phi \tag{14b}$$

with $\Phi \in [0, 2\pi]$ and the usual parametrization

$$\frac{x^2}{A^2 \sinh^2\mu} + \frac{y^2}{A^2 \cosh^2\mu} = 1. \tag{15}$$

For an ellipse of fixed size, A and μ are obtained from the relations $A \sinh\mu = a, A\cosh\mu = b$. Note that in Equations (14a-15), we swapped $\cosh\mu$ and $\sinh\mu$ because, as we show below, our ellipses are prolate, i.e., $a < b$.

Here, we aim to create a grid based on "concentric" ellipses of the same aspect ratio starting with the outermost one (orange curve in the left panel of Figure 3). With A and μ fixed, we create an elliptical grid

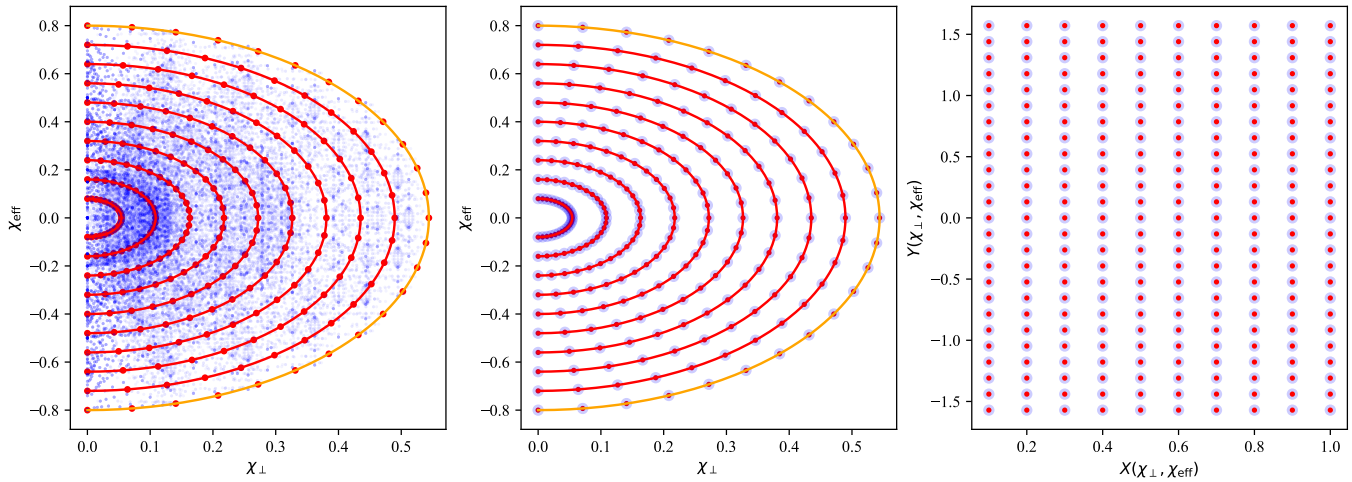

Figure 3 | Construction of the two dimensional elliptical and rectangular grids shown for the $\{x, y\} = \{\chi_\perp, \chi_{\rm eff}\}$ pairing. The blue dots in the left panel represent the $\sim$ 50000 points obtained from a regularly spaced grid in the intrinsic parameter space $\{q, \bar{S}_1, \bar{S}_2, \theta_1, \theta_2, \Delta\phi\}$. The orange curve is the outermost prolate ellipse whose axes are determined via Equations (13). The red curves are concentric ellipses retaining the same aspect ratio as the orange one. These are parametrized by r, s in Equations (16a, 16b). The red and orange dots situate the elements of the 10×25 elliptical grid where the azimuthal coordinate is sampled in steps of $\pi/24$ from $-\pi/2$ to $\pi/2$. The middle panel shows the same elliptical grid of the left panel overlaid, in faint blue, with the actual grid points that we determine by finding the roots of Equations (17a, 17b). On average, each blue dot is offset by 10^{-12} from the nearest, exactly positioned red dot. Finally, in the right panel, we show the corresponding rectangular grid. The red dots are exact, hence simply given by the coordinates $X = 0.1, 0.2, \ldots, 1, Y = -\pi/2, -\pi/2 + \pi/24, \ldots, \pi/2$ and the blue dots from the mapping of the blue dots of the middle panel via Equations (19a, 19b).

of our choosing via

$$x_{rs} = \frac{r}{N_r} A \sinh\mu \cos\left(\frac{\pi}{N_s}s - \frac{\pi}{2}\right), \tag{16a}$$

$$y_{rs} = \frac{r}{N_r} A \cosh\mu \sin\left(\frac{\pi}{N_s}s - \frac{\pi}{2}\right) \tag{16b}$$

with $r = 1, \ldots, N_r$ and $s = 0, \ldots N_s$. We show such a grid for $\{x, y\} = \{\chi_\perp, \chi_{\rm eff}\}$ in the left and middle panels of Figure 3 represented by the red dots with $N_r = 10, N_s = 24$, *i.e.*, a grid of $10 \times 25 = 250$ points. The grid over r builds concentric ellipses with the same aspect ratio and s angularly goes along each ellipse in steps of π/N_s.

The intrinsic parameters we seek should be chosen such that the corresponding values for $\{x, y\}$ yield points on the elliptical grid, *i.e.*, the red dots in the left panel of Figure 3. We start by finding the nearest point from the set of 50,000 points (blue dots in the left panel) to each grid point (red dot). For the $k^{\rm th}$ grid point with coordinates $\{x_k, y_k\}$, we find the nearest blue dot with coordinates $\{x_k^n, y_k^n\}$ generated from the intrinsic parameters $\{q^n, \bar{S}_1^n, \bar{S}_2^n, \theta_1^n, \theta_2^n, \Delta\phi^n\}$. We use these values as initial guesses in a rootfinding algorithm that translates to solving the following system

$$x_k - x(q, \bar{S}_1, \bar{S}_2, \theta_1, \theta_2, \Delta\phi) = 0, \tag{17a}$$

$$y_k - y(q, \bar{S}_1, \bar{S}_2, \theta_1, \theta_2) = 0 \tag{17b}$$

with the caveat that the used q values are consistent with our aforementioned η_j grid.

As this is numerical root finding, we replace the right hand sides of Equations (17a, 17b) with a threshold of 10^{-12}. We perform this root finding procedure for every single elliptical grid point. The end result is shown in the middle panel of Figure 3, where we place over each red dot a faint blue dot representing the grid points that our algorithm found. On average, each numerically determined grid point is offset by $\leq 10^{-12}$ from the exact grid (red) point. For a grid of 250 points, this amounts a total grid offset of $\lesssim 3 \times 10^{-9}$. We actually find this number to be 1.5×10^{-8} for the elliptical $\{\chi_\perp, \chi_{\rm eff}\}$ grid of Figure 3 because we had to relax our strict tolerance from 10^{-12} to 10^{-10} for certain grid points to speed up the procedure. [JT: kinda funny that we write the text as if people could distinguish 10^{-8} discrepancies in the points. I doubt the DPI of the plot could even show 10^{-3} levels of errors.] SA: Call it artistic freedom then? I am not disagreeing with you, perhaps Jake can make a version without the dots and we'll see As we show further below, a grid offset of $\sim 10^{-8}$ is much smaller than the average fit faithfulness that we obtain, $\lesssim 10^{-2}$, thus completely acceptable.

We repeated the same procedure to also obtain an elliptical grid in the $\chi_\perp$-$\chi_\parallel$ plane. In the interest of expediency, we used a tolerance of 10^{-8} resulting in an overall grid offset of 5×10^{-6}. Let us add that a few of the intrinsic coordinates for the grid points exceed NRSUR7DQ4's training limit of $\bar{S}_i = 0.8$ for spin magnitudes, but only by ~ 0.01 which is not severe.

As is well known, rectangular domains are often best suited for constructing fits to data, and we go one step further to transform the elliptical coordinates into a rectangular ones via

$$x = X\, A \sinh\mu \cos(Y), \tag{18a}$$

$$y = X\, A \cosh\mu \sin(Y), \tag{18b}$$

where $X \in [0, 1]$ and $Y \in [-\pi/2, \pi/2]$. Correspondingly, we have the following inverse relations

$$X = \frac{1}{A}\, \text{csch}\mu\, \text{sech}\mu \sqrt{x^2 \cosh^2\mu + y^2 \sinh^2\mu}, \tag{19a}$$

$$Y = \tan^{-1}\left(\frac{y \tanh\mu}{x}\right). \tag{19b}$$

Comparing Equation (16a) with (18a), and (16b) with (18b) gives the $N_r \times N_s$ rectangular grid $\{X_r, Y_s\}$ with $r = 1, \ldots, N_r, s = 0, \ldots N_s$, which we show in the right panel of Figure 3. Overall, we have the following four dimensional grid for the fitting: $\{X_r, Y_s, \eta_j, M_k\}$ with $j = 1, \ldots, 5$ and $k = 1, 2, 3$. As a final step, we introduce the rescaled variables $Z = 4\eta, V = M/(75 M_\odot)$.

After much trial and error, we settled on the following fitting func-

tion

$$\mathcal{F}(X,Y,Z,V) = \sum_{i=0}^{n_i}\sum_{j=0}^{n_j} \frac{\sum_{k=0}^{n_k}\sum_{l=0}^{1} c_{ijkl}\, Z^k V^l}{\sum_{k=0}^{n_k}\sum_{l=2}^{3} |c_{ijkl}|\, Z^k V^{l-2}}\, X^i Y^j \,. \tag{20}$$

We chose this particular form to better curb the fit's extrapolation behavior in parts of the $\{Z, V\}$ (mass ratio, total mass) space outside of the training region $Z < 0.64$ ($\eta < 0.16$) and $V < 1 \cup V > 2$ corresponding to $M < 75M_\odot \cup M > 150M_\odot$. We use two dimensional polynomials in the $\{X, Y\}$ subspace of the fit training domain because, as a result of our elliptical grid design, only rare combinations of intrinsic parameters yield points just outside our outermost ellipse. The maximum values of $\{n_i, n_j, n_k\}$ in the triple summation of Equation (20) are chosen such that we have at most roughly the same number of fit parameters as the total number of grid points used in the $\{x, y, \eta\}$ subspace, which is $10 \times 25 = 250$. Note that in the denominator of Equation (20), we take the absolute value of the fit coefficients c_{ijkl} to ensure that there are no singularities. We also set $c_{ij02} = 1$, which is the leading term in the denominator, a choice standard for Padé type fits. Our general procedure is as follows:
(i) start with a large ensemble of intrinsic parameters $\{q_i, \mathbf{S}_{1,i}, \mathbf{S}_{2,i}\}$ for $i = 1, \ldots, \mathcal{O}(10^4)$.
(ii) Impose an elliptical grid of size $N = N_r \times (N_s + 1)$ with the grid coordinates given by Equations (16a) and (16b).
(iii) Determine the set of intrinsic parameters $\{q_I, \mathbf{S}_{1,I}, \mathbf{S}_{2,I}\}$ yielding this grid to some tolerance, *e.g.*, 10^{-12}.
(iv) Compute the mismatches $\mathcal{M}_{K,L}$ of L models to NRSUR7DQ4 for the set $\{M_K, q_K, \mathbf{S}_{1,K}, \mathbf{S}_{2,K}\}$ where $K = 3I$ for the three distinct values of M that we use.
(v) Transform to the rectangular grid $\{X_K, Y_K, Z_K, V_K\}$.
(vi) For each model L, perform the fitting to the set $\{X_K, Y_K, Z_K, V_K, \log_{10}\mathcal{M}_{K,L}\}$ using MATHEMATICA's `NonlinearModelFit` function and store the coefficients $c_{ijkl,L}$ of Equation (20).

Our routine picks as the final fit the one for which the values of $\{n_i, n_j, n_k\}$ of Equation (20) yield the lowest relative difference with respect to the data set. For this purpose, we compute two relative differences. The first is the average relative distance between the fit and the data

$$\Delta^{(1)}_{\text{rel}} := \frac{1}{3N}\left(\sum_{k=1}^{3N}\left|1 - \frac{\mathcal{F}(X_k, Y_k, Z_k, V_k)}{\log_{10}\mathcal{M}_k}\right|^2\right)^{1/2}, \tag{21}$$

and the second is the signed relative difference

$$\Delta^{(2)}_{\text{rel}} := \frac{1}{3N}\sum_{k=1}^{3N}\left(1 - \frac{\mathcal{F}(X_k, Y_k, Z_k, V_k)}{\log_{10}\mathcal{M}_k}\right). \tag{22}$$

We pick the values for $\{n_i, n_j, n_k\}$ that simultaneously minimize both of the above relative differences which we treat as the most important fit attribute here as the model mismatches to NR are the main driver of the model selection in our PE runs.

Once the fit "training" is complete via the above optimization of $\{n_i, n_j, n_k\}$, we check fit performance over an appropriate verfication set. In Figure 4, we show a contour plot of the faithfulness of the fit for the $\log_{10}$ of the NRSUR7DQ4-SEOBNRV5PHM mismatches to the verification data set. The contours represent the absolute value of the relative difference between the fit and the data. The fit is trained over the $y = \chi_\parallel$ set (black dots) which, by design, trace concentric prolate ellipses in the $\{x, y\} = \{\chi_\perp, \chi_\parallel\}$ plane. The white dots mark the $\{x, y\}$ coordinates of the verification data. From the figure, we see that in a large portion of the space, the relative difference is 0.05 or less. Note that this quantity is not $\Delta^{(2)}_{\text{rel}}$ applied to the verification set, but rather the absolute value of the summand in Equation (22). [JT: I would maybe move this paragraph above the previous one and cut most of the text from that paragraph. It's very long-winded and this paragraph effectively summarizes most of it anyway.] SA: Done, with a new sentence added at the top and a new sentence below The fits for the NRSUR7DQ4-IMRPHENOMTPHM and the NRSUR7DQ4-IMRPHENOMXPHM mismatches also yield similar level of agreement as do the fits trained by the $y = \chi_{\text{eff}}$ set.

Similar figures for the fits to the IMRPHENOMTPHM, IMRPHENOMXPHM mismatches show a comparable overall magnitude to Figure 4 as well as the $\{x, y\} = \{\chi_\perp, \chi_{\text{eff}}\}$ counterparts to these figures. We summarize these results and provide additional metrics for all the fits in Table 2 where we see that the average relative distance (21) between each fit and the corresponding verification data is always less than 4×10^{-3} and the average relative difference (22) has magnitude less than 0.02. Interestingly, we observe that $\Delta^{(2),\text{ver}}_{\text{rel}}$ is negative for most fits indicating that the fits are slightly overestimating the data. The value of $1 - \tilde{R}^2$ is $\lesssim 0.01$ for all our fits, where $\tilde{R}^2$ is the reduced R square and for most cases we observe $\chi^2/\text{DoF} \approx \mathcal{O}(1)$. The cases for which this quantity is about an order of magnitude lower stem from the fact that we overestimate our errors bars. Recall that these are actually the standard deviations of an ensemble of mismatches (over a grid of certain extrinsic parameters per a given set of intrinsic parameters) whose average we take to be our individual data points.

As an additional check of the fits, we investigate their behavior in the extrapolation region corresponding to $\eta < 0.16$ (i.e., $q < 1/4$), $\bar{S}_{1,2} > 0.8$ and $M < 75M_\odot \cup M > 150M_\odot$. Recall that we chose the particular form of Equation (20) for the fitting function to better control unwanted extrapolation behavior such as blow-ups common to polynomial fitting. Specifically, the Padé type dependence on η and M was adopted so that the fits would not produce any nonsensical results such as $\log_{10}\mathcal{M} > 0$ in regions of the $\{\eta, M\}$ space quite distant from the training (interpolation) regime. On the other hand, since the relevant 2D cut of the training region covers most of the $\{x, y\}$ space, polynomial extrapolation should not cause issues.

We illustrate all of this in Figure 5, where we plot the fit in Equation (20) to the $\log_{10}$ of the NRSUR7DQ4-SEOBNRV5PHM mismatches as a function of M, evaluated at various extrapolated values of $\{\eta, \bar{S}_1 = \bar{S}_2\}$. The blue ellipse in the $\chi_\perp$-χ_{eff} plane traces the values $\{\eta, \bar{S}_{1,2}\} = \{0.139, 0.85\}$ ($q = 1/5$) with other intrinsic parameters chosen accordingly. Similarly, the red ellipse traces $\{\eta, \bar{S}_{1,2}\} = \{0.122, 0.9\}$ ($q = 1/6$) set and the orange ellipse the edge of the fit training region with $\{\eta, \bar{S}_{1,2}\} = \{0.16, 0.8\}$ ($q = 1/4$), which was already shown in Figure 3. The blue, red and orange dots mark the positions of seven cases along each corresponding ellipse in angular steps of $\pi/6$. An inset pointing to each dot displays the plot of the fit from $M = 50M_\odot$ to $200M_\odot$, but at each separate elliptical coordinate, hence the blue, red, orange colored curves. The shaded gray region in each inset marks the training range of $M \in [75, 150]M_\odot$ for the fits, only actually relevant to the orange curves as the blue and the red are, by definition, outside the training region. Thanks to the specific functional form of the fit, the extrapolation does not exhibit any worrying pathologies. Additionally, we note that the blue curves mostly lay between the orange and the red ones as we would expect. We should, however, caution that we are merely demonstrating that the extrapolation is not pathological. This does not mean that the fits are expected be faithful to the data in this regime. As such we recommend their use in the regime $q \leq 5, \bar{S}_i \leq 0.85$.

As a further test of the fit's performance in its extrapolation regime, we performed yet another recovery of an NR injection, this time `SXS:BBH:1156`, with $M = 100M_\odot, q = 0.228$ which is outside the $q \geq 1/4$ training regime of our fits. It also has $|\mathbf{S}_{2,\perp}|/m_2^2 \approx 0.76$

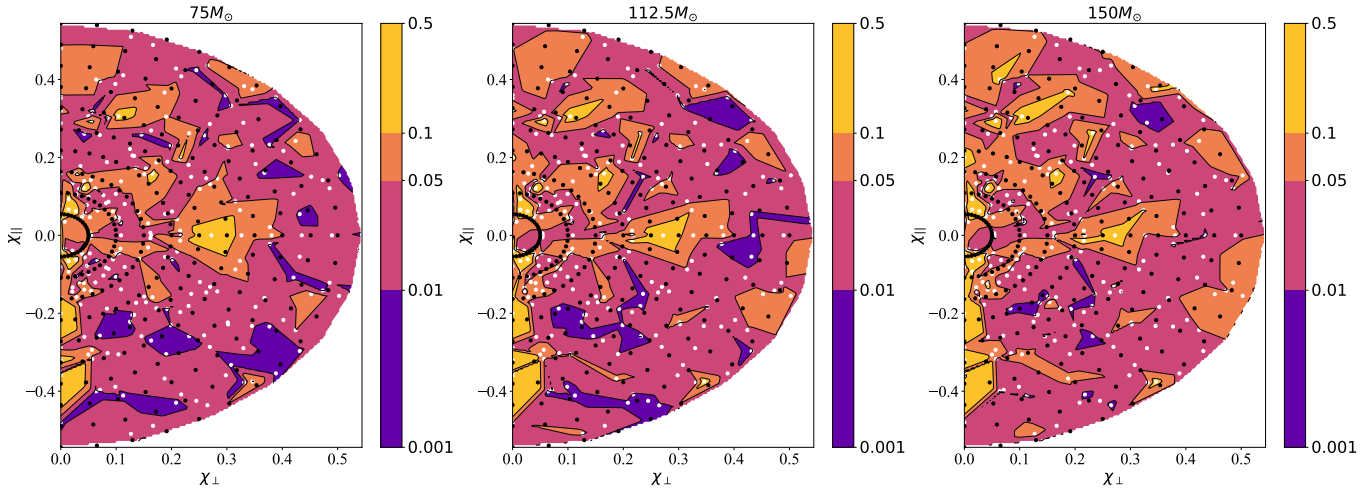

Figure 4 | A sample of fit faithfulness over the verification set. We plot the absolute value of the relative difference between the data for the $\log_{10}$ of the NRSUR7DQ4-SEOBNRV5PHM mismatches and the fit (20) to it with Equations (19a, 19b) substituted. The panels from left to right correspond to total masses of $75M_\odot, 112.5M_\odot, 150M_\odot$. $\chi_\perp$ and $\chi_\parallel$ are defined in Equations (11, 12), respectively. The white dots mark the verification set data points while the black dots mark the training set points. The color scale is logarithmic. Overall, we observe an absolute relative difference of less than 0.05 for most of the parameter space.

Model	y	$\{n_i, n_j, n_k\}$	# Params	$\Delta^{(1),\text{ver}}_{\text{rel}}$	$\Delta^{(2),\text{ver}}_{\text{rel}}$	$1-\tilde{R}^2$	χ^2/DoF
SEOBNRV5PHM	χ_{eff}	$\{5,4,3\}$	220	0.0032	-0.013	0.0049	0.050
SEOBNRV5PHM	$\chi_\parallel$	$\{6,4,3\}$	264	0.0026	0.0041	0.0052	1.3
IMRPHENOMTPHM	χ_{eff}	$\{6,4,3\}$	264	0.0039	-0.012	0.012	0.99
IMRPHENOMTPHM	$\chi_\parallel$	$\{5,4,3\}$	220	0.0039	-0.0090	0.0077	0.076
IMRPHENOMXPHM-ST	χ_{eff}	$\{5,4,3\}$	220	0.0038	-0.019	0.010	1.2
IMRPHENOMXPHM-ST	$\chi_\parallel$	$\{5,4,3\}$	220	0.0032	-0.0021	0.0078	0.74

Table 2 | Relevant metrics for the fits to the $\log_{10}$ of the mismatches between NRSUR7DQ4 and the models listed in the first column. The fitting function is given in Equation (20) with the transformation from the $\{X, Y\}$ coordinates to the elliptical $\{x, y\}$ done via Equations (19a, 19b). Column two lists our choice for y representing whether we used χ_{eff} or $\chi_\parallel$ when constructing the elliptical fit training grid, e.g., shown for χ_{eff} in Figure 3. Column three lists the upper limits of the triple summation used in Equation (20) which then determines the total number of fitting parameters c_{ijkl} displayed in column four. Columns five and six present the values for the two relative difference measures of Equations (21, 22) applied to the verification sets. Finally, column seven lists $1-\tilde{R}^2$, where $\tilde{R}^2$ is the reduced R square and column eight the chi square over the degrees of freedom. Overall, we see that the fits based on the $y = \chi_\parallel$ grid are slightly more faithful to the verification data and $\tilde{R}^2 \gtrsim 0.99$ for all fits.

while the primary has negligible planar spin. We show the results of our method applied to this PE run in Fig. 7 in the Supplementary Material, where we show that both m_1 and m_2 recovered well the injected values of $81.4M_\odot, 18.6M_\odot$, respectively. Such a good recovery of the masses, hence the mass ratio, is an indirect testament to the robustness of the fit (20) at least in regions slightly outside its mass ratio training regime. We additionally show that the primary spin is also well recovered.

We conclude this section by briefly returning to one issue regarding the fit construction already mentioned, that the fitting variables $\{\eta, x, y\}$ are not fully independent of each other, as each one is a function of the mass ratio q. However, as we explained already, the q-scaled spin projections yield more faithful fits to the data. As a check, we computed the correlation coefficients $\mathcal{C}_{mn}$ between the above parameters of the training sets $K_1(y = \chi_{\text{eff}})$, $K_2(y = \chi_\parallel)$. For K_1, we obtain $\mathcal{C}_{\chi_\perp\text{-}\chi_{\text{eff}}} = 0.09$, $\mathcal{C}_{\eta\text{-}\chi_{\text{eff}}} = 0.02$ and $\mathcal{C}_{\eta\text{-}\chi_\perp} = -0.315$. For K_2, we have $\mathcal{C}_{\chi_\perp\text{-}\chi_\parallel} = -0.1$, $\mathcal{C}_{\eta\text{-}\chi_\perp} = 0.03$ and $\mathcal{C}_{\eta\text{-}\chi_\parallel} = 0.216$. For both fit-training parameter sets, we have either uncorrelated pairings of fit variables or weakly correlated pairings. The fact that $|\mathcal{C}_{\eta\text{-}\chi_\parallel}|$ of set K_2 is less than $|\mathcal{C}_{\eta\text{-}\chi_\perp}|$ of set K_1 may partly explain why we observe a slightly better performance from the fits constructed from K_2 as indicated in Table 2.

1. Veitch, J. *et al.* Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library. *Phys. Rev.* **D91**, 042003 (2015). 1409.7215.
2. Ramos-Buades, A. *et al.* Next generation of accurate and efficient multipolar precessing-spin effective-one-body waveforms for binary black holes. *Phys. Rev. D* **108**, 124037 (2023). 2303.18046.
3. Thompson, J. E. *et al.* PhenomXO4a: a phenomenological gravitational-wave model for precessing black-hole binaries with higher multipoles and asymmetries. *Phys. Rev. D* **109**, 063012 (2024). 2312.10025.
4. Yelikar, A. B., Shaughnessy, R. O., Lange, J. & Jan, A. Z. Waveform systematics in gravitational-wave inference of signals from binary neutron star merger models incorporating higher order modes information (2024). 2404.16599.
5. Mac Uilliam, J., Akcay, S. & Thompson, J. E. Survey of four precessing waveform models for binary black hole systems. *Phys. Rev. D* **109**, 084077 (2024). 2402.06781.
6. Mac Uilliam, J., Akcay, S. & Hoy, C. Waging a Campaign: Results from an Injection/Recovery Study involving 30 numerical Relativity Simulations and three Waveform Models (2024). In preparation.
7. Dhani, A. *et al.* Systematic Biases in Estimating the Properties of Black Holes Due to Inaccurate Gravitational-Wave Models (2024). 2404.05811.
8. Pürrer, M. & Haster, C.-J. Gravitational waveform accuracy requirements for future ground-based detectors. *Phys. Rev. Res.* **2**, 023151 (2020). 1912.10055.
9. Moore, C. J., Finch, E., Buscicchio, R. & Gerosa, D. Testing general relativity with gravitational-wave catalogs: The insidious nature of waveform systematics. *iScience* **24**, 102577 (2021). URL https://www.sciencedirect.com/science/article/pii/S2589004221005459.

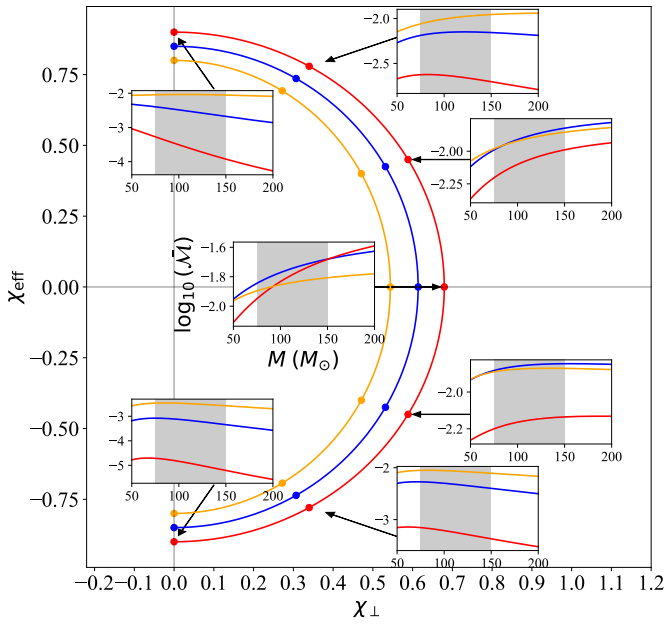

Figure 5 | The extrapolation of the fits. We show the performance of the fit to the $\log_{10}$ of the NRSur7dq4-SEOBNRv5PHM mismatches extrapolated to $\{q, \bar{S}_{1,2}\} = \{1/5, 0.85\}$ (blue) and $\{1/6, 0.9\}$ (red) as well as the fit at the edge of its training region with $\{q, \bar{S}_{1,2}\} = \{1/4, 0.8\}$ (orange). The red, blue, orange dots accordingly mark the positions of seven cases along each corresponding ellipse with the above $\{q, \bar{S}_{1,2}\}$ values and the remaining intrinsic parameters chosen such that the dots trace each ellipse in angular steps of $\pi/6$. The smaller orange ellipse was shown previously in Figure 3. Pointing to each Φ = constant dot is an inset showing the plot of the fit from $M = 50M_\odot$ to $200M_\odot$, but at each separate elliptical coordinate. The shaded gray region in each inset has been placed to remind the reader our training range of $M \in [75, 150]M_\odot$. From the insets, we see that the extrapolation does not appear to be pathological and the blue curves mostly lay between the red and orange ones as expected.

10. Kapil, V., Reali, L., Cotesta, R. & Berti, E. Systematic bias from waveform modeling for binary black hole populations in next-generation gravitational wave detectors. *Phys. Rev. D* **109**, 104043 (2024). 2404.00090.
11. Hamilton, E. *et al.* Catalog of precessing black-hole-binary numerical-relativity simulations. *Phys. Rev. D* **109**, 044032 (2024). 2303.05419.
12. Mroue, A. H. *et al.* A catalog of 174 binary black-hole simulations for gravitational-wave astronomy. *Phys.Rev.Lett.* **111**, 241104 (2013). 1304.6077.
13. Boyle, M. *et al.* The SXS Collaboration catalog of binary black hole simulations. *Class. Quant. Grav.* **36**, 195006 (2019). 1904.04831.
14. Healy, J., Lousto, C. O., Zlochower, Y. & Campanelli, M. The RIT binary black hole simulations catalog. *Class. Quant. Grav.* **34**, 224001 (2017). 1703.03423.
15. Healy, J. *et al.* Second RIT binary black hole simulations catalog and its application to gravitational waves parameter estimation. *Phys. Rev. D* **100**, 024021 (2019). 1901.02553.
16. Healy, J. & Lousto, C. O. Third RIT binary black hole simulations catalog. *Phys. Rev. D* **102**, 104018 (2020). 2007.07910.
17. Healy, J. & Lousto, C. O. Fourth RIT binary black hole simulations catalog: Extension to eccentric orbits. *Phys. Rev. D* **105**, 124010 (2022). 2202.00018.
18. Jani, K. *et al.* Georgia Tech Catalog of Gravitational Waveforms. *Class. Quant. Grav.* **33**, 204001 (2016). 1605.03204.
19. Bohé, A. *et al.* Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors. *Phys. Rev.* **D95**, 044028 (2017). 1611.03703.
20. Cotesta, R. *et al.* Enriching the Symphony of Gravitational Waves from Binary Black Holes by Tuning Higher Harmonics. *Phys. Rev.* **D98**, 084028 (2018). 1803.10701.
21. Cotesta, R., Marsat, S. & Pürrer, M. Frequency domain reduced order model of aligned-spin effective-one-body waveforms with higher-order modes. *Phys. Rev. D* **101**, 124040 (2020). 2003.12079.
22. Ossokine, S. *et al.* Multipolar Effective-One-Body Waveforms for Precessing Binary Black Holes: Construction and Validation. *Phys. Rev. D* **102**, 044055 (2020). 2004.09442.
23. Babak, S., Taracchini, A. & Buonanno, A. Validating the effective-one-body model of spinning, precessing binary black holes against numerical relativity. *Phys. Rev.* **D95**, 024010 (2017). 1607.05661.
24. Pan, Y. *et al.* Inspiral-merger-ringdown waveforms of spinning, precessing black-hole binaries in the effective-one-body formalism. *Phys.Rev.* **D89**, 084006 (2014). 1307.6232.
25. Husa, S. *et al.* Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal. *Phys. Rev.* **D93**, 044006 (2016). 1508.07250.
26. Khan, S. *et al.* Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era. *Phys. Rev.* **D93**, 044007 (2016). 1508.07253.
27. London, L. *et al.* First higher-multipole model of gravitational waves from spinning and coalescing black-hole binaries. *Phys. Rev. Lett.* **120**, 161102 (2018). 1708.00404.
28. Hannam, M. *et al.* Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms. *Phys. Rev. Lett.* **113**, 151101 (2014). 1308.3271.
29. Nagar, A. *et al.* Time-domain effective-one-body gravitational waveforms for coalescing compact binaries with nonprecessing spins, tides and self-spin effects. *Phys. Rev.* **D98**, 104052 (2018). 1806.01772.
30. Khan, S., Chatziioannou, K., Hannam, M. & Ohme, F. Phenomenological model for the gravitational-wave signal from precessing binary black holes with two-spin effects. *Phys. Rev.* **D100**, 024059 (2019). 1809.10113.
31. Khan, S., Ohme, F., Chatziioannou, K. & Hannam, M. Including higher order multipoles in gravitational-wave models for precessing binary black holes (2019). 1911.06050.
32. Pratten, G. *et al.* Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for nonprecessing quasicircular black holes. *Phys. Rev. D* **102**, 064001 (2020). 2001.11412.
33. García-Quirós, C. *et al.* Multimode frequency-domain model for the gravitational wave signal from nonprecessing black-hole binaries. *Phys. Rev. D* **102**, 064002 (2020). 2001.10914.
34. Pratten, G. *et al.* Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes. *Phys. Rev. D* **103**, 104056 (2021). 2004.06503.
35. Estellés, H. *et al.* Phenomenological time domain model for dominant quadrupole gravitational wave signal of coalescing binary black holes. *Phys. Rev. D* **103**, 124060 (2021). 2004.08302.
36. Estellés, H. *et al.* Time-domain phenomenological model of gravitational-wave subdominant harmonics for quasicircular nonprecessing binary black hole coalescences. *Phys. Rev. D* **105**, 084039 (2022). 2012.11923.
37. Estellés, H. *et al.* New twists in compact binary waveform modeling: A fast time-domain model for precession. *Phys. Rev. D* **105**, 084040 (2022). 2105.05872.
38. Gamba, R., Akçay, S., Bernuzzi, S. & Williams, J. Effective-one-body waveforms for precessing coalescing compact binaries with post-Newtonian twist. *Phys. Rev. D* **106**, 024020 (2022). 2111.03675.
39. Nagar, A. *et al.* Analytic systematics in next generation of effective-one-body gravitational waveform models for future observations. *Phys. Rev. D* **108**, 124018 (2023). 2304.09662.
40. Varma, V. *et al.* Surrogate models for precessing binary black hole simulations with unequal masses. *Phys. Rev. Research.* **1**, 033015 (2019). 1905.09300.
41. Varma, V. *et al.* Surrogate model of hybridized numerical relativity binary black hole waveforms. *Phys. Rev.* **D99**, 064045 (2019). 1812.07865.
42. Owen, B. J. Search templates for gravitational waves from inspiraling binaries: Choice of template spacing. *Phys. Rev.* **D53**, 6749–6761 (1996). gr-qc/9511032.
43. Abbott, R. *et al.* GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run (2021). 2111.03606.
44. Ashton, G. & Khan, S. Multiwaveform inference of gravitational waves. *Phys. Rev. D* **101**, 064037 (2020). 1910.09138.
45. Ashton, G. & Dietrich, T. The use of hypermodels to understand binary neutron star collisions. *Nature Astron.* **6**, 961–967 (2022). 2111.09214.
46. Hoy, C. Accelerating multimodel Bayesian inference, model selection, and systematic studies for gravitational wave astronomy. *Phys. Rev. D* **106**, 083003 (2022). 2208.00106.
47. Moore, C. J. & Gair, J. R. Novel Method for Incorporating Model Uncertainties into Gravitational Wave Parameter Estimates. *Phys. Rev. Lett.* **113**, 251101 (2014). 1412.3657.

48. Doctor, Z., Farr, B., Holz, D. E. & Pürrer, M. Statistical Gravitational Waveform Models: What to Simulate Next? *Phys. Rev.* **D96**, 123011 (2017). 1706.05408.
49. Williams, D., Heng, I. S., Gair, J., Clark, J. A. & Khamesra, B. A Precessing Numerical Relativity Waveform Surrogate Model for Binary Black Holes: A Gaussian Process Regression Approach (2019). 1903.09204.
50. Read, J. S. Waveform uncertainty quantification and interpretation for gravitational-wave astronomy. *Class. Quant. Grav.* **40**, 135002 (2023). 2301.06630.
51. Khan, S. Probabilistic model for the gravitational wave signal from merging black holes. *Phys. Rev. D* **109**, 104045 (2024). 2403.11534.
52. Blackman, J. *et al.* Numerical relativity waveform surrogate model for generically precessing binary black hole mergers. *Phys. Rev. D* **96**, 024058 (2017). 1705.07089.
53. Apostolatos, T. A., Cutler, C., Sussman, G. J. & Thorne, K. S. Spin induced orbital precession and its modulation of the gravitational wave forms from merging binaries. *Phys. Rev.* **D49**, 6274–6297 (1994).
54. Fairhurst, S., Green, R., Hoy, C., Hannam, M. & Muir, A. The two-harmonic approximation for gravitational waveforms from precessing binaries (2019). 1908.05707.
55. Fairhurst, S., Green, R., Hannam, M. & Hoy, C. When will we observe binary black holes precessing? *Phys. Rev. D* **102**, 041302 (2020). 1908.00555.
56. Aasi, J. *et al.* Advanced LIGO. *Class. Quant. Grav.* **32**, 074001 (2015). 1411.4547.
57. Acernese, F. *et al.* Advanced virgo: a second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity* **32**, 024001 (2014).
58. Akutsu, T. *et al.* Overview of KAGRA: Detector design and construction history. *PTEP* **2021**, 05A101 (2021). 2005.05574.
59. Hoy, C., Fairhurst, S. & Mandel, I. Precession and higher order multipoles in binary black holes (and lack thereof) (2024). 2408.03410.
60. Colleoni, M. *et al.* in preparation (2024).
61. Goldberg, J. N., MacFarlane, A. J., Newman, E. T., Rohrlich, F. & Sudarshan, E. C. G. Spin s spherical harmonics and edth. *J. Math. Phys.* **8**, 2155 (1967).
62. LIGO Scientific Collaboration and Virgo Collaboration. Noise curves used for simulations in the update of the observing scenarios paper. DCC (2022). URL https://dcc.ligo.org/LIGO-T2000012/public.
63. Speagle, J. S. dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society* **493**, 3132?3158 (2020). URL http://dx.doi.org/10.1093/mnras/staa278.
64. Ashton, G. *et al.* BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy. *Astrophys. J. Suppl.* **241**, 27 (2019). 1811.02042.
65. Abbott, R. *et al.* The population of merging compact binaries inferred using gravitational waves through GWTC-3 (2021). 2111.03634.
66. Pürrer, M., Hannam, M. & Ohme, F. Can we measure individual black-hole spins from gravitational-wave observations? *Phys. Rev. D* **93**, 084042 (2016). 1512.04955.
67. Ajith, P. *et al.* Inspiral-merger-ringdown waveforms for black-hole binaries with non-precessing spins. *Phys.Rev.Lett.* **106**, 241101 (2011). 0909.2867.
68. Ajith, P. Addressing the spin question in gravitational-wave searches: Waveform templates for inspiralling compact binaries with nonprecessing spins. *Phys.Rev.* **D84**, 084037 (2011). 1107.1267.
69. Baird, E., Fairhurst, S., Hannam, M. & Murphy, P. Degeneracy between mass and spin in black-hole-binary waveforms. *Phys. Rev. D* **87**, 024035 (2013). 1211.0546.
70. Toubiana, A. & Gair, J. R. Indistinguishability criterion and estimating the presence of biases (2024). 2401.06845.
71. Skilling, J. Nested sampling for general bayesian computation. *Bayesian Anal.* **1**, 833–859 (2006). URL https://doi.org/10.1214/06-BA127.
72. Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97–109 (1970).
73. Jan, A. Z., Yelikar, A. B., Lange, J. & O'Shaughnessy, R. Assessing and marginalizing over compact binary coalescence waveform systematics with RIFT. *Phys. Rev. D* **102**, 124069 (2020). 2011.03571.
74. Harry, I., Privitera, S., Bohé, A. & Buonanno, A. Searching for Gravitational Waves from Compact Binaries with Precessing Spins. *Phys. Rev.* **D94**, 024012 (2016). 1603.02444.
75. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Info. Theor.* **37**, 145–151 (1991).
76. Abbott, B. P. *et al.* GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs. *Phys. Rev.* **X9**, 031040 (2019). 1811.12907.
77. Abbott, R. *et al.* GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. *Phys. Rev. X* **11**, 021053 (2021). 2010.14527.
78. Schmidt, P., Ohme, F. & Hannam, M. Towards models of gravitational waveforms from generic binaries II: Modelling precession effects with a single effective precession parameter. *Phys. Rev.* **D91**, 024043 (2015). 1408.1810.
79. Damour, T. Coalescence of two spinning black holes: An effective one-body approach. *Phys. Rev.* **D64**, 124013 (2001). gr-qc/0103018.
80. Racine, E. Analysis of spin precession in binary black hole systems including quadrupole-monopole interaction. *Phys. Rev.* **D78**, 044021 (2008). 0803.1820.
81. Pürrer, M., Hannam, M., Ajith, P. & Husa, S. Testing the validity of the single-spin approximation in inspiral-merger-ringdown waveforms. *Phys. Rev. D* **88**, 064007 (2013). 1306.2320.
82. Thomas, L. M., Schmidt, P. & Pratten, G. New effective precession spin for modeling multimodal gravitational waveforms in the strong-field regime. *Phys. Rev. D* **103**, 083022 (2021). 2012.02209.
83. Hamilton, E. *et al.* Model of gravitational waves from precessing black-hole binaries through merger and ringdown. *Phys. Rev. D* **104**, 124027 (2021). 2107.08876.
84. Akcay, S., Gamba, R. & Bernuzzi, S. A hybrid post-Newtonian – effective-one-body scheme for spin-precessing compact-binary waveforms. *Phys. Rev. D* **103**, 024014 (2021). 2005.05338.
85. Gerosa, D. *et al.* A generalized precession parameter $\chi_{\rm P}$ to interpret gravitational-wave data. *Phys. Rev. D* **103**, 064067 (2021). 2011.11948.
86. Hoy, C., Akcay, S., Mac Uilliam, J. & Thompson, J. E. Incorporating multi-model uncertainty into gravitational-wave Bayesian inference - public data release (2024). In preparation.

Supplementary Material

In this work we also analysed the SXS:BBH:0143 [13,52] and SXS:BBH:1156 [13,52] numerical relativity simulations produced by the SXS collaboration. For the case of SXS:BBH:0143, we found largely overlapping posteriors between methods, with most one-dimensional marginalized 90% confidence intervals containing the true value. Here, we provide further details.

In Figure 6 we show the results obtained with our method, *NR informed*, and the *Evidence informed* and *Standard* analyses for the analysis of the SXS:BBH:0143 numerical relativity simulation. We see that in general, all methods give largely overlapping posteriors. The biggest difference is seen in the inferred secondary spin magnitude of the binary: our method prefers a rapidly spinning black hole, while the other methods prefer a spin ~ 0.25. The reason is because our method is primarily using the SEOBNRv5PHM model, while the Evidence informed result is primarily using IMRPhenomTPHM. In this region of the parameter space we find that SEOBNRv5PHM is the most accurate model, hence why it is favoured in our method: we find that SEOBNRv5PHM is $\sim 1.4\times$ more accurate than IMRPhenomTPHM and $\sim 1.8\times$ more accurate than IMRPhenomXPHM. Given that the individual spin components are difficult to measure at present-day detector sensitivities [66], it is not surprising that we the biggest difference between methods is seen for the secondary spin.

Data Availability

The aligned- and generic-spin match interpolants, the posterior samples from the analyses performed in this work, and our modifications to Bilby [64] are available in the data release [86].

Acknowledgements

We would like to thank (THIS COULD BE YOU!) for valuable comments on this manuscript. We are also grateful to Mark Hannam, Laura Nuttall and (SOMEONE ELSE WE COULDN'T THINK OF) for discussions throughout this project. We thank Alvin Chua and the California Institute of Technology for hosting CH and SA in March 2024, giving the authors time to discuss and develop this project. CH thanks the UKRI Future Leaders Fellowship for support through the

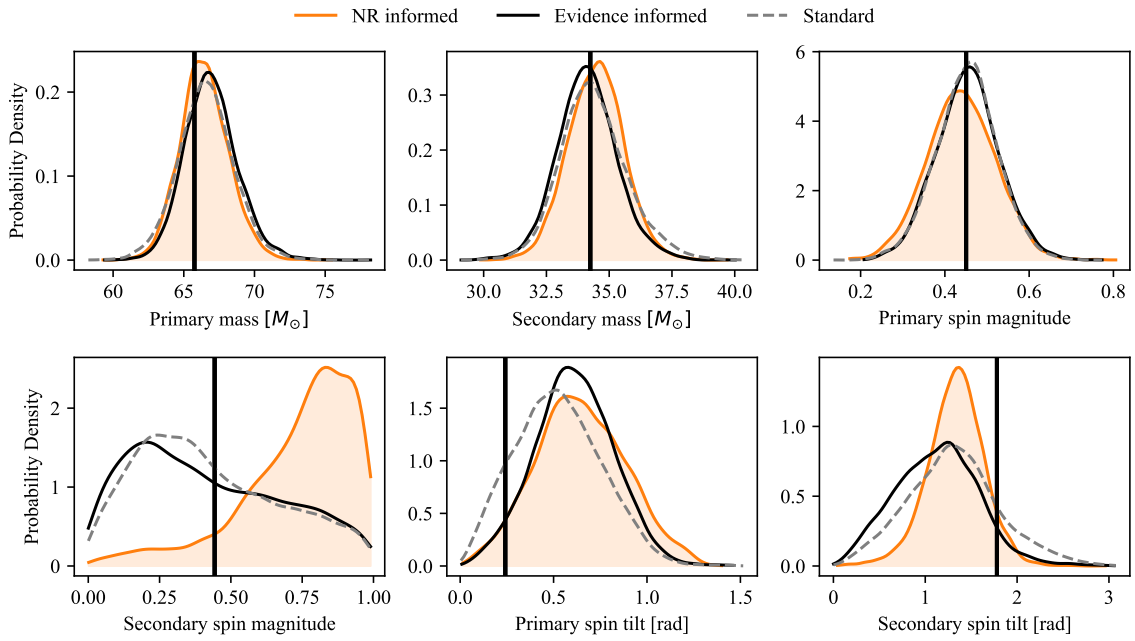

Figure 6 | One-dimensional posterior probabilities obtained in our analysis of the `SXS:BBH:0143` numerical relativity simulation. We show the measurement of binary component masses, as well as the individual spins (decomposed by the magnitude and tilt angle). The black vertical line indicates the true value.

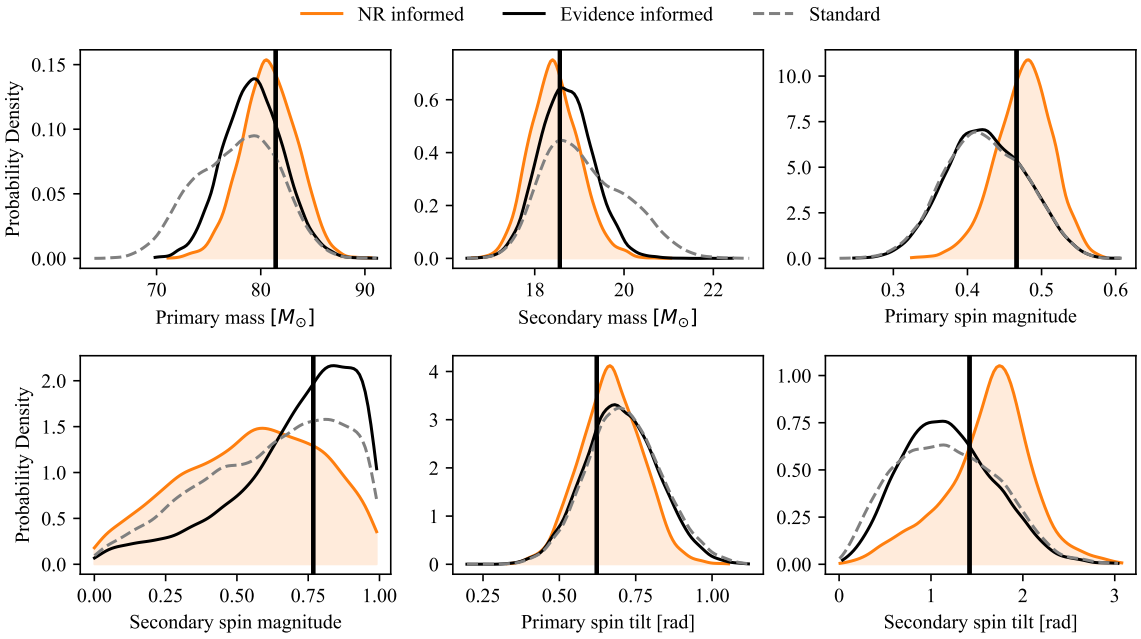

Figure 7 | One-dimensional posterior probabilities obtained in our analysis of the `SXS:BBH:1156` numerical relativity simulation. We show the measurement of binary component masses, as well as the individual spins (decomposed by the magnitude and tilt angle). The black vertical line indicates the true value.

grant MR/T01881X/1, SA and JMU acknowledge support from the University College Dublin Ad Astra Fellowship, and JT acknowledges support from the NASA LISA Preparatory Science grant 20-LPS20-0005. This work used the computational resources provided by the ICG, SEPNet and the University of Portsmouth, supported by STFC grant ST/N000064.

Author contributions

CH conceptualised the idea of sampling over multiple models, with priors dictated by their mismatch to numerical relativity simulations, as a method to incorporate model uncertainty into GW bayesian inference. SA and JT developed the idea of building mismatch interpolants for this application. SA and JT initiated and formed the project team. CH implemented the method presented here into `Bilby` [64] and performed all parameter estimation analyses. JT generated the aligned-

spin mismatches and the aligned-spin interpolant. CH and JT investigated choices for the model conditional prior. JMU generated the majority of generic-spin mismatches, with CH and SA generating a subset. SA produced the generic-spin interpolant. All authors contributed to the interpretation of the results and wrote the paper.

Competing interests

The authors declare that they have no competing financial interests.